Master Internship 2024

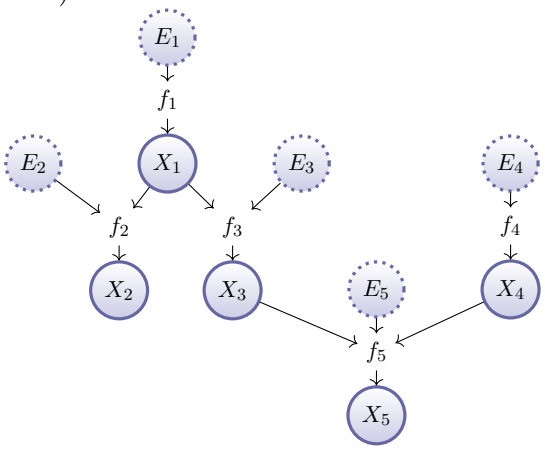
A Divide and Conquer strategy for Observational Causal Discovery

Topic: Causal Inference, Neural Network Identifiability of Causal Graph
Team: AO/TAU, CNRS − INRIA − LISN, Université Paris-Saclay
Advisors: Michèle Sebag (sebag@lri.fr)
Shuyu Dong (shuyu.dong@inria.fr)
Johanne Cohen (jcohen@lri.fr)
Duration: 6 months
Location: LISN, Paris-Saclay University – Building 660 – Shannon
Level: Master M2

1 Context & Motivation

The field of causal modeling is rapidly developing for a few years in Machine Learning, as it makes the promise to deliver robust models upon which interventions can soundly be based.

Causal models can be viewed as graphical models, relating each variable to its parents (immediate causes) augmented with an independent noise variable (reflecting all other unknown causes for this variable):



$$\begin{cases} X_1 = f_1(E_1) \\ X_2 = f_2(X_1, E_2) \\ X_3 = f_3(X_1, E_3) \\ X_4 = f_4(E_4) \\ X_5 = f_5(X_3, X_4, E_5) \end{cases}$$

Causal discovery aims to identify the causal graph $\mathcal{G}$ ($\mathcal{G}$ includes edge $X_i \to X_j$ if and only if X_i is an immediate cause of X_j) and the causal mechanisms f_i explaining a variable X_i from its parents $\mathrm{Pa}(X_i)$ and noise variable E_i:

$$X_i \sim f_i(\mathrm{Pa}(X_i), E_i).$$

The identification of $\mathcal{G}$ faces quite significant difficulties. Firstly, $\mathcal{G}$ is usually required to be a DAG (no cycle), to allow for its feedforward evaluation. Secondly, and mostly, the space of graphs (and DAGs) involving d variables is doubly exponential in d, and its identification boils down to a combinatorial optimization problem. Causal discovery algorithms rely on independence tests among variables, formulating constraints on edges based on independence and conditional independence tests, searching in the DAG space using score optimization (e.g. using Maximum Likelihood scores w.r.t. the observational data), and hybrid methods [1]. The current approaches notoriously hardly scale up beyond a few hundred variables.

2 Goal of the Internship

The main objective of the internship is to investigate how to tackle discovery using two foundational strategies in computer science and Machine Learning. The first one, "Divide and Conquer", aims to

break the set of variables into subsets (not necessarily disjoint) associated with each variable X_i and forming the Markov blanket of X_i, with $MB(X_i)$ including all immediate causes and effects of X_i and its spouses (i.e. variables sharing a same effect variable). The second is "Ensemble learning", where a set of (possibly partial) solutions to e.g. a classification or regression problem are combined to form a complete solution, different and expectedly better than its parts (the partial solutions).

Formally, it is shown that under some assumptions, the covariance matrix of the observational data yields the Markov blankets of each variable. The partial solutions are obtained by applying a fast algorithm on each data subset (involving all samples in the data, restricted to the variables in $MB(X_i)$), yielding a candidate subgraph relating X_i and the variables in $MB(X_i)$.

The challenge then is:

- Formulating the assumptions needed to make the problem affordable (for instance, the maximum number of causes for a variable is bounded; and/or if two variables are spouse, they cannot be cause of each other);
- Formulating the aggregation of the local graphs in terms of a constraint satisfaction problem (e.g. if $X_i \to X_j$ and $X_k \to X_k$ then X_i and X_k are spouse; if X_j and X_k are spouses they are not causes of each other);
- Examining the global graph solution of the above constraints (see for instance [**?**]), and
- Formulating how to repair the global graph as a machine learning problem.

3 Profile

The internship requires some curiosity about causality, besides excellent mathematical and machine learning skills + programming expertise. PhD grant available to continue the study.

References

[1] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.