Exercise Sheet: Causal Graph Analysis

1 Path Analysis

Consider the following DAG: $A \to B \to D \leftarrow C \leftarrow E$.

1. Identify the Structures:
   - What is the relationship between A, B, and D?
   - What is the relationship between B, D, and C?
2. Independence Testing:
   - Are A and E d-separated? Why or why not?
   - Does conditioning on D make B and C independent or dependent? Explain.

2 The Markov Blanket

Imagine a medical study where we observe:

- Diet (D): Whether a patient eats healthily.
- Exercise (E): Physical activity levels.
- Fitness Score (F): A combined metric.
- Blood Pressure (B): The outcome.

The graph is $D \to F$, $E \to F$, and $F \to B$.

1. What is the Markov Blanket of F?
2. If we condition on F, does information flow between D and E?
3. If we only observe D and B, can we identify the causal effect of D on B?

3 Lab Exercise: Causal Discovery with PC and FCI, with Python

3.1 Objectives

- Understand the difference between Causal Sufficiency and Latent Confounding.
- Implement the PC and FCI algorithms using the causal-learn library.
- Compare independence tests for Continuous (Gaussian) vs. Discrete (Categorical) data.

3.2 The Case Study: The Lung Cancer Model

We will investigate the relationship between environmental factors, habits, and cancer (K. B. Korb, A. E. Nicholson. Bayesian Artificial Intelligence):

A patient has been suffering from shortness of breath (called dyspnoea) and visits the doctor, worried that he has lung cancer. The doctor knows that other diseases, such as tuberculosis and bronchitis, are possible causes, as well as lung cancer. She also knows that other relevant information includes whether or not the patient is a smoker (increasing the chances of cancer and bronchitis) and what sort of air pollution he has been exposed to. A positive X-ray would indicate either TB or lung cancer.

The True Underlying DAG relate the following variables:

- Pollution (P): Exogenous cause.
- Smoking (S): Exogenous cause.
- Cancer (U): A hidden/latent variable caused by both Pollution and Smoking.
- X-ray (C1): Result influenced by the presence of Cancer.
- Dyspnoea (C2): Shortness of breath influenced by the presence of Cancer.

Crucial Note: In our dataset, the variable Cancer (U) is unobserved. We only see the causes (Pollution, Smoker) and the symptoms (X-ray, Dyspnoea).

3.3 Part 1: Continuous Case (Gaussian)

In this section, we assume all health metrics are measured on a continuous scale with Gaussian noise.

Step 1.1: Generate a dataset of $N = 5000$ independent samples as follows:

- $P \sim \mathcal{N}(0, 1)$
- $S \sim \mathcal{N}(0, 1)$
- $U = 0.6 \cdot P + 0.8 \cdot S + \mathcal{N}(0, 0.5)$
- $C1 = 0.7 \cdot U + \mathcal{N}(0, 0.4)$
- $C2 = 0.9 \cdot U + \mathcal{N}(0, 0.4)$

Step 1.2: Using `causal-learn`, compare the PC and FCI algorithms on the generated data (removing U, which is not observed) and visualize the results.

3.4 Part 2: Discrete Case (Categorical)

In medical reality, variables are often discrete. In particular, let us consider the following binary discretization:

- $P \in \{\text{low}, \text{high}\}$
- $S \in \{\text{T}, \text{F}\}$
- $U \in \{\text{T}, \text{F}\}$
- $C1 \in \{\text{pos}, \text{neg}\}$
- $C2 \in \{\text{T}, \text{F}\}$

This can be modelled by using Bernoulli distributions instead of Gaussian, as reported below:

```
A = np.random.binomial(1, 0.3, n) # Low/High Pollution
B = np.random.binomial(1, 0.5, n) # Non-smoker/Smoker

# Probabilistic logic for Cancer (U)
prob_U = 0.1 + 0.3 * A + 0.5 * B
U = np.random.binomial(1, np.clip(prob_U, 0, 1), n)

# X-ray and Dyspnoea results based on Cancer
C1 = np.random.binomial(1, np.where(U == 1, 0.8, 0.1), n)
C2 = np.random.binomial(1, np.where(U == 1, 0.7, 0.2), n)
```

Step 2.1: Generate a new dataset ($N = 5000$ samples) according to the above Bernoulli distribution.

Step 2.2: When data is discrete, the Fisher-Z test is no longer valid. We must switch to the Chi-Squared (G^2) test. Modify your PC and FCI call to handle the discrete data file with a correct independence test, and visualize the result.

3.5 Lab Questions

1. In the PC output, how is the relationship between Xray and Dyspnoea represented? Does PC suggest one causes the other, or that they are just correlated? Why is this technically "wrong" given the true DAG?

2. Look at the FCI output for the edge between Xray and Dyspnoea. Do you see a bidirected edge ($\leftrightarrow$) or circles at the ends of the edges (o-o)? What does this notation tell a researcher about the existence of Cancer (U)?

3. In the true DAG, are Pollution and Smoker independent? Does conditioning on the symptoms (Xray or Dyspnoea) make them dependent?

4. Check your algorithm outputs: did the orientation of the arrows correctly reflect this collider structure?

5. Compare the stability of the graphs. Did the discrete version require a larger sample size (N) to find the same structures as the Gaussian version?