Causal discovery

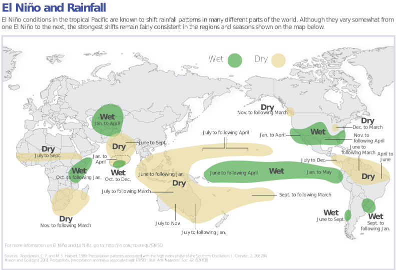
Gherardo Varando

IPL

09 May 2023

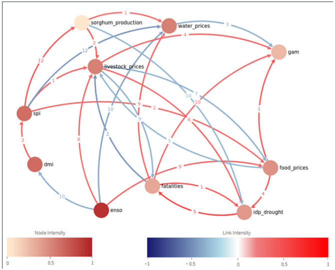
Causal Discovery

- Which are the regions most affected by ENSO?



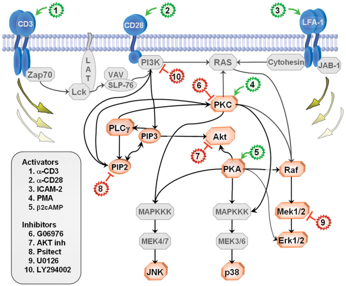
Causal Discovery

- Which are the regions most affected by ENSO?
- Which are the causal drivers and the causal relationships involving food insecurity?



Causal Discovery

- Which are the regions most affected by ENSO?
- Which are the causal drivers and the causal relationships involving food insecurity?
- Retrieve the reaction network between genes and proteins from single-cell data



Causal Discovery

- Learn the structure/graph of the causal relationships between variables

Causal Discovery

- Learn the structure/graph of the causal relationships between variables
- If only two variables are considered: bivariate causal discovery

Causal Discovery

- Learn the structure/graph of the causal relationships between variables
- If only two variables are considered: bivariate causal discovery
- it always require assumptions

Causal Discovery

- Learn the structure/graph of the causal relationships between variables
- If only two variables are considered: bivariate causal discovery
- it always require assumptions
- can use only observational data, or even interventions

Causal Discovery

- Learn the structure/graph of the causal relationships between variables
- If only two variables are considered: bivariate causal discovery
- it always require assumptions
- can use only observational data, or even interventions
- usually the output is not a complete description of the causal relationships (depends on the assumptions, the type of data etc..)

Causal discovery methods

- Constrained Based, Score based, Asymmetry based
- Methods for i.i.d. data or for time-series
- Bivariate vs multivariate
- Observational and/or Interventional data

In this session

- Causal discovery for BN and SCM
- Causal discovery taxonomy
- Some methods
- Examples

We will see only methods that employ observational data . . . thus we will need assumptions

Structural learning for BN

Estimate the structure of a BN from observational data. That is, recovering of G from (i.i.d.) observations sampled from a probability P such that (G, P) is a BN

- it is a model selection problem
- no causality involved (yet)
- Problem: a probability P can be *comaptible* with multiple BN!!

Markov equivalence class

- $\mathcal{M}(G) = \{P$ s.t. satisfies global (or local) Markov w.r.t. $G\}$

Markov equivalence class

- $\mathcal{M}(G) = \{P$ s.t. satisfies global (or local) Markov w.r.t. $G\}$
- If $\mathcal{M}(G_1) = \mathcal{M}(G_2)$ then we say that G_1 and G_2 are Markov equivalent. This means G_1 and G_2 represent the same set of d-separations and thus entails the same set of conditional independence statements

Markov equivalence class

- $\mathcal{M}(G) = \{P \text{ s.t. satisfies global (or local) Markov w.r.t. } G\}$
- If $\mathcal{M}(G_1) = \mathcal{M}(G_2)$ then we say that G_1 and G_2 are Markov equivalent. This means G_1 and G_2 represent the same set of d-separations and thus entails the same set of conditional independence statements
- A **Markov equivalence class** contains all graph which are Markov equivalent

Markov equivalence class

- $\mathcal{M}(G) = \{P \text{ s.t. satisfies global (or local) Markov w.r.t. } G\}$
- If $\mathcal{M}(G_1) = \mathcal{M}(G_2)$ then we say that G_1 and G_2 are Markov equivalent. This means G_1 and G_2 represent the same set of d-separations and thus entails the same set of conditional independence statements
- A **Markov equivalence class** contains all graph which are Markov equivalent

Theorem (Verma and Pearl [1991])

Two DAGs G_1 and G_2 are Markov equivalent if and only if they have the same skeleton and the same immoralities

Markov equivalence class

- $\mathcal{M}(G) = \{P$ rkov w.r.t. $G\}$
- If $\mathcal{M}(G_1) =$ G_2 are Markov equivalent. he same set of d-separations conditional independenc

Using observational data alone, Can we hope to learn something more than the Markov equivalence class?

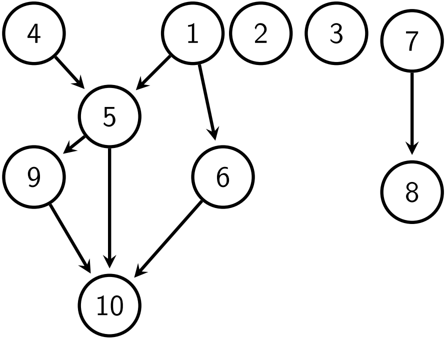
- A **Markov equivalence class** contains all graph which are Markov equivalent

Theorem (Verma and Pearl [1991])

Two DAGs G_1 and G_2 are Markov equivalent if and only if they have the same skeleton and the same immoralities

Theorem (Verma and Pearl [1991])

Two DAGs G_1 and G_2 are Markov equivalent if and only if they have the same skeleton and the same immoralities



we can represent the Markov equivalence class with a partial directed acyclic graph (PDAG)

Taxonomy of methods for BN

- Constrained based: PC algorithm

Taxonomy of methods for BN

- Constrained based: PC algorithm
- Score based or score+search: GES, tabu search, hill-climbing

Taxonomy of methods for BN

- Constrained based: PC algorithm
- Score based or score+search: GES, tabu search, hill-climbing
- hybrid: max-min-hill-climbing (mmhc)

Constrained based

- Start from the full dependency graph

Glymour et al. [2019], Spirtes et al. [2000]

Constrained based

- Start from the full dependency graph
- prune edges by iteratively test (conditional) independences

Glymour et al. [2019], Spirtes et al. [2000]

Constrained based

- Start from the full dependency graph
- prune edges by iteratively test (conditional) independences
- use some rules to orient edges

Glymour et al. [2019], Spirtes et al. [2000]

Constrained based

- Start from the full dependency graph
- prune edges by iteratively test (conditional) independences
- use some rules to orient edges
- obtain a PDAG (partial directed acyclic graph) representing the Markov equivalence class

Glymour et al. [2019], Spirtes et al. [2000]

The PC algorithm starts from a fully connected undirected graph and consists of three phases:

1. The *skeleton phase* uses statistical (conditional) independence tests to infer the adjacencies of the underlying causal graph. If two variables X and Y are found to be independent conditional on a (possibly empty) set of variables $\mathbf{Z}$, then the edge between X and Y is removed.
2. The *collider orientation phase* then orients all *collider motifs*, that is, motifs of the form $X \to Y \leftarrow Z$ where X and Z are non-adjacent. These orientations can be inferred because collider motifs impose a particular pattern of (conditional) (in-)dependencies.
3. The *orientation phase* finally uses graphical rules [Meek, 1995] to infer the orientation of as many remaining unoriented edges as possible using the acyclicity assumption and the fact that all colliders have been found in the previous step.

Score based

- find the **best scoring graph** with respect to some scoring criteria (e.g. BIC)

Chickering [2002]

Score based

- find the **best scoring graph** with respect to some scoring criteria (e.g. BIC)
- space of DAGs is huge .... we need heuristic searches

Chickering [2002]

Score based

- find the **best scoring graph** with respect to some scoring criteria (e.g. BIC)
- space of DAGs is huge .... we need heuristic searches
- hill-climbing, tabu search

Chickering [2002]

Score based

- find the **best scoring graph** with respect to some scoring criteria (e.g. BIC)
- space of DAGs is huge .... we need heuristic searches
- hill-climbing, tabu search
- GES (greedy equivalent search) in the space of CPDAG (the space of Markov equivalence classes) [under some assumpotions it is proven to recover the true Markov equivalence class]

Chickering [2002]

Causal discovery methods

- Constrained Based, Score based, Asymmetry based
- Methods for i.i.d. data or for time-series
- Bivariate vs multivariate
- Observational and/or Interventional data

Asymmetry based methods

- Use assumptions on the SCM to infer causal structure

Asymmetry based methods

- Use assumptions on the SCM to infer causal structure
- Can distinguish graphs in the same Markov equivalence class

Asymmetry based methods

- Use assumptions on the SCM to infer causal structure
- Can distinguish graphs in the same Markov equivalence class
- Developed especially in the bivariate setting (but also in multivariate case)

Asymmetry based methods

- Use assumptions on the SCM to infer causal structure
- Can distinguish graphs in the same Markov equivalence class
- Developed especially in the bivariate setting (but also in multivariate case)
- example, Linear non-Gaussian acyclic model (LiNGAM) [Shimizu et al., 2006]

Time series

- Causal discovery in time series can be *easier* then i.i.d. case
- The *arrow of time* helps in orienting some edges

Time series

- Causal discovery in time series can be *easier* then i.i.d. case
- The *arrow of time* helps in orienting some edges
- Granger causality, convergent cross-mapping (CCM), PC-based methods, score-based methods, (VAR)LinGAM, TiMINO, ...

Granger causality

Granger causality

$A(t)$ is linearly Granger-cause of $B(t)$ with respect to a fixed time-lag m if the null hypothesis $\{\alpha_\ell = 0 \text{ for } \ell = 1, \ldots, m\}$ is rejected for the linear AR model

$$B(t) = \sum_{\ell=1}^{m} \beta_\ell B(t-\ell) + \sum_{\ell=1}^{m} \alpha_\ell A(t-\ell) + \beta_0 + \epsilon.$$

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov): 507–554, 2002.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Christopher Meek. Causal Inference and Causal Explanation with Background Knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Uai'95, page 403–410, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 255–270, 1991.