

Causality and Graphical Models — Tutorial booklet (M2 GENIOMHE)

Student version

Christophe Ambroise

Table of contents

Introduction	3
I Tutorial sheets	4
1 Session 1 — Causality and directed graphical models	5
Objectives	5
Theory	5
Practice	6
2 Session 2 — Biological networks	8
Objectives	8
Theory	8
Practice	8
3 Session 3 — Interventions, Mendelian randomisation and causal discovery	10
Objectives	10
Theory	10
Practice	11
Appendices	12
A Key notions	12
Session 1 — Causality and directed graphical models	12
Session 2 — Biological networks	12
Session 3 — Interventions and causal discovery	12

Introduction

This booklet accompanies the three tutorial sessions of the course Causality and Graphical Models for the M2 GENIOMHE. Its exercises are taken from the booklet of the M2 Data Science course.

Each sheet has a **theory** part (pen and paper) and a **practice** part in Python. The practice exercises use the course environment, created at the root of the course with

```
uv sync
```

It provides numpy, pandas, scikit-learn, networkx, pgmpy, causal-learn, lingam and otlingam.

Data files are in `data/`: `sachs.csv` (flow cytometry, Sachs *et al.*, 2005) and `raw_ndhB.csv`, `raw_ndhD.csv` (chloroplast RNA editing, Guilcher *et al.*, 2021).

This version contains the exercises only.

Part I

Tutorial sheets

1 Session 1 — Causality and directed graphical models

Objectives

This sheet uses the notions of session 1: Simpson’s paradox and confounding, DAG factorisation, d-separation, explaining away.

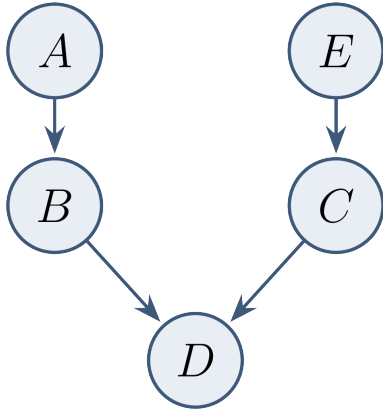
Theory

Exercise 1.1. Simpson’s paradox: smoking and sprint times. In a study of 100 m sprint times, smokers appear faster on average than non-smokers, yet non-smokers are faster within each gender. The mean times (in seconds) and the group sizes are:

Group	Smokers: mean	Smokers: size	Non-smokers: mean	Non-smokers: size
Men	11.5	80	11.0	20
Women	13.5	20	13.0	80

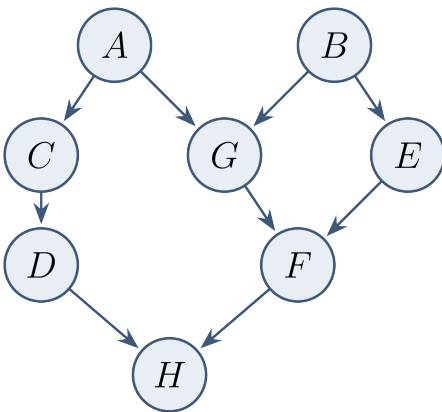
1. Compute the overall mean time of smokers and of non-smokers. Check that the paradox occurs.
2. Write each overall mean as a weighted average of the stratified means. Which weights differ between the two groups?
3. Draw a causal DAG linking gender G , smoking S and time T . Which variable is a confounder?
4. Which comparison answers the question “does smoking slow a runner down”? Compute the average causal effect $\mathbb{E}[T \mid do(S = 1)] - \mathbb{E}[T \mid do(S = 0)]$ using the gender distribution of the whole population.

Exercise 1.2. Factorisation of a gene network. Five genes A, B, C, D, E are either expressed (1) or not (0). Their regulation is described by the DAG below.



1. Write the factorisation of the joint distribution $p(a, b, c, d, e)$ implied by the DAG.
2. List the parameters of the model and count them. Compare with the number of parameters of an unrestricted joint distribution of five binary variables.
3. Give a topological order of the nodes. Is it unique?
4. Describe an algorithm that draws one sample (a, b, c, d, e) from the model using only Bernoulli draws (ancestral sampling).

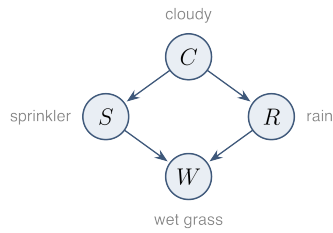
Exercise 1.3. d-separation in an eight-node DAG.



1. List all colliders of the graph.
2. Using d-separation, list all variables that are independent of A given D .
3. Is A independent of B ? Of B given H ? Justify with paths.

Practice

Exercise 1.4. The sprinkler network. Four binary variables: cloudy C , sprinkler S , rain R , wet grass W .



$P(C = 1) = 0.5$; $P(S = 1 \mid C = 1) = 0.1$, $P(S = 1 \mid C = 0) = 0.5$; $P(R = 1 \mid C = 1) = 0.8$, $P(R = 1 \mid C = 0) = 0.2$; $P(W = 1 \mid S + R = k) = 0.1, 0.9, 0.99$ for $k = 0, 1, 2$.

1. Simulate $n = 5000$ observations by ancestral sampling and store them in a **pandas** data frame.
2. Estimate the conditional probability tables by counting, and compare with the true values.
3. Compute exactly, by summing the joint distribution, $P(R = 1)$, $P(R = 1 \mid W = 1)$ and $P(R = 1 \mid W = 1, S = 1)$. Compare with the empirical frequencies. Interpret in terms of explaining away.
4. With `pgmpy.estimators.CITests.chi_square`, test $S \perp R$, $S \perp R \mid C$ and $S \perp R \mid \{C, W\}$. Relate the results to d-separation.

```

import numpy as np, pandas as pd
rng = np.random.default_rng(0)
n = 5000
C = rng.random(n) < 0.5
S = rng.random(n) < np.where(C, 0.1, 0.5)
...

```

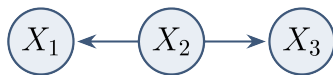
2 Session 2 — Biological networks

Objectives

This sheet uses the notions of session 2: precision matrix and partial correlations, graphical lasso, PC algorithm, on a protein signalling data set.

Theory

Exercise 2.1. Reading a precision matrix. Consider the DAG $X_1 \leftarrow X_2 \rightarrow X_3$ with Gaussian conditional distributions.



1. Which conditional independence does the DAG encode? Which undirected graph has the same independences?
2. Which of the following matrices could be the precision matrix K of (X_1, X_2, X_3) ? Justify (two conditions are needed).

$$A = \begin{pmatrix} 9 & 3 & 1 \\ 3 & 9 & 3 \\ 1 & 3 & 9 \end{pmatrix}, \quad B = \begin{pmatrix} 8 & -3 & 1 \\ -3 & 9 & -3 \\ 1 & -3 & 8 \end{pmatrix}, \quad C = \begin{pmatrix} 9 & 3 & 0 \\ 3 & 9 & 3 \\ 0 & 3 & 9 \end{pmatrix}, \quad D = \begin{pmatrix} 9 & -3 & 0 \\ -3 & 10 & -3 \\ 0 & -3 & 9 \end{pmatrix}.$$

3. For the admissible matrices, compute the partial correlation of X_1 and X_2 given X_3 .

Practice

Exercise 2.2. Protein signalling network from flow cytometry. The file `data/sachs.csv` contains $n = 7466$ single-cell measurements of 11 phosphorylated proteins and phospholipids (Sachs *et al.*, 2005, *Science*), as distributed with *The Elements of Statistical Learning*.

1. Load the data, look at the marginal distributions, and apply the transformation $x \mapsto \log(x - \min x + 1)$ followed by standardisation.
2. Fit `GraphicalLassoCV`. How many edges are selected?
3. Compute the partial correlations and list the ten strongest. Draw the estimated network with `networkx`.

4. Compare with the known pathway links Raf–Mek, Mek–Erk, Erk–Akt, Plcg–PIP2, PIP2–PIP3, PKC–P38, PKC–Jnk. Can the graphical lasso tell which protein activates which?

```
import numpy as np, pandas as pd
from sklearn.covariance import GraphicalLassoCV
d = pd.read_csv("../data/sachs.csv")
...
```

Exercise 2.3. PC on flow cytometry data. Use the log-transformed Sachs data of `data/sachs.csv`.

1. Run PC with the Fisher z test and $\alpha = 0.01$.
2. Compare the estimated skeleton with the consensus network reported by Sachs *et al.* (2005): PKC→PKA, PKC→Raf, PKC→Mek, PKC→P38, PKC→Jnk, PKA→Raf, PKA→Mek, PKA→Erk, PKA→Akt, PKA→P38, PKA→Jnk, Raf→Mek, Mek→Erk, Erk→Akt, Plcg→PIP2, Plcg→PIP3, PIP3→PIP2. How many consensus edges are recovered? How many estimated edges are absent from it?
3. Comment on the orientations. Which assumptions of PC are questionable here? (The data pool several experimental conditions.)

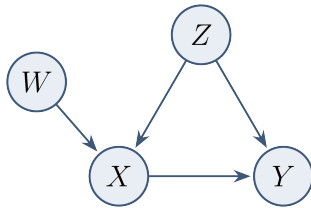
3 Session 3 — Interventions, Mendelian randomisation and causal discovery

Objectives

This sheet uses the notions of session 3: backdoor adjustment, instrumental variables and Mendelian randomisation, LiNGAM and OT-LiNGAM on RNA editing data.

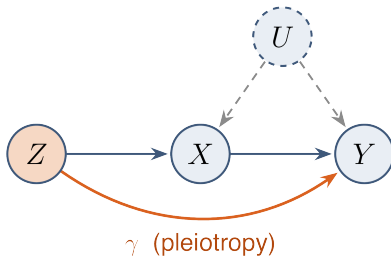
Theory

Exercise 3.1. Backdoor adjustment. X is a treatment, Y an outcome, Z a common cause of X and Y , W a cause of X only. All variables are observed.



1. Explain why $P(Y | X = x) \neq P(Y | do(X = x))$ in general.
2. List the backdoor paths from X to Y .
3. Which sets among $\emptyset$, $\{W\}$, $\{Z\}$, $\{Z, W\}$ satisfy the backdoor criterion?
4. Write $P(Y | do(X = x))$ with the backdoor formula for the smallest valid set. Is it useful to add W to the adjustment set?

Exercise 3.2. Instrumental variable assumptions. In the linear model of exercise 3.3 with a pleiotropic effect,



$$X = \beta_{ZX}Z + \gamma_X U + \varepsilon_X, \quad Y = \alpha X + \gamma Z + \gamma_Y U + \varepsilon_Y, \quad Z \perp (U, \varepsilon_X, \varepsilon_Y).$$

1. Compute $\text{Cov}(Z, X)$ and $\text{Cov}(Z, Y)$. Deduce the limit of the Wald ratio.
2. Under which condition is the Wald ratio consistent for α ? Express each IV assumption (relevance, independence, exclusion) as a property of the graph.
3. Compute the limit of the OLS slope of Y on X when $\gamma = 0$. When is it biased?

Practice

Exercise 3.3. Mendelian randomisation by simulation. Let $Z \sim \mathcal{B}(2, 0.3)$ be a genotype, $U \sim \mathcal{N}(0, 1)$ an unobserved confounder, and

$$X = \beta_{ZX}Z + U + \varepsilon_X, \quad Y = 0.2X + 1.5U + \varepsilon_Y,$$

with independent standard Gaussian errors.

1. With $\beta_{ZX} = 0.5$ and $n = 2000$, compute the OLS slope of Y on X and the Wald ratio $\widehat{\text{Cov}}(Z, Y) / \widehat{\text{Cov}}(Z, X)$.
2. Repeat 500 times and compare the distributions of both estimators (bias, standard deviation).
3. Same with a weak instrument, $\beta_{ZX} = 0.05$. What happens to the Wald ratio?
4. Add a direct effect γZ in the equation of Y (pleiotropy), $\gamma = 0.1$, $\beta_{ZX} = 0.5$. Measure the bias of the Wald ratio.

Exercise 3.4. Chronology of RNA editing in the chloroplast. The file `data/raw_ndhD.csv` (separator `;`) describes nanopore reads of the chloroplast *ndhD* transcript of *Arabidopsis thaliana* (Guilcher *et al.*, 2021). Each row is a read; each column `ed1–ed5` an editing site: `True` (edited), `False` (not edited), `Err` (sequencing error), empty (site not covered by the read).

1. Load the data. Recode `True` $\rightarrow$ 1, `False` and `Err` $\rightarrow$ 0, empty $\rightarrow$ missing. Report the number of reads, the proportion of complete reads and the editing frequency of each site.
2. On the complete reads (standardised), estimate a causal order of the five sites with DirectLiNGAM, Exhaustive OT-LiNGAM and Greedy OT-LiNGAM.
3. Repeat on all reads after imputation of the missing values with `sklearn.impute.IterativeImputer`. Are the orders stable?
4. Discuss: can these orders be read as a chronology of editing? List the assumptions of LiNGAM that are questionable here.

A Key notions

Session 1 — Causality and directed graphical models

- Ladder of causation: association $P(y | x)$, intervention $P(y | do(x))$, counterfactual.
- Simpson's paradox: an association reverses after aggregation when a confounder changes the weights of the strata.
- DAG factorisation: $p(x) = \prod_t p(x_t | x_{pa(t)})$.
- Chain and fork: conditioning on the middle node blocks; collider: conditioning on it (or a descendant) opens.
- d-separation $\Rightarrow$ conditional independence; Markov blanket: parents, children, co-parents.
- Markov equivalence: same skeleton and same v-structures; class represented by a CPDAG (completed partially directed acyclic graph, or essential graph): a single graph standing for all DAGs that encode the same conditional independences.

Session 2 — Biological networks

- Gaussian graphical model: $K = \Sigma^{-1}$, $K_{ij} = 0 \iff X_i \perp X_j | X_{\text{rest}}$, partial correlation $-K_{ij} / \sqrt{K_{ii}K_{jj}}$.
- Graphical lasso: $\min_K -\log \det K + \text{tr}(SK) + \lambda \|K\|_1$.
- PC algorithm: skeleton by conditional independence tests, v-structures from separating sets, Meek rules $\rightarrow$ CPDAG.
- Structural Hamming distance.

Session 3 — Interventions and causal discovery

- SCM: $V_j := f_j(V_{pa(j)}, U_j)$; $do(X = x)$ replaces the equation of X (graph surgery).
- Backdoor criterion and formula $P(y | do(x)) = \sum_z P(y | x, z)P(z)$; bad controls.
- Mendelian randomisation: genotype as instrument; Wald ratio $\text{Cov}(Z, Y) / \text{Cov}(Z, X)$.
- LiNGAM: linear SEM with independent non-Gaussian noises $\rightarrow$ causal order identifiable.
- OT-LiNGAM: non-Gaussianity $\mathcal{W}_2(\text{std}(X), \mathcal{N}(0, 1))^2$ of the residuals, maximised over orders.