

Causality and Graphical Models — Tutorial booklet

Student version

Christophe Ambroise

Table of contents

Introduction	4
I Tutorial sheets	5
1 Tutorial 1 — Causality and directed graphical models	6
Objectives	6
Theory	6
Practice	7
Further practice	8
2 Tutorial 2 — d-separation and Markov equivalence	9
Objectives	9
Theory	9
Practice	10
Further practice	11
3 Tutorial 3 — Markov random fields and Gaussian graphical models	12
Objectives	12
Theory	12
Practice	13
Further practice	14
4 Tutorial 4 — Interventions and identification	15
Objectives	15
Theory	15
Practice	16
Further practice	17
5 Tutorial 5 — Structure learning	19
Objectives	19
Theory	19
Practice	20
Further practice	21
6 Tutorial 6 — LiNGAM and OT-LiNGAM	22
Objectives	22
Theory	22
Practice	23

Further practice	23
Appendices	26
A Key notions	26
Lecture 1 — Causality and directed graphical models	26
Lecture 2 — d-separation and Markov equivalence	26
Lecture 3 — Markov random fields and Gaussian graphical models	26
Lecture 4 — Interventions, SCMs and identification	27
Lecture 5 — Structure learning	27
Lecture 6 — LiNGAM and OT-LiNGAM	27

Introduction

This booklet accompanies the tutorial sessions of the course Causality and Graphical Models.

Each sheet contains more exercises than a session can cover. The remaining exercises are for self-study.

The appendix “Key notions” summarises, lecture by lecture, what must be known.

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 Tutorial 1 — Causality and directed graphical models

Objectives

This sheet uses the notions of lecture 1: Simpson’s paradox and confounding, DAG factorisation, parameter counting, Gaussian Bayesian networks and their covariance.

Exercises 1 to 4 form the core of the session. Exercises 5 and 6 are for practice.

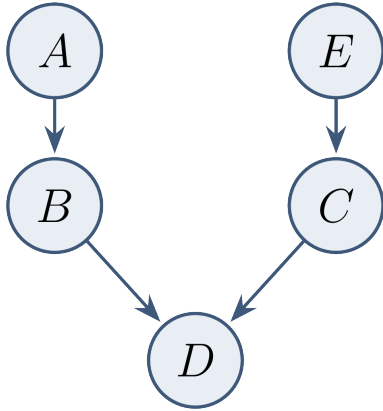
Theory

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. A Gaussian Bayesian network. Consider the chain $X_1 \rightarrow X_2 \rightarrow X_3$ defined by the structural equations

$$X_1 = Z_1, \quad X_2 = 1 + X_1 + Z_2, \quad X_3 = 2 + \frac{1}{2}(X_2 - 1) + Z_3,$$

where Z_1, Z_2, Z_3 are independent $\mathcal{N}(0, 1)$ variables.

1. Write the model in the form $x - \mu = W(x - \mu) + Sz$: give μ , W and S .
2. Compute the covariance matrix $\Sigma = (I - W)^{-1}S^2(I - W)^{-\top}$.
3. Compute the precision matrix $K = \Sigma^{-1}$, using $K = (I - W)^{\top}S^{-2}(I - W)$. Which entry of K is zero? Interpret this zero in terms of conditional independence.
4. Show that the DAG $X_1 \leftarrow X_2 \leftarrow X_3$ can produce the same covariance matrix. What does this imply for learning the direction of the arrows from Gaussian data?

Practice

Exercise 1.4. Simulating and estimating a linear SEM. This exercise uses the model of exercise 1.3.

1. Write a function `simulate(n, rng)` that returns an $n \times 3$ array drawn from the model by ancestral sampling (first X_1 , then X_2 , then X_3).
2. For $n = 10\,000$, compare the empirical covariance matrix with Σ , and the inverse of the empirical covariance with K .
3. Estimate the parameters of each conditional distribution: regress X_2 on X_1 and X_3 on X_2 by least squares, and estimate the residual variances. Compare with the true values.

4. Regress X_3 on (X_1, X_2) . What is the coefficient of X_1 ? Explain.

```
import numpy as np
rng = np.random.default_rng(0)

def simulate(n, rng):
    z = rng.standard_normal((n, 3))
    x1 = ...
    x2 = ...
    x3 = ...
    return np.column_stack([x1, x2, x3])
```

Further practice

Exercise 1.5. Simpson’s paradox by simulation. Simulate the sprint study of exercise 1.1: 100 men and 100 women; a man smokes with probability 0.8, a woman with probability 0.2; the time is $T = 11 + 2\mathbb{1}(\text{woman}) + 0.5S + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 0.2^2)$.

1. Compute the mean time of smokers and non-smokers, overall and by gender.
2. Fit the linear regressions $T \sim S$ and $T \sim S + G$. Compare the coefficients of S with the causal effect 0.5.
3. Which regression corresponds to the backdoor adjustment on G ?

Exercise 1.6. Reading independences in a table (after Bishop, ex. 8.3–8.4). Three binary variables have the joint distribution

a	b	c	$p(a, b, c)$	a	b	c	$p(a, b, c)$
0	0	0	0.192	1	0	0	0.192
0	0	1	0.144	1	0	1	0.064
0	1	0	0.048	1	1	0	0.048
0	1	1	0.216	1	1	1	0.096

1. Show that a and b are not independent.
2. Show that $a \perp b \mid c$.
3. Compute $p(a)$, $p(c \mid a)$ and $p(b \mid c)$, and check that $p(a, b, c) = p(a)p(c \mid a)p(b \mid c)$. Draw the corresponding DAG.
4. Give two other DAGs that are compatible with the same independences.

2 Tutorial 2 — d-separation and Markov equivalence

Objectives

This sheet uses the notions of lecture 2: chains, forks and colliders, d-separation, Markov blanket, Markov equivalence and CPDAG, parameter learning with a known graph.

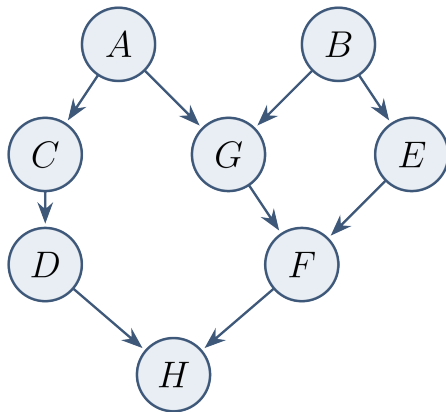
Exercises 1 to 4 form the core of the session. Exercises 5 and 6 are for practice.

Theory

Exercise 2.1. The three elementary structures. Let A, B, C be discrete variables.

1. Chain $A \rightarrow C \rightarrow B$: using the factorisation, show that $A \perp B \mid C$.
2. Fork $A \leftarrow C \rightarrow B$: show that $A \perp B \mid C$.
3. Collider $A \rightarrow C \leftarrow B$: show that $A \perp B$.
4. Collider again: let A and B be independent fair coins and $C = A \oplus B$ (exclusive or). Compute $P(A = 1 \mid C = 0)$, $P(B = 1 \mid C = 0)$ and $P(A = 1, B = 1 \mid C = 0)$. Conclude.

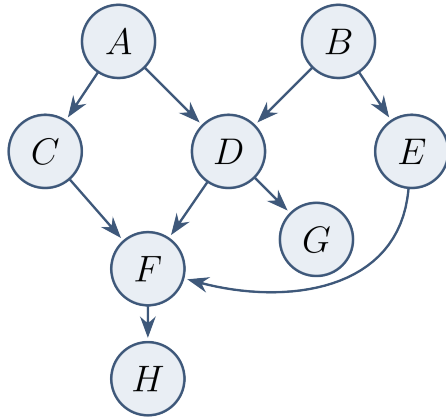
Exercise 2.2. d-separation in an eight-node DAG.



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.

Exercise 2.3. d-separation table.

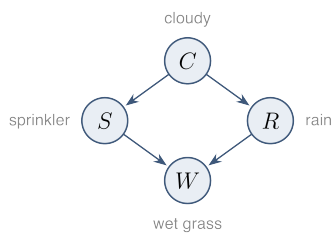


1. Identify the colliders.
2. For each line, say whether X and Y are d-separated by Z , and justify by exhibiting an active path or by showing that every path is blocked.

X	Y	Z
A	B	$\emptyset$
A	B	$\{H\}$
C	E	$\emptyset$
C	E	$\{D\}$
A	F	$\{C\}$
D	G	$\{F\}$

Practice

Exercise 2.4. 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)
...
```

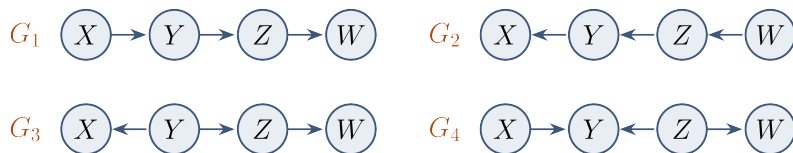
Further practice

Exercise 2.5. Markov blanket. In a DAG, show that the full conditional distribution of a node X_i is

$$p(x_i \mid x_{-i}) \propto p(x_i \mid x_{pa(i)}) \prod_{j \in ch(i)} p(x_j \mid x_{pa(j)}),$$

and deduce that X_i is independent of all other variables given its parents, children and co-parents.

Exercise 2.6. Markov equivalence. Consider the four DAGs on X, Y, Z, W :



1. Which of these DAGs are Markov equivalent? Use the Verma–Pearl theorem.
2. For each equivalence class, give one conditional independence that distinguishes it from the other class.
3. Draw the CPDAG of G_4 and list the DAGs of its equivalence class.

3 Tutorial 3 — Markov random fields and Gaussian graphical models

Objectives

This sheet uses the notions of lecture 3: factorisation and conditional independence, Hammersley–Clifford, Ising model and Gibbs sampling, precision matrix, partial correlations, graphical lasso.

Exercises 1 to 5 form the core of the session. Exercises 6 and 7 are for practice.

Theory

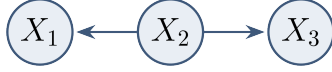
Exercise 3.1. Factorisation and conditional independence. Let X, Y, Z be discrete variables with a strictly positive joint distribution.

1. Show that if $p(x, y, z) = F(x, z)G(y, z)$ for two positive functions F and G , then $X \perp Y \mid Z$.
2. Conversely, show that $X \perp Y \mid Z$ implies such a factorisation.
3. Binary case: $F(x, z) = e^{\alpha x z}$ and $G(y, z) = e^{\beta y z}$ with $x, y, z \in \{0, 1\}$. Compute the normalising constant and $P(X = 1 \mid Y = 1, Z = 1)$.
4. Which undirected graph corresponds to this factorisation? Relate to the Hammersley–Clifford theorem.

Exercise 3.2. Schur complement and conditional Gaussian. Let $X = (X_A, X_B) \sim \mathcal{N}(\mu, \Sigma)$ with blocks $\Sigma_{AA}, \Sigma_{AB}, \Sigma_{BB}$, and $K = \Sigma^{-1}$ with blocks K_{AA}, K_{AB}, K_{BB} .

1. By solving the block system $\Sigma K = I$, show that $K_{AA} = (\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})^{-1}$ and $K_{AB} = -K_{AA}\Sigma_{AB}\Sigma_{BB}^{-1}$.
2. Write the quadratic form $(x - \mu)^\top K(x - \mu)$ in blocks and complete the square in x_A . Deduce that $X_A \mid X_B = b \sim \mathcal{N}(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(b - \mu_B), K_{AA}^{-1})$.
3. Take $A = \{1, 2\}$ and $B = \{3, \dots, p\}$. Show that the partial correlation of X_1 and X_2 given X_B equals $-K_{12}/\sqrt{K_{11}K_{22}}$.

Exercise 3.3. 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).

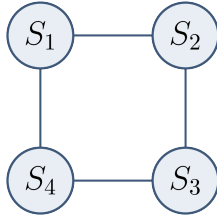
$$A = \begin{pmatrix} 9 & 3 & 1 \\ 3 & 9 & 3 \\ 1 & 3 & 9 \end{pmatrix}, \quad B = \begin{pmatrix} 8 & -3 & 1 \\ -3 & 9 & -3 \\ 1 & -3 & 8 \end{pmatrix}, \quad C = \begin{pmatrix} 9 & 3 & 0 \\ 3 & 9 & 3 \\ 0 & 3 & 9 \end{pmatrix}, \quad D = \begin{pmatrix} 9 & -3 & 0 \\ -3 & 10 & -3 \\ 0 & -3 & 9 \end{pmatrix}.$$

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

Practice

Exercise 3.4. Ising model on a 4-cycle: exact computation and Gibbs sampling.

Four spins $S_1, \dots, S_4 \in \{-1, +1\}$ on a cycle:



$$p(s) = \frac{1}{Z} \exp \left(\beta \sum_{i \sim j} s_i s_j + \alpha \sum_i s_i \right), \quad \alpha = 0.2, \beta = 0.5.$$

1. Show that $P(S_1 = +1 \mid s_{-1}) = \sigma(2(\alpha + \beta(s_2 + s_4)))$ with $\sigma(t) = 1/(1 + e^{-t})$.
2. Compute Z and $\mathbb{E}[S_1]$ exactly by enumerating the 16 configurations.
3. Implement a Gibbs sampler (5 000 sweeps, burn-in 500) and estimate $\mathbb{E}[S_1]$. Compare with the exact value.
4. Multiply the exponent by $1/T$ (temperature T). Describe the distribution when $T \rightarrow 0$ and $T \rightarrow \infty$, and check with your sampler.

Exercise 3.5. Graphical lasso and ROC curves.

1. Draw an Erdős–Rényi graph on $p = 15$ nodes with edge probability 0.2. Build a precision matrix K with $K_{ij} = -0.3$ on edges, 0 elsewhere, and a diagonal large enough to make K positive definite.
2. Simulate $n \in \{10, 100, 1000\}$ observations from $\mathcal{N}(0, K^{-1})$.
3. For each n and a grid of penalties λ , estimate K with `sklearn.covariance.graphical_lasso` on the empirical correlation matrix. Compute the true and false positive rates of edge detection.

4. Draw the three ROC curves and compute their areas. Comment on the role of n .

```
import numpy as np
from sklearn.covariance import graphical_lasso
rng = np.random.default_rng(2)
p = 15
A = np.triu(rng.random((p, p)) < 0.2, 1); A = A | A.T
...
```

Further practice

Exercise 3.6. Pseudo-likelihood of a binary Ising model. Let $X \in \{0, 1\}^p$ follow $p(x; \theta) = \exp(\theta \sum_{j \sim k} x_j x_k - \Phi(\theta))$ on a graph G .

1. Compute $P(X_j = 1 \mid x_{-j})$ and $P(X_j = 0 \mid x_{-j})$.
2. For an i.i.d. sample $x^{(1)}, \dots, x^{(n)}$, write the log-pseudo-likelihood $\sum_i \sum_j \log P(x_j^{(i)} \mid x_{-j}^{(i)})$. Why is it easier to maximise than the likelihood?

Exercise 3.7. 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")
...
```

4 Tutorial 4 — Interventions and identification

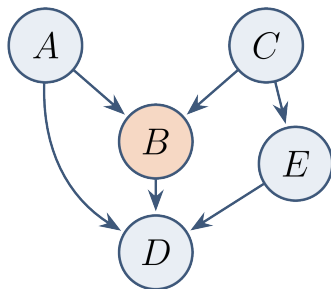
Objectives

This sheet uses the notions of lecture 4: structural causal models, interventions and truncated factorisation, backdoor and frontdoor adjustment, instrumental variables and Mendelian randomisation, counterfactuals.

Exercises 1 to 5 form the core of the session. Exercises 6 to 8 are for practice.

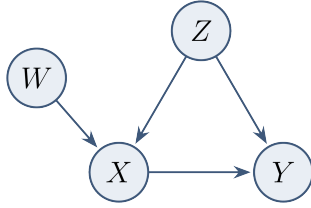
Theory

Exercise 4.1. Truncated factorisation and adjustment sets.



1. Write the observational factorisation of $p(a, b, c, d, e)$.
2. Write the interventional distribution $p(a, b, c, d, e \mid do(B = b_0))$ by truncated factorisation, then $p(d \mid do(B = b_0))$ as a function of observational conditionals.
3. List the backdoor paths from B to D .
4. Which of the following sets satisfy the backdoor criterion relative to (B, D) : $\emptyset$, $\{A\}$, $\{C\}$, $\{A, C\}$, $\{A, E\}$, $\{A, C, E\}$? Give the adjustment formula for one minimal valid set.

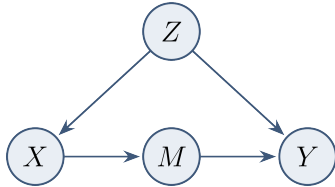
Exercise 4.2. Backdoor adjustment. X is a treatment, Y an outcome, Z a common cause of X and Y , W a cause of X only. All variables are observed.



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 4.3. A linear SCM: effects, regressions and counterfactuals. Consider the SCM, with U_Z, U_X, U_M, U_Y independent $\mathcal{N}(0, 1)$:

$$Z := U_Z, \quad X := 2Z + U_X, \quad M := 0.5X + U_M, \quad Y := 3M + Z + U_Y.$$



1. Compute $\mathbb{E}[Y | do(X = x)]$. What is the causal effect of a unit increase of X ?
2. Compute the slope of the regression of Y on X alone, $\text{Cov}(X, Y) / \text{Var}(X)$. Explain the difference with question 1.
3. What is the coefficient of X in the regression of Y on (X, Z) ? On (X, Z, M) ? Which regression estimates the causal effect?
4. One individual is observed with $z = 1$, $x = 3$, $m = 2$, $y = 8$. Compute the counterfactual $Y_{X=1}$ for this individual (abduction, action, prediction). Compare with $\mathbb{E}[Y | do(X = 1)]$.

Practice

Exercise 4.4. Checking exercise 4.3 by simulation.

1. Simulate $N = 100\,000$ draws of the SCM of exercise 4.3.
2. Fit the three regressions $Y \sim X$, $Y \sim X + Z$, $Y \sim X + Z + M$ and compare with your answers.
3. Implement the intervention $do(X = x)$ by replacing the equation of X , keeping the same noises, for $x = 0$ and $x = 1$. Estimate $\mathbb{E}[Y | do(X = 1)] - \mathbb{E}[Y | do(X = 0)]$.
4. Write a function `counterfactual(z, x, m, y, x_new)` implementing the three steps, and check the value found in exercise 4.3.

Exercise 4.5. 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

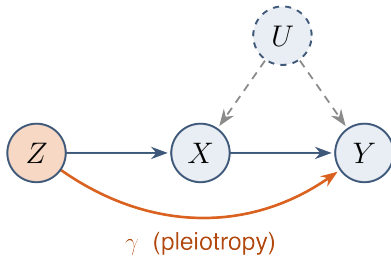
$$X = \beta_{ZX}Z + U + \varepsilon_X, \quad Y = 0.2X + 1.5U + \varepsilon_Y,$$

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.

Further practice

Exercise 4.6. Instrumental variable assumptions. In the linear model of exercise 4.5 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?

Exercise 4.7. Frontdoor adjustment. Binary variables: $U \sim \mathcal{B}(0.5)$ unobserved, $X \sim \mathcal{B}(0.2 + 0.6U)$, $M \sim \mathcal{B}(0.1 + 0.7X)$, $Y \sim \mathcal{B}(0.1 + 0.5M + 0.3U)$.

1. Compute exactly $P(Y = 1 \mid do(X = 1)) - P(Y = 1 \mid do(X = 0))$.
2. Simulate $n = 100\,000$ observations of (X, M, Y) only. Compute the naive difference $P(Y = 1 \mid X = 1) - P(Y = 1 \mid X = 0)$.
3. Estimate the causal effect with the frontdoor formula and compare.

Exercise 4.8. Simulating a structural causal model and an intervention. Consider the SCM of the lecture, with U_Z, U_X, U_Y independent $\mathcal{N}(0, 1)$:

$$Z := U_Z, \quad X := 2Z + U_X, \quad Y := 3X + Z + U_Y.$$

1. Write a function `simulate(n, rng, x_do=None)` returning a data frame with columns `Z`, `X`, `Y` drawn by ancestral sampling (first Z , then X , then Y). When `x_do` is given, the equation of X is replaced by $X := x_do$: this simulates the intervention $do(X = x_{do})$.
2. Draw $n = 100\,000$ observations. Fit the linear regressions $Y \sim X$ and $Y \sim X + Z$ and report the coefficient of X in each.
3. Simulate $do(X = 0)$ and $do(X = 1)$ and compute the difference of the means of Y . Which regression of question 2 recovers it?
4. In the observational sample, keep the units with $|X - 1| < 0.1$, then those with $|X| < 0.1$. Compute the difference of the means of Y between these two groups, and the mean of Z in each group. Why does this difference not match question 3?
5. Simulate a randomised experiment: the equation of X is replaced by $X := V$, with $V \sim \mathcal{N}(0, 5)$ drawn independently of everything else. Fit $Y \sim X$ on these data. Conclude on the role of randomisation.

5 Tutorial 5 — Structure learning

Objectives

This sheet uses the notions of lecture 5: mutual information and Chow–Liu trees, decomposable scores, the PC algorithm, CPDAGs and the structural Hamming distance.

Exercises 1 to 5 form the core of the session. Exercises 6 and 7 are for practice.

Theory

Exercise 5.1. Mutual information of a Gaussian pair. Let (X_s, X_t) be a Gaussian vector with correlation ρ . Show that

$$I(X_s; X_t) = \mathbb{E} \left[\log \frac{p(x_s, x_t)}{p(x_s)p(x_t)} \right] = -\frac{1}{2} \log(1 - \rho^2).$$

Deduce that, for Gaussian data, the Chow–Liu tree is the maximum spanning tree for the weights $|\rho_{st}|$.

Exercise 5.2. Chow–Liu by hand. Four Gaussian variables have the correlation matrix

$$R = \begin{pmatrix} 1 & 0.8 & 0.5 & 0.3 \\ 0.8 & 1 & 0.6 & 0.4 \\ 0.5 & 0.6 & 1 & 0.7 \\ 0.3 & 0.4 & 0.7 & 1 \end{pmatrix}.$$

1. Compute the six mutual informations $I(X_s; X_t)$.
2. Run Kruskal’s algorithm and give the Chow–Liu tree.
3. Orient the tree from root X_1 and write the corresponding factorisation. How many parameters does the tree model have, compared with the full Gaussian model?

Exercise 5.3. PC algorithm with an independence oracle. Five variables A, B, C, D, E are generated by an unknown DAG, and their distribution is faithful to it. An oracle answers conditional independence queries exactly. The pairs that can be separated, with the smallest separating set, are:

Pair	Separating set
A, B	$\emptyset$
A, D	$\{C\}$
A, E	$\{C\}$
B, D	$\{C\}$
B, E	$\{C\}$
C, E	$\{D\}$

All other pairs are dependent given every subset of the remaining variables.

1. Run the skeleton phase of PC. Give the skeleton and the separating sets.
2. Orient the v-structures.
3. Apply Meek's rules. Is the DAG fully identified?

Practice

Exercise 5.4. PC and GES on simulated data. Use the DAG recovered in exercise 5.3 with linear Gaussian equations, e.g. $C = 0.8A + 0.8B + \varepsilon_C$, $D = 0.9C + \varepsilon_D$, $E = 0.9D + \varepsilon_E$.

1. Write a function simulating n observations.
2. Run `causallearn.search.ConstraintBased.PC.pc` (Fisher z test, $\alpha = 0.05$) and `causallearn.search.ScoreBased.GES.ges` (BIC score) for $n \in \{50, 200, 1000, 5000\}$, 10 replications each.
3. Write a function computing the structural Hamming distance between two CPDAGs, and report its mean for each method and each n .
4. Study the effect of α on PC for $n = 200$.

The causal-learn graph matrix `G.graph` encodes $i \rightarrow j$ as $G[i, j] = -1$, $G[j, i] = 1$, and $i - j$ as $G[i, j] = G[j, i] = -1$.

Exercise 5.5. 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.)

Further practice

Exercise 5.6. Score equivalence of BIC. Two Gaussian variables X, Y , sample of size n .

1. Write the maximised log-likelihood of the models $G_1 : X \rightarrow Y$ and $G_2 : Y \rightarrow X$, and count their parameters.
2. Show that $\text{BIC}(G_1) = \text{BIC}(G_2)$. Compare with the empty graph.
3. Why is score equivalence a desirable property? Why does it prevent any score-based method from orienting $X - Y$?

Exercise 5.7. Chow–Liu tree of the signalling network. On the log-transformed Sachs data, compute the Gaussian mutual informations, build the maximum spanning tree with `networkx.maximum_spanning_tree`, and compare its edges with the graphical lasso network of exercise [3.7](#).

6 Tutorial 6 — LiNGAM and OT-LiNGAM

Objectives

This sheet uses the notions of lecture 6: identifiability of linear non-Gaussian models, ICA-LiNGAM and DirectLiNGAM, Wasserstein non-Gaussianity, the OT-LiNGAM order objective and its solvers, the disorder of an estimated order.

Exercises 1 to 5 form the core of the session. Exercises 6 to 9 are for practice.

Theory

Exercise 6.1. Two-variable LiNGAM. Let $X = e_1$ and $Y = bX + e_2$, with e_1, e_2 independent, centred, of unit variance, and $b \neq 0$.

1. Compute the least-squares coefficient c of the regression of X on Y , and the residual $r = X - cY$ as a linear combination of e_1 and e_2 . Check that $\text{Cov}(r, Y) = 0$.
2. Gaussian case: show that (r, Y) is a Gaussian vector and deduce that $r \perp Y$. Conclude that the model $Y \rightarrow X$ fits the data as well as $X \rightarrow Y$.
3. Non-Gaussian case: using the Darmois–Skritovich theorem, show that r and Y are dependent if e_1 is not Gaussian. Which direction then yields an independent residual?
4. Uniform noises, $b = 1$: compute $\mathbb{E}[r^2 Y^2]$ and $\mathbb{E}[r^2] \mathbb{E}[Y^2]$ (recall $\mathbb{E}[e^4] = 9/5$ for a standardised uniform). Conclude.

Exercise 6.2. Wasserstein non-Gaussianity of the uniform distribution. Let U be uniform on $[-\sqrt{3}, \sqrt{3}]$ (mean 0, variance 1).

1. Using the quantile formula $\mathcal{W}_2(\mu, \nu)^2 = \int_0^1 (F_\mu^{-1}(u) - F_\nu^{-1}(u))^2 du$, show that

$$\mathcal{W}_2(U, \mathcal{N}(0, 1))^2 = 2 - 2 \int_0^1 \sqrt{3} (2u - 1) \Phi^{-1}(u) du.$$

2. Show that $\int_0^1 u \Phi^{-1}(u) du = \mathbb{E}[Z\Phi(Z)] = \mathbb{E}[\phi(Z)]$ for $Z \sim \mathcal{N}(0, 1)$, and compute it (Stein's lemma: $\mathbb{E}[Zg(Z)] = \mathbb{E}[g'(Z)]$).
3. Deduce that $\mathcal{W}_2(U, \mathcal{N}(0, 1))^2 = 2 - 2\sqrt{3/\pi} \approx 0.0456$.
4. Check numerically with the order-statistics formula of the lecture on a sample of size 10^5 .

Exercise 6.3. Scale invariance and varsortability.

1. Consider the chain $X_1 \rightarrow X_2 \rightarrow X_3$ with $X_1 = e_1$, $X_2 = 2X_1 + e_2$, $X_3 = 2X_2 + e_3$, unit noise variances. Compute $\text{Var } X_j$. Which naive rule recovers the causal order?
2. Show that this rule fails once each variable is standardised, while the causal order is unchanged.
3. Show that the OT-LiNGAM objective $G(\sigma)$ is unchanged when each X_j is multiplied by a constant $c_j \neq 0$.

Practice

Exercise 6.4. Comparing LiNGAM estimators by simulation.

1. Write a function drawing a random linear SEM on $p = 6$ variables: random causal order, each possible edge present with probability 0.4, weights uniform on $\pm[0.5, 1.5]$.
2. Write a function sampling n observations with independent noises of type Gaussian, uniform or Laplace, with random scales in $[0.5, 2]$; standardise the data.
3. For each noise type and $n \in \{200, 1000\}$, run 10 replications of `lingam.DirectLiNGAM`, `lingam.ICALiNGAM`, `otlingam.ExhaustiveOTLiNGAM` and `otlingam.GreedyOTLiNGAM`, and compute the disorder of the estimated order with `otlingam.utils.disorder`.
4. Summarise the results in a table. Comment, in particular, on the Gaussian case.

The `lingam` convention is $B[j, k] = \text{effect of } X_k \text{ on } X_j$, used also by `disorder(B_true, causal_order)`.

Exercise 6.5. 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.

Further practice

Exercise 6.6. Cost of exact order search.

1. How many causal orders are there for d variables? How many local scores $\hat{g}_u(P)$, $u \notin P$, does the dynamic programme of Exhaustive OT-LiNGAM evaluate?
2. Compare these numbers for $d = 10, 20, 30$.
3. Describe a situation in which Greedy OT-LiNGAM makes an irreversible error that the exhaustive search avoids.

Exercise 6.7. Non-Gaussianity of binary variables.

1. Let $X \sim \mathcal{B}(p)$ and $\tilde{X}$ its standardised version. Show that

$$\mathcal{W}_2(\tilde{X}, \mathcal{N}(0, 1))^2 = 2 - \frac{2\phi(z_p)}{\sqrt{p(1-p)}}, \quad z_p = \Phi^{-1}(1-p).$$

2. Plot this quantity as a function of p . Which editing sites will OT-LiNGAM tend to place first?
3. Estimate the order of the 13 events of `data/raw_ndhB.csv` (complete reads) and compare it with the ranking of the sites by editing frequency. Assess the stability of the order on 50 bootstrap samples.

Exercise 6.8. Cause or effect? The direction between two variables. Let $X = e_1$ and $Y = 0.8X + e_2$, with e_1, e_2 independent, centred, of unit variance, all Gaussian, all uniform or all Laplace.

1. Write a function `pair(n, noise, rng)` returning an $n \times 2$ array whose columns are (Y, X) , in this order: the column order gives no hint on the direction.
2. Uniform noises, $n = 1\,000$: plot the residual of the regression of Y on X against X , and the residual of the regression of X on Y against Y . Which plot shows independence?
3. For each noise type and $n \in \{200, 1\,000\}$, run 100 replications and report the proportion of runs in which `lingam.DirectLiNGAM` and `otlingam.GreedyOTLiNGAM` place X (column 1) before Y (column 0), i.e. `causal_order_ == [1, 0]`.
4. Comment on the results, in particular in the Gaussian case.

Exercise 6.9. A complete LiNGAM analysis. The code below simulates $n = 1\,000$ observations of a linear SEM on six variables with non-Gaussian noises (Laplace for the first three, uniform for the last three), then shuffles the columns: the causal order has to be found.

```
import numpy as np
rng = np.random.default_rng(2026); n = 1000
B = np.zeros((6, 6)) # B[j, k]: effect of x_k on x_j
B[1, 0], B[2, 0], B[3, 1], B[3, 2] = 0.8, 0.6, -0.7, 0.5
B[4, 3], B[5, 2], B[5, 4] = 0.9, -0.6, 0.5
E = np.c_[rng.laplace(size=(n, 3)), rng.uniform(-1.7, 1.7, size=(n, 3))]
X = E @ np.linalg.inv(np.eye(6) - B).T # solution of X = B X + E
perm = rng.permutation(6) # column i = variable x_perm[i]
X, B_true = X[:, perm], B[np.ix_(perm, perm)]
```

1. For each column, compute the excess kurtosis (`scipy.stats.kurtosis`) and the p-value of the Shapiro–Wilk test (`scipy.stats.shapiro`). Which variables are closest to Gaussian, and why?
2. Fit `lingam.DirectLiNGAM` and `otlingam.ExhaustiveOTLiNGAM`. Compare the estimated orders and adjacency matrices with `B_true`, using `disorder`, `shd` and `f1_score` from `otlingam.utils`.
3. Run 100 bootstrap resamples of `DirectLiNGAM` (`model.bootstrap`, on a new instance: `bootstrap` refits the model in place) and compute the frequency of each edge with `get_probabilities(min_causal_effect=0.05)`. List the edges found in more than half of the resamples, and compare with the true edges.
4. Print the p-values of independence between the residuals (`get_error_independence_p_values`). Then remove the column of x_2 (a hidden variable) and refit. Which pair of residuals becomes dependent? Interpret.
5. Estimate the total effect of x_0 on x_5 with `estimate_total_effect` and compare it with the true total effect, read in $(I - B)^{-1}$.
6. Declare x_0 exogenous and x_5 a sink with `lingam.utils.make_prior_knowledge` and refit `DirectLiNGAM` with this prior knowledge.
7. Replace all the noises by standard Gaussian ones and repeat questions 2 and 4. What do the residual tests tell, and what do they miss?

A Key notions

Lecture 1 — Causality and directed graphical models

- Ladder of causation: association $P(y \mid x)$, intervention $P(y \mid do(x))$, counterfactual $P(y_{x'} \mid x, y)$.
- Simpson's paradox: an association reverses after aggregation when a confounder changes the weights of the strata.
- Conditional independence: $X \perp Y \mid Z \iff p(x, y \mid z) = p(x \mid z)p(y \mid z)$.
- DAG factorisation: $p(x) = \prod_t p(x_t \mid x_{pa(t)})$; equivalent to the ordered Markov property.
- Gaussian Bayesian network: $x - \mu = W(x - \mu) + Sz$, $\Sigma = (I - W)^{-1}S^2(I - W)^{-\top}$.

Lecture 2 — d-separation and Markov equivalence

- Chain and fork: conditioning on the middle node blocks; collider: conditioning on it (or a descendant) opens.
- d-separation: every path blocked; global Markov property: d-separation $\Rightarrow$ conditional independence.
- Markov blanket in a DAG: parents, children, co-parents.
- Faithfulness: independences of $p =$ d-separations of G .
- Markov equivalence (Verma–Pearl): 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.
- Parameter learning with a known graph: one estimation problem per node (counts, regressions).

Lecture 3 — Markov random fields and Gaussian graphical models

- Undirected graph: separation $\Rightarrow$ independence; Markov blanket = neighbours.
- Hammersley–Clifford: for $p > 0$, Markov property $\iff p(x) = \frac{1}{Z} \prod_c \psi_c(x_c)$ over cliques.
- Moralisation: a DAG distribution factorises on its moral graph.
- Ising model, conditional $P(x_i = 1 \mid x_{-i}) = \sigma(2(\alpha + \beta m_i))$, Gibbs sampling.
- Gaussian graphical model: $K = \Sigma^{-1}$, $K_{ij} = 0 \iff X_i \perp X_j \mid X_{\text{rest}}$, partial correlation $-K_{ij}/\sqrt{K_{ii}K_{jj}}$.
- Graphical lasso: $\min_K -\log \det K + \text{tr}(SK) + \lambda \|K\|_1$.

Lecture 4 — Interventions, SCMs and identification

- SCM: $V_j := f_j(V_{pa(j)}, U_j)$; intervention $do(X = x)$ replaces the equation of X (graph surgery).
- Truncated factorisation: $P(v \mid do(x)) = \mathbb{1}(x_v = x) \prod_{V_j \neq X} P(v_j \mid v_{pa(j)})$.
- Backdoor criterion: no descendant of X , every backdoor path blocked; $P(y \mid do(x)) = \sum_z P(y \mid x, z)P(z)$.
- Bad controls: mediators, colliders.
- Frontdoor formula: $P(y \mid do(x)) = \sum_m P(m \mid x) \sum_{x'} P(y \mid x', m)P(x')$.
- Instrumental variable and Mendelian randomisation: Wald ratio $\text{Cov}(Z, Y) / \text{Cov}(Z, X)$.
- Counterfactuals: abduction, action, prediction.

Lecture 5 — Structure learning

- Search space: super-exponential number of DAGs.
- Chow–Liu: maximum spanning tree for mutual informations; Gaussian $I = -\frac{1}{2} \log(1 - \rho^2)$.
- Decomposable scores (BIC, BDe), score equivalence; hill climbing, GES.
- PC algorithm: skeleton by conditional independence tests, v-structures from separating sets, Meek rules $\rightarrow$ CPDAG.
- Structural Hamming distance.

Lecture 6 — LiNGAM and OT-LiNGAM

- LiNGAM: $X = BX + \varepsilon$, acyclic B , independent noises, at most one Gaussian $\rightarrow$ causal order identifiable.
- Darmois–Skitovich; ICA-LiNGAM; DirectLiNGAM (independent residuals).
- Non-Gaussianity: $\mathcal{W}_2(\text{std}(X), \mathcal{N}(0, 1))^2$, exact from order statistics.
- OT-LiNGAM objective $G(\sigma) = \sum_j \mathcal{W}_2(\text{std}(R_j(\sigma)), \mathcal{N}(0, 1))^2$; its maximisers are the causal orders.
- Solvers: exhaustive dynamic programme ($d 2^{d-1}$ local scores), greedy search; disorder of an order.
- LiNGAM analysis in practice: non-Gaussianity check, bootstrap frequencies of the edges, independence tests between residuals (hidden confounders), prior knowledge (exogenous variables, sinks), total causal effects.