Inférence de réseaux

Cours 3 - graphes dirigés

Olivier Goudet

LERIA, Université d'Angers

27 janvier 2023

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.

Section 1

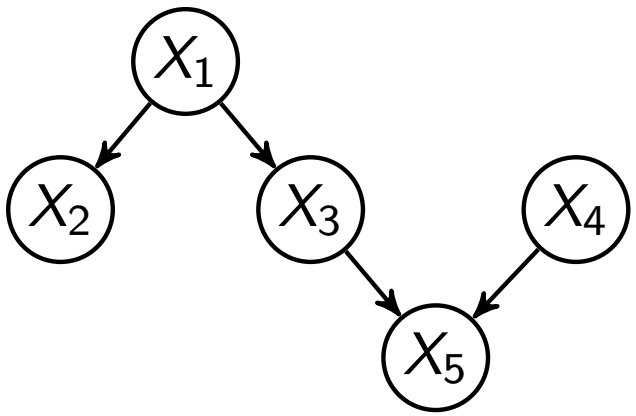
Modèles graphiques dirigés

Modèles graphiques dirigés

- Dans cette partie, on considère uniquement des graphes dirigés sans cycles, appelés DAG pour *Directed Acyclic Graph*.
- 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 dirigés du graphe. On note $(X_i, X_j) \in E$ une paire ordonnée qui correspond à un lien dirigé de X_i vers X_j.
  - On définit l'ensemble des parents de X_j dans le graphe $\mathcal{G}$ comme $pa_{\mathcal{G}}(X_j) = \{X_i \in V : (X_i, X_j) \in E\}$.
  - Ensemble des descendants de X_j dans $\mathcal{G}$: $de_{\mathcal{G}}(X_j) = \{X_i \in V : X_i = X_j \text{ ou } X_j \to \cdots \to X_i \in \mathcal{G}\}$.
  - Ensemble des non-descendants de X_j dans $\mathcal{G}$: $nd_{\mathcal{G}}(X_j) = V \backslash De_{\mathcal{G}}(X_j)$.
  - Ensemble des ancêtres de X_j dans $\mathcal{G}$: $an_{\mathcal{G}}(X_j) = \{X_i \in V : X_i = X_j \text{ ou } X_i \to \cdots \to X_j \in \mathcal{G}\}$.

Représentation du graphe dirigé par une matrice d'adjacence

- Un graphe dirigé avec d variables peut être représenté par une matrice binaire A représentant tous les liens dirigés entre les variables.
- $A_{ij} = 1$ si et seulement si il y a un lien dirigé du noeud X_i vers le noeud X_j dans le graphe $\mathcal{G}$, $A_{ij} = 0$ sinon.
- Remarque : pour un DAG, la matrice A peut être mise sous la forme d'une matrice triangulaire supérieure stricte (moyennant éventuellement un réindiçage des variables). On a de plus toujours $A^d = 0$.



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

Chemins dans un graphe dirigé

- On dit que les noeuds X_i et X_j sont adjacents dans le DAG $\mathcal{G} = (V, E)$ si $(X_i, X_j) \in E$ ou $(X_j, X_i) \in E$.
- Un chemin est une séquence de noeuds tels que les noeuds successifs sont adjacents.
- Si $\pi = (X_0, X_1, ..., X_k)$ est un chemin alors on dit que X_0 et X_k sont les points de terminaison du chemin π.
- Une variable X_i qui n'est pas un point de terminaison est un *collider* de π si $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$ est un sous-chemin de π. Sinon X_i est un *non-collider* de π.

Notions de d-connexion et de d-separation

- Deux noeuds X_i et X_j dans un DAG $\mathcal{G} = (V, E)$ sont dits *d-connectés* étant donné un ensemble de variables $C \subseteq V \backslash \{X_i, X_j\}$ si $\mathcal{G}$ contient un chemin π avec les points de terminaison X_i et X_j tel que :
  - Tous les *colliders* de π sont dans $an_{\mathcal{G}}(C)$
  - Il n'y a pas de *non-collider* de π dans C.
- On dit que deux ensembles de variables disjoints $A, B \subset V$ sont *d-connectés* étant donné $C \subseteq V \backslash (A \cup B)$ si il existe deux noeuds $X_i \in A$ et $X_j \in B$ qui sont *d-connectés* étant donné C.
- Si ce n'est pas le cas, alors C *d-sépare* A et B.

Hypothèses sur les données pour les graphes dirigés (1/3) - Propriétés de Markov

- Un vecteur de variable aléatoire $\mathbf{X} = (X_i : X_i \in V)$ satisfait la propriété de Markov locale par rapport au DAG $\mathcal{G}$ si pour tout $X_i \in V$:

$$X_i \perp\!\!\!\perp X_{nd_{\mathcal{G}}(X_i) \setminus pa_{\mathcal{G}}(X_i)} | X_{pa_{\mathcal{G}}(X_i)} \tag{2}$$

- $\mathbf{X}$ satisfait la propriété de Markov globale par rapport au graphe $\mathcal{G}$ si $X_A \perp\!\!\!\perp X_B | X_C$ pour tous les triplets d'ensembles de variables disjoints $A, B, C \subset V$ tels que C *d-separe* A et B dans $\mathcal{G}$, ce qui peut se noter $A \perp\!\!\!\perp_{\mathcal{G}} B | C$.
- Si $\mathcal{G}$ est un DAG et que $\mathbf{X}$ admet une densité de probabilité continue et positive, les propriétés de Markov locale et globale sont équivalentes (Lauritzen et al., 1990).
- Si $\mathbf{X}$ satisfait la propriété de Markov globale par rapport au graphe $\mathcal{G}$ alors $\mathcal{G}$ est appelé une *independence map* de $\mathbf{X}$.

Hypothèses sur les données pour les graphes dirigés - *faithfullness* (2/3)

- Un DAG est une *perfect map* de $\mathbf{X}$ si pour tout les ensembles de variables disjoints deux à deux $A, B, C \subseteq V$: $A \perp\!\!\!\perp_{\mathcal{G}} B|C$ si et seulement si $X_A \perp\!\!\!\perp X_B|X_C$.
- Une *perfect map* nécessite donc la propriété de Markov globale mais aussi son implication inverse appelé *faithfulness* : $X_A \perp\!\!\!\perp X_B|X_C \Rightarrow A \perp\!\!\!\perp_{\mathcal{G}} B|C$.

Hypothèses sur les données pour les graphes dirigés - *causal sufficiency* (3/3)

- Hypothèse de *causal sufficiency* : il n'existe pas de paire de variable $\{X_i, X_j\}$ de $\mathbf{X}$ ayant une cause commune qui n'est pas dans $\mathbf{X}_{\setminus i,j}$.
- Cette hypothèse est souvent faite à cause d'effets confondants cachés.
- En effet, si une variable "cachée" cause à la fois X_i et X_j, ces deux variables peuvent devenir dépendantes, alors qu'il n'y pas d'arc dirigé entre les deux (cf. exemple chocolat et prix Nobel).

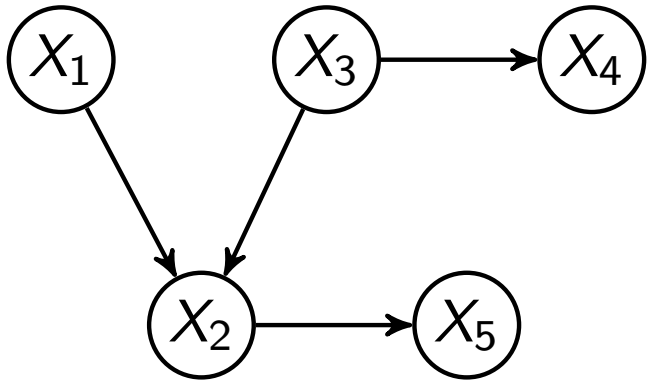
Factorisation de la distribution jointe suivant le graphe

- On dit que la distribution jointe $p(\mathbf{x})$ de $\mathbf{X}$ se factorise suivant le DAG $\mathcal{G} = (V, E)$ si on peut écrire :

$$p(\mathbf{x}) = \prod_{j=1}^{d} p(x_j | pa_{\mathcal{G}}(x_j)) \tag{3}$$

- On parle dans ce cas de **réseau bayésien**.
- Cette propriété de factorisation est équivalente à la propriété de Markov locale et globale dans le cas d'un DAG (Verma and Pearl, 1990).
- Intérêt de cette factorisation : la fonction de vraisemblance du modèle se factorise en d fonctions de vraisemblance locales qui peuvent se calculer séparément.

Exemple de DAG



- Le noeud X_2 est un *collider* du chemin $X_1 \to X_2 \leftarrow X_3 \to X_4$, tandis que ce n'est pas un *collider* du chemin $X_1 \to X_2 \to X_5$.
- Pour le noeud X_4, la propriété de Markov locale requiert $X_4 \perp\!\!\!\perp (X_1, X_2, X_5) | X_3$.
- X_1 et X_4 sont d-connectés étant donné $C = \{X_2\}$, $C = \{X_5\}$ et $C = \{X_2, X_5\}$, mais d-séparés étant donnés tous les autres sous-ensembles C' de $C = \{X_2, X_3, X_5\}$, la propriété de Markov globale impose que $X_1 \perp\!\!\!\perp X_4 | C'$ pour tous ces sous-ensembles C'.

Exemple

- On remarque que contrairement à la séparation dans les graphes non dirigés, la d-separation dans les DAG n'est pas monotone, dans le sens où $A \perp\!\!\!\perp_{\mathcal{G}} B | C$ n'implique pas que $A \perp\!\!\!\perp_{\mathcal{G}} B | C'$ pour tout ensemble C' tel que $C \subsetneq C'$.
- La factorisation de $p(\mathbf{x})$ suivant $\mathcal{G}$ prend la forme :

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1, x_3)p(x_3)p(x_4|x_3)p(x_5|x_2)$$

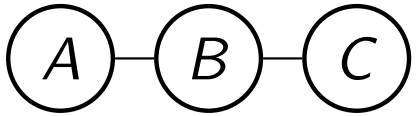
Graphes Markov equivalents

- Deux graphes $\mathcal{G}$ et $\mathcal{G}'$ sont Markov equivalents si $A \perp\!\!\!\perp_{\mathcal{G}} B | C$ est équivalent à $A \perp\!\!\!\perp_{\mathcal{G}'} B | C$.
- Deux graphes sont Markov equivalents si ils ont le même squelette et les mêmes *v-structures* (Verma and Pearl, 1991).
- Une *v-structure* est un triplet de noeuds tels que $X_i \rightarrow X_k \leftarrow X_j$ avec X_i et X_j non adjacents.

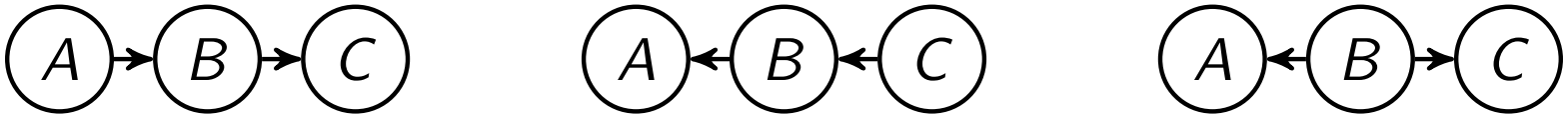
CPDAG : Completed Partially Directed Acyclic Graph

- Chaque classe d'équivalence de Markov peut être représentée par un CPDAG (*Completed Partially Directed Acyclic Graph*) qui a des liens dirigés et des liens non dirigés.
- Un CPDAG a le lien dirigé $X_i \to X_j$ si et seulement si cette arc $X_i \to X_j$ est commun à tous les DAGs de la classe d'équivalence.
- Si la classe contient un DAG avec $X_i \to X_j$ et un autre DAG avec $X_j \to X_i$ alors le CPDAG a le lien non dirigé $X_i - X_j$

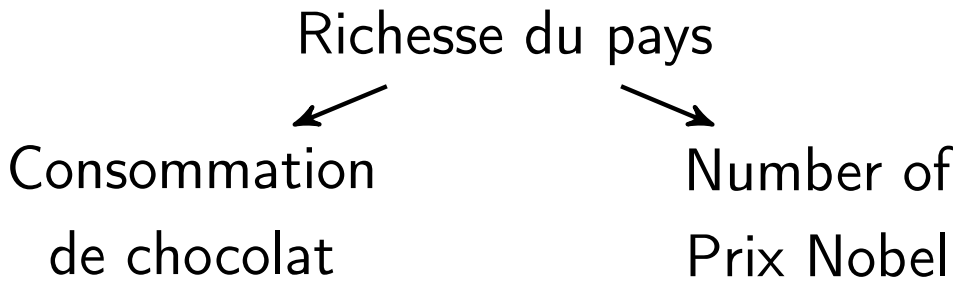
Exemple : classes d'équivalence de Markov

Squelette: 

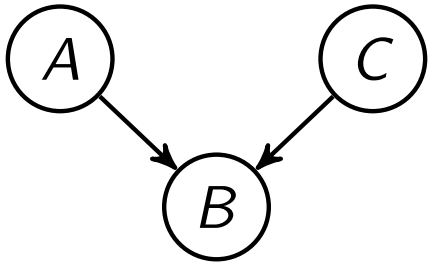
3 classes d'équivalence de Markov: $A \perp\!\!\!\perp C | B$



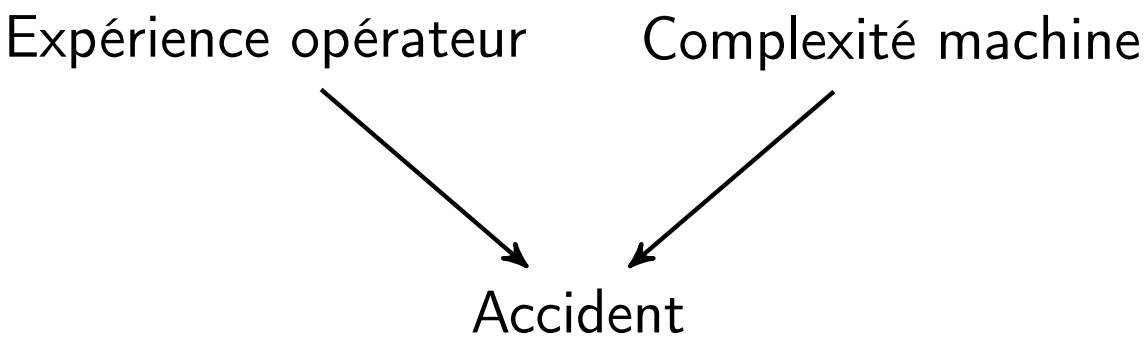
Exemple



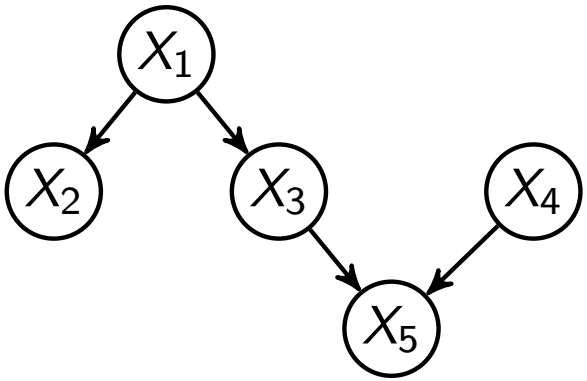
V-Structure: $A \not\perp\!\!\!\perp C | B$



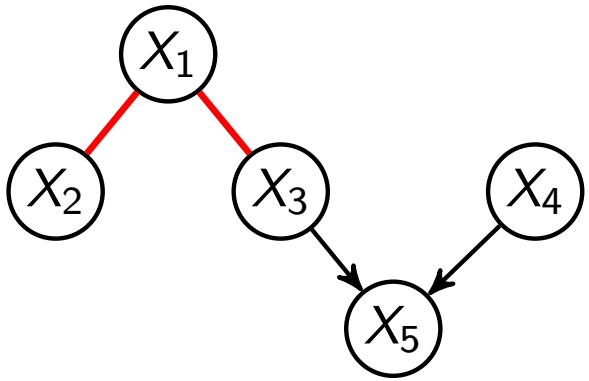
Exemple



Exemple CPDAG



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

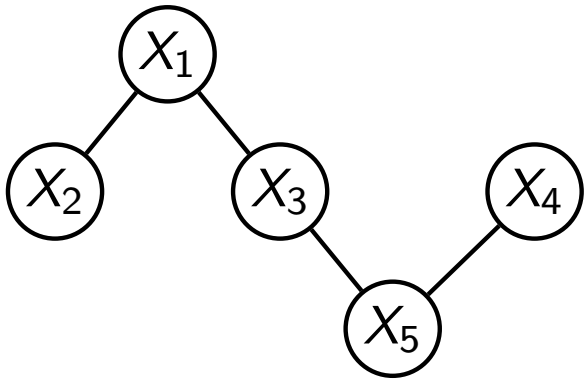


(b) Le CPDAG de $\mathcal{G}$.

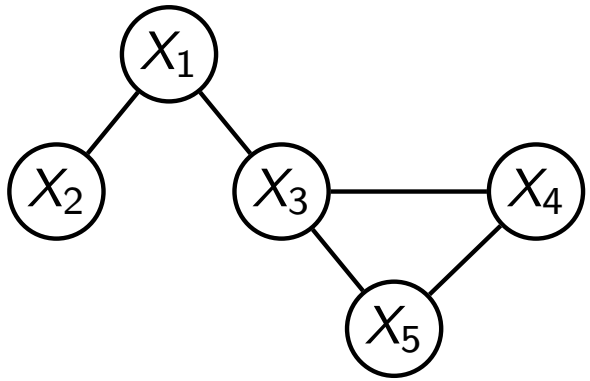
Lien entre squelette et graphe moral

- Le squelette d'un graphe dirigé est le graphe non dirigé obtenu en remplaçant tous les arcs dirigés par des arcs non dirigés.
- Le graphe moral $\mathcal{G}^m$ de $\mathcal{G}$ est construit en ajoutant un arc entre X_i et X_j pour chaque *v-structure* $X_i \rightarrow X_k \leftarrow X_j$ et en prenant le squelette du graphe résultant.
- Si G est une *perfect map* de $\mathbf{X}$ alors $\mathcal{G}^m$ est le graphe des indépendances conditionnelles de $\mathbf{X}$ (cf. cours 2).
- Le squelette d'un DAG est un sous-graphe de son graphe des indépendances conditionnelles.

Squelette et graphe moral



(a) Le squelette du DAG $\mathcal{G}$.



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

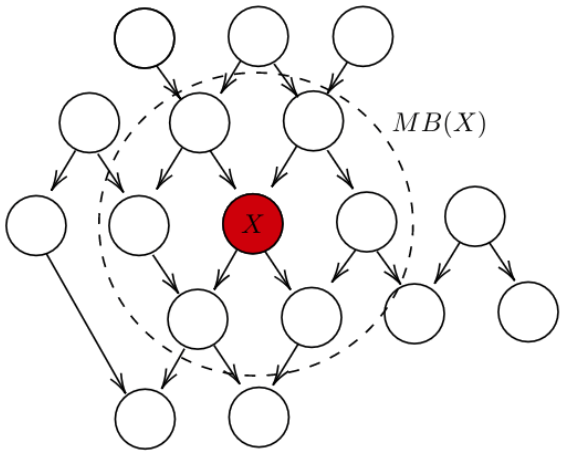
Couverture de Markov

- La couverture de Markov (*Markov Blanket*) de la variable X_i est le sous ensemble minimal de variables $MB(X_i)$ de $X_{\setminus i}$ tel que :

$$X_i \perp\!\!\!\perp X_{V \setminus (MB(X_i) \cup X_i)} | MB(X_i)$$

.

- Si $\mathcal{G}$ est une *perfect map* de $\mathbf{X}$ alors $MB(X_i)$ correspond à l'ensemble des voisins de X_i dans le graphe moral de $\mathcal{G}$: $MB(X_i) = nb_{\mathcal{G}^m}(X_i)$.
- L'identification de cette couverture de Markov est un problème général de sélection de variables (cf. fin du cours 2).



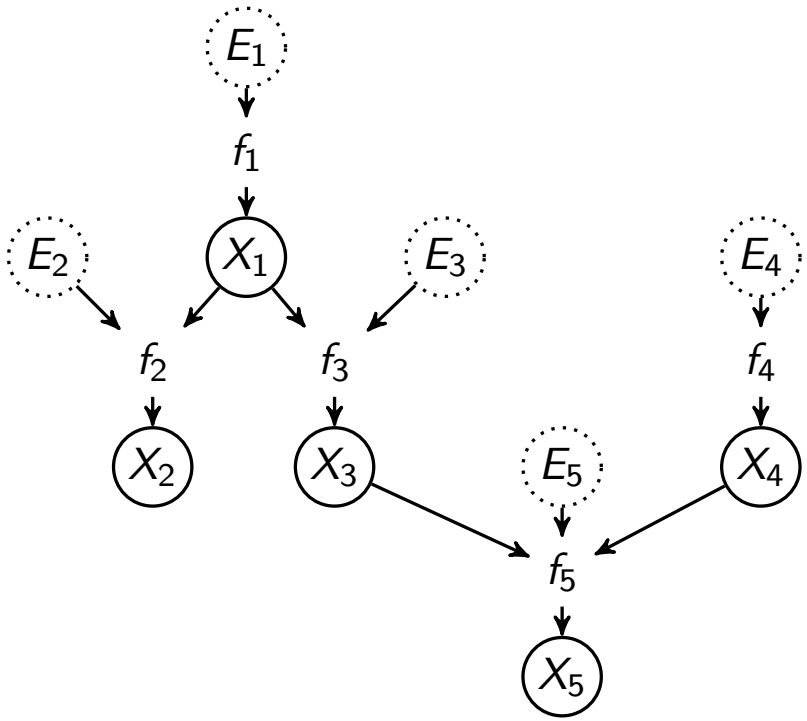
Modèle fonctionnel causal ou SEM (*Structural Equation Model*)

- Un DAG $\mathcal{G} = (V, E)$ peut aussi être vu comme un modèle fonctionnel causal (FCM : *Fonctional Causal Model*) ou modèle d'équations structurelles (SEM).
- Si $\mathbf{X}$ satisfait la propriété de Markov par rapport à $\mathcal{G}$ alors il existe des variables aléatoires indépendantes E_i et des fonctions f_i telles que :

$$X_i \leftarrow f_i(X_{\mathrm{Pa}(i;\mathcal{G})}, E_i), \text{ for } i = 1, \ldots, d \tag{4}$$

- De façon réciproque s'il existe un FCM satisfait par $\mathbf{X}$ alors $\mathbf{X}$ satisfait la propriété de Markov par rapport à $\mathcal{G}$.

Exemple FCM



$$\begin{cases} X_1 = f_1(E_1) \\ X_2 = f_2(X_1, E_2) \\ X_3 = f_3(X_1, E_3) \\ X_4 = f_4(E_4) \\ X_5 = f_5(X_3, X_4, E_5) \end{cases}$$

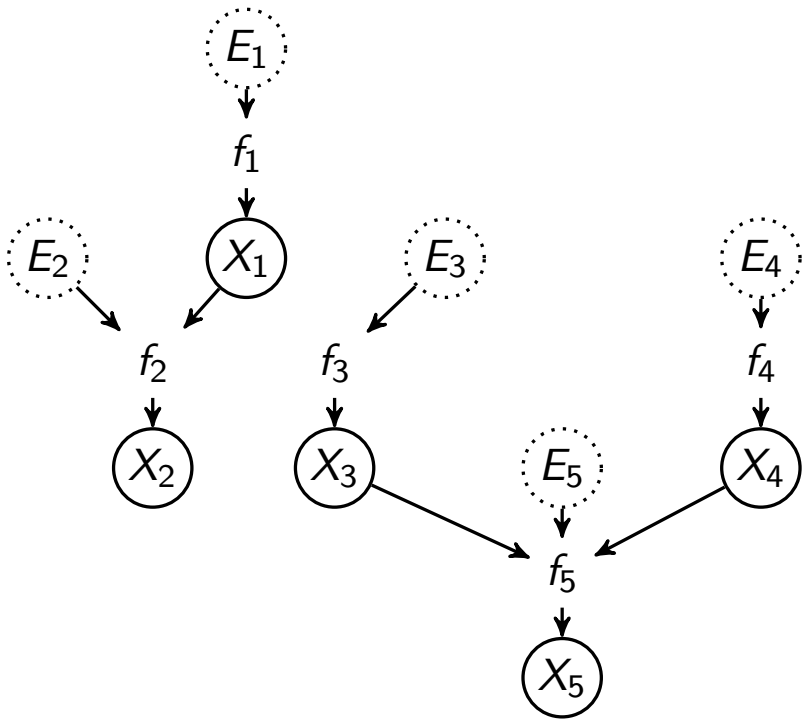
Figure 1: Exemple de modèle fonctionnel causal (FCM) avec $\mathbf{X} = [X_1, \ldots, X_5]$: A gauche : graphe causal $\mathcal{G}$; A droite : mécanismes causaux.

FCM et réseaux bayésiens

- Connaître le FCM, donne une factorisation de p suivant le graphe (réseau bayésien).
- Par contre connaître le réseau baysien et la factorisation de p ne donne pas l'expression du FCM.
- La représentation avec un FCM donne plus d'information que la représentation avec un réseau bayésien.
- La représentation avec un FCM permet de raisonner sur des interventions qui pourraient être effectuées sur le graphe causal.

Exemple d'intervention expérimentale

- A partir du graphe précédent, on peut choisir d'effectuer une intervention expérimentale sur X_3, de telle manière que X_3 soit générée à partir de E_3 uniquement et non plus à partir de X_1 et E_3 (on coupe le lien $X_1 \to X_3$).
- On peut en déduire la nouvelle distribution d'intervention résultante à partir du nouveau modèle causal :



$$\begin{cases} X_1 = f_1(E_1) \\ X_2 = f_2(X_1, E_2) \\ X_3 = f_3(E_3) \\ X_4 = f_4(E_4) \\ X_5 = f_5(X_3, X_4, E_5) \end{cases}$$

Lien entre FCM et modèle Gaussien

- Si toutes les fonctions f_i sont linéaires et que les variables aléatoires E_i sont indépendantes entre elles et suivent chacune une distribution normale centrée réduite, alors on peut écrire :

$$\mathbf{X} = \mathbf{X}B + \mathbf{E},$$

  avec la matrice $B \in \mathbb{R}^{d \times d}$ qui a des coefficient nuls $b_{i,j}$ si $X_i \notin pa_{\mathcal{G}}(X_j)$ et $\mathbf{E} = (E_1, ..., E_d)$.
- Ici $(I - B)$ est inversible car $\det(I - B) = 1$, car $\mathcal{G}$ est acyclique (B triangulaire supérieure stricte).
- La solution du FCM est $\mathbf{X} = \mathbf{E}(I - B)^{-1}$.
- $\mathbf{X}$ suit donc une distribution Gaussienne multivariée de matrice de covariance $\mathrm{Cov}(\mathbf{X}) = (I - B)^{-1^T} \mathrm{Cov}(\mathbf{E})(I - B)^{-1}$.

Apprentissage d'un graphe dirigé dans le cas Gaussien multivarié

- Dans le cas Gaussien multivarié, on peut utiliser le même type d'approche que la méthode du Glasso qui était utilisée pour apprendre un graphe non-dirigé.
- La matrice de précision de $\mathbf{X}$ est $K(B) = (I - B)K_E(I - B)^T$.
- Si on note $S = XX^T$ l'estimée empirique de Cov($\mathbf{X}$), l'estimateur de vraisemblance du modèle Gaussien graphique est (cf. cours 2) :

$$L(B) = \frac{1}{2} log(\det\ K(B)) - \frac{1}{2} tr(K(B)S) + cst \tag{5}$$

- Le problème a résoudre est :

$$\min_{B \in \mathbb{B}} \ -L(B) + \rho_\lambda(B) \tag{6}$$

- avec $\mathbb{B} \subset \mathbb{R}^{d \times d}$, l'ensemble des matrices d'adjacence représentant un graphe dirigé acyclique et $\rho_\lambda(B)$ un terme de régularisation ℓ_1 sur l'ensemble des coefficients de B.

Méthodes d'apprentissage locales à base de contraintes

- Idée générale de ces méthodes : exploiter les tests d'indépendance conditionnelle pour retrouver le DAG $\mathcal{G}$.
- D'après les hypothèses de Markov, de faithfullness et de *causal sufficiency*, moyennant le fait d'établir toutes les relations d'indépendance conditionnelle entre les variables, on retrouve le CPDAG de $\mathcal{G}$.
- Pour qu'une garantie théorique d'identifiabilité soit apportée, il est nécessaire de supposer que les tests d'indépendance conditionnelle sont "parfaits" (oracle).

Méthodes d'apprentissage locales à base de contraintes

- Ces algorithmes d'inférence se déroulent classiquement en trois étapes :
  1. retrouver le squelette du graphe,
  2. identifier les v-structures,
  3. appliquer des règles de propagation.
- Avantage de ces méthodes : tout type de test d'indépendance conditionnelle peut être utilisé (e.g. corrélation partielle ou tests non paramétriques comme KCI (Zhang et al., 2012)).
- Désavantage : en pratique difficile de faire ces tests en haute dimension + propagations d'erreurs possibles.

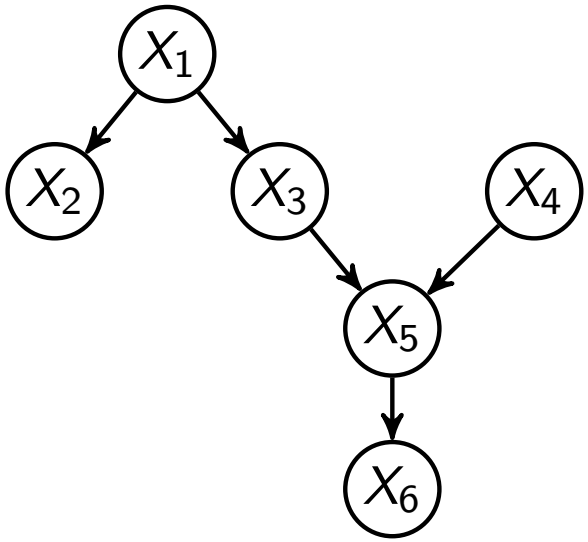
Retrouver le squelette

- Si $\mathcal{G}$ est une *perfect map*, on sait X_i et X_j sont adjacents dans $\mathcal{G}$ si et seulement si X_i et X_j sont conditionnellement dépendant étant donné tout sous-ensemble $C \in V \backslash \{X_i, X_j\}$.
- Une approche naive pour savoir si X_i et X_j sont adjacents est de faire touts les tests d'indépendance conditionnelle par rapport à tous les sous-ensembles de variables $C \in V \backslash \{X_i, X_j\}$ (Algorithmes SGS (Verma and Pearl, 1991)).
- Le nombre de tests à effecter augmente exponentiellement avec $card(V)$.
- Une amélioration a été proposée par Spirtes et al. (2000) pour réduire le nombre de tests à effectuer (Algorithme PC).

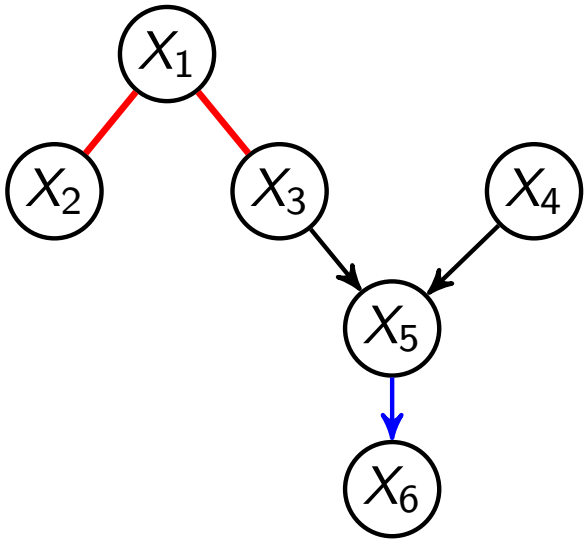
Exemple algorithme à base de contraintes pour 3 variables (Spirtes et al., 2000)

- On suppose qu'on a trois variables X_1, X_2 et X_3.
- On part du graphe complet entre toutes les variables.
- On effectue tout d'abord les tests d'independance entre chaque paire de variables. On trouve $X_1 \perp\!\!\!\perp X_3$, donc on retire l'arc entre X_1 et X_3.
- On ne trouve pas d'autres indépendances conditionnelles dans ce cas. Donc le squelette est $X_1 - X_2 - X_3$.
- Orientation : si on observe $X_1 \perp\!\!\!\perp X_3 | X_1$ alors il y a trois DAG équivalents possibles : $X_1 \rightarrow X_2 \rightarrow X_3$ ou $X_1 \leftarrow X_2 \leftarrow X_3$ ou $X_1 \leftarrow X_2 \rightarrow X_3$.
  Le CPDAG est $X_1 - X_2 - X_3$.
- Si on n'observe pas $X_1 \perp\!\!\!\perp X_3 | X_2$, alors on identifie un unique DAG : $X_1 \rightarrow X_2 \leftarrow X_3$.

Exemple, algorithme à base de contrainte : propagation



(a) The exact DAG of $\mathcal{G}$.



(b) The CPDAG of $\mathcal{G}$.

Algorithmes gloutons à base de score

- Ces méthodes recherchent les graphes qui minimisent un score globale.
- Dans le cas Gaussien le BIC score (Bayesian Information Criterion) est utilisé :

$$S(\mathcal{G}', \mathcal{D}) = -2 \sum_{\ell=1}^{n} \sum_{i=1}^{d} \log q(x_i^\ell | \mathbf{x}_{\setminus \mathbf{i}}^\ell, \tau_i^*) + k \log(\mathrm{n}) \tag{7}$$

- k est le nombre de paramètres du modèle. $\{\tau_i^*\}_{i=1}^d$ sont les jeux de paramètres qui minimise cette somme des scores de log-vraisemblance pour un graphe candidat $\mathcal{G}'$ donné.
- Idée : explorer l'espace des DAG $\mathcal{G}'$ possibles avec un algorithme glouton de façon à minimiser ce score globale (opérateur : ajouter un arc, enlever un arc, retourner un arc)
- Dans le cas Gaussien mulltivarié, si on fait les hypothèses de Markov, de *faithfullness* et de *causal sufficiency*, quand $n \to \infty$, si on minimise le BIC score, on retrouve le CPDAG de $\mathcal{G}$ (algorithme GES (Chickering, 2002))

Algorithmes hybrides

- D'autres algorithmes comme MMHC (Tsamardinos et al., 2006) font d'abord une recherche du graphe moral en identifiant toutes les couvertures de Markov pour chaque variable (cf. Cours 2).
- Une recherche du meilleur score avec un score BIC au sein du graphe moral est ensuite effectuée de façon à identifier le CPDAG.

Limites de ces algorithmes

- Toutes ces méthodes sont limitées à la recherche de la classe d'équivalence du DAG.
- En particulier si seulement deux variables sont observées, on ne peut pas identifier de v-structures.
- Est-il possible de retrouver quand même un DAG ?

Section 2

Causalité paire à paire

Causalité paire à paire

- Exploiter l'idée de simplicité du mécanisme causale dans une direction plutôt que l'autre ("School of Tuebingen" (Hoyer et al., 2009; Janzing and Schölkopf, 2010)).

Exemple

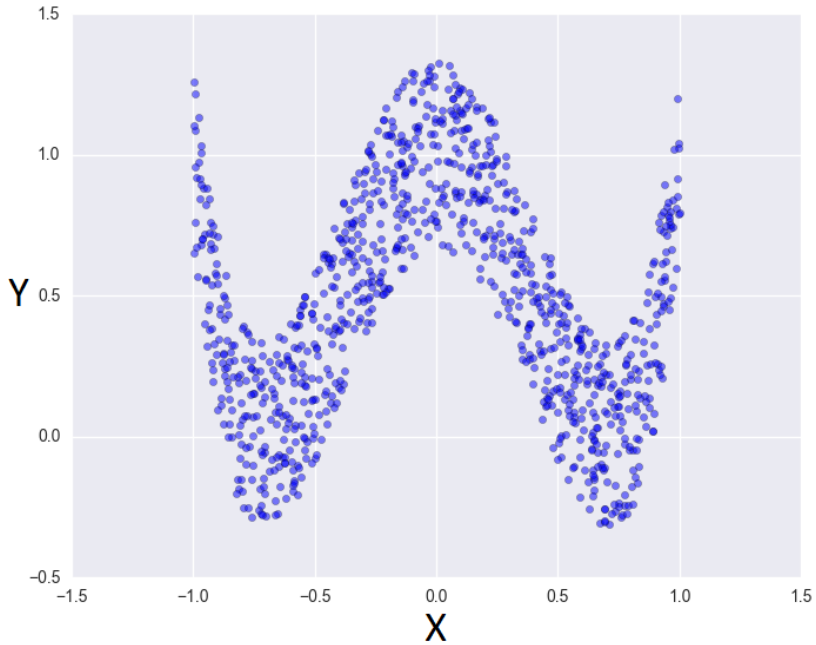


Figure 2

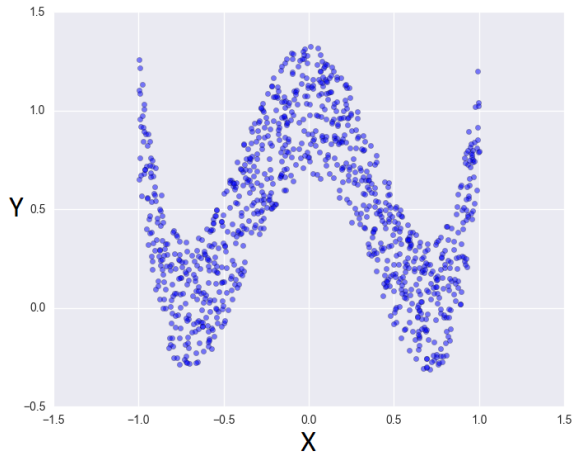
Coefficient de corrélation égal à 0 ! Mais plutôt facile de retrouver la direction causale...

Modèle génératif sous-jacent

Les données ont été générées de X vers Y avec ce modèle stochastique :

$$X \sim \mathcal{U}(-1, 1) \tag{8}$$
$$N_Y \sim \mathcal{U}(-1, 1)/3 \tag{9}$$
$$Y := 4 \times (X^2 - 0.5)^2 + N_Y, \tag{10}$$

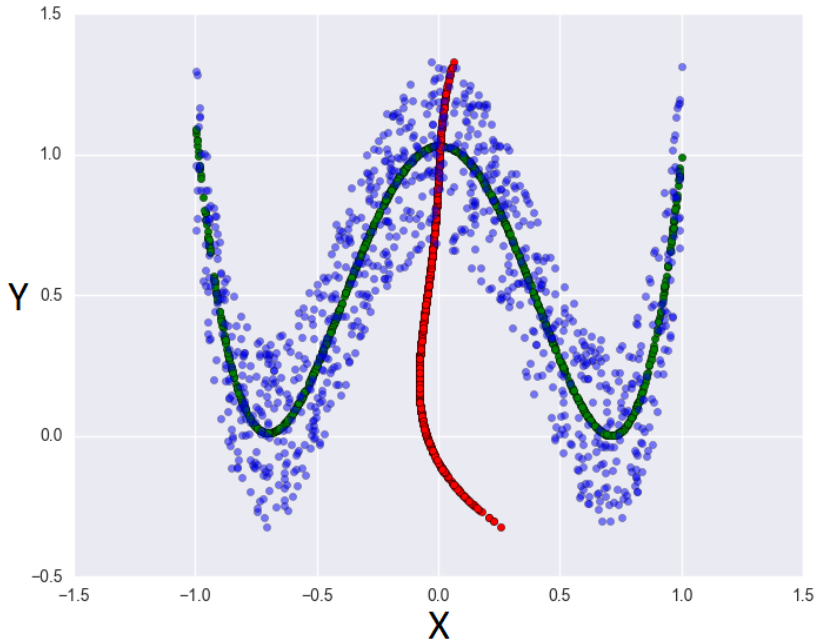
Modèle fonctionnel causal dans le cas bivarié



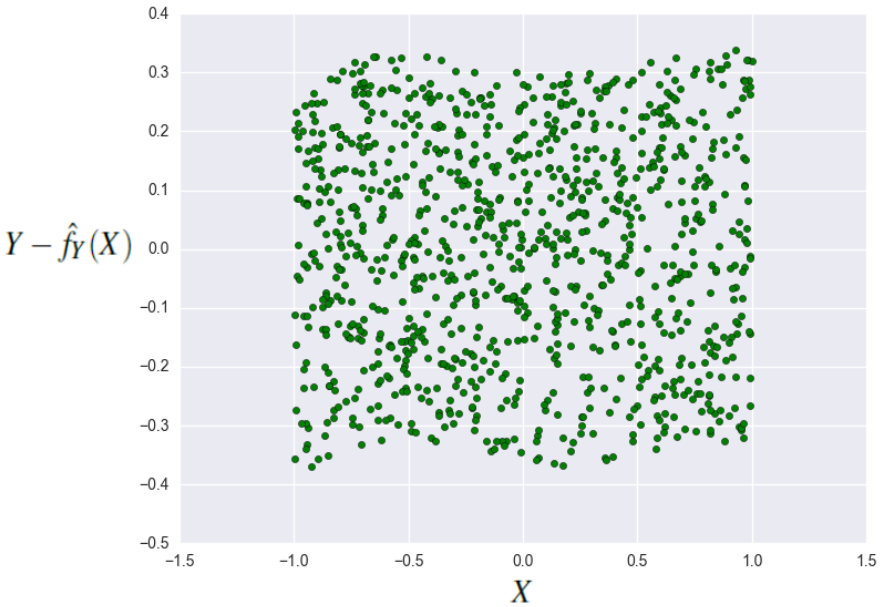
- Dépendence non-linéaire entre X et Y.
- On fait l'hypothèse que cette dépendance est due à $X \to Y$ ou $Y \to X$.
- Le cas $X \leftarrow Z \to Y$ est exclu.
- Une de ces deux hypothèses est vérifiée : (Mooij et al., 2016):
  1. $Y := f_Y(X, N_Y)$ avec $N_Y \perp\!\!\!\perp X$ (hypothèse 1, $X \to Y$),
  2. $X := f_X(Y, N_X)$ avec $N_X \perp\!\!\!\perp Y$ (hypothèse 2, $Y \to X$),

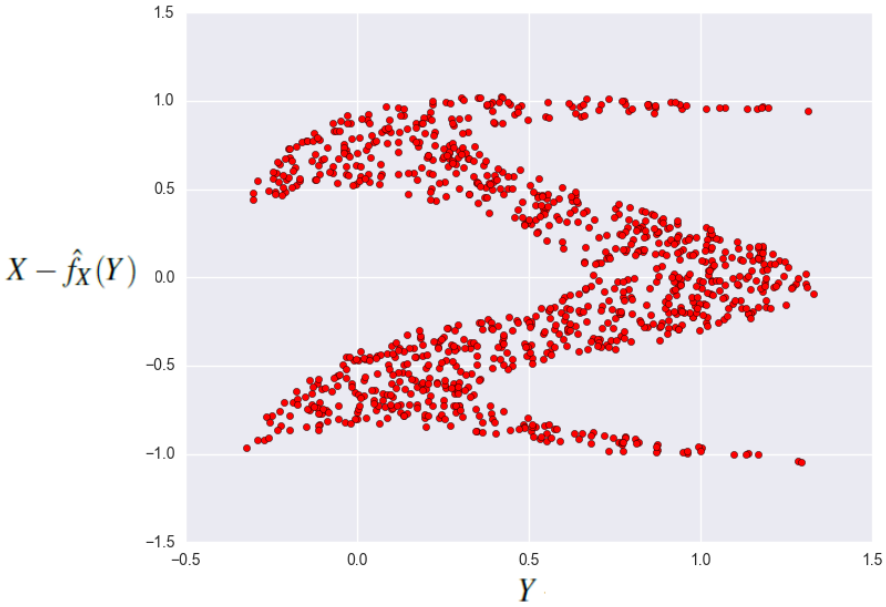
Retrouver le FCM dans le cas bivarié

Estimer deux modèles de régression polynomiale $\hat{f}_Y$ et $\hat{f}_X$ dans chaque direction :



Résidus de la régression



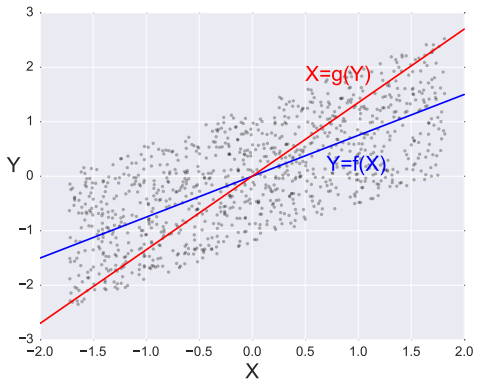


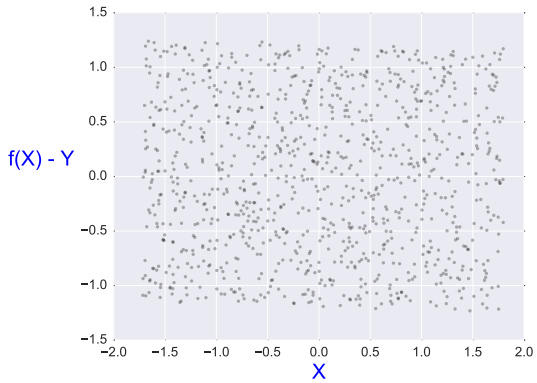
- Comme $(Y - \hat{f}_Y(X)) \perp\!\!\!\perp X$
  $\rightarrow$ modèle explicatif simple : $Y := \hat{f}_Y(X) + N_Y$ avec $N_Y \perp\!\!\!\perp X$ (Additive noise model (Hoyer et al., 2009))
- Comme $(X - \hat{f}_X(Y)) \not\perp\!\!\!\perp Y$
  $\rightarrow X := \hat{f}_X(Y) + N_X$ avec $N_X \perp\!\!\!\perp Y$ n'est pas satisfaisant.
- Un modèle explicatif $X := \hat{f}_X(Y, N_X)$ avec $N_X \perp\!\!\!\perp Y$ existe tout de même (Zhang and Hyvärinen, 2009), mais il est plus "complexe"...

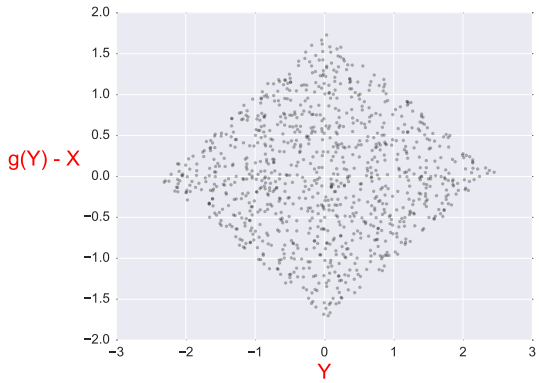
Causal additive noise model

- Causal additive noise model (ANM) Hoyer et al. (2009):
  $Y = f(X) + E$, avec $X \perp\!\!\!\perp E$
- Effectuer une régression et vérifier l'indépendance du résidu avec la cause.

Causal additive noise model (ANM) (Hoyer et al., 2009)





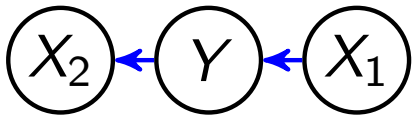


- Avec l'exemple 1 vu en introduction :

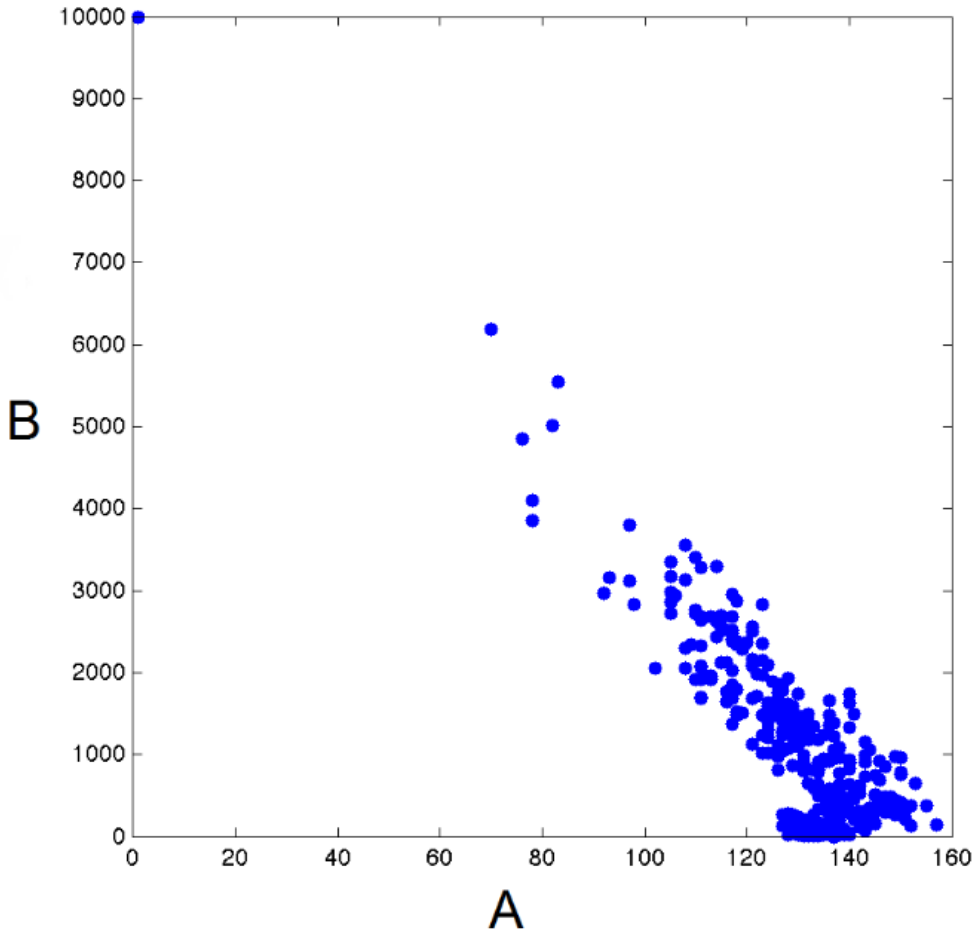
$$Y \leftarrow 0.5X_1 + E_{X_1},$$
$$X_2 \leftarrow Y + E_{X_2},$$

avec $X_1, E_1, E_2 \sim \text{Uniform}(0,1), X_1 \perp\!\!\!\perp E_1,\ Y \perp\!\!\!\perp E_2$

- En utilisant ANM on obtient $X_1 \to Y$ and $Y \to X_2$



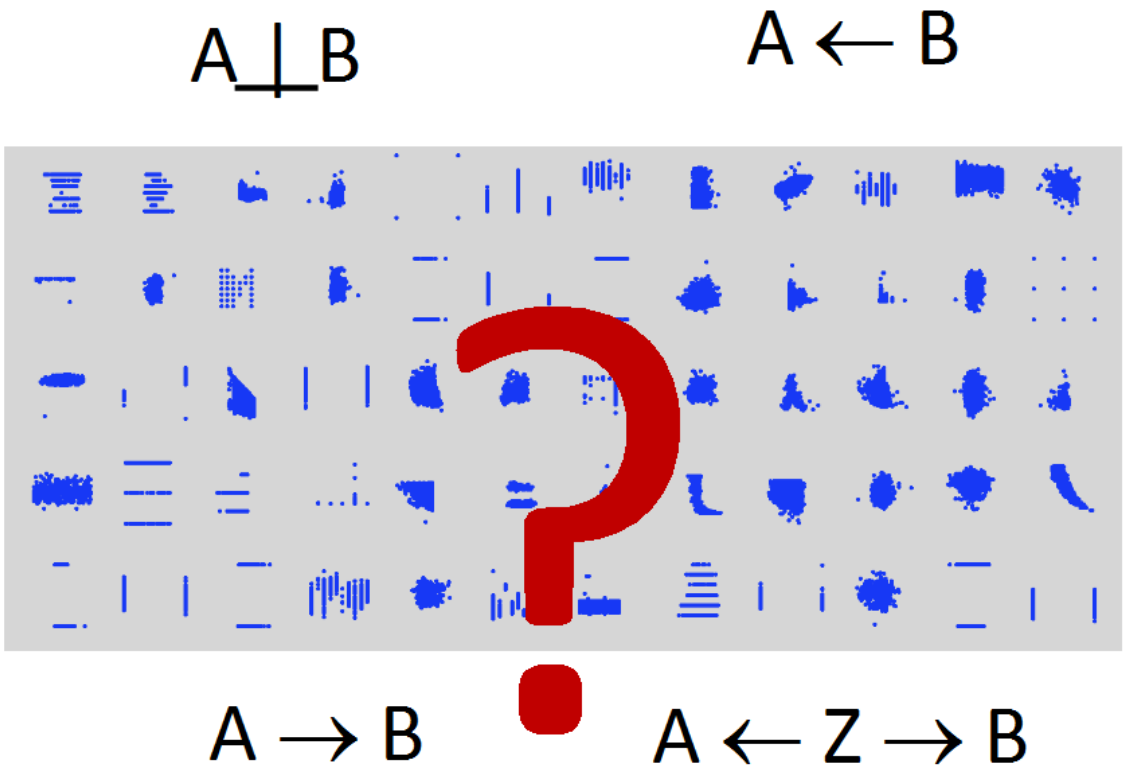
Quizz - paire réelle



Answer: A is temperature and B is altitude of German cities, hence B causes A (example provided by D. Janzig).

Plus sur le problème de causalité paire à paire

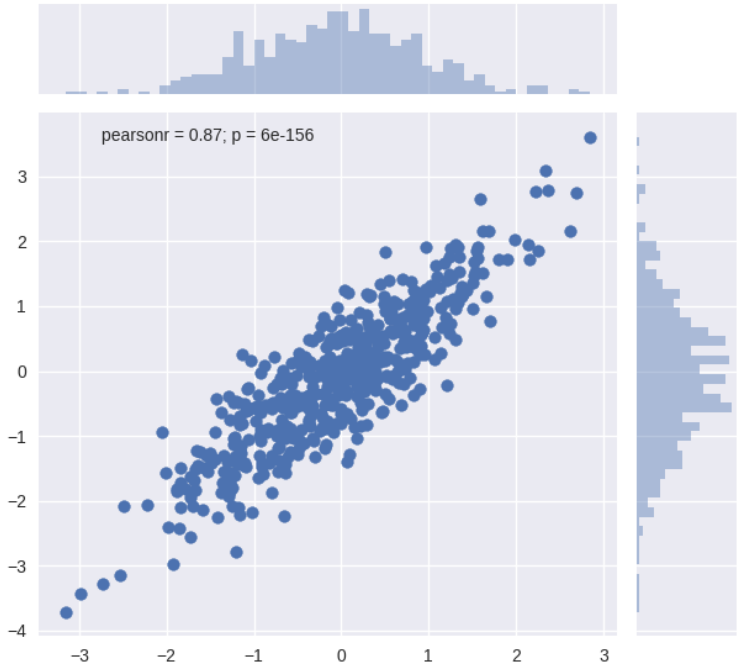
Cause effect pair challenge [Guyon 2013]



Livre : "Cause Effect Pairs in Machine Learning". Springer (2019)

Limitations des algorithmes de causalité paire à paire

- Cas non identifiables (ex: cas linéaire Gaussien) :

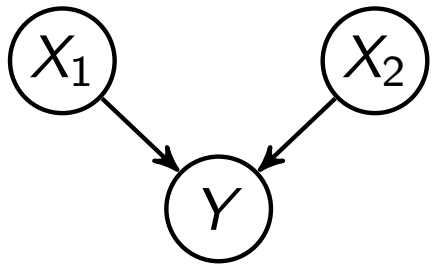


- Problème de généralité : Pour un grand nombre de paires réelles, dans les deux cas ($X \to Y$ et $Y \to X$) le modèle additif causal (ANM) ne correspond pas aux données.
- $\to$ des modèles plus généraux ont été proposés (Zhang et al., 2016): $y = g(f(X) + E)$ avec g une fonction inversible.

Limitations des algorithmes de causalité paire à paire

- Ne prennent pas en compte les relations d'indépendance conditionnelle. Par exemple :

$$X_1, X_2, E_{X_1} \sim \text{Gaussian}(0,1), X_1 \perp\!\!\!\perp E_{X_1},\ X_2 \perp\!\!\!\perp E_{X_1}$$
$$Y \leftarrow 0.5X_1 + X_2 + E_{X_1}$$



- (X_1, Y) et (X_2, Y) ne sont pas identifiables en paire à paire....
- ... alors que la v-structure $X_1 \not\perp\!\!\!\perp X_2 | Y$ peut être identifiée.

Section 3

Pour aller plus loin...

Structural agnostic model

- Apprentissage d'un modèle fonctionnel causal avec des réseaux de neurones génératifs.
- Modèle fonctionnel causal :

$$X_i \leftarrow f_i(X_{\mathrm{Pa}(i;\mathcal{G})}, E_i), \text{ for } i = 1, \ldots, d \tag{11}$$

- Idée : modéliser chaque mécanisme f_i par un réseau de neurones génératif.
- Un réseau de neurone discriminant sera entraîné en simultané pour juger de la qualité de reproduction des données.
- Structural agnostic model (SAM) (Kalainathan et al., 2022).

General framework - Functional Causal Models (Pearl, 2003)

We assume that the data have been generated by a Functional Causal Models (Pearl, 2003) $FCM = (\mathcal{G}, f, \mathcal{E})$ with $f = (f_1, \ldots, f_d)$:

$$X_j := f_j(X_{\mathrm{Pa}(j;\mathcal{G})}, E_j), \text{ with } E_j \sim \mathcal{N}(0,1) \text{ for } j = 1, \ldots, d. \tag{12}$$

1. $\mathcal{G}$ is assumed to be a Directed Acyclic Graph (DAG).
2. Each f_j is a deterministic and potentially a non-linear function from $\mathbb{R}^{|\mathrm{Pa}(j;\mathcal{G})|+1}$ to $\mathbb{R}$.
3. The noise variable E_j are independent from each other.

Goals

1. Unified framework - exploit all information available:
   - Exploit conditional independences (Key Idea 1) $\rightarrow$ *structure*
   - Exploit simplicity of mechanisms (Key Idea 2) $\rightarrow$ *mechanisms*
2. Agnostic approach:
   - No prior knowledge on the type of distribution (e.g. Gaussian)
   - No prior knowledge on the mechanisms (e.g. linear)
   - No prior knowledge on the interactions between noise and variables (e.g. additive noise)
3. Build a realistic simulator of the data: can be used to simulated intervention, perform counterfactual reasoning...

SAM : modelling FCM with generative neural networks

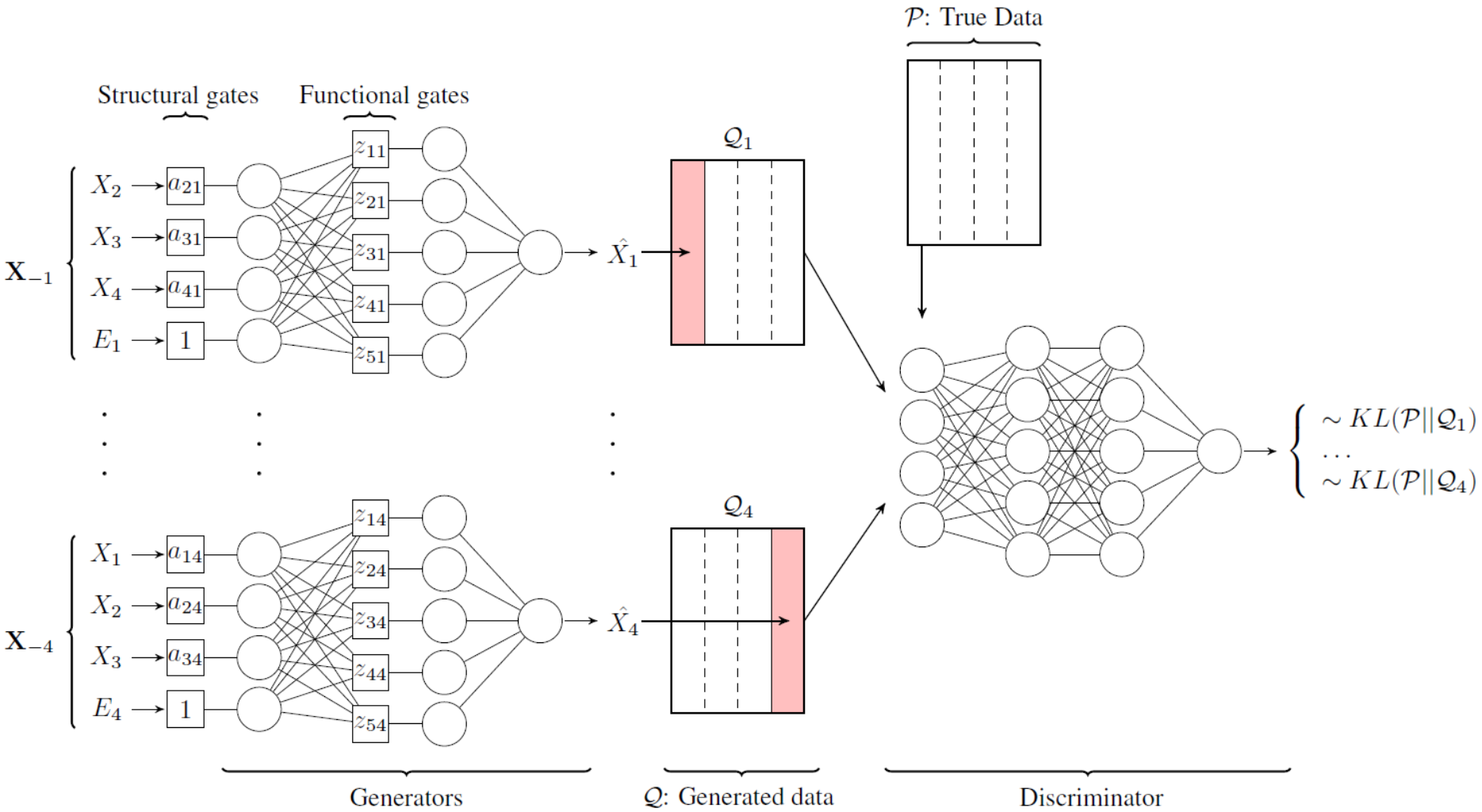
- Candidate model $\widehat{M} = (\widehat{\mathcal{G}}, \hat{f})$. For $j \in [[1, d]]$:

$$\hat{X}_j = \hat{f}_j(\mathbf{X}, E_j) \tag{13}$$

$$\hat{X}_j = m_j^\top \tanh\left(W_j^\top (a_j \odot \mathbf{X}) + n_j E_j + b_j\right) \odot z_j + \beta_j, \tag{14}$$

- **Structural gate** Boolean vector a_j.
  $a_{i,j} \in \{0, 1\}$ the i-th coefficient of the vector a_j. If $a_{i,j} = 1$, the directed edge $X_i \to X_j$ is in $\widehat{\mathcal{G}}$ (with $a_{i,i}$ always set at 0).
  $A = \{a_{i,j}\}_{i,j=1}^d$ is the adjacency matrix of the directed graph $\widehat{\mathcal{G}}$.
- **Functional gate** Boolean vector z_j.
  $z_{j,h} \in \{0, 1\}$ corresponds to the activation of the h-th hidden units of the neural network $\hat{f}_j$.

Structural agnostic model (SAM)



Sampling data with the model

- Each value of noise variable E_i is drawn independently from $\mathcal{N}(0, 1)$ at every evaluation.
- Each generator j sample the conditional generative distribution $q(x_j|x_{\text{Pa}(j;\hat{\mathcal{G}})}, \theta_j)$ with $\text{Pa}(j; \hat{\mathcal{G}}) = \{i \in [1, \ldots, d], a_{i,j} = 1\}$

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696.

Janzing, D. and Schölkopf, B. (2010). Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194.

Kalainathan, D., Goudet, O., Guyon, I., Lopez-Paz, D., and Sebag, M. (2022). Structural agnostic modeling: Adversarial learning of causal graphs. *Journal of Machine Learning Research*, 23(219):1–62.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed markov fields. *Networks*, 20(5):491–505.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016). Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102.

Pearl, J. (2003). Causality: models, reasoning and inference. *Econometric Theory*, 19(675-685):46.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78.

Verma, T. and Pearl, J. (1990). Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, pages 69–76. Elsevier.

Verma, T. and Pearl, J. (1991). *Equivalence and synthesis of causal models*. UCLA, Computer Science Department.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press.

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*.

Zhang, K., Wang, Z., Zhang, J., and Schölkopf, B. (2016). On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):13.