

Common Roots from High Stochastic Dependence

Bastian Steudel and Nihat Ay

MPI for Mathematics in the Sciences, Leipzig, Germany



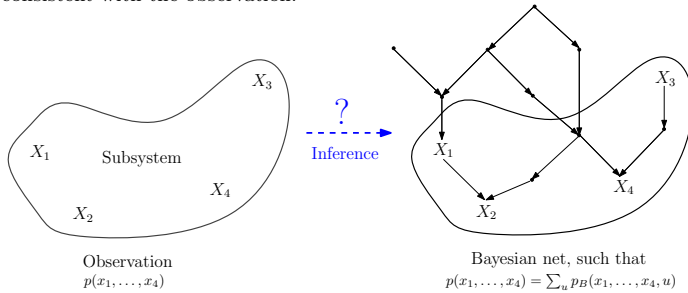
MAX-PLANCK-GESSELLSCHAFT

Setting

We use Bayesian nets as a formalization of the probabilistic and causal relations of a system and present a result that describes how information theoretic means can contribute to the causal inference process.

Task of Causal Inference

Starting from an observation of a subsystem in terms of a probability distribution of random variables, determine the class of Bayesian nets that are consistent with the observation.



Graphical Models and Information Theory

For a given directed acyclic graph G whose nodes are discrete random variables X_1, \dots, X_n , denote by $P(G)$ the family of joint probability distributions which factor according to G . Then for an arbitrary distribution p of the X_i

$$I_p(X_1, \dots, X_n) = D(p \parallel P(G)) + \sum_{i=1}^n I_p(X_i, \text{parents}(X_i)), \quad (*)$$

where

- $D(p \parallel P(G)) := \inf_{q \in P(G)} D(p \parallel q)$ is the distance of p from the family of distributions $P(G)$, measured in terms of *Kullback-Leibler divergence* $D(p \parallel q) = \sum p \log p/q$.

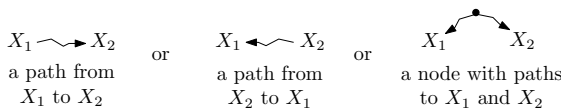
- I_p is the (generalized) mutual information

$$I_p(X_1, \dots, X_n) = D(p \parallel p(x_1) \otimes \dots \otimes p(x_n)).$$

General Question: Relation (*) only holds for distributions p defined on all nodes of the graph, so if the whole system has been observed. What can be said in cases of incomplete knowledge, that is if there are unobserved variables?

Inference of Common Roots

Common Roots of Two Variables: If X_1 and X_2 are stochastically dependent, then in every causal model we have:



Stochastic dependence of two variables \implies existence of a common ancestor

Definition. Let $\mathcal{X} = \{X_1, \dots, X_k\}$ be nodes of a Bayesian net B . A node which is an ancestor of at least c nodes of \mathcal{X} is called a **common root of order c** .

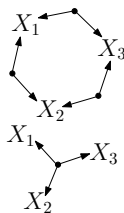
? What is a sufficient condition for the existence of common roots of more than two variables?

Example:

Two causal models of a subsystem $\{X_1, X_2, X_3\}$:

- no (conditional) independencies on the subsystem are enforced in either model
- model at the top has no common root of all three variables

\rightsquigarrow common roots can not be inferred from stochastic independencies alone.



Remarks

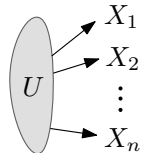
- Reformulation:** Common roots of order c can be inferred if

$$I_c := \frac{1}{c-1} \sum_{i=1}^n H(X_i) - H(X_1, \dots, X_n) > 0.$$

- The entropy of all common roots of order c is at least $\frac{c-1}{n-c+1} I_c$.

Example (Synchronized States): Assume that

- there are no causal interactions among the components of the observed subsystem and
- the mutual information of the subsystem is maximal and all variables have equal entropy



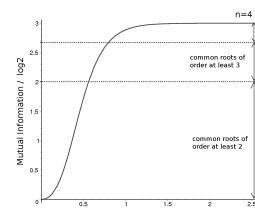
$$\rightsquigarrow H(U) \geq H(X_1) = \dots = H(X_n)$$

Example (Maximal Interaction): Distributions of binary variables of the form

$$p_a(x_1, \dots, x_n) \sim \exp(a x_n \dots x_k) \quad (a \in \mathbb{R}, x_i \in \{-1, 1\})$$

can be generated using only common roots of order two.

- Result holds also in the algorithmic causal setting introduced in [4] when substituting Kolmogorov complexity for entropy.



High mutual information is sufficient for common roots, but **not necessary**. Consider n random variables with values in $\{-1, 1\}$ and

$$p_a(x_1, \dots, x_n) \sim \exp\left(a \sum_{i=1}^n x_i x_j\right)$$

with $a \in \mathbb{R}$.

Theorem (Inference of Common Roots). Consider n random variables X_1, \dots, X_n taking on values in a finite set and a number c , $2 \leq c \leq n$. If the mutual information satisfies

$$I(X_1, \dots, X_n) > \left(1 - \frac{1}{c-1}\right) \sum_{i=1}^n H(X_i),$$

then there exist common roots of order c with positive entropy.

Future Work

1. How far can one go in characterizing causal models using only entropy-like quantities?

2. Consider the decomposition of mutual information into terms originating from the projection of p onto l -interaction spaces:

$$I(p) = \sum_{l=2}^k D(p^{(l)} \parallel p^{(l-1)})$$

Do causal interpretations for these interaction terms exist?

3. Derive heuristics for causal inference algorithms from information theoretic results as above.

References

- B. Steudel and N. Ay: *Inferring Common Causes from High Multi-Information*, submitted, 2008.
- N. Ay: *A Refinement of the Common Cause Principle*, SFI Working Paper, 2008.
- Campos L.: *A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests*, J. Mach. Learn. Res. Vol. 7, 2006.
- D. Janzing and B. Schoelkopf: *Causal inference using the algorithmic Markov condition*, arxiv:0804.3678, 2008.