ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

Exploring the causal order of binary variables via exponential hierarchies of Markov kernels

Xiaohai Sun[1] and Dominik Janzing[2]

1- Max Planck Institute for Biological Cybernetics
72076 Tübingen - Germany

2- Universität Karlsruhe (TH) - Institute for Algorithms and Cognitive Systems
76128 Karlsruhe - Germany

Abstract. We propose a new algorithm for estimating the causal structure that underlies the observed dependence among n ($n \geq 4$) binary variables $X_1, \ldots, X_n$. Our inference principle states that the factorization of the joint probability into conditional probabilities for X_j given $X_1, \ldots, X_{j-1}$ often leads to simpler terms if the order of variables is compatible with the directed acyclic graph representing the causal structure. We study joint measures of OR/AND gates and show that the complexity of the conditional probabilities (the so-called Markov kernels), defined by a hierarchy of exponential models, depends on the order of the variables. Some toy and real-data experiments support our inference rule.

1 Introduction

Discovering causal relationships is one of the most relevant challenges of science. Inferring causal relations from statistical data becomes more and more relevant as more and better data have become available in recent years. Until the early nineties, it was widely considered impossible to discover causal structures in observational data without using any controlled experiments. The seminal works of Pearl [1] and Spirtes et al. [2] showed that, under reasonable assumptions, it is possible to get hints on causal relations from non-experimental statistical data.

Their well-known approach for generating causal hypotheses, formalized by a directed acyclic graph (DAG), is based on the Markov condition and the faithfulness assumption: Among all graphs that contain enough causal arrows to explain *all* conditional statistical dependences, one prefers those structures which allow *only* these conditional dependences. One version of a causal inference algorithm based on these principles is the inductive causation (IC) algorithm[1] [1]. However, if few or no conditional independent relations are observed, a large set of structures is equivalent with respect to the implied conditional dependence and undistinguishable by this conventional approach. Additional inference rules are therefore desirable. The goal of our approach is to gain *additional* hints on causal directions from the *simplicity* or "plausibility" of the corresponding conditional probabilities (the so-called "Markov kernels").

[1] A refinement of the IC is implemented as the PC algorithm in TETRAD program (http://www.phil.cmu.edu/projects/tetrad/).

2 Principle of plausible Markov kernels

With respect to an order $X_1, \ldots, X_n$, the joint probability measure of n variables may be factorized into $P(x_1, \ldots, x_n) = P(x_1)P(x_2|x_1)\cdots P(x_n|x_1, \ldots, x_{n-1}) = \prod_{j=1}^{n} P(x_j|an_j)$. The shorthand $an_j := (x_1, \ldots, x_{j-1})$ denotes the values of all $j-1$ ancestors $AN_j := (X_1, \ldots, X_{j-1})$ of X_j. Obviously, any reordering $X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)}$, where $\pi \in \mathbf{S}_n$ is a permutation, defines another factorization. Moreover, if P satisfies the Markov condition with respect to a DAG $\mathcal{G}$, we can decompose the joint measure into $P(x_1, \ldots, x_n) = \prod_{j=1}^{n} P(x_j|pa_j)$, where pa_j is a tuple of values of all k_j parents of X_j in $\mathcal{G}$. We may consider the factorization above as the special case where $\mathcal{G}$ is the unique complete acyclic graph that corresponds to the ordering $X_1, \ldots, X_n$, i.e. $\mathcal{G}$ has arrows from each X_i to every X_j with $i < j$. The conditional probabilities $P(x_j|pa_j)$ are called the *Markov kernels* corresponding to $\mathcal{G}$. Likewise, we call $P(x_j|an_j)$ the Markov kernels corresponding to π.

Our assumption of plausible Markov kernels (short: plMK) states that the Markov kernels of real-world probability distributions tend to be simpler and smoother with respect to those complete graphs (orderings) which contain the true causal structure as a subgraph[2]. We define the most plMK as the unique solution of the following constrained entropy maximization. Let the value set $\mathcal{X}_j$ of each variable X_j be given and furthermore $P(X_1, \ldots, X_{j-1})$ be already known. Let α_j, β_{ij} denote the observed mean value of X_j and X_iX_j, respectively. Then the most plMK given the observed first and second moments is the conditional probability measure $P(X_j|X_1, \ldots, X_{j-1}) = P(X_j|AN_j)$ that maximizes the conditional entropy $\mathcal{H}(X_j|X_1, \ldots, X_{j-1})$ subject to $E(X_j) = \alpha_j$ and $E(X_iX_j) = \beta_{ij}$, for all $i \leq j$. This definition of the most plMK captures the linear interactions if the variables have the entire $\mathbb{R}$ as domain, which the inference principle of Shimizu et al. [3] is actually also based on. However, their causal inference principle is only justified for real-valued variables, since linear effects do not exist in the general case. For discrete/categorical or hybrid causal structures, our method obtains conditional probabilities which are smooth in an intuitive sense. Some examples of real-world data with various bounded continuous domains can be found in [4].

In the current paper, we focus on variables with binary domains. The motivation to study binary variables and the induced probability measure is that they allow us to define causal relations by some Boolean functions like OR and AND, which are simplified models for many causal relations in real life. Let the value set of every binary variable be $\{0, 1\}$. On can show that the most plMK of a binary X_j according to the definition above can be written as

$$P(X_j{=}1 \,|\, an_j) = \frac{1}{2} + \frac{1}{2}\tanh(\lambda + \sum_{i=1}^{j-1} \lambda_i x_i) \quad \text{for} \quad i = 1, \ldots, j-1\,.$$

[2]It is certainly hard to justify our assumption theoretically. Actually only extensive experiments with real-life data can really decide whether such kind of simplicity principle provides a reasonable causal inference rule or not.

The causal influence of each ancestor X_i $(i<j)$ on X_j can be characterized by the parameter λ_i. If λ_i negative, X_i has a repressive effect on the occurrence of X_j (*independent* of the value assignment of the other ancestors) and if λ_i positive, X_i is conducive to X_j. Such a unique separation into repressive and conductive variables should clearly be a feature of *the simplest* potential cause-effect relations. Based on this observation, it is natural to consider the most plMK as part of a hierarchy of exponential models as follows. Since the tanh function is invertible in the open interval $(0,1)$ we may represent every strictly positive Markov kernel of a binary X_j with ancestors AN_j by $P(X_j=1\,|\,an_j) = \frac{1}{2} + \frac{1}{2}\tanh(F_j(an_j))$ with $F_j(an_j) = \lambda + \sum_{i_1=1}^{j-1}\lambda_{i_1}x_i + \sum_{i_1,i_2=1}^{j-1}\lambda_{i_1i_2}x_{i_1}x_{i_2} + \cdots + \sum_{i_1,\ldots,i_{j-1}=1}^{j-1}\lambda_{i_1\ldots i_{j-1}}x_{i_1}\cdots x_{i_{j-1}}$. We define $\mathcal{K}_k^{(j)}$ as the set of conditional probability distributions $P(X_j|AN_j)$ for which all coefficients λ in F_j with more than k indices vanish and shall drop the superscript j when this will lead to no confusion. We obtain the hierarchy $\mathcal{K}_0 \subset \mathcal{K}_1 \subset \cdots \subset \mathcal{K}_{j-1}$. One can verify that the constrained entropy maximization leads to terms in $\mathcal{K}_k$ if the set of constraints is extended by terms up to the moments $E(X_{i_1}X_{i_2}\cdots X_{i_k})$. We will therefore consider the above hierarchy as a natural definition of the complexity of the Markov kernels and observe that our "most plMK" are in $\mathcal{K}_1$ which is the first non-trivial class, since for all kernels in $\mathcal{K}_0$ the variables AN_j do not influence X_j at all. We define $\mathcal{M}_1^{X_1,\ldots,X_n}$ as the set of joint measures on $X_1,\ldots,X_n$ for which all Markov kernels $P(x_j|an_j)$ are in $\mathcal{K}_1^{(j)}$. The *asymmetry* of the set $\mathcal{M}_1$ with respect to a reordering of variables is decisive in this paper. Based on the above assumption of plMK, the following are the two main steps of our plMK *causal order discovery* algorithm.

Step 1 According to each of the altogether $n!$ hypothetical orders $X_{\pi(1)}, \cdots, X_{\pi(n)}$, compute the most plMK $P(X_{\pi(1)}), \cdots, P(X_{\pi(n)}|X_{\pi(1)},\ldots,X_{\pi(n-1)})$ and then calculate the corresponding joint measure P_π.

Step 2 Evaluate the goodness of fit to given data within resulting $n!$ joint measures P_π and find out the orders corresponding to the winners.

We propose to apply the maximum log-likelihood approach or the so-called Scheffé tournament[3] to evaluate the goodness of fit to data in step 2.

3 OR/AND gates with random inputs

First, we would like to show that the Markov kernels describing OR/AND gates are both in the closure of the class $\mathcal{K}_1$. To see this, let binary $X_1,\ldots,X_{n-1}$ correspond to the input bits of an OR gate and X_n the output. The Markov kernel of X_n can be defined by $P(X_n=1\,|\,an_n) := 1-\prod_{i=1}^{n-1}(1-x_i)$. Defining $P_k(X_n=1\,|\,an_n) := \frac{1}{2}+\frac{1}{2}\tanh(-k+2k\sum_{i=1}^{n-1}x_i)$, we have $\lim_{k\to\infty}P_k(x_n|an_n) = P(x_n|an_n)$, i.e. $P(X_n|AN_n)$ is in $\mathcal{K}_1^{(n)}$. Now we consider the joint distribution $P(X_1,\ldots,X_n)$ that is generated by the OR gate when the inputs $X_1,\ldots,X_{n-1}$

[3]The Scheffé tournament is a kind of minimum distance estimate, see [5, 6]. The advantage of this method compared to the maximum log-likelihood approach is that the latter is more sensitive to small deviations in the regions of small probability.

are randomly chosen with uniform distribution, i.e. $P(x_j|an_j)=1/2$ for all $j<n$. Clearly the joint measure P is in $\mathcal{M}_1^{X_1,\ldots,X_n}$. Now we consider an ordering of the variables where the output is not at the end. Without loss of generality we consider the order $X_2,\ldots,X_n,X_1$. We have

$$P(X_1=1\,|\,x_2,x_3,\ldots,x_{n-1},X_n=1)=\begin{cases}1 & \text{for } x_2=x_3=\cdots=x_{n-1}=0\\ 1/2 & \text{otherwise}\end{cases} \quad (1)$$

$$\text{and}\quad P(X_1=1\,|\,X_2=\cdots=X_n=0)=0\,. \quad (2)$$

Note that the event $X_n=0$ and $X_i=1$ for some $i\in\{2,\ldots,n{-}1\}$ does not occur and the corresponding conditional probabilities need not to be specified. We will show that there is no Markov kernel in the closure of $\mathcal{K}_{n-3}$ that satisfies equation (1). We are particularly interested in the Markov kernel of X_1 since it depends on $n{-}1$ variables and is therefore the natural candidate for being the most complex Markov kernel. We write $P(X_1=1\,|\,x_2,\ldots,x_n)=\frac{1}{2}+\frac{1}{2}\tanh(F(x_2,\ldots,x_n)$, where F is an appropriate function. Define a function $\tilde{F}$ with $n-2$ arguments by $\tilde{F}(x_2,\ldots,x_{n-1}):=F(x_2,\ldots,x_{n-1},X_n=1)$. If the kernel (1) was in the closure of $\mathcal{M}_{n-3}$, there would exist a sequence F_k with polynomials of degree $n{-}3$ and a corresponding sequence $\tilde{F}_k$ of degree $n{-}3$ such that $\tilde{F}_k(x_2,\ldots,x_{n-1})$ tended to infinity for $x_2=x_3=\cdots=x_{n-1}=0$ and to zero otherwise. Elementary linear algebra arguments show that the space of polynomials of degree $n{-}3$ would then contain the element g with

$$g(x_2,\ldots,x_{n-1})=\begin{cases}1 & \text{for } x_2=x_3=\cdots=x_{n-1}=0\\ 0 & \text{otherwise}\end{cases}$$

This is however not true since the unique function g satisfying these constraints is given by $g(x_2,\ldots,x_{n-1})=\prod_{i=2}^{n-1}(1-x_i)$, which is a polynomial of degree $n{-}2$. The lower bound on the degree is tight because there is indeed a sequence of polynomials of degree $n-2$ that induce Markov kernels which satisfy the constraints (1) and (2) in the limit. The sequence $(F_k)_{k\in\mathbb{N}}$ of functions, given by $F_k(x_2,\ldots,x_n):=k\big(2(x_n-1)+\prod_{i=2}^{n-1}(1-x_i)\big)$, tends to $-\infty$ for $x_n=0$ and the induced Markov kernel satisfies therefore the constraint in equation (2). Moreover, conditions (1) are also satisfied.

This shows that the OR gate induces a joint measure that is in $\mathcal{M}_1$ when considered with respect to the *correct* causal order. By inverting input and output one can see that this is also true for an AND gate. The results above show that for $n\geq 4$ the set $\mathcal{M}_1$ is not invariant with respect to a reordering of n variables and kernels in $\mathcal{K}_1$ can lead to joint distributions defining kernels which are in $\mathcal{K}_{n-2}$ but not in $\mathcal{K}_{n-3}$. This implies that our proposed algorithm can in principle identify the output of an OR/AND gate as the effect and its random inputs as causes whenever the number of inputs is at least 3, provided that the number of data points in the sample is large enough to allow a reliable estimation of the joint measure.

4 Toy and real-world data experiments

To test our proposed algorithm, we sampled toy data with 3-bit inputs and 1-bit output of a noisy OR gate. The Markov kernel of an n-bit leaky noisy OR gate can be generalized by $P(X_{n+1} = 1 \,|\, x_1, \ldots, x_n) = (1 - r)(1 - q^{x_1 + \cdots + x_n}) + r$ with $q \in [0, 1]$ and[4] $r \in [0, 1]$. Repeated experiments have shown that in case of appropriate sample size[5] our algorithm can identify the output as the effect reliably.

In particular, we demonstrate here an experiment with a dataset of 10^5 points where the inputs are strongly correlated[6]. We observed in our example a quite balanced correlation structure[7]. We check that no conditional independence are present and the IC provides a complete undirected graph as result and is incapable of learning anything about causal directions. When applied our plMK algorithm, the most plMK yields 4 distinct joint measures P_i $(i = 1, \ldots, 4)$, depending on which of the 4 variables X_i appears at the end. According to both the Scheffé tournament and the log-likelihood score, P_4 achieves the best fit to the data. Hence we can indeed identify X_4 as output and obtain a useful hint about the true causal order. The positive results with OR/AND gates should not suggest that our method could also identify outputs/inputs generated by all imaginable logical gates. On the other hand, some complex gates like parity, for instance, are much more unlikely to be a model for causal relations in real-life.

A weakness of our algorithm is the 2^n computational dimension and the $n!$ search space. For large n, a straightforward implementation is very expensive, time consuming and thus only feasible on structures with few variables. The plMK algorithm could be nevertheless helpful in considering small dimension problem, e.g. exploring local structures, in particular, if there are very few constraints of independence available. Our results were positive in the sense that for large samples the algorithm has indeed generated reasonable causal structures even on data sets *without any independent constraints*. In the following, we make use of a real-world dataset to demonstrate how one can benefit from the advantages of both IC and plMK algorithm. We propose to combine them and consider our plMK algorithm as an extension for the conventional constraint-based IC algorithm. A pre-selection of causal hypotheses via IC may reduce the search space for plMK drastically. Our inference rule can distinguish between causal graphs that generate the same set of stochastic dependence. The conventional constraint-based approach tends to prefer directed graphs with small number of arrows. Our inference rule can additionally prefer those hypotheses where the causes influence their effect in a simple manner.

[4] q can be interpreted as specifying the probability of suppressing the input 1; r can be interpreted as a spontaneous inversion of the truth assignment at the output.

[5] For relatively independent inputs, we need a moderate sample size of $200 - 500$. In case of strongly correlated inputs, the sample size required is much larger.

[6] The variable X_1 is sampled according to an unbiased distribution. X_2 is with probability 0.4 given by the inverse of X_1 and with probability 0.6 given by uniform noise. X_3 is the output of a 2-bit noisy OR gate ($q = 0.2$, $r = 0.4$) with inputs $X_{1,2}$. Then we feed a 3-bit noisy OR ($q = 0.2, r = 0.3$) with $X_{1,2,3}$ as inputs and get X_4 as the output.

[7] It is "balanced" in the sense that all correlation coefficients are between $[0.12, 0.19]$.

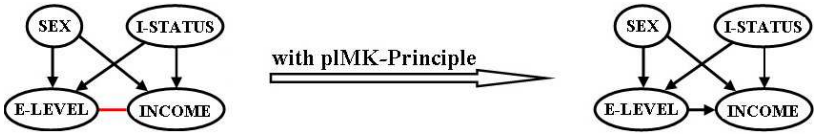

Fig. 1: Graphical representation of the causal hypothesis (CPS data) generated by PC (left), extended with the plMK algorithm (right).

We study causal relations between sex (SEX), immigrant status (I-STATUS), educational level (E-LEVEL), and annual income (INCOME)[8] with a real dataset from current population surveys (CPS) 1995, containing 112164 records (age 16 and over). We assume that gender, immigrant status and educational level affect the income, not vice versa. The causal hypotheses generated by the plMK algorithm were indeed consistent with this prior knowledge.

If we apply PC to CPS data, the result (Fig. 1, left) is incapable of making any statement about the orientation of the causal connection between E-LEVEL and INCOME, which means that a causal arrow from INCOME to E-LEVEL cannot be excluded. The plMK algorithm is here more specific since it allows only INCOME as the effect, i.e. the arrow from INCOME to E-LEVEL is excluded. Note, however, that plMK is in other respects less specific than PC since it cannot specify the structure of first three variables in the ordering. Recall, for instance, that the results of plMK did not show that SEX is not an effect of any other variables; the latter statement is only consistent with the class of preferred causal structures. Combining PC with the plMK algorithm we may then orient the edge from E-LEVEL to INCOME as done in Fig. 1, right. This shows that a combination of PC and plMK leads to the intuitively reasonable result that INCOME is not a cause of any other variable and SEX is not the effect.

References

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

[2] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search (Lecture Notes in Statistics)*. Springer-Verlag, New York, NY, 1993.

[3] S. Shimizu, A. Hyärinen, Y. Kano, and P.O. Hoyer. Discovery of Non-Gaussian Linear Causal Models Using ICA. In *Proc. of the 21st Conf. on Uncertainty in Artificial Intelligence*, pages 526–533, Edinburgh, UK, 2005.

[4] X. Sun, D. Janzing, and B. Schölkopf. Causal Inference by Choosing Graphs with Most Plausible Markov Kernels. In *Proc. of Ninth International Symposium on Artificial Intelligence and Mathematics*, Fort Lauderdale, FL, 2006.

[5] H. Scheffé. A Useful Convergence Theorem for Probability Distributions. *The Annals of Mathematical Statistics*, 18(5):434–438, 1947.

[6] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York, NY, 2001.

[8] The variables were transformed into binary ones, which stand for male/female, whether being native born in the US or not, whether having more than a Bachelor's degree or less, whether having an annual income of more than \$50K or less.