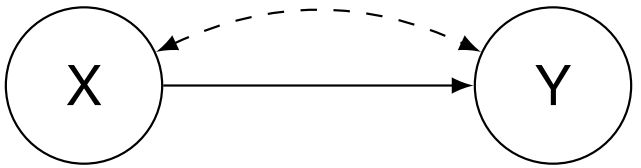
Soutenance
Habilitation à Diriger des Recherches

Apprentissage statistique
pour des données structurées en grande dimension

Emilie Devijver

CNRS, Université Grenoble Alpes, France

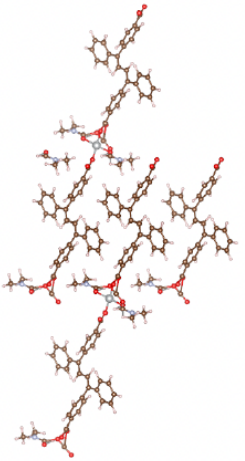
10 décembre 2024

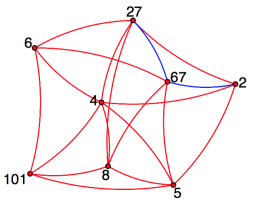


Motivation

Explorer et résoudre des **défis méthodologiques** en **apprentissage statistique**, guidé par des **applications concrètes**.

- Comment construire les données d'apprentissage ?
- Comment quantifier l'incertitude dans les données ?
- Comment modéliser l'interdépendance dans les données ?
- Comment limiter l'instabilité des méthodes ?
- Comment déterminer les structures causales sous-jacentes d'un processus ?





Thèmes de recherche

1. Méthodes pour la régression en grande dimension
2. Inférence de réseaux avec des modèles graphiques gaussiens
3. Inférence causale pour les séries temporelles
4. Apprentissage semi-supervisé
5. Application en science des matériaux

Thèmes de recherche

1. Méthodes pour la régression en grande dimension
2. Inférence de réseaux avec des modèles graphiques gaussiens
3. Inférence causale pour les séries temporelles
4. Apprentissage semi-supervisé
5. Application en science des matériaux

Cette présentation :

Etude des données interdépendantes avec les modèles graphiques

Modèles graphiques

Données : variables aléatoires $Y_1, \ldots, Y_p$.
Souvent ces variables sont dépendantes : on considère globalement

Y vecteur de taille p

On cherche à construire un graphe $G = (V, E)$ où

- Les noeuds $V = \{1, \ldots, p\} \Leftrightarrow$ variables aléatoires
- Les arêtes $E \Leftrightarrow$ dépendances **directes et significatives** entre les variables

Outils :

- **Partie 1.** Modèles graphiques non orientés
- **Partie 2.** Modèles graphiques orientés

Modèles graphiques

Données : variables aléatoires $Y_1, \ldots, Y_p$.
Souvent ces variables sont dépendantes : on considère globalement

$$\mathbf{Y} \text{ vecteur de taille } p$$

On cherche à construire un graphe $G = (V, E)$ où

- Les noeuds $V = \{1, \ldots, p\} \Leftrightarrow$ variables aléatoires
- Les arêtes $E \Leftrightarrow$ dépendances **directes et significatives** entre les variables

Outils :

- **Partie 1.** Modèles graphiques non orientés
- **Partie 2.** Modèles graphiques orientés

Modèles graphiques

Données : variables aléatoires $Y_1, \ldots, Y_p$.
Souvent ces variables sont dépendantes : on considère globalement

$$\mathbf{Y} \text{ vecteur de taille } p$$

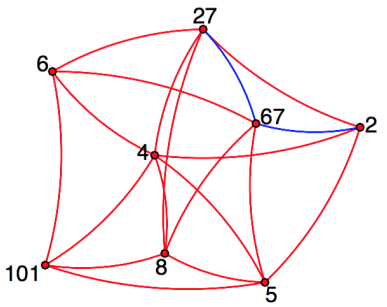
On cherche à construire un graphe $G = (V, E)$ où

- Les noeuds $V = \{1, \ldots, p\} \Leftrightarrow$ variables aléatoires
- Les arêtes $E \Leftrightarrow$ dépendances **directes et significatives** entre les variables

Outils :

- **Partie 1.** Modèles graphiques non orientés
- **Partie 2.** Modèles graphiques orientés

Partie 1.
MODÈLE GRAPHIQUE GAUSSIEN



Modèle graphique gaussien

Modèle

$$\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$$

$$\text{Arête dans le graphe} \Leftrightarrow \Theta_{j_1,j_2} = \Sigma^{-1}_{j_1,j_2} \neq 0$$

Données : $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ iid
But : estimer Θ
Estimateur du Maximum de vraisemblance

$$\widehat{\Theta} := \underset{\Theta}{\operatorname{argmin}} \{\log\det(\Theta) - \operatorname{tr}(S\Theta)\}$$

où S est la matrice de covariance empirique calculée sur $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$
Estimateur du Graphical Lasso [1] **:**

$$\widehat{\Theta}^{\mathrm{GL}}_{\lambda} := \underset{\Theta}{\operatorname{argmin}} \{\log\det(\Theta) - \operatorname{tr}(S\Theta) - \lambda||\Theta||_1\}$$

1. Friedman, Hastie, Tibshirani, 2008, Biostatistics

Modèle graphique gaussien

Modèle

$$\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$$

$$\text{Arête dans le graphe} \Leftrightarrow \Theta_{j_1,j_2} = \Sigma^{-1}_{j_1,j_2} \neq 0$$

Données : $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ iid

But : estimer Θ

Estimateur du Maximum de vraisemblance

$$\widehat{\Theta} := \underset{\Theta}{\text{argmin}} \{\log\det(\Theta) - \text{tr}(S\Theta)\}$$

où S est la matrice de covariance empirique calculée sur $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$

Estimateur du Graphical Lasso [1] :

$$\widehat{\Theta}^{\text{GL}}_{\lambda} := \underset{\Theta}{\text{argmin}} \{\log\det(\Theta) - \text{tr}(S\Theta) - \lambda||\Theta||_1\}$$

1. Friedman, Hastie, Tibshirani, 2008, Biostatistics

Modèle graphique gaussien

Modèle

$$\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$$

$$\text{Arête dans le graphe} \Leftrightarrow \Theta_{j_1,j_2} = \Sigma^{-1}_{j_1,j_2} \neq 0$$

Données : $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ iid

But : estimer Θ

Estimateur du Maximum de vraisemblance

$$\widehat{\Theta} := \underset{\Theta}{\text{argmin}} \left\{ \log\det(\Theta) - \text{tr}(S\Theta) \right\}$$

où S est la matrice de covariance empirique calculée sur $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$

Estimateur du Graphical Lasso [1] **:**

$$\widehat{\Theta}^{\text{GL}}_{\lambda} := \underset{\Theta}{\text{argmin}} \left\{ \log\det(\Theta) - \text{tr}(S\Theta) - \lambda \|\Theta\|_1 \right\}$$

1. Friedman, Hastie, Tibshirani, 2008, Biostatistics

Etat de l'art

Estimateur du Graphical Lasso [2] **:**

$$\widehat{\Theta}_{\lambda}^{GL} := \underset{\Theta}{\operatorname{argmin}} \{\log\det(\Theta) - \operatorname{tr}(S\Theta) - \lambda||\Theta||_1\}$$

⚠ Problème complexe d'optimisation

Implémentation efficace de $\widehat{\Theta}_{\lambda}^{GL}$: décomposition en 2 étapes [3]

1. Seuillage de $|S|$ au niveau $\lambda \Rightarrow$ détection d'une structure par blocs
   $\Leftrightarrow$ classification hiérarchique avec lien simple [4]
   💡 Remplacer le lien simple par le lien moyen et sélectionner K par validation croisée
2. Graphical Lasso de paramètre de régularisation λ dans chaque bloc

2. Friedman, Hastie, Tibshirani, 2008, Biostatistics
3. Mazumder et Hastie, 2012, JMLR
4. Tan, Witten et Shojaie, 2015, CSDA

Etat de l'art

Estimateur du Graphical Lasso [2] **:**

$$\widehat{\Theta}_{\lambda}^{GL} := \underset{\Theta}{\operatorname{argmin}} \{\log\det(\Theta) - \operatorname{tr}(S\Theta) - \lambda||\Theta||_1\}$$

⚠ Problème complexe d'optimisation

Implémentation efficace de $\widehat{\Theta}_{\lambda}^{GL}$: décomposition en 2 étapes [3]

1. Seuillage de $|S|$ au niveau $\lambda \Rightarrow$ détection d'une structure par blocs
   - $\Leftrightarrow$ classification hiérarchique avec lien simple [4]
   - 💡 Remplacer le lien simple par le lien moyen et sélectionner K par validation croisée
2. Graphical Lasso de paramètre de régularisation λ dans chaque bloc

2. Friedman, Hastie, Tibshirani, 2008, Biostatistics
3. Mazumder et Hastie, 2012, JMLR
4. Tan, Witten et Shojaie, 2015, CSDA

Etat de l'art

Estimateur du Graphical Lasso [2] **:**

$$\widehat{\Theta}_{\lambda}^{GL} := \underset{\Theta}{\operatorname{argmin}} \{\log\det(\Theta) - \operatorname{tr}(S\Theta) - \lambda||\Theta||_1\}$$

⚠ Problème complexe d'optimisation

Implémentation efficace de $\widehat{\Theta}_{\lambda}^{GL}$: décomposition en 2 étapes [3]

1. Seuillage de $|S|$ au niveau $\lambda \Rightarrow$ détection d'une structure par blocs
   - $\Leftrightarrow$ classification hiérarchique avec lien simple [4]
   - 💡 Remplacer le lien simple par le lien moyen et sélectionner K par validation croisée
2. Graphical Lasso de paramètre de régularisation λ dans chaque bloc

2. Friedman, Hastie, Tibshirani, 2008, Biostatistics
3. Mazumder et Hastie, 2012, JMLR
4. Tan, Witten et Shojaie, 2015, CSDA

Etat de l'art

Estimateur du Graphical Lasso [2] **:**

$$\widehat{\Theta}_{\lambda}^{GL} := \underset{\Theta}{\operatorname{argmin}} \{\log\det(\Theta) - \operatorname{tr}(S\Theta) - \lambda||\Theta||_1\}$$

⚠ Problème complexe d'optimisation

Implémentation efficace de $\widehat{\Theta}_{\lambda}^{GL}$: décomposition en 2 étapes [3]

1. Seuillage de $|S|$ au niveau $\lambda \Rightarrow$ détection d'une structure par blocs
   - $\Leftrightarrow$ classification hiérarchique avec lien simple [4]
   - 💡 Remplacer le lien simple par le lien moyen et sélectionner K par validation croisée
2. Graphical Lasso de paramètre de régularisation λ dans chaque bloc

2. Friedman, Hastie, Tibshirani, 2008, Biostatistics
3. Mazumder et Hastie, 2012, JMLR
4. Tan, Witten et Shojaie, 2015, CSDA

Shock [5]

Slope **h**euristic for bl**ock**-diagonal covariance structure detection for network inference

Soit $\mathbf{B} = (B_1, \ldots, B_K)$ la partition des variables.

$$F_{\mathbf{B}} = \left\{ f_{\mathbf{B}} = \phi_p(0, \Sigma_{\mathbf{B}}) \text{ avec } \Sigma_{\mathbf{B}} \in \mathbb{S}_p^{++}(\mathbb{R}) \;\middle|\; \Sigma_{\mathbf{B}} = P_\sigma \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_K \end{pmatrix} P_\sigma^{-1} \right\}$$

5. D. et Gallopin, 2018, JASA

Shock [5]

Slope **h**euristic for bl**ock**-diagonal covariance structure detection for network inference

Soit $\mathbf{B} = (B_1, \ldots, B_K)$ la partition des variables.

$$F_{\mathbf{B}} = \left\{ f_{\mathbf{B}} = \phi_p(0, \Sigma_{\mathbf{B}}) \text{ avec } \Sigma_{\mathbf{B}} \in \mathbb{S}_p^{++}(\mathbb{R}) \;\middle|\; \Sigma_{\mathbf{B}} = P_\sigma \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_K \end{pmatrix} P_\sigma^{-1} \right\}$$

$\mathcal{B}$: ensemble des partitions possibles avec p variables

Cardinal : nombre de Bell

par exemple, pour $p = 10$, $|\mathcal{B}| = 21147$

⚠ Exploration exhaustive de $\mathcal{B}$ impossible

5. D. et Gallopin, 2018, JASA

Shock [5]

Slope **h**euristic for bl**ock**-diagonal covariance structure detection for network inference

Soit $\mathbf{B} = (B_1, \ldots, B_K)$ la partition des variables.

$$F_{\mathbf{B}} = \left\{ f_{\mathbf{B}} = \phi_p(0, \Sigma_{\mathbf{B}}) \text{ avec } \Sigma_{\mathbf{B}} \in \mathbb{S}_p^{++}(\mathbb{R}) \;\middle|\; \Sigma_{\mathbf{B}} = P_\sigma \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_K \end{pmatrix} P_\sigma^{-1} \right\}$$

1. Détection de la structure par blocs
   - (1) Calcul de la matrice de covariance empirique S
   - (2) Construction des partitions $\mathcal{B}_\Lambda = (\mathbf{B}_\lambda)_{\lambda \in \Lambda}$ avec la classification hiérarchique avec lien simple
   - (3) Pour chaque partition $\mathbf{B} \in \mathcal{B}_\Lambda$, calculer l'estimateur du maximum de vraisemblance $\hat{f}_{\mathbf{B}}$.
   - (4) Sélectionner $\hat{\mathbf{B}}$ avec un critère de sélection de modèles

$$\hat{\mathbf{B}} = \underset{\mathbf{B} \in \mathcal{B}_\Lambda}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(\hat{f}_{\mathbf{B}}(\mathbf{y}_i)) + \operatorname{pen}(\mathbf{B}) \right\}$$

2. Graphical Lasso de paramètre de régularisation λ dans chaque bloc

5. D. et Gallopin, 2018, JASA

Shock en théorie

Procédure **adaptative minimax** pour apprendre la structure diagonale par blocs.

Théorème (D. et Gallopin, 2018, JASA)

Soit $c > 0$. Il existe $\kappa > 0$ et C_1 tels que, dès que

$$\mathrm{pen}(\mathbf{B}) \geq \kappa \frac{D_{\mathbf{B}}}{n}\left[2c^2 + \log\left(\frac{p^4}{D_{\mathbf{B}}(\frac{D_{\mathbf{B}}}{n}c^2 \wedge 1)}\right)\right],$$

alors

$$\hat{\mathbf{B}} = \underset{\mathbf{B}\in\mathcal{B}_\Lambda}{\mathrm{argmin}}\left\{-\frac{1}{n}\sum_{i=1}^{n}\log(\hat{f}_{\mathbf{B}}(\mathbf{y}_i)) + \mathrm{pen}(\mathbf{B})\right\}$$

vérifie une inégalité oraculaire et une borne inférieur minimax.

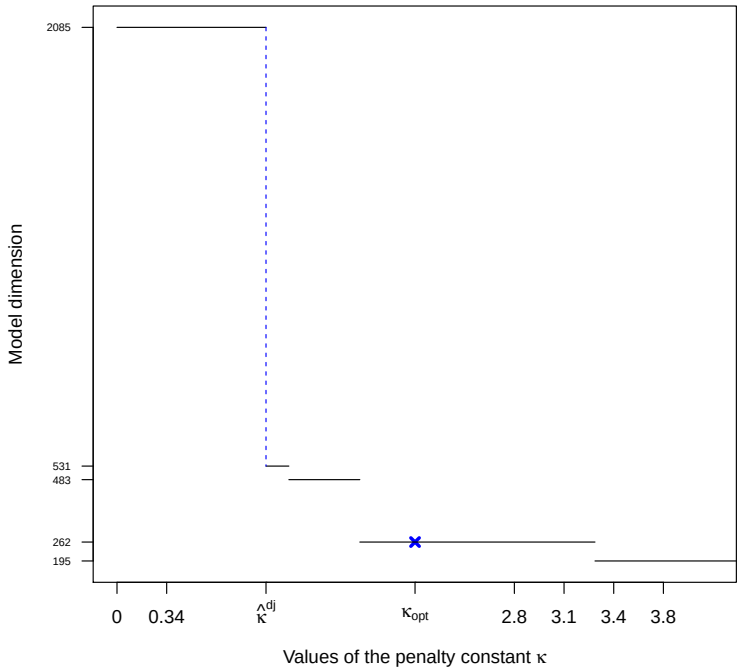
Shock en pratique

- On ne peut pas utiliser la pénalité théorique
- On simplifie la forme de la pénalité

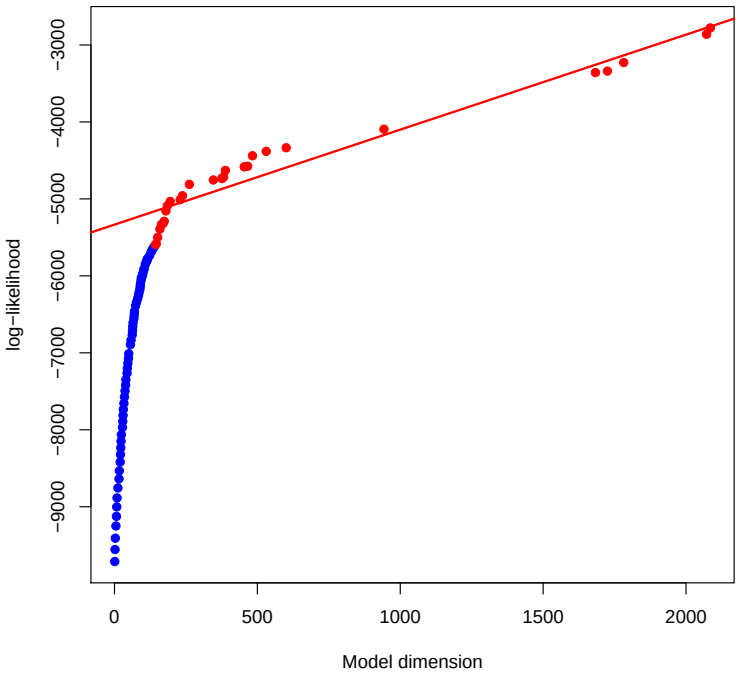
$$\mathrm{pen}(B) = \kappa D_B$$

- On calibre la constante κ à partir des données, en utilisant l'heuristique de pentes [6]

Méthode 1 : Saut de dimension



Méthode 2 : Régression robuste



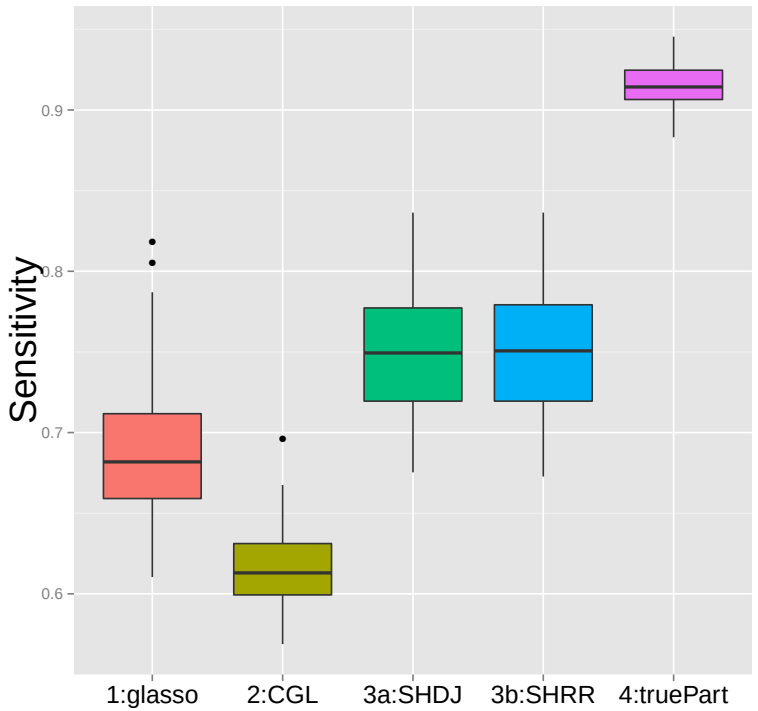
6. Birgé et Massart, 2007, PTRF

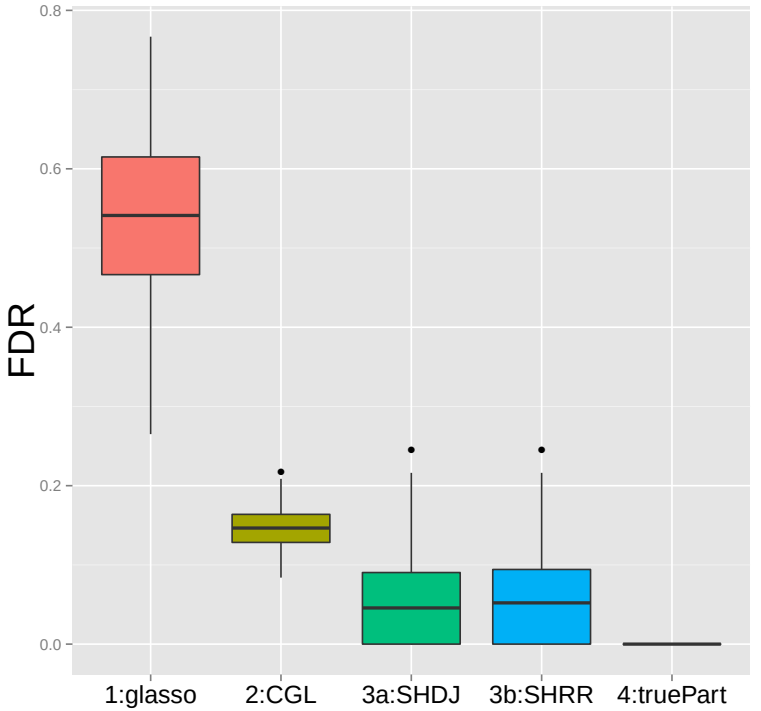
Performance sur les données simulées

Données simulées : $p = 100$, $n = 70$ et Σ est diagonale par blocs avec $K^\star = 15$.
Résultats sur 100 simulations.

$$\text{Sensitivité} = \frac{TP}{TP + FN}$$

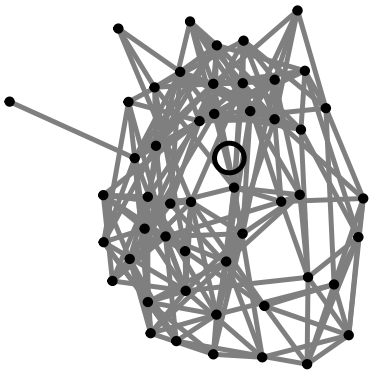
$$FDR = \frac{FP}{TP + FP}$$

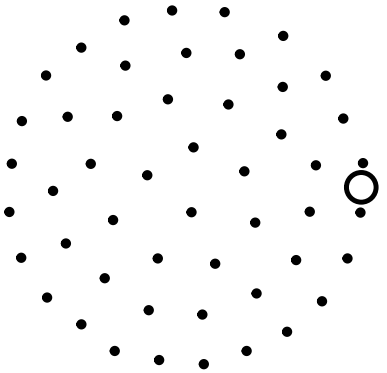


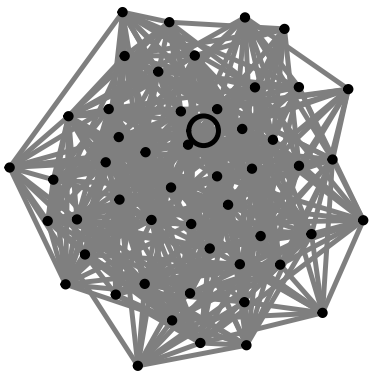


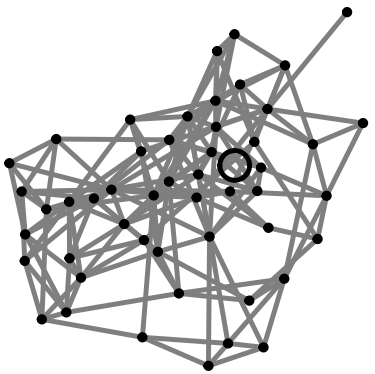
Stabilité de l'inférence de réseaux

4 jeux de données simulées - $n = 30, p = 50$









Etat de l'art

⚠ Comment rendre stable la sélection de variables ?
💡 En ré-échantillonnant !

- Bootstrap Lasso [7] , pour la régression linéaire pénalisée
- Sélection stable [8], en considérant le chemin de régularisation, et choix des hyperparamètres [9]
- StARS [10], sélectionne un modèle parcimonieux et stable
- ESCV [11] adapte la validation croisée pour assurer la stabilité

A-t-on besoin de ré-échantilloner pour stabiliser une méthode ?

- Mesure d'influence des observations [12]

7. Bach, 2008, ICML
8. Buhlmann et Meinshausen, 2010, JRSS B
9. Bodinier, Filippi, Nøst, Chiquet et Chadeau-Hyam, 2023, JRSS C
10. Liu, Roeder et Wasserman, 2010, NeurIPS
11. Lim et Yu, 2016, JCGS
12. Bar-Hen et Poggi, 2016, JMVA

Etat de l'art

⚠ Comment rendre stable la sélection de variables ?
💡 En ré-échantillonnant !

- Bootstrap Lasso [7] , pour la régression linéaire pénalisée
- Sélection stable [8], en considérant le chemin de régularisation, et choix des hyperparamètres [9]
- StARS [10], sélectionne un modèle parcimonieux et stable
- ESCV [11] adapte la validation croisée pour assurer la stabilité

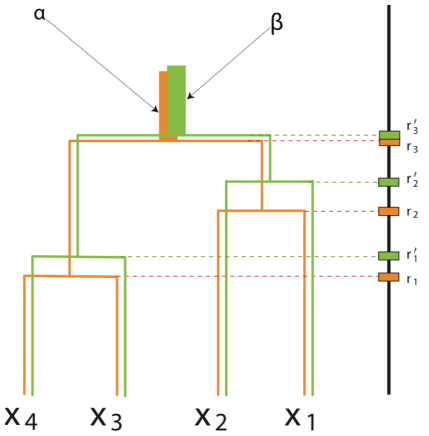
A-t-on besoin de ré-échantilloner pour stabiliser une méthode ?

- Mesure d'influence des observations [12]

7. Bach, 2008, ICML
8. Buhlmann et Meinshausen, 2010, JRSS B
9. Bodinier, Filippi, Nøst, Chiquet et Chadeau-Hyam, 2023, JRSS C
10. Liu, Roeder et Wasserman, 2010, NeurIPS
11. Lim et Yu, 2016, JCGS
12. Bar-Hen et Poggi, 2016, JMVA

Stabilité de shock [14]

La décomposition en deux étapes stabilise la méthode.
θ_{A_1} et θ_{A_2} deux dendrogrammes de lien simple basés sur les matrices de distance A_1 et A_2,
$d_{\text{coph}}(\theta_{A_1}, \theta_{A_2})$ la distance cophénétique.



Stabilité de la classification hiérarchique avec lien simple [13]

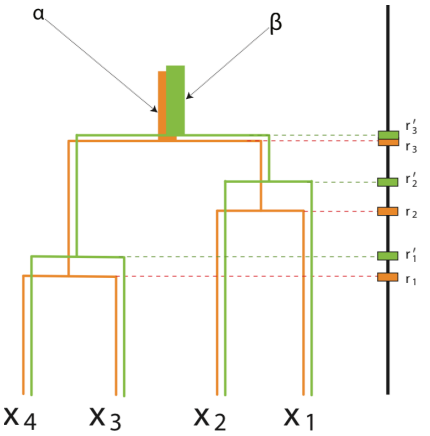
Théorème (D., Gallopin et Molinier, 2024+)

Soit deux échantillons $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ et $(\mathbf{y}_1, \ldots, \tilde{\mathbf{y}}_i, \ldots, \mathbf{y}_n)$, où $\tilde{\mathbf{y}}_i \sim \mathbf{y}_i \sim \mathbf{Y}$, et soit S et $\tilde{S}$ les matrices de covariance empiriques respectives. Alors, pour $\alpha \in (0,1)$, avec probabilité $1-\alpha$,

$$d_{coph}(\theta_{\mathbf{1}-|S|}, \theta_{\mathbf{1}-|\tilde{S}|}) \leq \frac{2p}{(n-1)\sqrt{\alpha}}.$$

13. Carlsson et Mémoli, 2010, JMLR
14. D., Gallopin et Molinier, 2024, arXiv

Stabilité de shock [14]



La décomposition en deux étapes stabilise la méthode.
θ_{A_1} et θ_{A_2} deux dendrogrammes de lien simple basés sur les matrices de distance A_1 et A_2,
$d_{\text{coph}}(\theta_{A_1}, \theta_{A_2})$ la distance cophénétique.

Stabilité de la classification hiérarchique avec lien simple [13]

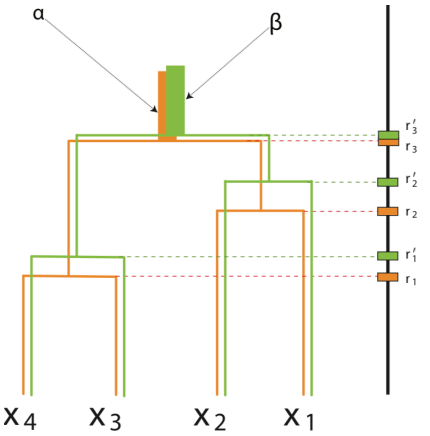
Théorème (D., Gallopin et Molinier, 2024+)

Soit deux échantillons $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ et $(\mathbf{y}_1, \ldots, \tilde{\mathbf{y}}_i, \ldots, \mathbf{y}_n)$, où $\tilde{\mathbf{y}}_i \sim \mathbf{y}_i \sim \mathbf{Y}$, et soit S et $\tilde{S}$ les matrices de covariance empiriques respectives. Alors, pour $\alpha \in (0,1)$, avec probabilité $1-\alpha$,

$$d_{coph}(\theta_{\mathbf{1}-|S|}, \theta_{\mathbf{1}-|\tilde{S}|}) \leq \frac{2p}{(n-1)\sqrt{\alpha}}.$$

13. Carlsson et Mémoli, 2010, JMLR
14. D., Gallopin et Molinier, 2024, arXiv

Stabilité de shock [14]



La décomposition en deux étapes stabilise la méthode.
θ_{A_1} et θ_{A_2} deux dendrogrammes de lien simple basés sur les matrices de distance A_1 et A_2,
$d_{\text{coph}}(\theta_{A_1}, \theta_{A_2})$ la distance cophénétique.

Stabilité de la classification hiérarchique avec lien simple [13]

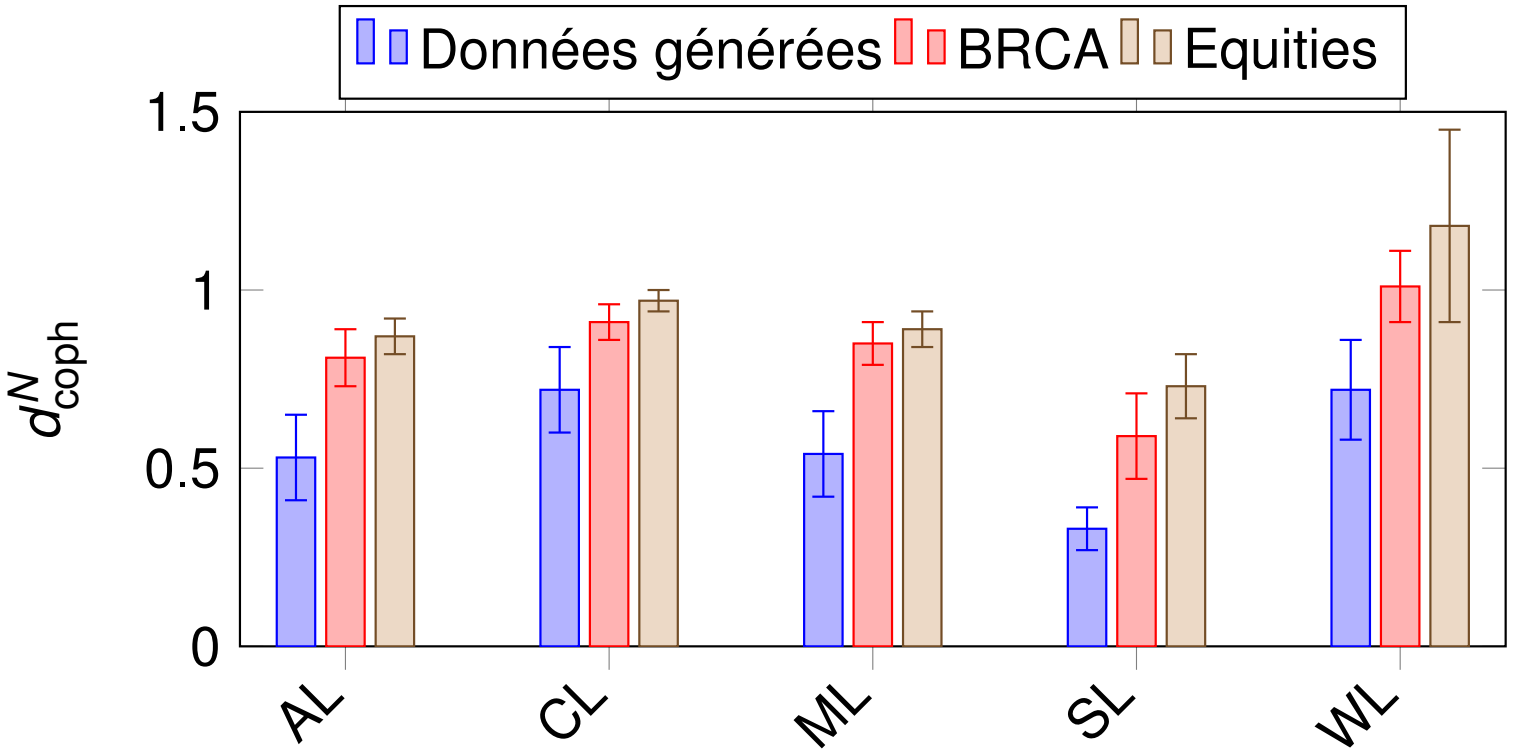
Théorème (D., Gallopin et Molinier, 2024+)

Soit deux échantillons $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ *et* $(\mathbf{y}_1, \ldots, \tilde{\mathbf{y}}_i, \ldots, \mathbf{y}_n)$, *où* $\tilde{\mathbf{y}}_i \sim \mathbf{y}_i \sim \mathbf{Y}$, *et soit* S *et* $\tilde{S}$ *les matrices de covariance empiriques respectives. Alors, pour* $\alpha \in (0,1)$, *avec probabilité* $1-\alpha$,

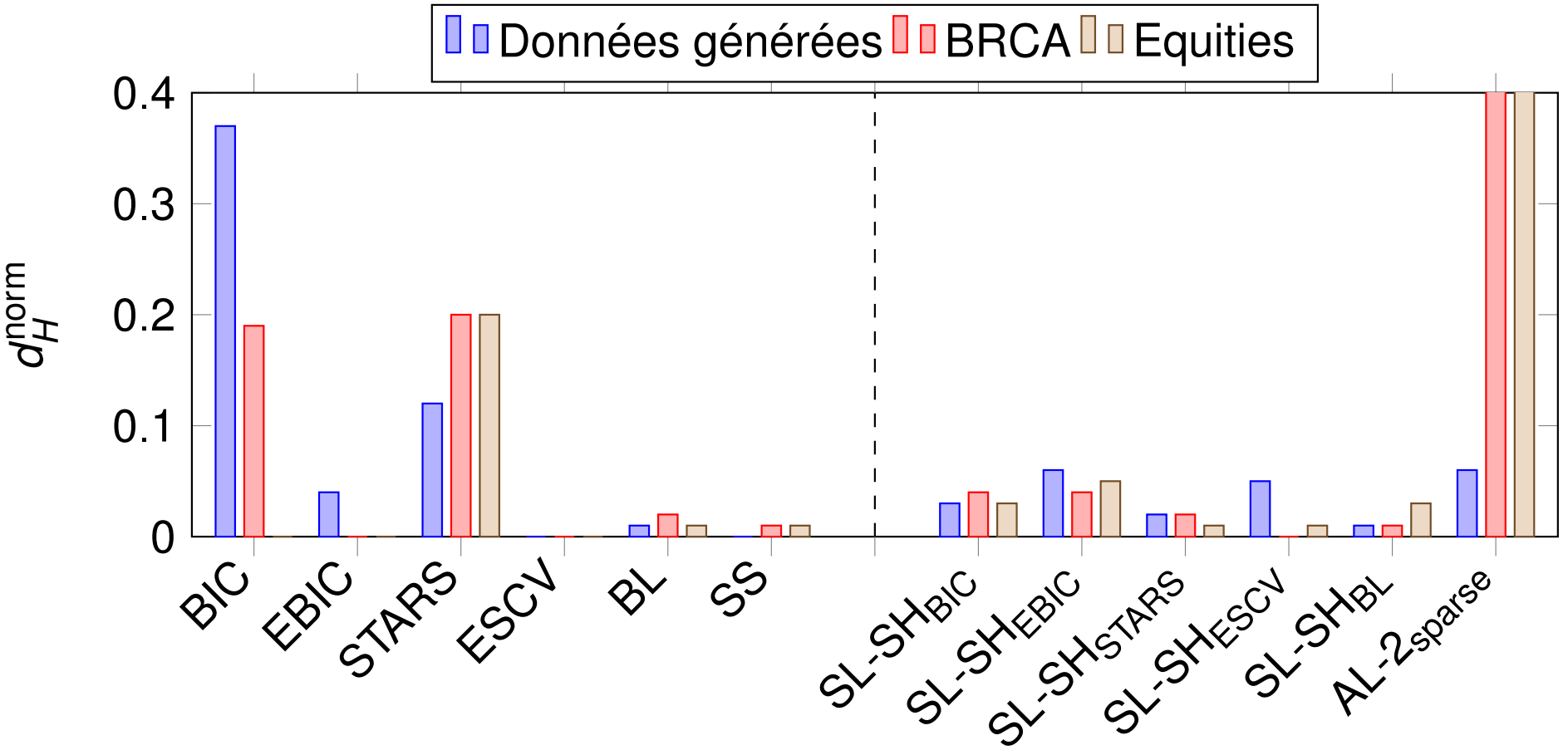
$$d_{coph}(\theta_{\mathbf{1}-|S|}, \theta_{\mathbf{1}-|\tilde{S}|}) \leq \frac{2p}{(n-1)\sqrt{\alpha}}.$$

13. Carlsson et Mémoli, 2010, JMLR
14. D., Gallopin et Molinier, 2024, arXiv

En pratique : quelle fonction de lien ?



En pratique : performance en stabilité des réseaux



Conclusion de la partie 1

- Méthode d'inférence de réseaux en grande dimension
- Décomposition en deux étapes pour stabiliser l'inférence

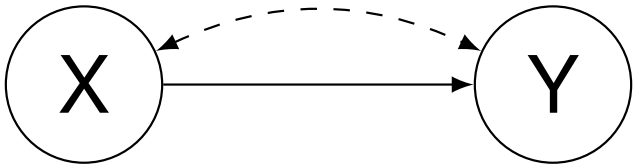
- Structure en blocs : variables cachées influentes ?
- Inconvénient de la modélisation : hypothèse gaussienne

Conclusion de la partie 1

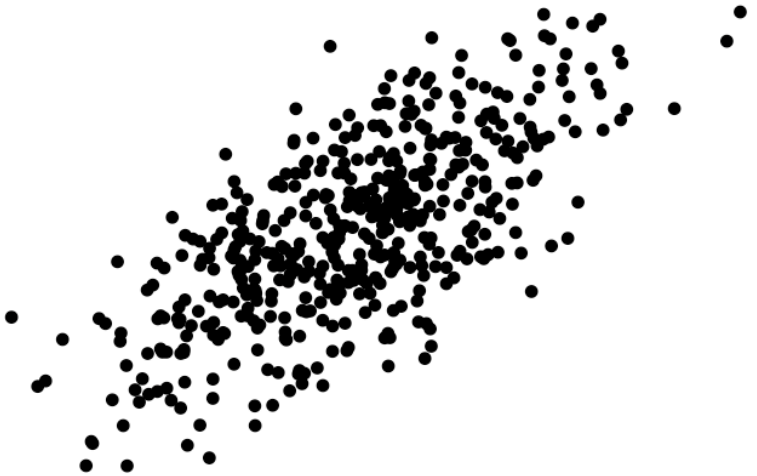
- Méthode d'inférence de réseaux en grande dimension
- Décomposition en deux étapes pour stabiliser l'inférence

- Structure en blocs : variables cachées influentes ?
- Inconvénient de la modélisation : hypothèse gaussienne

Partie 2.
CAUSALITÉ

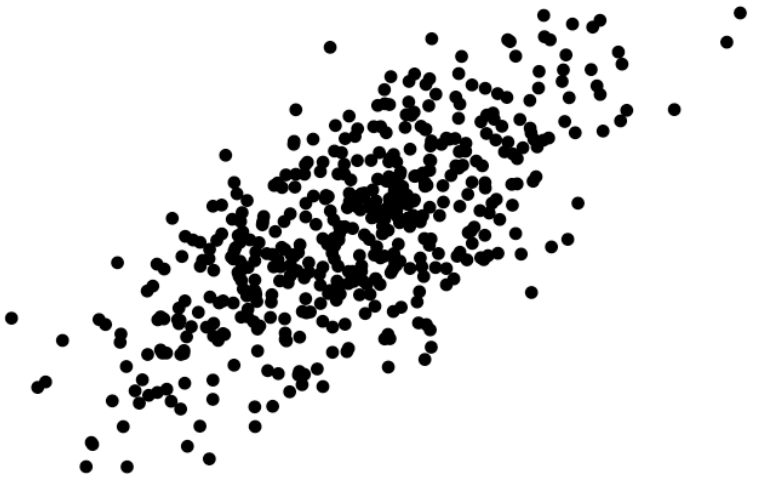


Graphes orientés - cas gaussien [15]



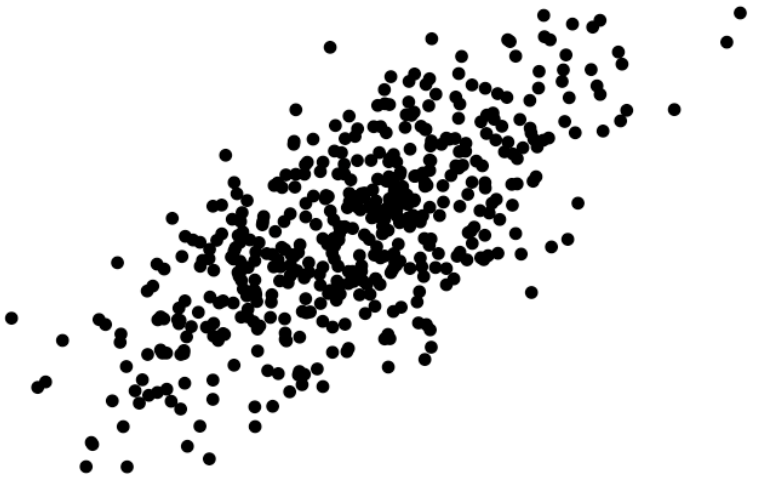
15. Shimizu, Hoyer, Hyvärinen et Kerminen, 2006, JMLR

Graphes orientés - cas gaussien [15]



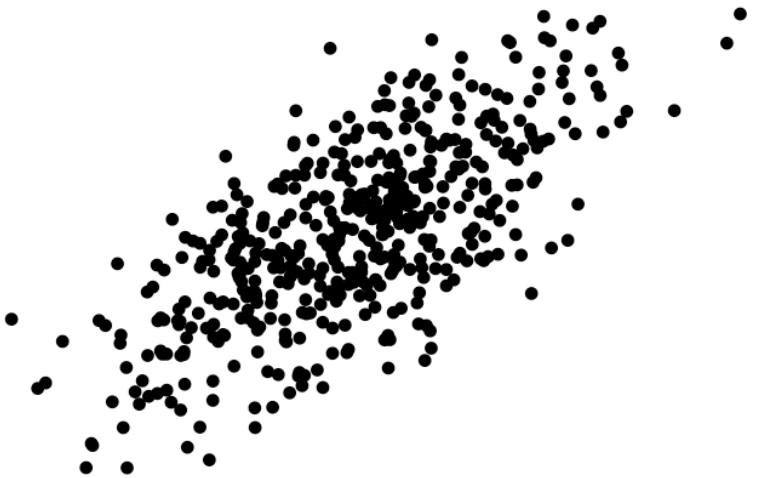
15. Shimizu, Hoyer, Hyvärinen et Kerminen, 2006, JMLR

Graphes orientés - cas gaussien [15]



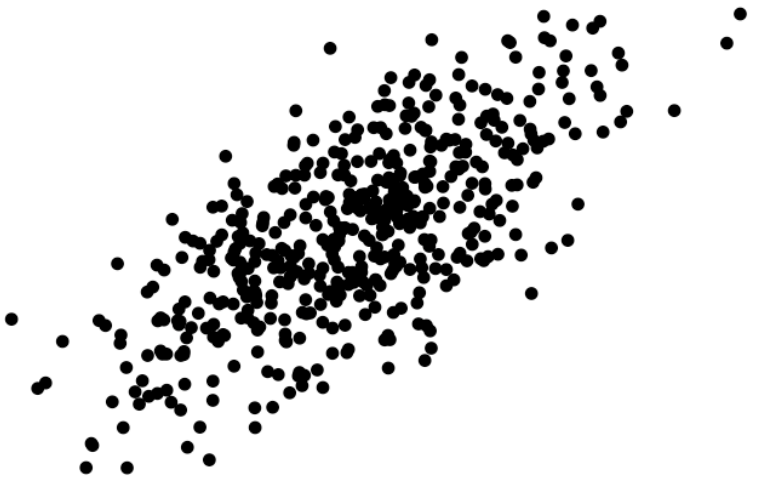
15. Shimizu, Hoyer, Hyvärinen et Kerminen, 2006, JMLR

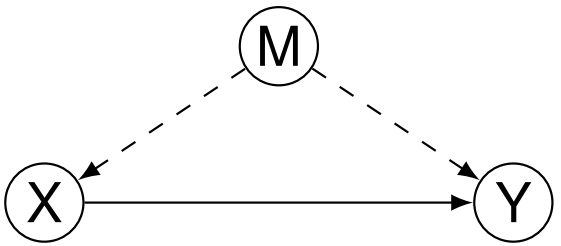
Graphes orientés - cas gaussien [15]



15. Shimizu, Hoyer, Hyvärinen et Kerminen, 2006, JMLR

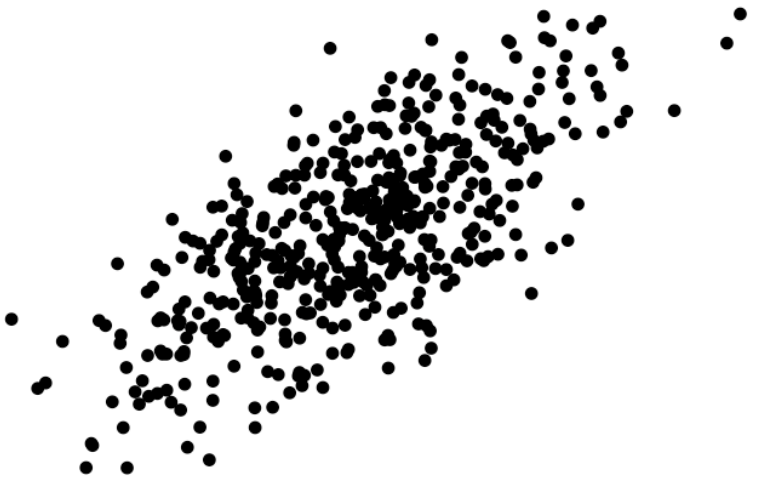
Graphes orientés - cas gaussien [15]

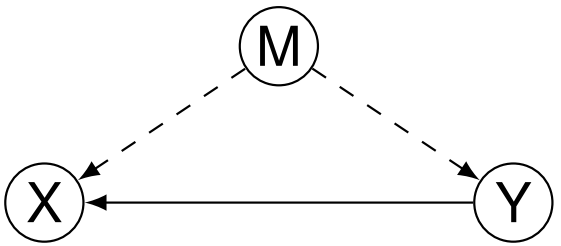




15. Shimizu, Hoyer, Hyvärinen et Kerminen, 2006, JMLR

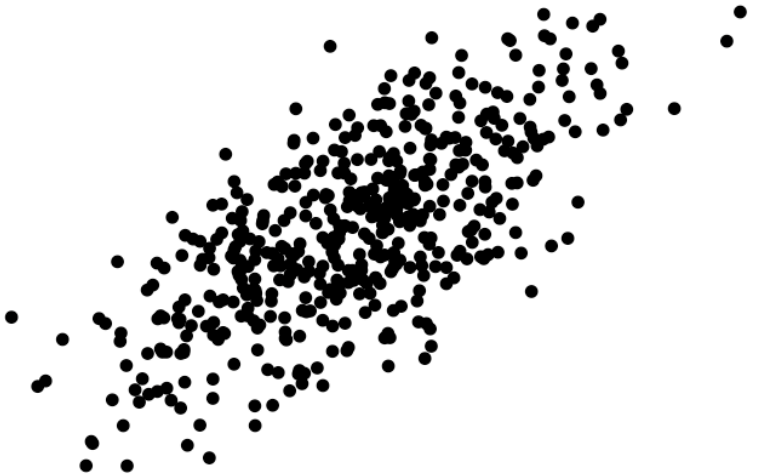
Graphes orientés - cas gaussien [15]





15. Shimizu, Hoyer, Hyvärinen et Kerminen, 2006, JMLR

Graphes orientés - cas gaussien [15]



⚠ Magie des vecteurs gaussiens

15. Shimizu, Hoyer, Hyvärinen et Kerminen, 2006, JMLR

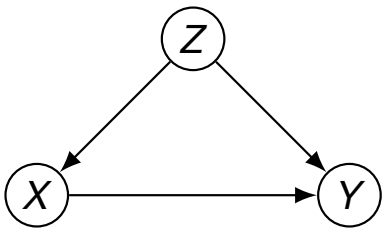
Causalité et interventions [16]

- **Modèle structurel causal**

$$X_j = f_j(\mathrm{Pa}(X_j), \varepsilon_j), \text{ et } (\varepsilon_j)_j \text{ mutuellement indépendants}$$

- **Représentation** : graphes dirigés acycliques (DAGs)

$$Z = \varepsilon_Z$$
$$X = f_X(Z, \varepsilon_X)$$
$$Y = f_Y(X, Z, \varepsilon_Y)$$



16. Pearl, 2000

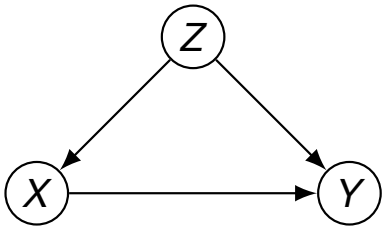
Causalité et interventions [16]

- **Modèle structurel causal**

$$X_j = f_j(\mathrm{Pa}(X_j), \varepsilon_j), \text{ et } (\varepsilon_j)_j \text{ mutuellement indépendants}$$

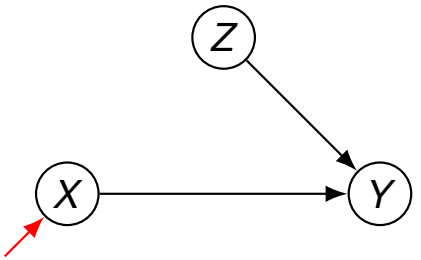
- **Représentation** : graphes dirigés acycliques (DAGs)

$$Z = \varepsilon_Z$$
$$X = f_X(Z, \varepsilon_X)$$
$$Y = f_Y(X, Z, \varepsilon_Y)$$



- **Intervention** : manipulation du système, encodée par l'opérateur do.

$$do(X = x)$$
$$Z = \varepsilon_Z$$
$$X = x$$
$$Y = f_Y(X, Z, \varepsilon_Y)$$



16. Pearl, 2000

Causalité et interventions [16]

- **Modèle structurel causal**

$$X_j = f_j(\mathrm{Pa}(X_j), \varepsilon_j), \text{ et } (\varepsilon_j)_j \text{ mutuellement indépendants}$$

- **Représentation** : graphes dirigés acycliques (DAGs)
- **Intervention** : manipulation du système, encodée par l'opérateur do.

$$P(y|do(x))$$

16. Pearl, 2000

Causalité et interventions [16]

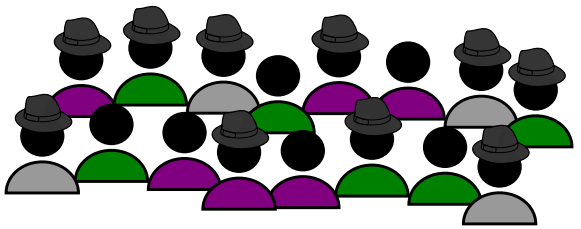
- **Modèle structurel causal**

$$X_j = f_j(\mathrm{Pa}(X_j), \varepsilon_j), \text{ et } (\varepsilon_j)_j \text{ mutuellement indépendants}$$

- **Représentation** : graphes dirigés acycliques (DAGs)
- **Intervention** : manipulation du système, encodée par l'opérateur do.

$$P(y|do(x))$$

$$\text{⚠ } P(y|do(x)) \neq P(y|x)$$



16. Pearl, 2000

Causalité et interventions [16]

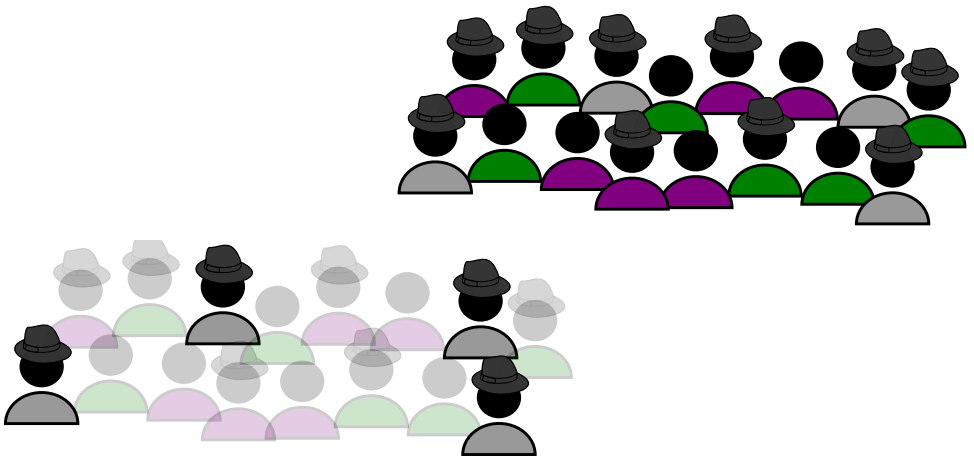
- **Modèle structurel causal**

$$X_j = f_j(\mathrm{Pa}(X_j), \varepsilon_j), \text{ et } (\varepsilon_j)_j \text{ mutuellement indépendants}$$

- **Représentation** : graphes dirigés acycliques (DAGs)
- **Intervention** : manipulation du système, encodée par l'opérateur do.

$$P(y|do(x))$$

$$\text{⚠ } P(y|do(x)) \neq P(y|x)$$



$$\mathbb{P}(\text{chapeau}|\text{T-shirt gris}) = 1$$

16. Pearl, 2000

Causalité et interventions [16]

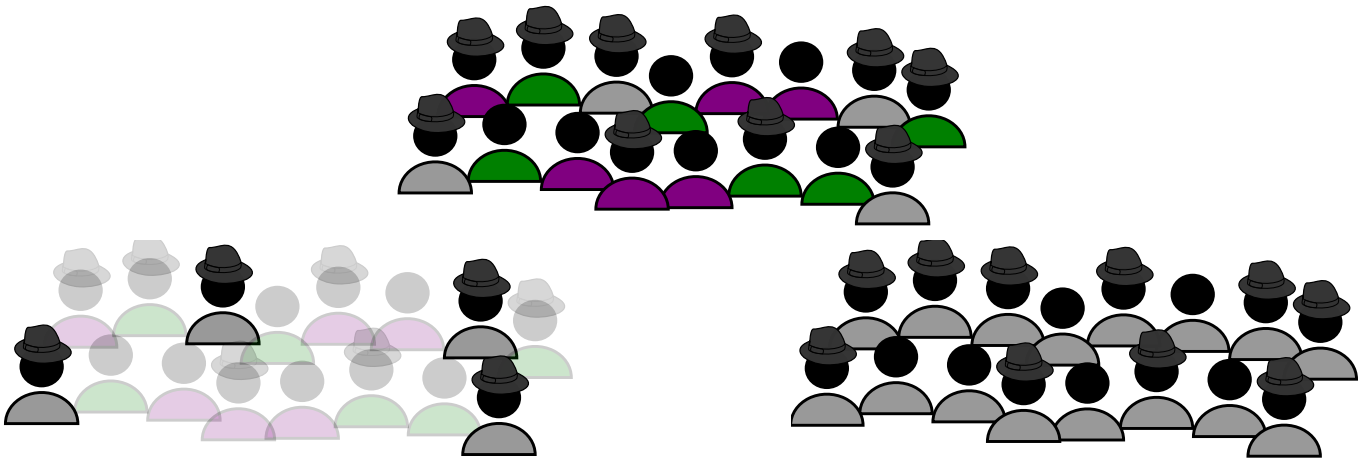
- **Modèle structurel causal**

$$X_j = f_j(\mathrm{Pa}(X_j), \varepsilon_j), \text{ et } (\varepsilon_j)_j \text{ mutuellement indépendants}$$

- **Représentation** : graphes dirigés acycliques (DAGs)
- **Intervention** : manipulation du système, encodée par l'opérateur do.

$$P(y|do(x))$$

$$\text{⚠ } P(y|do(x)) \neq P(y|x)$$



$$\mathbb{P}(\text{chapeau}|\text{T-shirt gris}) = 1 \qquad \mathbb{P}(\text{chapeau}|do(\text{T-shirt gris})) = 10/16$$

16. Pearl, 2000

Découverte causale

Etant données des données observationnelles, peut-on déterminer le graphe causal ?

Hypothèse : le graphe encode exactement (ni plus, ni moins) les dépendances (conditionnelles) des données observées

Etat de l'art

- Méthodes basées sur le bruit [17] [18]
- Méthodes basées sur les contraintes [19] [20]

17. Shimizu, Hoyer, Hyvärinen et Kerminen, 2006, JMLR
18. Zhang et Hyvärinen, 2009, UAI
19. Spirtes, Glymour et Scheines, 2000
20. Colombo et Maathuis, 2014, JMLR

Découverte causale

Etant données des données observationnelles, peut-on déterminer le graphe causal ?

Hypothèse : le graphe encode exactement (ni plus, ni moins) les dépendances (conditionnelles) des données observées

Etat de l'art

- Méthodes basées sur le bruit [17] [18]
  - Dans le modèle additif, si au plus un des bruits est Gaussien !
  - Si bruit non additif, il existe des extensions
- Méthodes basées sur les contraintes [19] [20]

17. Shimizu, Hoyer, Hyvärinen et Kerminen, 2006, JMLR
18. Zhang et Hyvärinen, 2009, UAI
19. Spirtes, Glymour et Scheines, 2000
20. Colombo et Maathuis, 2014, JMLR

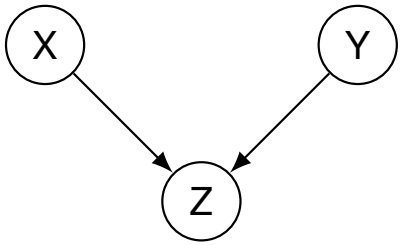
Découverte causale

Etant données des données observationnelles, peut-on déterminer le graphe causal ?

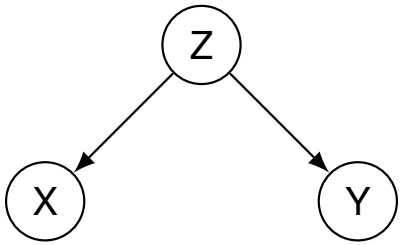
Hypothèse : le graphe encode exactement (ni plus, ni moins) les dépendances (conditionnelles) des données observées

Etat de l'art

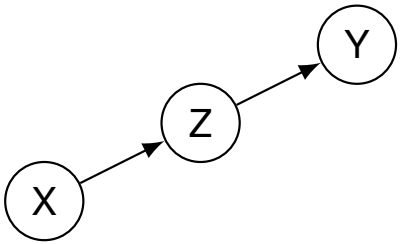
- Méthodes basées sur le bruit [17] [18]
- Méthodes basées sur les contraintes [19] [20]
  1. Squelette : construit par des tests d'indépendance, potentiellement non paramétrique
  2. Orientation par des règles



v-structure



Fourchette



Chaine

⚠ On obtient la classe d'équivalence de Markov du vrai graphe

17. Shimizu, Hoyer, Hyvärinen et Kerminen, 2006, JMLR
18. Zhang et Hyvärinen, 2009, UAI
19. Spirtes, Glymour et Scheines, 2000
20. Colombo et Maathuis, 2014, JMLR

Raisonnement causal

Etant donnés le graphe causal et des données observationnelles, peut-on répondre à une question causale ?

But : estimer $P(y|do(x))$

- Données expérimentales ? Essai de contrôle randomisé. Coûteux, non éthique ou même infaisable.
- Directement à partir des données observationnelles ? Identifiabilité

Etat de l'art

- Le critère backdoor identifie l'effet total dans les DAGs [21]

$$P(Y = y|do(X = x)) = \sum_{z \in \Omega(Z)} P(Y = y|X = x, Z = z)P(Z = z)$$

- Le critère backdoor généralisé identifie l'effet total dans la classe d'équivalence de Markov [22]
- Autres ? Critère frontdoor, do-calculus, algorithm ID, . . .

21. Pearl, 2000
22. Maathuis et Colombo, 2015, AoS

Raisonnement causal

Etant donnés le graphe causal et des données observationnelles, peut-on répondre à une question causale ?

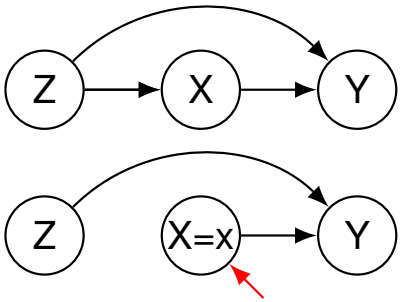
But : estimer $P(y|do(x))$

- Données expérimentales ? Essai de contrôle randomisé. Coûteux, non éthique ou même infaisable.
- Directement à partir des données observationnelles ? Identifiabilité

Etat de l'art

- Le critère backdoor identifie l'effet total dans les DAGs [21]

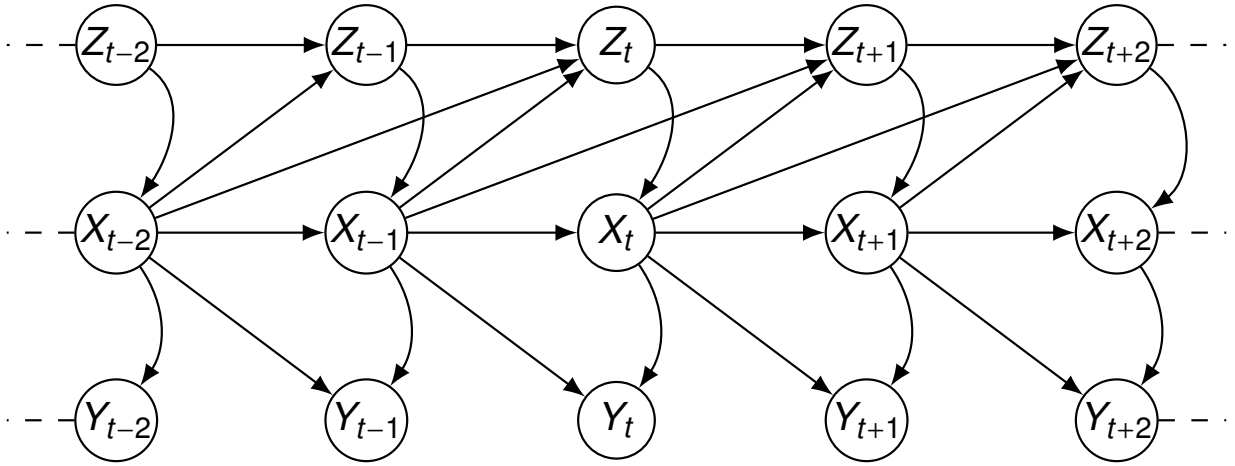
$$P(Y = y|do(X = x))$$
$$= \sum_{z \in \Omega(Z)} P(Y = y|X = x, Z = z)P(Z = z)$$



- Le critère backdoor généralisé identifie l'effet total dans la classe d'équivalence de Markov [22]
- Autres ? Critère frontdoor, do-calculus, algorithm ID, . . .

21. Pearl, 2000
22. Maathuis et Colombo, 2015, AoS

Graphes causaux pour les séries temporelles



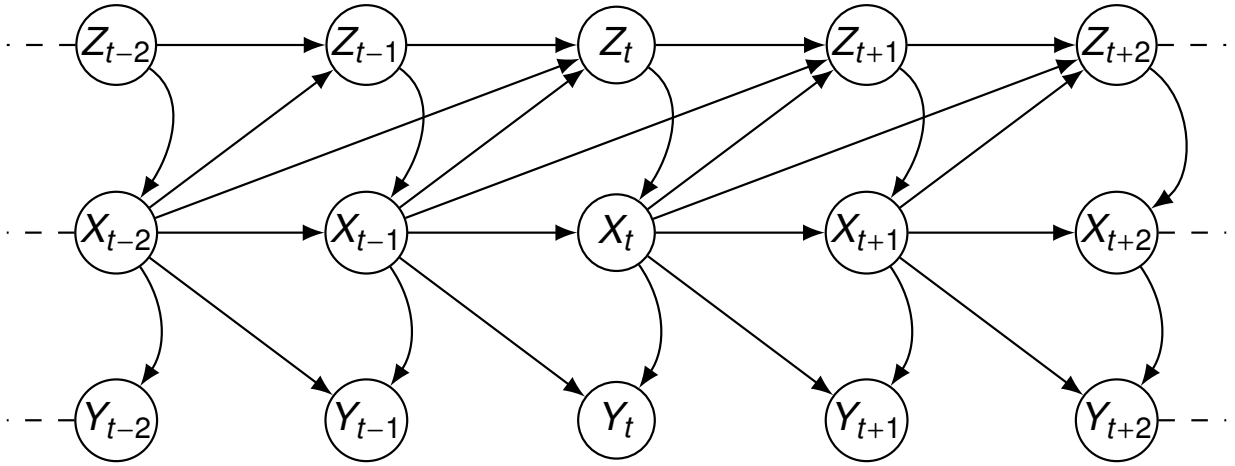
Graphe causal déplié

Graphe causal à fenêtre

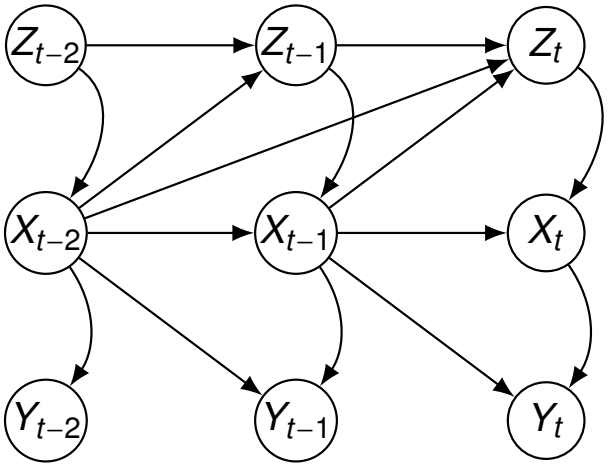
Graphe causal résumé
⚠ potentiellement cyclique

Graphe causal résumé
étendu

Graphes causaux pour les séries temporelles



Graphe causal déplié

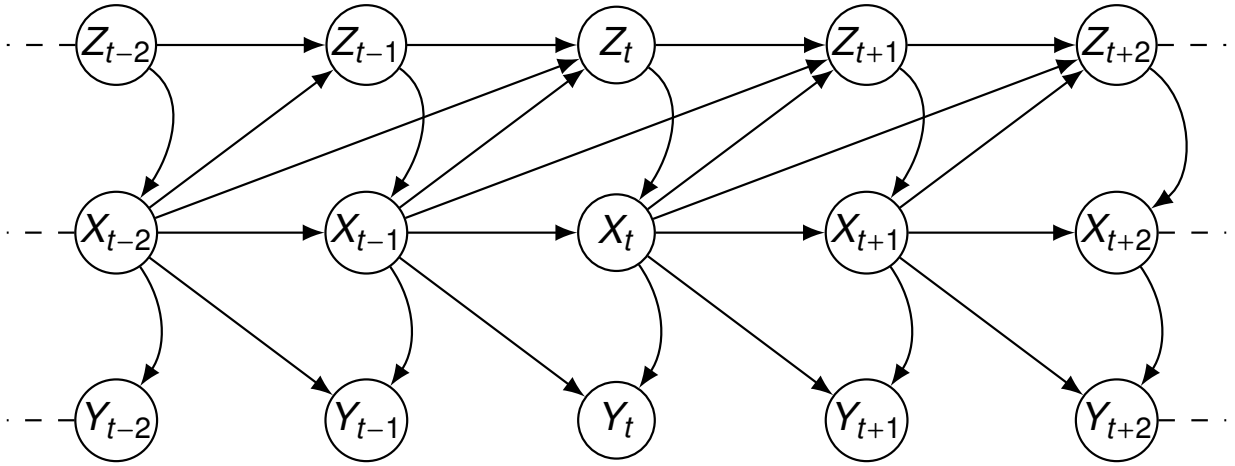


Graphe causal à fenêtre

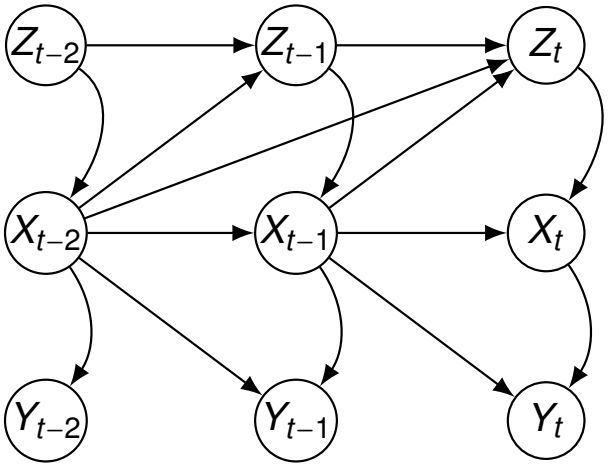
Graphe causal résumé
⚠ potentiellement cyclique

Graphe causal résumé
étendu

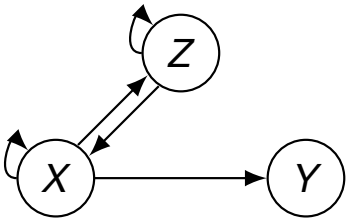
Graphes causaux pour les séries temporelles



Graphe causal déplié



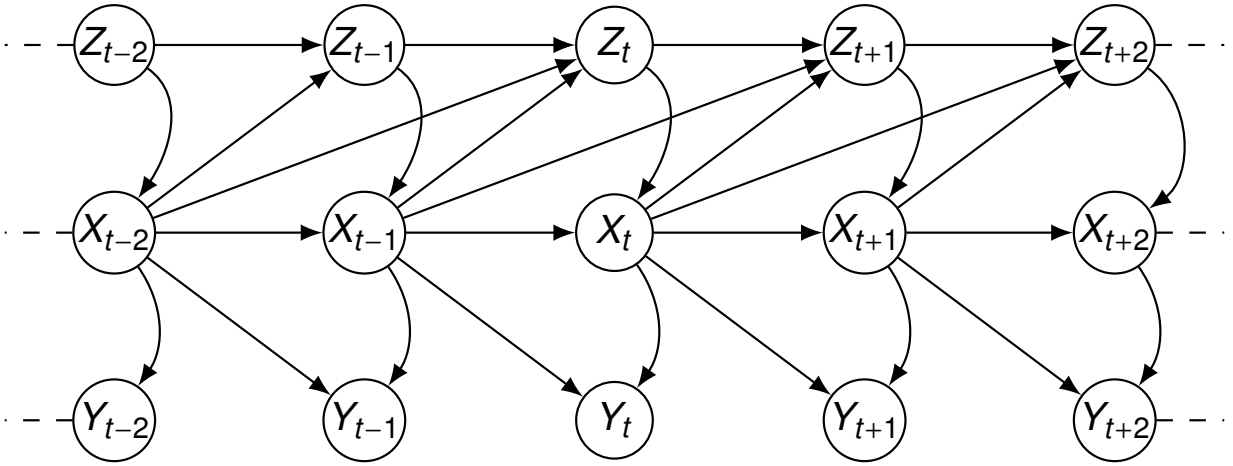
Graphe causal à fenêtre



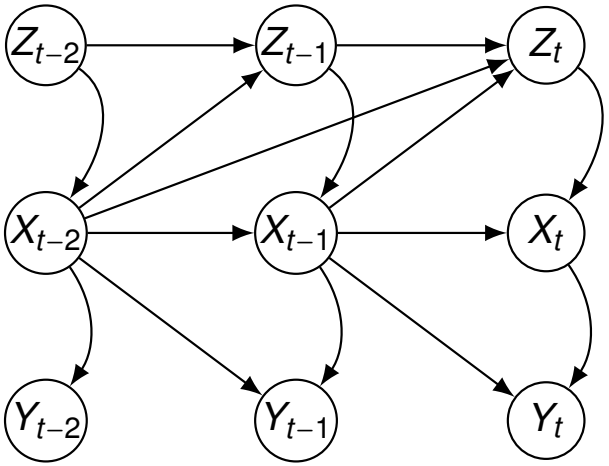
Graphe causal résumé
⚠ potentiellement cyclique

Graphe causal résumé étendu

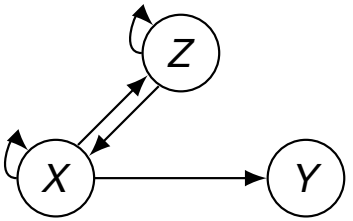
Graphes causaux pour les séries temporelles



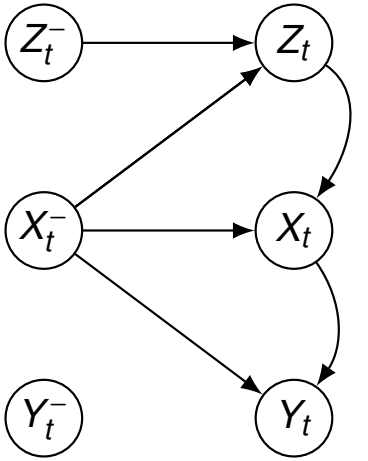
Graphe causal déplié



Graphe causal à fenêtre



Graphe causal résumé
⚠ potentiellement cyclique



Graphe causal résumé étendu

Découverte causale pour les séries temporelles

- Thèse Charles Assaad, 2021
- Thèse Lei Zan, 2024

- Etat de l'art : quels graphes ? Quelles méthodes ? [23]
- Introduction du graphe causal résumé étendu [24]
- Tests d'indépendance adaptées aux séries temporelles et aux données mixtes [25]
- Méthode hybride, profitant des avantages des différentes classes [26]

23. Assaad, D. et Gaussier, 2022, JAIR
24. Assaad, D. et Gaussier, 2022, UAI
25. Zan, Meynaoui, Assaad, D. et Gaussier, 2022, Entropy
26. Bystrova, Assaad, Arbel, D., Gaussier et Thuillier, 2024, TMLR

Raisonnement causal pour les séries temporelles

Etant donné un graphe causal pour les séries temporelles, en supposant qu'on observe toutes les variables, peut-on calculer

$$P(Y_t = y_t | do(X_{t-\gamma} = x_{t-\gamma}))?$$

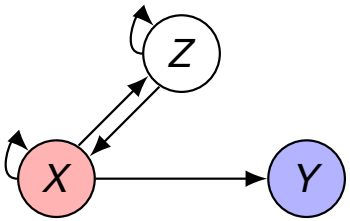
- Dans le graphe causal déplié [27] [28] : le critère backdoor est complet pour l'identification de l'effet total.
- Dans un graphe résumé ou résumé étendu : que signifie être identifiable ?
  - La même formule fonctionne pour tous les graphes admissibles
  - Dans le graphe résumé étendu [29], le critère backdoor commun est complet pour l'identification de l'effet total.
  - Dans le graphe résumé, c'est plus complexe, à cause des cycles potentiels.

27. Blondel, Arias et Gavaldà, 2017, IJDSA
28. Shpitser et Pearl, 2008, JMLR
29. Assaad, D., Gaussier, Goessler, Meynaoui, 2024, UAI

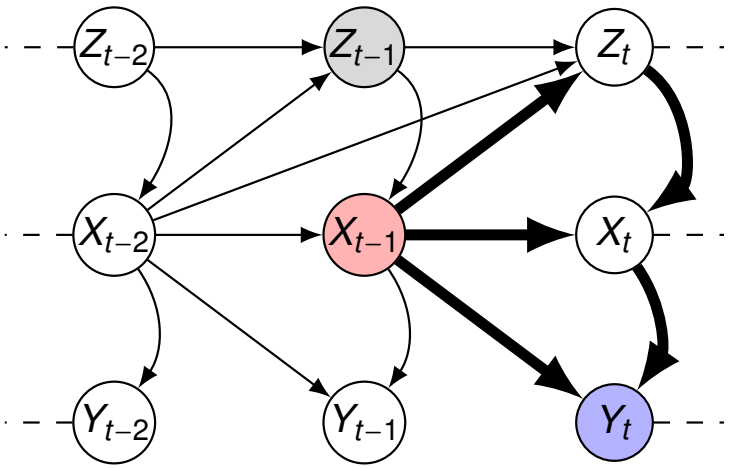
Raisonnement causal pour les séries temporelles

Théorème (Yvernes, Assaad, D., Gaussier, 2024+)

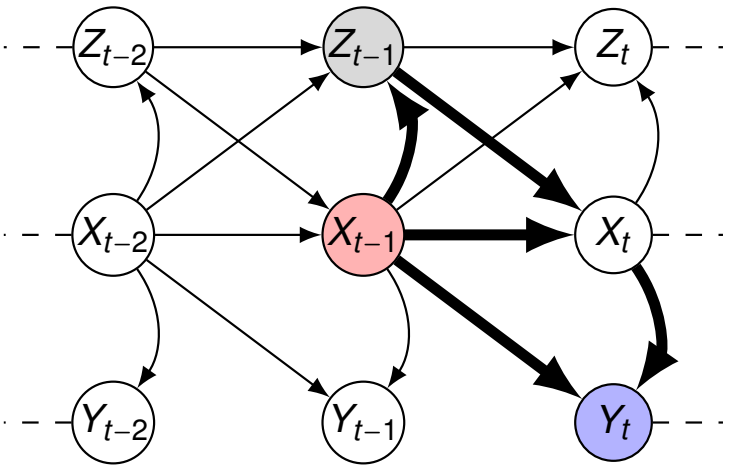
Soit un graphe causal résumé. L'effet total $P(y_t \mid do(x^1_{t-\gamma_1}), \ldots, do(x^m_{t-\gamma_m}))$ *est identifiable par backdoor commun si et seulement si pour tout* $X^j_{t-\gamma_j}$ *et pour tout graphe déplié admissible, il n'existe pas de chemin backdoor sans collider de* $X^i_{t-\gamma_i}$ *à* Y_t *qui reste dans le cône de descendance.*



Graphe causal résumé, et $P(y_t|do(x_{t-1}))$.



Un graphe déplié.



Un autre graphe déplié.

Sélection de variables causales

- En statistiques, la sélection de variables est utilisée pour
  - Réduire la dimension
  - Améliorer l'interprétabilité
  - Améliorer la généralisation des modèles

⚠ Peut-on mettre un sens causal aux variables sélectionnées ?

💡 Sous les hypothèses causales, les variables les plus pertinentes pour prédire une réponse sont ses parents, au sens minimax [30]

- La couverture de Markov (parents, enfants et époux) est le plus petit ensemble qui donne toute l'information

? Comment comprendre la distinction ?

? Que se passe-t-il si on suppose que certaines variables sont cachées ?

? Peut-on utiliser la sélection causale dans les méthodes classiques de prédiction en apprentissage statistique ?

🎓 Thèse Théotime Le Goff, en cours

30. Peters, Buhlmann et Meinshausen, 2016, JRSS B

Sélection de variables causales

- En statistiques, la sélection de variables est utilisée pour
  - Réduire la dimension
  - Améliorer l'interprétabilité
  - Améliorer la généralisation des modèles
- ⚠ Peut-on mettre un sens causal aux variables sélectionnées ?
- 💡 Sous les hypothèses causales, les variables les plus pertinentes pour prédire une réponse sont ses parents, au sens minimax [30]
- La couverture de Markov (parents, enfants et époux) est le plus petit ensemble qui donne toute l'information
- ? Comment comprendre la distinction ?
- ? Que se passe-t-il si on suppose que certaines variables sont cachées ?
- ? Peut-on utiliser la sélection causale dans les méthodes classiques de prédiction en apprentissage statistique ?
- 🎓 Thèse Théotime Le Goff, en cours

30. Peters, Buhlmann et Meinshausen, 2016, JRSS B

Abstraction de graphes causaux

- Connaissance causale à un plus haut niveau de granularité que les variables observées
- La classe d'équivalence de Markov est une abstraction du vrai DAG, la seule que l'on puisse atteindre en découverte sans hypothèse paramétrique
- Les graphes résumés et résumés étendus sont une abstraction du graphe causal déplié
- Proposition pour des groupes de variables [31]
- ? Découverte causale ? Raisonnement causal ?
- ? Comment choisir le bon niveau de granularité pour pouvoir estimer un effet causal ?
- Thèse Clément Yvernes, en cours

31. Anand, Ribeiro, Tian et Bareinboim, 2023, AAAI

Abstraction de graphes causaux

- Connaissance causale à un plus haut niveau de granularité que les variables observées
- La classe d'équivalence de Markov est une abstraction du vrai DAG, la seule que l'on puisse atteindre en découverte sans hypothèse paramétrique
- Les graphes résumés et résumés étendus sont une abstraction du graphe causal déplié
- Proposition pour des groupes de variables [31]
- ? Découverte causale ? Raisonnement causal ?
- ? Comment choisir le bon niveau de granularité pour pouvoir estimer un effet causal ?
- Thèse Clément Yvernes, en cours

31. Anand, Ribeiro, Tian et Bareinboim, 2023, AAAI

Graphe causal différentiel pour deux populations

- Réseaux différentiels [32] : large littérature, notamment avec les modèles graphiques gaussiens
- Pour les graphes causaux ? Littérature plus restreinte
- Découverte causale : modèles paramétriques [33]
- Intérêt pratique : collaboration avec le CHU
- ? Découverte causale non paramétrique ?
- ? Relaxer l'hypothèse d'ordre causal commun ?
- ? Pouvoir raisonner à partir de ces graphes ?

32. Shojaie, 2020, Computational Statistics
33. Chen, Bello, Aragam, et Ravikumar, 2023, NeurIPS

Graphe causal différentiel pour deux populations

- Réseaux différentiels [32] : large littérature, notamment avec les modèles graphiques gaussiens
- Pour les graphes causaux ? Littérature plus restreinte
- Découverte causale : modèles paramétriques [33]
- Intérêt pratique : collaboration avec le CHU
- ? Découverte causale non paramétrique ?
- ? Relaxer l'hypothèse d'ordre causal commun ?
- ? Pouvoir raisonner à partir de ces graphes ?

32. Shojaie, 2020, Computational Statistics
33. Chen, Bello, Aragam, et Ravikumar, 2023, NeurIPS

Conclusion de la partie 2

- Causalité au sens de Pearl
- Travaux sur les graphes causaux pour les séries temporelles (découverte et raisonnement)
- Différents intérêts pratiques :
  - découverte causale à partir de données observationnelles : lien entre la connaissance des experts et ce qui est encodé dans les données ?
  - répondre à des questions interventionnelles à partir de données observationnelles

Conclusion de l'exposé

- **Modélisation de l'interdépendance dans les données via les modèles graphiques**
  - Etude des modèles graphiques gaussiens
  - Exploration des modèles causaux, avec un focus particulier sur les séries temporelles
- **L'apprentissage statistique au service des applications**
  - La stabilité est essentielle pour garantir des résultats fiables et robustes
  - Potentiel d'interprétation grâce à des modèles paramétriques ou des représentation graphiques
  - Conclusions robustes dans les analyses par les experts
- **Perspectives**

Conclusion générale

Extrait de mes travaux...

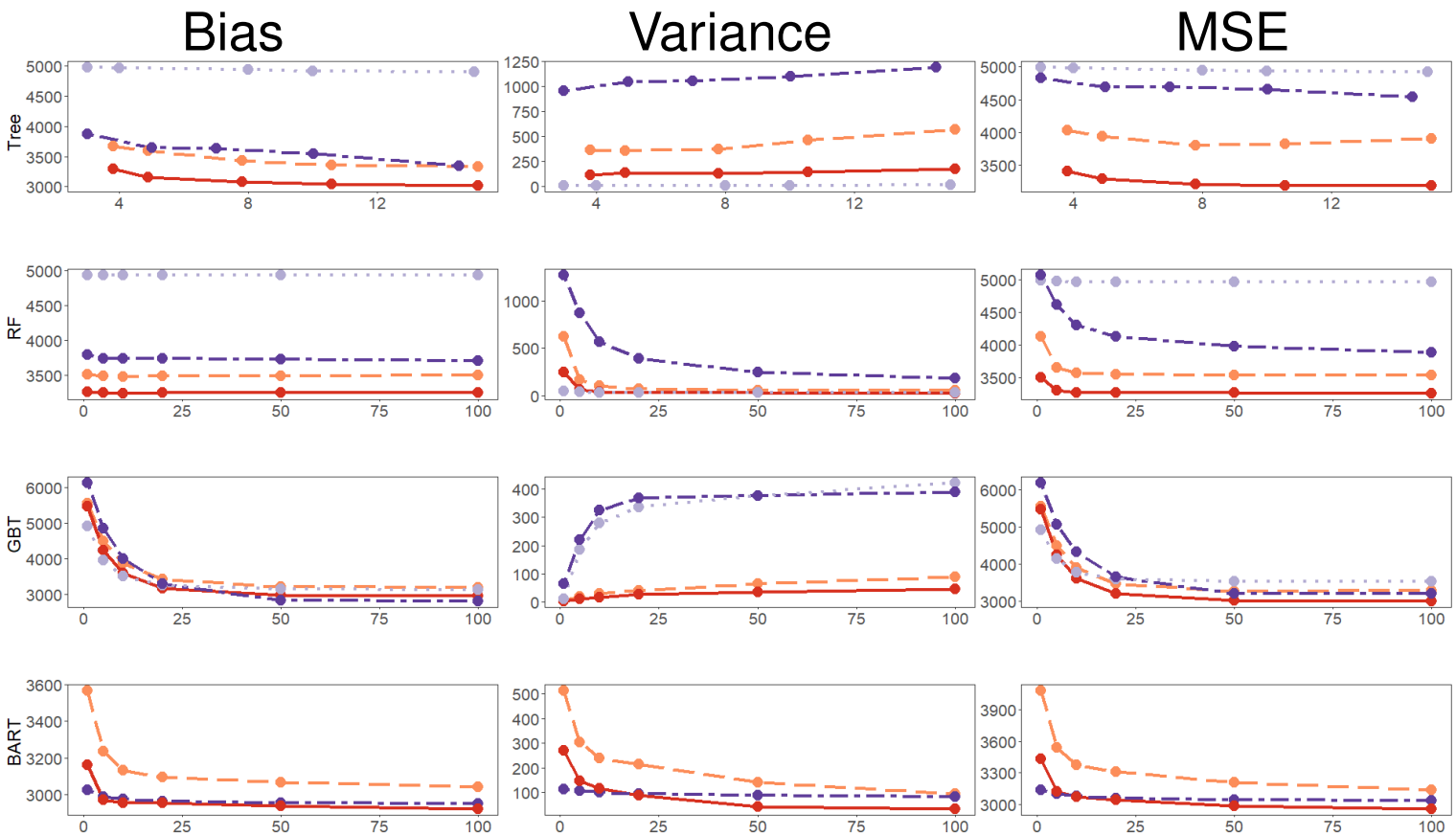
Conclusion générale

Extrait de mes travaux... plus de détails dans le manuscrit !
Merci aux doctorants et post-doctorants qui m'ont fait confiance avant l'heure !

Probabilistic regression trees and their ensemble extensions

$$f_{\mathrm{PR}}(;\Theta) = \sum_{k=1}^{K} \gamma_k \Psi(;\mathcal{R}_k, \sigma).$$

$$\lim_{N \to +\infty} \mathbb{E}\left[\left|f_{\mathrm{PR}}(\mathbf{X}; \hat{\Theta}_N) - \mathbb{E}(Y|\mathbf{X})\right|^2\right] = 0.$$



Nonlinear mixed effects modeling and warping for functional data

$$Y_i(t_{i,j}) = \mu\left\{w^{-1}(t_{i,j};\theta_i)\right\} + U_i\left\{w^{-1}(t_{i,j};\theta_i)\right\} + \varepsilon_{i,j}, \tag{1}$$

Theorem

Let $i \in \{1,\dots,n\}$ *be given. Let* $\theta_i \sim \mathcal{N}_r(\theta_0, \Sigma^\theta)$ *and* $\tilde{\theta}_i \sim \mathcal{N}_r(\tilde{\theta}_0, \Sigma^{\tilde{\theta}})$ *be used to define two warping functions* $w^{-1}(.;\theta_i)$ *and* $w^{-1}(.;\tilde{\theta}_i)$*, and let* X_i *and* $\tilde{X}_i$ *be the corresponding warped functions, such that*

$$Y_i(t) = X_i\{w^{-1}(t;\theta_i)\} = \tilde{X}_i\{w^{-1}(t;\tilde{\theta}_i)\}.$$

Then model (1) *is identifiable if and only if*

$$\mathbf{B}_i^\mu = \mathbb{E}_{\theta_i,\tilde{\theta}_i}\left\{(\mathbf{B}_i^\mu)^{\theta_i,\tilde{\theta}_i}\right\};$$

$$(\mathbf{B}_i^U)^\top \Sigma^{U_i} \mathbf{B}_i^U = \mathbb{V}_{\theta_i,\tilde{\theta}_i}\left\{(\mathbf{B}_i^\mu)^{\theta_i,\tilde{\theta}_i}\boldsymbol{\alpha}^\mu\right\} + \mathbb{E}_{\theta_i,\tilde{\theta}_i}\left[\left\{(\mathbf{B}_i^U)^{\theta_i,\tilde{\theta}_i}\right\}^\top \Sigma^{U_i}(\mathbf{B}_i^U)^{\theta_i,\tilde{\theta}_i}\right].$$

and at least one of the three condition is not satisfied :

1. $(\mathbf{B}_i^U)^\top \mathbf{B}_i^U \neq \mathbf{0}_{m_U}$;

Theorem

Fix N and $\mathbf{T}$. Suppose $(\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$ is a sequence of iid random variables satisfying the functional nonlinear mixed model (1) observed on fixed time points : for $i = 1, \ldots, N$, for $j = 1, \ldots, T_i$, $[\mathbf{Y}_i]_j = Y_i(t_{i,j})$. Moreover, suppose that the model is identifiable and that the update of $\boldsymbol{\theta}_i$ is a contraction mapping. Then, $((\hat{\boldsymbol{\alpha}}^{\mu})^{(\infty)}, (\hat{\sigma}_{\varepsilon})^{(\infty)}, (\Sigma^{U})^{(\infty)}, \hat{\boldsymbol{\theta}}_0^{(\infty)}, (\hat{\Sigma}^{\theta})^{(\infty)})$ exists and is unique, and the algorithm converges to this solution with a geometric rate with respect to the Euclidean distance.

Theorem

Suppose $(\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$ is a sequence of iid random variables satisfying the functional nonlinear mixed model observed on fixed time points : for $i = 1, \ldots, N$, for $j = 1, \ldots, T_i$, $[\mathbf{Y}_i]_j = Y_i(t_{i,j})$. We first make the following assumption, to avoid having to theoretically deal with a modeling bias. We assume that the functions μ, $(U_i)_{i=1,\ldots,N}$ and w belong to the space spanned by the considered spline basis, and that $\sigma_{\varepsilon} \underset{\min T_i \to \infty}{\to} 0$.

Suppose that the model is identifiable and that the update of $\boldsymbol{\theta}_i$ is a contraction mapping. We assume the existence and positive definiteness of $\mathcal{I}$, which is the limit of minus the expected Hessian matrix of the log-likelihood function based on the model. We also assume that for all $i = 1, \ldots, N$, the

Mixture of segmentation

$$Y_{ih}(t_j) = f_{k\ell h}(t_j) + \eta_{ijh},$$

$$[\mathbf{A}_i]_{j.}|(z_{ik} = 1, W_{j\ell} = 1) = \mu_{k\ell} + \varepsilon_{ij} \tag{2}$$

Theorem (Identifiability of (2))

Assume that :

- *For every* $k \in \{1, \ldots, K\}$ *and* $\ell \in \{0, \ldots, L_k\}$, *there exists at least one* $r \in \{1, \ldots, p\}$ *such that* $\sigma_{k\ell r} \neq \sigma_{k,\ell+1,r}$ *or* $\mu_{k\ell r} \neq \mu_{k,\ell+1,r}$.
- *We have* $D \geq \max_{k \in \{1,\ldots,K\}} L_k + 1$.
- *If there exists* $k \neq k'$ *such that* $L_k = L_{k'}$ *then :*

  there exists $\ell \in \{0, \ldots, L_k\}$ *such that* $T_{k\ell} \neq T_{k',\ell}$,
  or there exists $\ell \in \{0, \ldots, L_k\}$ *and* $r \in \{1, \ldots, p\}$ *such that :* $\sigma_{k\ell r} \neq \sigma_{k',\ell,r}$ *or* $\mu_{k\ell r} \neq \mu_{k',\ell,r}$.
- *For every* $k \in \{1, \ldots, K\}$, $\pi_k > 0$.

Under these assumptions, the model (2) *is identifiable.*

Theorem

Let $\mathbf{A}$ *be a matrix of a* $n \times T$ *observations of the model* (2) *with true parameter* $\theta, \mathbf{T}$ *where the number of clusters* K *and the number of segments* $(L_k)_{1 \leq k \leq K}$ *are known. We assume that there exists* $M > 0$ *such that for all* $k \in \{1, \dots, K\}$ *and* $\ell \in \{0, \dots, L_k\}$,

$$\mu_{k\ell} \in [-M; M];$$

that there exists $\tau_{\min} > 0$ *such that for all* $k \in \{1, \dots, K\}$ *and* $\ell \in \{0, \dots, L_k\}$,

$$T_{k,\ell+1} - T_{k\ell} > \tau_{\min} d.$$

We also assume that $\log(N)/d \underset{n,d \to +\infty}{\longrightarrow} 0$. *If there exists* $k \neq k'$ *such that* $L_k = L_{k'}$ *then we assume that there exists at least* $\tau_{\min} d$ *coordinates* j *such that the distribution of* $Y_{ij} | z_{ik} = 1$ *is different from the distribution of* $Y_{ij} | z_{ik'} = 1$. *Finally, we assume that there exists a constant* $c > 0$ *such that for every* $k \in \{1, \dots, K\}$, $\pi_k > c$, *and Assumption I. Then,*

$$\left(\widehat{\theta}, \widehat{\mathbf{T}}\right) \underset{n,d \to +\infty}{\overset{\mathbb{P}}{\to}} \left(\theta^\star, \mathbf{T}^\star\right).$$

Simultaneous confidence bands

Theorem

Set a probability $\alpha \in [0,1]$. Then, we have

$$\mathbb{P}\left(\forall t \in [0,1], \left|\hat{\underline{f}}^{L,L^*}(t) - \underline{f}^{L,L^*}(t)\right| \leq \hat{d}^L(t)\right) = 1-\alpha$$

$$\text{with } \hat{d}^L(t) = \hat{q}^L \sqrt{\hat{C}^{L,L^*}(t,t)/N};$$

and $\hat{q}^L$ defined as the solution of the following equation, seen as a function of q^L :

$$\alpha = \mathbb{P}(|t_{N-1}| > q^L) + \frac{\|\tau^L\|_1}{\pi}\left(1 + \frac{(q^L)^2}{N-1}\right)^{-(N-1)/2},$$

with $(\tau^L)^2(t) = \partial_{12}c(t,t) = Var(Z_L(t))'$ where we denote $\partial_{12}c(t,t)$ the partial derivatives of a function $c(t,s)$ in the first and second coordinates and then evaluated at $t = s$.

We can thus deduce a confidence band of level $1-\alpha$ for $\underline{f}^{L,L^*}$:

$$CB_1(\underline{f}^{L,L^*}) = \{\forall t \in [0,1], [\hat{\underline{f}}^{L,L^*}(t) - \hat{d}^L(t); \hat{\underline{f}}^{L,L^*}(t) + \hat{d}^L(t)]\}.$$

A natural criteria to select the best L is then

$$\widetilde{crit}(L) = \|\hat{q}^{L_{\max}} - \hat{q}^{L}\|^2 + \lambda \frac{L}{N}.$$

We define $\tilde{L} = \arg\min_L \widetilde{crit}(L)$, and center the band around $\underline{\hat{f}}^{\tilde{L},L^*}$:

$$CB_2(\underline{f}^{L^*}) = CB_1(\underline{f}^{\tilde{L},L^*})$$

Tree-based Active Learning

$$n_k^* = n_{\text{act}} \frac{\sqrt{\pi_k \hat{\sigma}_k^2}}{\sum_{\ell=1}^{K} \sqrt{\pi_\ell \hat{\sigma}_\ell^2}};$$

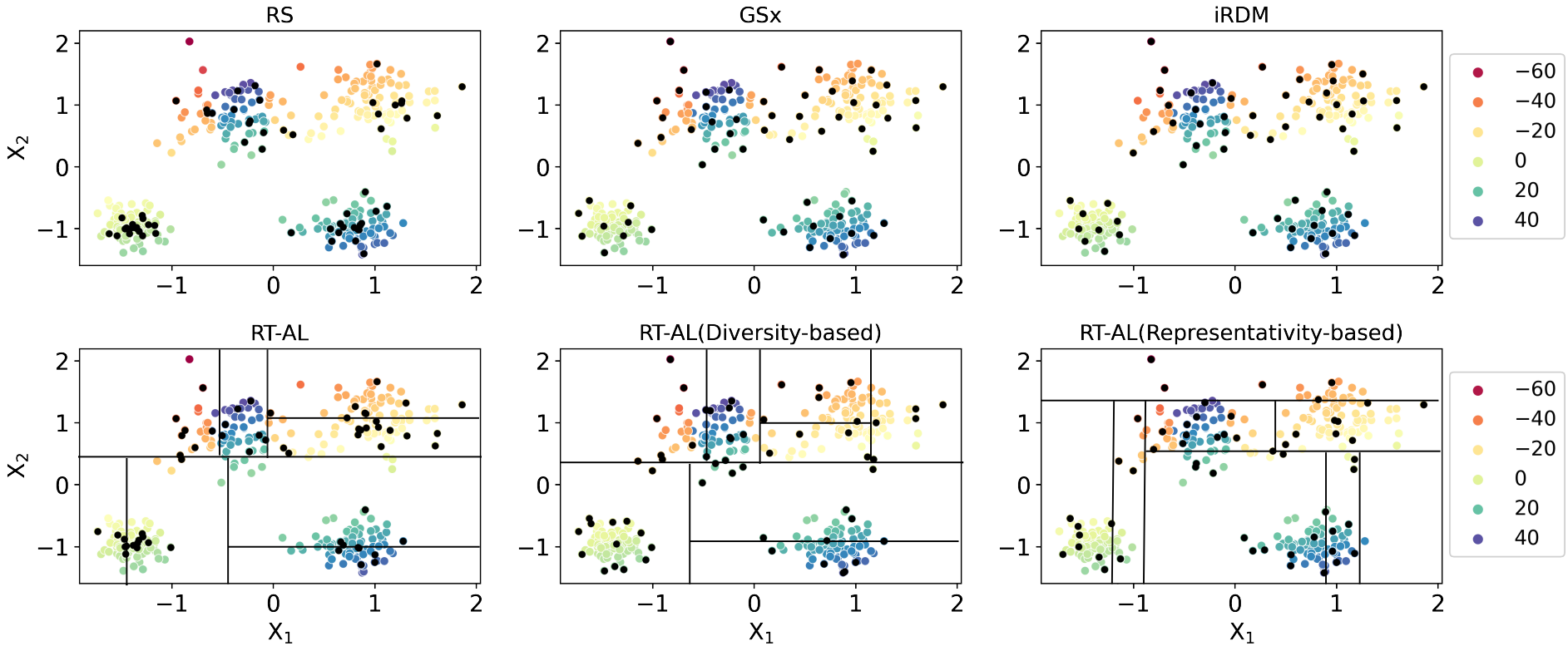


Figure – Comparison of the samples selected for labeling (shown as black dots) by our method from a generated dataset with 2 features and 500 samples, using different query criteria (labeled as RT-AL, RT-AL(Diversity-based) and RT-AL(Representativity-based)), with passive sampling and model-free AL methods

Multi-class classification with partially labeled data

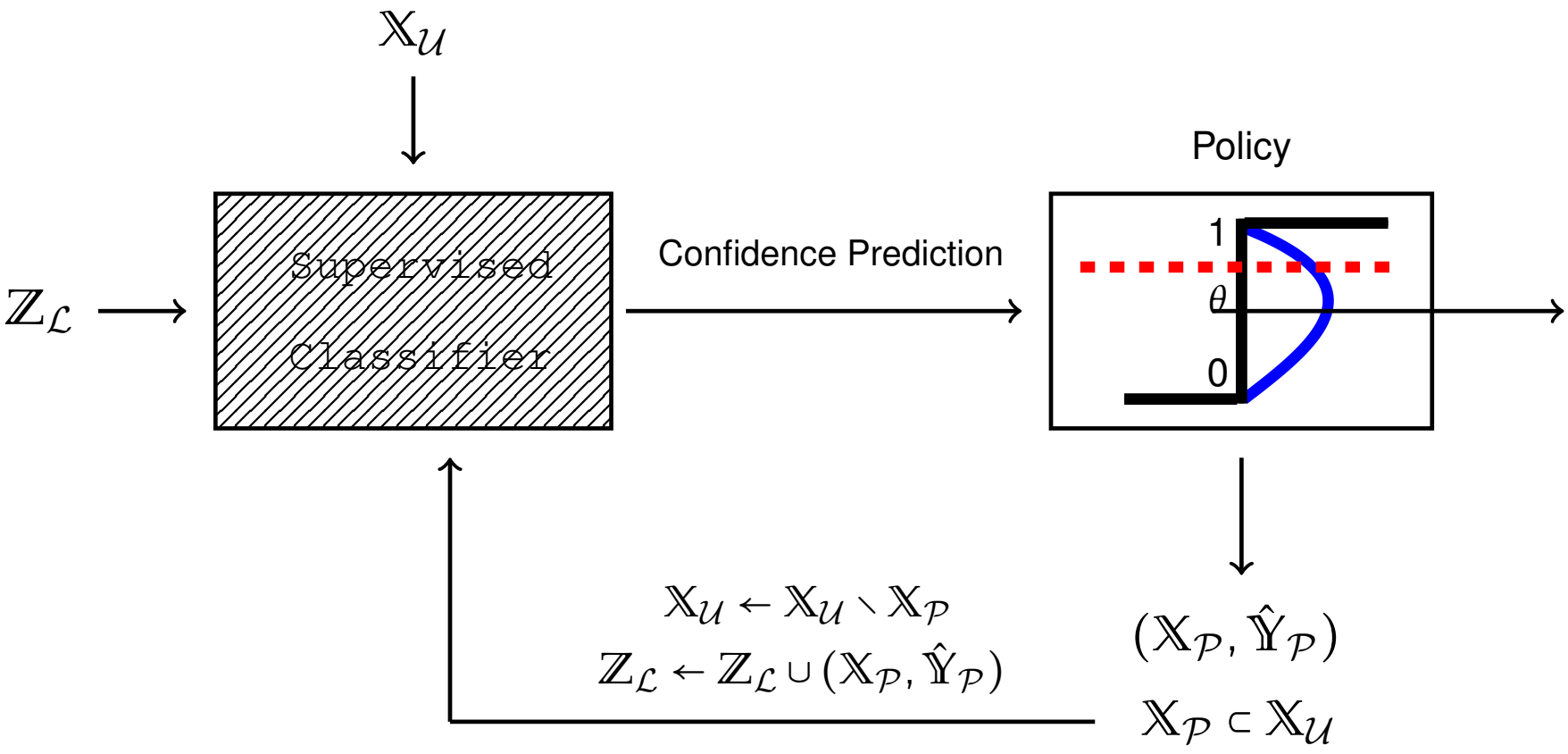


Figure – Self-training Algorithm : A supervised classifier is learned from the labeled set and is used on the unlabeled set to provide predictions for the classes. A policy determines which predictions are trusted (typically the most confident ones), and pseudo-labels are assigned to those observations based on their predictions. This process is iterated until no more points remain unlabeled.