Inférence de réseaux

Causalité *pairwise*

Olivier Goudet

Université d'Angers

22 janvier 2026

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.
4. Cours 4 - O. Goudet
   - Causalité *pairwise*.
   - Méthodes d'inférence de réseaux utilisées en bioinformatique.

Plan

1 Introduction - causalité *pairwise*

2 Méthodes qui testent l'indépendance entre le bruit et la cause

3 Méthodes génératives

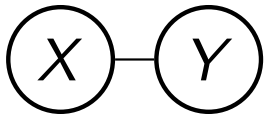
4 Méthodes discriminantes avec des classifieurs

Section 1

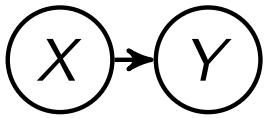
Introduction - causalité *pairwise*

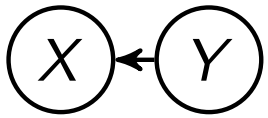
Causalité *pairwise*

- On observe uniquement deux variables X et Y.
- Si $X \not\perp\!\!\!\perp Y$, alors X et Y correspondent à des sommets adjacents dans le graphe causal composé de ces deux variables.
- Si de plus, on fait l'hypothèse de *causal sufficiency* (cf. cours précédent), alors on a forcément $X \to Y$ ou $Y \to X$.
- Mais comment retrouver ce sens causal si on ne peut pas identifier de v-structures dans ce graphe (en effet il faut au moins 3 variables pour identifier des v-structures) ?

Squelette: 

2 graphes Markov équivalents possibles:





Exemple

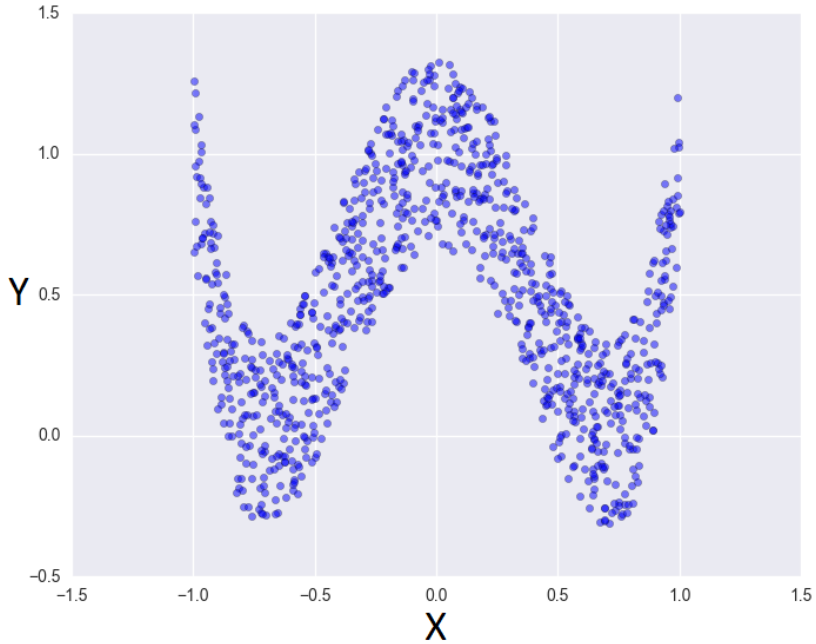


Figure 1

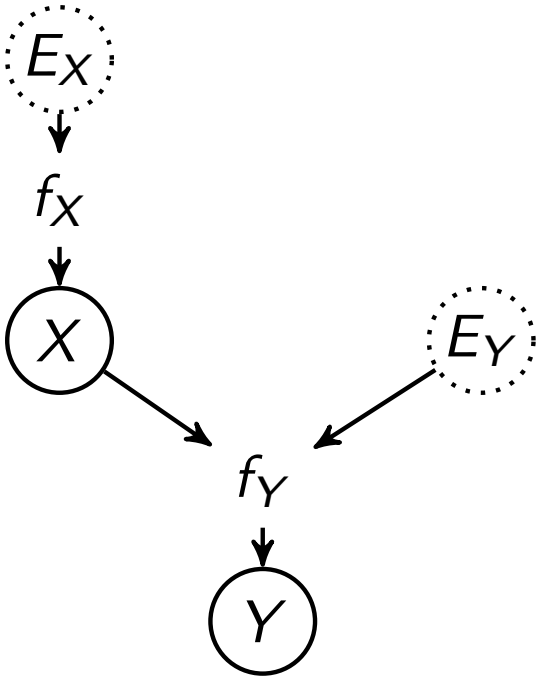
- Coefficient de corrélation égal à 0 !
- Mais dépendance forte mesurée avec un calcul d'information mutuelle.
- Plutôt facile de retrouver la direction causale dans ce cas.
- A votre avis ?

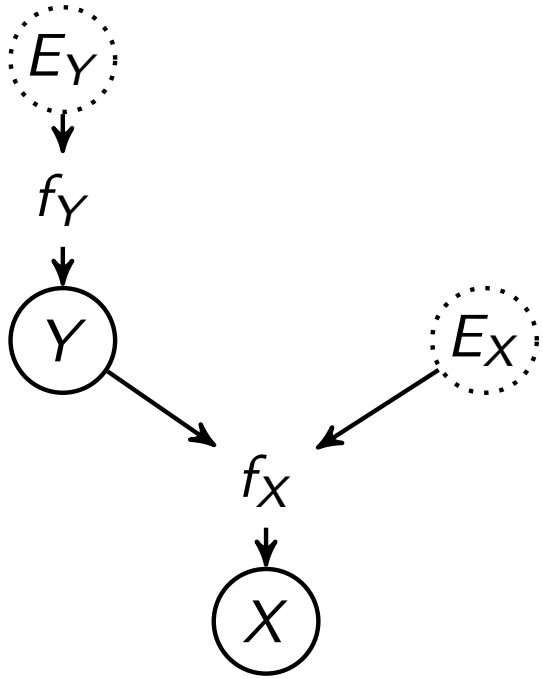
Une réponse possible

- Réponse plausible : $X \to Y$.
- En effet, les données ont été générées de X vers Y avec ce modèle stochastique :

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

Modèles fonctionnels causaux possibles dans le cas bivarié





$$\begin{cases} X = f_X(E_X) \\ Y = f_Y(X, E_Y) \\ \text{avec } E_X \perp\!\!\!\perp E_Y \end{cases}$$

$$\begin{cases} Y = f_Y(E_Y) \\ X = f_X(Y, E_X) \\ \text{avec } E_X \perp\!\!\!\perp E_Y \end{cases}$$

Figure 2: Deux hypothèses possibles de modèle fonctionnel causal dans le cas bivarié (si on exclue le cas des effets confondants cachés).

Causalité *pairwise*

- Des méthodes spécifiques pour la causalité dites *pairwise* ont été développées dans la littérature.
- "School of Tuebingen" (Hoyer et al., 2009; Janzing and Schölkopf, 2010)
- Voir le livre *Cause Effect Pairs in Machine Learning* (Guyon et al., 2019).

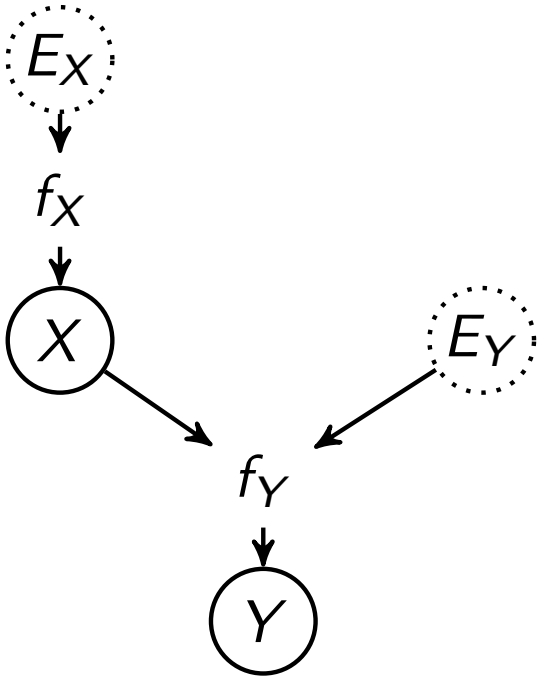
Différents type de méthode pour la causalité *pairwise*.

- Méthodes qui exploitent l'idée que la variable de bruit E_Y doit être **indépendante de la cause** X dans un modèle $Y = f(X, E_Y)$ si $X \to Y$.
- **Méthodes génératives** : si X cause Y, il est a priori plus simple de générer des données dans le sens $X \to Y$, plutôt que $Y \to X$.
- **Méthodes d'apprentissage supervisé**. Étant donné un ensemble de paires $\{X_i, Y_i\}$ observés et étiquetées, il s'agit de construire un modèle qui prédit si $X \to Y$ ou $Y \to X$ pour de nouvelles paires observées. On se ramène à un problème de **classification** dans ce cas.

Section 2

Méthodes qui testent l'indépendance entre le bruit et la cause

Il y a en fait une v-structure implicite



$$\begin{cases} X = f_X(E_X) \\ Y = f_Y(X, E_Y) \\ \text{avec } E_X \perp\!\!\!\perp E_Y \end{cases}$$

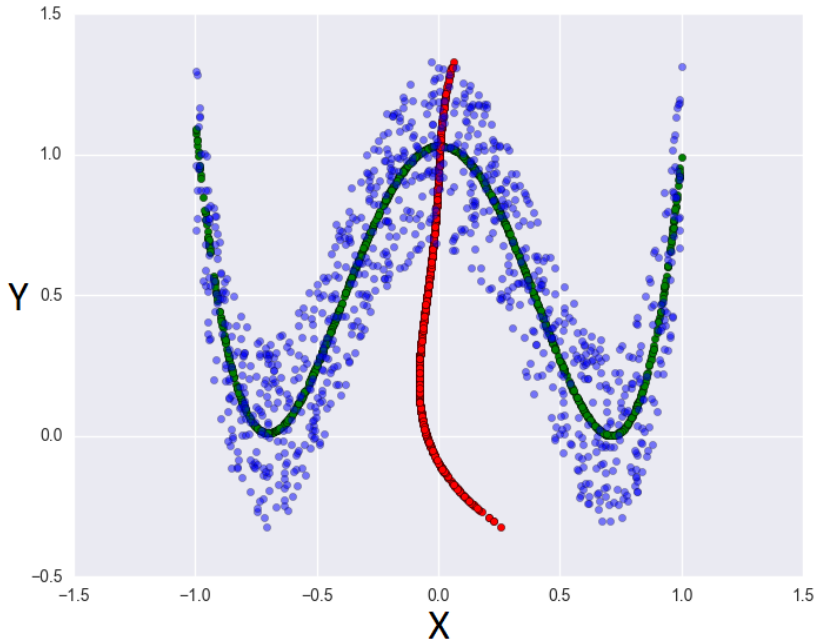
- Comme $E_X \perp\!\!\!\perp E_Y$, on doit avoir en fait aussi $X \perp\!\!\!\perp E_Y$.
- Il y a une "v-structure" implicite $X \rightarrow Y \leftarrow E_Y$.

Tester l'indépendance entre la cause et le bruit

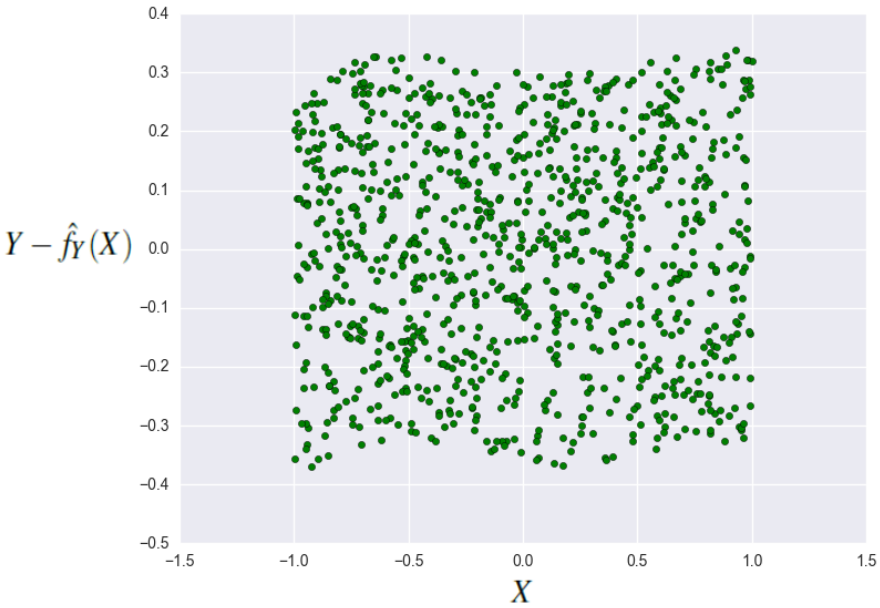
- **Idée :** apprendre des modèles $\hat{f_X}$ et $\hat{f_Y}$ en cherchant parmi des classes simples de fonctions, avec des techniques de régression.
- Deux modèles à fitter à partir des données $Y = \hat{f_Y}(X, E_Y)$ et $X = \hat{f_X}(Y, E_X)$.
- Ensuite pour chacun des deux modèles, effectuer un test d'indépendance statistique entre la cause et le résidu de la régression.

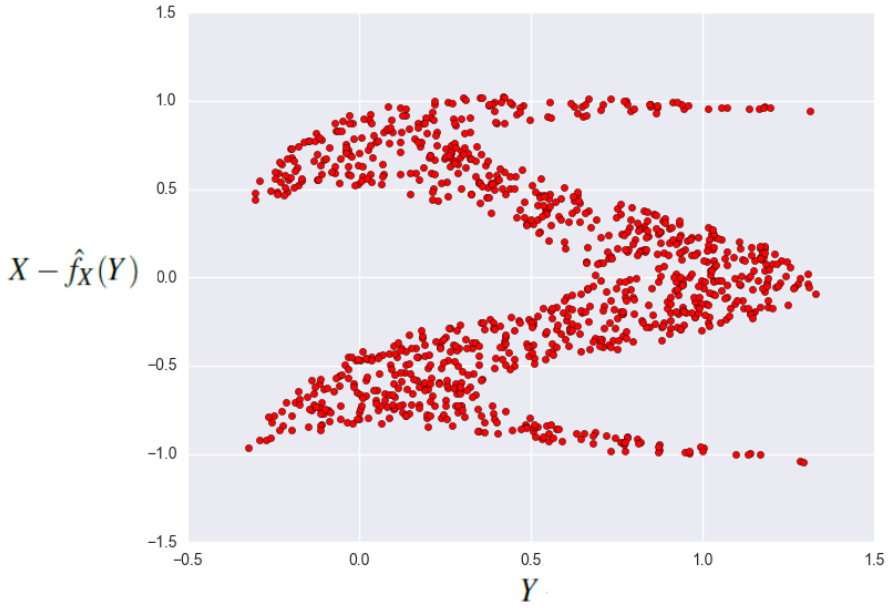
Exemple

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



Résidus de la régression



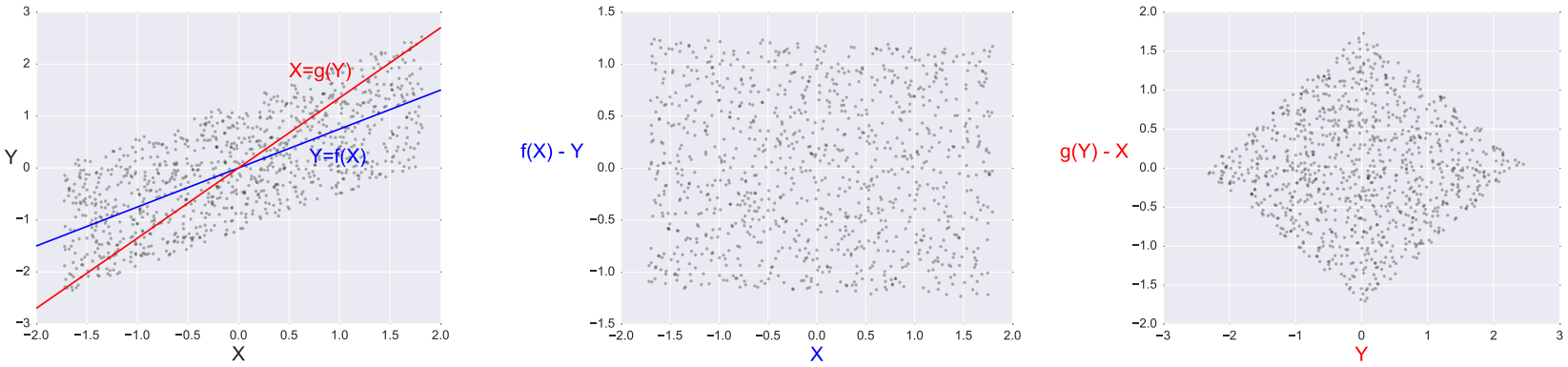


- Comme $(Y - \hat{f}_Y(X)) \perp\!\!\!\perp X$
  $\rightarrow$ **modèle explicatif simple** : $Y := \hat{f}_Y(X) + E_Y$ avec $E_Y \perp\!\!\!\perp X$ (Additive noise model (Hoyer et al., 2009))
- Comme $(X - \hat{f}_X(Y)) \not\perp\!\!\!\perp Y$
  $\rightarrow$ Il n'est pas possible d'ajuster un modèle $X := \hat{f}_X(Y) + E_X$ avec $E_X \perp\!\!\!\perp Y$.
- Un modèle explicatif $X := \hat{f}_X(Y, E_X)$ avec $E_X \perp\!\!\!\perp Y$ 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)

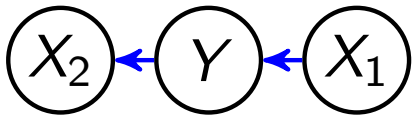


- Pour ce modèle fonctionnel causal avec trois variables :

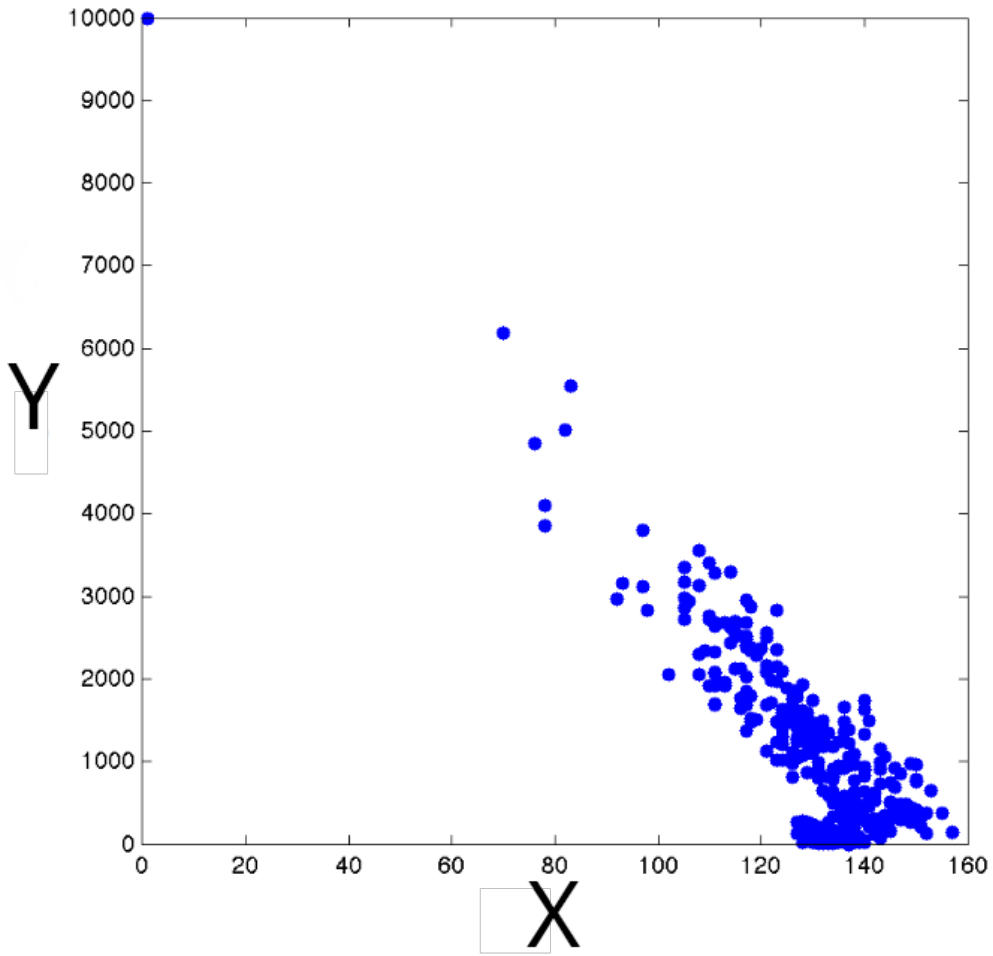
$$Y \leftarrow 0.5X_1 + E_{X_1},$$
$$X_2 \leftarrow Y + E_{X_2},$$

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



Réponse

- C'est $Y \to X$.
- Variable X : Température T mesurée dans des stations de montagne en Allemagne.
- Variable Y : Altitude Z de ces stations.
- C'est l'altitude (Z) qui cause la température (T) et pas l'inverse.

Application de la méthode ANM dans le sens $Z \to T$

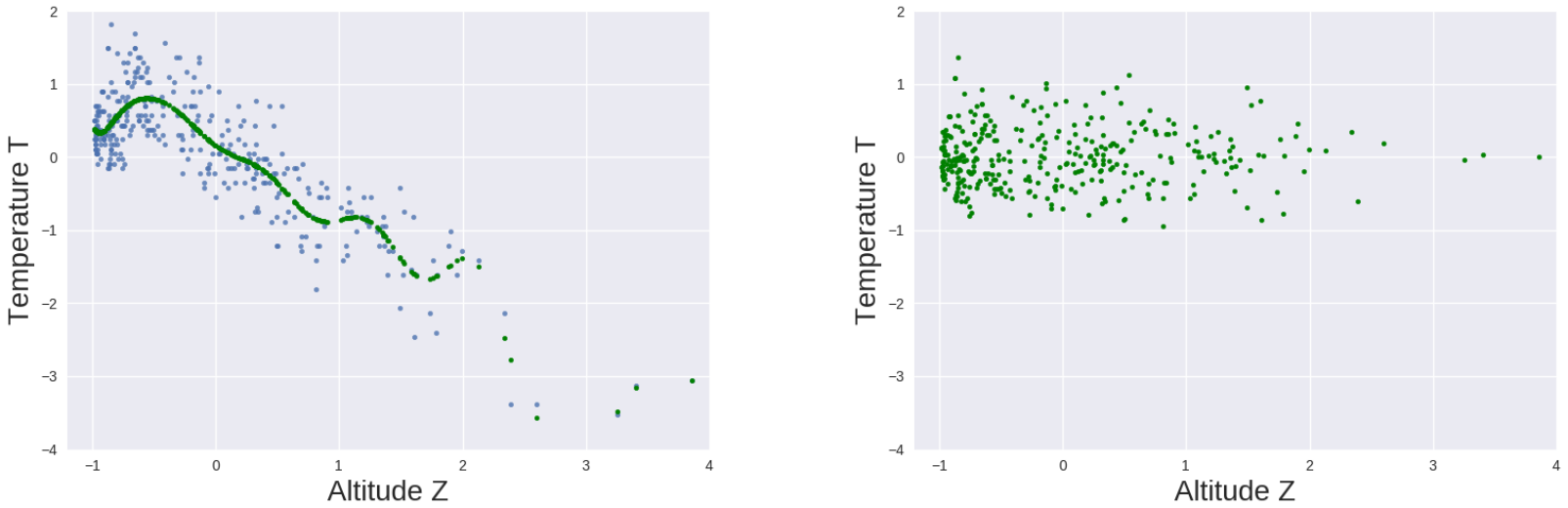


Figure 3: l'hypothèse causale $T := \hat{f}_T(Z) + E_T$ avec $E_T \perp\!\!\!\perp Z$ peut être valable.

Application de la méthode ANM dans le sens $T \to Z$

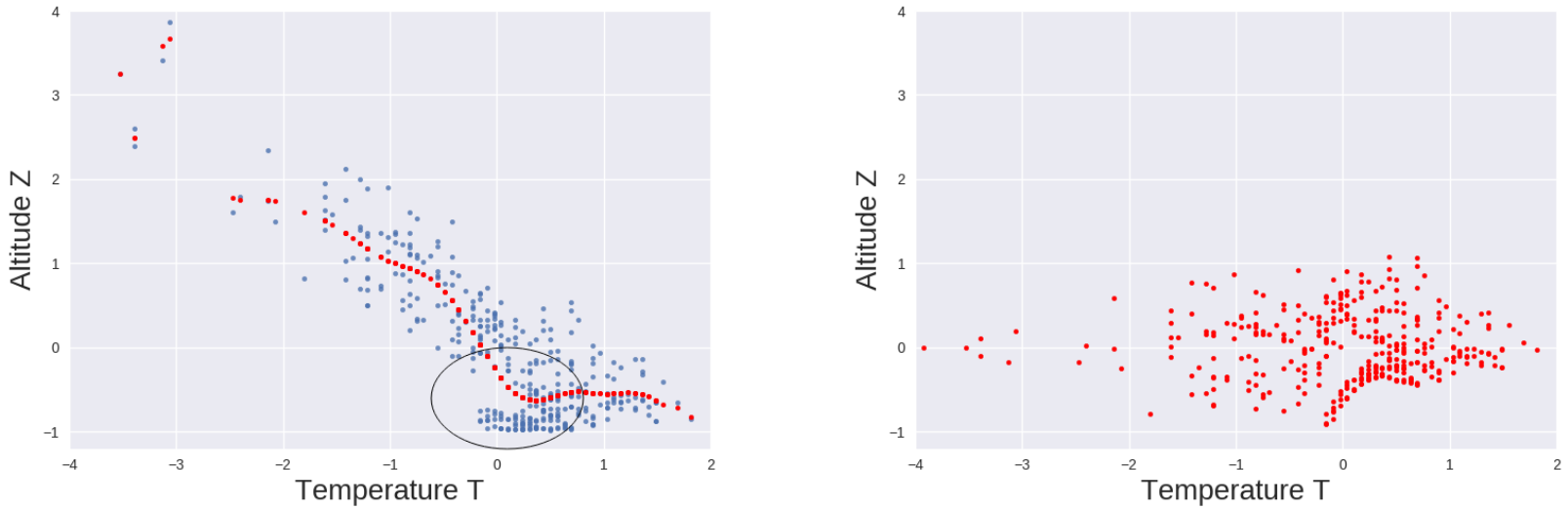
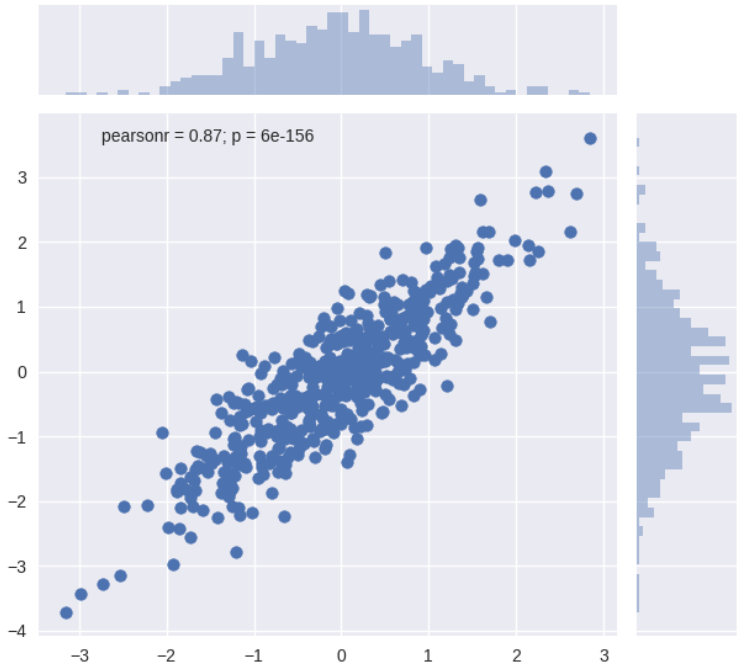


Figure 4: Dans une certaine mesure, ce résidu n'est pas indépendant de la température T. Par conséquent, l'hypothèse causale $Z := \hat{f}_Z(T) + E_Z$ avec $E_Z \perp\!\!\!\perp T$ ne tient pas.

Limitations de cette méthode ANM - cas non identifiable

- Le cas linéaire gaussien est non identifiable :



- Il a été montré dans (Hoyer et al., 2009) que c'est d'ailleurs le seul cas non identifiable, c'est à dire pour lequel on peut ajuster un modèle ANM dans les deux sens.

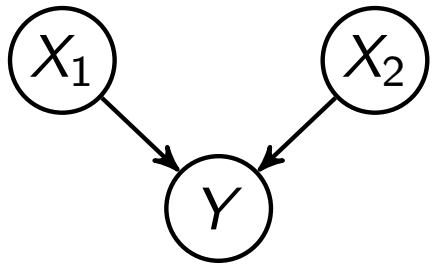
Problème de généralité

- **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.
- 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 *pairwise* pour orienter un graphe

- La méthode ANM ne prend 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 peut être identifiée, car on peut vérifier facilement dans ce cas avec un test de corrélation partielle que $X_1 \not\perp\!\!\!\perp X_2 | Y$.

Section 3

Méthodes génératives

Score de log-vraisemblance global

- Ici on va noter les variables X_1 et X_2 au lieu de X et Y (cf. notations cours 2).
- Il y a seulement deux graphes candidats $\hat{\mathcal{G}}$ correspondant à $X_1 \to X_2$ ou $X_2 \to X_1$.
- On note $\tau = (\tau_1, \tau_2)$ l'ensemble des paramètres des 2 Markov kernels à apprendre conjointement.
- Pour notre échantillon de données $\mathcal{D}$ supposées iid, pour chaque DAG candidat $\hat{\mathcal{G}}$, en utilisant la propriété de factorisation de Markov, la log vraisemblance négative pénalisée que l'on souhaite minimiser est

$$\mathcal{L}^n(\widehat{\mathcal{G}}, \tau, D) = -\frac{1}{n}\sum_{\ell=1}^{n}\sum_{j=1}^{2} \log q_j(x_j^{(\ell)} | x_{\mathrm{Pa}(j;\hat{\mathcal{G}})}^{(\ell)}, \tau_j) + \lambda_S |\widehat{\mathcal{G}}|, \qquad (4)$$

avec λ_S un hyperparamètre de la méthode et $|\widehat{\mathcal{G}}| = 1$ car il y a toujours un seul arc dans les graphes candidats.

Décomposition de la log-vraisemblance

On peut décomposer la log vraisemblance de chaque Markov kernel de la façon suivant (cf. cours précédents) :

$$\log q_j(x_j^{(\ell)}|x_{\mathrm{Pa}(j;\hat{\mathcal{G}})}^{(\ell)}, \tau_j) = \log \frac{q_j(x_j^{(\ell)}|x_{\mathrm{Pa}(j;\hat{\mathcal{G}})}^{(\ell)}, \tau_j)}{p(x_j^{(\ell)}|x_{\mathrm{Pa}(j;\hat{\mathcal{G}})}^{(\ell)})} - \hat{I}^n(X_j, X_{\overline{\mathrm{Pa}}(j;\hat{\mathcal{G}})}|X_{\mathrm{Pa}(j;\hat{\mathcal{G}})}) + \log p(x_j^{(\ell)}|x_{-j}^{(\ell)}) \quad (5)$$

Calcul des termes d'information mutuelle conditionnelle dans le cas de deux variables

- Pour le graphe $X_1 \to X_2$, on a

$$\sum_{j=1}^{d} \hat{I}^n(X_j, X_{\overline{\mathrm{Pa}}(j;\hat{\mathcal{G}})} | X_{\mathrm{Pa}(j;\hat{\mathcal{G}})}) = \hat{I}^n(X_1, X_{\overline{\mathrm{Pa}}(1;\hat{\mathcal{G}})} | X_{\mathrm{Pa}(1;\hat{\mathcal{G}})}) \quad (6)$$

$$+ \hat{I}^n(X_2, X_{\overline{\mathrm{Pa}}(2;\hat{\mathcal{G}})} | X_{\mathrm{Pa}(2;\hat{\mathcal{G}})}) \quad (7)$$

$$= \hat{I}^n(X_1, X_2) \quad (8)$$

- Pour le graphe $X_2 \to X_1$, on a de même : $\sum_{j=1}^{d} \hat{I}^n(X_j, X_{\overline{\mathrm{Pa}}(j;\hat{\mathcal{G}})} | X_{\mathrm{Pa}(j;\hat{\mathcal{G}})}) = \hat{I}^n(X_2, X_1)$
- Donc les termes en $\sum_{j=1}^{d} \hat{I}^n(X_j, X_{\overline{\mathrm{Pa}}(j;\hat{\mathcal{G}})} | X_{\mathrm{Pa}(j;\hat{\mathcal{G}})})$ peuvent être supprimés car ils égaux pour les deux graphes candidats.

Sélection de modèle dans le cas bivarié

- Pour faire la selection de modèle, il ne reste plus que les termes fonctionnels.
- On doit trouver le graphe qui maximise :

$$\mathcal{L}^n(\widehat{\mathcal{G}}, \tau, D) \sim \frac{1}{n} \sum_{\ell=1}^{n} \sum_{j=1}^{2} \log \frac{q_j(x_j^{(\ell)} | x_{\mathrm{Pa}(j;\hat{\mathcal{G}})}^{(\ell)}, \tau_j)}{p(x_j^{(\ell)} | x_{\mathrm{Pa}(j;\hat{\mathcal{G}})}^{(\ell)})} \tag{9}$$

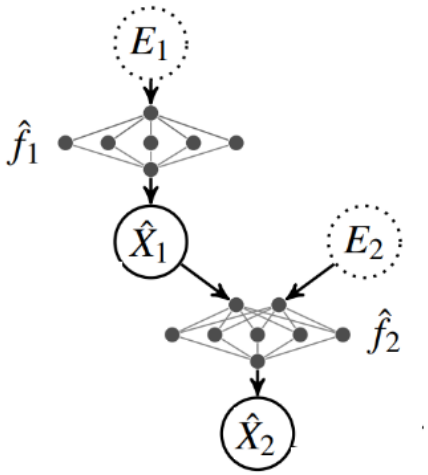
- Ce qui revient juste à comparer les score des deux modèles de régression dans chaque sens et voir lequel est le plus grand.

$$C_{X_1 \to X_2} = \frac{1}{n} \sum_{\ell=1}^{n} [\log\, q_1(x_1^{(\ell)}, \tau_1^*) + \log\, q_2(x_2^{(\ell)} | x_1^{(\ell)}, \tau_2^*)] \tag{10}$$

$$C_{X_2 \to X_1} = \frac{1}{n} \sum_{\ell=1}^{n} [\log\, q_2(x_2^{(\ell)}, \tau_2^*) + \log\, q_1(x_1^{(\ell)} | x_2^{(\ell)}, \tau_1^*)] \tag{11}$$

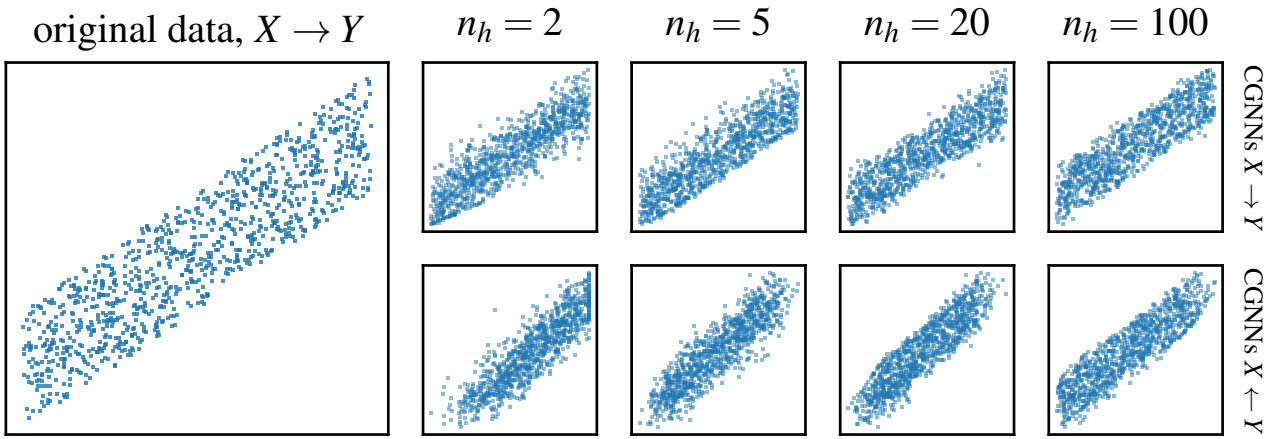
Modèle CGNN dans le cas de deux variables (Goudet et al., 2017)

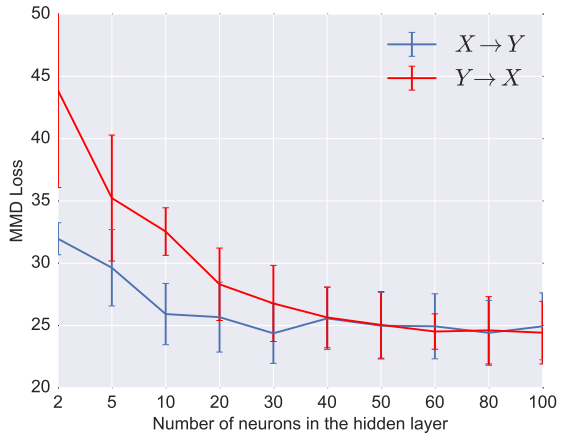
- $q1$ et $q2$ sont des réseaux de neurones génératifs (cf. cours précédent).
- n_h neurones dans chaque couche cachée.
- On apprend un modèle dans les deux sens et on garde celui avec le meilleur score.
- Dans CGNN, il s'agissait d'un score de *Maximum Mean Discrepancy* (MMD).



$$\hat{X}_1 = \hat{f}_1(E_1)$$
$$\hat{X}_2 = \hat{f}_2(\hat{X}_1, E_2)$$

Expériences avec CGNN - compromis complexité/reproduction des données

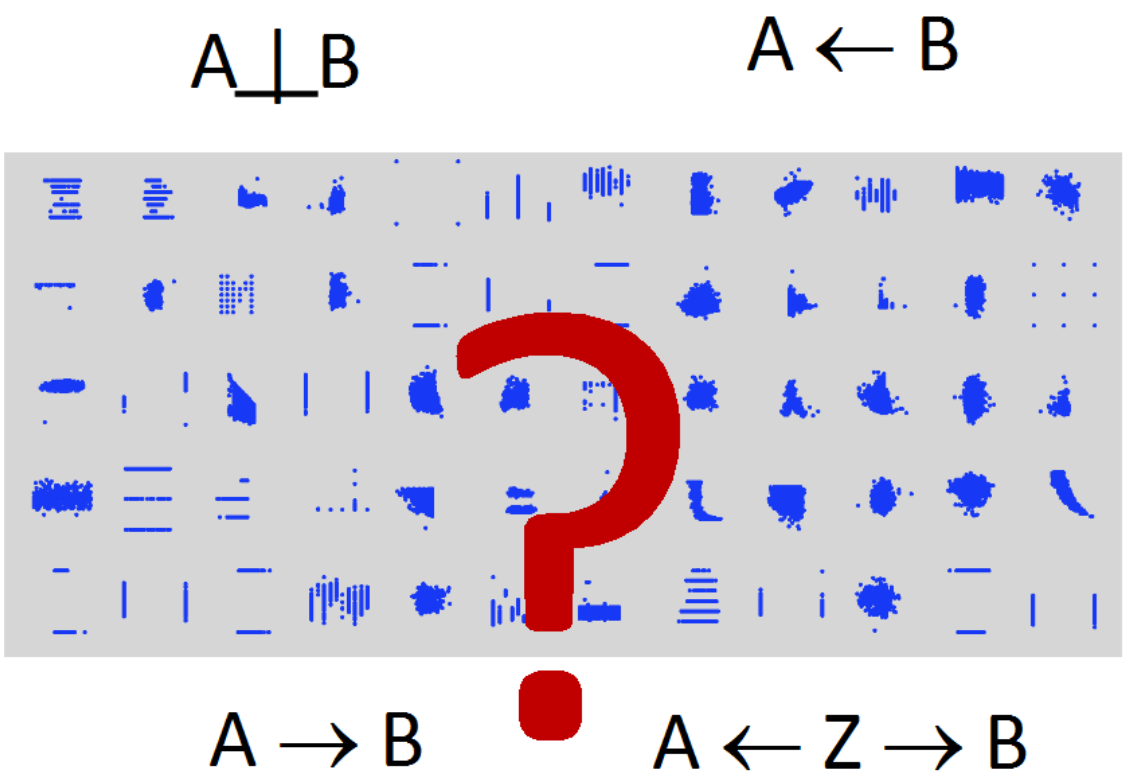




Section 4

Méthodes discriminantes avec des classifieurs

Cause effect pair challenge [Guyon 2013]



Inférence de la causalité comme un problème d'apprentissage supervisé

- Ensemble de données d'entraînement $((d_1, g_1), (d_2, g_2), \ldots, (d_n, g_n))$.
- Chaque entrée d_j est elle-même un ensemble de données de paires $(x_{1j}, y_{1j}), \ldots, (x_{pj}, y_{pj})$.
- C'est **une distribution de distribution** ! *Mother distribution*.
- L'étiquette g_i représente le mécanisme causal à l'œuvre dans l'ensemble de données d_i.
- Quatres classes différentes dans ce challenge :
  1. $X \rightarrow Y$
  2. $X \leftarrow Y$
  3. $X \leftarrow Z \rightarrow Y$
  4. $X \perp\!\!\!\perp Y$

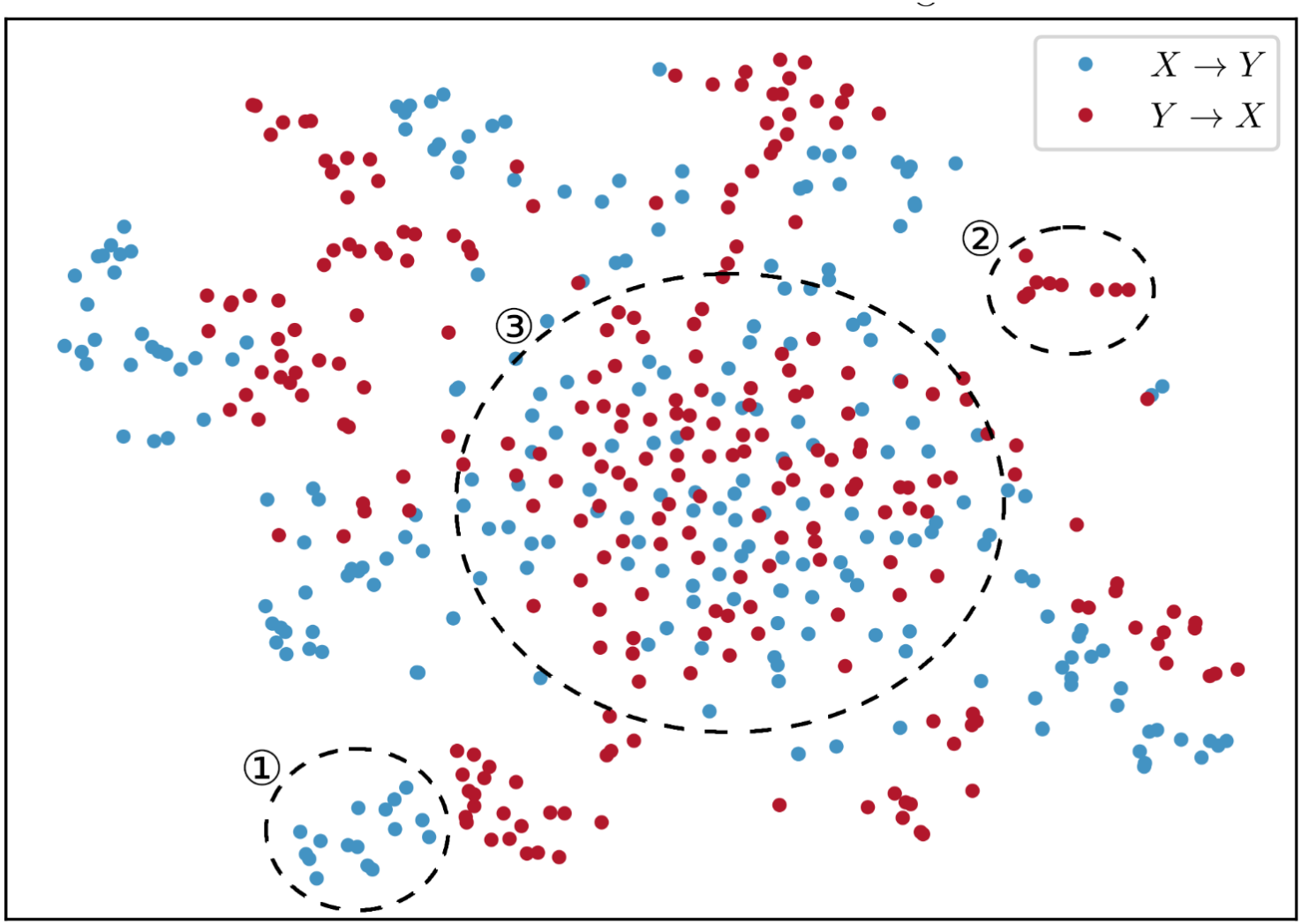
Méthode avec des kernels Lopez-Paz et al. (2015)

- Dataset d'entraînement avec les paires observées : $S = \{(x_{ij}, y_{ij})_{j=1}^{n_i}\}_{i=1}^{n}$.
- Projection de chaque paire observée $S_j = (x_{ij}, y_{ij})_{j=1}^{n_i}$ dans $\mathbb{R}^m$:

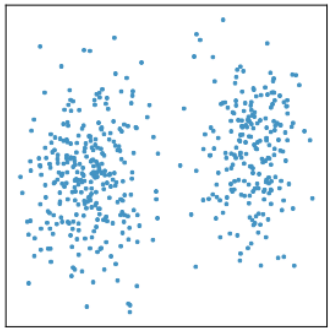
$$\mu_{k,m}(P_{S_j}) = \frac{2C_k}{|S|} \sum_{x_{ij}, y_{ij} \in S_j} (\cos(w_j^x * x_{ij} + w_j^y * y_{ij} + b_j))_{j=1}^{m} \in \mathbb{R}^m \quad (12)$$

- Puis entraînement d'un classifieur classique $cl : \mathbb{R}^m \to \{0, 1, 2, 3\}$ pour prédire l'une des quatres classes associée à la paire observée.

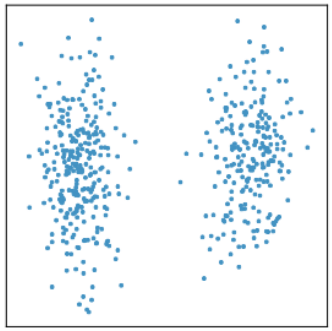
Projection en deux dimensions des paires observées



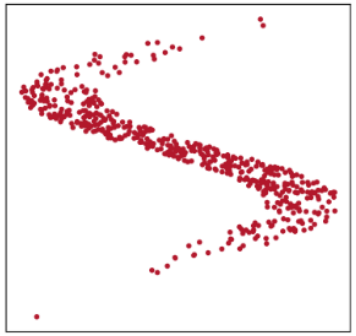
Classification des paires observées



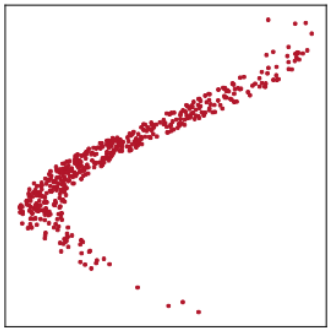
(1) $X \to Y$ pair in (1)



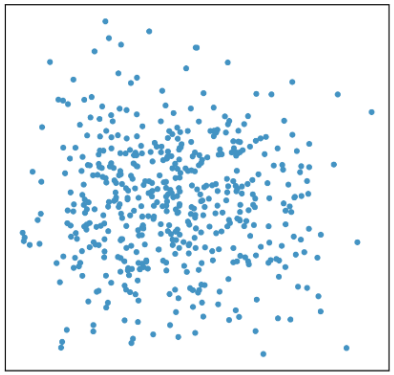
(2) $X \to Y$ pair in (1)



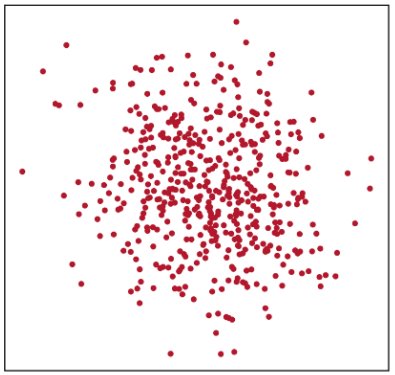
(3) $Y \to X$ pair in (2)



(4) $Y \to X$ pair in (2)



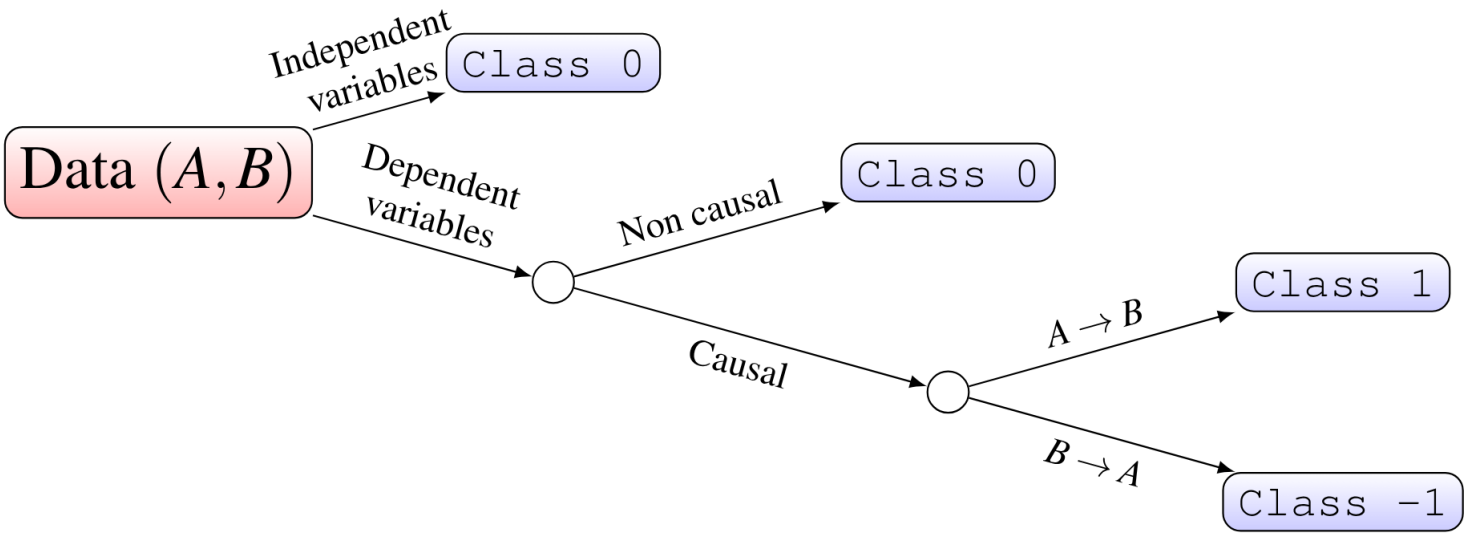
(1) $X \to Y$ pair in (3)



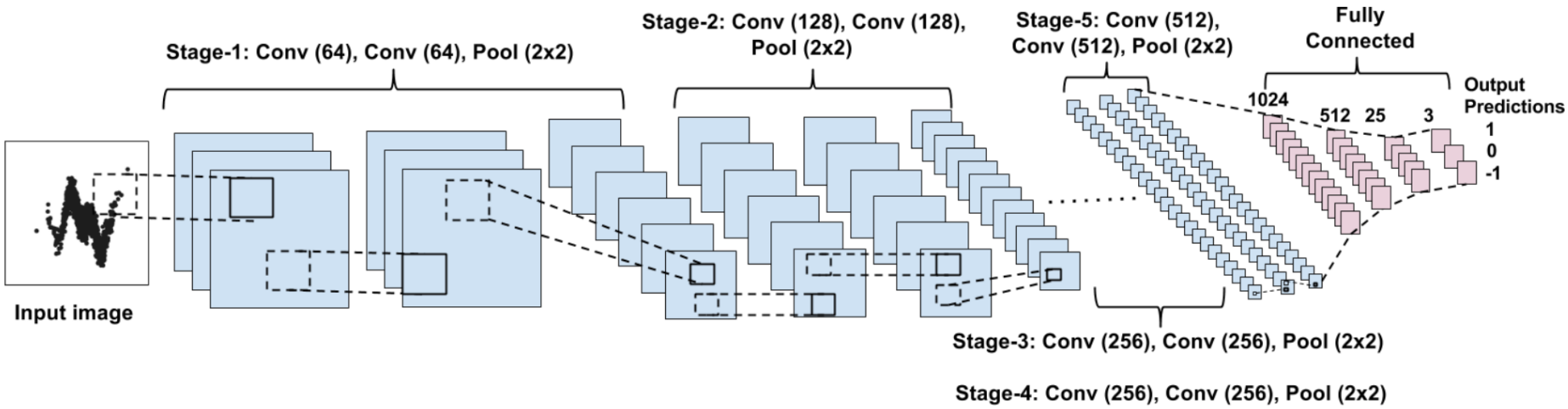
(2) $Y \to X$ pair in (3)

Méthodes avec des arbres de décision

- Modèle de Fonollosa (2019) qui a gagné le challenge sur la causalité *pairwise*.
- Exemple d'arbre de décisions simplifié :



Méthodes avec des CNN (Singh et al., 2017)



Section 5

Références

Fonollosa, J. A. (2019). Conditional distribution variability measures for causality detection. In *Cause Effect Pairs in Machine Learning*, pages 339–347. Springer.

Goudet, O., Kalainathan, D., Caillou, P., Lopez-Paz, D., Guyon, I., Sebag, M., Tritas, A., and Tubaro, P. (2017). Learning functional causal models with generative neural networks. *arXiv preprint arXiv:1709.05321*.

Guyon, I., Statnikov, A., and Batu, B. B. (2019). *Cause effect pairs in machine learning*. Springer.

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.

Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. O. (2015). Towards a learning theory of cause-effect inference. In *ICML*, pages 1452–1461.

Singh, K., Gupta, G., Vig, L., Shroff, G., and Agarwal, P. (2017). Deep convolutional neural networks for pairwise causality. *arXiv preprint arXiv:1701.00597*.

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., 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.