

Causal Inference Lab

1 Objective

In this lab, we estimate the causal effect of a treatment A on an outcome Y when the causal DAG is unknown using two different datasets "data1.csv" and "data2.csv".

2 Datasets

2.1 Dataset 1:

A set of continuous variables: Mental Health (A), Immune Function (Y), Sleep, Comorbidity Index, Exercise, Inflammation, Healthcare-seeking, Cortisol, Overall health self-assessment, Socioeconomic status, Quality of life score (EQ-5D).

2.2 Dataset 2:

Binary exposure and outcome: Smoking during pregnancy (A), Pre-term birth (Y), and continuous: Previous births, Maternal age, Cortisol.

3 Tasks

For both datasets, the causal DAG is unknown. The goal of the lab is therefore to follow the pipeline below:

1. Learn a CPDAG using the PC algorithm with the Fisher-Z test.
2. Identify, when possible, a valid adjustment set for estimating the effect of A on Y using the back-door criterion.
3. Estimate the causal effect using G-computation:
 - for the first dataset, use a linear regression model;
 - for the second dataset, use a non-linear classifier, e.g., a random forest.

4 R vs Python

The lab can be completed either in R or Python. For Python users, we recommend using the `pyciphod` package; installation instructions and the required modules are provided below. For R users, we encourage you to explore and identify the appropriate packages to perform the different steps of the analysis.

5 Python setup

To install the package:

```
python -m pip install "git+https://github.com/CIPHOD/pyCIPHOD"
```

Imports:

```
import pandas as pd
import numpy as np

from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestClassifier

from pyciphod.causal_estimation import GComputation
from pyciphod.causal_reasoning import back_door_criterion
from pyciphod.causal_discovery import PC
```

Dataset 1 true causal effect: -0.014

Dataset 2 true causal effect: 0.3