cours 4
Apprentissage de structure de réseau bayésien

Master SID — Raisonnement dans l'incertain

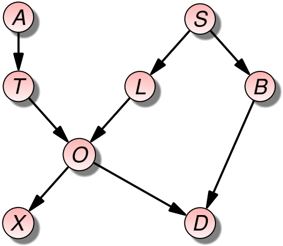
1. Algorithme fondé sur les contraintes

L'algorithme PC – Phase 1 : le squelette

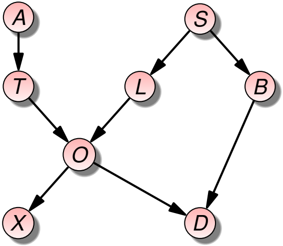
Apprentissage du squelette

$\mathcal{G} \leftarrow$ graphe non orienté complet
$d \leftarrow 0$ // taille de l'ensemble de conditionnement
$SepSet_{XY} \leftarrow \emptyset$ pour tout couple de nœuds X, Y
répéter
    pour chaque *couple* (X, Y) *t.q.* $X - Y \in \mathcal{G}$ *et* $|\mathbf{Adj}(X) \backslash \{Y\}| \geq d$ **faire**
        répéter
            Choisir un $\mathbf{Z} \subseteq \mathbf{Adj}(X) \backslash \{Y\}$ t.q. $|\mathbf{Z}| = d$
            si $X \perp\!\!\!\perp_P Y | \mathbf{Z}$ **alors**
                Supprimer l'arête $X - Y$ de $\mathcal{G}$
                $SepSet_{XY} \leftarrow \mathbf{Z}$
                break
        jusqu'à *Tous les* **Z** *de taille d ont été testés*;
    $d \leftarrow d + 1$
jusqu'à $|\mathbf{Adj}(X)| \leq d$ *pour tout nœud* X;

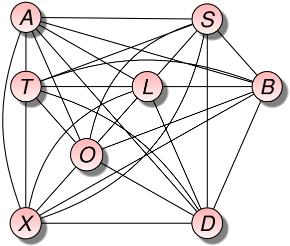
Exemple d'application (1/3)



RB réel

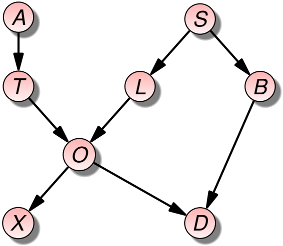


RB réel

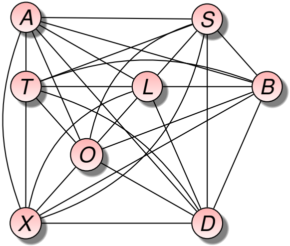


$\mathcal{G}$ complet

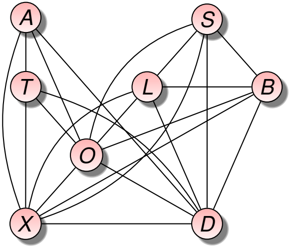
Exemple d'application (1/3)



RB réel



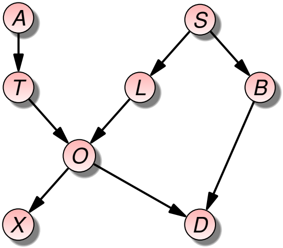
$\mathcal{G}$ complet



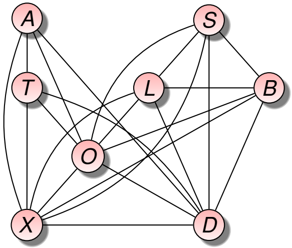
$\perp\!\!\!\perp$ ordre $d = 0$

- $A \perp\!\!\!\perp B$, $A \perp\!\!\!\perp L$, $A \perp\!\!\!\perp S$
- $B \perp\!\!\!\perp T$
- $L \perp\!\!\!\perp T$
- $S \perp\!\!\!\perp T$

Exemple d'application (2/3)

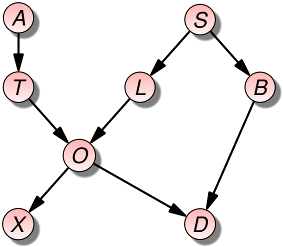


RB réel

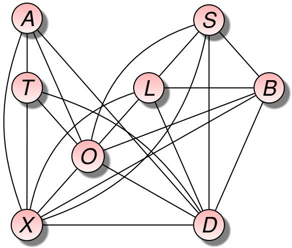


$\perp\!\!\!\perp$ ordre $d = 0$

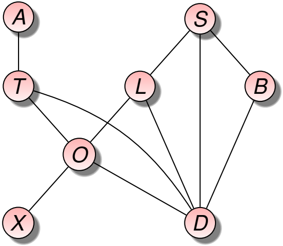
Exemple d'application (2/3)



RB réel



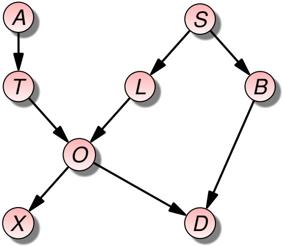
$\perp\!\!\!\perp$ ordre $d = 0$



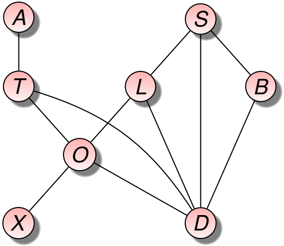
$\perp\!\!\!\perp$ ordre $d = 1$

- $A \perp\!\!\!\perp D | T$, $A \perp\!\!\!\perp O | T$
- $A \perp\!\!\!\perp X | T$, $B \perp\!\!\!\perp L | S$
- $B \perp\!\!\!\perp O | S$, $B \perp\!\!\!\perp X | S$
- $D \perp\!\!\!\perp X | O$, $L \perp\!\!\!\perp X | O$
- $O \perp\!\!\!\perp S | L$, $S \perp\!\!\!\perp X | L$
- $T \perp\!\!\!\perp X | O$

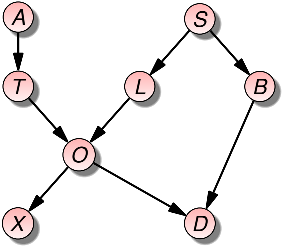
Exemple d'application (3/3)



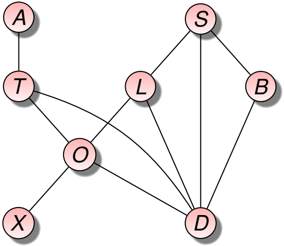
RB réel



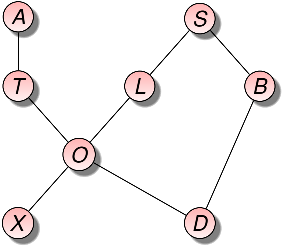
$\perp\!\!\!\perp$ ordre $d = 1$



RB réel



$\perp\!\!\!\perp$ ordre $d = 1$



$\perp\!\!\!\perp$ ordre $d = 2$

- $D \perp\!\!\!\perp L | \{O, B\}$
- $D \perp\!\!\!\perp S | \{O, B\}$
- $D \perp\!\!\!\perp T | \{O, S\}$

Définition

- unshielded triple : triplet $\langle X, Z, Y\rangle$ de nœuds de $\mathcal{G}$ t.q. :
  - $X - Z - Y \in \mathcal{G}$
  - $X - Y \notin \mathcal{G}$
- équivalent non orienté d'une v-structure

L'algorithme PC – Phase 2 : l'orientation (1/2)

Définition

- unshielded triple : triplet $\langle X, Z, Y \rangle$ de nœuds de $\mathcal{G}$ t.q. :
  - $X - Z - Y \in \mathcal{G}$
  - $X - Y \notin \mathcal{G}$
- équivalent non orienté d'une v-structure

Orientation des arêtes

// 1. Orientation des v-structures
$\mathcal{G}_{PDAG} \leftarrow \mathcal{G}$
pour chaque *unshielded triple* $\langle X, Z, Y \rangle$ *de* $\mathcal{G}$ **faire**
  si $Z \notin SepSet_{XY}$ **alors**
    *R*0 : dans $\mathcal{G}_{PDAG}$, remplacer $X - Z$ et $Z - Y$ par $X \rightarrow Z$ et $Z \leftarrow Y$

L'algorithme PC – Phase 2 : l'orientation (1/2)

Définition

- unshielded triple : triplet $\langle X, Z, Y \rangle$ de nœuds de $\mathcal{G}$ t.q. :
  - $X - Z - Y \in \mathcal{G}$
  - $X - Y \notin \mathcal{G}$
- équivalent non orienté d'une v-structure

Orientation des arêtes

// 1. Orientation des v-structures
$\mathcal{G}_{PDAG} \leftarrow \mathcal{G}$
pour chaque *unshielded triple* $\langle X, Z, Y \rangle$ *de* $\mathcal{G}$ **faire**
  si $Z \notin SepSet_{XY}$ **alors**
    $R0$: dans $\mathcal{G}_{PDAG}$, remplacer $X - Z$ et $Z - Y$ par $X \rightarrow Z$ et $Z \leftarrow Y$

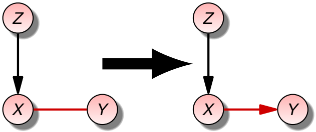
$\Longrightarrow \mathcal{G}_{PDAG} =$ squelette $+$ v-structures = pattern

L'algorithme PC – Phase 2 : les propagations (1/2)

4 règles de propagation :

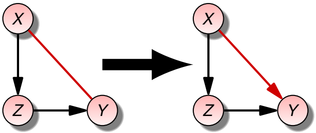
pas de v-structure :

R1 :



pas de circuit :

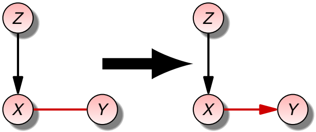
R2 :



4 règles de propagation :

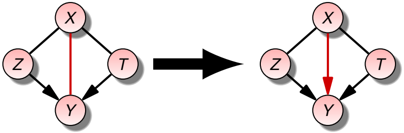
pas de v-structure :

R1 :



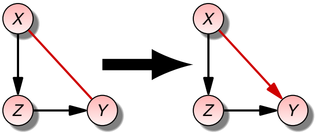
ni circuit ni v-structure :

R3 :



pas de circuit :

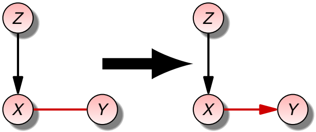
R2 :



4 règles de propagation :

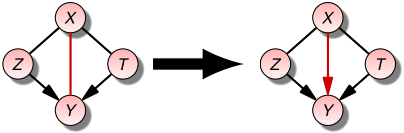
pas de v-structure :

R1 :



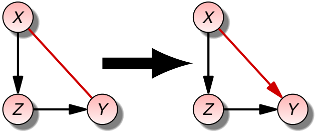
ni circuit ni v-structure :

R3 :



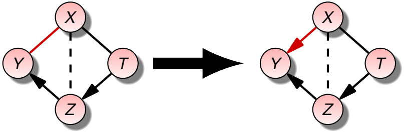
pas de circuit :

R2 :



ni circuit ni v-structure :

R4 :



L'algorithme PC – Phase 2 : les propagations (2/2)

- CPDAG $=$ completed partially directed acyclic graph
  $=$ représentant de la classe d'équivalence de Markov

L'algorithme PC – Phase 2 : les propagations (2/2)

- CPDAG $=$ completed partially directed acyclic graph
  $=$ représentant de la classe d'équivalence de Markov

Théorème – Orientation soundness – Meek(95)

- Les règles R1 à R4 sont sûres (soundness) :
  pas respectées $\Longrightarrow \exists$ nouvelle v-structure ou un circuit

L'algorithme PC – Phase 2 : les propagations (2/2)

- CPDAG $=$ completed partially directed acyclic graph
  $=$ représentant de la classe d'équivalence de Markov

Théorème – Orientation soundness – Meek(95)

- Les règles R1 à R4 sont sûres (soundness) :
  pas respectées $\Longrightarrow$ $\exists$ nouvelle v-structure ou un circuit

Théorème – Orientation completeness – Meek(95)

- Appliquer R1, R2, R3 sur un pattern $\Longrightarrow$ CPDAG

L'algorithme PC – Phase 2 : les propagations (2/2)

- CPDAG $=$ completed partially directed acyclic graph
  $=$ représentant de la classe d'équivalence de Markov

Théorème – Orientation soundness – Meek(95)

- Les règles R1 à R4 sont sûres (soundness) :
  pas respectées $\Longrightarrow$ $\exists$ nouvelle v-structure ou un circuit

Théorème – Orientation completeness – Meek(95)

- Appliquer R1, R2, R3 sur un pattern $\Longrightarrow$ CPDAG

Théorème – Completeness with background knowledge – Meek(95)

- Background knowledge $\mathcal{K}$: ensemble d'arcs interdits $+$
  ensemble d'arcs oblogatoires
- Appliquer R1 à R4 sur un pattern et orienter les arêtes
  selon $\mathcal{K}$ $\Longrightarrow$ CPDAG

Orientation des arêtes

// Orientation des v-structures

$\mathcal{G}_{PDAG} \leftarrow \mathcal{G}$

pour chaque *unshielded triple* $\langle X, Z, Y \rangle$ *de* $\mathcal{G}$ **faire**

  si $Z \notin SepSet_{XY}$ **alors**

    *R*0 : dans $\mathcal{G}_{PDAG}$, remplacer $X - Z$ et $Z - Y$ par $X \rightarrow Z$ et $Z \leftarrow Y$

L'algorithme PC – Phase 2 : l'orientation (2/2)

Orientation des arêtes

// Orientation des v-structures

$\mathcal{G}_{PDAG} \leftarrow \mathcal{G}$

pour chaque *unshielded triple* $\langle X, Z, Y \rangle$ *de* $\mathcal{G}$ **faire**

  si $Z \notin SepSet_{XY}$ **alors**

    $R0$: dans $\mathcal{G}_{PDAG}$, remplacer $X - Z$ et $Z - Y$ par $X \rightarrow Z$ et $Z \leftarrow Y$

// Propagations

répéter

  pour chaque *arête* $X - Y \in \mathcal{G}_{PDAG}$ **faire**

    si $(X - Y)$ *arête rouge d'une règle R1, R2, R3 ou R4* **alors**

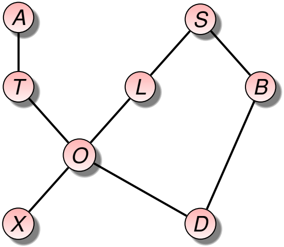
      Orienter l'arête selon la règle

jusqu'à *ce que plus aucune arête ne puisse être orientée*;

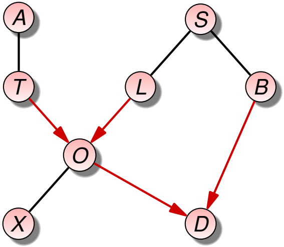
pour chaque *unshielded triple* $\langle X, Z, Y \rangle$ *de* $\mathcal{G}$ **faire**
  si $Z \notin SepSet_{XY}$ **alors**
    $R0$: dans $\mathcal{G}_{PDAG}$, remplacer $X - Z$ et $Z - Y$ par $X \rightarrow Z$ et $Z \leftarrow Y$

Sepsets :

- $S_{AB} = S_{AL} = S_{AS} = \emptyset$ $S_{AD} = S_{AO} = S_{AX} = \{T\}$
- $S_{BT} = \emptyset$ $S_{BL} = S_{BO} = S_{BX} = \{S\}$
- $S_{DL} = S_{DS} = \{O, B\}$ $S_{DT} = \{O, S\}$ $S_{DX} = \{O\}$
- $S_{LT} = \emptyset$ $S_{LX} = \{O\}$ $S_{OS} = \{L\}$
- $S_{ST} = \emptyset$ $S_{SX} = \{L\}$ $S_{TX} = \{O\}$

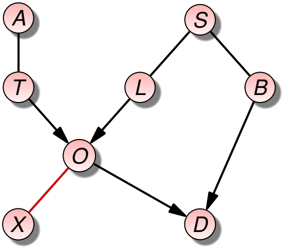


Après phase 1

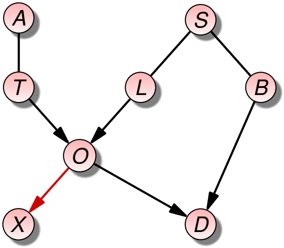


Après $R0$

// Propagation sans rajouter de v-structure
pour chaque *arête* $(X, Y) \in \mathcal{G}_{PDAG}$ **faire**
  si $(X - Y)$ *arête rouge de R1 ($Z \rightarrow X$ et $X - Y$)* **alors**
    remplacer $X - Y$ par $X \rightarrow Y$



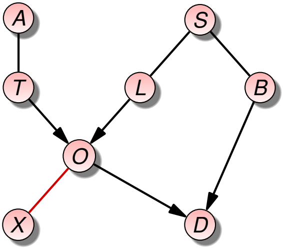
Après $R0$



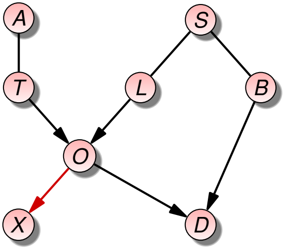
Après $R1$

Exemple d'application (2/2)

// Propagation sans rajouter de v-structure
pour chaque *arête* $(X, Y) \in \mathcal{G}_{PDAG}$ **faire**
    si $(X - Y)$ *arête rouge de R1* $(Z \rightarrow X$ *et* $X - Y)$ **alors**
        remplacer $X - Y$ par $X \rightarrow Y$



Après $R0$

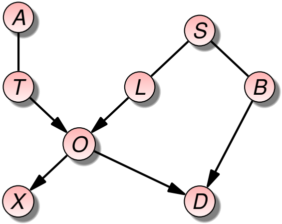


Après $R1$

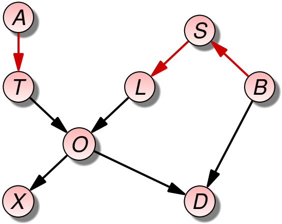
Résultat : représentant de la classe d'équivalence de Markov

- CPDAG : completed partially directed acyclic graph

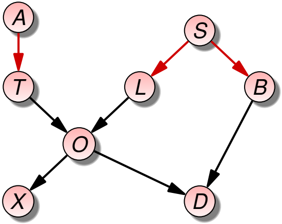
6 possibilités :



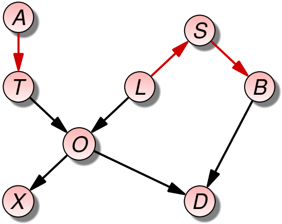
Après $R1$



1ère possibilité



2ème possibilité



3ème possibilité

- 3 autres possibilités similaires mais avec $T \rightarrow A$

Tests d'indépendance conditionnelle (1/2)

- X, Y : 2 variables aléatoires, $\mathbf{Z}$: ensemble de variables
- $N_{xy\mathbf{z}}$: nombre d'occurrences de $(X = x, Y = y, \mathbf{Z} = \mathbf{z})$ dans $\mathbf{D}$

Tests d'indépendance conditionnelle (1/2)

- X, Y : 2 variables aléatoires, $\mathbf{Z}$: ensemble de variables
- $N_{xy\mathbf{z}}$: nombre d'occurrences de $(X = x, Y = y, \mathbf{Z} = \mathbf{z})$ dans **D**
- $N_{x\mathbf{z}} = \sum_{y \in \Omega_Y} N_{xy\mathbf{z}}$ $\quad N_{y\mathbf{z}} = \sum_{x \in \Omega_X} N_{xy\mathbf{z}}$ $\quad N_{\mathbf{z}} = \sum_{x \in \Omega_X} N_{x\mathbf{z}}$

Test du χ^2

$$\chi^2_{statistics}(X, Y|\mathbf{Z}) = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} \sum_{\mathbf{z} \in \Omega_{\mathbf{Z}}} \frac{\left(N_{xy\mathbf{z}} - \frac{N_{x\mathbf{z}} N_{y\mathbf{z}}}{N_{\mathbf{z}}}\right)^2}{\frac{N_{x\mathbf{z}} N_{y\mathbf{z}}}{N_{\mathbf{z}}}}$$

- X, Y : 2 variables aléatoires, $\mathbf{Z}$: ensemble de variables
- $N_{xy\mathbf{z}}$: nombre d'occurrences de $(X = x, Y = y, \mathbf{Z} = \mathbf{z})$ dans $\mathbf{D}$
- $N_{x\mathbf{z}} = \sum_{y \in \Omega_Y} N_{xy\mathbf{z}}$ $\qquad N_{y\mathbf{z}} = \sum_{x \in \Omega_X} N_{xy\mathbf{z}}$ $\qquad N_{\mathbf{z}} = \sum_{x \in \Omega_X} N_{x\mathbf{z}}$

Test du χ^2

$$\chi^2_{statistics}(X, Y|\mathbf{Z}) = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} \sum_{\mathbf{z} \in \Omega_{\mathbf{Z}}} \frac{\left(N_{xy\mathbf{z}} - \frac{N_{x\mathbf{z}} N_{y\mathbf{z}}}{N_{\mathbf{z}}}\right)^2}{\frac{N_{x\mathbf{z}} N_{y\mathbf{z}}}{N_{\mathbf{z}}}}$$

- Nb degrés de liberté : $df = (|\Omega_X| - 1) \times (|\Omega_Y| - 1) \times |\Omega_{\mathbf{Z}}|$
- α : niveau de risque (souvent 5%)
- Si $\chi^2_{statistics}(X, Y|\mathbf{Z}) \leq \chi^2(df, \alpha)$ alors $X \perp\!\!\!\perp Y|\mathbf{Z}$

Règle usuelle : ne faire le test que si $N_{xy\mathbf{z}} \geq 5$ pour tout $x, y, \mathbf{z}$

Test du G^2

$$G^2_{statistics}(X, Y|\mathbf{Z}) = 2 \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} \sum_{\mathbf{z} \in \Omega_{\mathbf{Z}}} N_{xy\mathbf{z}} \ln \frac{N_{xy\mathbf{z}} N_{\mathbf{z}}}{N_{x\mathbf{z}} N_{y\mathbf{z}}}$$

Tests d'indépendance conditionnelle (2/2)

Test du G^2

$$G^2_{statistics}(X, Y|\mathbf{Z}) = 2 \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} \sum_{\mathbf{z} \in \Omega_{\mathbf{Z}}} N_{xy\mathbf{z}} \ln \frac{N_{xy\mathbf{z}} N_{\mathbf{z}}}{N_{x\mathbf{z}} N_{y\mathbf{z}}}$$

- Nb degrés de liberté : $df = (|\Omega_X| - 1) \times (|\Omega_Y| - 1) \times |\Omega_{\mathbf{Z}}|$
- α : niveau de risque (e.g. 5%)
- Si $G^2_{statistics}(X, Y|\mathbf{Z}) \leq \chi^2(df, \alpha)$ alors $X \perp\!\!\!\perp Y|\mathbf{Z}$

- En pratique, tests du G^2 plus robustes que ceux du χ^2

Problèmes de l'algorithme PC

- Résultat dépendant de l'ordre des calculs :

pour chaque *arête* $X - Y$ *t.q.* $|\textbf{Adj}(X)\backslash\{Y\}| \geq d$ **faire**
    $\ldots$
    Supprimer l'arête $X - Y$ de $\mathcal{G}$
    $\ldots$

Problèmes de l'algorithme PC

- Résultat dépendant de l'ordre des calculs :

pour chaque *arête* $X - Y$ *t.q.* $|\mathbf{Adj}(X)\backslash\{Y\}| \geq d$ **faire**
    ...
    Supprimer l'arête $X - Y$ de $\mathcal{G}$
    ...

$\Longrightarrow$ PC-stable, variante de Colombo et Maathuis (2014)

Problèmes de l'algorithme PC

- Résultat dépendant de l'ordre des calculs :

  pour chaque *arête* $X - Y$ *t.q.* $|\mathbf{Adj}(X) \backslash \{Y\}| \geq d$ **faire**

  ...

  Supprimer l'arête $X - Y$ de $\mathcal{G}$

  ...

  $\Longrightarrow$ PC-stable, variante de Colombo et Maathuis (2014)

- $|\mathbf{D}|$ petite $\Longrightarrow$ tests χ^2 et G^2 peu fiables

Problèmes de l'algorithme PC

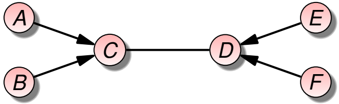
- Résultat dépendant de l'ordre des calculs :

  pour chaque *arête* $X - Y$ *t.q.* $|\mathbf{Adj}(X) \backslash \{Y\}| \geq d$ **faire**
  ...
  Supprimer l'arête $X - Y$ de $\mathcal{G}$
  ...

  $\Longrightarrow$ PC-stable, variante de Colombo et Maathuis (2014)

- $|\mathbf{D}|$ petite $\Longrightarrow$ tests χ^2 et G^2 peu fiables

- Phase 3 : pas d'orientation possible :



Problèmes de l'algorithme PC

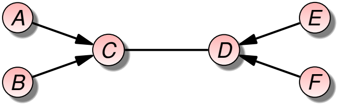
- Résultat dépendant de l'ordre des calculs :

  pour chaque *arête* $X - Y$ *t.q.* $|\mathbf{Adj}(X) \backslash \{Y\}| \geq d$ **faire**
  $\quad$ ...
  $\quad$ Supprimer l'arête $X - Y$ de $\mathcal{G}$
  $\quad$ ...

  $\Longrightarrow$ PC-stable, variante de Colombo et Maathuis (2014)

- $|\mathbf{D}|$ petite $\Longrightarrow$ tests χ^2 et G^2 peu fiables

- Phase 3 : pas d'orientation possible :



  - cause 1 : pas de DAG-faithfulness (relations déterministes ?)
    Luo (2006), Rodrigues de Morais *et al.* (2008), Mabrouk *et al.* (2014)

Problèmes de l'algorithme PC

- Résultat dépendant de l'ordre des calculs :

  pour chaque *arête* $X - Y$ *t.q.* $|\mathbf{Adj}(X) \backslash \{Y\}| \geq d$ **faire**

  ...

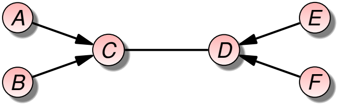
  Supprimer l'arête $X - Y$ de $\mathcal{G}$

  ...

  $\Longrightarrow$ PC-stable, variante de Colombo et Maathuis (2014)

- $|\mathbf{D}|$ petite $\Longrightarrow$ tests χ^2 et G^2 peu fiables

- Phase 3 : pas d'orientation possible :



  - cause 1 : pas de DAG-faithfulness (relations déterministes ?)
    Luo (2006), Rodrigues de Morais *et al.* (2008), Mabrouk *et al.* (2014)
  - cause 2 : erreurs dans les tests statistiques

Problèmes de l'algorithme PC

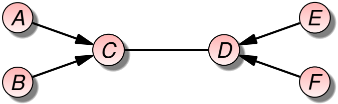
- Résultat dépendant de l'ordre des calculs :

  pour chaque *arête* $X - Y$ *t.q.* $|\mathbf{Adj}(X)\backslash\{Y\}| \geq d$ **faire**
  $\quad$ ...
  $\quad$ Supprimer l'arête $X - Y$ de $\mathcal{G}$
  $\quad$ ...

  $\Longrightarrow$ PC-stable, variante de Colombo et Maathuis (2014)

- $|\mathbf{D}|$ petite $\Longrightarrow$ tests χ^2 et G^2 peu fiables

- Phase 3 : pas d'orientation possible :



  - cause 1 : pas de DAG-faithfulness (relations déterministes ?) Luo (2006), Rodrigues de Morais *et al.* (2008), Mabrouk *et al.* (2014)
  - cause 2 : erreurs dans les tests statistiques
  - cause 3 : présence de variables latentes (non observées)

  $\Longrightarrow$ IC* Verma (1993), FCI Spirtes, Glymour et Scheines (2000)

2 Apprentissage à base de scores

Apprentissage fondé sur les scores

- Meilleure structure : celle qui colle le mieux aux données

Apprentissage fondé sur les scores

Meilleure structure : celle qui colle le mieux aux données

$\Longrightarrow$ 1 vraisemblance : $\mathcal{G}^* = \text{Argmax}_{\mathcal{G}}\, \mathcal{L}(\mathcal{G} : \mathbf{D})$

Apprentissage fondé sur les scores

Meilleure structure : celle qui colle le mieux aux données

$\Longrightarrow$ 1. vraisemblance : $\mathcal{G}^* = \text{Argmax}_{\mathcal{G}}\, \mathcal{L}(\mathcal{G} : \mathbf{D})$

2. $\mathcal{G}^*$ = structure choisie avec le moins de « regret »

Apprentissage fondé sur les scores

Meilleure structure : celle qui colle le mieux aux données

$\Longrightarrow$ 1 vraisemblance : $\mathcal{G}^* = \text{Argmax}_{\mathcal{G}}\, \mathcal{L}(\mathcal{G} : \mathbf{D})$

2 $\mathcal{G}^* =$ structure choisie avec le moins de « regret »

$\Longrightarrow$ différents scores

Apprentissage fondé sur les scores

Meilleure structure : celle qui colle le mieux aux données

$\Longrightarrow$ 1. vraisemblance : $\mathcal{G}^* = \text{Argmax}_{\mathcal{G}}\, \mathcal{L}(\mathcal{G} : \mathbf{D})$

2. $\mathcal{G}^* =$ structure choisie avec le moins de « regret »

$\Longrightarrow$ différents scores

Propriétés souhaitables des scores

- **Rasoir d'Occam :**
  Privilégier les $\mathcal{G}$ simples plutôt que complexes

Apprentissage fondé sur les scores

Meilleure structure : celle qui colle le mieux aux données

$\Longrightarrow$ ❶ vraisemblance : $\mathcal{G}^* = \text{Argmax}_{\mathcal{G}}\, \mathcal{L}(\mathcal{G} : \mathbf{D})$

❷ $\mathcal{G}^* =$ structure choisie avec le moins de « regret »

$\Longrightarrow$ différents scores

Propriétés souhaitables des scores

- **Rasoir d'Occam :**
  Privilégier les $\mathcal{G}$ simples plutôt que complexes
- **Consistance locale :**
  Ajouter un arc « utile » devrait augmenter le score
  Ajouter un arc « inutile » devrait diminuer le score

Meilleure structure : celle qui colle le mieux aux données

$\Longrightarrow$ ❶ vraisemblance : $\mathcal{G}^* = \text{Argmax}_{\mathcal{G}} \mathcal{L}(\mathcal{G} : \mathbf{D})$

❷ $\mathcal{G}^*$ = structure choisie avec le moins de « regret »

$\Longrightarrow$ différents scores

Propriétés souhaitables des scores

- **Rasoir d'Occam :**
  Privilégier les $\mathcal{G}$ simples plutôt que complexes
- **Consistance locale :**
  Ajouter un arc « utile » devrait augmenter le score
  Ajouter un arc « inutile » devrait diminuer le score
- **Score équivalence :**
  2 RB Markov-équivalents devraient avoir le même score

Apprentissage fondé sur les scores

Meilleure structure : celle qui colle le mieux aux données

$\Longrightarrow$ ❶ vraisemblance : $\mathcal{G}^* = \text{Argmax}_{\mathcal{G}}\, \mathcal{L}(\mathcal{G} : \mathbf{D})$

❷ $\mathcal{G}^* =$ structure choisie avec le moins de « regret »

$\Longrightarrow$ différents scores

Propriétés souhaitables des scores

- **Rasoir d'Occam :**
  Privilégier les $\mathcal{G}$ simples plutôt que complexes
- **Consistance locale :**
  Ajouter un arc « utile » devrait augmenter le score
  Ajouter un arc « inutile » devrait diminuer le score
- **Score équivalence :**
  2 RB Markov-équivalents devraient avoir le même score
- **Décomposition locale :**
  L'ajout/retrait d'un arc $\Longrightarrow$ score mis à jour en ne regardant que la partie de $\mathcal{G}$ autour de l'arc

En route vers Greedy Hill Climbing

Voisinage d'un DAG $\mathcal{G}$

Voisinage de $\mathcal{G}$ = ensemble des graphes $\mathcal{G}'$ t.q. :

- $\mathcal{G}'$ est un DAG
- $\mathcal{G}'$ s'obtient en appliquant à $\mathcal{G}$ un opérateur parmi :
  - l'ajout d'un arc
  - la suppression d'un arc
  - le retournement d'un arc

On note $\mathcal{N}(\mathcal{G})$ le voisinage de $\mathcal{G}$

En route vers Greedy Hill Climbing

Voisinage d'un DAG $\mathcal{G}$

Voisinage de $\mathcal{G}$ = ensemble des graphes $\mathcal{G}'$ t.q. :

- $\mathcal{G}'$ est un DAG
- $\mathcal{G}'$ s'obtient en appliquant à $\mathcal{G}$ un opérateur parmi :
  - l'ajout d'un arc
  - la suppression d'un arc
  - le retournement d'un arc

On note $\mathcal{N}(\mathcal{G})$ le voisinage de $\mathcal{G}$

1. Parcourir l'espace des DAG en partant d'un graphe $\mathcal{G}_0$

En route vers Greedy Hill Climbing

Voisinage d'un DAG $\mathcal{G}$

Voisinage de $\mathcal{G}$ = ensemble des graphes $\mathcal{G}'$ t.q. :

- $\mathcal{G}'$ est un DAG
- $\mathcal{G}'$ s'obtient en appliquant à $\mathcal{G}$ un opérateur parmi :
  - l'ajout d'un arc
  - la suppression d'un arc
  - le retournement d'un arc

On note $\mathcal{N}(\mathcal{G})$ le voisinage de $\mathcal{G}$

1. Parcourir l'espace des DAG en partant d'un graphe $\mathcal{G}_0$
2. en trouvant dans son voisinage le graphe $\mathcal{G}'$ de score plus élevé

En route vers Greedy Hill Climbing

Voisinage d'un DAG $\mathcal{G}$

Voisinage de $\mathcal{G}$ = ensemble des graphes $\mathcal{G}'$ t.q. :

- $\mathcal{G}'$ est un DAG
- $\mathcal{G}'$ s'obtient en appliquant à $\mathcal{G}$ un opérateur parmi :
  - l'ajout d'un arc
  - la suppression d'un arc
  - le retournement d'un arc

On note $\mathcal{N}(\mathcal{G})$ le voisinage de $\mathcal{G}$

1. Parcourir l'espace des DAG en partant d'un graphe $\mathcal{G}_0$
2. en trouvant dans son voisinage le graphe $\mathcal{G}'$ de score plus élevé
3. en itérant le processus avec $\mathcal{G}'$ jusqu'à un optimum (local)

En route vers Greedy Hill Climbing

Voisinage d'un DAG $\mathcal{G}$

Voisinage de $\mathcal{G}$ = ensemble des graphes $\mathcal{G}'$ t.q. :

- $\mathcal{G}'$ est un DAG
- $\mathcal{G}'$ s'obtient en appliquant à $\mathcal{G}$ un opérateur parmi :
  - l'ajout d'un arc
  - la suppression d'un arc
  - le retournement d'un arc

On note $\mathcal{N}(\mathcal{G})$ le voisinage de $\mathcal{G}$

1. Parcourir l'espace des DAG en partant d'un graphe $\mathcal{G}_0$
2. en trouvant dans son voisinage le graphe $\mathcal{G}'$ de score plus élevé
3. en itérant le processus avec $\mathcal{G}'$ jusqu'à un optimum (local)

$\Longrightarrow$ Algorithme « Greedy Hill Climbing »

Greedy Hill Climbing

Algorithme d'apprentissage « glouton » :

// étape 1
$\mathcal{G}_{best} \leftarrow$ graphe orienté vide (sans arc) // meilleur graphe trouvé
$sc_{best} \leftarrow$ Score($\mathcal{G}_{best}$) // score du meilleur DAG

// parcours de l'espace des DAG – étape 3
répéter
    // recherche du meilleur voisin
    $\mathcal{N} \leftarrow$ voisinage de $\mathcal{G}_{best}$ // calcul du voisinage
    trouvé $\leftarrow$ false // meilleur voisin pas encore trouvé
    // parcours du voisinage – étape 2
    pour chaque $\mathcal{G}' \in \mathcal{N}$ **faire**
        $sc' \leftarrow$ Score($\mathcal{G}'$)
        si $sc' > sc_{best}$ **alors**
            $\mathcal{G}_{best} \leftarrow \mathcal{G}'$, $sc_{best} \leftarrow sc'$
            trouvé $\leftarrow$ true
jusqu'à *trouvé* $=$ *false*;
Retourner $\mathcal{G}_{best}$ // graphe localement optimal

Exemple d'application

- Étape 1 : $\mathcal{G}_{best}$ = A B C $sc_{best} = 0$

Exemple d'application

- Étape 1 : $\mathcal{G}_{best} = A \quad B \quad C$ $\quad sc_{best} = 0$

- Voisinage $\mathcal{N}$:

$A \rightarrow B \quad C$	Score=1	$A \leftarrow B \quad C$	Score=3
$A \quad B \rightarrow C$	Score=0	$A \quad B \leftarrow C$	Score=-2
$A \rightarrow C \quad (B)$	Score=2	$A \leftarrow C \quad (B)$	Score=1

Exemple d'application

- Étape 1 : $\mathcal{G}_{best} = A \quad B \quad C \qquad sc_{best} = 0$

- Voisinage $\mathcal{N}$:

$A \rightarrow B \quad C$	Score=1	$A \leftarrow B \quad C$	Score=3
$A \quad B \rightarrow C$	Score=0	$A \quad B \leftarrow C$	Score=-2
$A \rightarrow C \quad B$	Score=2	$A \leftarrow C \quad B$	Score=1

- Nouveau graphe $\mathcal{G}_{best} = A \leftarrow B \quad C \qquad sc_{best} = 3$

Exemple d'application

- Étape 1 : $\mathcal{G}_{best} = A \quad B \quad C$ $\quad sc_{best} = 0$

- Voisinage $\mathcal{N}$:

$A \rightarrow B \quad C$	Score=1	$A \leftarrow B \quad C$	Score=3
$A \quad B \rightarrow C$	Score=0	$A \quad B \leftarrow C$	Score=-2
$A \rightarrow C, \quad B$	Score=2	$A \leftarrow C, \quad B$	Score=1

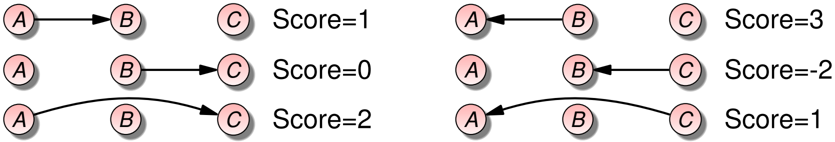
- Nouveau graphe $\mathcal{G}_{best} = A \leftarrow B \quad C$ $\quad sc_{best} = 3$

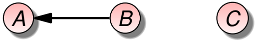
- Nouveau voisinage $\mathcal{N}$:

$A \leftarrow B \rightarrow C$	Score=4	$A \leftarrow B \leftarrow C$	Score=2
$A \leftarrow B, \quad A \rightarrow C$	Score=1	$A \leftarrow B, \quad A \leftarrow C$	Score=3
$A \rightarrow B \quad C$	Score=1	$A \quad B \quad C$	Score=0

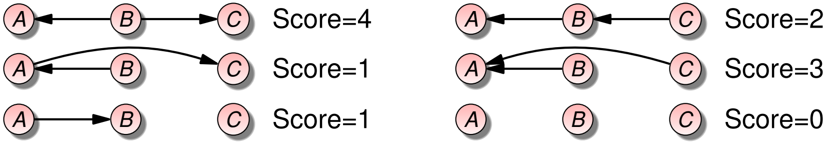
- Étape 1 : $\mathcal{G}_{best} =$  $sc_{best} = 0$

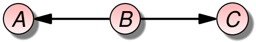
- Voisinage $\mathcal{N}$:



- Nouveau graphe $\mathcal{G}_{best} =$  $sc_{best} = 3$

- Nouveau voisinage $\mathcal{N}$:



- Nouveau graphe $\mathcal{G}_{best} =$  $sc_{best} = 4$

Exemple d'application

- Étape 1 : $\mathcal{G}_{best} = A \quad B \quad C$ $\quad sc_{best} = 0$

- Voisinage $\mathcal{N}$:

$A \rightarrow B \quad C$	Score=1	$A \leftarrow B \quad C$	Score=3
$A \quad B \rightarrow C$	Score=0	$A \quad B \leftarrow C$	Score=-2
$A \rightarrow C, \quad B$	Score=2	$A \leftarrow C, \quad B$	Score=1

- Nouveau graphe $\mathcal{G}_{best} = A \leftarrow B \quad C$ $\quad sc_{best} = 3$

- Nouveau voisinage $\mathcal{N}$:

$A \leftarrow B \rightarrow C$	Score=4	$A \leftarrow B \leftarrow C$	Score=2
$A \leftarrow B, \quad A \rightarrow C$	Score=1	$A \leftarrow B, \quad A \leftarrow C$	Score=3
$A \rightarrow B \quad C$	Score=1	$A \quad B \quad C$	Score=0

- Nouveau graphe $\mathcal{G}_{best} = A \leftarrow B \rightarrow C$ $\quad sc_{best} = 4$

- Nouveau voisinage $\mathcal{N}$:

  ......

*Score BD($\mathcal{G}$|**D**)* *[Heckerman, Geiger and Chickering (1995)]*

$$\text{Score}_{BD}(\mathcal{G}|\mathbf{D}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij} + \alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})}$$

- $\Gamma(\cdot)$: généralisation continue de la factorielle

  $\Gamma(n) = (n-1)!$ pour tout $n \in \mathbb{N}^*$

Le score BD (1/2)

*Score BD($\mathcal{G}$|**D**)* *[Heckerman, Geiger and Chickering (1995)]*

$$\mathrm{Score}_{BD}(\mathcal{G}|\mathbf{D}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij}+\alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk}+\alpha_{ijk})}{\Gamma(\alpha_{ijk})}$$

- $\Gamma(\cdot)$: généralisation continue de la factorielle

  $\Gamma(n) = (n-1)!$ pour tout $n \in \mathbb{N}^*$

- $\mathrm{Score}_{BD}(X_i|\mathbf{Pa}(X_i), \mathbf{D}) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij}+\alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk}+\alpha_{ijk})}{\Gamma(\alpha_{ijk})}$

- $\mathrm{Score}_{BD}(\mathcal{G}|\mathbf{D}) = \prod_{i=1}^{n} \mathrm{Score}_{BD}(X_i|\mathbf{Pa}(X_i), \mathbf{D})$

Le score BD (2/2)

⚠ En pratique, on calcule plutôt $\log(\text{Score}_{BD}(\mathcal{G}|\mathbf{D}))$

En pratique, on calcule plutôt $\log(\text{Score}_{BD}(\mathcal{G}|\mathbf{D}))$

Score BD($\mathcal{G}|\mathbf{D}$) *[Heckerman, Geiger and Chickering (1995)]*

- $\log(\text{Score}_{BD}(X_i|\mathbf{Pa}(X_i),\mathbf{D}))$
$$= \sum_{j=1}^{q_i} \log\Gamma(\alpha_{ij}) - \log\Gamma(N_{ij}+\alpha_{ij}) + \sum_{k=1}^{r_i} \log\Gamma(N_{ijk}+\alpha_{ijk}) - \log\Gamma(\alpha_{ijk})$$
- $\log(\text{Score}_{BD}(\mathcal{G}|\mathbf{D})) = \sum_{i=1}^{n} \log(\text{Score}_{BD}(X_i|\mathbf{Pa}(X_i),\mathbf{D}))$

Le score K2

Score K2 *[Cooper et Herskovits (1992)]*

- $\text{Score}_{K2}(\mathcal{G}|\mathbf{D}) = \log(\text{Score}_{BD}(\mathcal{G}|\mathbf{D}))$ avec :

  $\alpha_{ijk} = 1$ pour tout i, j, k

Score K2 *[Cooper et Herskovits (1992)]*

- $\text{Score}_{K2}(\mathcal{G}|\mathbf{D}) = \log(\text{Score}_{BD}(\mathcal{G}|\mathbf{D}))$ avec :

  $\alpha_{ijk} = 1$ pour tout i, j, k

  $\Longrightarrow$ revient à rajouter 1 dans chaque cellule des tableaux de contingence (de comptage)

Le score K2

Score K2 *[Cooper et Herskovits (1992)]*

- $\text{Score}_{K2}(\mathcal{G}|\mathbf{D}) = \log(\text{Score}_{BD}(\mathcal{G}|\mathbf{D}))$ avec :

  $\alpha_{ijk} = 1$ pour tout i, j, k

  $\Longrightarrow$ revient à rajouter 1 dans chaque cellule des tableaux de contingence (de comptage)

  $\Longrightarrow$ revient à rajouter $r_i q_i$ observations dans **D**

Le score K2

Score K2 *[Cooper et Herskovits (1992)]*

- $\text{Score}_{K2}(\mathcal{G}|\mathbf{D}) = \log(\text{Score}_{BD}(\mathcal{G}|\mathbf{D}))$ avec :

  $\alpha_{ijk} = 1$ pour tout i, j, k

  $\Longrightarrow$ revient à rajouter 1 dans chaque cellule des tableaux de contingence (de comptage)

  $\Longrightarrow$ revient à rajouter $r_i q_i$ observations dans $\mathbf{D}$

- $$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i), \mathbf{D}) = \sum_{j=1}^{q_i} \left[\log \left(\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \right) + \sum_{k=1}^{r_i} \log(N_{ijk}!) \right]$$

- $$\text{Score}_{K2}(\mathcal{G}|\mathbf{D}) = \sum_{i=1}^{n} \text{Score}_{K2}(X_i|\mathbf{Pa}(X_i), \mathbf{D})$$

Exemple d'application (1/2)

$$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i), \mathbf{D}) = \sum_{j=1}^{q_i} \left[\log\left(\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \right) + \sum_{k=1}^{r_i} \log(N_{ijk}!) \right]$$

Exemple d'application (1/2)

$$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i), \mathbf{D}) = \sum_{j=1}^{q_i} \left[\log \left(\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \right) + \sum_{k=1}^{r_i} \log(N_{ijk}!) \right]$$

A B C

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

$$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i),\mathbf{D}) = \sum_{j=1}^{q_i}\left[\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right) + \sum_{k=1}^{r_i}\log(N_{ijk}!)\right]$$

A B C

- $\mathbf{D} \Longrightarrow r_i = 2$ pour tout i ($|\Omega_A| = |\Omega_B| = |\Omega_C| = 2$)

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

Exemple d'application (1/2)

$$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i),\mathbf{D}) = \sum_{j=1}^{q_i}\left[\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right) + \sum_{k=1}^{r_i}\log(N_{ijk}!)\right]$$

- $\mathbf{D} \Longrightarrow r_i = 2$ pour tout i ($|\Omega_A| = |\Omega_B| = |\Omega_C| = 2$)
- $\mathcal{G} \Longrightarrow q_i = 1$ (pas de parent) et $N_{ij} = |\mathbf{D}| = 7\ \forall i$

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

Exemple d'application (1/2)

$$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i),\mathbf{D}) = \sum_{j=1}^{q_i}\left[\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right)+\sum_{k=1}^{r_i}\log(N_{ijk}!)\right]$$

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\mathbf{D} \Longrightarrow r_i = 2$ pour tout i ($|\Omega_A| = |\Omega_B| = |\Omega_C| = 2$)
- $\mathcal{G} \Longrightarrow q_i = 1$ (pas de parent) et $N_{ij} = |\mathbf{D}| = 7\ \forall i$
- $i = 1 \quad (A)$:

a_1	a_2
3	4

$r_1 = |\Omega_A| = 2$

$\mathbf{Pa}(A) = \emptyset \Longrightarrow q_i = 1$

Exemple d'application (1/2)

$$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i),\mathbf{D}) = \sum_{j=1}^{q_i}\left[\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right) + \sum_{k=1}^{r_i}\log(N_{ijk}!)\right]$$

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\mathbf{D} \Longrightarrow r_i = 2$ pour tout i $(|\Omega_A| = |\Omega_B| = |\Omega_C| = 2)$
- $\mathcal{G} \Longrightarrow q_i = 1$ (pas de parent) et $N_{ij} = |\mathbf{D}| = 7\ \forall i$
- $i = 1 \quad (A)$:

a_1	a_2
3	4

$r_1 = |\Omega_A| = 2$
$\mathbf{Pa}(A) = \emptyset \Longrightarrow q_i = 1$

$$\text{Score}_{K2}(A) = \log\left(\frac{(2-1)!}{(7+2-1)!}\right) + \log(3!) + \log(4!) \approx -5,6348$$

Exemple d'application (1/2)

$$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i),\mathbf{D}) = \sum_{j=1}^{q_i}\left[\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right) + \sum_{k=1}^{r_i}\log(N_{ijk}!)\right]$$

A B C

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\mathbf{D} \Longrightarrow r_i = 2$ pour tout i ($|\Omega_A| = |\Omega_B| = |\Omega_C| = 2$)
- $\mathcal{G} \Longrightarrow q_i = 1$ (pas de parent) et $N_{ij} = |\mathbf{D}| = 7\ \forall i$
- $i = 1 \quad (A)$:

a_1	a_2
3	4

$r_1 = |\Omega_A| = 2$
$\mathbf{Pa}(A) = \emptyset \Longrightarrow q_i = 1$

$$\text{Score}_{K2}(A) = \log\left(\frac{(2-1)!}{(7+2-1)!}\right) + \log(3!) + \log(4!) \approx -5,6348$$

- $i = 2 \quad (B)$:

b_1	b_2
3	4

$\text{Score}_{K2}(B) \approx -5,6348$

Exemple d'application (1/2)

$$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i),\mathbf{D}) = \sum_{j=1}^{q_i}\left[\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right)+\sum_{k=1}^{r_i}\log(N_{ijk}!)\right]$$

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\mathbf{D} \Longrightarrow r_i = 2$ pour tout i $(|\Omega_A| = |\Omega_B| = |\Omega_C| = 2)$
- $\mathcal{G} \Longrightarrow q_i = 1$ (pas de parent) et $N_{ij} = |\mathbf{D}| = 7\ \forall i$
- $i = 1 \quad (A)$:

a_1	a_2
3	4

$r_1 = |\Omega_A| = 2$
$\mathbf{Pa}(A) = \emptyset \Longrightarrow q_i = 1$

$$\text{Score}_{K2}(A) = \log\left(\frac{(2-1)!}{(7+2-1)!}\right) + \log(3!) + \log(4!) \approx -5,6348$$

- $i = 2 \quad (B)$:

b_1	b_2
3	4

$\text{Score}_{K2}(B) \approx -5,6348$

- $i = 3 \quad (C)$:

c_1	c_2
4	3

$\text{Score}_{K2}(C) \approx -5,6348$

Exemple d'application (1/2)

$$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i),\mathbf{D}) = \sum_{j=1}^{q_i}\left[\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right) + \sum_{k=1}^{r_i}\log(N_{ijk}!)\right]$$

A B C

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\mathbf{D} \Longrightarrow r_i = 2$ pour tout i $(|\Omega_A| = |\Omega_B| = |\Omega_C| = 2)$
- $\mathcal{G} \Longrightarrow q_i = 1$ (pas de parent) et $N_{ij} = |\mathbf{D}| = 7\ \forall i$
- $i = 1 \quad (A)$:

a_1	a_2
3	4

$r_1 = |\Omega_A| = 2$
$\mathbf{Pa}(A) = \emptyset \Longrightarrow q_i = 1$

$$\text{Score}_{K2}(A) = \log\left(\frac{(2-1)!}{(7+2-1)!}\right) + \log(3!) + \log(4!) \approx -5,6348$$

- $i = 2 \quad (B)$:

b_1	b_2
3	4

$\text{Score}_{K2}(B) \approx -5,6348$

- $i = 3 \quad (C)$:

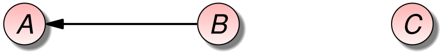
c_1	c_2
4	3

$\text{Score}_{K2}(C) \approx -5,6348$

- $\text{Score}_{K2}(\mathcal{G}) \approx -5,6348 - 5,6348 - 5,6348 = -16.9044$

Exemple d'application (2/2)

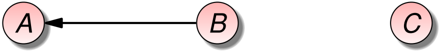
$$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i), \mathbf{D}) = \sum_{j=1}^{q_i} \left[\log\left(\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \right) + \sum_{k=1}^{r_i} \log(N_{ijk}!) \right]$$



A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

Exemple d'application (2/2)

$$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i),\mathbf{D}) = \sum_{j=1}^{q_i}\left[\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right) + \sum_{k=1}^{r_i}\log(N_{ijk}!)\right]$$



- Score$_{K2}(B)$ et Score$_{K2}(C)$ inchangés $\approx -5,6348$

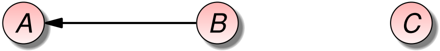
A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $(A|B)$:

	a_1	a_2	N_{ij}
b_1	2	1	3
b_2	1	3	4

$N_{ij} = \sum$ sur chaque ligne

$$\mathrm{Score}_{K2}(X_i|\mathbf{Pa}(X_i),\mathbf{D}) = \sum_{j=1}^{q_i}\left[\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right)+\sum_{k=1}^{r_i}\log(N_{ijk}!)\right]$$



- $\mathrm{Score}_{K2}(B)$ et $\mathrm{Score}_{K2}(C)$ inchangés $\approx -5,6348$

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $(A|B)$:

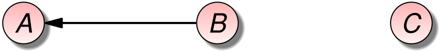
	a_1	a_2	N_{ij}
b_1	2	1	3
b_2	1	3	4

$N_{ij} = \sum$ sur chaque ligne

$$\mathrm{Score}_{K2}(A|B) = \log\left(\frac{(2-1)!}{(3+2-1)!}\right)+\log(2!)+\log(1!) \quad \text{(ligne } b_1\text{)}$$
$$+\log\left(\frac{(2-1)!}{(4+2-1)!}\right)+\log(1!)+\log(3!) \quad \text{(ligne } b_2\text{)}$$
$$\approx -6,1738$$

Exemple d'application (2/2)

$$\text{Score}_{K2}(X_i|\mathbf{Pa}(X_i),\mathbf{D}) = \sum_{j=1}^{q_i}\left[\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right) + \sum_{k=1}^{r_i}\log(N_{ijk}!)\right]$$



A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\text{Score}_{K2}(B)$ et $\text{Score}_{K2}(C)$ inchangés $\approx -5,6348$
- $(A|B)$:

	a_1	a_2	N_{ij}
b_1	2	1	3
b_2	1	3	4

$N_{ij} = \sum$ sur chaque ligne

$$\begin{aligned}\text{Score}_{K2}(A|B) &= \log\left(\tfrac{(2-1)!}{(3+2-1)!}\right) + \log(2!) + \log(1!) \quad \textcolor{red}{(\text{ligne } b_1)}\\ &+ \log\left(\tfrac{(2-1)!}{(4+2-1)!}\right) + \log(1!) + \log(3!) \quad \textcolor{red}{(\text{ligne } b_2)}\\ &\approx -6,1738\end{aligned}$$

- $\text{Score}_{K2}(\mathcal{G}) \approx -6,1738 - 5,6348 - 5,6348 = -17.4434$

$\Longrightarrow$ Graphe moins probable que celui du slide précédent

Le score BDeu

- Problème : Score K2 non score-équivalent :

  2 RB Markov-équivalents n'ont pas forcément le même score !

Score BDeu *[Buntine (1991)]*

- $\text{Score}_{BDeu}(\mathcal{G}|\mathbf{D}) = \log(\text{Score}_{BD}(\mathcal{G}|\mathbf{D}))$ avec :

  $\alpha_{ijk} = \frac{N'}{r_i q_i}$ pour tout i, j, k $\quad N' =$ « *effective sample size* »

  $\Longrightarrow$ revient à rajouter $\frac{N'}{r_i q_i}$ dans chaque cellule des tableaux de contingence (de comptage)

Le score BDeu

- Problème : Score K2 non score-équivalent :
  2 RB Markov-équivalents n'ont pas forcément le même score !

Score BDeu *[Buntine (1991)]*

- $\text{Score}_{BDeu}(\mathcal{G}|\mathbf{D}) = \log(\text{Score}_{BD}(\mathcal{G}|\mathbf{D}))$ avec :

  $\alpha_{ijk} = \frac{N'}{r_i q_i}$ pour tout i, j, k $\quad N' =$ « *effective sample size* »

  $\Longrightarrow$ revient à rajouter $\frac{N'}{r_i q_i}$ dans chaque cellule des tableaux de contingence (de comptage)

  $\Longrightarrow$ revient à rajouter N' observations dans $\mathbf{D}$

Le score BDeu

- **Problème :** Score K2 non score-équivalent :

  2 RB Markov-équivalents n'ont pas forcément le même score !

Score BDeu *[Buntine (1991)]*

- $\text{Score}_{BDeu}(\mathcal{G}|\mathbf{D}) = \log(\text{Score}_{BD}(\mathcal{G}|\mathbf{D}))$ avec :

  $\alpha_{ijk} = \frac{N'}{r_i q_i}$ pour tout i, j, k $\quad N' =$ « *effective sample size* »

  $\Longrightarrow$ revient à rajouter $\frac{N'}{r_i q_i}$ dans chaque cellule des tableaux de contingence (de comptage)

  $\Longrightarrow$ revient à rajouter N' observations dans $\mathbf{D}$

- $\text{Score}_{BDeu}(X_i|\mathbf{Pa}(X_i), \mathbf{D})$

$$= \sum_{j=1}^{q_i} \log\Gamma\left(\frac{N'}{q_i}\right) - \log\Gamma\left(N_{ij} + \frac{N'}{q_i}\right) + \sum_{k=1}^{r_i} \log\Gamma\left(N_{ijk} + \frac{N'}{r_i q_i}\right) - \log\Gamma\left(\frac{N'}{r_i q_i}\right)$$

- $\text{Score}_{BDeu}(\mathcal{G}|\mathbf{D}) = \sum_{i=1}^{n} \text{Score}_{BDeu}(X_i|\mathbf{Pa}(X_i), \mathbf{D})$

Le score BDeu

- Problème : Score K2 non score-équivalent :

  2 RB Markov-équivalents n'ont pas forcément le même score !

Score BDeu *[Buntine (1991)]*

- $\text{Score}_{BDeu}(\mathcal{G}|\mathbf{D}) = \log(\text{Score}_{BD}(\mathcal{G}|\mathbf{D}))$ avec :

  $\alpha_{ijk} = \frac{N'}{r_i q_i}$ pour tout i, j, k $\quad N' =$ « *effective sample size* »

  $\Longrightarrow$ revient à rajouter $\frac{N'}{r_i q_i}$ dans chaque cellule des tableaux de contingence (de comptage)

  $\Longrightarrow$ revient à rajouter N' observations dans $\mathbf{D}$

- $\text{Score}_{BDeu}(X_i|\mathbf{Pa}(X_i), \mathbf{D})$

$$= \sum_{j=1}^{q_i} \log\Gamma\left(\frac{N'}{q_i}\right) - \log\Gamma\left(N_{ij} + \frac{N'}{q_i}\right) + \sum_{k=1}^{r_i} \log\Gamma\left(N_{ijk} + \frac{N'}{r_i q_i}\right) - \log\Gamma\left(\frac{N'}{r_i q_i}\right)$$

- $\text{Score}_{BDeu}(\mathcal{G}|\mathbf{D}) = \sum_{i=1}^{n} \text{Score}_{BDeu}(X_i|\mathbf{Pa}(X_i), \mathbf{D})$

- BDeu : score équivalent [Heckerman, Geiger, Chickering (1995)]

Scores issus de la théorie de l'information (1/3)

- Score BD($\mathcal{G}$) : $\displaystyle\int_\Theta P(\mathbf{D}|\mathcal{G},\Theta)\pi(\Theta|\mathcal{G})d\Theta$

Scores issus de la théorie de l'information (1/3)

- Score BD($\mathcal{G}$) : $\int_\Theta P(\mathbf{D}|\mathcal{G},\Theta)\pi(\Theta|\mathcal{G})d\Theta$

  $\Longrightarrow$ vraisemblance moyenne sur $\{$RB de structure $\mathcal{G}\}$

  $\Longrightarrow$ moyenne sur tous les paramètres

Scores issus de la théorie de l'information (1/3)

- Score BD($\mathcal{G}$) : $\displaystyle\int_\Theta P(\mathbf{D}|\mathcal{G},\Theta)\pi(\Theta|\mathcal{G})d\Theta$

  $\Longrightarrow$ vraisemblance moyenne sur $\{$RB de structure $\mathcal{G}\}$

  $\Longrightarrow$ moyenne sur tous les paramètres

- Et si on utilisait les paramètres optimaux plutôt que la moyenne ?

  Score($\mathcal{G}$) : $\max_\Theta P(\mathbf{D}|\mathcal{G},\Theta)$

Scores issus de la théorie de l'information (1/3)

- Score BD($\mathcal{G}$) : $\int_\Theta P(\mathbf{D}|\mathcal{G},\Theta)\pi(\Theta|\mathcal{G})d\Theta$

  $\Longrightarrow$ vraisemblance moyenne sur $\{$RB de structure $\mathcal{G}\}$

  $\Longrightarrow$ moyenne sur tous les paramètres

- Et si on utilisait les paramètres optimaux plutôt que la moyenne ?

  Score($\mathcal{G}$) : $\max_\Theta P(\mathbf{D}|\mathcal{G},\Theta)$

  $\Longrightarrow$ Paramètres qui « collent » le plus aux données

Scores issus de la théorie de l'information (1/3)

- Score BD($\mathcal{G}$) : $\displaystyle\int_\Theta P(\mathbf{D}|\mathcal{G},\Theta)\pi(\Theta|\mathcal{G})d\Theta$

  $\Longrightarrow$ vraisemblance moyenne sur $\{$RB de structure $\mathcal{G}\}$

  $\Longrightarrow$ moyenne sur tous les paramètres

- Et si on utilisait les paramètres optimaux plutôt que la moyenne ?

  Score($\mathcal{G}$) : $\max_\Theta P(\mathbf{D}|\mathcal{G},\Theta)$

  $\Longrightarrow$ Paramètres qui « collent » le plus aux données

  $\Longrightarrow$ Apprentissage de paramètres par max de vraisemblance

Scores issus de la théorie de l'information (1/3)

- Score BD($\mathcal{G}$) : $\displaystyle\int_{\Theta} P(\mathbf{D}|\mathcal{G},\Theta)\pi(\Theta|\mathcal{G})d\Theta$

  $\Longrightarrow$ vraisemblance moyenne sur $\{$RB de structure $\mathcal{G}\}$

  $\Longrightarrow$ moyenne sur tous les paramètres

- Et si on utilisait les paramètres optimaux plutôt que la moyenne ?

  Score($\mathcal{G}$) : $\max_{\Theta} P(\mathbf{D}|\mathcal{G},\Theta)$

  $\Longrightarrow$ Paramètres qui « collent » le plus aux données

  $\Longrightarrow$ Apprentissage de paramètres par max de vraisemblance

Fondement des scores issus de la théorie de l'information

Scores issus de la théorie de l'information (2/3)

- Optimisation des paramètres de $\mathcal{G}$ par max de vraisemblance
  $\Longrightarrow$ score « log-likelihood »

- Optimisation des paramètres de $\mathcal{G}$ par max de vraisemblance $\Longrightarrow$ score « log-likelihood »

Score log-likelihood (LL)

- $\text{Score}_{LL}(\mathcal{G}|\mathbf{D}) = \sum_{i=1}^{n} \text{Score}_{LL}(X_i|\mathbf{Pa}(X_i), \mathbf{D})$
- $\text{Score}_{LL}(X_i|\mathbf{Pa}(X_i), \mathbf{D}) = \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log\left(\frac{N_{ijk}}{N_{ij}}\right)$

- Optimisation des paramètres de $\mathcal{G}$ par max de vraisemblance
  $\Longrightarrow$ score « log-likelihood »

Score log-likelihood (LL)

- $\text{Score}_{LL}(\mathcal{G}|\mathbf{D}) = \sum_{i=1}^{n} \text{Score}_{LL}(X_i|\mathbf{Pa}(X_i), \mathbf{D})$
- $\text{Score}_{LL}(X_i|\mathbf{Pa}(X_i), \mathbf{D}) = \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log\left(\frac{N_{ijk}}{N_{ij}}\right)$

⚠ En pratique, apprend des graphes trop denses
$\Longrightarrow$ « Sur-apprentissage »
$\Longrightarrow$ Rajouter une pénalité s'il y a trop d'arcs

Score MDL (minimum description length)

- $\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \frac{1}{2}\log(N)|\mathcal{G}|$
- $|\mathcal{G}| = \sum_{i=1}^{n}(r_i - 1) \times q_i$

Score MDL (minimum description length)

- $\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \frac{1}{2}\log(N)|\mathcal{G}|$
- $|\mathcal{G}| = \sum_{i=1}^{n}(r_i - 1) \times q_i$: nb de paramètres θ_{ijk} à choisir

Scores issus de la théorie de l'information (3/3)

Score MDL (minimum description length)

- $\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \frac{1}{2}\log(N)|\mathcal{G}|$
- $|\mathcal{G}| = \sum_{i=1}^{n}(r_i - 1) \times q_i$: nb de paramètres θ_{ijk} à choisir
- $-\frac{1}{2}\log(N)|\mathcal{G}|$: pénalité

  Permet de minimiser la taille mémoire pour stocker le RB

Scores issus de la théorie de l'information (3/3)

Score MDL (minimum description length)

- $Score_{MDL}(\mathcal{G}|\mathbf{D}) = Score_{LL}(\mathcal{G}|\mathbf{D}) - \frac{1}{2}\log(N)|\mathcal{G}|$
- $|\mathcal{G}| = \sum_{i=1}^{n}(r_i - 1) \times q_i$: nb de paramètres θ_{ijk} à choisir
- $-\frac{1}{2}\log(N)|\mathcal{G}|$: pénalité
  Permet de minimiser la taille mémoire pour stocker le RB

Score BIC (Bayesian Information Criterion)

- S'appuie sur le critère BIC [Schwarz (1978)]
- $Score_{BIC}(\mathcal{G}|\mathbf{D}) = Score_{MDL}(\mathcal{G}|\mathbf{D})$

Scores issus de la théorie de l'information (3/3)

Score MDL (minimum description length)

- $\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \frac{1}{2}\log(N)|\mathcal{G}|$
- $|\mathcal{G}| = \sum_{i=1}^{n}(r_i - 1) \times q_i$: nb de paramètres θ_{ijk} à choisir
- $-\frac{1}{2}\log(N)|\mathcal{G}|$: pénalité

  Permet de minimiser la taille mémoire pour stocker le RB

Score BIC (Bayesian Information Criterion)

- S'appuie sur le critère BIC [Schwarz (1978)]
- $\text{Score}_{BIC}(\mathcal{G}|\mathbf{D}) = \text{Score}_{MDL}(\mathcal{G}|\mathbf{D})$

Score AIC (Akaike Information Criterion)

- S'appuie sur le critère d'information d'Akaike (1973)
- $\text{Score}_{AIC}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - |\mathcal{G}|$

Exemple d'application (1/2)

$$\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \frac{1}{2}\log(N)|\mathcal{G}| \quad \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) = \sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log\left(\frac{N_{ijk}}{N_{ij}}\right)$$

Exemple d'application (1/2)

$$\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \frac{1}{2}\log(N)|\mathcal{G}| \quad \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) = \sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log\left(\frac{N_{ijk}}{N_{ij}}\right)$$

A B C

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

$$\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \frac{1}{2}\log(N)|\mathcal{G}| \quad \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) = \sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log\left(\frac{N_{ijk}}{N_{ij}}\right)$$

A B C

- $\mathbf{D} \Longrightarrow r_i = 2$, $\mathcal{G} \Longrightarrow q_i = 1$ (pas de parent)

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

Exemple d'application (1/2)

$$\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \frac{1}{2}\log(N)|\mathcal{G}| \quad \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) = \sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log\left(\frac{N_{ijk}}{N_{ij}}\right)$$

A B C

- $\mathbf{D} \Longrightarrow r_i = 2$, $\mathcal{G} \Longrightarrow q_i = 1$ (pas de parent)

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

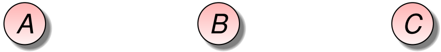
- $i = 1 \quad (A)$:

a_1	a_2	N_{ij}
3	4	7

$N_{ij} = \text{total ligne} = 3 + 4$

Exemple d'application (1/2)

$$\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \frac{1}{2}\log(N)|\mathcal{G}| \quad \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) = \sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk} \log\left(\frac{N_{ijk}}{N_{ij}}\right)$$



- $\mathbf{D} \Longrightarrow r_i = 2$, $\mathcal{G} \Longrightarrow q_i = 1$ (pas de parent)

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $i = 1 \quad (A)$:

a_1	a_2	N_{ij}
3	4	7

N_{ij} = total ligne $= 3 + 4$

$$\text{Score}_{LL}(A) = 3\log\left(\frac{3}{7}\right) + 4\log\left(\frac{4}{7}\right) \approx -4,78$$

Exemple d'application (1/2)

$$\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \frac{1}{2}\log(N)|\mathcal{G}| \quad \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) = \sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log\left(\frac{N_{ijk}}{N_{ij}}\right)$$

A B C

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\mathbf{D} \Longrightarrow r_i = 2$, $\mathcal{G} \Longrightarrow q_i = 1$ (pas de parent)
- $i = 1 \quad (A)$:

a_1	a_2	N_{ij}
3	4	7

$N_{ij} = \text{total ligne} = 3 + 4$

$$\text{Score}_{LL}(A) = 3\log\left(\frac{3}{7}\right) + 4\log\left(\frac{4}{7}\right) \approx -4,78$$

- $i = 2 \quad (B)$:

b_1	b_2
3	4

$\text{Score}_{LL}(B) \approx -4,78$

Exemple d'application (1/2)

$$\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \frac{1}{2}\log(N)|\mathcal{G}| \quad \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) = \sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log\left(\frac{N_{ijk}}{N_{ij}}\right)$$

A B C

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\mathbf{D} \Longrightarrow r_i = 2$, $\mathcal{G} \Longrightarrow q_i = 1$ (pas de parent)
- $i = 1 \quad (A)$:

a_1	a_2	N_{ij}
3	4	7

$N_{ij} =$ total ligne $= 3 + 4$

$\text{Score}_{LL}(A) = 3\log\left(\frac{3}{7}\right) + 4\log\left(\frac{4}{7}\right) \approx -4,78$

- $i = 2 \quad (B)$:

b_1	b_2
3	4

$\text{Score}_{LL}(B) \approx -4,78$

- $i = 3 \quad (C)$:

c_1	c_2
4	3

$\text{Score}_{LL}(C) \approx -4,78$

Exemple d'application (1/2)

$$\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \tfrac{1}{2}\log(N)|\mathcal{G}| \quad \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) = \sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log\left(\frac{N_{ijk}}{N_{ij}}\right)$$

A B C

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\mathbf{D} \Longrightarrow r_i = 2$, $\mathcal{G} \Longrightarrow q_i = 1$ (pas de parent)
- $i = 1 \quad (A)$:

a_1	a_2	N_{ij}
3	4	7

  $N_{ij} = \text{total ligne} = 3 + 4$

  $\text{Score}_{LL}(A) = 3\log\left(\frac{3}{7}\right) + 4\log\left(\frac{4}{7}\right) \approx -4,78$
- $i = 2 \quad (B)$:

b_1	b_2
3	4

  $\text{Score}_{LL}(B) \approx -4,78$
- $i = 3 \quad (C)$:

c_1	c_2
4	3

  $\text{Score}_{LL}(C) \approx -4,78$
- $|\mathcal{G}| = \sum_{i=1}^{3}(r_i - 1)\times q_i = (1\times 1) + (1\times 1) + (1\times 1) = 3$

Exemple d'application (1/2)

$$\text{Score}_{MDL}(\mathcal{G}|\mathbf{D}) = \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) - \tfrac{1}{2}\log(N)|\mathcal{G}| \quad \text{Score}_{LL}(\mathcal{G}|\mathbf{D}) = \sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log\left(\frac{N_{ijk}}{N_{ij}}\right)$$

A B C

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\mathbf{D} \Longrightarrow r_i = 2, \mathcal{G} \Longrightarrow q_i = 1$ (pas de parent)
- $i = 1 \quad (A)$:

a_1	a_2	N_{ij}
3	4	7

  $N_{ij} =$ total ligne $= 3 + 4$

  $\text{Score}_{LL}(A) = 3\log\left(\frac{3}{7}\right) + 4\log\left(\frac{4}{7}\right) \approx -4,78$
- $i = 2 \quad (B)$:

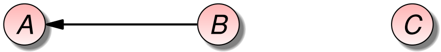
b_1	b_2
3	4

  $\text{Score}_{LL}(B) \approx -4,78$
- $i = 3 \quad (C)$:

c_1	c_2
4	3

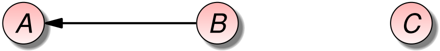
  $\text{Score}_{LL}(C) \approx -4,78$
- $|\mathcal{G}| = \sum_{i=1}^{3}(r_i - 1)\times q_i = (1\times 1) + (1\times 1) + (1\times 1) = 3$
- $\text{Score}_{MDL}(\mathcal{G}) \approx -4,78\times 3 - 0,5\times\log(7)\times 3 \approx -17.259$

$$\text{Score}_{MDL}(X_i|\mathbf{Pa}(X_i), \mathbf{D}) = \sum_{k=1}^{r_i} \left[N_{ijk} \log \left(\frac{N_{ijk}}{N_{ij}} \right) \right] - \frac{1}{2} \log(N) \times (r_i - 1) \times q_i$$



A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

$$\text{Score}_{MDL}(X_i|\mathbf{Pa}(X_i),\mathbf{D}) = \sum_{k=1}^{r_i}\left[N_{ijk}\log\left(\frac{N_{ijk}}{N_{ij}}\right)\right] - \frac{1}{2}\log(N)\times(r_i-1)\times q_i$$



A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\text{Score}_{MDL}(B)$ et $\text{Score}_{MDL}(C) \approx -4,78 - \frac{1}{2}\log(7) \approx -5.753$
- $(A|B)$:

	a_1	a_2	N_{ij}
b_1	2	1	3
b_2	1	3	4

$N_{ij} = \sum$ sur chaque ligne

Exemple d'application (2/2)

$$\text{Score}_{MDL}(X_i|\mathbf{Pa}(X_i), \mathbf{D}) = \sum_{k=1}^{r_i} \left[N_{ijk} \log \left(\frac{N_{ijk}}{N_{ij}} \right) \right] - \frac{1}{2} \log(N) \times (r_i - 1) \times q_i$$

$$\begin{tikzcd} A & & B \arrow[ll] & & C \end{tikzcd}$$

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\text{Score}_{MDL}(B)$ et $\text{Score}_{MDL}(C) \approx -4,78 - \frac{1}{2}\log(7) \approx -5.753$
- $(A|B)$:

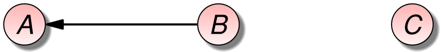
	a_1	a_2	N_{ij}
b_1	2	1	3
b_2	1	3	4

$N_{ij} = \sum$ sur chaque ligne

$$\begin{aligned} \text{Score}_{MDL}(A|B) &= 2 \times \log\left(\tfrac{2}{3}\right) + 1 \times \log\left(\tfrac{1}{3}\right) && \text{(ligne } b_1\text{)} \\ &+ 1 \times \log\left(\tfrac{1}{4}\right) + 3 \times \log\left(\tfrac{3}{4}\right) && \text{(ligne } b_2\text{)} \\ &- \tfrac{1}{2}\log(7) \times 1 \times 2 \approx -6,105 && \text{(pénalité)} \end{aligned}$$

Exemple d'application (2/2)

$$\mathrm{Score}_{MDL}(X_i|\mathbf{Pa}(X_i),\mathbf{D}) = \sum_{k=1}^{r_i}\left[N_{ijk}\log\left(\frac{N_{ijk}}{N_{ij}}\right)\right] - \frac{1}{2}\log(N)\times(r_i-1)\times q_i$$



A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_1	c_1

- $\mathrm{Score}_{MDL}(B)$ et $\mathrm{Score}_{MDL}(C) \approx -4,78 - \frac{1}{2}\log(7) \approx -5.753$
- $(A|B)$:

	a_1	a_2	N_{ij}
b_1	2	1	3
b_2	1	3	4

$N_{ij} = \sum$ sur chaque ligne

$$\begin{aligned}\mathrm{Score}_{MDL}(A|B) &= 2\times\log\left(\tfrac{2}{3}\right) + 1\times\log\left(\tfrac{1}{3}\right) && \text{(ligne } b_1\text{)}\\ &+ 1\times\log\left(\tfrac{1}{4}\right) + 3\times\log\left(\tfrac{3}{4}\right) && \text{(ligne } b_2\text{)}\\ &- \tfrac{1}{2}\log(7)\times 1\times 2 \approx -6,105 && \text{(pénalité)}\end{aligned}$$

- $\mathrm{Score}_{MDL}(\mathcal{G}) \approx -6,105 - 5.753 - 5.753 = -17,611$

$\Longrightarrow$ Graphe moins probable que celui du slide précédent

Bibliographie

- Akaike, H. (1973) « Information theory and an extension of the maximum likelihood principle », Proceedings of the 2nd International Symposium on Information Theory, 267–281
- Buntime W. (1991) « Theory refinement on Bayesian networks », Proceedings of Uncertainty in Artificial Intelligence, 52–60
- Colombo D. et Maathuis M.H. (2014) « Order-Independent Constraint-Based Causal Structure Learning », Journal of Machine Learning Research, 15 :3921–3962
- Geiger D. et Heckerman D. (1997) « A Characterization of the Dirichlet Distribution through Global and Local Parameter Independence », The Annals of Statistics, 25(3) :1344–1369
- Heckerman D., Geiger D. et Chickering D. (1995) « Learning Bayesian Networks : The Combination of Knowledge and Statistical Data », Machine Learning, 20 :197–243

Bibliographie

- Luo W. (2006) « Learning Bayesian networks in semi-deterministic systems », Proceedings of the Canadian Conference on Artificial Intelligence, 230–241
- Mabrouk A., Gonzales C., Jabet-Chevalier K. et Chojnaki E. (2014) « An Efficient Bayesian Network Structure Learning Algorithm in the Presence of Deterministic Relations », Proceedings of the European Conference on Artificial Intelligence, 567–572
- Meek C. (1995) « Causal inference and causal explanation with background knowledge », Proceedings of the Conference on Uncertainty in Artificial Intelligence, 403—410
- Pearl J. et Verma T. (1991) « A theory of inferred causation », Proceedings of the 2nd International Conference on Knowledge Representation and Reasoning, 441–452

Bibliographie

- Rodrigues de Morais, S. Aussem, A. et Corbex M. (2008) ≪ Handling almost-deterministic relationships in constraint-based Bayesian network discovery : Application to cancer risk factor identification ≫, Proceedings of the European Symposium on Artificial Neural Networks, 101–106
- Schwarz, G.E. (1978) ≪ Estimating the dimension of a model ≫, Annals of Statistics, 6(2) :461–464
- Spirtes E., Glymour C. et Scheines R. (2000) Causation, Prediction and Search, 2nd edition, Springer-Verlag
- Verma T. (1993) ≪ Graphical aspects of causal models ≫, Technical report R-191, UCLA, Computer Science Department