ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Inferring the Causal Decomposition under the Presence of Deterministic Relations.

Jan Lemeire[1,2], Stijn Meganck[1,2], Francesco Cartella[1],
Tingting Liu[1] and Alexander Statnikov[3]

1-ETRO Department, Vrije Universiteit Brussel, Pleinlaan 2, Brussels, Belgium
2-Interdisciplinary Institute for Broadband Technology (IBBT), Belgium
{jan.lemeire, stijn.meganck, francesco.cartella, tingting.liu}@vub.ac.be
3-Center for Health Informatics and Bioinformatics, New York University Medical Center, USA; alexander.statnikov@nyumc.org

Abstract. The presence of deterministic relations pose problems for current algorithms that learn the causal structure of a system based on the observed conditional independencies. Deterministic variables lead to information equivalences; two sets of variables have the same information about a third variable. Based on information content, one cannot decide on the direct causes. Several edges model equally well the dependencies. We call them equivalent edges. We propose to select among the equivalent edges the one with the simplest descriptive complexity. This approach assumes that the descriptive complexity increases along a causal path. As confirmed by our experimental results, the accuracy of the method depends on the chance of accidental matches of complexities.

1 Introduction

The goal of this work is to infer from observations the causal structure of a system under the presence of deterministic relations. Inferring causality consists of (i) detecting the causal influences between the variables, i.e. the modularity aspect, and (ii) detecting the orientation of these influences. The pioneering work of Spirtes, Glymour and Scheines [1] showed how the causal graph can be learned from the conditional independencies observed in the system. Consult this work for the basic concepts of independence-based structure learning used in this text. For inferring the causal orientation, independence-based algorithms rely on a pattern of (in)dependencies: $X \rightarrow Z \leftarrow Y$ is recognized by $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y|Z$, which define a v-structure. This remains true if some of the relations are deterministic.

Inferring the causal decomposition (modularity) on the other hand, relies on the *stochastic nature* of the relations among the variables. Indirect causes can be identified since they can be screened off from the effect through direct causes. If $X \rightarrow Y \rightarrow Z$ reflect the true causal structure, X and Z are dependent, but become independent when conditioned on Y. In the probabilistic case, Y and Z remain dependent when conditioned on X. This let us decide that Y and Z must be adjacent. If, on the other hand, $Y = f(X)$ for some deterministic function f, then also $Y \perp\!\!\!\perp Z|X$ holds. So Z depends on both X and Y, but both X and Y become independent when conditioned on the other. One cannot decide on whether $X - Z$ or $Y - Z$[1] form a direct causal relation. The solution we propose is to compare the descriptive complexities of $X - Z$ and $Y - Z$

[1] With an undirected edge we denote that we have not yet identified the orientation.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

($P(Z|X)$ and $P(Z|Y)$ respectively) and identify the one with the least complexity as the direct cause.

The complexity criterion is motivated by the property that causal relations express independent mechanisms of the system. In our example, $P(Y|X)$ and $P(Z|Y)$ represent the causal mechanisms. $P(Z|X)$ is 'generated' by $P(Y|X)$ and $P(Z|Y)$. It follows that $P(Z|X)$ will be more complex than $P(Z|Y)$, unless the complexities of $P(Y|X)$ and $P(Z|Y)$ cancel out when generating $P(Z|X)$. Consider the quadratic function for f: $Y = X^2$. Only if $P(Z|Y)$ has the form of a square root function, as for instance $Z = \sqrt{Y} + U$ with U an independent disturbance term, then $P(Z|X)$ will be significantly simpler than $P(Z|Y)$. Consider systems for which the causal mechanisms are chosen independently. It follows that a match is a matter of chance: an accidental cancellation of the complexities. Then the criterion applies. As we will show experimentally, the probability of a match will always be lower than 0.5 and decreases drastically with wider ranges of curve forms.

We first discuss how deterministic relation lead to information equivalences and equivalent edges, then we show how and when complexities can help us to decide on the true causal edges. We provide an empirical estimator and use it to validate the method on simulated data.

2 The problem

Property 1 *Given f a deterministic function:* $Y = f(\boldsymbol{X}) \Rightarrow Y \perp\!\!\!\perp Z | \boldsymbol{X}, \boldsymbol{U} \quad \forall Z, \boldsymbol{U} \in \boldsymbol{V}$

Note that single stochastic variables are denoted by capital letters, sets of variables by boldface capital letters. If a conditional independence of Property 1 does not follow from Markov, it leads to violations of the *intersection condition*, one of the necessary conditions for faithfulness [2]. The non-Markovian independencies can be characterized as information equivalences.

2.1 Information equivalences

We call $\boldsymbol{X}$ and $\boldsymbol{Y}$ *information equivalent* with respect to Z, which we call the *target variable*, when:

$$\boldsymbol{X}, \boldsymbol{Y} \not\perp\!\!\!\perp Z \mid \boldsymbol{W} \quad \& \quad \boldsymbol{X} \perp\!\!\!\perp Z \mid \boldsymbol{W}, \boldsymbol{Y} \quad \& \quad \boldsymbol{Y} \perp\!\!\!\perp Z \mid \boldsymbol{W}, \boldsymbol{X}. \tag{1}$$

where W is disjoint with $\boldsymbol{X}$ and $\boldsymbol{Y}$, and not containing Z. The interpretation is that knowledge of either $\boldsymbol{X}$ or $\boldsymbol{Y}$ is completely equivalent from the viewpoint of Z. We write an information equivalence as $Z|\boldsymbol{X}\ eq\ \boldsymbol{Y}$. The motivation is that $P(Z|\boldsymbol{Y}) = P(Z|f(\boldsymbol{X})) = P(Z|\boldsymbol{X})$ [3].

If for an information equivalence, no set $\boldsymbol{U}$ exists that screens of one of the equivalent sets from the target variable without also being information equivalent, we call it a *basic information equivalence*. To exclude special cases [3], we assume the following:

Assumption 1 *For a basic information equivalence, Eq. 1 holds for any $\boldsymbol{W}$ disjoint with $\boldsymbol{X}$ and $\boldsymbol{Y}$, and not containing Z.*

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

The next property shows that if a set $\boldsymbol{U}$ exists that screens off one of the equivalent sets from the target variable, it also screens off the other set.

Property 2 $Z|\boldsymbol{X}\ eq\ \boldsymbol{Y}:\ \ \boldsymbol{X}\perp\!\!\!\perp Z|\boldsymbol{U}\ \ \Leftrightarrow\ \ \boldsymbol{Y}\perp\!\!\!\perp Z|\boldsymbol{U}$

$$P(\boldsymbol{Y}|Z,\boldsymbol{U}) = \sum_{\boldsymbol{X}} P(\boldsymbol{Y}|\boldsymbol{X},Z,\boldsymbol{U}).P(\boldsymbol{X}|Z,\boldsymbol{U}) \tag{2}$$

$$= \sum_{\boldsymbol{X}} P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}).P(\boldsymbol{X}|\boldsymbol{U}) = P(\boldsymbol{Y}|\boldsymbol{U}) \tag{3}$$

The first factor of Eq. 2 leads to the first factor of Eq. 3 by the independence following from the information equivalence which holds for any $\boldsymbol{U}$ by Assumption 1. Z can be removed from the conditioning set of the second factor by the given independence $\boldsymbol{X}\perp\!\!\!\perp Z|\boldsymbol{U}$. ∎

The next property is necessary for the theorem.

Property 3 $Z|\boldsymbol{X}\ eq\ \boldsymbol{Y}\ \ \&\ \ \boldsymbol{X}\perp\!\!\!\perp\boldsymbol{Y}|\boldsymbol{U}\ \ \Rightarrow\ \ \boldsymbol{X}\perp\!\!\!\perp\boldsymbol{Z}|\boldsymbol{U},\boldsymbol{W}\ \ \ \forall\boldsymbol{W}$

Now we show that basic information equivalences lead to edges that equally well represent the distribution.

2.2 Equivalent edges

Definition 1 *Consider distribution P and two disjoint sets of edges $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$. Both sets are called equivalent edge sets if for any DAG G which is not Markovian for P:*

$$\begin{aligned} & G \cup \boldsymbol{E}_1 \ \textit{ is Markovian for } P \\ \Leftrightarrow\ & G \cup \boldsymbol{E}_2 \ \textit{ is Markovian for } P \end{aligned} \tag{4}$$

and that this is not true if we would remove an edge from $\boldsymbol{E}_1$ or $\boldsymbol{E}_2$.

Theorem 4 *Given a distribution P that can be modeled by a Markovian DAG. Consider $\boldsymbol{X}$ and $\boldsymbol{Y}$ a basic information equivalence for Z in P, then $\boldsymbol{E}_X = \cup_{X\in\boldsymbol{X}} X \to Z$ and $\boldsymbol{E}_Y = \cup_{Y\in\boldsymbol{Y}} Y \to Z$ form equivalent edge sets for P.*

Proof Assume G any DAG Markovian to P including $E_{\boldsymbol{X}}$ and $E_{\boldsymbol{Y}}$. Then the DAG $G_{/E_{\boldsymbol{X}}\cup E_{\boldsymbol{Y}}}$ (G without edges $E_{\boldsymbol{X}}$ and $E_{\boldsymbol{Y}}$) is not Markovian to P, since $\boldsymbol{X}$ and $\boldsymbol{Y}$ should be d-connected to Z when conditioned on all other variables (Assumption 1). We prove that if $G_{/E_{\boldsymbol{Y}}}$ is Markovian then $G_{/E_{\boldsymbol{X}}}$ is Markovian. The reverse is proven similarly. When each dependency in P is modeled by a d-connection in $G_{/E_{\boldsymbol{Y}}}$, this should also be modeled in $G_{/E_{\boldsymbol{X}}}$. We can limit us to dependencies that are represented in $G_{/E_{\boldsymbol{Y}}}$ but not in $G_{/E_{\boldsymbol{X}}\cup E_{\boldsymbol{Y}}}$ since dependencies present in the last one are by definition present in $G_{/E_{\boldsymbol{X}}}$. Assume $A\not\perp\!\!\!\perp B|\boldsymbol{C}$ and thus $A\not\perp\!\!\!\perp B|\boldsymbol{C}$ in $G_{/E_{\boldsymbol{Y}}}$, but not in $G_{/E_{\boldsymbol{X}}\cup E_{\boldsymbol{Y}}}$. This means that all active paths go through edges of $E_{\boldsymbol{X}}$. It follows that for one such edge $X_i \to Z$, $X_i\not\perp\!\!\!\perp Z|\boldsymbol{C}$ (*) must hold since otherwise the d-connection does not give a dependency. For this dependency $\boldsymbol{Y} \not\subseteq \boldsymbol{C}$ must hold (Assumption 1). It follows that $A\not\perp\!\!\!\perp B|\boldsymbol{C}$ in $G_{/E_{\boldsymbol{X}}}$:

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

path $Z \leftarrow Y_j - X_i - C$ is created which gives a d-connection due to the following. Since (1) at least one Y_j in $\boldsymbol{Y}$ is not in $\boldsymbol{C}$, and (2) all members of $\boldsymbol{X}$ are d-connected to all members of $\boldsymbol{Y}$. If there is a set d-separating them, it cannot be a subset of $\boldsymbol{C}$ since it would turn dependency (*) into an independency by Property 3. And, finally, (3) a v-structure in X_i on the path can be excluded. By the independencies $B \perp\!\!\!\perp Z | Y_i, \boldsymbol{C}$, $Y_i \perp\!\!\!\perp Z | \boldsymbol{X}, \boldsymbol{C}$ and $B \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{C}$ (v-structure) it follows that $Z \perp\!\!\!\perp B | \boldsymbol{C}$, which excludes the possibility of dependency $A \not\perp\!\!\!\perp B | \boldsymbol{C}$. ∎

3 The complexity criterion

In [4] we provided an extension to the PC algorithm to detect information equivalences based on the conditional independencies. The algorithm returns sets of equivalent edges. Because of the information equivalence, the amount of information that one variable conveys about another does not give us a criterion to decide upon adjacency. We introduce the *descriptive complexity of relationships* as a criterion to decide which one gives the direct causal relation. We define it very general with the conditional Kolmogorov complexity $K(x \mid y)$ which is defined as the length of the shortest program that given y as input prints x and then halts [5]. The complexity of the relationship between X and Y, written as $K(X \mid Y)$, can then be quantified by estimating $K(x^n \mid y^n)$, where x^n and y^n are the vectors of the observed data, with n the sample size.

3.1 Complexity Increase

The complexity criterion makes sense by the assumption that the complexities of the relations "do not decrease along a causal path".

Assumption 2 (Complexity Increase Assumption) *Given a set of variables **V** whose causal structure can be represented by a DAG G, for all disjoint subsets **X**, **Y**, **Z** of **V**:*

$$\boldsymbol{X} \perp \boldsymbol{Z} \mid \boldsymbol{Y} \; in \; G \;\Rightarrow\; K(\boldsymbol{X} \mid \boldsymbol{Z}) \geq K(\boldsymbol{X} \mid \boldsymbol{Y}). \tag{5}$$

3.2 Complexity extimator for structural equations containing continuous variables

A regression analysis is used for estimating $K(x^n \mid y^n)$. It seeks the most appropriate function that fits the data, such that the function minimizes

$$f_{min} = \underset{f \in \mathcal{F}}{argmin} \{K(f) + K(e^n)\}, \tag{6}$$

with $\mathcal{F}$ the set of admissible functions and e^n the error vector defined as $e_i = x_i - f(y_i)$ with i from 1 to n. The model class $\mathcal{F}$ is populated with the mononomials, polynomials and root functions up to degree 5, the inverse, the power, the square root and the step function. The description of the hypothesis then contains the values of the function's parameters, each needing d bits (the precision), and the function type, for which we count 1 byte for each operation (addition, subtraction, multiplication, division, power,

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

square root and logarithm) in the function[2]. A floating-point value is encoded with d bits, whereas an integer value i requires $\log(i)$ bits.

It is shown that the optimal precision d for each parameter is given by $d = 1/2 \log_2 n +$ c, with n the sample size and c some constant [6]. Hence

$$K(f) = \#parameters.\frac{\log_2(n)}{2} + 8.\#operations + K \tag{7}$$

with K a constant term that does not depend on f. Therefore it does not play any role in finding the minimal description. The second part of Eq. 6, $K(e^n)$, reflects the goodness-of-fit of the curve $Y = f(\mathbf{X})$. By choosing the normal distribution as probability distribution of the errors (the deviances of the data with respect to the curve), $K(e^n)$ equals the sum of squared errors:

$$K(e^n) = \sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{8}$$

A regression analysis thus has to minimize the sum of Eq. 7 and Eq. 8.

3.3 Linearity

If two variables X and Y are related by a linear bijection, the relation of X and Y with *any other* variable will be completely similar, qualitatively and quantitatively. Both variables contain the same information about any other variable and in the same form. So, in the absence of background knowledge and when no considering scaling, they represent equivalent quantities. The variables are indistinguishable, they are redundant and one could be removed from the data.

4 Experiments

We considered 100 artificial data sets generated by sets of randomly created structural equations. For each dataset, we first generate 5 variables $X_1, \ldots, X_5$ with a distribution randomly chosen from two options with equal probability: either the uniform distribution on $[0, 1]$ or a Gaussian mixture distribution. Then 10 variables $X_6, \ldots, X_{15}$ were defined according to the structural equation $X_i = f_i(X_j, \ldots, X_{j+k}) + \lambda_i E_i$. The variables $X_j, \ldots, X_{j+k}$ are chosen randomly from $X_1, \ldots, X_{i-1}$ ("the causally preceding variables"). We only used $k \geq 0$ when the linear function is chosen for f_i. For non-linear functions we only consider one dependent variable. λ_i has the probability of 0.5 to be zero, and is otherwise chosen uniformly between $[0, 0.1]$, where 0.1 means that the Gaussian noise term E_i has a standard deviation which is 10% of the maximal value of the function. We did experiments by choosing f_i randomly from a set of functions (the first column in the table below) and randomly chosen parameters. For the root, mononomial and polynomial functions, the maximal exponent is set to 5.

[2]This choice of description method attributes shorter description lengths for simpler function, but nevertheless is somewhat arbitrary. The objectivity of the Kolmogorov complexity is based on the *Invariance Theorem*. The shortest programs that output a given string written in different universal computer languages are of equal length *up to a certain constant* [5]. A complete objective measure does not exist.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Our inference method is applied for each basic information equivalence found in the data. We assumed that all conditional independencies are correctly identified. The inferred orientation is checked with the true orientation. When the complexities differed by less than 8 bits, no decision was taken. Except when the term of Eq. 8 is the same, then the complexities should differ by at least one bit. The experimental results are shown in the Table 1. The second column gives the total number of equivalences that were found, the last column gives the percentage of undecided edges that were actually correct inferences.

	Number	Correct	Undecided	Undec. OK
Quadratic and square root	630	51%	7%	53%
Monomials and root functions	819	75%	25%	73%
Polynomials	609	86%	32%	61%
Polynomials and linear functions	440	73%	48%	46%

Table 1: Accuracy results for the experiments with simulated data.

The first row results in equivalences with a 50% chance of complexity cancellation. This explains the accuracy. The following rows show what happens when the variation and the complexity of the functions increases. The presence of linear functions increases the number of indecisions, but also decreases the accuracy.

5 Conclusions

Deterministic relations generate conditional independencies not coming from Markov. It results in graphs that equally well represent the independencies of the distribution. The graphs can be characterized by equivalent edges; edges that can be interchanged. We proposed to compare the descriptive complexity of relations to choose among equivalent edges the one that reflects the direct causal relation. A better-than-random guess is provided for non-linear functions.

References

[1] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer Verlag, 2nd edition, 1993.

[2] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA, Morgan Kaufman Publishers, 1988.

[3] Jan Lemeire. *Learning Causal Models of Multivariate Systems and the Value of it for the Performance Modeling of Computer Programs*. PhD thesis, Vrije Universiteit Brussel, 2007.

[4] Jan Lemeire, Stijn Meganck, and Francesco Cartella. Robust independence-based causal structure learning in absence of adjacency faithfulness. In *Procs of European Workshop on Probabilistic Graphical Models (PGM), Helsinki, Finland*, 2010.

[5] Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 1997.

[6] J. Rissanen. *Stochastic Complexity in Statistical Enquiry*. World Scientific, Singapore, 1989.