Inférence de réseaux

Cours 2 - graphes non dirigés

Olivier Goudet

LERIA, Université d'Angers

19 janvier 2022

Organisation générale du cours

4 séances de 2 heures de CM :

1. Cours 1 - S. Aubourg
   - Introduction aux réseaux de gènes.
2. Cours 2 - O. Goudet
   - Introduction à la causalité.
   - Notions d'indépendance entre différentes variables.
   - Graphes non dirigés.
3. Cours 3 - O. Goudet
   - Graphes dirigés.
   - Causalité paire à paire.
4. Cours 4 - O. Goudet
   - Méthodes d'inférence de réseaux utilisées en bioinformatique.

Objectif général du cours

- Présenter la notion d'inférence de graphes non dirigés et dirigés à partir de données.
- C'est en lien avec le cours que vous avez eu cette année :
  - Régression linéaire en Grande Dimension - Fabien Panloup
- Conférence 03/02 - institut pasteur - modèles graphiques.
- Mettre en perspective certaines méthodes utilisées dans la littérature pour l'inférence de réseaux biologiques.
- Application en pratique de ces méthodes sur un challenge avec des données génomiques simulées et réelles (4 séances de TP de 2 heures). Plate-forme CodaLab.

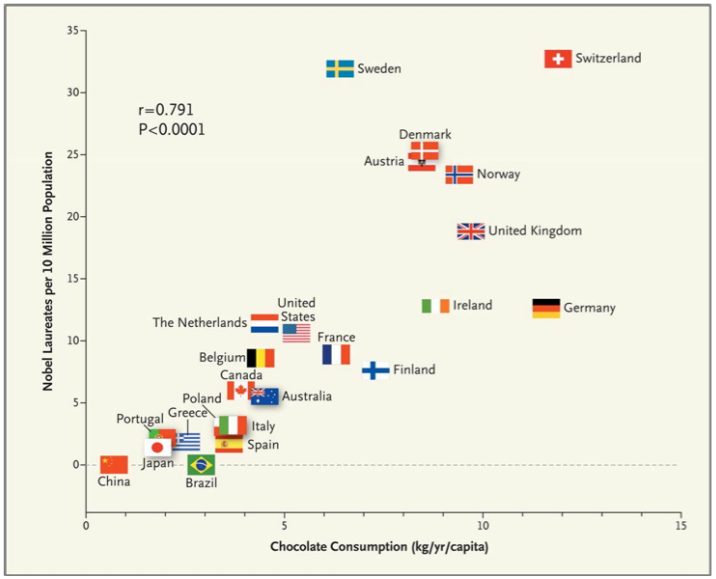
Section 1

Introduction

Buts des modèles causaux

- Enjeu dans de nombreux domaines : comprendre le processus génératif des données avec des modèles causaux.
  - Va au delà de la prévision et de la simple notion de dépendance statistique : **corrélation n'implique pas causalité !**
  - En biologie : **réseau de co-expression (GCN)** $\rightarrow$ **réseau de régulation (GRN)**
  - Permet de mieux expliquer des phénomènes.
  - Permet de réaliser des actions qui vont avoir des impacts sur des variables cibles.
- Par exemple en biologie : est-ce que modifier le fonctionnement d'un gène donné va avoir un impact sur la résistance de la plante à des éléments pathogènes ?
- On va voir quelques exemples où des prescriptions qui ne s'appuieraient pas sur la connaissance du "vrai" graphe causal sous-jacent peuvent mener à des interprétations erronées.

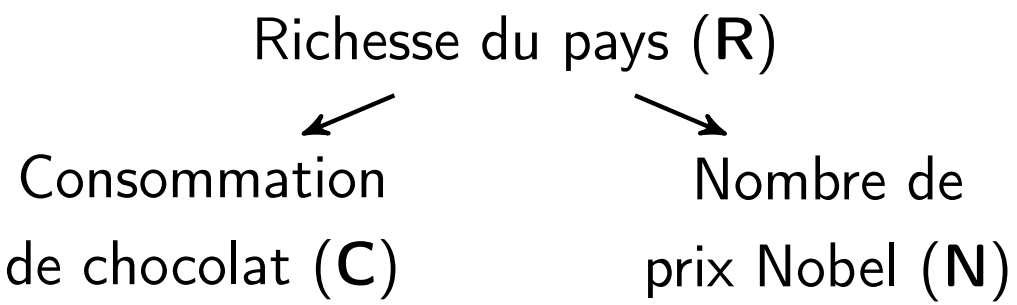
Exemple 1 (I.Guyon) : faut-il distribuer du chocolat aux chercheurs pour obtenir plus de prix Nobel ?



F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

- Corrélation positive et très significative entre la consommation de chocolat (en abscisses) et le nombre de prix Nobel (en ordonnées) pour cet ensemble de pays.
- Est-ce qu'il existe une relation causale entre ces deux variables ?
- Sinon quelle explication proposez vous ?

Une explication causale possible



- Ici en fait les deux variables **C** et **N** sont dépendantes, probablement car elles sont toutes les deux la cause d'une variable commune cachée, la richesse du pays (**R**): $C \not\perp\!\!\!\perp N$.
- Par contre, connaissant la richesse du pays, ces deux variables sont sûrement indépendantes : $C \perp\!\!\!\perp N | R$.

Pour le vérifier : effectuer une expérience randomisée !

Faire un test avec des chercheurs pendant 10 ans :

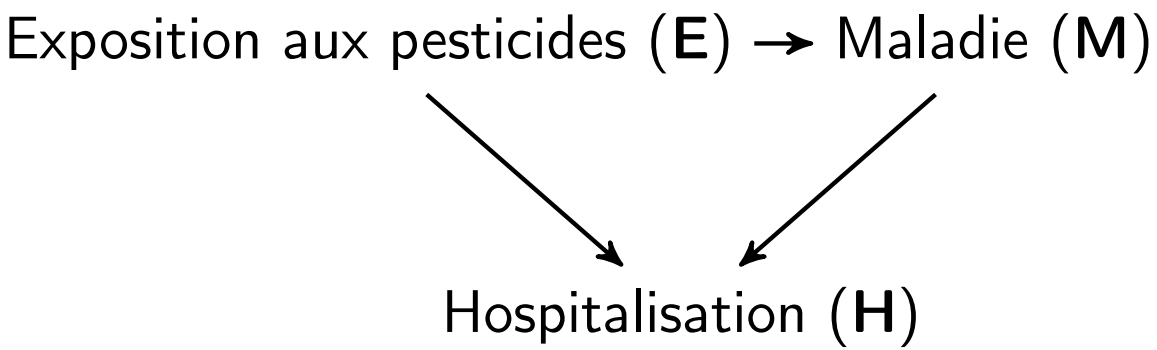
- Choisir aléatoirement la moitié d'entre eux et leur donner du chocolat.
- Donner des pommes aux autres.
- Comparer les nombres de prix Nobel obtenus dans les deux groupes.

Exemple 2 (A. Aussem) : paradoxe de Berkson (1946)

	Exposé		Non exposé	
	Malade	non malade	Malade	non malade
Hospitalisé	22	600	8	200
Non Hospitalisé	8	400	12	800

- La prévalence d'un cancer est de 2.9% $((22+8)/(22+8+600+400))$ parmi les personnes exposées aux pesticides et de 1.9% parmi les personnes non-exposées.
- Parmi les personnes exposées environ 60.4% des personnes sont hospitalisées, et parmi les personnes non exposés 20.4% des personnes sont hospitalisées.
- Parmi les personnes malades 60% sont hospitalisées. Parmi les personnes non malades 40% sont hospitalisées.
- Parmi les personnes hospitalisées, la prévalence de la maladie est de 3.5% parmi les personnes exposées au risque et de 3.8% parmi les patients non-exposés.
- En regardant uniquement les personnes à l'hôpital et non pas le groupe entier, on pourrait conclure que le fait d'être exposé aux pesticides réduit en fait le risque de cancer...

Explication causale

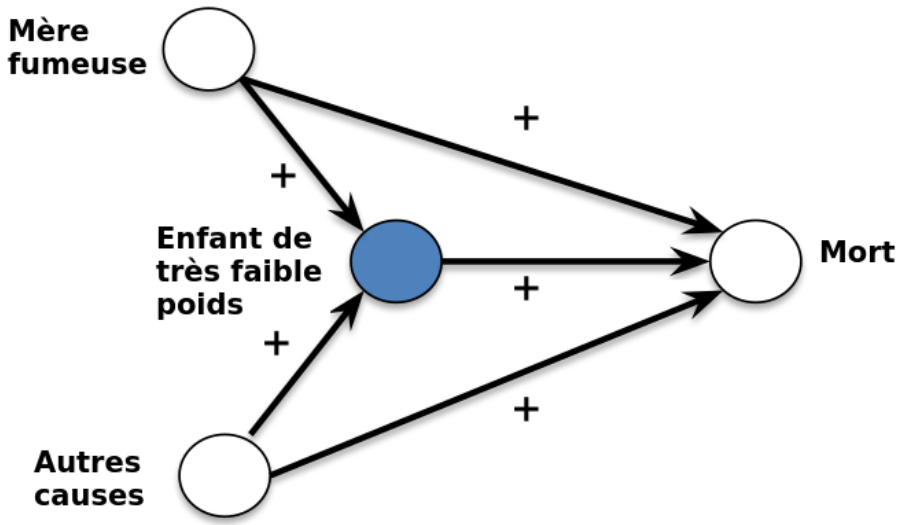


- Conditionnellement au fait d'être hospitalisé (**H**) les variables Maladie (**M**) et Exposition aux pesticides (**E**) deviennent dépendantes négativement.
- Alors que dans la population générale (si on ne conditionne pas sur (**H**)), il peut y avoir un impact positif de (**E**) sur (**M**).

Exemple 3 (A. Aussem) : paradoxe du poids des naissances (1967)

- Les enfants de mères fumeuses sont plus susceptibles de donner naissance à un enfant de très faible poids.
- Les enfants de très faible poids ont un taux de mortalité beaucoup plus important que les autres.
- Contrairement à ce qu'on pourrait penser, les enfants de très faible poids, mais nés de mères fumeuses ont un taux de mortalité plus bas que les autres enfants de très faible poids.
- Conclusion : avoir une mère fumeuse est bénéfique pour la santé de l'enfant !

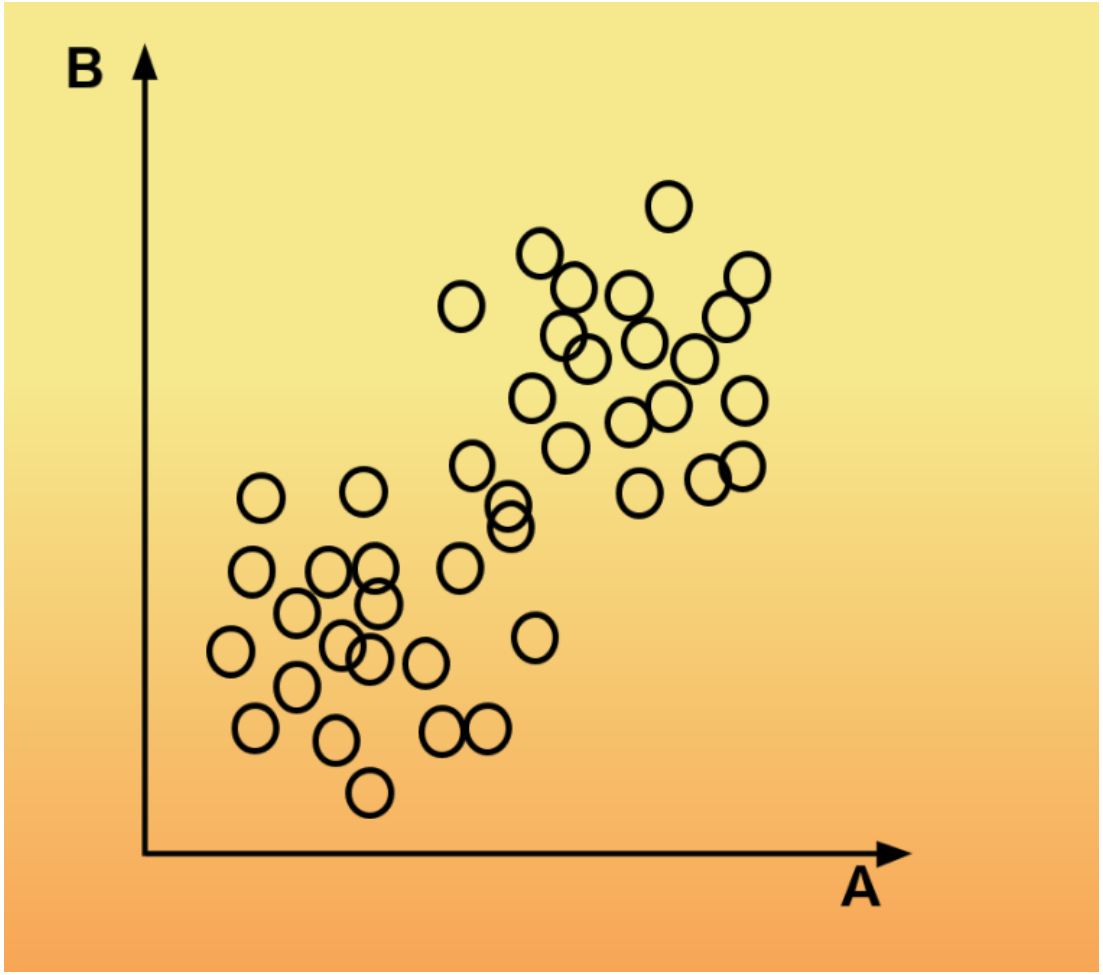
Explication causale



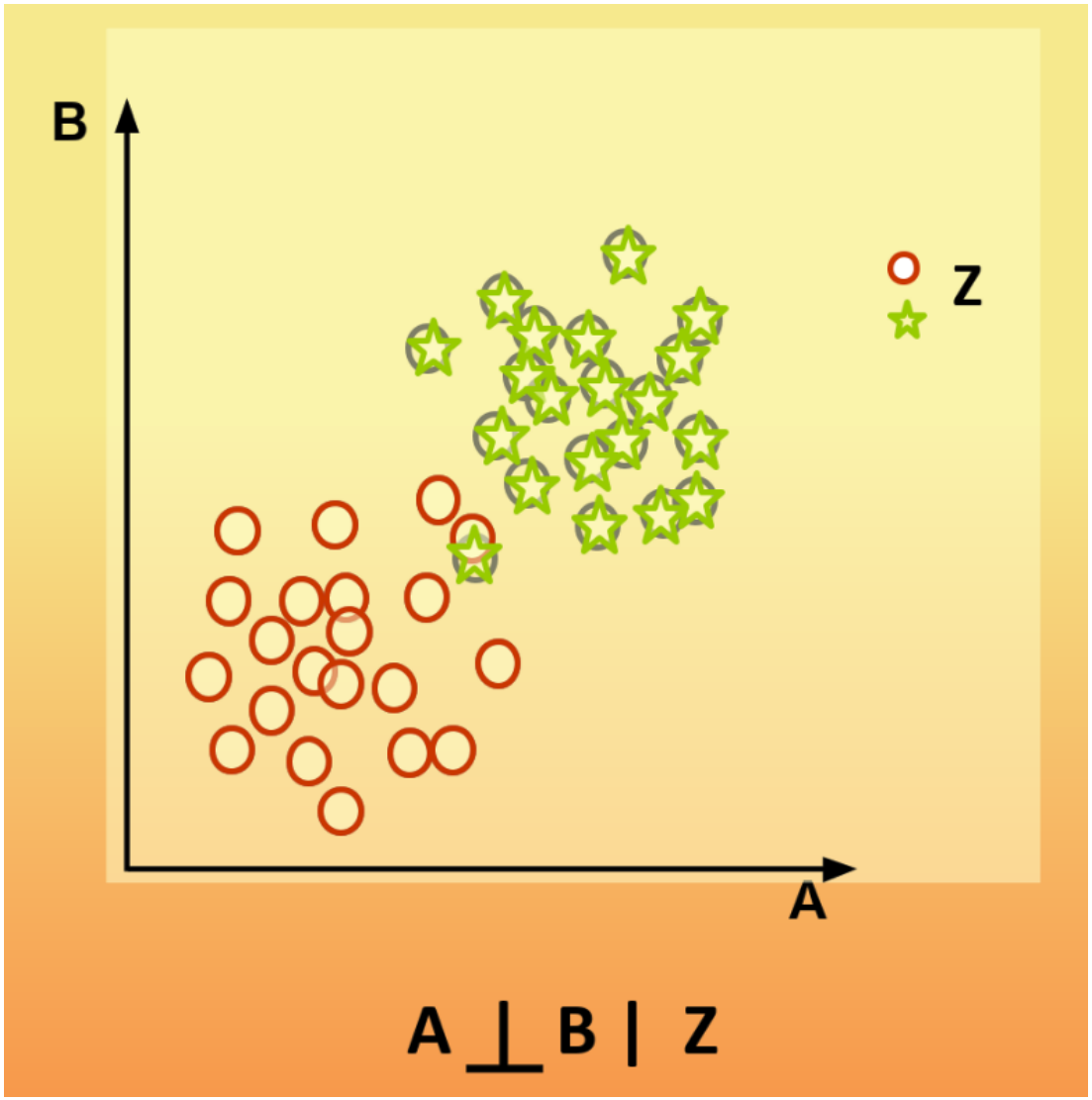
- Fumer peut être nuisible dans le sens où cela contribue au risque de sous-poids de l'enfant, mais d'autres causes de sous-poids peuvent être encore plus néfastes sur le taux de survie de l'enfant (ex : graves problèmes génétiques).
- Si on considère un enfant de faible poids, savoir que sa mère est fumeuse réduit en fait la probabilité qu'une autre cause soit présente.

Exemple 4 (I. Guyon) : paradoxe de Simpson (1899)

Est-ce que vous voyez un lien de dépendance statistique entre A et B ?
Est-ce qu'il existe un lien de cause à effet entre A et B ?

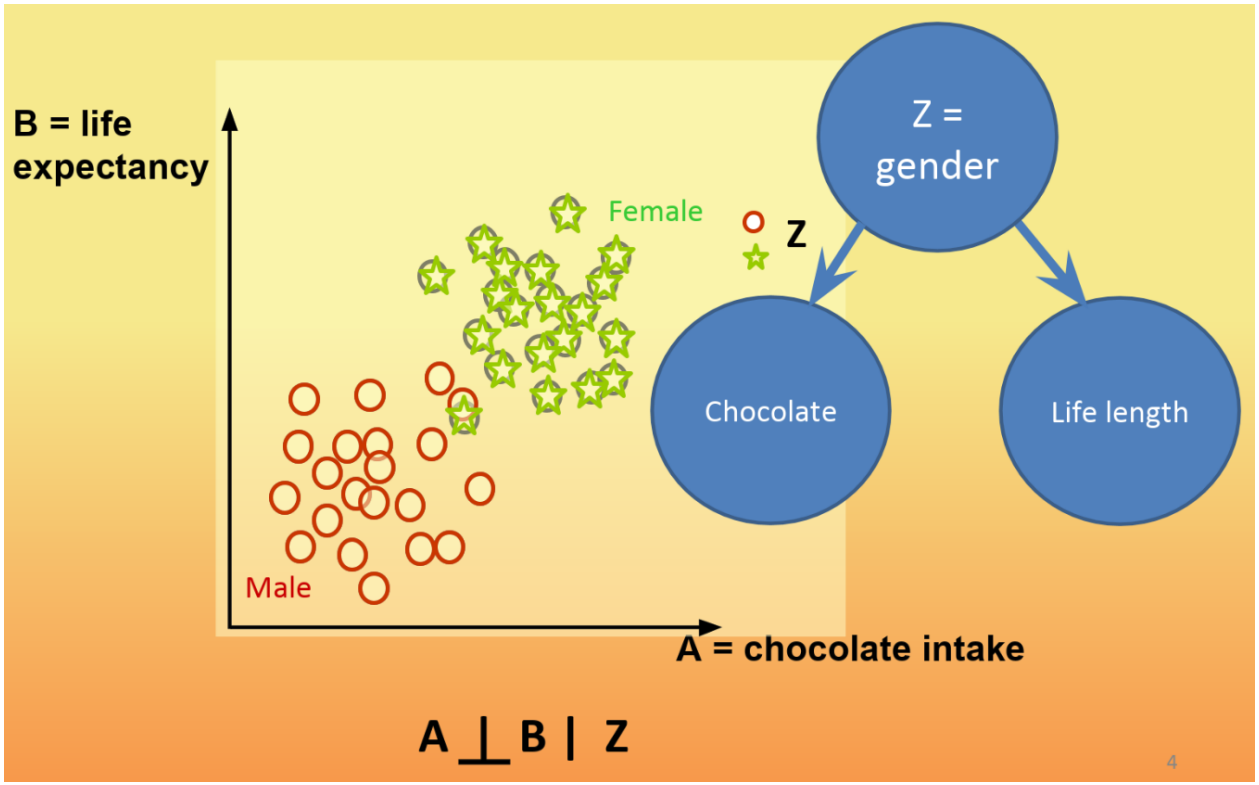


Paradoxe de Simpson



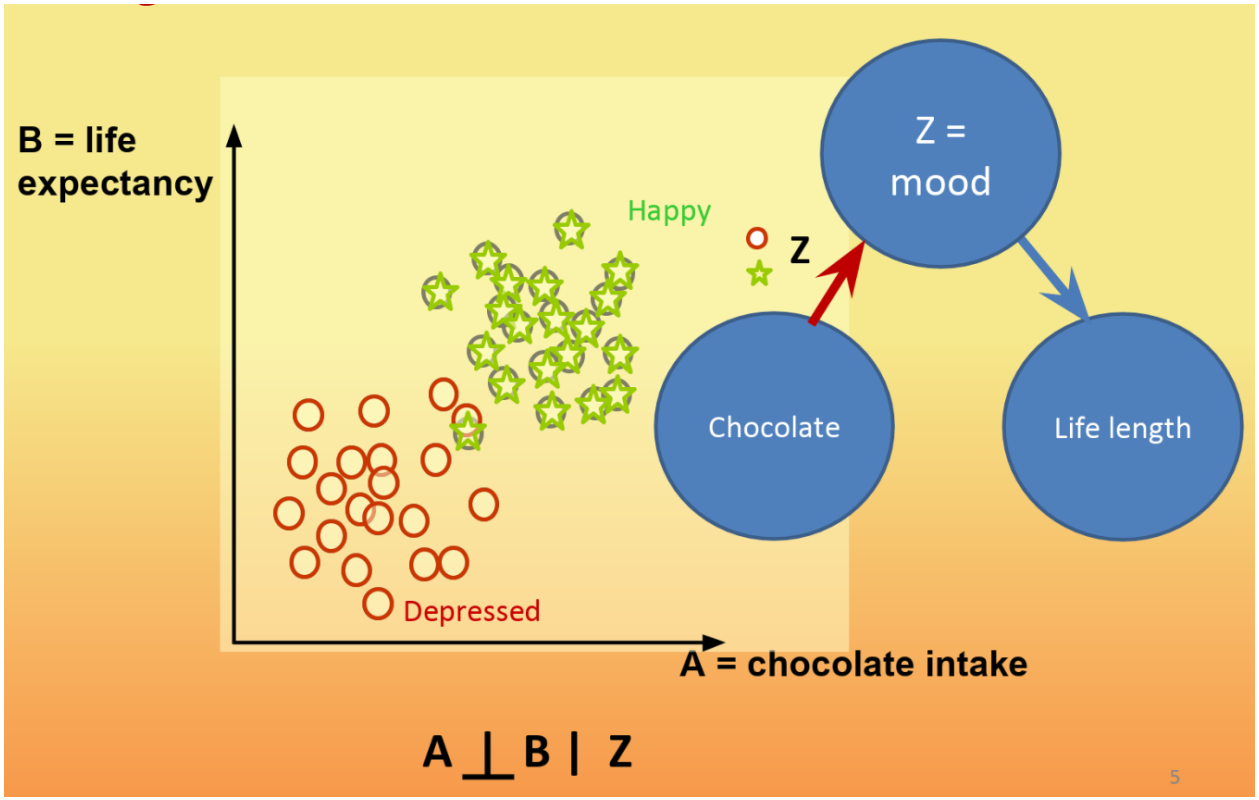
Paradoxe de Simpson

Exemple de cas où il n'y a pas de relation de cause à effet entre A et B.



Paradoxe de Simpson

Exemple de cas où il y a effectivement une relation de cause à effet entre A et B.



Paradoxe de Simpson

- Ils existent plusieurs modèles causaux explicatifs cohérents avec les tests d'indépendance statistiques effectués : $A \not\perp\!\!\!\perp B$ et $A \perp\!\!\!\perp B|Z$.
- On verra plus tard cette notion de classes d'équivalence de Markov: $A \to Z \to B$, $B \to Z \to A$ et $A \leftarrow Z \to B$.
- On verra aussi qu'il existe quand même un moyen de retrouver le vrai graphe causal en cherchant le modèle le plus "simple" qui permet d'expliquer les données (Rasoir d'Ockham).
- **Rasoir d'Ockham** : "Les hypothèses suffisantes les plus simples doivent être préférées" (principe heuristique fondamental en science).

Comment inférer un graphe causal ?

- La voie royale :
  - Interventions et essais randomisés.
  - Ce qui est fait pour tester les effets des médicaments par exemple.
- Mais dans beaucoup de domaines, les interventions sont :
  - impossibles (climat)
  - non éthique (convaincre les gens de fumer)
  - trop chères (économie)
  - longues à mettre en place expérimentalement (biologie)

Deux principaux problèmes dans le domaine de la causalité

Si on ne s'appuie pas sur des interventions, il y a deux grands problèmes en causalité explorés dans la littérature scientifique :

1. **Problème 1** (*Causal discovery* ou *Structure learning*): Il s'agit de trouver le graphe causal représentant les relations de cause à effet entre des variables observées à partir de données d'observation. Parfois on se limite à la découverte d'une classe d'équivalence de graphe. On donne aussi souvent en pratique des scores de confiance à chaque relation causale que l'on peut inférer.

2. **Problème 2** (*Do-calculus*): Etant donné des données d'observation et un graphe causal qui relie des variables observées, trouver la distribution d'intervention d'une variable aléatoire X_j en réponse à la modification d'une autre variable X_i par une intervention externe (manipulation) en contrôlant les effets des autres variables observées $\mathbf{X}_{\setminus ij} := X_{\{1,\dots,d\}\setminus i,j}$:

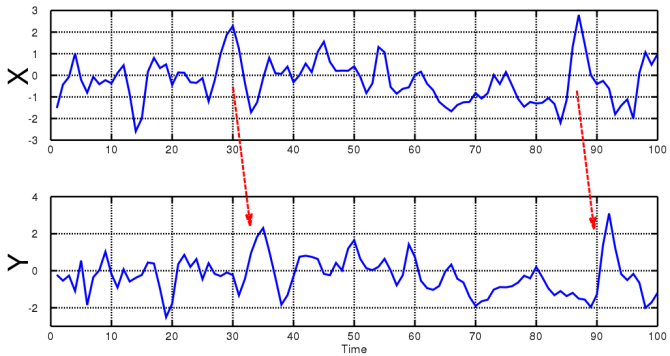
$$P_{X_j|\mathrm{do}(X_i=x,\mathbf{X}_{\setminus ij}=\mathbf{c})} \tag{1}$$

Inférence de la structure d'un graphe causal à partir de données

- Apprendre un graphe de causalité sans avoir besoin de faire de nouvelles expériences.
- Modèles fondées sur des données déjà acquises.
- Construire un modèle plausible qui explique la façon dont les données ont pu être générées.
  $\rightarrow$ sélection de modèles.
- Inférence d'un graphe causal.
  $\rightarrow$ tâche de "rétro-ingénierie".

Différents types de données disponibles pour l'inférence d'un graphe causal (1/2)

- Cas 1 : des séries temporelles $[x_{1,t}, x_{2,t}, ..., x_{d,t}]$. cf. notion de causalité au sens de Granger. On ne parlera pas de ce cas ici.



- Cas 2 : des données d'intervention. On sait qu'une ou plusieurs variables précises ont été modifiées et on a mesuré leurs effets sur les autres variables étudiées.
- Cas 3 : des données observées sans indications de temps mais supposées échantillonnées de façon iid suivant une distribution jointe inconnue.

Différents types de données disponibles pour l'inférence d'un graphe causal (2/2)

Types de variables observées :

- Cas A : variables continues - par exemple, la mesure de l'expression d'un gène relevé sur une puce à ADN.
- Cas B : variables discrètes - par exemple, le contexte de l'expérience (1 jour, 0 nuit)

Cas étudié dans ce cours

- Certaines méthodes de la littérature combinent ces différents types de données: Cas 1, 2 et 3, mais aussi Cas A et B (données mixtes).
- Dans ce cours, on considère que l'on n'a pas d'information de temps, qu'il n'y a pas d'interventions connues et que toutes les variables sont continues (Cas 3A).
- C'est le cas qui nous intéresse dans ce cours pour l'étude des matrices d'expression des gènes.

Inférence d'un graphe causal de régulation génétique à partir de données d'observation continues

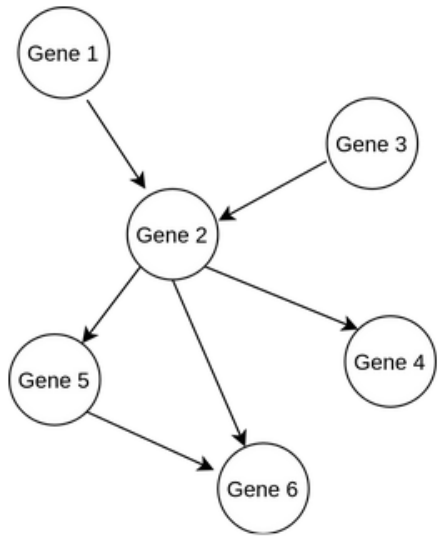
Matrice d'expression du génome de la plante *Arabidopsis thaliana*

26 374 gènes

1042 Expériences

	Gène 1	Gène 2	Gène 3	Gène 4	Gène 5	Gène 6	...
Expérience 1	1,13	1,49	0,78	1,71	2,11	5,09	
Expérience 2	1,10	0,58	1,25	0,37	1,04	4,56	
Expérience 3	2,13	1,40	0,40	1,56	2,03	5,48	
Expérience 4	0,70	0,78	0,85	0,62	1,51	4,25	
Expérience 5	1,52	1,28	0,88	0,85	1,35	5,55	
Expérience 6	2,01	2,04	-0,04	0,10	3,14	4,27	
Expérience 7	0,74	0,54	1,45	0,20	2,52	5,92	
Expérience 8	0,23	0,85	1,61	0,15	2,14	6,04	
Expérience 9	1,64	1,64	1,65	1,52	3,53	6,25	
Expérience 10	1,10	1,37	1,28	0,97	3,02	5,71	
Expérience 11	0,77	0,92	0,83	0,73	3,14	6,16	
Expérience 12	1,39	1,73	0,93	1,24	3,49	5,88	
Expérience 13	1,41	1,65	1,32	1,18	3,88	7,02	
Expérience 14	1,57	2,25	0,61	0,80	1,74	4,80	
Expérience 15	0,98	1,67	0,42	0,13	1,96	4,64	
Expérience 16	2,23	1,71	0,83	0,85	1,91	7,39	
...							

Exemple de graphe de régulation génétique inféré



Données simulées et données réelles

Pendant les TP, on va exploiter deux types de données :

- **Des données simulées** : on connaît le "vrai" graphe qui a généré les données. Le but est de le retrouver.
  - $\rightarrow$ Permet de comparer des méthodes d'inférence de réseau avec un score global.
- **Des données réelles** : on ne connaît rien du vrai graphe (ou très peu de choses). On cherche à inférer un graphe qui permettra de suggérer des relations causales entre les gènes qui seront à valider par des expériences.
  - Base de mesures d'expression des gènes relevées pour un grand nombre d'expériences menées au cours des dernières années à l'IRHS.
  - Suggestion de relations entre les gènes (pour faire de nouvelles découvertes ?...)

En plus des données...

En pratique, tenir compte des connaissances du domaine peut aussi être utile.

- Par exemple : restreindre la recherche des causes de certains gènes parmi une liste connue de facteurs de transcription. Cas des challenges Dream4 (2009) et Dream5 (2010).
- Restreindre la recherche des mécanismes causaux à certaines classes de fonctions ou de processus si on a des a priori dessus.
  - Exemple : on sait que certains mécanismes physiologiques correspondent à des effets linéaires ou bien se déclenchent au delà d'un certain seuil.

Plan du cours aujourd'hui

1. Notions d'indépendance.
2. Notions d'indépendance conditionnelle
3. Modèles graphiques non dirigés.

Section 2

Notions d'indépendance entre deux variables

Rappel : indépendances entre deux variables

Si (X, Y) est une paire de variable aléatoire de densité de probabilité jointe $p(x, y)$, X et Y sont indépendante ($X \perp\!\!\!\perp Y$) si et seulement si $p(x, y)$ peut être décomposée comme le produit des deux densités marginales :

$$X \perp\!\!\!\perp Y \iff p(x, y) = p(x)p(y)$$

.

Information mutuelle entre deux variables

- L'information mutuelle entre deux variables continues X et Y est définie par :

$$I(X,Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} d_x d_y \tag{2}$$

- L'information mutuelle mesure l'information partagée par X et Y, ou bien à quel point connaissant une des deux variables on peut inférer la seconde.
- L'information mutuelle $I(X,Y)$ est égale à zéro si et seulement si X et Y sont indépendantes.
- L'information mutuelle peut aussi être exprimée avec la divergence de Kullback-Leibler entre $p(x,y)$ et $p(x)p(y)$:

$$I(X,Y) = D_{KL}(p(x,y) \,\|\, p(x)p(y)) \tag{3}$$

Information mutuelle en fonction de l'entropie

- L'information mutuelle entre deux variables continues X et Y peut s'exprimer en terme d'entropie :

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

- L'entropie d'une variable aléatoire continue X avec une densité de probabilité $p(x)$ est :

$$H(X) = -\int_{\mathbb{R}} p(x) \log p(x) dx \tag{4}$$

Calcul de l'entropie dans le cas Gaussien

- Entropie d'une variable aléatoire X qui suit une loi normale, $X \sim \mathcal{N}(0, \sigma^2)$. $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{x^2}{2\sigma^2})$.

$$h(X) = -\int_{\mathbb{R}} p(x) \log p(x) dx \tag{5}$$

$$= -\int_{\mathbb{R}} p(x) \log \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{x^2}{2\sigma^2}) dx \tag{6}$$

$$= -\int_{\mathbb{R}} p(x) \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{x^2}{2\sigma^2} \right] dx \tag{7}$$

$$\tag{8}$$

- On sait que : $\int_{\mathbb{R}} p(x) = 1$ et que $\int_{\mathbb{R}} x^2 p(x) = \mathbb{E}[X^2] = \sigma^2$ (moment d'ordre 2).
- Donc $h(X) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{\sigma^2}{2\sigma^2} = \frac{1}{2}\log(2\pi e\sigma^2)$
- Même valeur de l'entropie pour $X \sim \mathcal{N}(\mu, \sigma^2)$ (exercice)

Calcul de l'entropie pour une loi normale multivariée

- Densité de probabilité d'un vecteur Gaussien $\mathbf{X} = (X_1, ..., X_d)$ de dimension d. $\mathbf{X} = \mathcal{N}(0, \Sigma)$:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x}} \tag{9}$$

- (Exercice) Retrouver la formule de l'entropie de $\mathbf{X}$:

$$h(\mathbf{X}) = \frac{1}{2} \log((2\pi e)^d |\Sigma|) \tag{10}$$

- Dans le cas bivarié :

$$h(X, Y) = \frac{1}{2} \log((2\pi e)^2 (\sigma_x^2 \sigma_y^2 - \sigma_{xy}^2)) \tag{11}$$

- σ_{xy} est la covariance entre X et Y.

Calcul de l'information mutuelle dans le cas Gaussien

- Pour deux variables Gaussiennes $X \sim \mathcal{N}(\mu_x, \sigma_x)$ et $Y \sim \mathcal{N}(\mu_y, \sigma_y)$:

$$
\begin{aligned}
I(X,Y) &= H(X) + H(Y) - H(X,Y) \\
&= \frac{1}{2}\log\left(\frac{\sigma_x^2\sigma_y^2}{\sigma_x^2\sigma_y^2 - \sigma_{xy}^2}\right) \\
&= -\frac{1}{2}\log\left(\frac{\sigma_y^2 - \sigma_{xy}^2/\sigma_x^2}{\sigma_y^2}\right) \\
&= -\frac{1}{2}\log\left(1 - (\frac{\sigma_{xy}}{\sigma_x\sigma_y})^2\right) \\
&= -\frac{1}{2}\log(1-\rho_{x,y}^2)
\end{aligned}
$$

- $\rho_{x,y}$ est le coefficient de corrélation de Pearson entre X and Y:

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Comment calculer l'information mutuelle entre deux variables continues dans le cas non Gaussien ? (1/2)

- Dans le cas non Gaussien, il peut être difficile d'estimer l'information mutuelle pour des variables continues. Il n'y a généralement pas de formule analytique simple.
- Une façon directe et répandue pour estimer cette information mutuelle est de découper les supports des variables X et Y en intervalles de taille finie et d'approximer $I(X, Y)$ par:

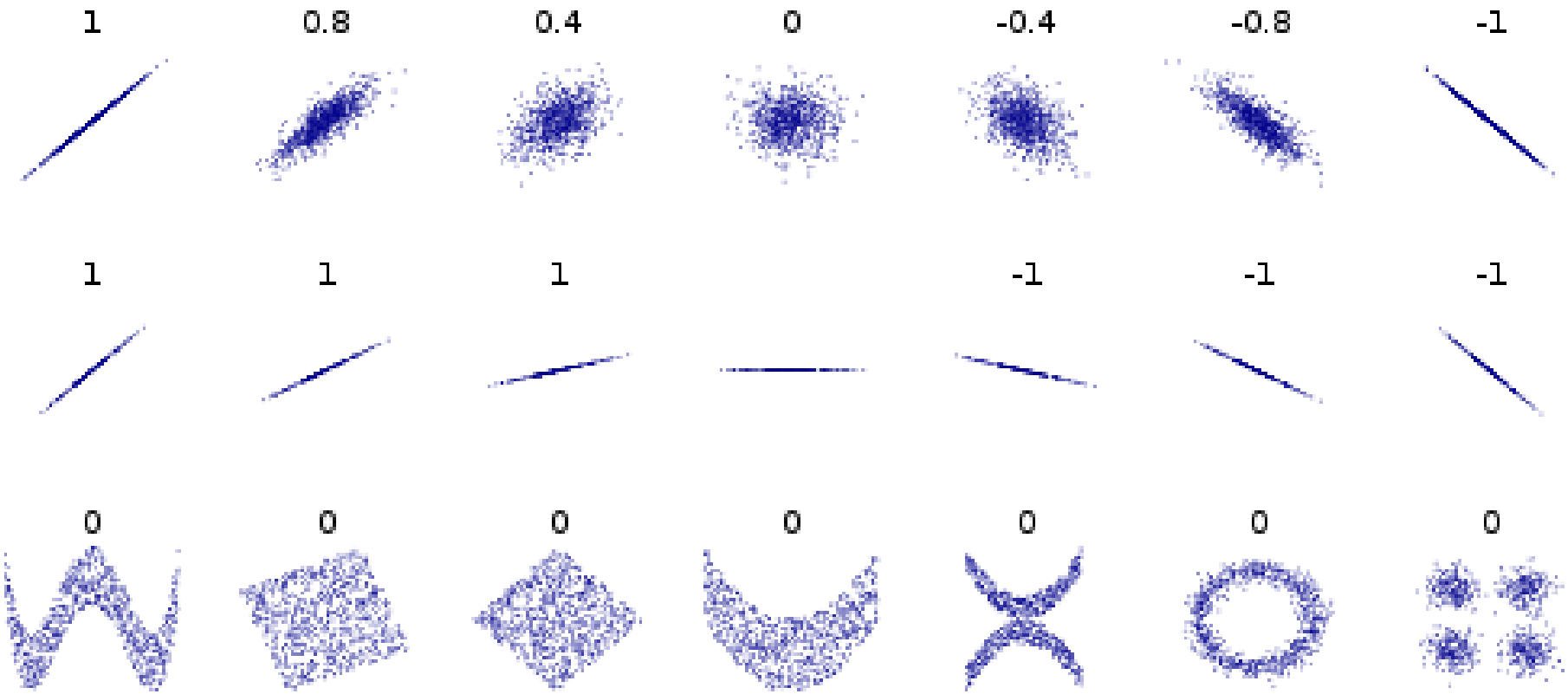
$$I(X, Y) \approx I_{binned}(X, Y) = \sum_{i,j} f(i, j) \log \frac{f(i, j)}{f_x(i) f_y(j)} \tag{12}$$

Avec $f_x(i) = \int_i p(x)dx$, $f_y(j) = \int_j p(y)dy$ et $f(i, j) = \int_i \int_j p(x, y)dxdy$. $\int_i$ correspond à l'intégrale sur l'intervalle i.

Comment calculer l'information mutuelle entre deux variables continues dans le cas non Gaussien ? (2/2)

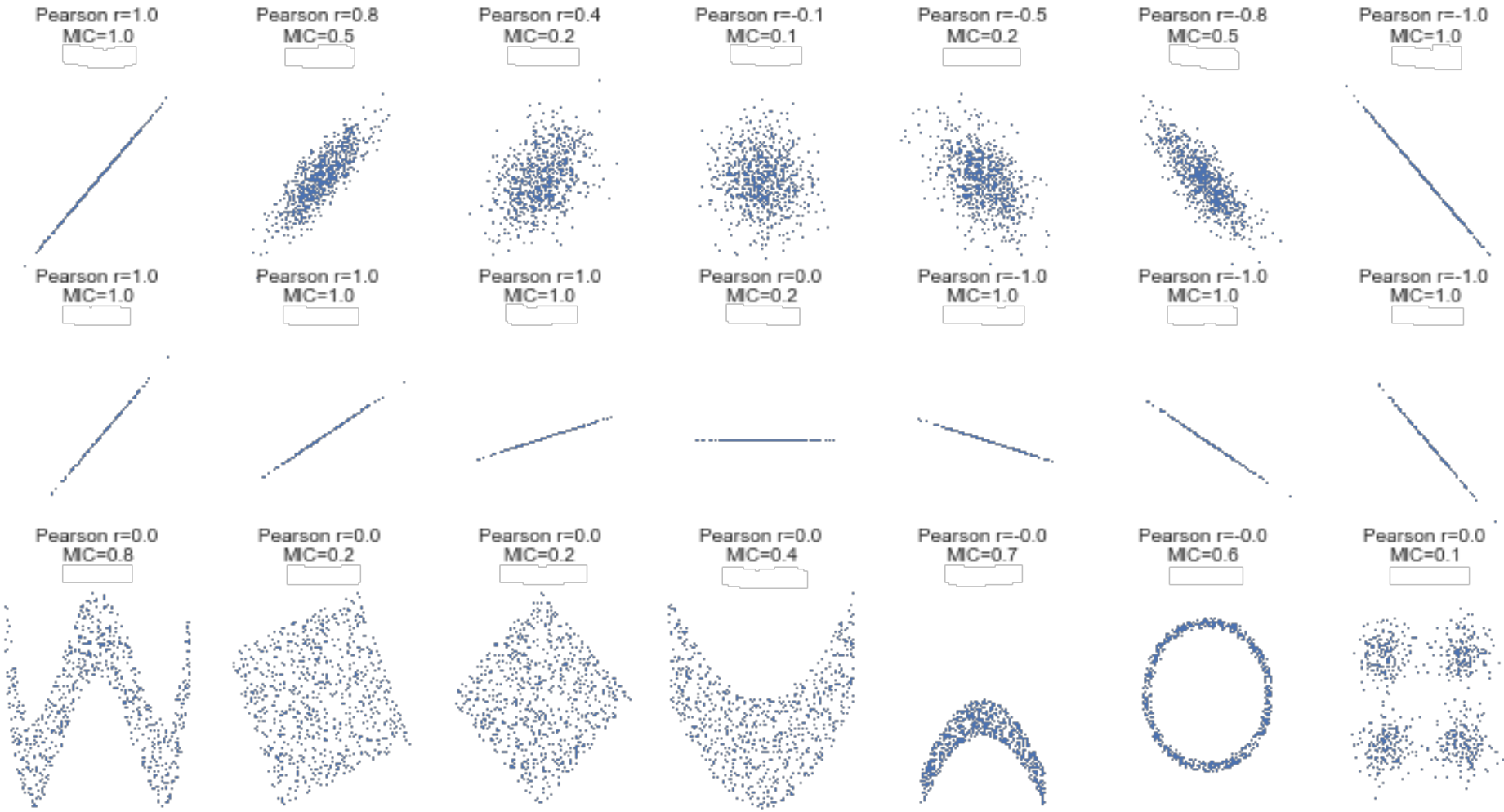
- Une estimation de $I_{binned}(X, Y)$ est ensuite obtenue en comptant le nombre de points dans chaque boîte.
- Si $n_x(i)$ (resp. $n_y(j)$ est le nombre de points dans l'intervalle i de X (resp. j de Y) et $n(i, j)$ le nombre de points dans l'intersection.
- On peut approximer $f_x(i) \approx n_x(i)/N$, $f_y(j) = n_y(j)/N$ et $f(i, j) = n(i, j)/N$.
- Des estimateurs optimisés (Fraser and Swinney, 1986; Darbellay and Vajda, 1999) utilisent des tailles adaptatives d'intervalles pour découper X et Y, de façon à avoir autant de points dans chaque boîte (i, j).

En pratique : coefficients de corrélation dans différents cas

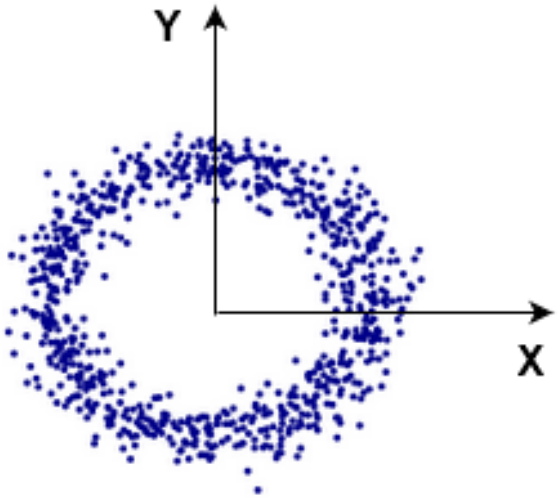


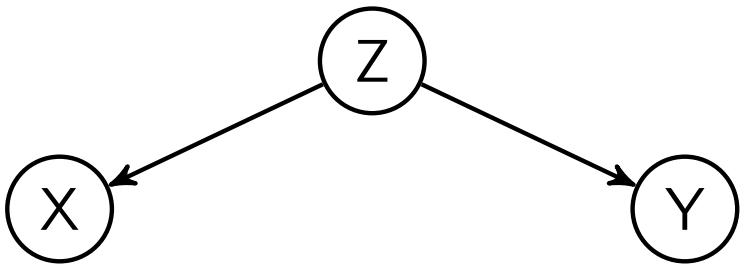
Source : wikipedia

Estimation de l'info mutuelle pour ces différents cas



Exemple de paire avec une dépendance non linéaire





Modèle génératif sous-jacent:

$$Z, E_X E_Y \sim \text{Uniform}(0,1),$$
$$X \leftarrow sin(\pi * Z) + 0.1 * E_X,$$
$$Y \leftarrow cos(\pi * Z) + 0.1 * E_Y$$

Tests d'indépendance statistiques :

- $\rho_{x,y} = 0$
- $I_{binned}(X, Y) \approx 0.6$

Section 3

Notions d'indépendance conditionnelle

Indépendance conditionnelle

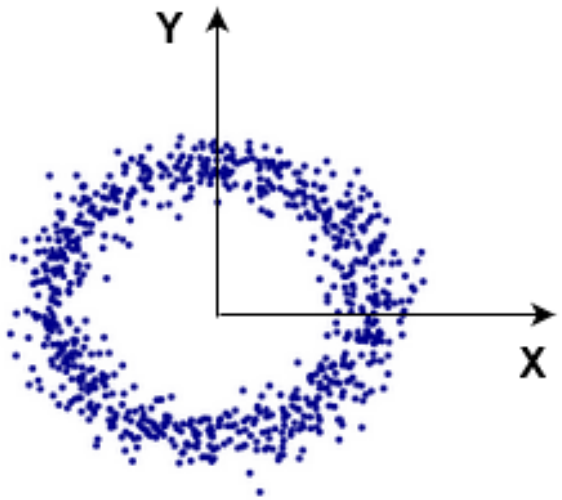
X et Y sont indépendantes conditionnellement à un ensemble de variables Z : notation $X \perp\!\!\!\perp Y|Z$.

$$X \perp\!\!\!\perp Y|Z \iff p(x,y|z) = p(x|z)p(y|z) \tag{13}$$

Autre caractérisation utile par la suite :

$$X \perp\!\!\!\perp Y|Z \iff \exists(g,h)\ , p(x,y,z) = g(x,z)h(y,z) \tag{14}$$

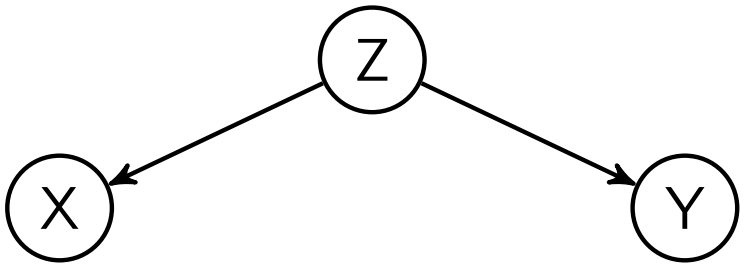
Si on reprend le cas du rond



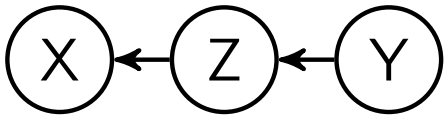
$$p(x, y|z) = \frac{p(x, y, z)}{p(z)}$$
$$= \frac{p(x|z)p(y|z)p(z)}{p(z)}$$
$$= p(x|z)p(y|z)$$

Donc :

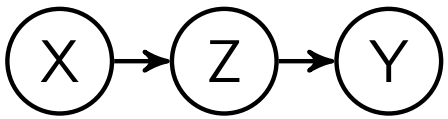
$$X \perp\!\!\!\perp Y|Z$$



Deux autres cas de trois variables avec $X \perp\!\!\!\perp Y|Z$

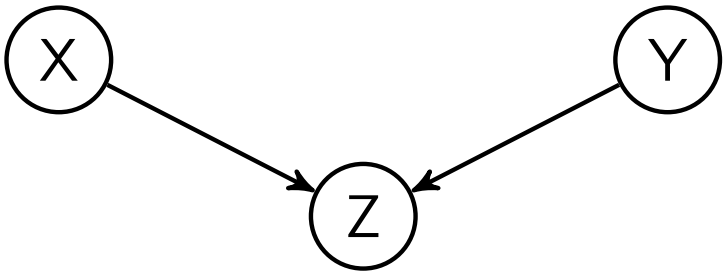


$$
\begin{aligned}
p(x,y|z) &= \frac{p(x,y,z)}{p(z)} \\
&= \frac{p(x|z)p(z|y)p(y)}{p(z)} \\
&= \frac{p(x|z)p(y|z)p(z)}{p(z)} \\
&= p(x|z)p(y|z)
\end{aligned}
$$



Idem dans l'autre sens en permutant X et Y

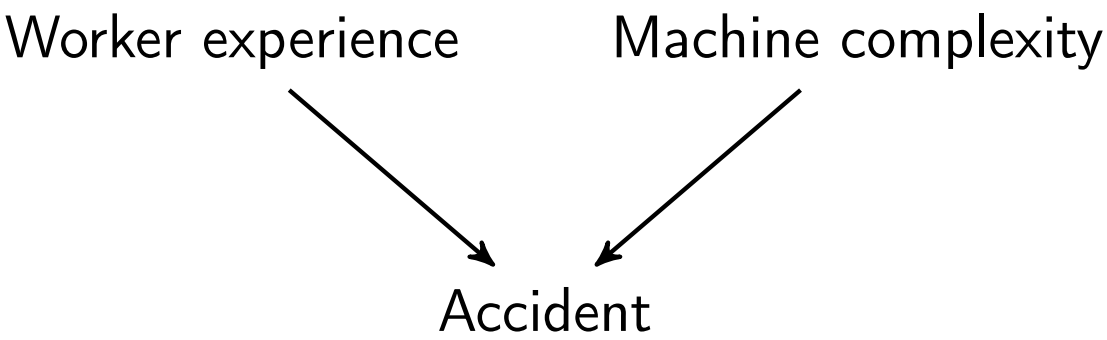
Cas de la v-structure $X \perp\!\!\!\perp Y$ e $X \not\perp\!\!\!\perp Y|Z$



$$p(x, y|z) = \frac{p(x, y, z)}{p(z)}$$
$$= \frac{p(x)p(y)p(z|x, y)}{p(z)}$$

Dans ce cas, $X \not\perp\!\!\!\perp Y|Z$

Example



Information mutuelle conditionnelle

- L'information mutuelle entre deux variables X et Y conditionnellement à un ensemble de variables Z est:

$$I(X, Y|Z) = \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} dxdydz \qquad (15)$$

- L'information mutuelle conditionnelle mesure l'information partagée par X et Y connaissant Z.
- $I(X, Y|Z)$ est égale à zéro si et seulement si $p(x, y|z) = p(x|z)p(y|z)$, si et seulement si X et Y sont indépendantes conditionnellement à Z.

Cas gaussien

Dans le cas où X et Y sont des variables gaussiennes, on a :

$$I(X, Y|Z) = -\frac{1}{2}\log[1 - \rho^2_{xy|z}] \tag{16}$$

Avec $\rho^2_{xy|z}$ le coefficient de corrélation partielle entre X et Y connaissant Z.
Dans le cas où Z est une variable unique, on a :

$$\rho_{xy|z} = \frac{\rho_{xy} - \rho_{xz}\rho_{zy}}{\sqrt{1 - \rho^2_{xz}}\sqrt{1 - \rho^2_{zy}}} \tag{17}$$

Comment faire un test d'indépendance conditionnelle dans le cas non Gaussien ?

Pour aller plus loin. Voir les articles suivants :

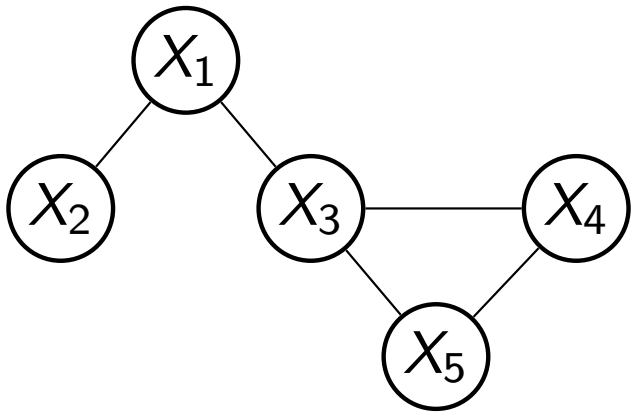
- Test d'indépendance conditionnelle avec des kernels : *Kernel Conditional Independence test (KCI)* (Zhang et al., 2012).
- Test d'indépendance conditionnelle basé sur les plus proches voisins : *Conditional Mutual Information Test (CMIT)* (Runge, 2017)

Section 4

Modèles graphiques non dirigés

Qu'est-ce qu'un modèle graphique non dirigé ?

- Un modèle graphique non dirigé capture des relations de dépendances statistiques entre les variables $X_1, \ldots X_d$ (Lauritzen, 1996).
- Chaque modèle correspond à un graphe $\mathcal{G} = (V, E)$, avec:
  - V, l'ensemble des noeuds du graphe (chaque noeud correspond à une variable de $\mathbf{X}$). Pour des raisons de simplicité, on notera X_i la variable et X_i son noeud associé dans le graphe $\mathcal{G}$.
  - E, l'ensemble des liens non dirigés du graphe. On note $\{X_i, X_j\} \in E$ un lien non dirigé entre X_i et X_j.
- On définit l'ensemble des voisins de X_i, $nb_{\mathcal{G}}(X_i) = \{X_j \in V : \{X_i, X_j\} \in E\}$.

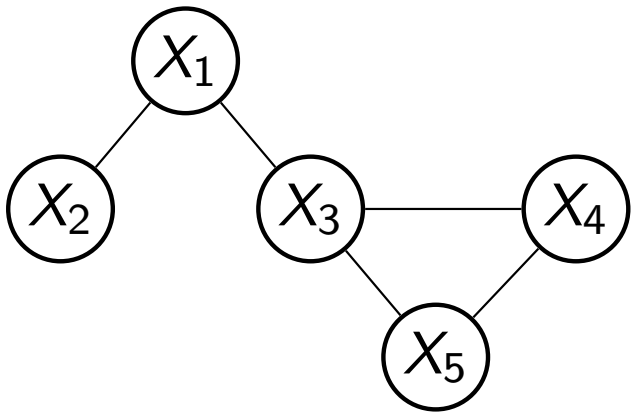


Quel est l'intérêt de retrouver un graphe non dirigé ?

- En biologie, on est parfois simplement intéressé par des relation de co-expressions qu'on peut vérifier avec la *Gene Ontology* (GO), pour savoir si deux gènes participent à la même fonction.
- Un graphe non dirigé est en pratique plus facile à inférer qu'un graphe dirigé.
- Retrouver ce graphe non dirigé peut être un point de départ pour trouver le graphe dirigé (causal). Le squelette du vrai graphe causal est en fait contenu dans ce graphe non dirigé.
- En pratique, on connaît aussi parfois déjà les causes possibles (facteurs de transcription). Il suffit donc de retrouver pour un gène donné quel sont ses meilleurs prédicteurs parmi ces facteurs de transcription (Cas des challenge Dream4 et Dream5). Une méthode de sélection de variables ou d'inférence de graphes non dirigés suffit donc dans ce cas.

Représentation du graphe non dirigé par une matrice d'adjacence symétrique

- Un graphe non dirigé avec d variables peut être représenté par une matrice binaire symétrique A représentant tous les liens entre les variables.
- $A_{ij} = 1$ si et seulement si il y a un lien entre les noeuds X_i et X_j dans le graphe $\mathcal{G}$, $A_{ij} = 0$ sinon.

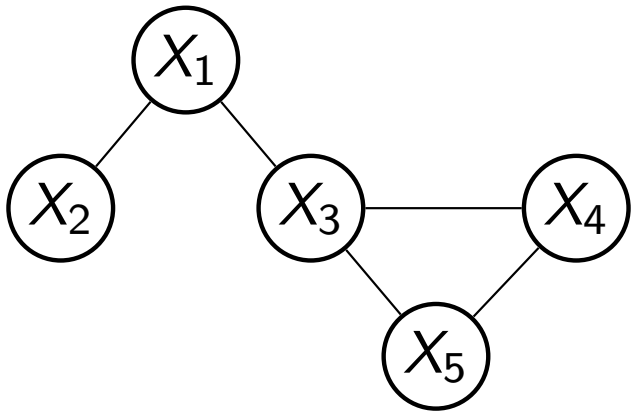


$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \quad (18)$$

Propriétés de Markov

- $\mathbf{X}$ satisfait la *propriété de Markov par paire* par rapport au graphe $\mathcal{G}$ si: $\{X_i, X_j\} \notin E \implies X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j}$.
- $\mathbf{X}$ satisfait la *propriété de Markov locale* par rapport au graphe $\mathcal{G}$ si pour chaque $X_i \in V$, $X_i \perp\!\!\!\perp X_{V \setminus (nb_{\mathcal{G}}(X_i) \cup X_i)} | X_{nb_{\mathcal{G}}(X_i)}$
- $\mathbf{X}$ satisfait la *propriété de Markov globale* par rapport au graphe $\mathcal{G}$ si $X_A \perp\!\!\!\perp X_B | X_C$ pour tout triplé d'ensemble de variables disjoints deux à deux $A, B, C \subset V$ tels que C *sépare* A et B dans $\mathcal{G}$, c'est à dire que tous les chemins entre des noeuds de A et de B passent par au moins un noeud de C.

Exemple d'indépendances conditionnelles induites par un graphe non dirigé



- Si $\mathbf{X}$ satisfait la *propriété de Markov par paire* par rapport au graphe ci-dessus :
  - comme $\{X_1, X_4\} \notin E$, alors $X_1 \perp\!\!\!\perp X_4 | (X_2, X_3, X5)$.
  - comme $\{X_2, X_4\} \notin E$, alors $X_2 \perp\!\!\!\perp X_4 | (X_1, X_3, X5)$.

  .
- *Propriété de Markov locale* pour le noeud X_4: $X_1 \perp\!\!\!\perp (X_4, X_5) | (X_2, X_3)$.
- La *propriété de Markov globale* implique d'autres relations d'indépendances comme $X_5 \perp\!\!\!\perp X_2 | (X_3, X_4)$, etc...

Hypothèse de *faithfulness*

$\mathbf{X}$ est fidèle au graph $\mathcal{G}$ si toutes les relations d'indépendances entre les d variables X_i de $\mathbf{X}$ sont représentées sur le graphe $\mathcal{G}$.
Notamment :

- Si $X_A \perp\!\!\!\perp X_B | X_C$, alors l'ensemble de variables C sépare les ensembles de variables A et B dans le graphe $\mathcal{G}$.
- On a aussi

$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j} \implies \{X_i, X_j\} \notin E,$$

- et en prenant la contraposée de cette implication :

$$\{X_i, X_j\} \in E \implies X_i \not\perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j}$$

Exemple de cas "non fidèle" qui peut arriver en pratique dans des cas réels

- Malheureusement, en pratique, il peut exister des cas particuliers où l'hypothèse de faithfullness n'est pas vérifée.
- Par exemple, si le gène X_i régule le gène X_j directement. Dans le graphe d'interaction, on s'attend à avoir un lien entre X_i et X_j.
- Or, si l'effet de X_i sur X_j est très faible, on risque de mesurer $X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j}$.
- C'est une des raisons qui rend le problème d'inférence de réseau difficile en pratique !

Hypothèses de Markov et *faithfulness* réunies

Si on suppose la *propriété de Markov* ainsi que l'hypothèse de *faithfulness*, on a l'équivalence suivante :

$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j} \iff \{X_i, X_j\} \notin E \tag{19}$$

Inférence d'un graphe non dirigé à partir de données d'observation

- On note $\mathbf{X} = [X_1, \dots X_d]$ un vecteur de d variables continues, avec une distribution de probabilité jointe inconnue $p(\mathbf{x})$. On dispose d'un échantillon iid de n points tirés suivant $p(\mathbf{x})$, noté $D = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$, with $\mathbf{x}^{(\ell)} = (x_1^{(\ell)}, \dots, x_d^{(\ell)})$, avec $x_j^{(l)}$ le ℓ-ième échantillon de X_j.
- On fait l'hypothèse que $\mathbf{X}$ satisfait les propriétés de Markov par rapport à un graphe $\mathcal{G}$.
- But : retrouver $\mathcal{G}$.

Approches "tests d'indépendance conditionnelle"

Approches qui exploitent la *propriété de Markov paire à paire* et de *faithfulness* :

- On part d'un graphe complet et on fait des tests d'indépendance conditionnelle pour vérifier pour chaque paire de variable $\{X_i, X_j\} \in V$ si $X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j}$.
- Si $X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j}$ alors on enlève l'arc $\{X_i, X_j\}$.
- $\binom{d}{2}$ tests d'indépendances à effectuer.

Cas gaussien multivarié

- Pour un vecteur Gaussien $\mathbf{X} = [X_1, ..., X_d]$ de dimension d.
  $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ avec Σ inversible.
- Densité de probabilité de $\mathbf{X}$:

$$p(\mathbf{X}) = \frac{1}{\sqrt{(2\pi)^d det\Sigma}} \exp(-1/2(\mathbf{X} - \mu)^T \Sigma^{-1}(\mathbf{X} - \mu)) \tag{20}$$

- La matrice $K = \Sigma^{-1}$ est appelée la matrice de précision de X.
  $K = (k_{ij})_{1 \leq i,j \leq d}$.

En dimension 3 par exemple

- Si $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$. En dimension 3. $K = (k_{ij})_{1 \leq i,j \leq 3}$

$$p(x_1, x_2, x_3) = \frac{1}{\sqrt{(2\pi)^d det\Sigma}} \exp(-1/2(k_{11}x_1^2 + k_{22}x_2^2 + k_{33}x_3^2 + +2k_{12}x_1x_2 + 2k_{13}x_1x_3 + 2k_{23}x_2x_3))$$

- Comme

$$X1 \perp\!\!\!\perp X2|X3 \iff \exists (g, h) \ , p(x_1, x_2, x_3) = g(x_1, x_3)h(x_2, x_3)$$

alors

$$X1 \perp\!\!\!\perp X2|X3 \iff k_{12} = 0.$$

Cas général pour d variables gaussiennes

$\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$. En dimension d. $K = (k_{ij})_{1 \leq i,j \leq d}$.

On note $\mathbf{X}_{\setminus i,j}$ l'ensemble des variables de $\mathbf{X}$ privé des variables X_i et X_j.

$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j} \iff k_{ij} = 0 \tag{21}$$

Inférence d'un graphe non dirigé dans le cas Gaussien

Dans le cas Gaussien, il existe une méthode globale avec la matrice de précision :

- Pour vérifier si $X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j}$ pour chaque paire de variable, il suffit d'estimer la matrice de précision K de $\mathbf{X}$, avec $K = \Sigma^{-1}$.
- Cette matrice de précision est usuellement retrouvée par maximisation de la log vraisemblance de la matrice des données d'observation.

Maximisation de la log vraisemblance pour retrouver K

- Densité de probabilité d'un vecteur Gaussien $\mathbf{X} = (X_1, ..., X_d)$ de dimension d. $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$:

$$p(\mathbf{X}) = \frac{1}{\sqrt{(2\pi)^d det\Sigma}} \exp(-1/2(\mathbf{X} - \mu)^T \Sigma^{-1}(\mathbf{X} - \mu)) \tag{22}$$

- Etant donné n points iid tirés de cette loi multivariée Gaussienne, la log-vraisemblance à maximiser est :

$$\begin{aligned} L(K) &= \sum_{i=1}^{n} \log\, p(\mathbf{X}^i) \\ &= -\frac{1}{2} log(\det\, \Sigma) - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{X}^i - \mu)^T \Sigma^{-1} (\mathbf{X}^i - \mu) + cst \end{aligned}$$

On a $K = \Sigma^{-1}$ et $\det(K) = \frac{1}{\det(\Sigma)}$. Donc :

$$L(K) = \frac{1}{2} log(\det\ K) - \frac{1}{2} \sum_{i=1}^{n} tr(K(\mathbf{X}^i - \mu)^T(\mathbf{X}^i - \mu)) + cst \tag{23}$$

$$L(K) = \frac{1}{2} log(\det\ K) - \frac{1}{2} tr(K \sum_{i=1}^{n} (\mathbf{X}^i - \mu)^T(\mathbf{X}^i - \mu)) + cst \tag{24}$$

$$L(K) = \frac{1}{2} log(\det\ K) - \frac{1}{2} tr(KS) + cst \tag{25}$$

Avec $S = \sum_{i=1}^{n} (\mathbf{X}^i - \mu)^T(\mathbf{X}^i - \mu)$ l'estimation empirique de la matrice de covariance calculée à partie des données d'observation.

Estimateur *Glasso* pour inférer un graphe non dirigé dans le cas Gaussien

- Banerjee et al. (2008) a proposé la méthode du *graphical lasso* (*glasso*) qui consiste à maximiser $L(K)$ mais en ajoutant une pénalisation ℓ_1 sur les coefficient de K de façon à forcer le plus de coefficient possible de K à tendre vers 0 :

$$\hat{K}^{gl} = \underset{K}{argmin}(-L(K) + \lambda \|K\|_1), \tag{26}$$

avec :

$$\|K\|_1 = \sum_{i,j} |k_{ij}| \tag{27}$$

Inférer un graphe non dirigé dans le cas non Gaussien en exploitant la propriété de Markov paire à paire (1/2)

- En pratique, dans le cas non Gaussien, vérifier si $X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j}$ n'est pas évident.
- On a vu qu'il existait des tests d'indépendance conditionnelle pour calculer $I(X, Y|Z)$, mais quand Z contient beaucoup de variables, le test est difficile à réaliser (malédiction de la dimension).

Inférer un graphe non dirigé dans le cas non Gaussien en exploitant la propriété de Markov paire à paire (2/2)

Des heuristiques utilisées en bioinformatique qui marchent plutôt bien vous seront présentées au cours de la dernière séance de cette UE, comme les méthodes ARACNE, CLR, etc. Ces méthodes procèdent comme suit dans les grandes lignes :

- On calcule pour chaque paire $I(X_i, X_j)$.
- On applique un seuil sur $I(X_i, X_j)$ pour faire un premier filtre. On enlève les paires de variables non dépendantes.
- On enlève ensuite des liens qui correspondent à des liens présumés indirectes. Par exemple dans la méthode ARACNE (Margolin et al., 2006), il est fait l'hypothèse que si $X_i \perp\!\!\!\perp X_j | Z$, avec Z une variable unique, alors on s'attend à ce que : $I(X_i, X_j) \leq min[I(X_i, Z); I(Z, X_j)]$
- Souvent, il est en effet constaté en pratique que les liens indirectes sont de plus faible intensité que les liens directes.

Autres approches pour inférer un graphe non dirigé

Approches qui exploitent les *propriétés de Markov locale* :

- Pour chaque variable $X_i \in V$, on détermine $nb_{\mathcal{G}}(X_i)$, ce qui permet de retrouver le graphe.
- Une méthode de sélection de variable comme la méthode lasso peut permettre de retrouver ces voisins $nb_{\mathcal{G}}(X_i)$ dans le graphe non dirigé.
- On effectue une régression de X_i par rapport à toutes les autres variables $\mathbf{X}_{\setminus i}$ en entrée. Avec la pénalisation ℓ_1, toutes les variables inutiles $X_{V \setminus (nb_{\mathcal{G}}(X_i) \cup X_i)}$ doivent être éliminées.
- d problèmes de régression à effectuer.

Inférer un graphe non dirigé grâce à des méthodes de sélection de variable

- On considère qu'inférer le graphe non dirigé $\mathcal{G}$ correspond à résoudre d problèmes de régression avec une sélection de variables (un problème pour chaque variable/noeud X_i).
- D'après la propriété de Markov local, le "vrai" sous ensemble de variables de $\mathbf{X}_{\setminus i}$ dont dépend X_i est $nb_{\mathcal{G}}(X_i)$. Le problème bien sûr est qu'on ne connaît pas $\mathcal{G}$ donc on ne connaît pas l'ensemble $nb_{\mathcal{G}}(X_i)$. Le but est de le retrouver.

Modèle de sélection de variable général

- Pour retrouver $nb_{\mathcal{G}}(X_i)$, on va construire un modèle de régression de X_i conditionnellement à l'ensemble des autres variables $\mathbf{X}_{\setminus i}$. On note $q(x_i|\mathbf{x}_{\theta_i}, \tau_i)$ ou $q(x_i^\ell|\mathbf{x}_{\setminus \mathbf{i}}^\ell, \theta_i, \tau_i)$ la densité de ce modèle probabiliste, avec deux types de paramètres, θ_i correspondant aux indices des variables sélectionnées dans notre modèle et τ_i les paramètres utilisés pour prédire X_i.
- Notre problème est d'identifier le sous-ensemble minimal de variables dans $\mathbf{X}_{\setminus i}$ tel que l'on maximise la vraisemblance conditionnelle sur les données dont on dispose par rapport à ces paramètres θ_i et τ_i.

- Pour notre échantillon de données supposées iid, la log vraisemblance conditionnelle (normalisée) est égale à :

$$L = L(\theta_i, \tau_i | \mathcal{D}) = \frac{1}{n} \sum_{\ell=1}^{n} \log\, q(x_i^\ell | \mathbf{x}_{\setminus \mathbf{i}}^\ell, \theta_i, \tau_i) \tag{28}$$

- Si on multiplie et divise q par $p(x_i | x_{\theta_i})$, on obtient :

$$L = \frac{1}{n} \sum_{\ell=1}^{n} \log\, \frac{q(x_i^\ell | \mathbf{x}_{\setminus \mathbf{i}}^\ell, \theta_i, \tau_i)}{p(x_i^\ell | \mathbf{x}_{\theta_\mathbf{i}}^\ell)} + \frac{1}{n} \sum_{\ell=1}^{n} \log\, p(x_i^\ell | \mathbf{x}_{\theta_\mathbf{i}}^\ell) \tag{29}$$

- De la même manière si on multiplie et divise le deuxième terme de (29) par $p(x_i | \mathbf{x}_{\setminus \mathbf{i}})$. On obtient :

$$L = \frac{1}{n} \sum_{\ell=1}^{n} \log\, \frac{q(x_i^\ell | \mathbf{x}_{\setminus \mathbf{i}}^\ell, \theta_i, \tau_i)}{p(x_i^\ell | \mathbf{x}_{\theta_\mathbf{i}}^\ell)} + \frac{1}{n} \sum_{\ell=1}^{n} \log\, \frac{p(x_i^\ell | \mathbf{x}_{\theta_\mathbf{i}}^\ell)}{p(x_i^\ell | \mathbf{x}_{\setminus \mathbf{i}}^\ell)} + \frac{1}{n} \sum_{\ell=1}^{n} \log\, p(x_i^\ell | \mathbf{x}_{\setminus \mathbf{i}}^\ell) \tag{30}$$

Interprétation de ces trois termes - d'après (Brown et al., 2012)

- Asymptotiquement, on a :

$$\underset{n\to+\infty}{L} = \mathbb{E}_{\mathbf{x}}\left[\log \frac{q(x_i|x_{\setminus i},\theta_i,\tau_i)}{p(x_i|x_{\theta_i})}\right] + \mathbb{E}_{\mathbf{x}}\left[\log \frac{p(x_i|x_{\theta_i})}{p(x_i|x_{\setminus i})}\right] + \mathbb{E}_{\mathbf{x}}\left[\log p(x_i|x_{\setminus i})\right] \quad (31)$$

- On peut cherche à miniser $-L$ au lieu de maximiser L. On a :

$$\underset{n\to+\infty}{-L} = \mathbb{E}_{\mathbf{x}}\left[\log \frac{p(x_i|\mathbf{x}_{\theta_\mathbf{i}})}{q(x_i|\mathbf{x}_{\setminus\mathbf{i}},\theta_i,\tau_i)}\right] + \mathbb{E}_{\mathbf{x}}\left[\log \frac{p(x_i|\mathbf{x}_{\setminus\mathbf{i}})}{p(x_i|\mathbf{x}_{\theta_\mathbf{i}})}\right] - \mathbb{E}_{\mathbf{x}}\left[\log p(x_i|\mathbf{x}_{\setminus\mathbf{i}})\right] \quad (32)$$

- Le terme $-\mathbb{E}_{\mathbf{x}}\left[\log p(x_i|\mathbf{x}_{\setminus\mathbf{i}})\right]$ correspond à $H(X_i|\mathbf{X}_{\setminus\mathbf{i}})$, l'entropie de X_i conditionnellement à $\mathbf{X}_{\setminus\mathbf{i}}$.

- On note $\bar{\theta}_i$ les indices des variables candidates parmi $\mathbf{X}_{\setminus \mathbf{i}}$ non sélectionnées par notre modèle. On a donc $\mathbf{X}_{\setminus \mathbf{i}} = \{\mathbf{X}_{\theta_i}, \mathbf{X}_{\bar{\theta}_i}\}$. Le second terme de l'équation 32 devient donc :

$$
\begin{aligned}
\mathbb{E}_{\mathbf{x}}\left[\log \frac{p(x_i|\mathbf{x}_{\setminus \mathbf{i}})}{p(x_i|\mathbf{x}_{\theta_\mathbf{i}})}\right] &= \int_{\mathbb{R}} p(\mathbf{x}) \log \frac{p(x_i|\mathbf{x}_{\theta_\mathbf{i}}\mathbf{x}_{\bar{\theta}_i})}{p(x_i|\mathbf{x}_{\theta_\mathbf{i}})} d\mathbf{x} \\
&= \int_{\mathbb{R}} p(\mathbf{x}) \log \frac{p(x_i|\mathbf{x}_{\theta_\mathbf{i}}\mathbf{x}_{\bar{\theta}_i}) p(\mathbf{x}_{\bar{\theta}_i}|\mathbf{x}_{\theta_i})}{p(x_i|\mathbf{x}_{\theta_\mathbf{i}}) p(\mathbf{x}_{\bar{\theta}_i}|\mathbf{x}_{\theta_i})} d\mathbf{x} \\
&= \int_{\mathbb{R}} p(\mathbf{x}) \log \frac{p(\mathbf{x}_{\bar{\theta}_i} x_i|\mathbf{x}_{\theta_\mathbf{i}})}{p(x_i|\mathbf{x}_{\theta_\mathbf{i}}) p(\mathbf{x}_{\bar{\theta}_i}|\mathbf{x}_{\theta_i})} d\mathbf{x} \\
&= I(X_i, \mathbf{X}_{\bar{\theta}_i}|\mathbf{X}_{\theta_i})
\end{aligned}
$$

- Pour résumer, on a donc :

$$\underset{n\to+\infty}{-L} = \mathbb{E}_{\mathbf{x}}\left[\log \frac{p(x_i|\mathbf{x}_{\theta_i})}{q(x_i|\mathbf{x}_{\setminus \mathbf{i}}, \theta_i, \tau_i)}\right] + I(X_i, \mathbf{X}_{\bar{\theta}_i}|\mathbf{X}_{\theta_i}) + H(X_i|\mathbf{X}_{\setminus \mathbf{i}}) \tag{33}$$

- On peut faire les interprétations suivantes sur ces trois termes :
  - Le premier terme correspond à un ratio de vraisemblance entre la vrai distribution et la distribution donnée par le modèle étant donné les variables sélectionnées. La taille de ce terme dépend de la qualité d'approximation de notre modèle q pour approximer p compte tenu des variables θ_i qu'on a sélectionnées pour prédire la variable X_i. Souvent, si on choisit un modèle q adapté et/ou suffisamment "expressif" pour reconstruire p, on pourra trouver un jeu de paramètres τ_i tel que $q(x_i|\mathbf{x}_{\setminus \mathbf{i}}, \theta_i, \tau_i) \approx p(x_i|\mathbf{x}_{\theta_i})$, c'est à dire tel que $\mathbb{E}_{\mathbf{x}}\left[\log \frac{p(x_i|\mathbf{x}_{\theta_i})}{q(x_i|\mathbf{x}_{\setminus \mathbf{i}}, \theta_i, \tau_i)}\right] \approx 0$.
  - Le troisième terme $H(X_i|\mathbf{X}_{\setminus \mathbf{i}})$ est une constante du problème qui ne dépend ni de θ_i, ni de τ_i, on peut donc le négliger.
  - Il reste donc le second terme $I(X_i, \mathbf{X}_{\bar{\theta}_i}|\mathbf{X}_{\theta_i})$ qui est capital pour notre problème. Il correspond à l'information mutuelle entre la variable cible X_i et les variables $\mathbf{X}_{\bar{\theta}_i}$ que l'on n'a pas sélectionnées conditionnellement aux variables $\mathbf{X}_{\theta_i}$ que l'on a sélectionnées.

- Si pour tout θ_i, on suppose qu'on peut trouver un jeu de paramètres τ_i tel que $\mathbb{E}_{\mathbf{x}}\left[\log \frac{p(x_i|\mathbf{x}_{\theta_i})}{q(x_i|\mathbf{x}_{\setminus i},\theta_i,\tau_i)}\right] = 0$, quand $n \to +\infty$, on a donc :

$$\underset{\theta_i}{\text{argmax}}\ L(\theta_i|\mathcal{D}) = \underset{\theta_i}{\text{argmin}}\ I(X_i, \mathbf{X}_{\bar{\theta}_i}|\mathbf{X}_{\theta_i}) \tag{34}$$

- Or ce minimum est atteint si on sélectionne l'ensemble de variables $\mathbf{X}_{\theta_i} = nb_{\mathcal{G}}(X_i)$, car d'après la propriété de Markov locale, $I(X_i, X_{V\setminus(nb_{\mathcal{G}}(X_i)\cup X_i)}|X_{nb_{\mathcal{G}}(X_i)}) = 0$, car $X_i \perp\!\!\!\perp X_{V\setminus(nb_{\mathcal{G}}(X_i)\cup X_i)}|X_{nb_{\mathcal{G}}(X_i)}$ et de plus par définition de l'information mutuelle $I(X_i, \mathbf{X}_{\bar{\theta}_i}|\mathbf{X}_{\theta_i}) \geq 0$.
- Attention ce n'est pas suffisant pour retrouver $nb_{\mathcal{G}}(X_i)$. D'autres sous-ensemble de variables $\mathbf{X}_{\theta_i}$ peuvent correspondre à $I(X_i, \mathbf{X}_{\bar{\theta}_i}|\mathbf{X}_{\theta_i}) = 0$.

Besoin d'ajouter une pénalisation sur la complexité du modèle

1er problème : éviter de sélectionner des variables superflues.

- Si on sélectionne un sous-ensemble de variables plus grand que $nb_{\mathcal{G}}(X_i)$, c'est à dire tel que $nb_{\mathcal{G}}(X_i) \subset \mathbf{X}_{\theta_i}$, alors on a quand même $I(X_i, \mathbf{X}_{\bar{\theta}_i} | \mathbf{X}_{\theta_i}) = 0$.
- C'est pour cela notamment qu'on ajoute en générale une pénalisation sur la complexité du modèle au score de vraisemblance (par exemple avec une contrainte sur le nombre de paramètres du modèle dans le score BIC), de façon à ne pas sélectionner de variables inutiles et ainsi espérer retrouver le sous-ensemble de variable minimal $\mathbf{X}_{\theta_i} = nb_{\mathcal{G}}(X_i)$ tel que $I(X_i, \mathbf{X}_{\bar{\theta}_i} | \mathbf{X}_{\theta_i}) = 0$.

Modèle de régression avec une régularisation sur les paramètres.

- Un grand nombre de méthodes utilisées en bioinformatique qui marchent bien en pratique comme GENIE3 ou TIGRESS que vous verrez lors de la dernière séance font une régression de X_i par rapport à toutes les autres variables disponibles avec une méthode de régularisation de façon à retrouver ces meilleures prédicteurs $nb_{\mathcal{G}}(X_i)$ de X_i.
- Attention cependant, même si ces méthodes se présentent comme tel parfois, ce n'est pas de la causalité. Parmi les voisins de X_i certains n'ont même pas de relation causale directe avec X_i !

Importance de l'hypothèse de fidélité (*faithfullness*)

2ème problème : s'assurer que toutes les variables importantes soient sélectionnées.

- L'hypothèse de faithfullness : $\{X_i, X_j\} \in E \implies X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j}$
- On a besoin de cette hypothèse pour dire que si omet par exemple de sélectionner une variable X_k de $nb_{\mathcal{G}}(X_i)$ lors de la sélection de variable alors comme $X_k \in \mathbf{X}_{\bar{\theta}_i}$ on a :

$$I(X_i, \mathbf{X}_{\bar{\theta}_i} | \mathbf{X}_{\theta_i}) \geq I(X_i, X_k | \mathbf{X}_{\theta_i}) \tag{35}$$
$$\geq I(X_i, X_k | \mathbf{X}_{\setminus_i}) \tag{36}$$
$$\tag{37}$$

- Et comme $I(X_i, X_k | \mathbf{X}_{\setminus_i}) > 0$ d'après l'hypothèse de faithfullness alors $I(X_i, \mathbf{X}_{\bar{\theta}_i} | \mathbf{X}_{\theta_i})$ n'est pas le minimum.

Couverture de Markov

- Sous les hypothèse de faithfullness et Markov, le sous-ensemble minimal de variable tel que $I(X_i, \mathbf{X}_{\bar{\theta}_i} | \mathbf{X}_{\theta_i}) = 0$ correspond effectivement à $nb_{\mathcal{G}}(X_i)$.
- Ce sous ensemble correspond à la couverture de Markov (*Markov blanket*) pour les graphes dirigés qu'on verra au cours suivant !

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516.

Brown, G., Pocock, A., Zhao, M.-J., and Luján, M. (2012). Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of machine learning research*, 13(Jan):27–66.

Darbellay, G. A. and Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321.

Fraser, A. and Swinney, H. (1986). Using mutual information to find independent coordinates for strange attractors. *Phys. Rev. A*, 33:1134–1140.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, page S7. BioMed Central.

Runge, J. (2017). Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. *arXiv preprint arXiv:1709.01447*.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.