cours 5
Apprentissage fondé sur les scores

Aix*Marseille université Initiative d'excellence — Master SID — Raisonnement dans l'incertain

©CG(2026)

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 $\mathcal{G}$

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

❷ en trouvant dans son voisinage le graphe $\mathcal{G}'$ de score plus élevé

❸ en itérant le processus avec $\mathcal{G}'$ jusqu'à un optimum (local)

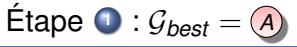
$\Longrightarrow$ Algorithme « Greedy Hill Climbing »

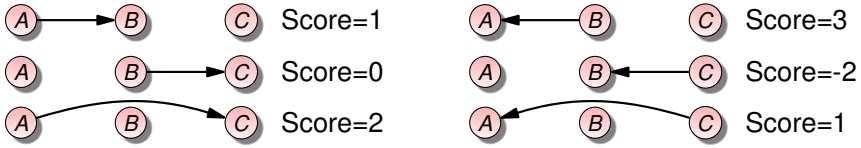
Greedy Hill Climbing

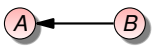
Algorithme d'apprentissage « glouton » :

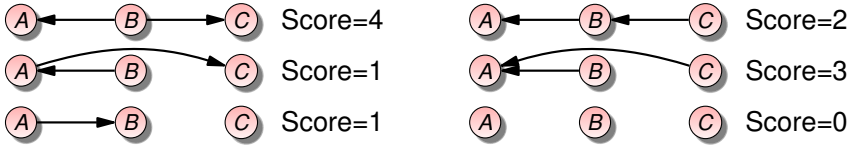
```
1  // étape ❶
2  G_best ← graphe orienté vide (sans arc) // meilleur graphe trouvé
3  sc_best ← Score(G_best) // score du meilleur DAG
4
5  // parcours de l'espace des DAG – étape ❸
6  répéter
7      // recherche du meilleur voisin
8      N ← voisinage de G_best // calcul du voisinage
9      trouvé ← false // meilleur voisin pas encore trouvé
10     // parcours du voisinage – étape ❷
11     pour chaque G' ∈ N faire
12         sc' ← Score(G')
13         si sc' > sc_best alors
14             G_best ← G', sc_best ← sc'
15             trouvé ← true
16 jusqu'à trouvé = false;
17 Retourner G_best // graphe localement optimal
```

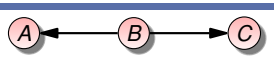
Exemple d'application

- 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 $\mathcal{G}_{best} =$  $sc_{best} = 4$
- Nouveau voisinage $\mathcal{N}$:

.....

Le score BD (1/2)

Score BD($\mathcal{G}|\mathbf{D}$) *[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 $n \in \mathbb{N}^*$

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

Le score BD (2/2)

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

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

  $\Longrightarrow$ 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)

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

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 $N_{ij} = |\mathbf{D}| = 7\ \forall i$
- $i = 1$ (A) :

a_1	a_2
3	4

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

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

- $i = 2$ (B) :

b_1	b_2
3	4

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

- $i = 3$ (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	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

$\text{Score}_{K2}(A|B) = \log\left(\frac{(2-1)!}{(3+2-1)!}\right) + \log(2!) + \log(1!)$ (ligne b_1)

$+ \log\left(\frac{(2-1)!}{(4+2-1)!}\right) + \log(1!) + \log(3!)$ (ligne b_2)

$\approx -6,1738$

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

$\Longrightarrow$ 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 $\quad N' = \ll$ *effective sample size* $\gg$

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

  $\Longrightarrow$ 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 $\ll$ collent $\gg$ 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 »

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

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)

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

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)
- $i = 1$ (A) :

a_1	a_2	N_{ij}
3	4	7

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

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

b_1	b_2
3	4

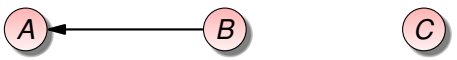
  $\text{Score}_{LL}(B) \approx -4,78$
- $i = 3$ (C) :

c_1	c_2
4	3

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

Exemple d'application (2/2)

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



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

  $\text{Score}_{MDL}(A|B) = 2 \times \log\left(\frac{2}{3}\right) + 1 \times \log\left(\frac{1}{3}\right)$ (ligne b_1)
  $+ 1 \times \log\left(\frac{1}{4}\right) + 3 \times \log\left(\frac{3}{4}\right)$ (ligne b_2)
  $- \frac{1}{2}\log(7) \times 1 \times 2 \approx -6,105$ (pénalité)
- $\text{Score}_{MDL}(\mathcal{G}) \approx -6,105 - 5.753 - 5.753 = -17,611$

  $\Longrightarrow$ 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
- 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

- Pearl J. et Verma T. (1991) ≪ A theory of inferred causation ≫, Proceedings of the 2nd International Conference on Knowledge Representation and Reasoning, 441–452
- Schwarz, G.E. (1978) ≪ Estimating the dimension of a model ≫, Annals of Statistics, 6(2) :461–464