Seance 3 pour le groupe Causalité

Sophie Donnet et Zacharie Naulet

March 31, 2025

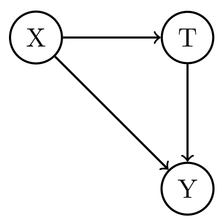
Vu aux séances 1 et 2.

- Séance 1: paradox de Simpson, potential outcome, counter factual, variables confondantes, hypothèses pour identifier les effets causaux $(Y(1) - Y(0))$ à partir des dobservations?
- Séance 2: travail sur les DAG ou réseaux bayésiens, lectures d'indépensances conditionnelles à partir de la structure du graphe, d-séparation.

Pour cette séance 3 : do-intervention

1 Do-intervention

Exemple Considérons 3 variables Y l'outcome (survie), T le traitement et X des covariables qui décrivent le patient. Voici leur DAG



Ainsi les covariables ont une influence à la fois sur la survie et sur le traitement donné au patient. Selon ce DAG, la loi jointe des variables s'écrit:

$$P(Y, T, X) = P(Y \mid T, X)P(T \mid X)P(X)$$

La *do*-intervention consiste à remplacer la loi de T dans la loi jointe ci-dessus par la Dirac $\delta_{\{T=t\}}$

$$P(Y \mid T = t, X)\delta_{\{T=t\}}P(X)$$

On note $Y(t)$ la variable dont la loi est définie par:

$$p_{Y(t)}(y) = \int_x p(y \mid T = t, X = x)p(x)dx$$

Typiquement on s'intéresse à $\mathbb{E}[Y(t)]$: effet causal de t sur l'outcome
On rencontrera la notation $Y \mid do(T = t)$ mais ce n'est pas une loi conditionnelle.

Définition Si on a $\mathbb{X} = X_1, \ldots, X_d$ et un DAG.

$$P(x_1, \ldots, x_d) = \prod_{i=1}^{d} P(x_i \mid pa(x_i))$$

où $pa(X_i)$ sont les parents des X_i dans le DAG. Soit S un sous ensemble de noeuds

$$P(\mathbb{X}_{\overline{S}} | do(S)) = \prod_{i \notin S} P(x_i \mid pa(x_i))$$

Do-intervention versus expérence conditionnelle $Y(t) \neq Y|T = t$. En effet,

$$\begin{aligned} p_{Y(t)}(y) &= \int_x p(y \mid T = t, X = x) p(X = x) dx \\ p_{Y|T=t}(y) &= \int_x p(y \mid T = t, X = x) p(X = x \mid T = t) dx \\ &= \int_x p(y \mid T = t, X = x) \frac{p(T = t, X = x)}{p(T = t)} dx \end{aligned}$$

$p(X = x) = p(X = x \mid T = t)$ si X et T sont indépendantes.

2 Modèle causal?

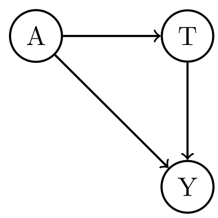
On cherche à comprendre en quoi le sens des flèches est important.

Considérons 3 variables:

- Y la betadiversité d'un ecosystème,
- A l'altitude et
- T la température.

On veut comprendre l'effet de T sur la betadiversité Y.

Modèle 1 L'altitude a un effet sur la température ET la biodiversité



On propose le modèle linéaire suivant:

$$\begin{aligned} Y &= \mu_Y + \alpha A + \beta T + \mathcal{N}(0, \sigma_Y^2) \\ T &= \mu_T + \gamma A + \mathcal{N}(0, \sigma_T^2) \\ A &\sim \mathcal{N}(\mu_A, \sigma_A^2) \end{aligned}$$

Dans ce cas, l'effet qui nous intéresse est

$$p(Y(t)) = \int_a p(y|T = t, A = a) p(A = a) da$$

$$Y(t) \mid T = t, A \sim \mu_Y + \alpha A + \beta t + \mathcal{N}(0, \sigma_Y^2)$$

et

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

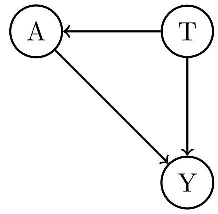
donc

$$Y(t) \sim \mu_Y + \alpha \mu_A + \beta t + \mathcal{N}(0, \alpha^2 \sigma_A^2 + \sigma_Y^2)$$

$$\mathbb{E}[Y(t)] = \mu_Y + \alpha \mu_A + \beta t$$

Effet causal de T sur Y (causal) $= \beta$

Modèle 2



La température a un effet sur l'altitude et la biodiversité. C'est idiot dit comme ça, mais en terme de modèles c'est valable aussi. Les deux modèles sont les mêmes, à reparamétrisation près.

$$\begin{aligned} Y &= \mu_Y + \alpha A + \beta T + \mathcal{N}(0, \sigma_Y^2) \\ A &= \nu_A + \delta T + \mathcal{N}(0, \xi_A^2) \\ T &\sim \mathcal{N}(\nu_T, \xi_T^2) \end{aligned}$$

La deuxième ligne peut être réecrite

$$T = \frac{\nu_A}{\delta} + \frac{1}{\delta} A + \frac{1}{\delta}\mathcal{N}(0, \xi_A^2) = \frac{\nu_A}{\delta} + \frac{1}{\delta} A + \mathcal{N}(0, \frac{\xi_A^2}{\delta^2})$$

et marginalement

$$A \sim \mathcal{N}(\nu_A + \delta\nu_T, \delta^2\xi_T^2 + \xi_A^2)$$

donc les modèles sont les mêmes avec la reparamétrisation suivante:

Model 1	Model 2
$\mu_Y, \alpha, \beta, \sigma_Y^2$	$\mu_Y, \alpha, \beta, \sigma_Y^2$
γ	$\frac{1}{\delta}$
μ_A	$\nu_A + \delta\nu_T$
μ_T	$\frac{\nu_A}{\delta}$
$\mu_T + \gamma\mu_A$	ν_T
σ_A^2	$\delta^2\xi_T^2 + \xi_A^2$
σ_T^2	$\frac{\xi_A^2}{\delta^2}$

$$\gamma := \frac{1}{\delta}$$

Dans ce cas, l'effet qui nous intéresse est

$$p(Y(t)) = \int_a p(y \mid T = t, A = a) p(A = a \mid T = t) da$$

$$Y \mid T = t, A \sim \mu_Y + \alpha A + \beta t + \mathcal{N}(0, \sigma_Y^2)$$

et

$$A \mid T = t \sim \mathcal{N}(\nu_A + \delta t, \xi_A^2)$$

donc

$$Y(t) \sim \mu_Y + \alpha\nu_A + \alpha\delta t + \beta t + \mathcal{N}(0, \alpha^2\xi_A^2 + \sigma_Y^2)$$

$$\mathbb{E}[Y(t)] = \mu_Y + \alpha\nu_A + (\alpha\delta + \beta)t$$

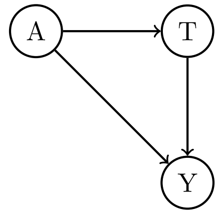
Effet causal de T sur $Y = \beta + \alpha\delta \neq \beta$. On ajoute l'effet direct ET l'effet indirect.

Ainsi, le DAG (=le sens des flèches) a une influence sur la quantité qui caractérise l'effet dit causal.

3 Identifabilité de l'effet causal

On se demande à quelle condition sur la structure du modèle / réseau bayésien l'effet causal est identifiable à partir de la loi des observations, qui peut seulement porter sur une partie seulement des variables.

Illustration Reprenons le modèle linéaire précédent dans sa version 1 .



$$
\begin{array}{rcl}
Y & = & \mu_Y + \alpha A + \beta T + \mathcal{N}(0, \sigma_Y^2) \\
T & = & \mu_T + \gamma A + \mathcal{N}(0, \sigma_T^2) \\
A & \sim & \mathcal{N}(\mu_A, \sigma_A^2)
\end{array}
$$

A est une variable confondante car elle joue à la fois directement sur Y mais aussi au travers de T. On a vu que si on observe l'altitude A, on peut obtenir l'effet causal β à partir de la loi des observations qui est la loi jointe, en faisant une régression prenant en compte A. A-t-on accès à cet effet causal β si on n'observe pas A?

Supposons que A soit une variable latente $\sim \mathcal{N}(\mu_A, \sigma_A^2)$ alors la loi des observations (Y, T) est encore gaussienne, il faut calculer ses moments;

$$
\begin{array}{lll}
\mathbb{E}(A) = \mu_A & \mathbb{E}(T) = \mu_T + \gamma\mu_A & \mathbb{E}(Y) = \mu_Y + \alpha\mu_A + \beta\mu_T \\
\mathbb{V}(A) = \sigma_A^2 & \mathbb{V}(T) = \gamma^2\sigma_A^2 + \sigma_T^2 & \mathbb{C}ov(T, A) = \mathbb{C}ov(\mu_T + \gamma A + \varepsilon_T, A) = \gamma\sigma_A^2
\end{array}
$$

En effet

$$
\begin{aligned}
\mathbb{V}(T) &= \mathbb{V}(\mu_T + \gamma A + \varepsilon_T) \\
&= \gamma^2 \mathbb{V}(A) + \sigma_T^2 \\
&= \gamma^2\sigma_A^2 + \sigma_T^2
\end{aligned}
$$

De plus:

$$
\begin{aligned}
\mathbb{V}(Y) &= \mathbb{V}(\mu_T + \alpha A + \beta T + \varepsilon_Y) \\
&= \mathbb{V}(\mu_T + \alpha A + \beta(\mu_T + \gamma A + \varepsilon_T) + \varepsilon_Y) \\
&= \mathbb{V}((\alpha + \beta\gamma)A + \beta\varepsilon_T + \varepsilon_Y) \\
&= (\alpha + \beta\gamma)^2\sigma_A^2 + \beta^2\sigma_T^2 + \sigma_Y^2
\end{aligned}
$$

$$
\begin{array}{rcl}
\mathbb{C}ov(T, Y) & = & \mathbb{C}ov(T, Y) \\
& = & \mathbb{C}ov(T, \mu_Y + \alpha A + \beta T + \varepsilon_Y) \\
& = & \alpha\mathbb{C}ov(T, A) + \beta\mathbb{C}ov(T, T) \\
& = & \alpha\mathbb{C}ov(T, A) + \beta(\gamma^2\sigma_A^2 + \sigma_T^2) \\
& = & \alpha\gamma\sigma_A^2 + \beta\gamma^2\sigma_A^2 + \beta\sigma_T^2 \\
& = & \gamma(\alpha + \beta\gamma)\sigma_A^2 + \beta\sigma_T^2
\end{array}
$$

$$
\begin{pmatrix} Y \\ T \\ A \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_Y + \alpha\mu_A + \beta\mu_T \\ \mu_T + \gamma\mu_A \\ \mu_A \end{pmatrix}, \begin{pmatrix} (\alpha + \beta\gamma)^2\sigma_A^2 + \beta^2\sigma_T^2 + \sigma_Y^2 & & \\ \gamma(\alpha + \beta\gamma)\sigma_A^2 + \beta\sigma_T^2 & \gamma^2\sigma_A^2 + \sigma_T^2 & \\ & \gamma\sigma_A^2 & \sigma_A^2 \end{pmatrix} \right)
$$

Ainsi, la loi marginale de (Y, T) est:

$$\begin{pmatrix} Y \\ T \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_Y + \alpha\mu_A + \beta\mu_T \\ \mu_T + \gamma\mu_A \end{pmatrix}, \begin{pmatrix} (\alpha+\beta\gamma)^2\sigma_A^2 + \beta^2\sigma_T^2 + \sigma_Y^2 & \\ \gamma(\alpha+\beta\gamma)\sigma_A^2 + \beta\sigma_T^2 & \gamma^2\sigma_A^2 + \sigma_T^2 \end{pmatrix}\right)$$

On voit difficilement comment identifier le paramètre β à partir de cette loi marginale. Ie. exprimer β en fonction des moments de cette loi? On peut creuser et exprimer la loi conditionnelle de $Y|T$

Rappel Si $(Z_1, Z_Z) \sim \mathcal{N}((\mu_1, \mu_2), \Sigma)$ alors $Z_1|Z_2 = z$ est aussi gaussienne de moyenne

$$m(z_2) = \mu_1 - \frac{\Sigma_{12}}{\Sigma_{22}}(z_2 - \mu_2)$$

et de variance

$$\overline{\Sigma} = \Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}}$$

Donc ici : une fois les A intégrés, $Y|T$ suit la loi:

$$\begin{aligned} Y &= \mu_Y + \alpha\mu_A + \beta\mu_T - \frac{\gamma(\alpha+\beta\gamma)\sigma_A^2 + \beta\sigma_T^2}{\gamma^2\sigma_A^2 + \sigma_T^2}(T - \mu_T + \gamma\mu_A) + \widetilde{\varepsilon} \\ &= \widetilde{\mu}_Y + \frac{\gamma(\alpha+\beta\gamma)\sigma_A^2 + \beta\sigma_T^2}{\gamma^2\sigma_A^2 + \sigma_T^2}T + \widetilde{\varepsilon} \\ &= \widetilde{\mu}_Y + \left(\beta + \frac{\gamma\alpha\sigma_A^2}{\gamma^2\sigma_A^2 + \sigma_T^2}\right)T + \widetilde{\varepsilon} \end{aligned}$$

Ainsi, si je fais une regression de Y contre T sans corriger par A, je vais estimer non pas β tout seul mais $\beta + \frac{\gamma\alpha}{\gamma^2\sigma_A^2+\sigma_T^2}$ donc je vais avoir un biais dans mon estimation.

Remarque qui rassure : si $\gamma = 0$ (pas d'effet de l'altitude sur la température), $\alpha = 0$ (par d'effet de l'altitude sur Y) ou $\sigma_A^2 = 0$ (A non aléatoire) , l'estimateur est sans biais.

Question Dans un modèle plus complexe contenant d variables $(X_1, \ldots, X_d)$ dont j'ai le DAG. Je veux identifier l'effet causal de X_1 sur X_d (au sens vu ci dessus). Supposons que je n'observe pas toutes les variables mais seulement un sous-ensemble de ces variables : $W \subset \{X_2, \ldots, X_{d-1}\}$

Cet ensemble W est-t-il suffisant pour identifier l'effet causal de X_1 sur X_d? au sens où est-ce que j'arrive à ecrire la loi de $X_d(x_1)$ à partir de la loi observationnelle $P(X_1, X_d, W)$?

4 Backdoor criterion

Establishes a sufficient condition for nonparametric identifiability of causal effects.

Recall from last session: path, blocked path.

Here we introduce backdoor path:

Definition 1 (backdoor path). *A backdoor path from X_k to X_ℓ is a path $(X_{i_1}, \ldots, X_{i_n})$ with $i_1 = k$, $i_n = \ell$, and an ingoing edge $X_{i_1} \leftarrow X_{i_2}$ (ie. the path "enters" X_k trough the backdoor).*

Example:

- $A \leftarrow B \rightarrow C$ is a backdoor path from A to C :
- $A \rightarrow B \rightarrow C$ is not a backdoor path.

Without loss of generality we consider random variables $(X_1, \ldots, X_d)$ and we wish to identify the causal effect of X_1 on X_d; ie establish that the "post-intervention" distribution of X_d can we written as a function of the "observational" distribution P.

First remark that if P is the joint distribution of $(X_1, \ldots, X_d)$ – ie. everything is observed – then we known all the conditional distributions of P and the causal effect is trivially identifiable.

Here we are interested in finding a subset $W \subset \{X_2, \ldots, X_{d-1}\}$ of variables for which the "post-intervention" distribution of X_d can be written as a function of the "observational" marginal distribution of $\{X_1, X_2\} \cup W$.

Definition 2 (Backdoor criterion)**.** *A set of variables $W \subset \{X_2, \dots, X_{d-1}\}$ satisfies the backdoor criterion relative to (X_1, X_d) if:*

- *W blocks all backdoor paths from X_1 to X_d ;*
- *W contains no descendent of X_1.*

Theorem 1. *If W satisfies the backdoor criterion, then the causal effect of X_1 on X_d is identifiable from the observational distribution of $\{X_1, X_d\} \cup W$.*

Proof. Let P be the observational distribution of $(X_1, \dots, X_d)$. By assumption P is Markov with respect to the DAG G. We write pa_i the set of indices of parents of X_i.

We assume that P admits conditional distributions of X_i given $(X_j)_{j \in \mathrm{pa}_i}$ that have densities wrt some measures given by $\mu_i(\cdot \mid (x_j)_{j \in \mathrm{pa}_i})$; so p has the density

$$p_{1:d}(x_1, \dots, x_d) = \prod_{i=1}^{d} \mu_i(x_i \mid (x_j)_{j \in \mathrm{pa}_i}).$$

We are interested in the intervention $\mathrm{do}(X_1 = x)$, which leads to the "post-interventional" density:

$$p^{\mathrm{do}}_{1:d}(x_1, \dots, x_d) = \mathbf{1}_{\{x_1 = x\}} \prod_{i=2}^{d} \mu_i(x_i \mid (x_j)_{j \in \mathrm{pa}_i}).$$

In particular we wish to show that the marginal $p^{\mathrm{do}}_{1:d}$ on X_d can be written as a function of the marginal of $p_{1:d}$ on $\{X_1, X_d\} \cup W$.

The main trick is the following. We introduce a new random variable $X_0 \in \{0, 1\}$ and a joint distribution P' for $(X_0, \dots, X_d)$ such that P' has the density

$$p'_{0:d}(x_0, \dots, x_d) = p'_0(x_0)\mu'_1(x_1 \mid (x_j)_{j \in \mathrm{pa}_1}, x_0) \prod_{i=2}^{d} \mu_i(x_i \mid (x_j)_{j \in \mathrm{pa}_i})$$

where

$$\mu'_1(x_1 \mid (x_j)_{j \in \mathrm{pa}_1}, x_0) = \begin{cases} \mu_1(x_1 \mid (x_j)_{j \in \mathrm{pa}_1}) & \text{if } x_0 = 0, \\ \mathbf{1}_{\{x_1 = x\}} & \text{if } x_0 = 1. \end{cases}$$

Observe that P' is Markov with respect to the graph G' obtained from G by adding a vertex X_0 and a directed edge $X_0 \to X_1$. Furthermore,

$$\begin{aligned} P'(X_1 \in \cdot, \dots, X_d \in \cdot \mid X_0 = 0) &= P \\ P'(X_1 \in \cdot, \dots, X_d \in \cdot \mid X_0 = 1) &= P^{\mathrm{do}}. \end{aligned}$$

Now using "obvious" notations:

$$\begin{aligned} p^{\mathrm{do}}_d(x_d) &= p'_{d|0}(x_d \mid X_0 = 1) \\ &= \int p'_{d|0,W}(x_d \mid X_0 = 1, W = w) p'_{W|0}(dw \mid X_0 = 1) \end{aligned}$$

Remark that under P', conditional to $X_0 = 1$ it is the case that $X_1 = x$ P'-almost-surely. Therefore the previous indeed equals

$$p^{\mathrm{do}}_d(x_d) = \int p'_{d|0:1,W}(x_d \mid X_0 = 1, X_1 = x, W = w) p'_{W|0}(dw \mid X_0 = 1).$$

Now we make the following two observations. Under P':

- Conditional to (W, X_1) the variables X_0 and X_d are independent. This is because in G', all the paths from X_0 to X_d must
  - either go through one of the backdoor path from X_1 to X_d, which is by assumption blocked by W ;

– either go through $X_0 \to X_1 \to$ some child of X_1, which is a chain where the middle node is conditioned on $\implies$ blocked by X_1.

It follows that

$$\begin{aligned} p'_{d|0:1,W}(x_d \mid X_0 = 1, X_1 = x, W = w) &= p'_{d|1,W}(x_d \mid X_1 = x, W = w) \\ &= p'_{d|0:1,W}(x_d \mid X_0 = 0, X_1 = x, W = w) \\ &= p_{d|1,W}(x_d \mid X_1 = x, W = w). \end{aligned}$$

- W and X_0 are independent. By assumption, no vertex of W is a descendent of X_1. This implies that all paths from X_0 to W contains a collider [draw picture!]. Hence

$$\begin{aligned} p'_{W|0}(w \mid X_0 = 1) &= p'_W(w) \\ &= p'_{W|0}(w \mid X_0 = 0) \\ &= p_W(w). \end{aligned}$$

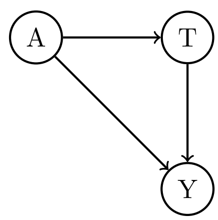
Gathering these two remarks:

$$p_d^{\mathrm{do}}(x_d) = \int p_{d|1,W}(x_d \mid X_1 = x, W = w) p_W(dw).$$

Note that the rhs of the last display is solely a function of the marginal of P on (X_1, X_d, W). $\square$

4.1 Applications

4.1.1 Application 1 : modèle (Y, T, A)



Effet causal de T sur Y.
Backdoor paths

- $T \leftarrow A \to Y$: open

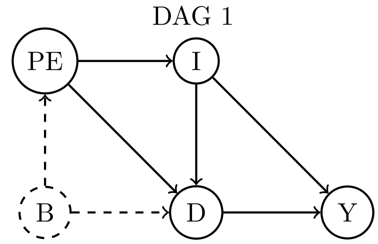
Si on choisit $W = \emptyset$, il existe un open backdoor ne passant pas par W. Le critère ne s'applique pas (mais il n'est pas suffisant:!)

4.1.2 Application 2 : modèle plus complexe

From site mixtape

- Y = earning
- D = level of education
- I = family income
- PE = parental education
- B = background : genetics, environment

On veut identifier l'effet causal de D sur Y.



Effet causal de $D->Y$
Backdoor paths

- $D \leftarrow I \rightarrow Y$
- $D \leftarrow PE \rightarrow I \rightarrow Y$
- $D \leftarrow B \rightarrow PE \rightarrow I \rightarrow Y$

Si on pose $W = \{I\}$: tous les open Backdoor paths passent pas W. Donc il suffit de prendre I en compte avoir une estimation "sans biais" de l'effet causal...

References

[Cam20] John Campbell. *Causation in Psychology.* Harvard University Press, 2020.

[GCL+21] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. A Survey of Learning Causality with Data: Problems and Methods. *ACM Computing Surveys*, 53(4):1–37, July 2021. arXiv:1809.09337 [cs].

[GZS19] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10:524, June 2019.

[HHG24] Uzma Hasan, Emam Hossain, and Md Osman Gani. A Survey on Causal Discovery Methods for I.I.D. and Time Series Data, March 2024. arXiv:2303.15027 [cs].

[IR15] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

[MSK+21] Raha Moraffah, Paras Sheth, Mansooreh Karami, Anchit Bhattacharya, Qianru Wang, Anique Tahir, Adrienne Raglin, and Huan Liu. Causal Inference for Time series Analysis: Problems, Methods and Evaluation, February 2021. arXiv:2102.05829 [cs].

[Nea] Brady Neal. Introduction to Causal Inference.

[NPR+22] Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. Methods and tools for causal discovery and causal inference. *WIREs Data Mining and Knowledge Discovery*, 12(2):e1449, March 2022.

[Pea00] Judea Pearl. *Causality.* Cambridge University Press, New York, 2000.

[Sha18] Ramalingam Shanmugam. Elements of causal inference: foundations and learning algorithms. *Journal of Statistical Computation and Simulation*, 88(16):3248–3248, November 2018.

[SJH+] Shohei Shimizu, Ism Ac Jp, Patrik O Hoyer, Patrik Hoyer, Aapo Hyvarinen, Aapo Hyvarinen, Antti Kerminen, and Antti Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery.

[Wag] Stefan Wager. STATS 361: Causal Inference.

[Zha] Qingyuan Zhao. Lecture Notes on Causal Inference.

[ZS23] Alessio Zanga and Fabio Stella. A Survey on Causal Discovery: Theory and Practice, May 2023. arXiv:2305.10032 [cs].