Causal discovery with Python

In this Lab you will explore several causal discovery methods and valuate them in differnt ways. We shall consider both synthetic data and real data

Synthetic data - (Lunceford & Davidian 2004)

We will perform the same simulations proposed in Lunceford & Davidian 2004.
Generative model
The response variable Y is generated according the following equation:

$$Y = \nu_0 + \nu_1 X_1 + \nu_2 X_2 + \nu_3 X_3 + \nu_4 A + \xi_1 V_1 + \xi_2 V_2 + \xi_3 V_3 + \varepsilon, \varepsilon \sim \mathcal{N}(0,1),$$

where $\nu = (\nu_0, \nu_1, \nu_2, \nu_3, \nu_4)^T = (0, -1, 1, -1, 2), \xi = (\xi_1, \xi_2, \xi_3) = (-1, 1, 1)$. The covariates are distributed as $X_3 \sim Bernoulli(0.2)$, and conditionnaly on X_3

- If $X_3 = 0$, $V_3 \sim Bernoulli(0.25)$ and $(X_1, V_1, X_2, V_2)^T \sim \mathcal{N}(\tau_0, \Sigma)$
- If $X_3 = 1$, $V_3 \sim Bernoulli(0.75)$ and $(X_1, V_1, X_2, V_2)^T \sim \mathcal{N}(\tau_1, \Sigma)$

with

$$\tau_0 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix}, \tau_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0.5 & -0.5 & -0.5 \\ 0.5 & 1 & -0.5 & -0.5 \\ -0.5 & -0.5 & 1 & 0.5 \\ -0.5 & -0.5 & 0.5 & 1 \end{pmatrix}$$

and the treatment A is generated as a Bernoulli of the propensity score

$$e(X, \beta) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3)}$$

with $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T = (0, 0.6, -0.6, 0.6)$.
The data generating process is implemented in the Python notebook given in Moodle.

1. Try several causal discovery methods : PC algorithm, LinGAM and ANM on this synthetic dataset
2. Can you recover the true causal graph that was simulated?
3. Evaluate the inferred graph and compare it to ground truth?
4. Are the discovry methods you usd robust with respect to noise? to subsampling?

Causal discovery with Python

A case study based on real data : Hotel Booking Cancellations

We shall study an example from the markting domain extracted from the DoWhy library about Hotel Booking Cancellations. We recall the description of the dataset

- Booking information for a city hotel and a resort hotel taken from a real hotel in Portugal
- Includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things
- All personally identifying information has been removed from the data

The reference of the study case was givn in Lecture 4 : `https://www.sciencedirect.com/science/artic`

1. Try several causal discovery methods : PC algorithm, LinGAM and ANM on this synthetic dataset
2. Can you recover the causal graph that was given in Lecture 4?
3. Evaluate the inferred graph and compare it to ground truth?
4. Impact of a misspecification of the causal graph for the estimation of effects as done in Lecture 4