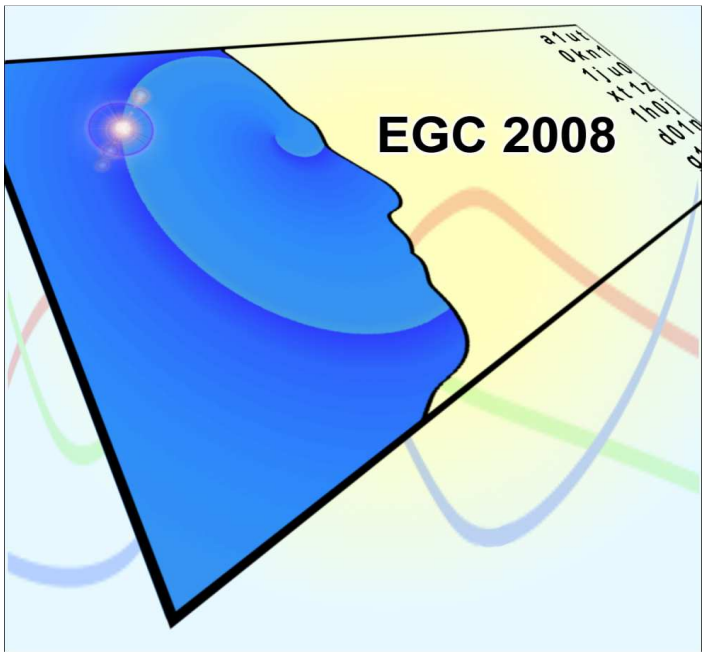
8èmes Journées Francophones
Extraction et Gestion des Connaissances
Sophia Antipolis 29 janvier 2008

Tutoriel



Réseaux Bayésiens

Introduction et apprentissage
Modélisation et découverte de connaissances

Organisateur :
- Philippe LERAY (Univ. Nantes)

Responsables des tutoriels EGC
Hicham Behja (INRIA, Sophia Antipolis)
Bernard Senach (INRIA, Sophia Antipolis)

http://www-sop.inria.fr/axis/egc08
EGC
INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
INRIA
centre de recherche SOPHIA ANTIPOLIS - MÉDITERRANÉE

Réseaux bayésiens - introduction et apprentissage modélisation et découverte de connaissances

Philippe LERAY
philippe.leray@univ-nantes.fr

Equipe COD
Laboratoire d'Informatique de Nantes Atlantique
Site Ecole Polytechnique de l'Université de Nantes
La Chantrerie - rue Christian Pauc - BP 50609
44306 Nantes Cedex 3

Résumé

La représentation des connaissances et le raisonnement à partir de ces représentations a donné naissance à de nombreux modèles. Les modèles graphiques probabilistes, et plus précisément les réseaux bayésiens (RB), initiés par Judea Pearl dans les années 1980, se sont révélés des outils très pratiques pour la représentation de connaissances incertaines et le raisonnement à partir d'informations incomplètes, dans de nombreux domaines comme la bio-informatique, la gestion du risque, le marketing, la sécurité informatique, le transport, etc.

La partie graphique des RB offre un outil intuitif inégalable et attractif dans de nombreuses applications où les utilisateurs ont besoin de "comprendre" ce que raconte le modèle qu'ils utilisent. La construction de ces modèles à partir de données permet aussi de découvrir des connaissances utiles aux experts, en allant – sous certaines réserves - jusqu'à la découverte de relations causales.

Ce tutoriel se propose tout d'abord de définir la notion de réseau bayésien puis de donner un apercu de l'utilisation de ces modèles pour répondre à différentes requêtes (notion d'inférence ou de raisonnement probabiliste). Nous aborderons ensuite le problème de l'apprentissage des réseaux bayésiens à partir de données complètes ou incomplètes, en commençant par la détermination des distributions de probabilité conditionnelles définies par un graphe donné (apprentissage des paramètres), et en essayant ensuite de déterminer le graphe même à partir des données (apprentissage de la structure). Pour finir, nous aborderons le cas plus particulier des réseaux bayésiens causaux, et verrons comment l'apprentissage de la structure de ces modèles peut mener à la découverte de relations causales.

Mots-clés

Réseaux bayésiens, apprentissage, données complètes, données incomplètes, découverte de causalité

Plan

Le tutoriel proposé est inspiré des formations réseaux bayésiens effectuées pour le réseau RISC du RISC en 2005 et 2006, et des cours dispensés en formation ingénieur à l'INSA de Rouen et en Master Recherche à l'Université de Rouen. Le plan est le suivant :

MATIN

- Réseaux bayésiens : définition et notion d'inférence
  - définition, notion de d-séparation
  - les réseaux bayésiens comme modèles génératifs
  - notion d'inférence, principe des principaux algorithmes (message passing, junction tree)
  - exemples d'utilisation
- Réseaux bayésiens : apprentissage des paramètres
  - maximum de vraisemblance vs. maximum a posteriori
  - données complètes vs. données incomplètes

APRES-MIDI

- Réseaux bayésiens : apprentissage de la structure
  - recherche d'indépendances conditionnelles vs. maximisation d'un score d'adéquation
  - les différents espaces de recherche
  - données complètes vs. données incomplètes
- Réseaux bayésiens et causalité
  - un réseau bayésien n'est pas forcément un modèle causal
  - définition d'un réseau bayésien causal
  - intervention/manipulation vs. observation
  - suffisance causale vs. variables latentes

Références

- Jensen, F. V. (1996). *An introduction to Bayesian Networks*. Taylor and Francis, London, United Kingdom.
- Maes, S., Meganck, S., and Leray, P. (2007). An integral approach to causal inference with latent variables. In Russo, F. and Williamson, J., editors, *Causality and Probability in the Sciences*. Texts In Philosophy series, London College Publications, pp 17-41.
- Misc. (2007). *Modèles graphiques probabilistes*. In Leray, P., editor, Revue d'Intelligence Artificielle, number 21:3/2007. Hermès.
- Naïm, P., Wuillemin, P.-H., Leray, P., Pourret, O., and Becker, A. (2004). *Réseaux bayésiens*. Eyrolles, Paris.
- Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, England.

Réseaux bayésiens
introduction et apprentissage
modélisation et découverte de connaissances

Philippe LERAY
philippe.leray@univ-nantes.fr

Equipe COnnaissances et Décision
Laboratoire d'Informatique de Nantes Atlantique – UMR 6241
Site de l'Ecole Polytechnique de l'université de Nantes

Au programme ...

Matin $\Longrightarrow$ **Notions générales**

- Définition, D-séparation, Notion d'inférence

Matin — **Apprentissage des paramètres**

- Maximum de vraisemblance / a posteriori
- Données complètes / incomplètes

Après-midi — **Apprentissage de la structure**

- Recherche d'indépendances / maximisation score
- Quel espace ? Données complètes / incomplètes

Après-midi — **RB et causalité**

- RB causal, intervention / observation, suffisance causale

Un peu d'histoire

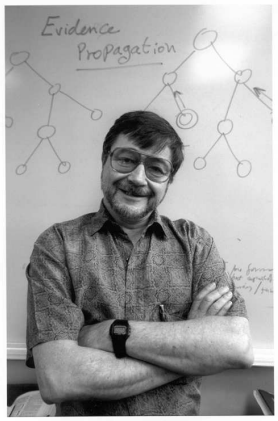
RULE037

IF the organism
1) stains grampos
2) has coccus shape
3) grows in chains

THEN
There is suggestive evidence (.7) that the identity of the organism is streptococcus.

1970-1990 : L'ère des systèmes experts

- systèmes à base de règles de production
  `si X=vrai et Y=absent alors Z=faux`
- moteur d'inférence (chainage avant, arrière)



Judea Pearl (1936–) : les réseaux bayésiens

- 1982 : *Reverend Bayes on inference engines: A distributed hierarchical approach*
  `P(X=vrai)=0.3 et P(Z=faux)=0.2 ...`
  `P(Y=absent)=?`
- 1988 : *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann

Rappels de probabilités

Probabilité conditionnelle

- A et M deux événements
- information a priori sur A : $P(A)$
- M s'est produit : $P(M) \neq 0$
- s'il existe un lien entre A et M, cet événement va modifier notre connaissance sur A
- information a posteriori : $P(A|M) = \frac{P(A,M)}{P(M)}$

Rappels de probabilités

Indépendance

- A et B sont indépendants ssi :
  $P(A, B) = P(A) \times P(B)$
  $P(A|B) = P(A)$
  $P(B|A) = P(B)$

Indépendance conditionnelle

- A et B sont indépendants conditionnellement à C ssi :
  $P(A|B, C) = P(A|C)$

Rappels de probabilités

$\{M_i\}$ ensemble complet d'événements mutuellement exclusifs

Marginalisation : $P(A) = \sum_i P(A, M_i)$

Théorème des probabilités totales

Un événement A peut résulter de plusieurs causes M_i. Quelle est la probabilité de A connaissant :

- *les probabilités élémentaires P(Mi) (a priori)*
- *les probabilités conditionnelles de A pour chaque Mi*

$$P(A) = \sum_i P(A|M_i)P(M_i)$$

mais comment répondre à la question inverse ?

Rappels de probabilités

$\{M_i\}$ ensemble complet d'événements mutuellement exclusifs

Théorème de Bayes

Un événement A s'est produit. Quelle est la probabilité que ce soit la cause M_i qui l'ait produit ?

$$P(M_i|A) = \frac{P(A|M_i) \times P(M_i)}{P(A)}$$

- $P(Mi|A)$: probabilité a posteriori
- $P(A)$: constante (pour chaque M_i) cf. th. probas totales

Théorème de Bayes généralisé (Chain rule)

$$P(A_1 \dots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \dots P(A_n|A_1 \dots A_{n-1})$$

Définition d'un réseau bayésien

Principe

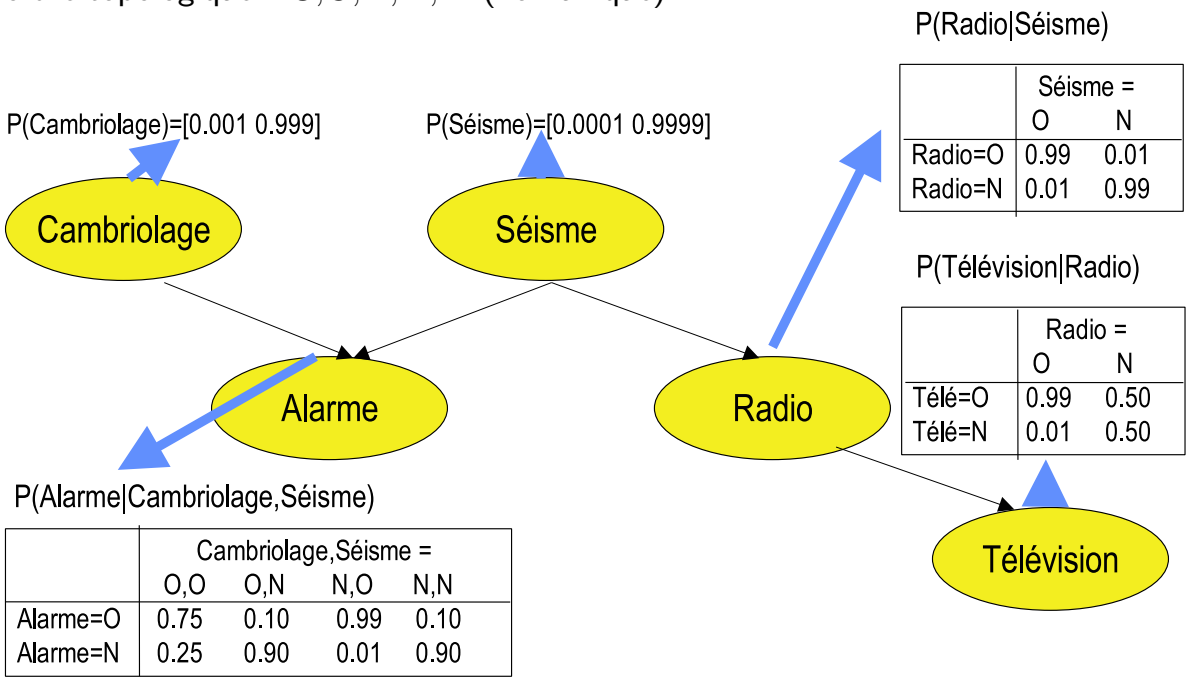
- prendre en compte les indépendances conditionnelles entre les variables pour simplifier la loi jointe donnée par le théorème de Bayes généralisé.

Définition

- Un réseau bayésien est défini par
  - la description qualitative des dépendances (ou des indépendances conditionnelles) entre des variables
    graphe orienté sans circuit (DAG)
  - la description quantitative de ces dépendances
    probabilités conditionnelles (CPD)

Exemple

ordre topologique : C, S, A, R, T (non unique)



P(Radio|Séisme)

	Séisme = O	N
Radio=O	0.99	0.01
Radio=N	0.01	0.99

P(Télévision|Radio)

	Radio = O	N
Télé=O	0.99	0.50
Télé=N	0.01	0.50

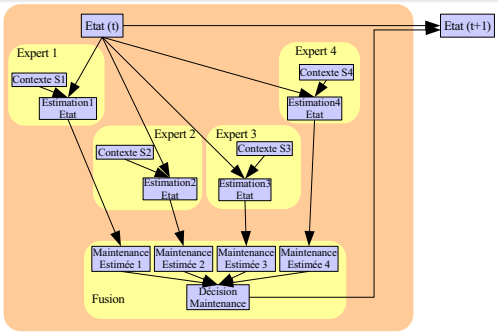
P(Alarme|Cambriolage,Séisme)

	Cambriolage,Séisme = O,O	O,N	N,O	N,N
Alarme=O	0.75	0.10	0.99	0.10
Alarme=N	0.25	0.90	0.01	0.90

Intérêts et motivation

Intérêts des réseaux bayésiens

- outil de **représentation** graphique des connaissances
- représentation de l'incertain
- raisonnement à partir de données incomplètes : **inférence**



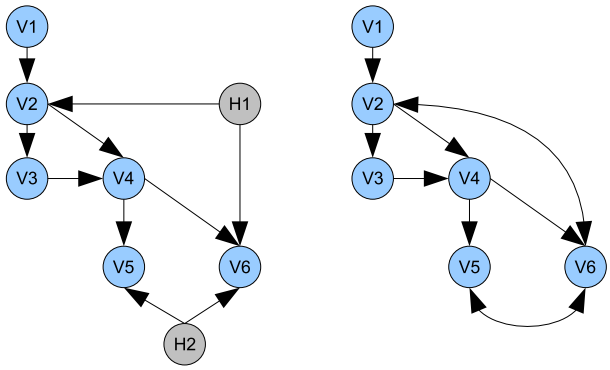
Motivation

- comment déterminer la structure, avec des données complètes ou incomplètes ?

Intérêts et motivation

Autre intérêt

- outil de **découverte** de connaissances à partir de données



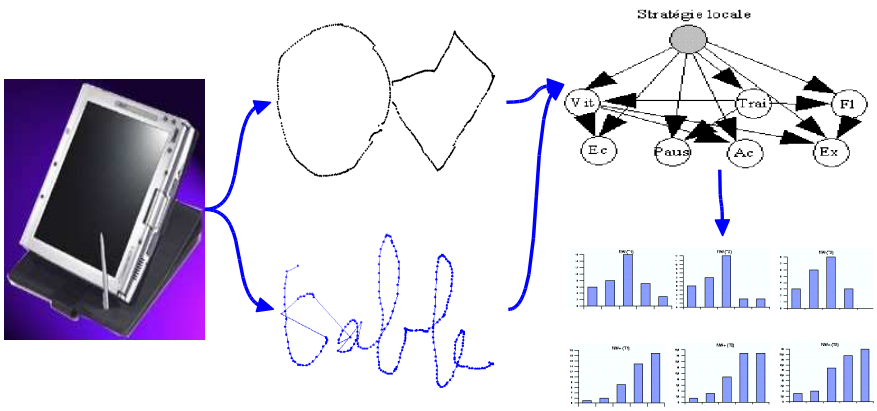
Motivation

- comment découvrir des connaissances : relations causales, variables latentes ?

Intérêts et motivation

Des domaines d'application variés

- diagnostic, fiabilité, maintenance, sécurité informatique
- psychologie, sciences de la cognition, maîtrise des risques



Motivation

- fournir des outils pour la modélisation de systèmes complexes

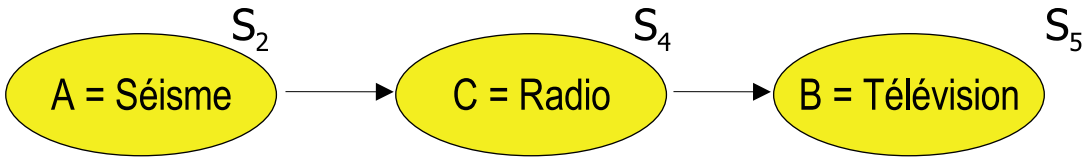
RB et indépendance conditionnelle

Les RB représentent graphiquement les indépendances conditionnelles

Exemple sur 3 nœuds

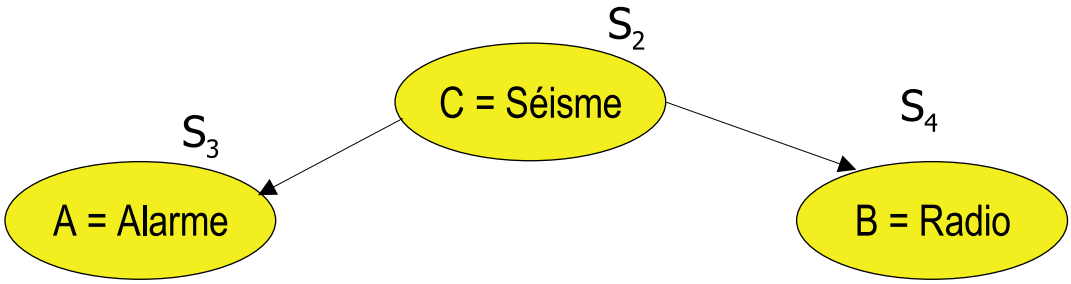
- 3 types de relations (simples) entre A, B et C :
  - $A \rightarrow C \rightarrow B$: connexion série
  - $A \leftarrow C \rightarrow B$: connexion divergente
  - $A \rightarrow C \leftarrow B$: connexion convergente (V-structure)

Connexion série



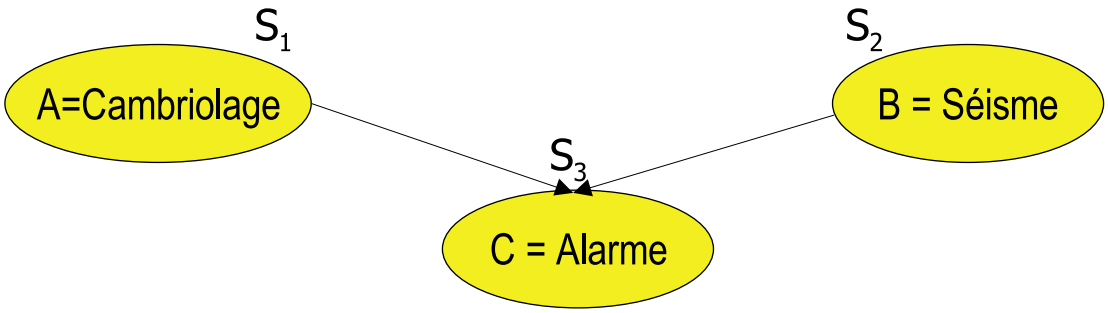
- A et B sont dépendants
- A et B sont indépendants conditionnellement à C
  - si C est connue, A n'apporte aucune information sur B
  - $P(S_5|S_4, S_2) = P(S_5|S_4) = P(S_5|parents(S_5))$

Connexion divergente



- A et B sont dépendants
- A et B sont indépendants conditionnellement à C
  - si C est connue, A n'apporte aucune information sur B
  - $P(S_4|S_2, S_3) = P(S_4|S_2) = P(S_4|parents(S_4))$

Connexion convergente – V-structure



- A et B sont indépendants
- A et B sont dépendants conditionnellement à C
  - si C est connue, A apporte une information sur B
  - $P(S_3|S_1, S_2) = P(S_3|parents(S_3))$

Conséquence

Rappel du théorème de Bayes généralisé

$$P(S) = P(S_1) \times P(S_2|S_1) \times P(S_3|S_1, S_2) \times \cdots \times P(S_n|S_1 \ldots S_{n-1})$$

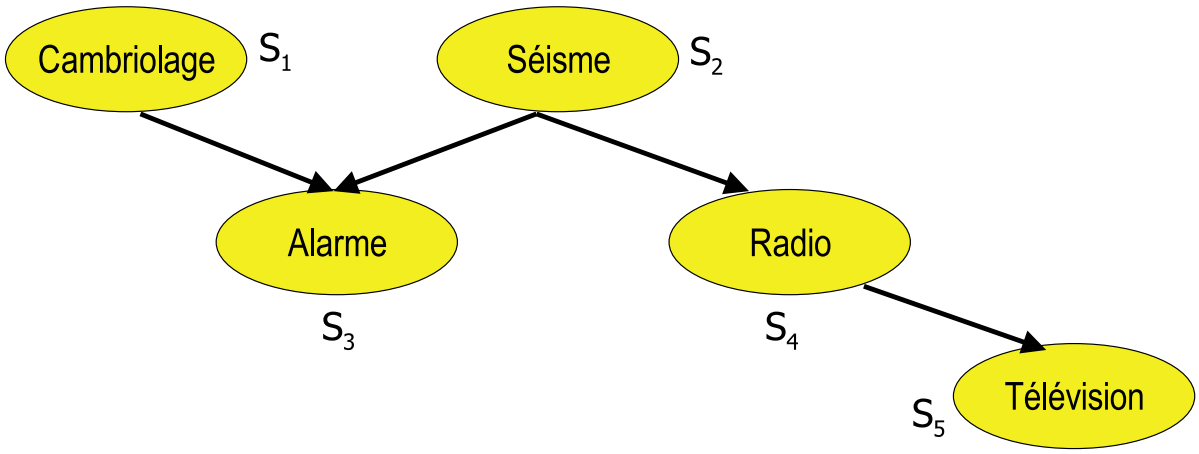
Conséquence dans un RB

- $P(S_i|S_1 \ldots S_{i-1}) = P(S_i|parents(S_i))$ d'où

$$P(S) = \Pi_{i=1}^{n} P(S_i|parents(S_i))$$

- La loi jointe (globale) se décompose en un produit de lois conditionnelles locales
- RB = représentation compacte de la loi jointe $P(S)$

Exemple



$$P(Cambriolage, Seisme, Alarme, Radio, Tele) =$$
$$P(S_1)P(S_2|S_1)P(S_3|S_1, S_2)P(S_4|S_1, S_2, S_3)P(S_5|S_1, S_2, S_3, S_4)$$
$$P(S_1) \quad P(S_2) \quad P(S_3|S_1, S_2) \quad P(S_4|S_2) \quad P(S_5|S_4)$$

D-séparation

Principe

- Déterminer si deux variables quelconques sont indépendantes conditionnellement à un ensemble de variables instantiées

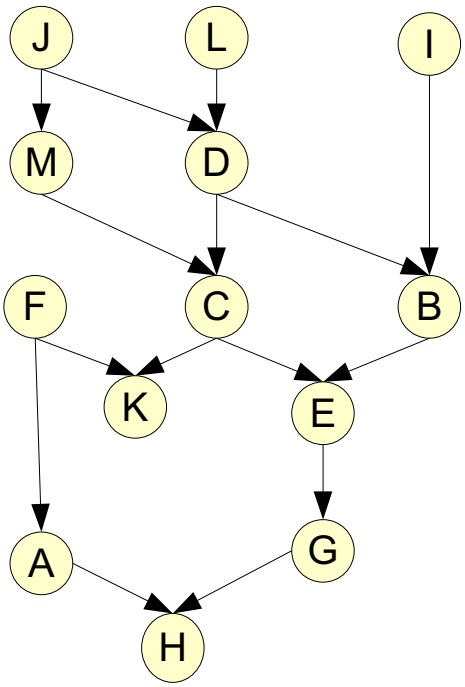
Définition

- Deux variables A et B sont d-séparées si pour tous les chemins entre A et B, il existe une variable intermédiaire V différente de A et B telle que l'une des deux propositions est vraie :
  - la connexion est série ou divergente et V est instancié
  - la connexion est convergente et ni V ni ses descendants ne sont instanciés
- Si A et B ne sont pas d-séparés, ils sont d-connectés

Exemple

D-séparation

- la connexion est série ou divergente et V est instancié
- la connexion est convergente et ni V ni ses descendants ne sont instanciés



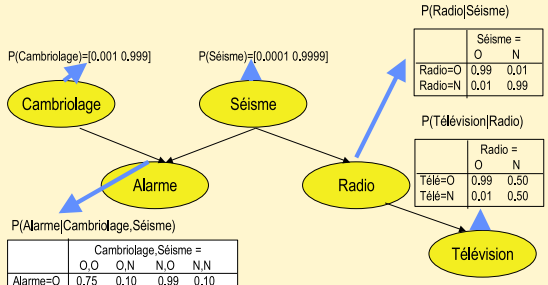
RB = modèle génératif

Principe

- RB = représentation compacte de la loi jointe $P(S)$
- Utilisation de méthodes d'échantillonnage pour générer des données qui suivent cette loi

Exemple : forward sampling

- si $rand1 < 0.001$, $C = O$, sinon N
- si $rand2 < 0.0001$, $S = O$, sinon N
- si $rand3 < P(A = O|C = ..., S = ...)$, $A = O$, sinon N
- ...



Notion d'inférence

Inférence

- calcul de n'importe quelle $P(S_i|S_j = x)$ (NP-complet)
- l'observation $\{S_j = x\}$ est appelée évidence

Algorithmes exacts

- Message Passing (Pearl 1988) pour les arbres
- Junction Tree (Jensen 1990)
- Shafer-Shenoy (1990)

Problème = explosion combinatoire de ces méthodes pour des graphes fortement connectés.

Algorithmes approchés

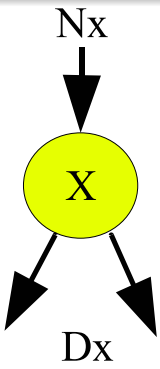
- Echantillonnage
- Méthodes variationnelles

Message Passing (Pearl 1988)

Principe

- Chaque nœud envoie des messages à ses voisins
- L'algorithme ne marche que dans le cas des arbres (mais est généralisable au cas des poly-arbres)

- E = ensemble de variables instanciées.
  $E = N_x \cup D_x$
- 2 types de messages λ et π serviront à calculer
  - $\lambda(X) \propto P(D_x|X)$
  - $\pi(X) \propto P(X|N_x)$



- et ensuite on peut montrer que

$$P(X|E = e) \propto \lambda(X)\pi(X)$$

Message Passing

Les messages λ

- Pour chaque enfant Y de X,

$$\lambda_Y(X = x) = \sum_y P(Y = y|X = x)\lambda(Y = y)$$

Comment calculer λ en chaque nœud ?

Calcul de λ

- Si X instancié, $\lambda(X) = [001 \dots 0]$
  (la position du 1 correspond à la valeur donnée à X)
- sinon
  - si X est une feuille, $\lambda(X) = [1 \dots 1]$
  - sinon $\lambda(X = x) = \Pi_{Y \in Enf(X)} \lambda_Y(X = x)$

Message Passing

Les messages π

- Pour Z l'unique parent de X,

$$\pi_X(Z = z) = \pi(Z = z) \prod_{U \in Enf(Z) \setminus \{X\}} \lambda_U(Z = z)$$

Comment calculer π en chaque nœud ?

Calcul de π

- Si X instancié, $\pi(X) = [001 \ldots 0]$
  (la position du 1 correspond à la valeur donnée à X)
- sinon
  - si X est la racine, $\pi(X) = P(X)$
  - sinon $\pi(X = x) = \sum_z P(X = x|Z = z)\pi_X(Z = z)$

Junction Tree (Jensen 1990)

- Message Passing ne s'applique bien qu'aux arbres
- Besoin d'un algorithme plus général

Principe

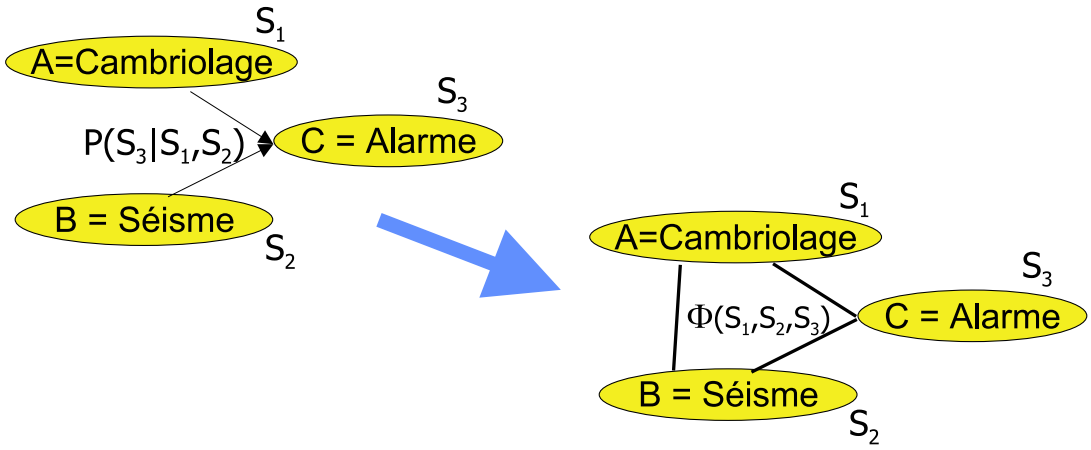
- Transformer le graphe en un arbre (non orienté)...
- Arbre = arbre de jonction des cliques maximales du graphe moralisé et triangulé

- Moralisation = ???
- Triangulation = ???
- Cliques = ???

Junction Tree

Moralisation

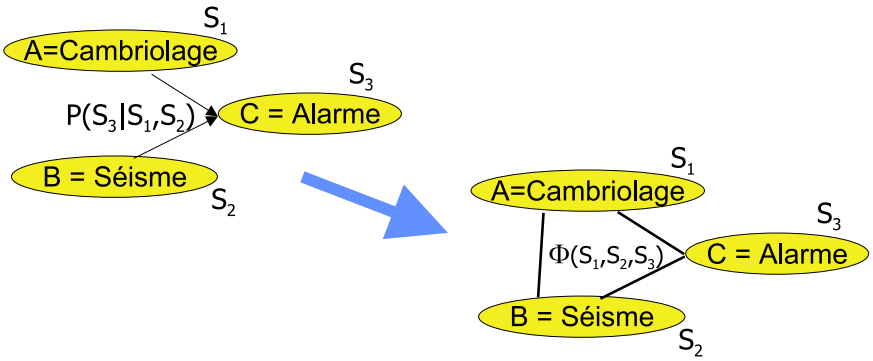
- marier les parents de chaque nœud



Junction Tree

Triangulation

- tout cycle de longueur au moins 4 doit contenir une corde (arête reliant deux sommets non consécutifs sur le cycle)
- (= aucun sous-graphe cyclique de longueur ≥ 4)
- Triangulation optimale pour des graphes non-dirigés = NP-difficile (comment choisir les meilleures cordes?)



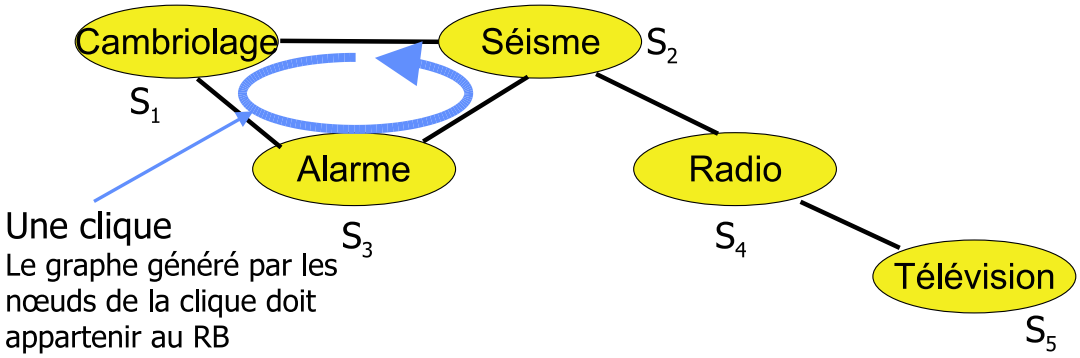
Junction Tree

Clique

- sous-graphe dont les nœuds sont complétement connectés

Clique maximale

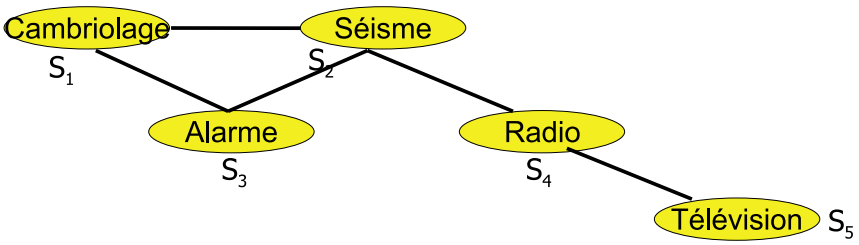
- l'ajout d'un autre nœud à cette clique ne donne pas une clique

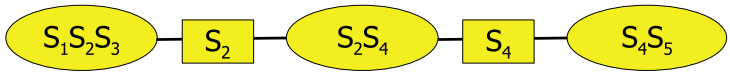


Junction Tree

Théorème

- Si le graphe est moralisé et triangulé, alors les cliques peuvent être organisées en un arbre de jonction





$$P(S) = \Phi(S1, S2, S3)\Phi(S2, S4)\Phi(S4, S5)$$

- L'inférence se fait au niveau des Φ

Références

- **Les Réseaux Bayésiens** - P. Naïm, P.H. Wuillemin, Ph. Leray, O. Pourret, A. Becker (Eyrolles) 2007
- **Probabilistic reasoning in Intelligent Systems: Networks of plausible inference** - J. Pearl (Morgan Kaufman) 1988
- **An introduction to Bayesian Networks** - F. Jensen (Springer Verlag) 1996
- **Probabilistic Networks and Expert Systems** - R.G. Cowell & al. (Springer Verlag) 1999
- **Learning Bayesian Networks** - R. Neapolitan (Prenctice Hall) 2003
- **Learning in Graphical Models** - Jordan M.I. ed. (Kluwer) 1998
- **An integral approach to causal inference with latent variables** - S. Maes et al. In Russo, F. and Williamson, J., editors, Causality and Probability in the Sciences. Texts In Philosophy series, London College Publications, pp 17-41. 2007

Réseaux bayésiens
introduction et apprentissage
modélisation et découverte de connaissances

Philippe LERAY
philippe.leray@univ-nantes.fr

Equipe COnnaissances et Décision
Laboratoire d'Informatique de Nantes Atlantique – UMR 6241
Site de l'Ecole Polytechnique de l'université de Nantes

Au programme ...

Matin — **Notions générales**

- Définition, D-séparation, Notion d'inférence

Matin — $\Longrightarrow$ **Apprentissage des paramètres**

- Maximum de vraisemblance / a posteriori
- Données complètes / incomplètes

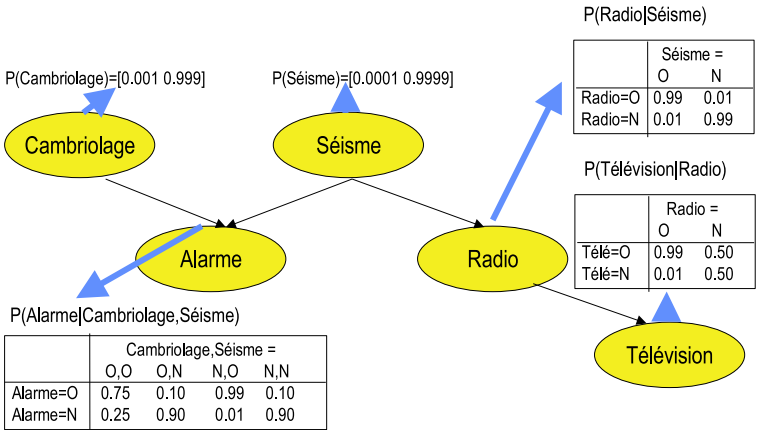
Après-midi — **Apprentissage de la structure**

- Recherche d'indépendances / maximisation score
- Quel espace ? Données complètes / incomplètes

Après-midi — **RB et causalité**

- RB causal, intervention / observation, suffisance causale

Définition d'un réseau bayésien



Un réseau bayésien est défini par

- la description qualitative des dépendances (ou des indépendances conditionnelles) entre des variables
  graphe orienté sans circuit (DAG)
- la description quantitative de ces dépendances
  probabilités conditionnelles (CPD)

Notion d'apprentissage

Construire un réseau bayésien

1. structure fixée, on cherche seulement les CPD
   - à partir d'expertises : élicitation de connaissances
   - à partir de données complètes / incomplètes
2. on cherche la structure
   - à partir de données complètes / incomplètes
   - dans quel espace ?
   - connaît-on toutes les variables ?

Apprentissage (données complètes)

Estimation de paramètres — Données complètes $\mathcal{D}$

- Déterminer les paramètres des différentes CPD à partir de $\mathcal{D}$
- Approche statistique classique = *max. de vraisemblance (MV)*

$$\hat{\theta}^{MV} = \text{argmax } P(\mathcal{D}|\theta)$$

- Probabilité d'un événement = fréquence d'apparition de l'événement

Maximum de vraisemblance (MV)

$$\hat{P}(X_i = x_k | Pa(X_i) = x_j) = \hat{\theta}^{MV}_{i,j,k} = \frac{N_{i,j,k}}{\sum_k N_{i,j,k}}$$

$N_{i,j,k}$ = nb d'occurences de $\{X_i = x_k \text{ et } Pa(X_i) = x_j\}$

Apprentissage (données complètes)

Autre approche

- Approche bayésienne = *max. à posteriori (MAP)*

$$\hat{\theta}^{MAP} = \text{argmax } P(\theta|\mathcal{D}) = \text{argmax } P(\mathcal{D}|\theta)P(\theta)$$

- besoin d'une loi a priori sur les paramètres $P(\theta)$
- souvent distribution *conjuguée* à la loi de X
- si $P(X)$ multinomiale, $P(\theta)$ conjuguée = Dirichlet :

$$P(\theta) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (\theta_{i,j,k})^{\alpha_{i,j,k}-1}$$

où $\alpha_{i,j,k}$ sont les cœfficients de la distribution de Dirichlet associée au coefficient $\theta_{i,j,k}$

Apprentissage (données complètes)

Maximum a Posteriori (MAP)

$$\hat{P}(X_i = x_k | Pa(X_i) = x_j) = \hat{\theta}^{MAP}_{i,j,k} = \frac{N_{i,j,k} + \alpha_{i,j,k} - 1}{\sum_k (N_{i,j,k} + \alpha_{i,j,k} - 1)}$$

Autre approche bayésienne

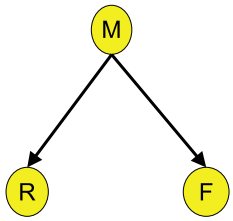
- *espérance à posteriori (EAP)* : calculer l'espérance a posteriori de $\theta_{i,j,k}$ au lieu du max.

$$\hat{P}(X_i = x_k | Pa(X_i) = x_j) = \hat{\theta}^{EAP}_{i,j,k} = \frac{N_{i,j,k} + \alpha_{i,j,k}}{\sum_k (N_{i,j,k} + \alpha_{i,j,k})}$$

Exemple

- Données complètes (MV)

$$\begin{aligned} \hat{P}(M = m_0) &= 6/15 = 0.4 \\ \hat{P}(M = m_1) &= 8/15 = 0.53 \\ \hat{P}(M = m_2) &= 1/15 = 0.07 \end{aligned}$$



$$\begin{aligned} \hat{P}(F = OK | M = m_0) &= 1/6 = 0.17 \\ \hat{P}(F = BAD | M = m_0) &= 5/6 = 0.83 \\ &\text{etc ...} \end{aligned}$$

M	F	R
m_0	BAD	O
m_0	BAD	O
m_0	BAD	O
m_0	BAD	O
m_0	BAD	N
m_0	OK	O
m_1	BAD	O
m_1	BAD	N
m_1	OK	O
m_1	OK	N
m_1	OK	O
m_1	OK	N
m_1	OK	O
m_1	OK	N
m_2	OK	N

- Problème :
  $\hat{P}(F = BAD | M = m_2) = 0/1$
  car cette configuration ne figure pas dans notre (petite) base d'exemples

Exemple

- Données complètes (EAP)
  - A priori de Dirichlet sur les $\theta_{i,j,k}$
  - $\approx$ pseudo tirage *a priori* de N^* mesures
- Exemples
  - A priori de Dirichlet sur M réparti sur m_0 et $m_1 = [50 \;\; 50 \;\; 0]$

$$\begin{aligned}\hat{P}(M = m_0) &= (6 + 50)/(15 + 100) = 0.487\\ \hat{P}(M = m_1) &= (8 + 50)/(15 + 100) = 0.5043\\ \hat{P}(M = m_2) &= (1 + 0)/(15 + 100) = 0.0087\end{aligned}$$

  - A priori de Dirichlet sur $(F|M = m_i)$ $= [9 \;\; 1]$

$$\hat{P}(F = BAD|M = m_2) = (0 + 1)/(1 + 10) = 0.09$$

M	F	R
m_0	BAD	O
m_0	BAD	O
m_0	BAD	O
m_0	BAD	O
m_0	BAD	N
m_0	OK	O
m_1	BAD	O
m_1	BAD	N
m_1	OK	O
m_1	OK	N
m_1	OK	O
m_1	OK	N
m_1	OK	O
m_1	OK	N
m_2	OK	N

Apprentissage (données incomplètes)

Plusieurs types de données incomplètes (Rubin, 1976)

- MCAR : *Missing Completly At Random*
  - absence de données = complétement aléatoire
  - comment estimer MV ou MAP ?
    - Complete / Available Case Analysis ...
- MAR : *Missing At Random*
  - probabilité qu'une donnée soit manquante dépend des variables observées
  - comment estimer MV ou MAP ?
    - *Expectation Maximisation* ...
- NMAR : *Not Missing At Random*
  - absence de données dépend de phénom. externes
  - besoin de connaissances supplém. dans le modèle

Complete / Available Case Analysis

Complete Case Analysis

- Extraire de la base de données incomplète les individus complètement mesurés
- Avantage : on retombe dans le cas des données complètes
- Inconvénient : taux d'incomplétude important $\Rightarrow$ peu de données complètes

Available Case Analysis

- Principe : pas besoin de savoir si C est mesuré pour estimer les paramètres de $P(A|B)$
- Pour estimer $P(A|B)$, extraire de la base de données incomplète les individus pour lesquels A et B sont mesurés
- Avantage : on retombe dans le cas des données complètes

Algorithme Expectation Maximisation

Algorithme très général (Dempster 1977)

- Algorithme général d'estimation de paramètres avec des données incomplètes

Principe

- Algorithme itératif
  - initialiser les paramètres $\theta^{(0)}$ (random, CCA / ACA)
  - E estimer les valeurs manquantes à partir des paramètres actuels $\theta^{(t)}$
    - = calculer $P(X_{\text{manquant}}|X_{\text{mesurés}})$ dans le RB actuel
    - = faire des inférences dans le RB muni des paramètres $\theta^{(t)}$
  - M ré-estimer les paramètres $\theta^{(t+1)}$ à partir des données complétées
    - en utilisant MV, MAP, ou EAP

Exemple

- Données manquantes (EM+MV)
  - Exemple sur l'estimation de $P(M)$
  - Initialisation $\hat{P}^{(0)}(M) = [1/3 \;\; 1/3 \;\; 1/3]$

M	F	R
m_0	BAD	O
m_0	BAD	O
?	BAD	O
m_0	BAD	O
?	BAD	N
m_0	OK	O
m_1	BAD	O
m_1	BAD	N
?	OK	O
m_1	OK	N
m_1	OK	O
m_1	OK	N
m_1	?	O
m_1	OK	N
m_2	OK	N

Exemple

M	F	R	$\hat{P}(M = m_0)$	$\hat{P}(M = m_1)$	$\hat{P}(M = m_2)$
m_0	BAD	O	1	0	0
m_0	BAD	O	1	0	0
?	BAD	O	1/3	1/3	1/3
m_0	BAD	O	1	0	0
?	BAD	N	1/3	1/3	1/3
m_0	OK	O	1	0	0
m_1	BAD	O	0	1	0
m_1	BAD	N	0	1	0
?	OK	O	1/3	1/3	1/3
m_1	OK	N	0	1	0
m_1	OK	O	0	1	0
m_1	OK	N	0	1	0
m_1	?	O	0	1	0
m_1	OK	N	0	1	0
m_2	OK	N	0	0	1
TOTAL			**5**	**8**	**2**

Itérat^{o}1

- [E]

Exemple

M	F	R	$\hat{P}(M = m_0)$	$\hat{P}(M = m_1)$	$\hat{P}(M = m_2)$
m_0	BAD	O	1	0	0
m_0	BAD	O	1	0	0
?	BAD	O	1/3	1/3	1/3
m_0	BAD	O	1	0	0
?	BAD	N	1/3	1/3	1/3
m_0	OK	O	1	0	0
m_1	BAD	O	0	1	0
m_1	BAD	N	0	1	0
?	OK	O	1/3	1/3	1/3
m_1	OK	N	0	1	0
m_1	OK	O	0	1	0
m_1	OK	N	0	1	0
m_1	?	O	0	1	0
m_1	OK	N	0	1	0
m_2	OK	N	0	0	1
TOTAL			**5**	**8**	**2**

Itérat^{o}1

- [E]
- [M] :

$\hat{P}^{(1)}(m_0) = 5/15 = 0.333$

$\hat{P}^{(1)}(m_1) = 8/15 = 0.533$

$\hat{P}^{(1)}(m_2) = 2/15 = 0.133$

Exemple

M	F	R	$\hat{P}(M = m_0)$	$\hat{P}(M = m_1)$	$\hat{P}(M = m_2)$
m_0	BAD	O	1	0	0
m_0	BAD	O	1	0	0
?	BAD	O	0.333	0.533	0.133
m_0	BAD	O	1	0	0
?	BAD	N	0.333	0.533	0.133
m_0	OK	O	1	0	0
m_1	BAD	O	0	1	0
m_1	BAD	N	0	1	0
?	OK	O	0.333	0.533	0.133
m_1	OK	N	0	1	0
m_1	OK	O	0	1	0
m_1	OK	N	0	1	0
m_1	?	O	0	1	0
m_1	OK	N	0	1	0
m_2	OK	N	0	0	1
TOTAL			**5**	**8.6**	**1.4**

Itérat^{o}2

- [E]

Exemple

M	F	R	$\hat{P}(M = m_0)$	$\hat{P}(M = m_1)$	$\hat{P}(M = m_2)$
m_0	BAD	O	1	0	0
m_0	BAD	O	1	0	0
?	BAD	O	0.333	0.533	0.133
m_0	BAD	O	1	0	0
?	BAD	N	0.333	0.533	0.133
m_0	OK	O	1	0	0
m_1	BAD	O	0	1	0
m_1	BAD	N	0	1	0
?	OK	O	0.333	0.533	0.133
m_1	OK	N	0	1	0
m_1	OK	O	0	1	0
m_1	OK	N	0	1	0
m_1	?	O	0	1	0
m_1	OK	N	0	1	0
m_2	OK	N	0	0	1
TOTAL			**5**	**8.6**	**1.4**

Itérat^{o}2

- [E]
- [M] :

$\hat{P}^{(2)}(m_0)$
$= 5/15$
$= 0.333$

$\hat{P}^{(2)}(m_1)$
$= 8.6/15$
$= 0.573$

$\hat{P}^{(2)}(m_2)$
$= 1.4/15$
$= 0.093$

Références

- **Les Réseaux Bayésiens** - P. Naïm, P.H. Wuillemin, Ph. Leray, O. Pourret, A. Becker (Eyrolles) 2007
- **Probabilistic reasoning in Intelligent Systems: Networks of plausible inference** - J. Pearl (Morgan Kaufman) 1988
- **An introduction to Bayesian Networks** - F. Jensen (Springer Verlag) 1996
- **Probabilistic Networks and Expert Systems** - R.G. Cowell & al. (Springer Verlag) 1999
- **Learning Bayesian Networks** - R. Neapolitan (Prenctice Hall) 2003
- **Learning in Graphical Models** - Jordan M.I. ed. (Kluwer) 1998
- **An integral approach to causal inference with latent variables** - S. Maes et al. In Russo, F. and Williamson, J., editors, Causality and Probability in the Sciences. Texts In Philosophy series, London College Publications, pp 17-41. 2007

Réseaux bayésiens
introduction et apprentissage
modélisation et découverte de connaissances

Philippe LERAY
philippe.leray@univ-nantes.fr

Equipe COnnaissances et Décision
Laboratoire d'Informatique de Nantes Atlantique – UMR 6241
Site de l'Ecole Polytechnique de l'université de Nantes

Au programme ...

Matin — Notions générales

- Définition, D-séparation, Notion d'inférence

Matin — Apprentissage des paramètres

- Maximum de vraisemblance / a posteriori
- Données complètes / incomplètes

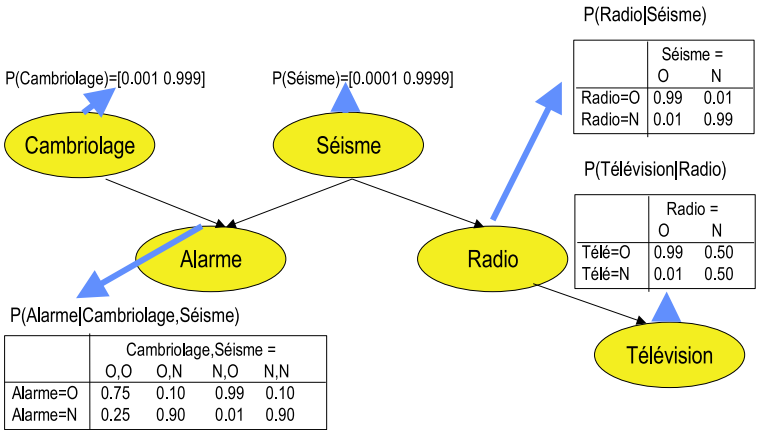
Après-midi $\Longrightarrow$ Apprentissage de la structure

- Recherche d'indépendances / maximisation score
- Quel espace ? Données complètes / incomplètes

Après-midi — RB et causalité

- RB causal, intervention / observation, suffisance causale

Définition d'un réseau bayésien



Un réseau bayésien est défini par

- la description qualitative des dépendances (ou des indépendances conditionnelles) entre des variables
  graphe orienté sans circuit (DAG)
- la description quantitative de ces dépendances
  probabilités conditionnelles (CPD)

Notion d'apprentissage

Construire un réseau bayésien

1. structure fixée, on cherche seulement les CPD
   - à partir d'expertises : élicitation de connaissances
   - à partir de données complètes / incomplètes
2. on cherche la structure
   - à partir de données complètes / incomplètes
   - dans quel espace ?
   - connaît-on toutes les variables ?

Problème complexe

Taille de l'espace de recherche

- le nombre de structures possibles à partir de n nœuds est super-exponentiel (Robinson 77)

$$NS(n) = \begin{cases} 1 & , \quad n = 0 \text{ ou } 1 \\ \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-1)} NS(n-i), & n > 1 \end{cases}$$

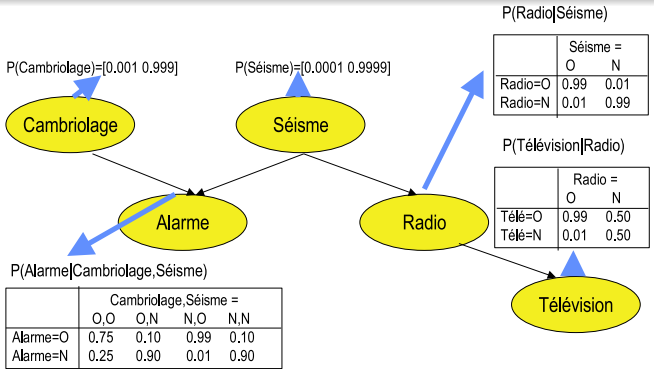
$$NS(5) = 29281 \qquad NS(10) = 4.2 \times 10^{18}$$

- recherche exhaustive impossible / taille de l'espace

Dimension d'un réseau bayésien

Définition

Nombre de paramètres (indépendants) nécessaires pour décrire l'ensemble des CPD associées au RB



Exemples

- $Dim(B) = 1 + 1 + 4 + 2 + 2$
- Graphe "vide" : $Dim(B_0) = ?$
- complétement connecté : $Dim(B_c) = ?$

Equivalence de Markov

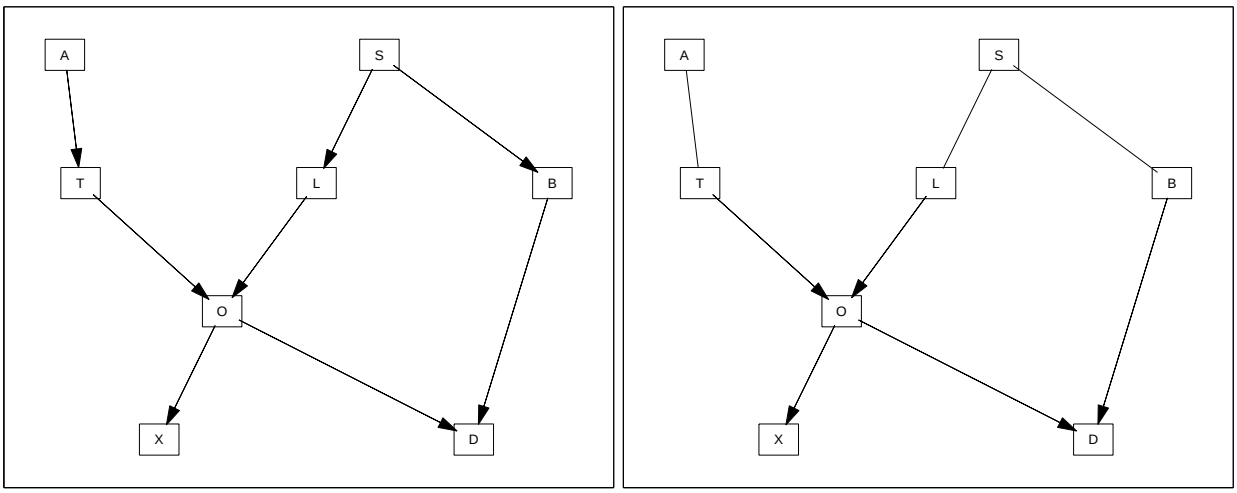
Définition

B_1 et B_2 sont équivalents au sens de Markov ssi ils ont le même squelette et décrivent les mêmes dépendances et indépendances conditionnelles

Conséquences

- B_1 et B_2 partagent les mêmes V-structures et "arcs inferés"
- tous les graphes équivalents peuvent être représentés par un graphe partiellement orienté (squelette, V-structure et arcs inferés) (CPDAG)
- on appelle ce CPDAG le représentant de la classe d'équivalence

Equivalence de Markov - exemple



Apprentissage (données complètes)

Recherche d'un bon réseau bayésien

- Un RB résume des dépendances et indépendances conditionnelles
- Trouver la structure == trouver ces infos dans les données

Recherche d'IC

Deux algorithmes de référence

- Pearl et Verma : IC et IC*
- Spirtes, Glymour et Scheines : SGS, PC, CI, FCI

Principe commun

- construire un graphe non dirigé contenant les relations entre les variables (tests du χ^2)
  - par ajout d'arêtes (Pearl et Verma)
  - par suppression d'arêtes (SGS)
- détecter les V-structures (idem)
- "propager" les orientations de certains arcs

Recherche d'IC

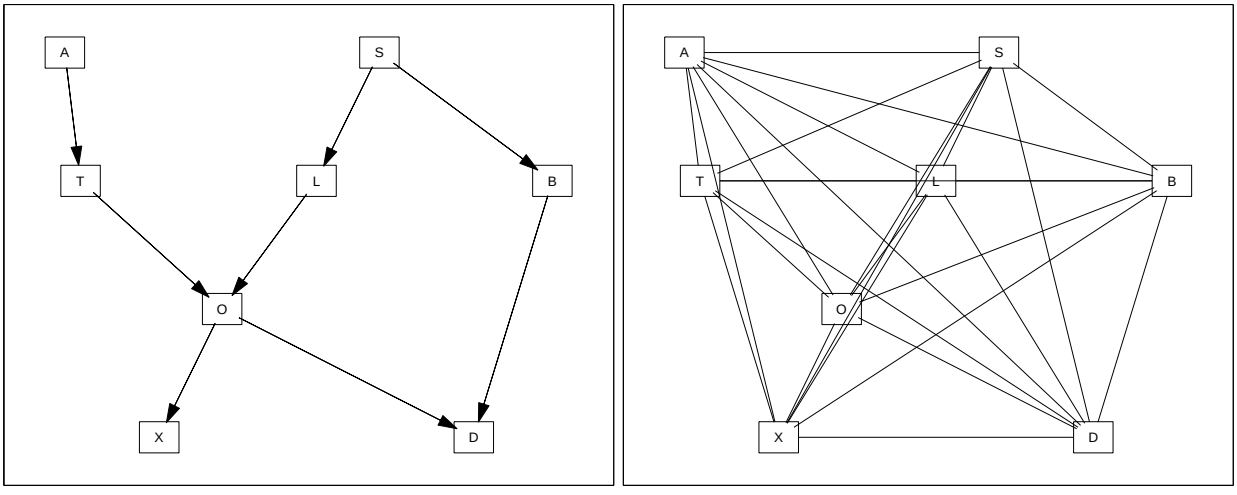
Problèmes principaux

- Fiabilité du test d'indépendance conditionnellement à un grand nb de variables (et avec un nb de données restreint)
  - Heuristique SGS : si $df < \frac{N}{10}$, alors dépendance
- Explosion du nb de tests à effectuer
  - Heuristique PC : commencer par l'ordre 0 ($X_A \perp X_B$) puis l'ordre 1 ($X_A \perp X_B \mid X_C$), etc ...

Algorithme PC

Etape 0 : Graphe non orienté reliant tous les nœuds

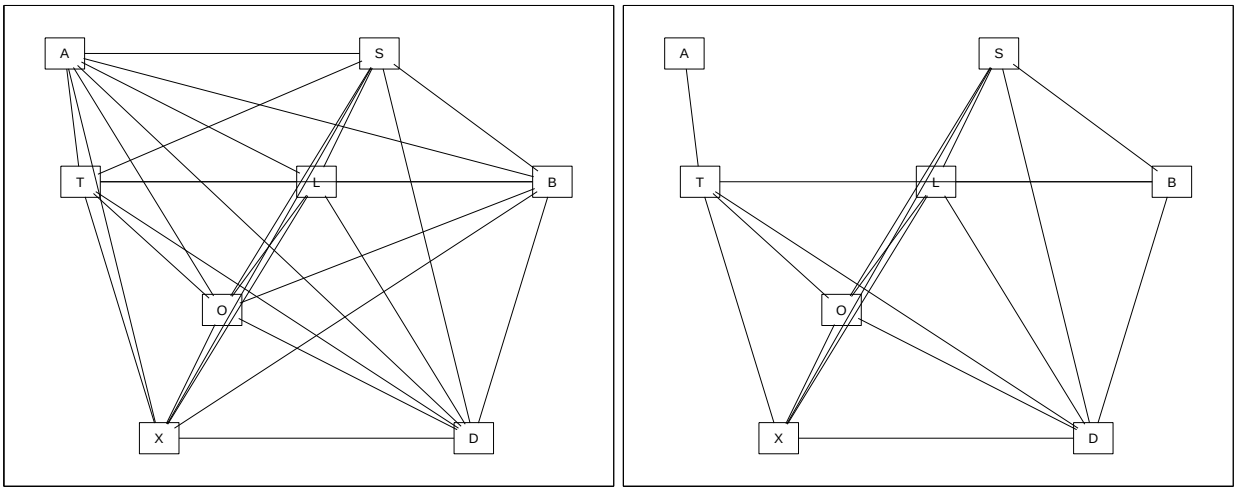
A gauche, le réseau "théorique" utilisé pour générer 5000 exemples.



Algorithme PC

Etape 1a : Suppression des IC d'ordre 0

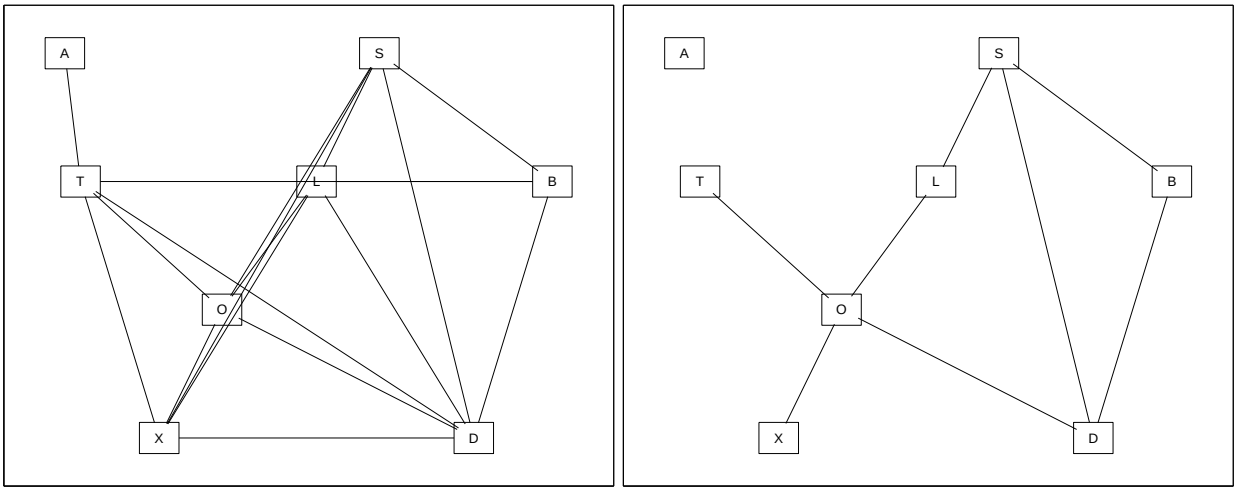
χ^2: $S \perp A$ $L \perp A$ $B \perp A$ $O \perp A$ $X \perp A$ $D \perp A$ $T \perp S$ $L \perp T$ $O \perp B$ $X \perp B$



Algorithme PC

Etape 1b : Suppression des IC d'ordre 1

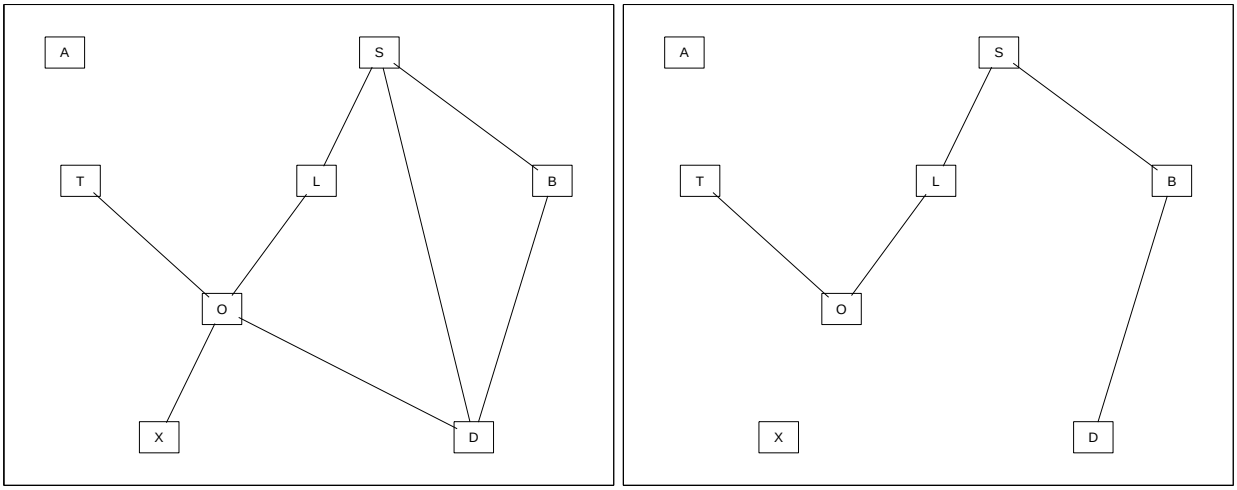
χ^2: $T \perp A|O$ $O \perp S|L$ $X \perp S|L$ $B \perp T|S$ $X \perp T|O$ $D \perp T|O$ ...



Algorithme PC

Etape 1c : Suppression des IC d'ordre 2

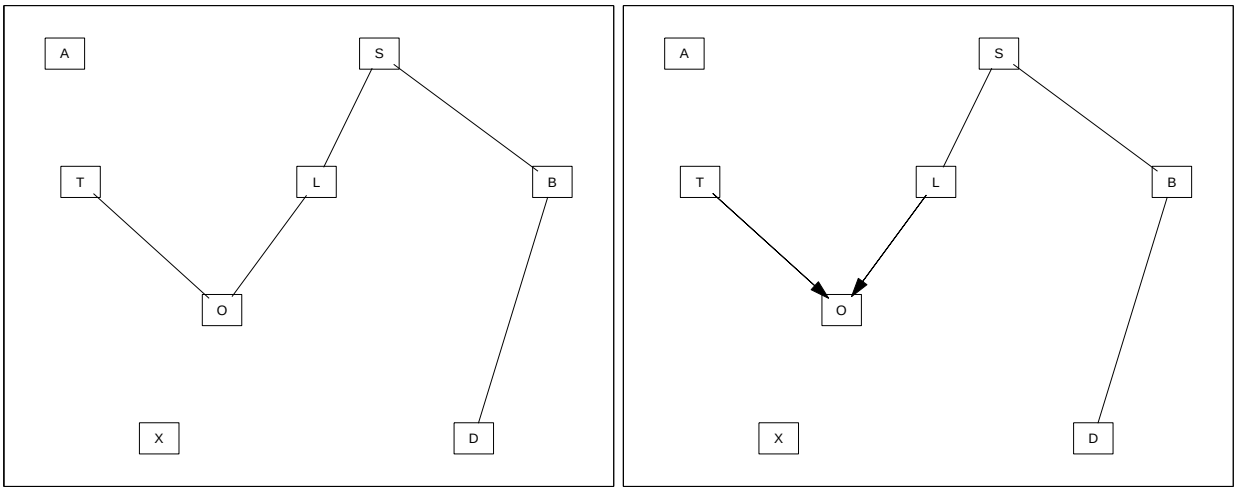
χ^2: $D \perp S | \{L, B\}$ $X \perp O | \{T, L\}$ $D \perp O | \{T, L\}$



Algorithme PC

Etape 2 : Recherche des V-structures

χ^2 : découverte de la V-structure $T \rightarrow O \leftarrow L$



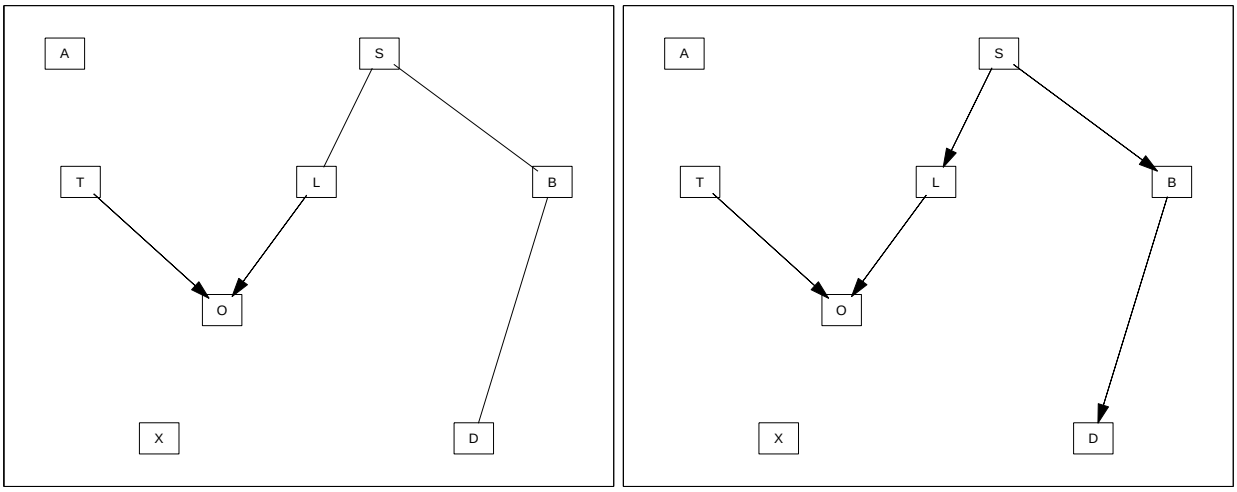
Etape 3 : Orientation récursive de certaines arêtes

aucune ici

Algorithme PC

Instanciation du PDAG

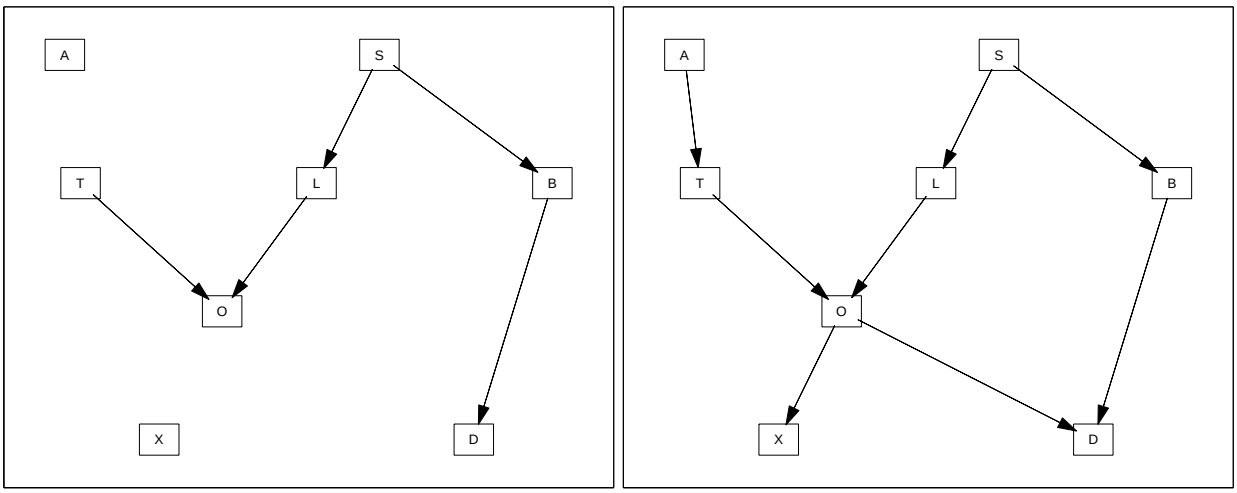
Orientation des arcs restants
(seule condition : ne pas introduire de nouvelle V-structure)



Algorithme PC

Réseau obtenu vs. théorique

Le test du χ^2 sur 5000 exemples n'a pas réussi à retrouver
$A \rightarrow T$, $O \rightarrow X$ et $O \rightarrow D$



Apprentissage (données complètes)

Recherche d'un bon réseau bayésien

- Première méthode : rechercher directement les indépendances conditionnelles
- Autre méthode : associer un "score" à chaque structure
  - calculable "rapidement" / décomposable localement

$$Score(B, \mathcal{D}) = \text{constante} + \sum_{i=1}^{n} score(X_i, pa_i)$$

  - notion de *score équivalence*
    - Un score S est dit *score equivalent* ssi pour deux structures B_1 et B_2 équivalentes on a $S(B_1, \mathcal{D}) = S(B_2, \mathcal{D})$.

Notion de score

Principe général : rasoir d'Occam

- *Pluralitas non est ponenda sine neccesitate*
  (La pluralité (des notions) ne devrait pas être posée sans nécessité)
- *Frustra fit per plura quod potest fieri per pauciora*
  (C'est en vain que l'on fait avec plusieurs ce que l'on peut faire avec un petit nombre)

= Principe de parcimonie = trouver le modèle

- qui représente le mieux les données $\mathcal{D}$:
  vraisemblance : $L(\mathcal{D}|\theta, B)$
- et qui soit le plus simple possible :
  nb de paramètres pour décrire B : $Dim(B)$

Exemples de score

AIC et BIC

- Compromis vraisemblance / complexité
- Application des critères AIC (Akaïke 70) et BIC (Schwartz 78)

$$S_{AIC}(B, \mathcal{D}) = \log L(\mathcal{D}|\theta^{MV}, B) - Dim(B)$$

$$S_{BIC}(B, \mathcal{D}) = \log L(\mathcal{D}|\theta^{MV}, B) - \frac{1}{2} Dim(B) \log N$$

Scores bayésiens : BD, BDe, BDeu

- $S_{BD}(B, \mathcal{D}) = P(B, \mathcal{D})$ (Cooper et Herskovits 92)
- $BDe = BD +$ *score équivalence* (Heckerman 94)

$$S_{BD}(B, \mathcal{D}) = P(B) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij} + \alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})}$$

Apprentissage (données complètes)

Recherche d'un bon réseau bayésien

- Heuristique de recherche :
  - espace $\mathcal{B}$
    - restriction aux arbres : Chow&Liu, MWST
    - ordonnancement des nœuds : K2
    - recherche gloutonne : Greedy Search
  - espace $\mathcal{E}$
    - Greedy Equivalence Search

Restriction à l'espace des arbres

Principe

- quel est le meilleur arbre passant par tous les nœuds, i.e. maximisant un score défini pour chaque arc possible ?

Réponse : Arbre de recouvrement maximal

- *MWST : Maximum Weight Spanning Tree*
  - (Chow et Liu 68) : information mutuelle :

$$W(X_A, X_B) = \sum_{a,b} \frac{N_{ab}}{N} \log \frac{N_{ab} N}{N_{a.} N_{.b}}$$

  - (Heckerman 94) : score local quelconque :

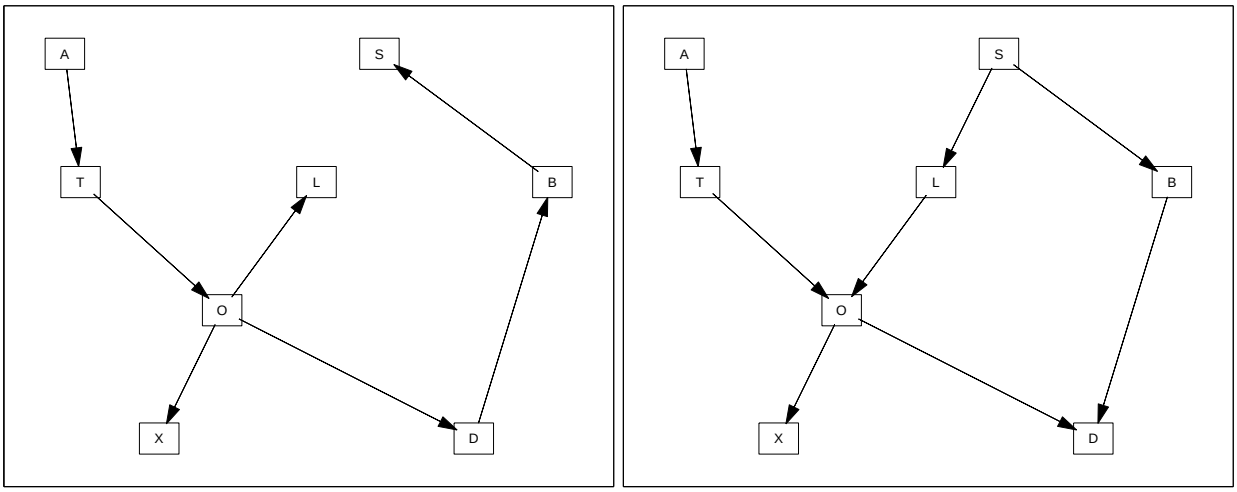
$$W(X_A, X_B) = score(X_A, Pa(X_A) = X_B) - score(X_A, \emptyset)$$

Restriction à l'espace des arbres

Déroulement

- MWST donne un arbre non orienté reliant toutes les variables.
- arbre non orienté = CPDAG représentant dans l'espace des équivalents de Markov de tous les arbres dirigés qui partagent cette même structure !
- transformation en arbre orienté en choisissant arbitrairement un nœud racine et en dirigeant chaque arête à partir de ce nœud.

Exemple : réseau obtenu vs. théorique



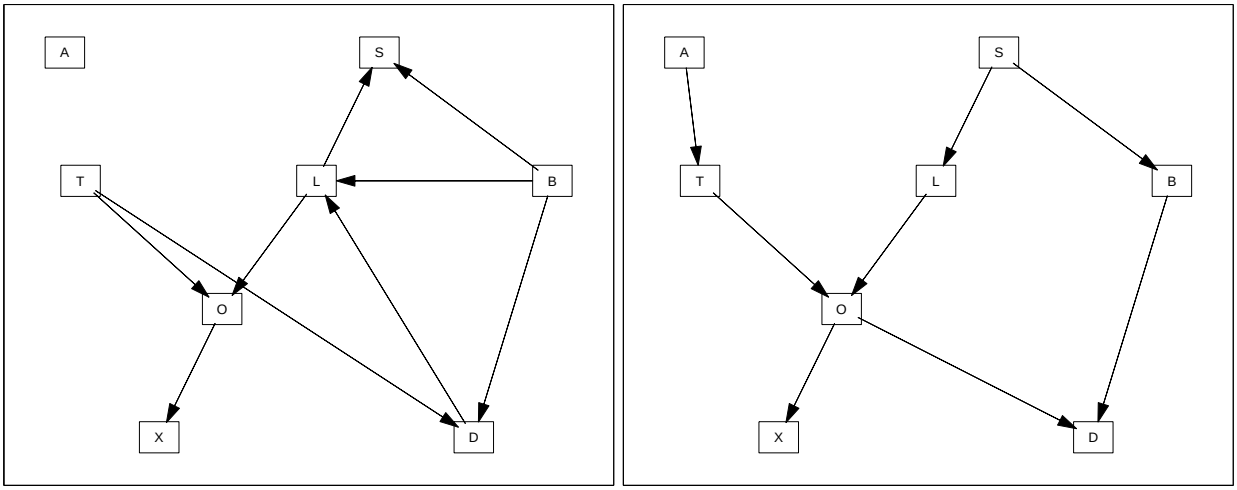
Ce type d'algorithme ne peut pas découvrir de V-structures, ni de cycles ...

Recherche gloutonne (greedy search)

Principe

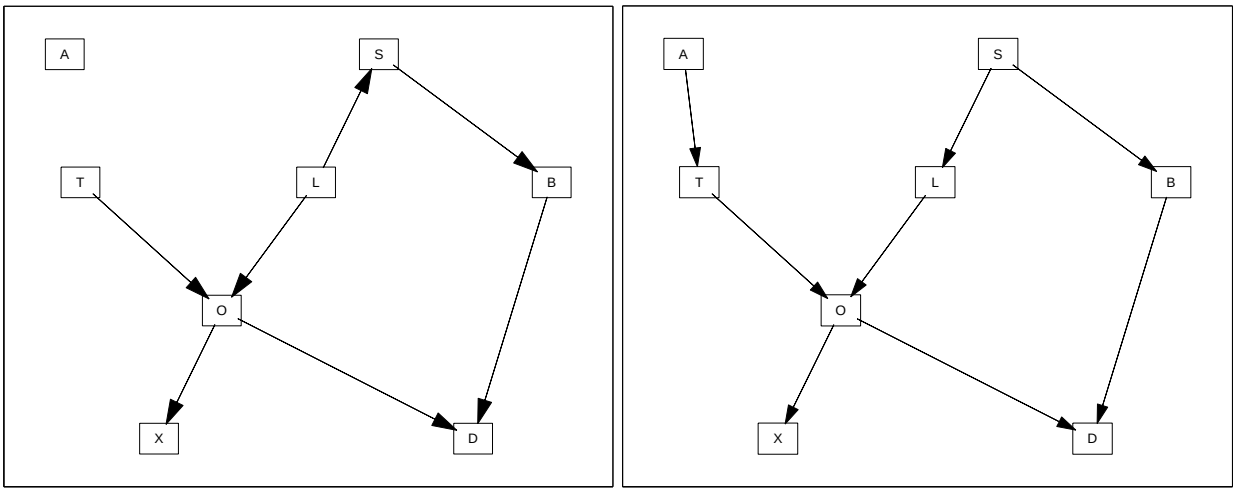
- Parcours de l'espace à l'aide d'opérateurs classiques :
  - ajout d'arc
  - inversion d'arc
  - suppression d'arc
- sous reserve que le graphe obtenu soit toujours un DAG (pas de circuit)
- possibilité de commencer à partir d'un graphe précis

Exemple : réseau obtenu vs. théorique



On tombe surement dans un optimum local

Exemple : réseau obtenu vs. théorique



Initialisation de la recherche par l'arbre obtenu par MWST :
on arrive à un meilleur résultat

Et avec des données incomplètes

Problème

= calculer le score lorsque les données sont incomplètes
$\mathcal{X} = \{\mathcal{D}, \mathcal{H}\}$

Une solution : Structural EM (Friedman 97)

$\simeq$ *Greedy Search* + *EM* sur les paramètres

- EM paramétrique pour améliorer $\theta^{(i)}$ pour un $B^{(i)}$ fixé
- recherche de $B^{(i+1)}$ parmi les voisins de $B^{(i)}$, avec des données complétées selon $\theta^{(i)}$
- et ainsi de suite ...

Et si on changeait d'espace de recherche

Remarques

- IC/PC : on obtient en réalité le PDAG représentant la classe d'équivalence de Markov
- MWST : idem (arbre non dirigé)
- La plupart des scores ne distinguent pas des réseaux équivalents, d'où des problèmes de convergence

Recherche dans $\mathcal{E}$

- $\mathcal{E}$ = espace des représentants des classes d'équiv. de Markov
- Meilleures propriétés : OUI
  - 2 structures équivalentes = une seule structure dans $\mathcal{E}$
- Meilleure taille : NON
  - $\mathcal{E}$ est quasiment de même taille que l'espace des RB (ratio asymptotique de 3,7 : Gillispie et Perlman 2001)

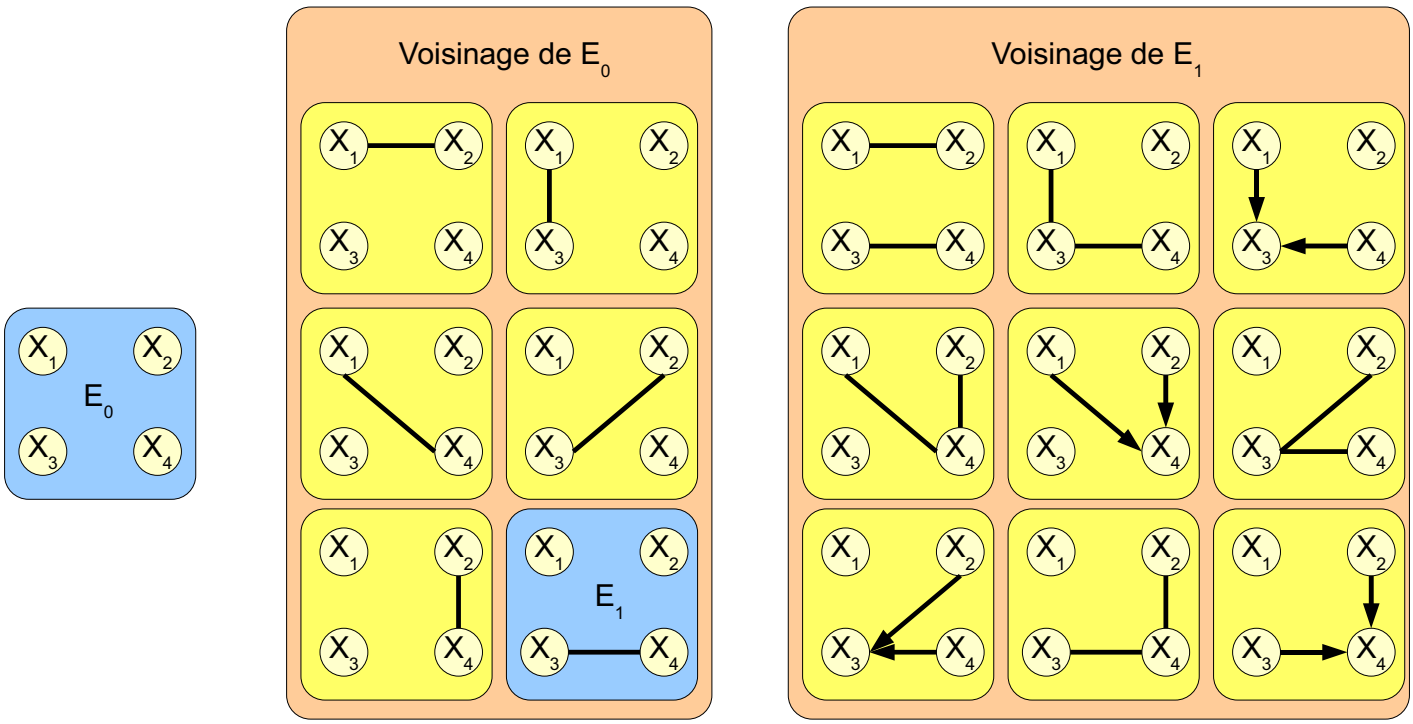
Greedy Equivalent Search

Principe (Chickering 2002)

- Recherche gloutonne dans $\mathcal{E}$
- Phase 1 : ajout d'arcs jusqu'à convergence
- Phase 2 : suppression d'arcs jusqu'à convergence

- Adaptation aux données incomplètes : GES-EM (Borchani et al. 2006)

Exemple d'ajout d'arcs dans $\mathcal{E}$



Références

- **Les Réseaux Bayésiens** - P. Naïm, P.H. Wuillemin, Ph. Leray, O. Pourret, A. Becker (Eyrolles) 2007
- **Probabilistic reasoning in Intelligent Systems: Networks of plausible inference** - J. Pearl (Morgan Kaufman) 1988
- **An introduction to Bayesian Networks** - F. Jensen (Springer Verlag) 1996
- **Probabilistic Networks and Expert Systems** - R.G. Cowell & al. (Springer Verlag) 1999
- **Learning Bayesian Networks** - R. Neapolitan (Prenctice Hall) 2003
- **Learning in Graphical Models** - Jordan M.I. ed. (Kluwer) 1998
- **An integral approach to causal inference with latent variables** - S. Maes et al. In Russo, F. and Williamson, J., editors, Causality and Probability in the Sciences. Texts In Philosophy series, London College Publications, pp 17-41. 2007

Réseaux bayésiens
introduction et apprentissage
modélisation et découverte de connaissances

Philippe LERAY
philippe.leray@univ-nantes.fr

Equipe COnnaissances et Décision
Laboratoire d'Informatique de Nantes Atlantique – UMR 6241
Site de l'Ecole Polytechnique de l'université de Nantes

Au programme ...

Matin — Notions générales

- Définition, D-séparation, Notion d'inférence

Matin — Apprentissage des paramètres

- Maximum de vraisemblance / a posteriori
- Données complètes / incomplètes

Après-midi — Apprentissage de la structure

- Recherche d'indépendances / maximisation score
- Quel espace ? Données complètes / incomplètes

Après-midi — $\Longrightarrow$ RB et causalité

- RB causal, intervention / observation, suffisance causale

Un RB n'est pas un modèle causal

- RB classique :
  - $A \rightarrow B$ ne signifie pas forcément causalité entre A et B,
  - seuls les arcs du CPDAG représentant de la classe d'équivalence de Markov représentent des causalités *

Confusion

- lorsque le graphe est construit par un expert, le graphe est souvent causal
- lorsque le graphe est appris avec des données, il n'a aucune raison d'être causal !

- Pas toujours grave ...
  - graphes équivalents $\Rightarrow$ même loi jointe, donc même résultat pour les algorithmes d'inférence (probabiliste)

  $\Rightarrow$ la causalité n'est pas utile pour l'inférence (probabiliste)

Réseau bayésien causal

Réseau bayésien causal

- chaque $A \rightarrow B$ représente une relation de causalité directe, i.e. le fait que A est bien la cause directe qui génère B

- si la causalité n'est pas utile pour l'inférence (probabiliste), à quoi peut servir un réseau bayésien causal ?

Intervention vs. Observation

- Inférence classique :
  - on observe $B = b$,
  - on calcule $P(A|B = b)$
- Inférence causale [Pearl 00]:
  - on agit/manipule/intervient sur B : $do(B = b)$

exemple avec $A \rightarrow B$

- $P(A|do(B = b)) = P(A)$,
- $P(B|do(A = a)) = P(B|A = a)$

exemple avec $A \leftarrow B$

- $P(A|do(B = b)) = P(A|B = b)$,
- $P(B|do(A = a)) = P(B)$

Manipulation Theorem

- Spécifier comment la loi jointe change après une manipulation $do(M = m)$

Version intuitive

- on oublie les causes "officielles" de M (ses parents dans le graphe)
- on garde le fait que $M = m$ pour les effets que cela déclenche (les enfants de M)

Version officielle [Spirtes et al. 00]

$$P(v|do(m)) = \left(\prod_{V_i \in V \setminus M} P(v_i|Pa(V_i)) \right)_{M=m}$$

Apprentissage d'une structure causale

- En général, utilisation de données *d'observation*
  - quelle que soit la méthode, résultat = représentant de la classe d'équivalence
  - détermination partielle des relations causales

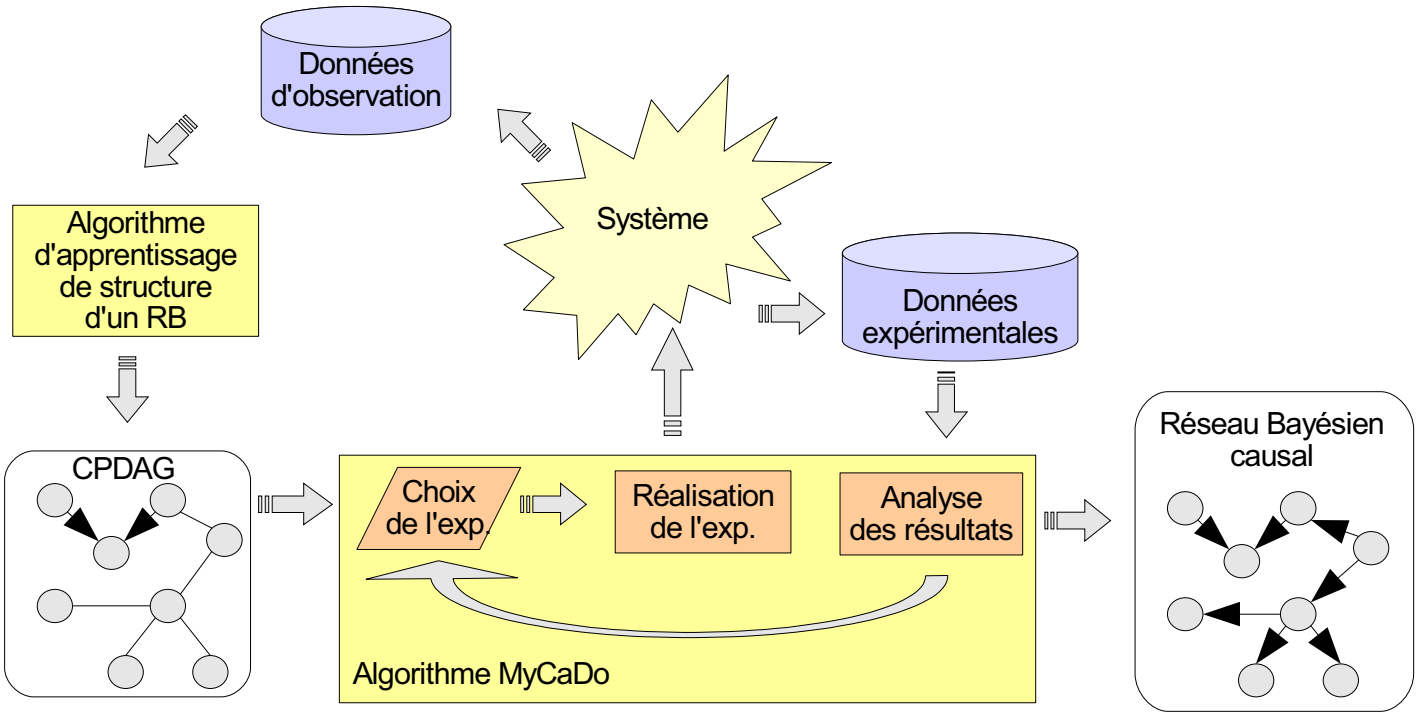
Solutions pour trouver un graphe complètement causal

- utiliser uniquement des données *d'expérimentation*, et décider au fur et à mesure quelle expérience sera la plus utile à réaliser (*active learning* [Murphy 01], ...)

Idée : algorithme MyCaDo [Meganck, Leray & Manderick 06]

- tirer partie des données d'observations souvent existantes et nombreuses
- utiliser des données d'expérimentation uniquement pour finir d'orienter le CPDAG

Algorithme MyCaDo



Algorithme MyCaDo

1. Choix de l'expérience = choix d'une variable M à manipuler
   - orientant potentiellement le plus d'arcs
   - en tenant compte d'éventuels coûts d'expérimentation et/ou d'observation des variables
2. Réalisation de l'expérience
   - $do(M = m)$ pour toutes les valeurs possibles m
   - observation des variables C candidates (C–M)
3. Analyse des résultats
   - $P(C|M)$ (observation) $\simeq P(C|do(M))$ (expérience) ?
   - si égalité, alors $C \leftarrow M$, sinon $M \leftarrow C$
   - propagation éventuelle de l'arc découvert

Mais ce n'est pas fini ...

Exemple simple, avec 2 variables

- S (la Seine déborde) et P (j'ai pris mon parapluie)
- Des données d'observation montrent que ces deux variables ne sont pas indépendantes :

S–P

- On décide d'agir sur S et d'observer P : pas de modification

$\Rightarrow$ S n'est pas la cause de P

- Faut-il en conclure que P est la cause de S ?
- En agissant aussi sur P, on aurait vu que P n'est pas la cause de S
- Intérêt = découverte d'une variable latente (il pleut...)

Suffisance Causale

- Les algorithmes précédents se basent tous sur l'hypothèse de suffisance causale

Suffisance causale

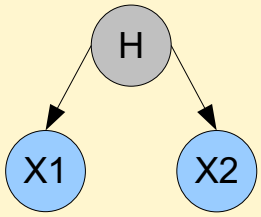
Toutes les variables nécessaires à la modélisation sont connues

- Abandonner l'hypothèse de suffisance causale = Essayer de découvrir des variables latentes lors de l'apprentissage de structure
  - de façon explicite (méthodes à base de score)
  - de façon implicite (SMCM vs. MAG)

Modélisation explicite vs. implicite

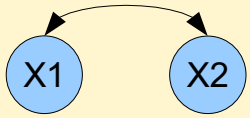
Modélisation explicite

- Adaptation de Structural EM
- Avantages
  - inférence probabiliste : OK
- Inconvénients
  - complexité de la méthode
  - inférence causale : NON (le graphe n'est pas causal)



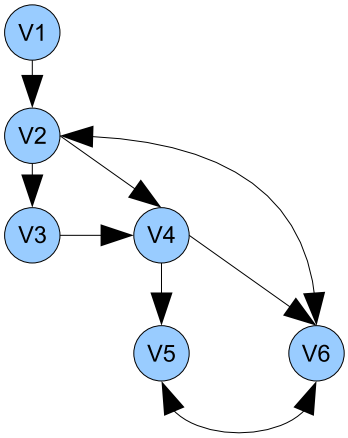
Modélisation implicite

- Modèle plus "léger"
  - pas besoin de déterminer la cardinalité de H
- Deux formalismes aux objectifs différents
  - inférence causale : SMCM, Semi Markovian Causal Model
  - apprentissage de la structure : MAG, Maximum Ancestral Graph



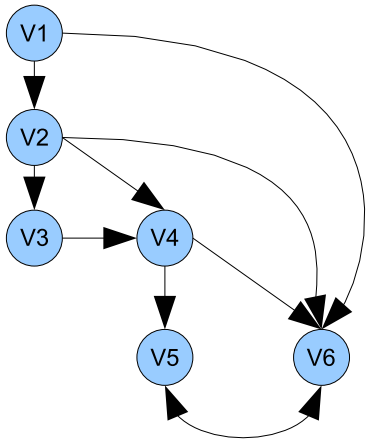
SMCM vs. MAG

- SMCM [Pearl 00]



- $A \leftrightarrow B$: cause commune latente
- $A \rightarrow B$: relation de causalité directe

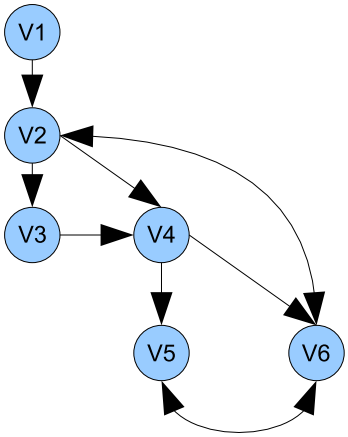
- MAG [Richardson & Spirtes 02]



- $A \leftrightarrow B$: cause commune latente
- $A \rightarrow B \Rightarrow$ dépendance entre A et B
- existence de chemins "induits"

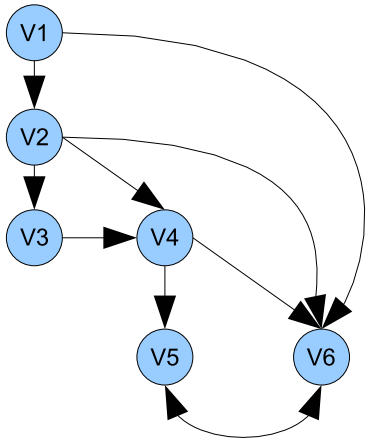
SMCM vs. MAG

- SMCM [Pearl 00]



- Inf. causale : en théorie
- Inférence prob. : NON
- Apprent. structure : NON

- MAG [Richardson & Spirtes 02]



- Inf. causale : partielle
- Inférence prob. : NON
- Apprent. structure : partielle

SMCM vs. MAG

- **Apprentissage** à partir d'observations : OK, mais obtention du représentant de la classe d'équivalence (CPAG)

  CPAG $\longrightarrow$ MAG : inutile, un MAG n'est pas causal

- **Inférence causale** : OK dans les SMCM
- **Inférence probabiliste** : il manque une paramétrisation efficace des SMCM

Une approche globale : MyCaDo++

- **Apprentissage** à partir d'observations : OK, mais obtention du représentant de la classe d'équivalence (CPAG)

Notre idée : [Meganck, Maes, Leray & Manderick 06]

passer directement du CPAG à un SMCM à partir de données d'expérimentation

- **Inférence causale** : OK dans les SMCM
- **Inférence probabiliste** : il manque une paramétrisation efficace des SMCM

Notre idée : [Meganck, Maes, Leray & Manderick 06]

proposer une paramétrisation efficace d'un SMCM

Références

- **Les Réseaux Bayésiens** - P. Naïm, P.H. Wuillemin, Ph. Leray, O. Pourret, A. Becker (Eyrolles) 2007
- **Causality: Models, Reasoning, and Inference** - J. Pearl (Cambridge University Press) 2000
- **An introduction to Bayesian Networks** - F. Jensen (Springer Verlag) 1996
- **Probabilistic Networks and Expert Systems** - R.G. Cowell & al. (Springer Verlag) 1999
- **Learning Bayesian Networks** - R. Neapolitan (Prenctice Hall) 2003
- **Learning in Graphical Models** - Jordan M.I. ed. (Kluwer) 1998
- **An integral approach to causal inference with latent variables** - S. Maes et al. In Russo, F. and Williamson, J., editors, Causality and Probability in the Sciences. Texts In Philosophy series, London College Publications, pp 17-41. 2007