Inférence de réseaux

Introduction à la causalité et graphes non dirigés

Olivier Goudet

Université d'Angers

15 janvier 2026

Plan

Section 1

Notions d'indépendance entre des variables aléatoires

Subsection 1

Indépendance entre deux variables

Rappel : indépendance entre deux variables

Si (X, Y) est une paire de variable aléatoire de densité de probabilité conjointe $p(x, y)$, 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 définies sur $\mathbb{R}$ est

$$I(X,Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} d_x d_y. \quad (1)$$

- 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 $p(x)p(y)$:

$$I(X,Y) = D_{KL}(p(x,y) \,\|\, p(x)p(y)) \quad (2)$$

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 :

$$H(X) = -\int_{\mathbb{R}} p(x) \log p(x) dx \tag{3}$$

Calcul de l'entropie dans le cas gaussien

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

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

$$= -\int_{\mathbb{R}} p(x) \log \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{x^2}{2\sigma^2}) dx \tag{5}$$

$$= -\int_{\mathbb{R}} p(x) \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{x^2}{2\sigma^2} \right] dx \tag{6}$$

$$\tag{7}$$

- 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 $X = (X_1, ..., X_d)$ de dimension d. $X \sim \mathcal{N}(0, \Sigma)$:

$$p(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2} x^T \Sigma^{-1} x} \tag{8}$$

- La formule de l'entropie dans ce cas est (cf. pdf annexe preuves) :

$$h(X) = \frac{1}{2} \log((2\pi e)^d |\Sigma|) \tag{9}$$

- Dans le cas bivarié, on a donc

$$h(X, Y) = \frac{1}{2} \log((2\pi e)^2 (\sigma_x^2 \sigma_y^2 - \sigma_{xy}^2)) \tag{10}$$

- 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:

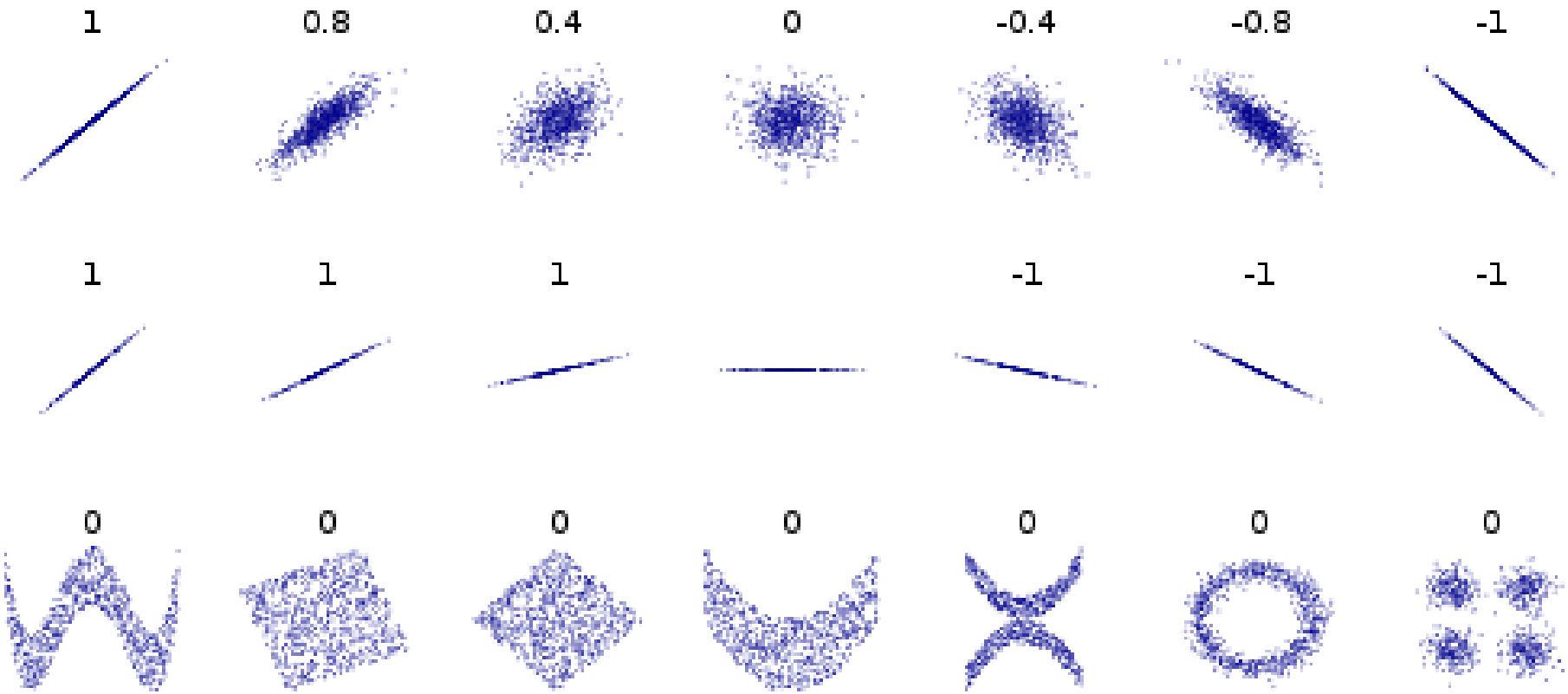
$$I(X, Y) \approx I_{binned}(X, Y) = \sum_{i,j} f(i,j) \log \frac{f(i,j)}{f_x(i) f_y(j)} \tag{11}$$

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)

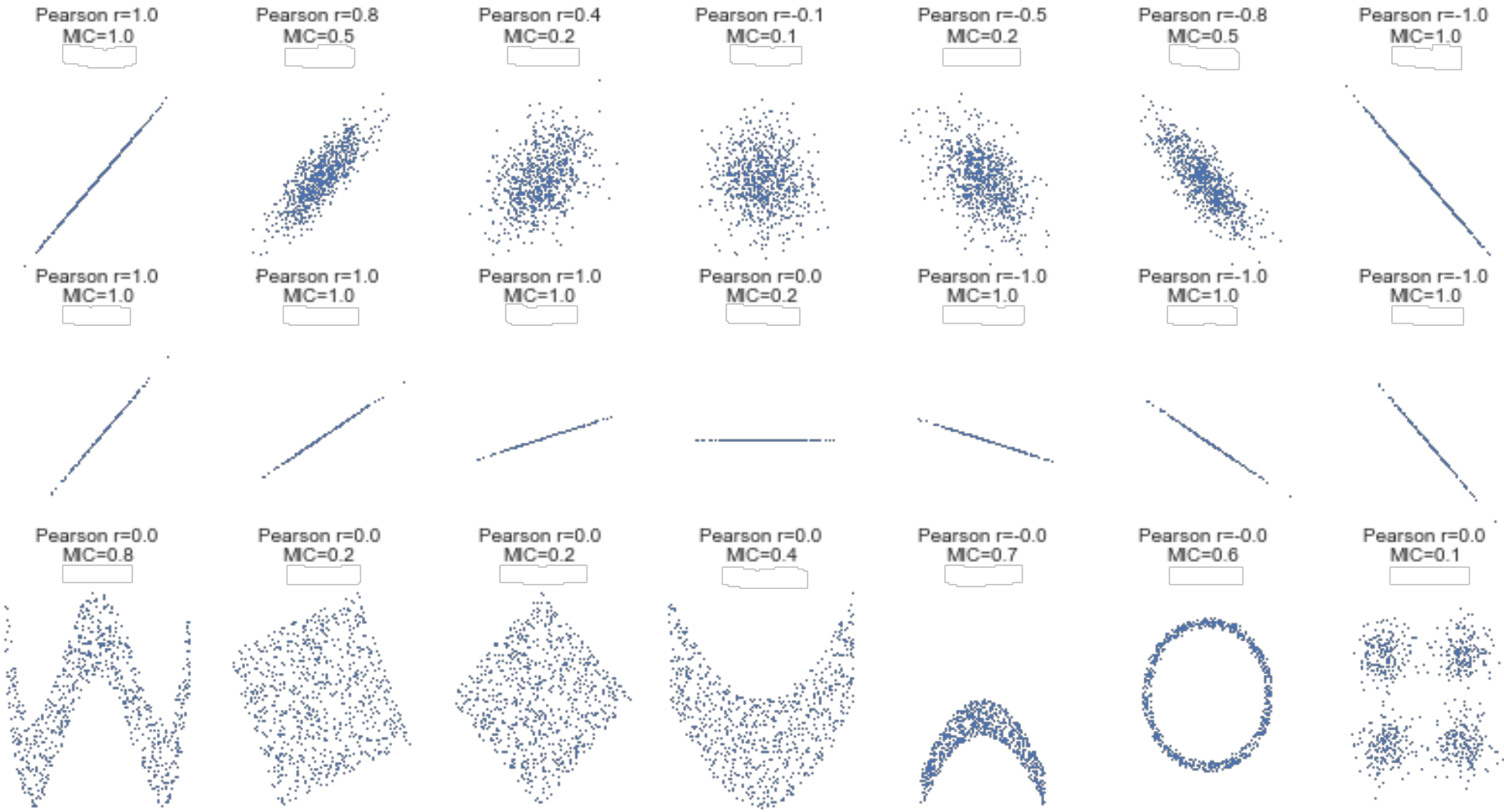
- Une estimation de $I_{binned}(X, Y)$ est ensuite obtenue en comptant le nombre de points dans chaque boîte.
- Si $n_x(i)$ (resp. $n_y(j)$ est le nombre de points dans l'intervalle i de X (resp. j de Y) et $n(i, j)$ le nombre de points dans l'intersection.
- On peut approximer $f_x(i) \approx n_x(i)/N$, $f_y(j) = n_y(j)/N$ 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

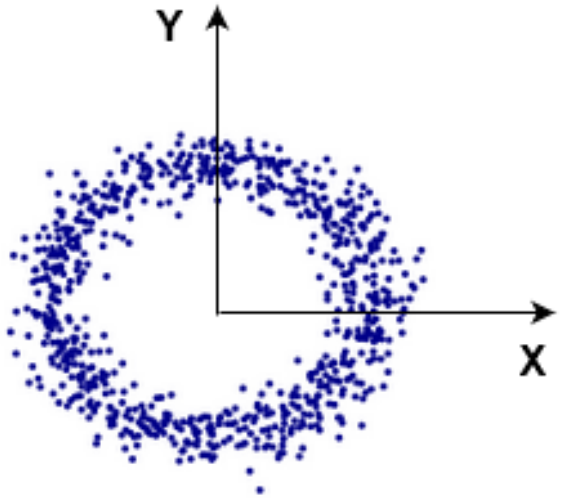


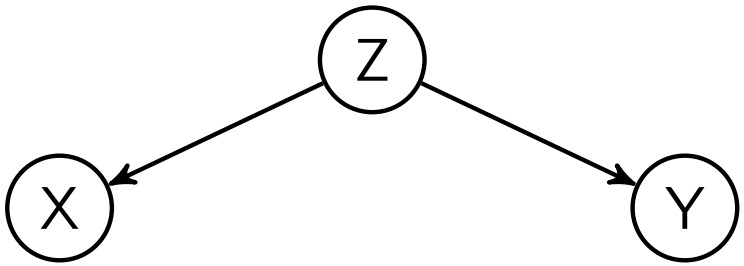
Source : wikipedia

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 :

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

Subsection 2

Notions d'indépendance conditionnelle

Indépendance conditionnelle

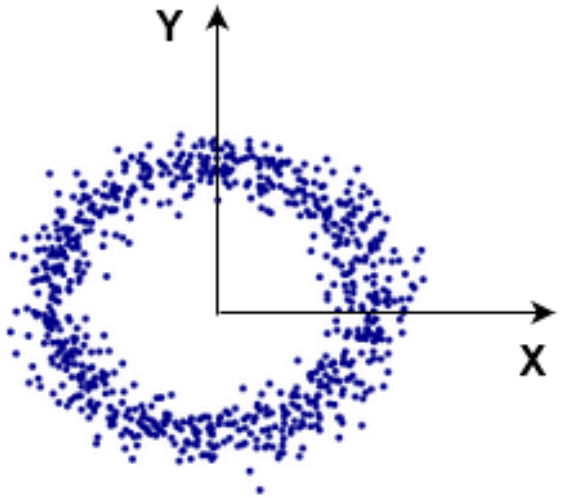
X et Y sont indépendantes conditionnellement à un ensemble de variables Z : notation $X \perp\!\!\!\perp Y|Z$.

$$X \perp\!\!\!\perp Y|Z \iff p(x, y|z) = p(x|z)p(y|z) \tag{12}$$

Autre caractérisation utile par la suite :

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

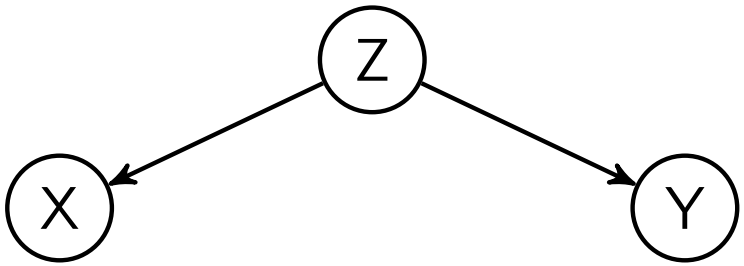
Si on reprend le cas du rond



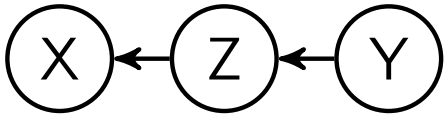
$$p(x,y|z) = \frac{p(x,y,z)}{p(z)}$$
$$= \frac{p(x|z)p(y|z)p(z)}{p(z)}$$
$$= p(x|z)p(y|z)$$

Donc :

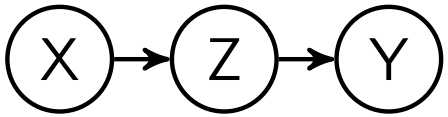
$$X \perp\!\!\!\perp Y|Z$$



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

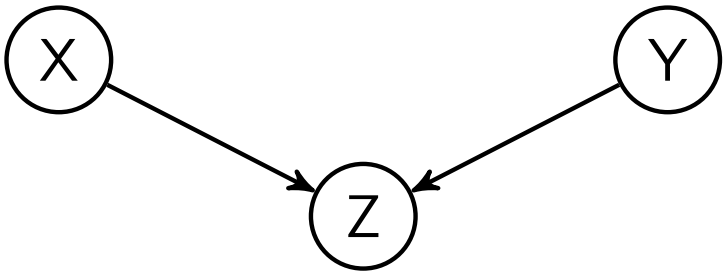


$$
\begin{aligned}
p(x,y|z) &= \frac{p(x,y,z)}{p(z)} \\
&= \frac{p(x|z)p(z|y)p(y)}{p(z)} \\
&= \frac{p(x|z)p(y|z)p(z)}{p(z)} \\
&= p(x|z)p(y|z)
\end{aligned}
$$



Idem dans l'autre sens en permutant X et Y

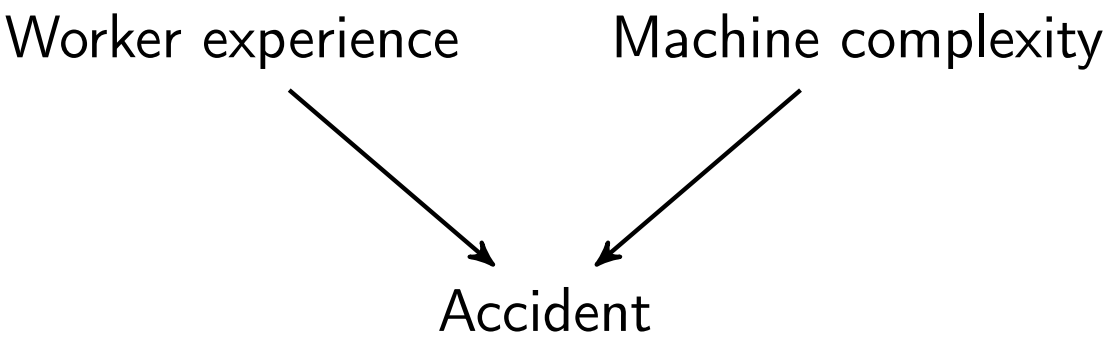
Cas de la v-structure $X \perp\!\!\!\perp Y$ e $X \not\perp\!\!\!\perp Y|Z$



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

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:

$$I(X, Y|Z) = \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} dxdydz \qquad (14)$$

- 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, Y et Z sont trois variables gaussiennes, on a :

$$I(X, Y|Z) = -\frac{1}{2}\log[1 - \rho^2_{xy|z}] \tag{15}$$

Avec $\rho^2_{xy|z}$ le coefficient de corrélation partielle entre X et Y connaissant Z :

$$\rho_{xy|z} = \frac{\rho_{xy} - \rho_{xz}\rho_{zy}}{\sqrt{1 - \rho^2_{xz}}\sqrt{1 - \rho^2_{zy}}} \tag{16}$$

Voir détail des calculs dans le pdf annexe du cours.

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 2

Modèles graphiques non dirigés

Subsection 1

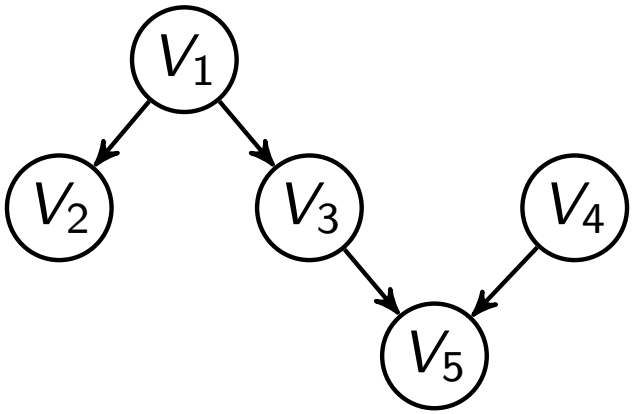
Notations et définitions

Notations

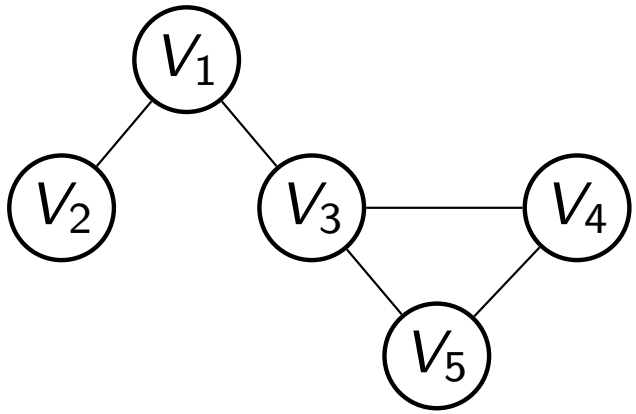
- Dans cette partie, on considère un ensemble de d variables aléatoires, notées $X_1, \dots X_d$.
- Le vecteur qui regroupe ces variables aléatoires est noté $X = (X_1, \dots X_d)$.
- La densité de X (loi conjointe) est notée $p(x)$.
- La densité de chaque variable X_i est noté $p(x_i)$ (loi marginale).
- On note $X_{\setminus i,j}$, le vecteur aléatoire constitué de toutes les variables de X sauf les variables X_i et X_j.

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

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



(a) Le vrai graphe causal $\mathcal{G}$.



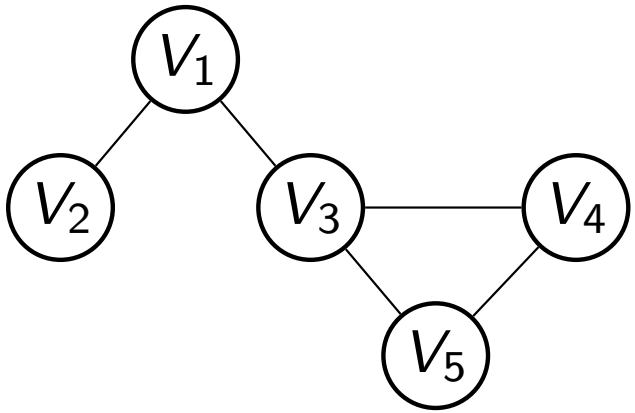
(b) Le graphe moral $\mathcal{G}^m$ de $\mathcal{G}$.

Relations de voisinage dans le graphe

- On définit l'ensemble des voisins de V_i dans le graphe $\mathcal{G}^m$ par $nb_{\mathcal{G}^m}(V_i) = \{V_j \in V : \{V_i, V_j\} \in E\}$.
- On note $\overline{nb_{\mathcal{G}^m}}(V_i) = nb_{\mathcal{G}^m}(V_i) \cup V_i$.
- Pour un ensemble de sommets E inclu dans V, on notera parfois X_E le vecteur aléatoire qui regroupe l'ensemble des variables correspondant aux sommets de E.
- Par exemple $X_{nb_{\mathcal{G}^m}(V_i)}$ correspond au vecteur aléatoire regroupant toutes les variables aléatoires associées aux voisins du sommet V_i dans le graphe $\mathcal{G}^m$.

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 V_i et V_j dans le graphe $\mathcal{G}^m$, $A_{ij} = 0$ sinon.



$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \quad (17)$$

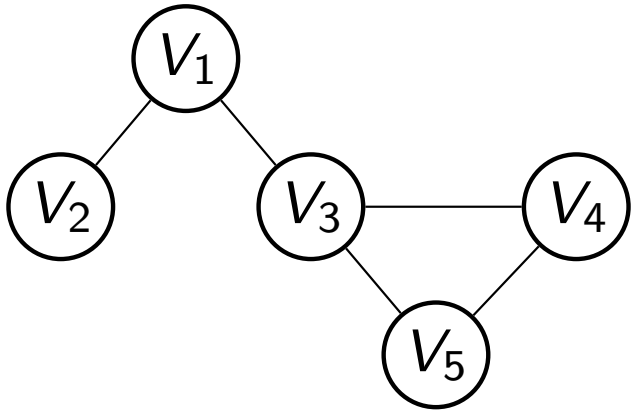
Subsection 2

Hypothèses et propriétés

Propriétés de Markov

- X satisfait la *propriété de Markov par paire* (P) par rapport au graphe $\mathcal{G}^m$ si: $\{V_i, V_j\} \notin E \implies X_i \perp\!\!\!\perp X_j | X_{\setminus i,j}$.
- X satisfait la *propriété de Markov locale* (L) par rapport au graphe $\mathcal{G}^m$ si pour chaque $V_i \in V$, $X_i \perp\!\!\!\perp X_{V \setminus \overline{nb_{\mathcal{G}^m}}(V_i)} | X_{nb_{\mathcal{G}^m}(V_i)}$
- X satisfait la *propriété de Markov globale* (G) par rapport au graphe $\mathcal{G}^m$ 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 $\mathcal{G}^m$, c'est à dire que tous les chemins entre des noeuds de A et de B passent par au moins un noeud de C.
- **Si $p(x) > 0$, ces trois propriétés de Markov, (P), (L) et (G), sont équivalentes [Lauritzen 1996]**
  Voir démonstration dans le pdf annexe du cours.

Exemple d'indépendances conditionnelles induites par un graphe non dirigé



- Si X satisfait la *propriété de Markov par paire* (P) par rapport au graphe ci-dessus :
  - comme $\{V_1, V_4\} \notin E$, alors $X_1 \perp\!\!\!\perp X_4 | (X_2, X_3, X5)$.
  - comme $\{V_2, V_4\} \notin E$, alors $X_2 \perp\!\!\!\perp X_4 | (X_1, X_3, X5)$.
- *Propriété de Markov locale* (L) pour le noeud V_4: $X_1 \perp\!\!\!\perp (X_4, X_5) | (X_2, X_3)$.
- La *propriété de Markov globale* (G) implique d'autres relations d'indépendances comme $X_5 \perp\!\!\!\perp X_2 | (X_3, X_4)$, etc...

Hypothèse de *faithfulness*

X est fidèle au graph $\mathcal{G}^m$ si toutes les relations d'indépendances entre les d variables X_i de X sont représentées sur le graphe $\mathcal{G}^m$.
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 $\mathcal{G}^m$.
- On a aussi

$$X_i \perp\!\!\!\perp X_j | X_{\setminus i,j} \implies \{V_i, V_j\} \notin E,$$

- et en prenant la contraposée de cette implication :

$$\{V_i, V_j\} \in E \implies X_i \not\perp\!\!\!\perp X_j | X_{\setminus i,j}$$

Exemple de cas "non fidèle" qui peut arriver en pratique avec des données réelles

- Malheureusement, en pratique, il peut exister des cas particuliers où l'hypothèse de *faithfullness* n'est pas vérifée.
- Exemple : le gène G_i dont le niveau d'expression est modélisé par la variable X_i **régule directement** le gène G_j dont le niveau d'expression est modélisé par la variable X_j . Dans le graphe d'interaction, on s'attend à avoir un lien entre V_i et V_j.
- Or, si l'effet de G_i sur G_j est très faible, on risque de mesurer $X_i \perp\!\!\!\perp X_j | X_{\setminus i,j}$.
- C'est une des raisons qui rend le problème d'inférence de réseaux difficile en pratique !

Hypothèses de Markov et *faithfulness* réunies

Si on suppose les *propriétés de Markov* ainsi que les hypothèses de *faithfulness*, on a notamment l'équivalence suivante de X par rapport au graphe $\mathcal{G}^m = (V, E)$:

$$X_i \perp\!\!\!\perp X_j | X_{\setminus i,j} \iff \{V_i, V_j\} \notin E \tag{18}$$

Section 3

Méthodes d'inférence d'un graphe non dirigé

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 peut suffire dans ce cas.

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

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

Méthodes d'inférence d'un graphe non dirigé

On distingue deux grandes familles de méthodes pour l'inférence d'un graphe non dirigé :

1. Méthodes qui exploitent la propriété de Markov par paire (P).
   - Elles consistent à effectuer des tests d'indépendance conditionnelle entre les variables (cf. section 3).
   - Dans le cas gaussien, il y a des techniques particulières.
2. Méthodes qui exploitent la propriété de Markov locale (L).
   - Elles consistent à retrouver les voisins de chaque variables dans le graphe.
   - Retrouver les voisins est souvent réalisé à l'aide de modèles de régression pénalisée (e.g. lasso) de chaque variable par rapport à toutes les autres.

Subsection 1

Méthodes qui exploitent la propriété de Markov par paire (P)

Inférer un graphe non dirigé en exploitant la propriété de Markov paire à paire

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 sommets $\{V_i, V_j\} \in V$ si $X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j}$ ou bien $X_i \not\perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j}$.
- Si on observe $X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j}$ alors on enlève l'arc $\{V_i, V_j\}$.
- $\binom{d}{2}$ tests d'indépendances à effectuer.

Tests d'indépendance conditionnelle

- En pratique, pour vérifier si $X_i \perp\!\!\!\perp X_j | X_{\setminus i,j}$, il faut vérifier si $I(X_i, X_j | X_{\setminus i,j}) = 0$ ou bien si $I(X_i, X_j | X_{\setminus i,j}) < \alpha$ avec α un seuil choisi expérimentalement.
- $I(X_i, X_j | X_{\setminus i,j})$ donne un score de confiance sur la présence ou non de l'arc $\{V_i, V_j\}$ dans le graphe.
- On a vu dans la section 2 qu'il existait différentes techniques pour estimer $I(X, Y|Z)$, mais quand Z contient beaucoup de variables, le test est difficile à réaliser (malédiction de la dimension).

Heuristiques utilisées en bioinformatique

Des heuristiques utilisées en bioinformatique qui marchent plutôt bien vous seront présentées plus en détail 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.

Cas gaussien multivarié

Dans le cas gaussien multivarié, il existe des techniques spéciales efficaces qui exploitent des propriétés particulières de la matrice de covariance.

- Soit $X = (X_1, ..., X_d)$ un vecteur gaussien de dimension d.
  $X \sim \mathcal{N}(\mu, \Sigma)$, avec μ le vecteur des moyennes et Σ la matrice de covariance.
- Densité de probabilité de X :

$$p(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-1/2(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{19}$$

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

Exemple : en dimension 3

- Si $X \sim \mathcal{N}(0, \Sigma)$ en dimension 3, on a

$$p(x_1, x_2, x_3) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \times e^{-1/2(k_{11}x_1^2 + k_{22}x_2^2 + k_{33}x_3^2 + 2k_{12}x_1x_2 + 2k_{13}x_1x_3 + 2k_{23}x_2x_3)} \tag{20}$$

- Comme on sait que

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

alors on a

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

Cas général pour d variables gaussiennes

- $X \sim \mathcal{N}(\mu, \Sigma)$. En dimension d. $K = (k_{ij})_{1 \leq i,j \leq d}$.
- On a l'équivalence suivante :

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

Voir preuve dans le pdf annexe du cours.

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 $X_i \perp\!\!\!\perp X_j | X_{\setminus i,j}$ pour chaque paire de variable, il suffit d'estimer la matrice de précision K de X, avec $K = \Sigma^{-1}$.
- D'après le slide précédent, les coefficients non nuls de K correspondent directement aux arcs du graphe.
- 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 $X = (X_1, ..., X_d)$ de dimension d. $X \sim \mathcal{N}(\mu, \Sigma)$:

$$p(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-1/2(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{22}$$

- Etant donné n points x^i iid tirés de cette loi multivariée gaussienne, on va chercher à estimer les paramètres de la matrice $K = \Sigma^{-1}$ qui maximisent la log-vraisemblance :

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

car $|K| = \frac{1}{|\Sigma|}$.

Comme $(x^i - \mu)^T K (x^i - \mu)$ est un scalaire, on a $(x^i - \mu)^T K (x^i - \mu) = \mathrm{Tr}((x^i - \mu) K (x^i - \mu)^T)$ et en utilisant la relation Tr(ABC) = Tr(CAB), on en déduit

$$L(K) = \frac{1}{2} log(|K|) - \frac{1}{2} \sum_{i=1}^{n} \mathrm{Tr}(K (x^i - \mu)(x^i - \mu)^T) + cst \tag{23}$$

$$L(K) = \frac{1}{2} log(|K|) - \frac{1}{2} \mathrm{Tr}(K \sum_{i=1}^{n} (x^i - \mu)(x^i - \mu)^T) + cst \tag{24}$$

$$L(K) = \frac{1}{2} log(|K|) - \frac{1}{2} \mathrm{Tr}(KS) + cst, \tag{25}$$

avec $S = \sum_{i=1}^{n} (x^i - \mu)(x^i - \mu)^T$ 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 :

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

Subsection 2

Méthodes qui exploitent la propriété de Markov locale (L)

Méthodes qui exploitent la propriété de Markov locale (L)

- Pour chaque noeud $V_i \in V$, on cherche à déterminer $nb_{\mathcal{G}^m}(V_i)$, ce qui permet de retrouver le graphe $\mathcal{G}^m$.
- C'est un ensemble de voisins tel que $X_i \perp\!\!\!\perp X_{V \setminus \overline{nb_{\mathcal{G}^m}}(V_i))} | X_{nb_{\mathcal{G}^m}(V_i)}$ (cf section précédente).

Utiliser des modèles de régression pénalisée

- Une méthode de sélection de variable comme la méthode lasso peut permettre de retrouver ces voisins $nb_{\mathcal{G}^m}(V_i)$ dans le graphe non dirigé.
- On effectue une régression pénalisée de X_i par rapport à toutes les autres variables $X_{\setminus i}$ en entrée.
- Avec la pénalisation ℓ_1, toutes les variables inutiles correspondant aux sommets dans $V \setminus \overline{nb_{\mathcal{G}^m}(V_i)}$ doivent être éliminées (cf. cours "régression linéaire en Grande Dimension" de Fabien Panloup)
- d problèmes de régression à effectuer.
- Note : ces d problèmes sont indépendants et peuvent donc être traités en parallèle.

Modèle de sélection de variable général

- Pour retrouver $nb_{\mathcal{G}^m}(V_i)$, on va construire un modèle de régression de X_i conditionnellement à l'ensemble des autres variables $\mathbf{X}_{\setminus i}$. On note $q(x_i|\mathrm{x}_{\theta_i}, \tau_i)$ ou $q(x_i^\ell|\mathrm{x}_{\setminus i}^\ell, \theta_i, \tau_i)$ 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.

Décomposition de la log-vraisemblance

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

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

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

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

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

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

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

- Asymptotiquement, on a :

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

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

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

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

Terme d'information mutuelle conditionnelle

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

$$
\begin{aligned}
\mathbb{E}_x\left[\log \frac{p(x_i|x_{\setminus i})}{p(x_i|x_{\theta_i})}\right] &= \int_{\mathbb{R}} p(x)\, \log \frac{p(x_i|x_{\theta_i} x_{\bar{\theta}_i})}{p(x_i|x_{\theta_i})} dx \\
&= \int_{\mathbb{R}} p(x)\, \log \frac{p(x_i|x_{\theta_i} x_{\bar{\theta}_i}) p(x_{\bar{\theta}_i}|x_{\theta_i})}{p(x_i|x_{\theta_i}) p(x_{\bar{\theta}_i}|x_{\theta_i})} dx \\
&= \int_{\mathbb{R}} p(x)\, \log \frac{p(x_{\bar{\theta}_i} x_i|x_{\theta_i})}{p(x_i|x_{\theta_i}) p(x_{\bar{\theta}_i}|x_{\theta_i})} dx \\
&= I(X_i, X_{\bar{\theta}_i}|X_{\theta_i})
\end{aligned}
$$

Pour résumer, on a donc :

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

Partie constante du score de vraisemblance

Le troisième terme $H(X_i|X_{\setminus i})$ est une constante du problème qui ne dépend ni de θ_i, ni de τ_i, on peut donc le négliger.

Partie "fonctionnelle" du score de vraisemblance

- $\mathbb{E}_{\mathrm{x}}\left[\log \frac{p(x_i|\mathrm{x}_{\theta_i})}{q(x_i|\mathrm{x}_{\setminus i},\theta_i,\tau_i)}\right]$ correspond à un ratio de vraisemblance entre la vrai distribution et la distribution donnée par le modèle étant donné les variables sélectionnées.
- La valeur 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 $q(x_i|\mathrm{x}_{\setminus i},\theta_i,\tau_i) \approx p(x_i|\mathrm{x}_{\theta_i})$, c'est à dire tel que $\mathbb{E}_{\mathrm{x}}\left[\log \frac{p(x_i|\mathrm{x}_{\theta_i})}{q(x_i|\mathrm{x}_{\setminus i},\theta_i,\tau_i)}\right] \approx 0$.

Partie "structurelle" du score de vraisemblance

Il reste donc le second terme $I(X_i, X_{\bar{\theta}_i} | X_{\theta_i})$ qui est capital pour notre problème. Il correspond à l'information mutuelle entre la variable cible X_i et les variables $X_{\bar{\theta}_i}$ que l'on n'a pas sélectionnées conditionnellement aux variables X_{θ_i} que l'on a sélectionnées.

Interprétation du problème de régression du point de vue de la théorie de l'information

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

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

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

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 $X_{nb_{\mathcal{G}^m}(V_i)}$, c'est à dire tel que $X_{nb_{\mathcal{G}^m}(V_i)} \subset \mathrm{X}_{\theta_i}$, alors on a quand même $I(X_i, \mathrm{X}_{\bar{\theta}_i} | \mathrm{X}_{\theta_i}) = 0$.
- 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 $\mathrm{X}_{\theta_i} = X_{nb_{\mathcal{G}^m}(V_i)}$ tel que $I(X_i, \mathrm{X}_{\bar{\theta}_i} | \mathrm{X}_{\theta_i}) = 0$.

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

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

- L'hypothèse de faithfullness : $\{V_i, V_j\} \in E \implies X_i \not\perp\!\!\!\perp X_j | \mathbf{X}_{\setminus i,j}$
- On a besoin de cette hypothèse pour dire que si omet par exemple de sélectionner une variable X_k de $X_{nb_{\mathcal{G}^m}(V_i)}$ lors de la sélection de variable alors comme $X_k \in \mathrm{X}_{\bar{\theta}_i}$ on a :

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

$$\geq I(X_i, X_k | \mathrm{X}_{\setminus i,k}) \tag{36}$$

$$\tag{37}$$

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

Couverture de Markov

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

Des méthodes d'inférence de réseaux mettant en jeu des modèles de régression

- Un grand nombre de méthodes utilisées en bioinformatique qui marchent bien en pratique comme GENIE3 (Huynh-Thu et al., 2010) ou TIGRESS (Haury et al., 2012) 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 $X_{nb_{\mathcal{G}^m}(V_i)}$ 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 V_i certains n'ont même pas de relation causale directe avec V_i !

Section 4

Références

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.

Haury, A.-C., Mordelet, F., Vera-Licona, P., and Vert, J.-P. (2012). Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6:1–17.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776.

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*.