Chapter 9

From Dependency to Causality: A Machine Learning Approach

Gianluca Bontempi, Maxime Flauder

Abstract The relationship between statistical dependency and causality lies at the heart of all statistical approaches to causal inference. Recent results in the ChaLearn cause-effect pair challenge have shown that causal directionality can be inferred with good accuracy also in Markov indistinguishable configurations thanks to data driven approaches. This paper proposes a supervised machine learning approach to infer the existence of a directed causal link between two variables in multivariate settings with $n > 2$ variables. The approach relies on the asymmetry of some conditional (in)dependence relations between the members of the Markov blankets of two variables causally connected. Our results show that supervised learning methods may be successfully used to extract causal information on the basis of asymmetric statistical descriptors also for $n > 2$ variate distributions.

Key words: causal inference, information theory, machine learning

Gianluca Bontempi
Machine Learning Group, Computer Science Department, Interuniversity Institute of Bioinformatics in Brussels (IB), ULB, Université Libre de Bruxelles, Brussels, Belgium e-mail: gbonte@ulb.ac.be

Maxime Flauder
Machine Learning Group, Computer Science Department, Interuniversity Institute of Bioinformatics in Brussels (IB), ULB, Université Libre de Bruxelles, Brussels, Belgium e-mail: max.flauder@gmail.com

9.1 Introduction

The relationship between statistical dependency and causality lies at the heart of all statistical approaches to causal inference and can be summarized by two famous statements: *correlation (or more generally statistical association) does not imply causation* and *causation induces a statistical dependency between causes and effects (or more generally descendants)* [Reichenbach, 1956]. In other terms it is well known that statistical dependency is a necessary yet not sufficient condition for causality. The unidirectional link between these two notions has been used by many formal approaches to causality to justify the adoption of statistical methods for detecting or inferring causal links from observational data. The most influential one is the Causal Bayesian Network approach, detailed in [Koller and Friedman, 2009] which relies on notions of independence and conditional independence to detect causal patterns in the data. Well known examples of related inference algorithms are the constraint-based methods like the PC algorithms [Spirtes et al., 2000] and IC [Pearl, 2000]. These approaches are founded on probability theory and have been shown to be accurate in reconstructing causal patterns in many applications [Pourret et al., 2008], notably in bioinformatics [Friedman et al., 2000]. At the same time they restrict the set of configurations which causal inference is applicable to. Such boundary is essentially determined by the notion of *distinguishability* which defines the set of Markov equivalent configurations on the basis of conditional independence tests. Typical examples of indistinguishability are the two-variable setting and the completely connected triplet configuration [Guyon et al., 2007] where it is impossible to distinguish between cause and effects by means of conditional or unconditional independence tests.

If on one hand the notion of indistinguishability is probabilistically sound, on the other hand it should not prevent us from addressing interesting yet indistinguishable causal patterns. In fact, indistinguishability results rely on two main aspects: i) they refer only to specific features of dependency (notably conditional or unconditional independence) and ii) they state the conditions (e.g. faithfulness) under which it is possible to distinguish (or not) *with certainty* between configurations. Accordingly, indistinguishability results do not prevent the existence of statistical algorithms able to *reduce the uncertainty about the causal pattern* even in indistinguishable configurations. This has been made evident by the appearance in recent years of a series of approaches which tackle the cause-effect pair inference, like ANM (Additive Noise Model) [Hoyer et al., 2009], IGCI (Information Geometry Causality Inference) [Daniusis et al., 2010, Janzing et al., 2012], LiNGAM (Linear Non Gaussian Acyclic Model) [Shimizu et al., 2006] and the algorithms described in [Mooij et al., 2010] and [Statnikov et al., 2012][1]. What is common to these approaches is that they use alternative statistical features of the data to detect causal patterns and reduce the uncertainty about their directionality. A further important step in this direction

[1] A more extended list of recent algorithms is available in `http://www.causality.inf.ethz.ch/cause-effect.php?page=help`.

has been represented by the recent organization of the ChaLearn cause-effect pair challenge [Guyon, 2014]. The good (and significantly better than random) accuracy obtained on the basis of observations of pairs of causally related (or unrelated) variables supports the idea that alternative strategies can be designed to infer with success (or at least significantly better than random) indistinguishable configurations.

It is worthy to remark that the best ranked approaches[2] in the ChaLearn competition share a common aspect: they infer from statistical features of the bivariate distribution the probability of the existence and then of the directionality of the causal link between two variables. The success of these approaches shows that the problem of causal inference can be successfully addressed as a supervised machine learning approach where the inputs are features describing the probabilistic dependency and the output is a class denoting the existence (or not) of a directed causal link. Once sufficient training data are made available, conventional feature selection algorithms [Guyon and Elisseeff, 2003] and classifiers can be used to return a prediction better than random.

The effectiveness of machine learning strategies in the case of pairs of variables encourages the extension of the strategy to configurations with a larger number of variables. In this paper we propose an original approach to learn from multivariate observations the probability that a variable is a direct cause of another. This task is undeniably more difficult because

- the number of parameters needed to describe a multivariate distribution increases rapidly (e.g. quadratically in the Gaussian case),
- information about the existence of a causal link between two variables is returned also by the nature of the dependencies existing between the two variables and the remaining ones.

The second consideration is evident in the case of a collider configuration $\mathbf{z}_1 \rightarrow \mathbf{z}_2 \leftarrow \mathbf{z}_3$: in this case the dependency (or independency) between $\mathbf{z}_1$ and $\mathbf{z}_3$ tells us more about the link $\mathbf{z}_1 \rightarrow \mathbf{z}_2$ than the dependency between $\mathbf{z}_1$ and $\mathbf{z}_2$. This led us to develop a machine learning strategy (described in Section 9.2) where descriptors of the relation existing between members of the Markov blankets of two variables are used to learn the probability (i.e. a score) that a causal link exists between two variables. The approach relies on the asymmetry of some conditional (in)dependence relations between the members of the Markov blankets of two variables causally connected. The resulting algorithm (called D2C and described in Section 9.3) predicts the existence of a direct causal link between two variables in a multivariate setting by (i) creating a set of of features of the relationship based on asymmetric descriptors of the multivariate dependency and (ii) using a classifier to learn a mapping between the features and the presence of a causal link.

[2] We took part in the ChaLearn challenge and we ranked 8th in the final leader board.

In Section 9.4 we report the results of a set of experiments assessing the accuracy of the D2C algorithm. Experimental results based on synthetic and published data show that the D2C approach is competitive and often outperforms state-of-the-art methods.

9.2 Learning the Relation between Dependency and Causality in a Configuration with $n > 2$ Variables.

This section presents an approach to learn, from a number of observations, the relationships existing between the n variate distribution of $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n]$ and the existence of a directed causal link between two variables $\mathbf{z}_i$ and $\mathbf{z}_j$, $1 \leq i \neq j \leq n$, in the case of no confounding, no selection bias and no feedback configurations. Several parameters may be estimated from data in order to represent the multivariate distribution of $\mathbf{Z}$, like the correlation or the partial correlation matrix. Some problems however arise in this case like: (i) these parameters are informative in case of Gaussian distributions only, (ii) identical (or close) causal configurations could be associated to very different parametric values, thus making difficult the learning of the mapping and (iii) different causal configurations may lead to identical (or close) parametric values.

In other terms it is more relevant to describe the distribution in structural terms (e.g. with notions of conditional dependence/independence) rather than in parametric terms. Two more aspects have to be taken into consideration. First since we want to use a learning approach to identify cause-effect relationships we need some quantitative features to describe the structure of the multivariate distribution. Second, since asymmetry is a distinguishing characteristic of a causal relationship, we expect that effective features should share the same asymmetric properties.

In this paper we will use information theory to represent and quantify the notions of (conditional) dependence and independence between variables and to derive a set of asymmetric features to reconstruct causality from dependency.

9.2.1 Notions of Information Theory

Let us consider three continuous random variables $\mathbf{z}_1$, $\mathbf{z}_2$ and $\mathbf{z}_3$ having a joint Lebesgue density[3]. Let us start by considering the relation between $\mathbf{z}_1$ and $\mathbf{z}_2$. The mutual information [Cover and Thomas, 1990] between $\mathbf{z}_1$ and $\mathbf{z}_2$ is defined in terms

[3] Boldface denotes random variables.

of their probabilistic density functions $p(z_1)$, $p(z_2)$ and $p(z_1,z_2)$ as

$$I(\mathbf{z}_1;\mathbf{z}_2) = \int\int \log \frac{p(z_1,z_2)}{p(z_1)p(z_2)} p(z_1,z_2) dz_1 dz_2 = H(\mathbf{z}_1) - H(\mathbf{z}_1|\mathbf{z}_2) \tag{9.1}$$

where H is the *entropy* and the convention $0 \log \frac{0}{0} = 0$ is adopted. This quantity measures the amount of stochastic dependence between $\mathbf{z}_1$ and $\mathbf{z}_2$ [Cover and Thomas, 1990]. Note that, if $\mathbf{z}_1$ and $\mathbf{z}_2$ are Gaussian distributed the following relation holds

$$I(\mathbf{z}_1;\mathbf{z}_2) = -\frac{1}{2}\log(1-\rho^2) \tag{9.2}$$

where ρ is the Pearson correlation coefficient between $\mathbf{z}_1$ and $\mathbf{z}_2$.

Let us now consider a third variable $\mathbf{z}_3$. The *conditional mutual information* [Cover and Thomas, 1990] between $\mathbf{z}_1$ and $\mathbf{z}_2$ once $\mathbf{z}_3$ is given is defined by

$$\begin{aligned} I(\mathbf{z}_1;\mathbf{z}_2|\mathbf{z}_3) = \int\int\int \log \frac{p(z_1,z_2|z_3)}{p(z_1|z_3)p(z_2|z_3)} p(z_1,z_2,z_3) dz_1 dz_2 dz_3 = \\ = H(\mathbf{z}_1|\mathbf{z}_3) - H(\mathbf{z}_1|\mathbf{z}_2,\mathbf{z}_3) \end{aligned} \tag{9.3}$$

The conditional mutual information is null if and only if $\mathbf{z}_1$ and $\mathbf{z}_2$ are conditionally independent given $\mathbf{z}_3$.

A structural notion which can be described in terms of conditional mutual information is the notion of Markov Blanket (MB). The Markov Blanket of variable $\mathbf{z}_i$ in an n dimensional distribution 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. In information theoretic terms let us consider a set $\mathbf{Z}$ of n random variables, a variable $\mathbf{z}_i$ and a subset $\mathbf{M}_i \subset \mathbf{Z} \setminus \mathbf{z}_i$. The subset $\mathbf{M}_i$ is said to be a *Markov blanket* of $\mathbf{z}_i$ if it is the minimal subset satisfying

$$I(\mathbf{z}_i;(\mathbf{Z} \setminus (\mathbf{M}_i \cup \mathbf{z}_i))|\mathbf{M}_i) = 0$$

Effective algorithms have been proposed in literature to infer a Markov Blanket from observed data [Tsamardinos et al., 2003b]. Feature selection algorithms are also useful to construct a Markov blanket of a given target variable once they rely on notions of conditional independence to select relevant variables [Meyer and Bontempi, 2014].

9.2.2 Causality and Asymmetric Dependency Relationships

The notion of causality is central in science and also an intuitive notion of everyday life. The remarkable property of causality which distinguishes it from dependency is asymmetry.

In probabilistic terms a variable $\mathbf{z}_i$ is dependent on a variable $\mathbf{z}_j$ if the density of $\mathbf{z}_i$, conditional on the observation $\mathbf{z}_j = z_j$, is different from the marginal one

$$p(z_i|\mathbf{z}_j = z_j) \neq p(z_i)$$

In information theoretic terms the two variables are dependent if $I(\mathbf{z}_i;\mathbf{z}_j) = I(\mathbf{z}_j;\mathbf{z}_i) > 0$. This implies that dependency is *symmetric*. If $\mathbf{z}_i$ is dependent on $\mathbf{z}_j$, then $\mathbf{z}_j$ is dependent on $\mathbf{z}_i$ too as shown by

$$p(z_j|\mathbf{z}_i = z_i) \neq p(z_j)$$

The formal representation of the notion of causality demands an extension of the syntax of the probability calculus as done by Pearl [1995] with the introduction of the operator do which allows to distinguish the observation of a value of $\mathbf{z}_j$ (denoted by $\mathbf{z}_j = z_j$) from the manipulation of the variable $\mathbf{z}_j$ (denoted by $\mathtt{do}(\mathbf{z}_\mathtt{j} = \mathtt{z}_\mathtt{j})$). Once this extension is accepted we say that a variable $\mathbf{z}_j$ is a cause of a variable $\mathbf{z}_i$ (e.g. "diseases cause symptoms") if the distribution of $\mathbf{z}_i$ is different from the marginal one when we set the value $\mathbf{z}_j = z_j$

$$p(z_i|\mathtt{do}(\mathbf{z}_j = z_j)) \neq p(z_i)$$

but not vice versa (e.g. "symptoms do not cause disease")

$$p(z_j|\mathtt{do}(\mathbf{z}_i = z_i)) = p(z_j)$$

The extension of the probability notation made by Pearl allows to formalize the intuition that causality is *asymmetric*. Another notation which allows to represent causal expression is provided by graphical models or more specifically by Directed Acyclic Graphs (DAG) [Koller and Friedman, 2009]. In this paper we will limit to consider causal relationships modeled by DAG, which proved to be convenient tools to understand and use the notion of causality. Furthermore we will make the assumption that the set of causal relationships existing between the variables of interest can be described by a Markov and faithful DAG [Pearl, 2000]. This means that the DAG is an accurate map of dependencies and independencies of the represented distribution and that using the notion of *d-separation* it is possible to read from the graph if two sets of nodes are (in)dependent conditioned on a third.

The asymmetric nature of causality suggests that if we want to infer causal links from dependency we need to find some features (or descriptors) which describe the

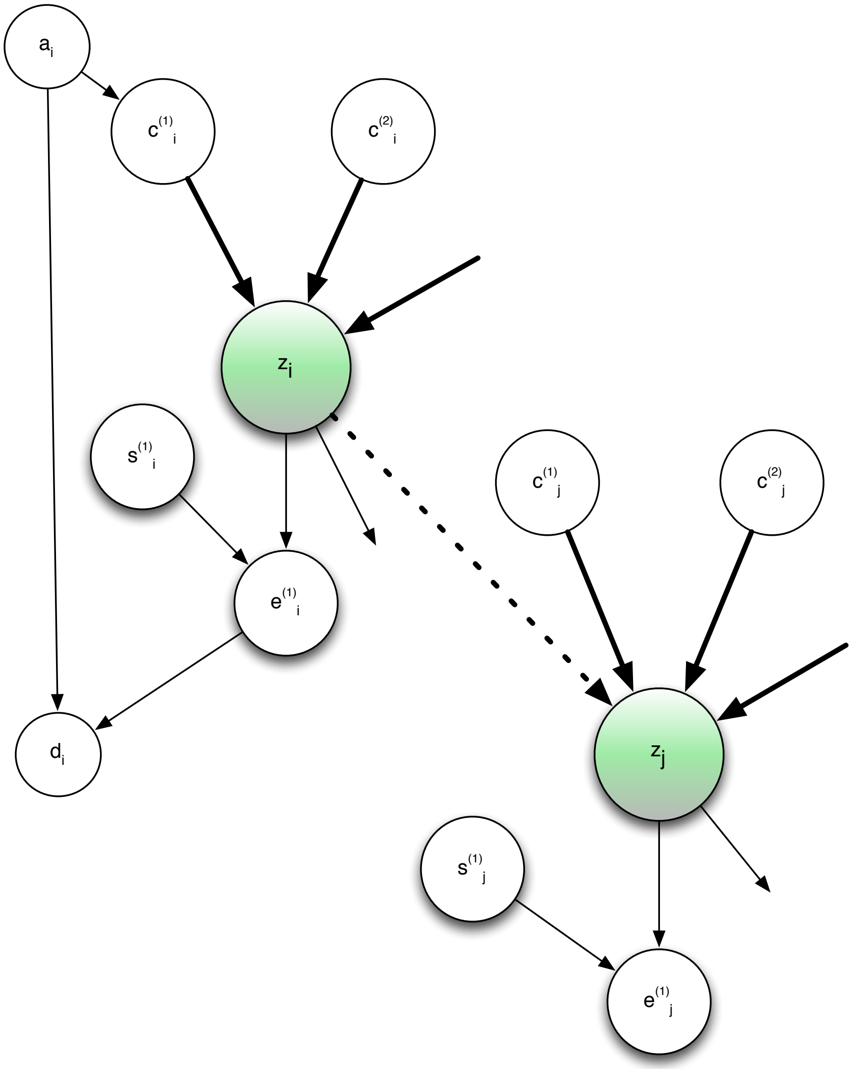

Fig. 9.1: Two causally connected variables and their Markov Blankets.

dependency and share with causality the property of asymmetry. Let us suppose that we are interested in predicting the existence of a directed causal link $\mathbf{z}_i \rightarrow \mathbf{z}_j$ where $\mathbf{z}_i$ and $\mathbf{z}_j$ are components of an observed n-dimensional vector $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$.

We define as *dependency descriptor* of the ordered pair $\langle i, j \rangle$ 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 $\mathbf{z}_k, i \neq j, j \neq k, i \neq k$.

We call a dependency descriptor *symmetric* if $d(i, j) = d(j, i)$ otherwise we call it *asymmetric*. Correlation and mutual information are symmetric descriptors since

$$d(i, j) = I(\mathbf{z}_i; \mathbf{z}_j) = I(\mathbf{z}_j; \mathbf{z}_i) = d(j, i)$$

Because of the asymmetric property of causality, if we want to maximize our chances to reconstruct causality from dependency we need to identify relevant asymmetric descriptors. In order to define useful asymmetric descriptors we have recourse to the Markov Blankets of the two variables $\mathbf{z}_i$ and $\mathbf{z}_j$.

Let us consider for instance the portion of a DAG represented in Figure 9.1 where the variable $\mathbf{z}_i$ is a direct cause of $\mathbf{z}_j$. The figure shows also the Markov Blankets of the two variables (denoted M_i and M_j respectively) and their components, i.e. the direct causes (denoted by $\mathbf{c}$), the direct effects ($\mathbf{e}$) and the spouses ($\mathbf{s}$) [Pellet and Elisseeff, 2008].

In what follows we will make two assumptions: (i) the only path between the sets $\mathbf{z}_i \cup M_i$ and $\mathbf{z}_j \cup M_j$ 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 $\mathbf{s}_i$ ($\mathbf{s}_j$). We will discuss these assumptions at the end of the section. Given these assumptions and because of d-separation [Geiger et al., 1990], a number of asymmetric conditional (in)dependence relations holds between the members of M_i and M_j (Table 9.1). For instance (first line of Table 9.1), by conditioning on the effect $\mathbf{z}_j$ we create a dependence between $\mathbf{z}_i$ and the direct causes of $\mathbf{z}_j$ while by conditioning on the $\mathbf{z}_i$ we d-separate $\mathbf{z}_j$ and the direct causes of $\mathbf{z}_i$.

The relations in Table 9.1 can be used to define the following set of asymmetric descriptors,

$$d_1^{(k)}(i, j) = I(\mathbf{z}_i; \mathbf{c}_j^{(k)} | \mathbf{z}_j), \tag{9.4}$$

$$d_2^{(k)}(i, j) = I(\mathbf{e}_i^{(k)}; \mathbf{c}_j^{(k)} | \mathbf{z}_j), \tag{9.5}$$

$$d_3^{(k)}(i, j) = I(\mathbf{c}_i^{(k)}; \mathbf{c}_j^{(k)} | \mathbf{z}_j), \tag{9.6}$$

$$d_4^{(k)}(i, j) = I(\mathbf{z}_i; \mathbf{c}_j^{(k)}), \tag{9.7}$$

whose asymmetry is given by

$$d_1^{(k)}(i,j) = I(\mathbf{z}_i;\mathbf{c}_j^{(k)}|\mathbf{z}_j) > 0, \quad d_1^{(k)}(j,i) = I(\mathbf{z}_j;\mathbf{c}_i^{(k)}|\mathbf{z}_i) = 0, \tag{9.8}$$

$$d_2^{(k)}(i,j) = I(\mathbf{e}_i^{(k)};\mathbf{c}_j^{(k)}|\mathbf{z}_j) > 0, \quad d_2^{(k)}(j,i) = I(\mathbf{e}_j^{(k)};\mathbf{c}_i^{(k)}|\mathbf{z}_i) = 0, \tag{9.9}$$

$$d_3^{(k)}(i,j) = I(\mathbf{c}_i^{(k)};\mathbf{c}_j^{(k)}|\mathbf{z}_j) > 0, \quad d_3^{(k)}(j,i) = I(\mathbf{c}_j^{(k)};\mathbf{c}_i^{(k)}|\mathbf{z}_i) = 0, \tag{9.10}$$

$$d_4^{(k)}(i,j) = I(\mathbf{z}_i;\mathbf{c}_j^{(k)}) = 0, \quad d_4^{(k)}(j,i) = I(\mathbf{z}_j;\mathbf{c}_i^{(k)}) > 0. \tag{9.11}$$

Relation i,j	Relation j,i
$\forall k \quad \mathbf{z}_i \not\perp\!\!\!\perp \mathbf{c}_j^{(k)} \mid \mathbf{z}_j$	$\forall k \quad \mathbf{z}_j \perp\!\!\!\perp \mathbf{c}_i^{(k)} \mid \mathbf{z}_i$
$\forall k \quad \mathbf{e}_i^{(k)} \not\perp\!\!\!\perp \mathbf{c}_j^{(k)} \mid \mathbf{z}_j$	$\forall k \quad \mathbf{e}_j^{(k)} \perp\!\!\!\perp \mathbf{c}_i^{(k)} \mid \mathbf{z}_i$
$\forall k \quad \mathbf{c}_i^{(k)} \not\perp\!\!\!\perp \mathbf{c}_j^{(k)} \mid \mathbf{z}_j$	$\forall k \quad \mathbf{c}_j^{(k)} \perp\!\!\!\perp \mathbf{c}_i^{(k)} \mid \mathbf{z}_i$
$\forall k \quad \mathbf{z}_i \perp\!\!\!\perp \mathbf{c}_j^{(k)}$	$\forall k \quad \mathbf{z}_j \not\perp\!\!\!\perp \mathbf{c}_i^{(k)}$

Table 9.1: Asymmetric (un)conditional (in)dependance relationships between members of the Markov Blankets of $\mathbf{z}_i$ and $\mathbf{z}_j$ in Figure 9.1.

Relation i,j	Relation j,i
$\forall k \quad \mathbf{z}_i \not\perp\!\!\!\perp \mathbf{e}_j^{(k)}$	$\forall k \quad \mathbf{z}_j \not\perp\!\!\!\perp \mathbf{e}_i^{(k)}$
$\forall k \quad \mathbf{z}_i \perp\!\!\!\perp \mathbf{s}_j^{(k)}$	$\forall k \quad \mathbf{z}_j \perp\!\!\!\perp \mathbf{s}_i^{(k)}$
$\forall k \quad \mathbf{z}_i \perp\!\!\!\perp \mathbf{e}_j^{(k)} \mid \mathbf{z}_j$	$\forall k \quad \mathbf{z}_j \perp\!\!\!\perp \mathbf{e}_i^{(k)} \mid \mathbf{z}_i$
$\forall k \quad \mathbf{z}_i \perp\!\!\!\perp \mathbf{s}_j^{(k)} \mid \mathbf{z}_j$	$\forall k \quad \mathbf{z}_j \perp\!\!\!\perp \mathbf{s}_i^{(k)} \mid \mathbf{z}_i$
$\forall k \quad \mathbf{e}_i^{(k)} \perp\!\!\!\perp \mathbf{e}_j^{(k)} \mid \mathbf{z}_i$	$\forall k \quad \mathbf{e}_j^{(k)} \perp\!\!\!\perp \mathbf{e}_i^{(k)} \mid \mathbf{z}_j$
$\forall k \quad \mathbf{e}_i^{(k)} \perp\!\!\!\perp \mathbf{s}_j^{(k)} \mid \mathbf{z}_j$	$\forall k \quad \mathbf{e}_j^{(k)} \perp\!\!\!\perp \mathbf{s}_i^{(k)} \mid \mathbf{z}_i$

Table 9.2: Symmetric (un)conditional (in)dependance relationships between members of the Markov Blankets of $\mathbf{z}_i$ and $\mathbf{z}_j$ in Figure 9.1.

At the same time we can write a set of symmetric conditional (in)dependence relations (Table 9.2) and the equivalent formulations in terms of mutual information terms:

$$I(\mathbf{z}_j;\mathbf{e}_i^{(k)}) > 0, \tag{9.12}$$

$$I(\mathbf{z}_i;\mathbf{e}_j^{(k)}) > 0, \tag{9.13}$$

$$I(\mathbf{z}_j;\mathbf{s}_i^{(k)}) = I(\mathbf{z}_i;\mathbf{s}_j^{(k)}) = 0, \tag{9.14}$$

$$I(\mathbf{z}_i;\mathbf{e}_j^{(k)}|\mathbf{z}_j) = I(\mathbf{z}_j;\mathbf{e}_i^{(k)}|\mathbf{z}_i) = I(\mathbf{z}_i;\mathbf{s}_j^{(k)}|\mathbf{z}_j) = I(\mathbf{z}_j;\mathbf{s}_i^{(k)}|\mathbf{z}_i) = 0, \tag{9.15}$$

$$I(\mathbf{e}_j^{(k)};\mathbf{e}_i^{(k)}|\mathbf{z}_i) = I(\mathbf{e}_i^{(k)};\mathbf{e}_j^{(k)}|\mathbf{z}_j) = I(\mathbf{e}_i^{(k)};\mathbf{s}_j^{(k)}|\mathbf{z}_j) = I(\mathbf{e}_j^{(k)};\mathbf{s}_i^{(k)}|\mathbf{z}_i) = 0. \tag{9.16}$$

9.2.3 From Asymmetric Relationships to Distinct Distributions

The asymmetric properties of the four descriptors (9.4)-(9.7) is encouraging if we want to exploit dependency related features to infer causal properties from data. However, this optimism is undermined by the fact that all the descriptors require already the capability of distinguishing between the causes (i.e. the terms **c**) and the effects (i.e. the terms **e**) of the Markov Blanket of a given variable. Unfortunately this discriminating capability is what we are looking for!

In order to escape this circularity problem we consider two solutions. The first is to have recourse to a preliminary phase that prioritizes the components of the Markov Blanket and then use this result as starting point to detect asymmetries and then improve the classification of causal links. This is for instance feasible by using a filter selection algorithm, like mIMR [Bontempi and Meyer, 2010, Bontempi et al., 2011], which aims to prioritize the direct causes in the Markov Blanket by searching for pairs of variables with high relevance and low interaction.

The second solution is related to the fact that the asymmetry of the four descriptors induces a difference in the distributions of some information theoretic terms which do not require the distinction between causes and effects within the Markov Blanket. The consequence is that we can replace the descriptors (9.4)-(9.7) with other descriptors (denoted with the letter D) that can be actually estimated from data.

Let $\mathbf{m}^{(k)}$ denote a generic component of the Markov Blanket with no distinction between cause, effect or spouse. It follows that a population made of terms depending on $\mathbf{m}^{(k)}$ is a mixture of three subpopulations, the first made of causes, the second made of effects and the third of spouses, respectively. It follows that the distribution of the population is a *finite mixture* [McLaughlan, 2000] of three distributions, the first related to the causes, the second to the effects and the third to the spouses. Since the moments of the finite mixture are functions of the moments of each component, we can derive some properties of the resulting mixture from the properties of each component. For instance if we can show that all the subpopulations but one are identical (e.g. all the elements of the third subpopulation in the first mixture are larger than the elements of the analogous subpopulation in the second mixture), we can derive that the two mixture distributions are different.

Consider for instance the quantity $I(\mathbf{z}_i;\mathbf{m}_j^{(k_j)}|\mathbf{z}_j)$ where $\mathbf{m}_j^{(k_j)}$, $k_j = 1,\dots,K_j$ is a member of the set $M_j \setminus \mathbf{z}_i$. From (9.8) and (9.15) it follows that the mixture distribution associated to the populations $D_1(i,j) = \{I(\mathbf{z}_i;\mathbf{m}_j^{(k_j)}|\mathbf{z}_j), k_j = 1,\dots,K_j\}$ and $D_1(j,i) = \{I(\mathbf{z}_j;\mathbf{m}_i^{(k_i)}|\mathbf{z}_i), k_i = 1,\dots,K_i\}$ are different since

$$\begin{cases} I(\mathbf{z}_i;\mathbf{m}_j^{(k_j)}|\mathbf{z}_j) > I(\mathbf{z}_j;\mathbf{m}_i^{(k_i)}|\mathbf{z}_i), & \text{if } \mathbf{m}_j^{(k_j)} = \mathbf{c}_j^{(k_j)} \wedge \mathbf{m}_i^{(k_i)} = \mathbf{c}_i^{(k_i)} \\ I(\mathbf{z}_i;\mathbf{m}_j^{(k_j)}|\mathbf{z}_j) = I(\mathbf{z}_j;\mathbf{m}_i^{(k_i)}|\mathbf{z}_i), & \text{else} \end{cases} \tag{9.17}$$

It follows that even if we are not able to distinguish between a cause $\mathbf{c}_j \in M_j$ and an effect $\mathbf{e}_j \in M_j$, we know that the distribution of the population $D_1(i,j)$ differs from the distribution of the population $D_1(j,i)$. We can therefore use the population $D_1(i,j)$ (or some of its moments) as descriptor of the causal dependency.

Similarly we can replace the descriptors (9.5), (9.6) with the distributions of the population $D_2(i,j) = \{I(\mathbf{m}_i^{(k_i)};\mathbf{m}_j^{(k_j)}|\mathbf{z}_j), k_j = 1,\dots,K_j, k_i = 1,\dots,K_i\}$. From (9.9), (9.10) and (9.16) we obtain that the distributions of the populations $D_2(i,j)$ and $D_2(j,i)$ are different.

If we make the additional assumption that $I(\mathbf{z}_j;\mathbf{e}_i^{(k)}) = I(\mathbf{z}_i;\mathbf{e}_j^{(k)}) > 0$ from (9.11) we obtain also that the distribution of the population $D_3(i,j) = \{I(\mathbf{z}_i;\mathbf{m}_j^{(k_j)}), k_j = 1,\dots,K_j\}$ is different from the one of $D_3(j,i) = \{I(\mathbf{z}_j;\mathbf{m}_i^{(k_i)}), k_i = 1,\dots,K_i\}$.

The previous results are encouraging and show that though we are not able to distinguish between the different components of a Markov Blanket, we can notwithstanding compute some quantities (in this case distributions of populations) whose asymmetry is informative about the causal relationships $\mathbf{z}_i \rightarrow \mathbf{z}_j$.

As a consequence by measuring from observed data some statistics (e.g. quantiles) related to the distribution of these asymmetric descriptors, we may obtain some insight about the causal relationship between two variables. This idea is made explicit in the algorithm described in the following section.

Though these results rely on the two assumptions made before (i.e. single path and no common ancestors), two considerations are worthy to be made. First, the main goal of the approach is to shed light on the existence of dependency asymmetries also in multivariate contributions. Secondly we expect that the second layer (based on supervised learning) will eventually compensate for configurations not compliant with the assumptions and take advantage of complementarity or synergy of the descriptors in discriminating between causal configurations.

9.3 The D2C Algorithm

The rationale of the D2C algorithm is to predict the existence of a causal link between two variables in a multivariate setting by (i) creating a set of features of the relationship between the members of the Markov Blankets of the two variables and (ii) using a classifier (e.g. a Random Forest as in our experiments) to learn a mapping between the features and the presence of a causal link.

We use two sets of features to summarize the relation between the two Markov blankets: the first one accounts for the presence (or the position if the MB is obtained by ranking) of the terms of M_j in M_i and vice versa. For instance it is evident that if $\mathbf{z}_i$ is a cause of $\mathbf{z}_j$ we expect to find $\mathbf{z}_i$ highly ranked between the causal terms of M_j but $\mathbf{z}_j$ absent (or ranked low) among the causes of M_i. The second set of features is based on the results of the previous section and is obtained by summarizing the distributions of the asymmetric descriptors with a set of quantiles.

We propose then an algorithm (D2C) which for each pair of measured variables $\mathbf{z}_i$ and $\mathbf{z}_j$:

1. infers from data the two Markov Blankets (e.g. by using state-of-the-art approaches) M_i and M_j and the subsets $M_i \setminus \mathbf{z}_j = \{\mathbf{m}^{(k_i)}, k_i = 1, \dots, K_i\}$ and $M_j \setminus \mathbf{z}_i = \{\mathbf{m}^{(k_j)}, k_j = 1, \dots, K_j\}$. Most of the existing algorithms associate to the Markov Blanket a ranking such that the most strongly relevant variables are ranked before.
2. computes a set of (conditional) mutual information terms describing the dependency between $\mathbf{z}_i$ and $\mathbf{z}_j$

$$I = [I(\mathbf{z}_i;\mathbf{z}_j), I(\mathbf{z}_i;\mathbf{z}_j|\mathrm{M}_j \setminus \mathbf{z}_i), I(\mathbf{z}_i;\mathbf{z}_j|\mathrm{M}_i \setminus \mathbf{z}_j)] \tag{9.18}$$

3. computes the positions $P_i^{(k_i)}$ of the members $\mathbf{m}^{(k_i)}$ of $M_i \setminus \mathbf{z}_j$ in the ranking associated to $M_j \setminus \mathbf{z}_i$ and the positions $P_j^{(k_j)}$ of the terms $\mathbf{m}^{(k_j)}$ in the ranking associated to $M_i \setminus \mathbf{z}_j$. Note that in case of the absence of a term of M_i in M_j, the position is set to $K_j + 1$ (respectively $K_i + 1$).
4. computes the populations based on the asymmetric descriptors introduced in Section 9.2.3:

   a. $D_1(i,j) = \{I(\mathbf{z}_i;\mathbf{m}_j^{(k_j)}|\mathbf{z}_j), k_j = 1, \dots, K_j\}$

   b. $D_1(j,i) = \{I(\mathbf{z}_j;\mathbf{m}_i^{(k_i)}|\mathbf{z}_i), k_i = 1, \dots, K_i\}$

   c. $D_2(i,j) = \{I(\mathbf{m}_i^{(k_i)};\mathbf{m}_j^{(k_j)}|\mathbf{z}_j), k_i = 1, \dots, K_i, k_j = 1, \dots, K_j\}$ and

   d. $D_2(j,i) = \{I(\mathbf{m}_j^{(k_j)};\mathbf{m}_i^{(k_i)}|\mathbf{z}_i), k_i = 1, \dots, K_i, k_j = 1, \dots, K_j\}$

   e. $D_3(i,j) = \{I(\mathbf{z}_i;\mathbf{m}_j^{(k_j)}), k_j = 1,\dots,K_j\}$,

   f. $D_3(j,i) = \{I(\mathbf{z}_j,\mathbf{m}_i^{(k_i)}), k_i = 1,\dots,K_i\}$

5. creates a vector of descriptors

$$x = [I, \mathscr{Q}(\hat{P}_i), \mathscr{Q}(\hat{P}_j), \mathscr{Q}(\hat{D}_1(i,j)), \mathscr{Q}(\hat{D}_1(j,i)), \\ \mathscr{Q}(\hat{D}_2(i,j)), \mathscr{Q}(\hat{D}_2(j,i)), \mathscr{Q}(\hat{D}_3(i,j)), \mathscr{Q}(\hat{D}_3(j,i))] \quad (9.19)$$

   where $\hat{P}_i$ and $\hat{P}_j$ are the empirical distributions of the populations $\{P_i^{(k_i)}\}$ and $\{P_j^{(k_j)}\}$, $\hat{D}_h(i,j)$ denotes the empirical distribution of the corresponding population $D_h(i,j)$ and $\mathscr{Q}$ returns a set of sample quantiles of a distribution (in the experiments we set the quantiles to 0.1, 0.25, 0.5, 0.75, 0.9).

The vector x can be then derived from observational data and used to create a vector of descriptors to be used as inputs in a supervised learning paradigm.

The rationale of the algorithm is that the asymmetries between M_i and M_j (e.g. Table 9.1) induce an asymmetry on the distributions $\hat{P}$ and $\hat{D}$ and that the quantiles of those distributions provide information about the directionality of causal link ($\mathbf{z}_i \to \mathbf{z}_j$ or $\mathbf{z}_j \to \mathbf{z}_i$.) In other terms we expect that the distribution of these variables should return useful information about which is the cause and the effect. Note that these distributions would be more informative if we were able to rank the terms of the Markov Blankets by prioritizing the direct causes (i.e. the terms $\mathbf{c}_i$ and $\mathbf{c}_j$) since these terms play a major role in the asymmetries of Table 9.1. The D2C algorithm can then be improved by choosing an appropriate Markov Blanket selector algorithms, like the mIMR filter.

In the experiments (Section 9.4) we derive the information terms as difference between (conditional) entropy terms (see Equations 9.1 and 9.3) which are themselves estimated by a Lazy Learning regression algorithm [Bontempi et al., 1999] by making an assumption of Gaussian noise. Lazy Learning returns a leave-one-out estimation of conditional variance which can be easily transformed in entropy under the normal assumption [Cover and Thomas, 1990]. The (conditional) mutual information terms are then obtained by using the relations (9.1) and (9.3).

9.3.1 Complexity Analysis

In this subsection we make a complexity analysis of the approach: first it is important to remark that since the D2C approach relies on a classifier, its learning phase can be time-consuming and dependent on the number of samples and dimension. However, this step is supposed to be performed only once and from the user perspective it is more relevant to consider the cost in the testing phase. Given two nodes for which a test of the existence of a causal link is required, three steps have to be performed:

1. computation of the Markov blankets of the two nodes. The information filters we used have a complexity $O(Cn^2)$ where C is the cost of the computation of mutual information [Meyer and Bontempi, 2014]. In case of very large n this complexity may be bounded by having the filter preceded by a ranking algorithm with complexity $O(Cn)$. Such ranking may limit the number of features taken into consideration by the filters to $n' < n$ reducing then considerably the cost.
2. once a number K_i (K_j) of members of MB_i (MB_j) have been chosen, the rest of the procedure has a complexity related to the estimation of a number $O(K_iK_j)$ of descriptors. In this paper we used a local learning regression algorithm to estimate the conditional entropies terms. Given that each regression involves at most three terms, the complexity is essentially related linearly to the number N of samples
3. the last step consists in the computation of the Random Forest predictions on the test set. Since the RF has been already trained, the complexity of this step depends only on the number of trees and not on the dimensionality or number of samples.

For each test, the resulting complexity has then a cost of the order $O(Cn + Cn'^2 + K_iK_jN)$. It is important to remark that an advantage of D2C is that, if we are interested in predicting the causal relation between two variables only, we are not forced to infer the entire adjacency matrix (as typically the case in constraint-based methods). This mean also that the computation of the entire matrix can be easily made parallel.

9.4 Experimental Validation

In this section the D2C (Section 9.3) algorithm is assessed in a set of synthetic experiments and published data sets.

9.4.1 Synthetic Data

This experimental session addresses the problem of inferring causal links from synthetic data generated for linear and non-linear DAG configurations of different sizes.

All the variables are continuous, and the dependency between children and parents is modelled by the additive relationship

$$x_i = \sum_{j \in par(i)} f_{i,j}(x_j) + \varepsilon_i, \qquad i = 1, \dots, n \tag{9.20}$$

where the noise $\varepsilon_i \sim N(0, \sigma_i)$ is Normal, $f_{i,j}(x) \in L(x)$ and three sets of continuous functions are considered:

- `linear`: $L(x) = \{f \mid f(x) = a_0 + a_1 x\}$
- `quadratic`: $L(x) = \{f \mid f(x) = a_0 + a_1 x + a_2 x^2\}$
- `sigmoid`: $L(x) = \{f \mid f(x) = \frac{1}{1+exp(a_0+a_1 x)}\}$

In order to assess the accuracy with respect to dimensionality, we considered three network sizes:

- `small`: number of nodes n is uniformly sampled in the interval $[20, 30]$,
- `medium`: number of nodes n is uniformly sampled in the interval $[100, 200]$,
- `large`: number of nodes n is uniformly sampled in the interval $[500, 1000]$,

The assessment procedure relies on the generation of a number of DAG structures[4] and the simulation, for each of them, of N (uniformly random in $[100, 500]$) node observations according to the dependency (9.20). In each data set we removed the observations of five percent of the variables in order to introduce unobserved variables.

For each DAG, on the basis of its structure and the data set of observations, we collect a number of pairs $\langle x_d, y_d \rangle$, where x_d is the descriptor vector returned by (9.19) and y_d is the class denoting the existence (or not) of the causal link in the DAG topology.

Several sizes of training set are considered. The largest D2C training set is made of $D = 60000$ pairs $\langle x_d, y_d \rangle$ and is obtained by generating DAGs and storing for each of them the descriptors associated to at most 4 positives examples (i.e. a pair where the node z_i is a direct cause of z_j) and at most 6 negatives examples (i.e. a pair where the node z_i is not a direct cause of z_j). A Random Forest classifier is trained on the balanced data set: we use the implementation from the R package `randomForest` [Liaw and Wiener, 2002] with default setting.

The test set is obtained by considering a number of independently simulated DAGs. We consider 190 DAGs for the small and medium configurations and 90 for

[4] We used the function `random_dag` from the R package gRbase [Dethlefsen and Højsgaard, 2005].

the large configuration. For each testing DAG we select 4 positives examples (i.e. a pair where the node z_i is a direct cause of z_j) and 6 negatives examples (i.e. a pair where the node z_i is not a direct cause of z_j). The predictive accuracy of the trained Random Forest classifier is then assessed on the test set.

The D2C approach is compared in terms of classification accuracy (Balanced Error Rate (BER)) to several state-of-the-art approaches:

- `ANM`: Additive Noise Model [Hoyer et al., 2009] using a Gaussian process with RBF kernel and the Hilbert-Schmidt Independence Criterion (pvalue=0.02)[5]
- `DAGL1`: DAG-Search score-based algorithm with potential parents selected with a L1 penalization [Schmidt et al., 2007][6].
- `DAGSearch`: unrestricted DAG-Search score-based algorithm (multiple restart greedy hill-climbing, using edge additions, deletions, and reversals) [Schmidt et al., 2007][6],
- `DAGSearchSparse`: DAG-Search score-based algorithm with potential parents restricted to the 10 most correlated features [Schmidt et al., 2007][6],
- `gs`: Grow-Shrink constraint-based structure learning algorithm [Margaritis, 2003][7],
- `hc`: hill-climbing score-based structure learning algorithm [Daly and Shen, 2007][7],
- `iamb`: incremental association MB constraint-based structure learning algorithm [Tsamardinos et al., 2003b][7],
- `mmhc`: max-min hill climbing hybrid structure learning algorithms [Tsamardinos et al., 2010][7],
- `PC`: Estimate the equivalence class of a DAG using the PC algorithm[8] (this method was used only for the small size configuration (Figure 9.3) for computational time reasons)
- `si.hiton.pc`: Semi-Interleaved HITON-PC local discovery structure learning algorithms [Tsamardinos et al., 2003a][7],
- `tabu`: tabu search score-based structure learning algorithm[7].

The BER of six versions of the D2C method are compared to the BER of state-of-the-art methods in Figures 9.3 (small), Figure 9.4 (medium) and Figure 9.5 (large).

[5] The code is available in `https://staff.fnwi.uva.nl/j.m.mooij/code/additive-noise.tar.gz`.

[6] The code is available in `http://www.cs.ubc.ca/~murphyk/Software/DAGlearn/`.

[7] The code is available in the R package `bnlearn` [Scutari, 2010].

[8] The code is available in the R package `pcalg` [Kalisch et al., 2012]

The six versions of D2C are obtained by considering two types of training data (i.e. one based on linear dependency and one based on the same dependency used for testing) and three training set sizes (equal to 400, 3000 and 60000 respectively) Each subfigure corresponds to the three types of stochastic dependency (top: linear, middle: quadratic, bottom: sigmoid).

A series of considerations can be made on the basis of the experimental results:

- the n-variate approach D2C obtains competitive results with respect to several state-of-the-art techniques in the linear case,
- the improvement of D2C wrt state-of-the-art techniques (often based on linear assumptions) tends to increase when we move to more nonlinear configurations, In particular the accuracy of the D2C algorithm is able to generalize to DAG with different number of nodes and different distributions also when trained only on data observed for linear DAGs (see accuracy of D2Cx$_{lin}$ in the second and third row of Figures 9.3, 9.4 and 9.5)
- the accuracy of the D2C approach improves by increasing the number of training examples,
- with a small number of examples (i.e. $N = 400$) it is already possible to learn a classifier D2C whose accuracy is competitive with state-of-the-art methods,
- the ANM approach is not able to return accurate information about causal dependency by taking into consideration only bivariate information,
- the analysis of the importance of the D2C descriptors (based on the Mean Decrease Accuracy of the Random Forest in Figure 9.2) shows that the most relevant variables in the vector (9.19) are the terms in I, D_1 and D_3.

The D2C code is available in the CRAN R package `D2C` [Bontempi et al., 2014].

9.4.2 Published Data

The second part of the assessment relies on the simulated and resimulated data sets proposed in Table 11 of [Aliferis et al., 2010]. These 103 data sets were obtained by simulating data from known Bayesian networks and also by resimulation, where real data is used to elicit a causal network and then data is simulated from the obtained network. We split the 103 data sets in two portions: a training portion (made of 52 sets) and a second portion (made of 51 sets) for testing. This was done in order to assess the accuracy of two versions of the D2C algorithm: the first uses as training set only 40000 synthetic samples generated as in the previous section, the second includes in the training set also the 52 data sets of the training portion. The goal

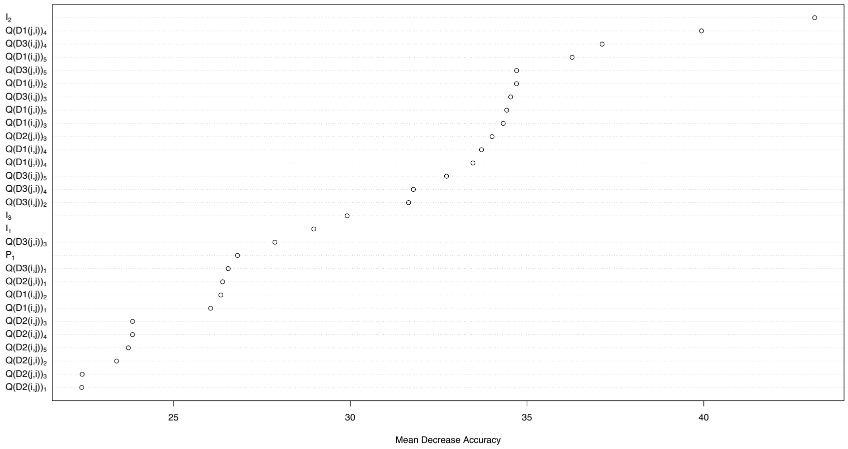

Fig. 9.2: Importance of D2C features returned by the Random Forest mean decrease accuracy. I_i denotes the ith component of the descriptor vector (9.18) while $Q(Dx(i,j))_k$ denotes the kth quantile of the population of descriptor $Dx(i,j)$.

is to assess the generalization accuracy of the D2C algorithm with respect to DAG distributions never encountered before and not included in the training set. In this section we compare D2C to a set of algorithms implemented by the *Causal Explorer* software [Aliferis et al., 2003][9]:

- `GS`: Grow/Shrink algorithm
- `IAMB`: Incremental Association-Based Markov Blanket
- `IAMBnPC`: IAMB with PC algorithm in the pruning phase
- `interIAMBnPC`: IAMB with PC algorithm in the interleaved pruning phase

and two filters based on information theory, mRMR [Peng et al., 2005] and mIMR [Bontempi and Meyer, 2010]. The comparison is done as follows: for each data set and for each node (having at least a parent) the causal inference techniques return the ranking of the inferred parents. The ranking is assessed in terms of the average of Area Under the Precision Recall Curve (AUPRC) and a t-test is used to assess if the set of AUPRC values is significantly different between two methods. Note that the higher the AUPRC the more accurate is the inference method.

[9] Note that we use *Causal Explorer* here because, unlike `bnlearn` which estimates the entire adjacency matrix, it returns a ranking of the inferred causes for a given node.

The summary of the paired comparisons is reported in Table 9.3 for the D2C algorithm trained on the synthetic data only and in Table 9.4 for the D2C algorithm trained on both synthetic data and the 52 training data sets.

	GS	IAMB	IAMBnPC	interIAMBnPC	mRMR	mIMR
W-L	48-3 (32-0)	43-8 (21-0)	46-5 (26-0)	46-5 (25-0)	42-9 (17-0)	34-17 (12-0)

Table 9.3: D2C trained on synthetic data only: number of data sets for which D2C has an AUPRC (significantly (pval < 0.05)) higher/lower than the method in the column. W-L stands for Wins-Losses.

	GS	IAMB	IAMBnPC	interIAMBnPC	mRMR	mIMR
W-L	49-2 (36-0)	49-2 (27-0)	49-2 (32-0)	49-2 (32-0)	42-9 (17-0)	46-5 (19-1)

Table 9.4: D2C trained on synthetic data and 52 training data sets: number of data sets for which the D2C has an AUPRC (significantly (pval < 0.05)) higher/lower than the method in the column. W-L stands for Wins-Losses.

It is worthy to remark that

- the D2C algorithm is extremely competitive and outperforms the other techniques taken into consideration,
- the D2C algorithm is able to generalize to DAG with different number of nodes and different distributions also when trained only on synthetic data simulated on linear DAGs,
- the D2C algorithm takes advantage from the availability of more training data and in particular of training data related to the causal inference task of interest, as shown by the improvement of the accuracy from Table 9.3 to Table 9.4,
- the two filters (mRMR and mIMR) algorithm appears to be the least inaccurate among the state-of-the-art algorithms,
- though the D2C is initialized with the results returned by the mIMR algorithm, it is able to improve its output and to significantly outperform it.

9.5 Conclusion

Two attitudes are common with respect to causal inference for observational data. The first is pessimistic and motivated by the consideration that *correlation (or de-*

pendency) does not imply causation. The second is optimistic and driven by the fact that *causation implies correlation (or dependency)*. This paper belongs evidently to the second school of thought and relies on the confidence that causality leaves footprints in the form of stochastic dependency and that these footprints can be detected to retrieve causality from observational data. The results of the ChaLearn challenge and the preliminary results of this paper confirm the potential of machine learning approaches in predicting the existence of causality links on the basis of statistical descriptors of the dependency. We are convinced that this will open a new research direction where learning techniques may be used to reduce the degree of uncertainty about the existence of a causal relationships also in indistinguishable configurations which are typically not addressed by conditional independence approaches.

Further work will focus on 1) discovering additional features of multivariate distributions to improve the accuracy 2) addressing and assessing other related classification problems (e.g. predicting if a variable is an ancestor or descendant of a given one) 3) extending the work to partial ancestral graphs [Zhang, 2008] (e.g. exploiting the logical relations presented in Claassen and Heskes [2011]) extending the validation to real data sets and configurations with a still larger number of variables (e.g. network inference in bioinformatics).

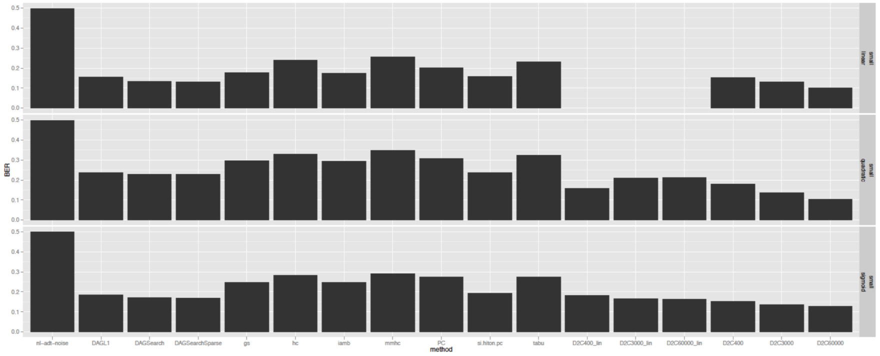

Fig. 9.3: Balanced Error Rate of the different methods for small size DAGs and three types of dependency (top: linear, middle: quadratic, bottom: sigmoid). The notation D2Cx stands for D2C with a training set of size *x* and where training and test sets are based on DAGs with the same type of dependency. The notation D2Cx_lin stands for D2C with a training set of size *x* based on DAGs with linear dependency only.

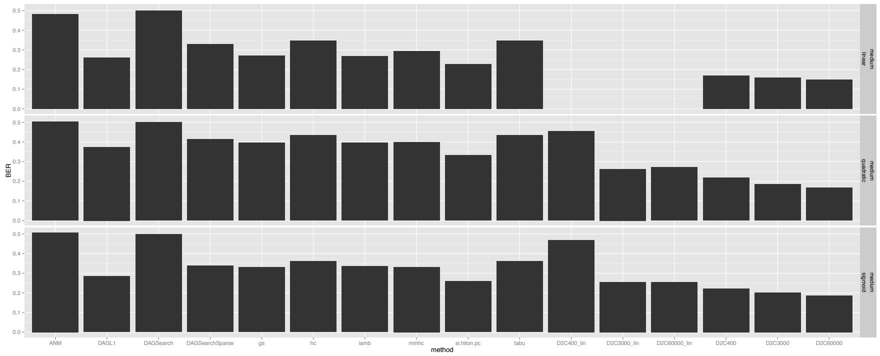

Fig. 9.4: Balanced Error Rate of the different methods for medium size DAGs and three types of dependency (top: linear, middle: quadratic, bottom: sigmoid). The notation D2Cx stands for D2C with a training set of size *x* and where training and test sets are based on DAGs with the same type of dependency. The notation D2Cx_lin stands for D2C with a training set of size *x* based on DAGs with linear dependency only.

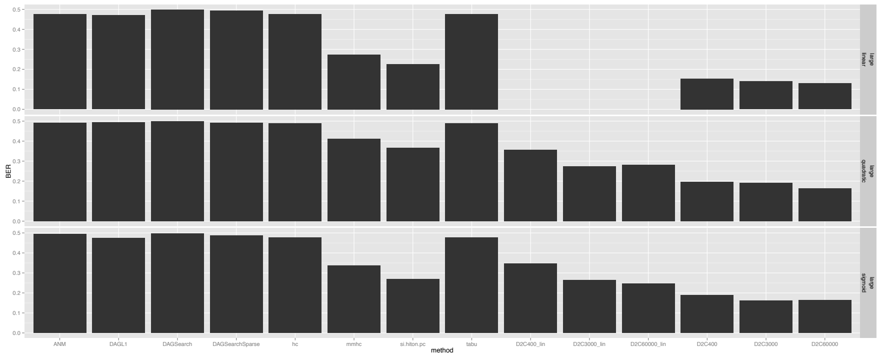

Fig. 9.5: Balanced Error Rate of the different methods for large size DAGs and three types of dependency (top: linear, middle: quadratic, bottom: sigmoid). The notation D2Cx stands for D2C with a training set of size *x* and where training and test sets are based on DAGs with the same type of dependency. The notation D2Cx_lin stands for D2C with a training set of size *x* based on DAGs with linear dependency only.

Acknowledgment

This work was supported by the ARC project "Discovery of the molecular pathways regulating pancreatic beta cell dysfunction and apoptosis in diabetes using functional genomics and bioinformatics" funded by the Communauté Française de Belgique and the BridgeIRIS project funded by INNOVIRIS, Brussels Region. The authors wishes to thank the editor and the anonymous reviewers for their insightful comments and remarks.

References

C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification. *Journal of Machine Learning Research*, 11:171–234, 2010.

C.F. Aliferis, I. Tsamardinos, and A. Statnikov. Causal explorer: A probabilistic network learning toolkit for biomedical discovery. In *Proceedings of METMBS*, 2003.

G. Bontempi and P.E. Meyer. Causal filter selection in microarray data. In *Proceedings of ICML*, 2010.

G. Bontempi, M. Birattari, and H. Bersini. Lazy learning for modeling and control design. *International Journal of Control*, 72(7/8):643–658, 1999.

G. Bontempi, B. Haibe-Kains, C. Desmedt, C. Sotiriou, and J. Quackenbush. Multiple-input multiple-output causal strategies for gene selection. *BMC Bioinformatics*, 12(1):458, 2011.

G. Bontempi, C. Olsen, and M. Flauder. *D2C: Predicting Causal Direction from Dependency Features*, 2014. URL http://CRAN.R-project.org/package=D2C. R package version 1.1.

T. Claassen and T. Heskes. A logical characterization of constraint-based causal discovery. In *Proceedings of UAI*, 2011.

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

R. Daly and Q. Shen. Methods to accelerate the learning of bayesian network structures. In *Proceedings of the UK Workshop on Computational Intelligence*, 2007.

P. Daniusis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Scholkopf. Inferring deterministic causal relations. In *Proceedings of UAI*, pages 143–150, 2010.

C. Dethlefsen and S. Højsgaard. A common platform for graphical models in R: The gRbase package. *Journal of Statistical Software*, 14(17):1–12, 2005. URL http://www.jstatsoft.org/v14/i17/.

N. Friedman, M. Linial, I. Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7, 2000.

D. Geiger, T. Verma, and J. Pearl. Identifying independence in bayesian networks. *Networks*, 20, 1990.

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.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

I. Guyon, C. Aliferis, and A. Elisseeff. *Computational Methods of Feature Selection*, chapter Causal Feature Selection, pages 63–86. Chapman and Hall, 2007.

PO Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Scholkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, pages 689–696, 2009.

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.

M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012. URL http://www.jstatsoft.org/v47/i11/.

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

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL http://CRAN.R-project.org/doc/Rnews/.

D. Margaritis. *Learning Bayesian Network Model Structure from Data*. PhD thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 2003.

G.J. McLaughlan. *Finite Mixture Models*. Wiley, 2000.

P.E. Meyer and G. Bontempi. *Biological Knowledge Discovery Handbook*, chapter Information-theoretic gene selection in expression data. IEEE Computer Society, 2014.

JM Mooij, O. Stegle, D. Janzing, K. Zhang, and B. Scholkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*, 2010.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669–710, 1995.

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

J.P. Pellet and A. Elisseeff. Using markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9:1295–1342, 2008.

H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency,max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

O. Pourret, P. Nam, and B. Marcot. *Bayesian Networks: A Practical Guide to Applications*. Wiley, 2008.

H. Reichenbach. *The Direction of Time*. University of California Press, Berkeley, 1956.

M. Schmidt, A. Niculescu-Mizil, and K. Murphy. Learning graphical model structure using l1-regularization paths. In *Proceedings of AAAI*, 2007.

Marco Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010. URL http://www.jstatsoft.org/v35/i03/.

S. Shimizu, P.O. Hoyer, A. Hyvrinen, and A.J. Kerminen. A linear, non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

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

A. Statnikov, M. Henaff, N.I. Lytkin, and C. F. Aliferis. New methods for separating causes from effects in genomics data. *BMC Genomics*, 13(S22), 2012.

I. Tsamardinos, CF Aliferis, and A Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of KDD*, pages 673–678, 2003a.

I. Tsamardinos, C.F. Aliferis, and A. Statnikov. Algorithms for large scale markov blanket discovery. In *Proceedings of FLAIRS*, 2003b.

I. Tsamardinos, LE Brown, and CF Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2010.

J. Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9: 1437–1474, 2008.