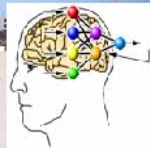


Apprentissage de la structure des réseaux bayésiens : de la construction d'un modèle graphique probabiliste à la recherche de causalité



Philippe LERAY

`philippe.leray@univ-nantes.fr`

Equipe COonnaissances et Décision

Laboratoire d'Informatique de Nantes Atlantique – UMR 6241

Site de l'Ecole Polytechnique de l'université de Nantes

Au programme ...

RB : Introduction

- Définition, Notion d'inférence, Apprentissage des paramètres

Apprentissage de la structure

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

RB et causalité

- RB causal, intervention / observation, suffisance causale

Un peu d'histoire

RULE037

IF the organism

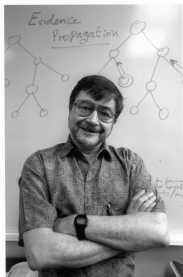
- 1) stains grampos
- 2) has coccus shape
- 3) grows in chains

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 $X=\text{vrai}$ et $Y=\text{absent}$ alors $Z=\text{faux}$
- moteur d'inférence (chainage avant, arrière)



Judea Pearl (1936–) : les réseaux bayésiens

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

Un peu d'histoire



2008 : Une reconnaissance

Judea Pearl obtient la Benjamin Franklin Medal in Computer and Cognitive Science pour

- la création des premiers algorithmes de calculs et de raisonnement avec des observations incertaines
- permettant la découverte des associations inconnues et des relations causales cachées parmi des millions d'observations
- avec un impact sur un large champ de problèmes en science et en ingénierie.

Définition d'un réseau bayésien

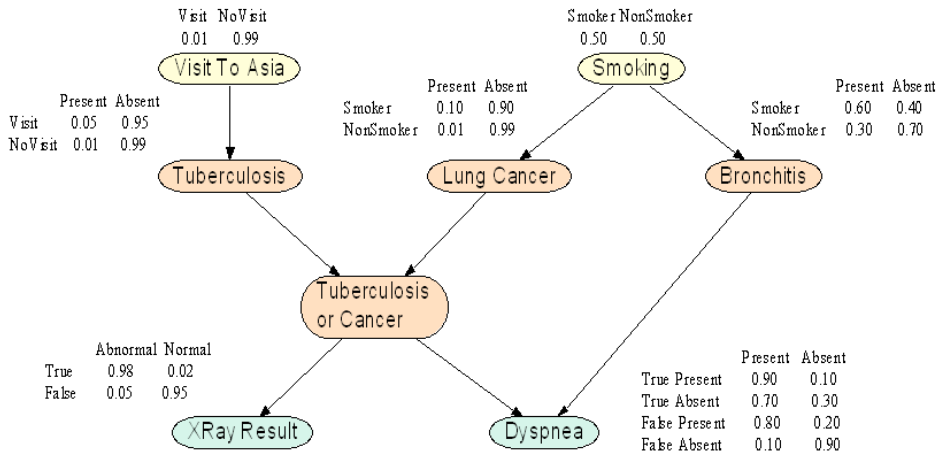
Définition

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

Conséquence

- $P(S) = \prod_{i=1}^n P(S_i | \text{parents}(S_i))$
- La loi jointe (globale) se décompose en un produit de lois conditionnelles locales
- RB = représentation compacte de la loi jointe $P(S)$

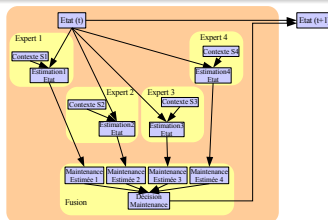
Exemple



Intérêts et motivation

Intérêts des réseaux bayésiens

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



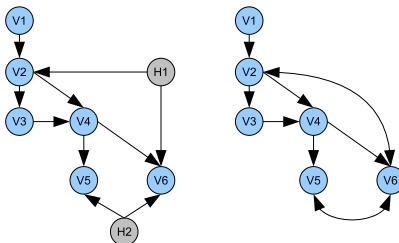
Motivation

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

Intérêts et motivation

Autre intérêt

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



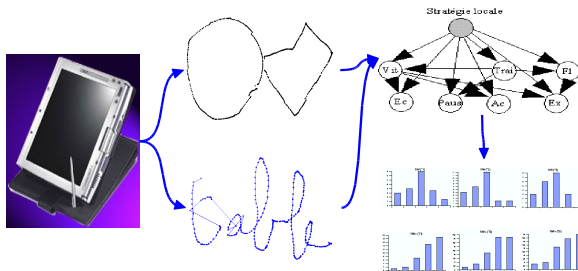
Motivation

- 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 = "modèle de dépendance"

d-séparation

- critère graphique pour déterminer si deux variables sont indépendantes conditionnellement à un ensemble de variables.
- (non décrit ici)

"sous-structure" remarquable: $X \longrightarrow Z \longleftarrow Y$ (V-structure)

- X et Y indépendants
- X et Y dépendants conditionnellement à Z

$X \longrightarrow Z \longrightarrow Y$ ou $X \longleftarrow Z \longrightarrow Y$

- X et Y dépendants
- X et Y indépendants conditionnellement à Z

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

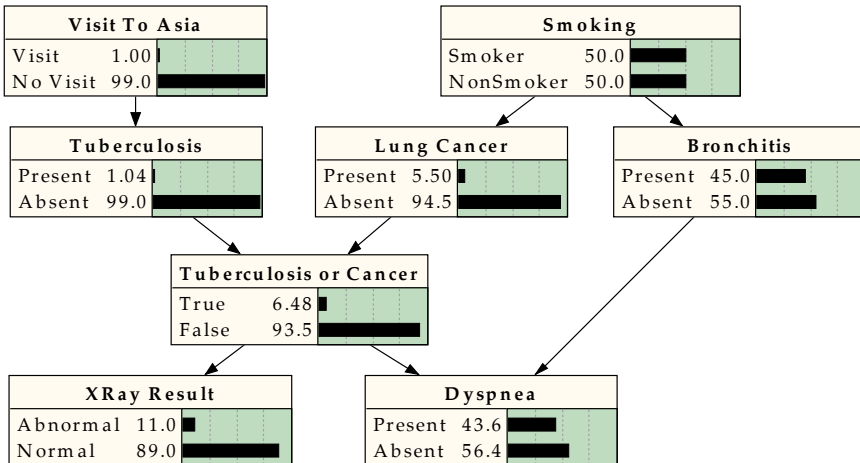
- Message Passing (Pearl 1988) pour les arbres
- Junction Tree (Jensen 1990)
- Shafer-Shenoy (1990)

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

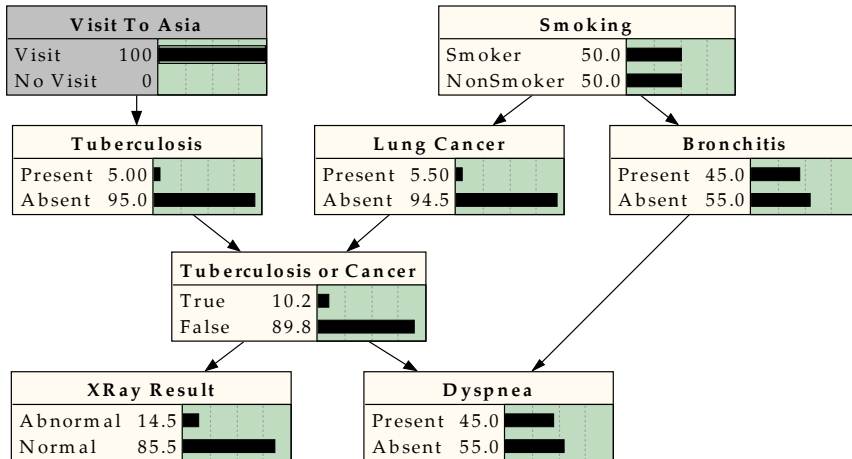
Algorithmes approchés

- Échantillonnage
- Méthodes variationnelles

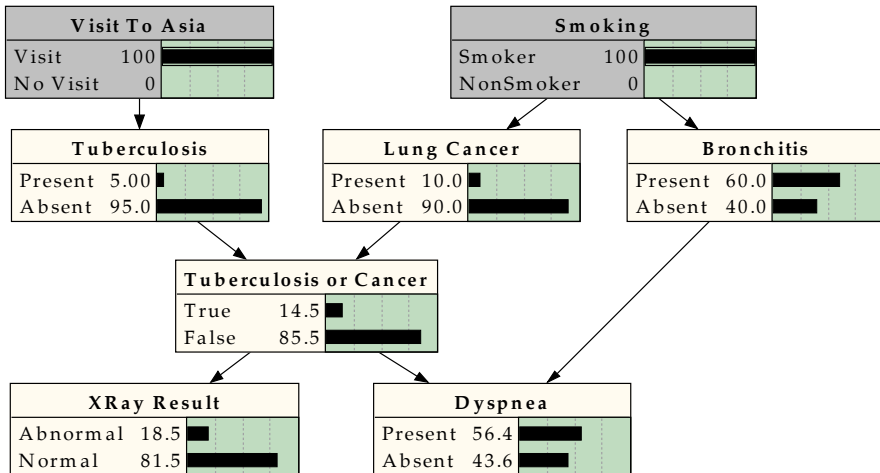
Exemple



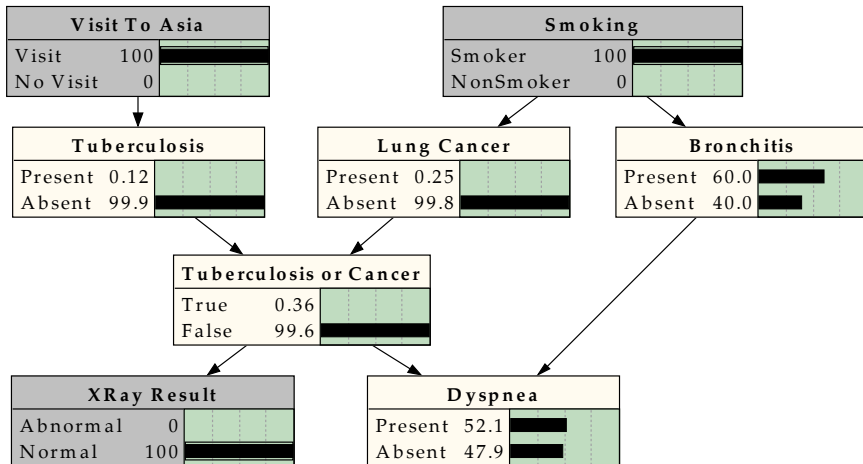
Exemple



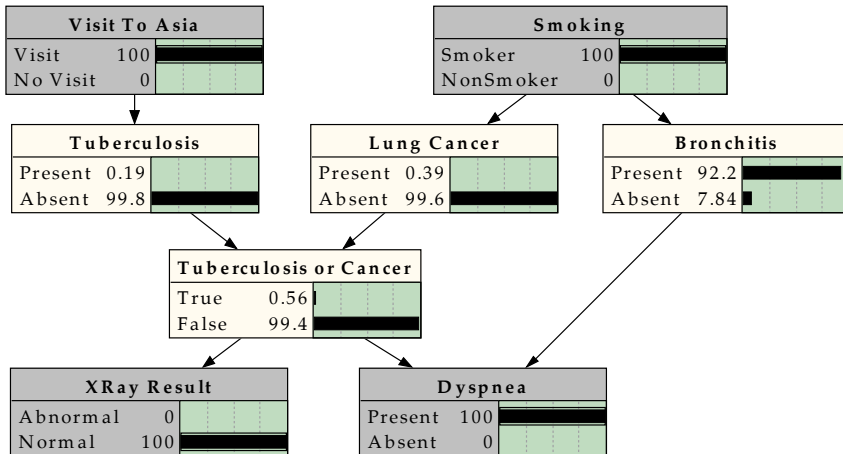
Exemple



Exemple



Exemple



Notion d'apprentissage des paramètres

structure fixée, on cherche seulement les CPD

à partir d'expertises

- élicitation de connaissances

à partir de données complètes

- maximum de vraisemblance
- maximum a posteriori

à partir de données incomplètes

- complete case analysis, available case analysis
- expectation maximisation

Au programme ...

RB : Introduction

- Définition, Notion d'inférence, Apprentissage des paramètres

Apprentissage de la structure

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

RB et causalité

- RB causal, intervention / observation, suffisance causale

Problème complexe

Taille de l'espace de recherche

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

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

$$NS(5) = 29281$$

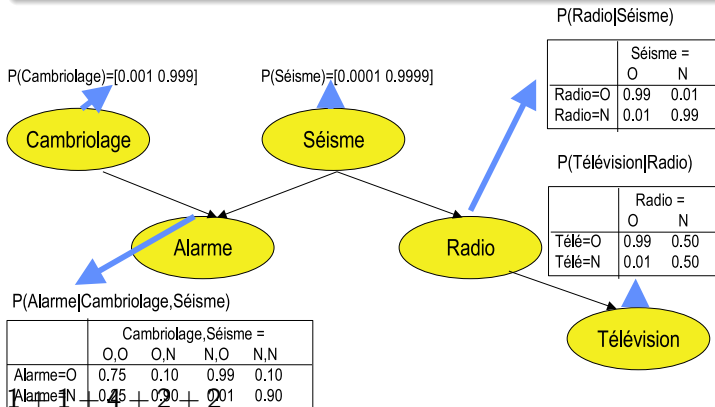
$$NS(10) = 4.2 \times 10^{18}$$

- 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



Equivalence de Markov

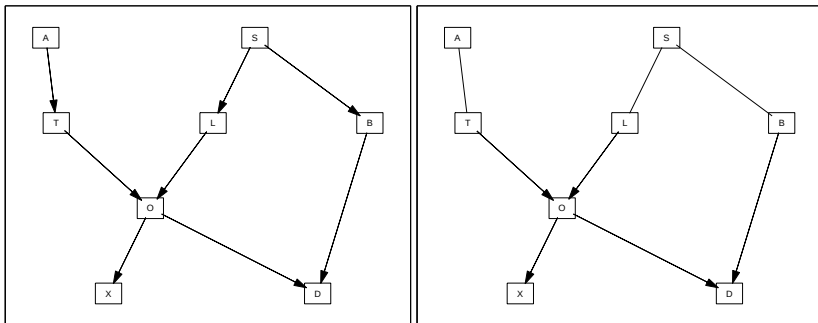
Définition

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

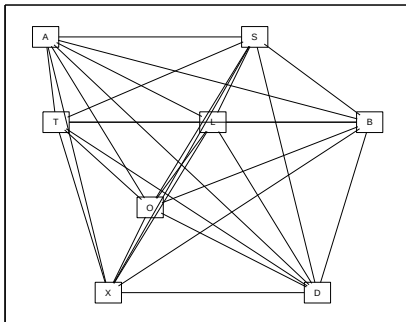
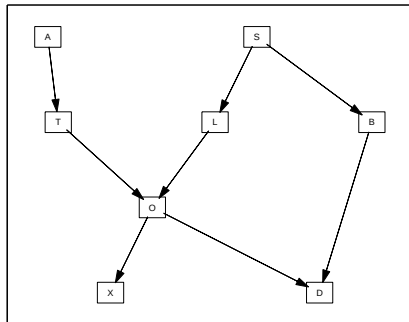
Problèmes principaux

- 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

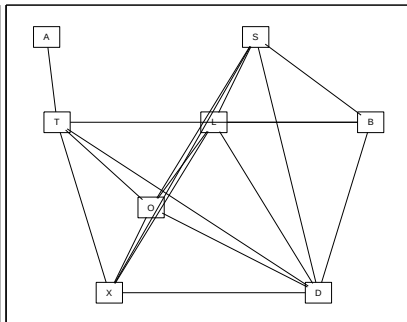
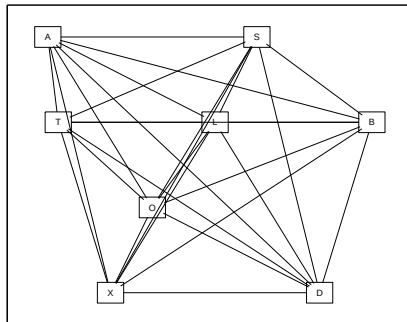
A gauche, le réseau "théorique" utilisé pour générer 5000 exemples.



Algorithme PC

Etape 1a : Suppression des IC d'ordre 0

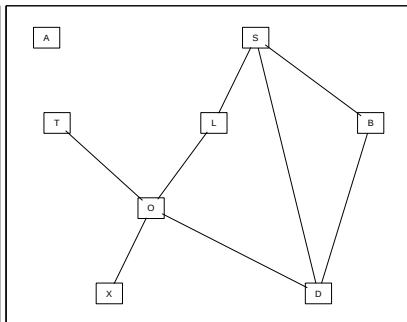
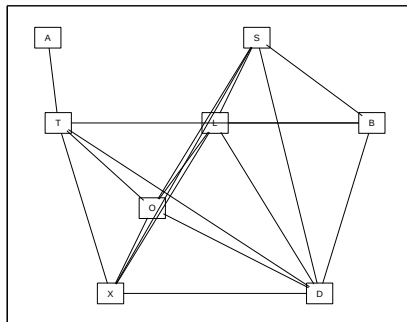
χ^2 : $S \perp A$ $L \perp A$ $B \perp A$ $O \perp A$ $X \perp A$ $D \perp A$ $T \perp S$ $L \perp T$ $O \perp B$ $X \perp B$



Algorithme PC

Etape 1b : Suppression des IC d'ordre 1

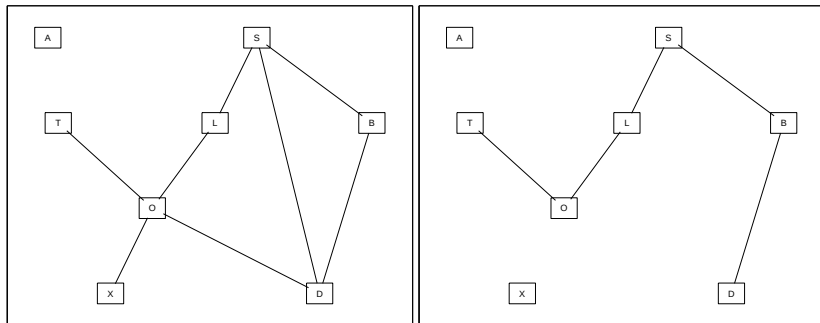
χ^2 : $T \perp A | O$ $O \perp S | L$ $X \perp S | L$ $B \perp T | S$ $X \perp T | O$ $D \perp T | O \dots$



Algorithme PC

Etape 1c : Suppression des IC d'ordre 2

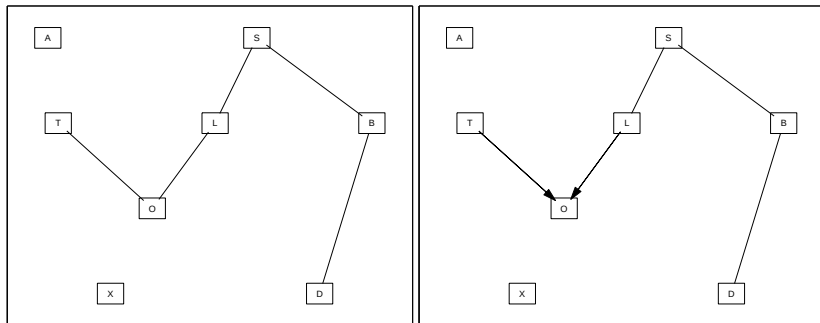
$\chi^2: D \perp S | \{L, B\} \quad X \perp O | \{T, L\} \quad D \perp O | \{T, L\}$



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

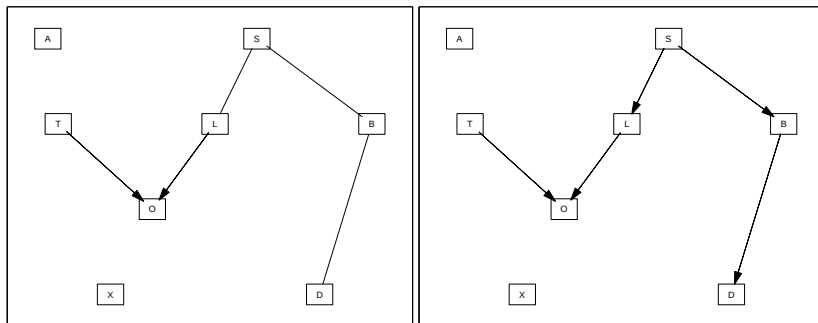
aucune ici

Algorithme PC

Instanciation du PDAG

Orientation des arcs restants

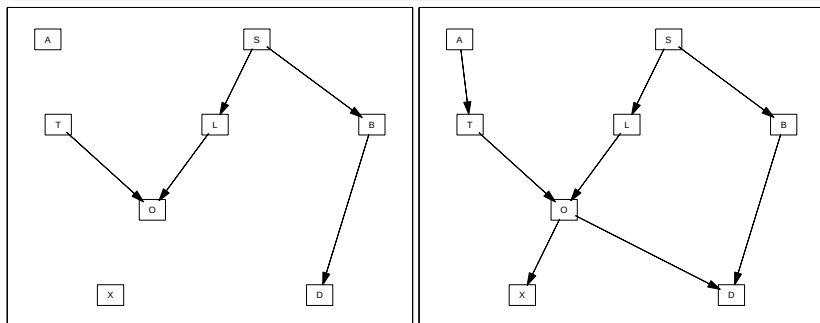
(seule condition : ne pas introduire de nouvelle V-structure)



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 équivalent* 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 (Akaike 70) et BIC (Schwartz 78)

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

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

Scores bayésiens : BD, BDe, BDeu

- $S_{BD}(B, \mathcal{D}) = P(B, \mathcal{D})$ (Cooper et Herskovits 92)
- $BDe = BD + \text{score équivalence}$ (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
 - algorithmes génétiques
 - 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

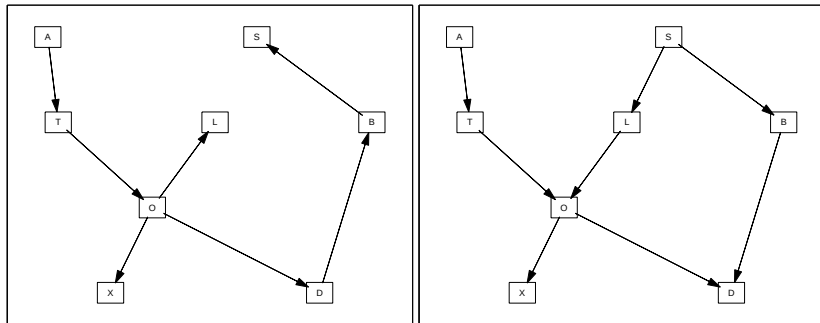
- *MWST : Maximum Weight Spanning Tree*
 - (Chow et Liu 68) : information mutuelle :

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

- (Heckerman 94) : score local quelconque :

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

Exemple : réseau obtenu vs. théorique



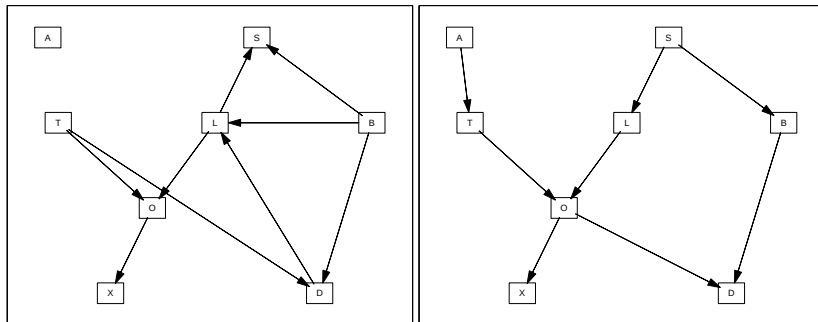
Ce type d'algorithme ne peut pas découvrir de V-structures, ni de cycles ...

Recherche gloutonne (greedy search)

Principe

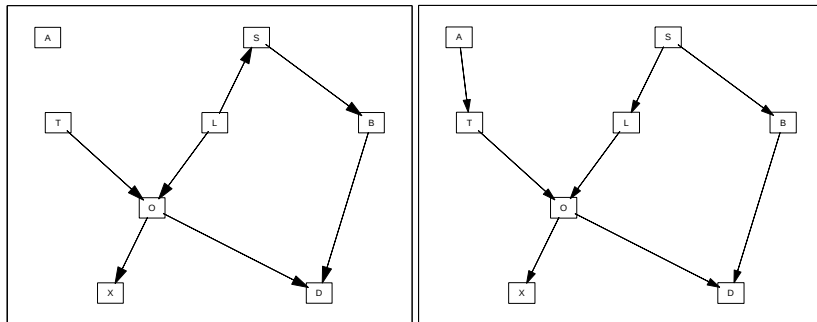
- Parcours de l'espace à l'aide d'opérateurs classiques :
 - ajout d'arc
 - inversion d'arc
 - suppression d'arc
- sous réserve que le graphe obtenu soit toujours un DAG (pas de circuit)
- possibilité de commencer à partir d'un graphe précis

Exemple : réseau obtenu vs. théorique



On tombe sûrement 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)

≈ *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 $\mathcal{E}$

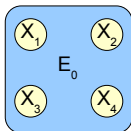
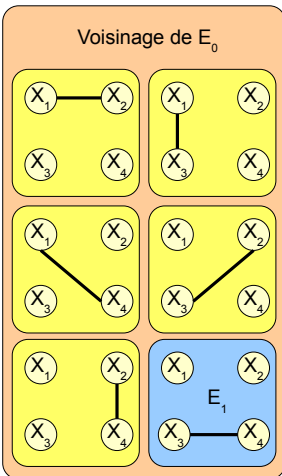
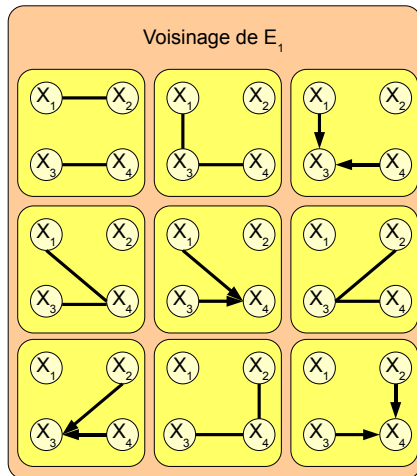
- $\mathcal{E}$ = espace des représentants des classes d'équiv. de Markov
- Meilleures propriétés : OUI
 - 2 structures équivalentes = une seule structure dans $\mathcal{E}$
- Meilleure taille : NON
 - $\mathcal{E}$ est quasiment de même taille que l'espace des RB (ratio asymptotique de 3,7 : Gillispie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

Voisinage de E_0 Voisinage de E_1 

Au programme ...

RB : Introduction

- Définition, Notion d'inférence, Apprentissage des paramètres

Apprentissage de la structure

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

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

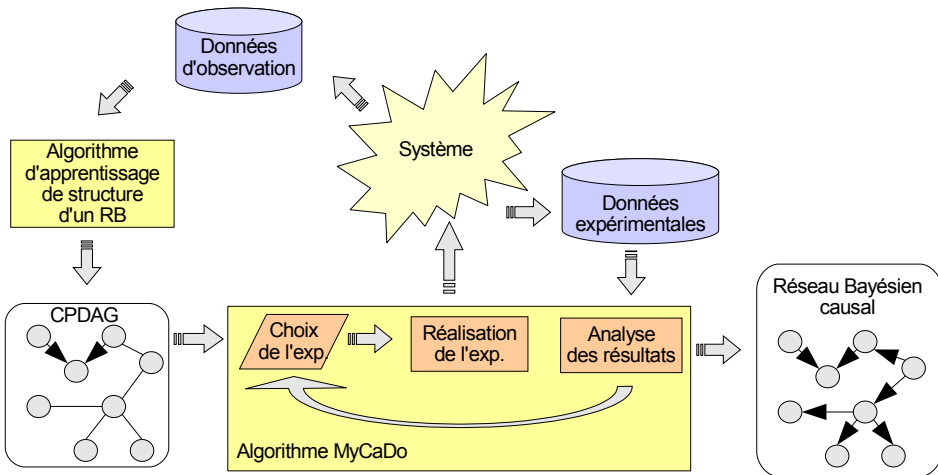
- utiliser uniquement des données *d'expérimentation*, et décider au fur et à mesure quelle expérience sera la plus utile à réaliser (*active learning* [Murphy 01], ...)

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

- ❶ 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
- ❷ Réalisation de l'expérience
 - $do(M = m)$ pour toutes les valeurs possibles m
 - observation des variables C candidates ($C-M$)
- ❸ 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 :

$$S \rightarrow P$$

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

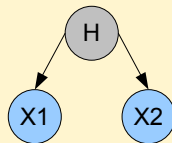
Toutes les variables nécessaires à la modélisation sont connues

- 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

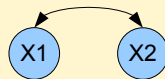
Modélisation explicite

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



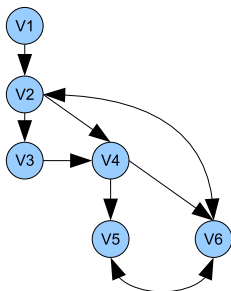
Modélisation implicite

- 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



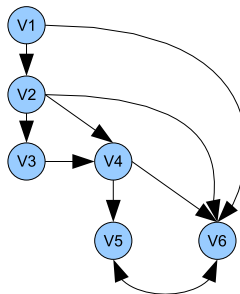
SMCM vs. MAG

• SMCM [Pearl 00]



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

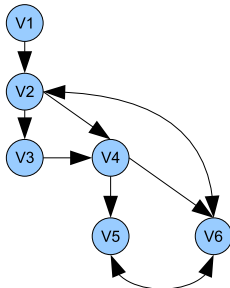
• MAG [Richardson & Spirtes 02]



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

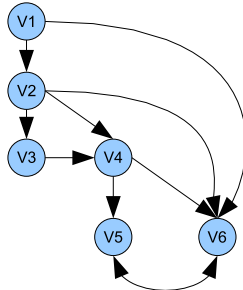
SMCM vs. MAG

• 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 $\rightarrow$ 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

Causalité et sélection de variables

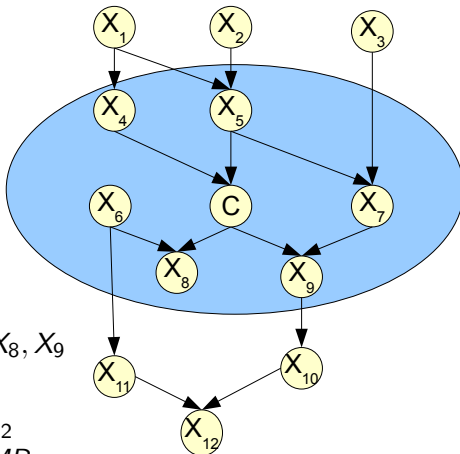
Principe

- Faire bénéficier la problématique de sélection de variables de connaissances causales sur la tâche à résoudre
- Ex : *Decès* très corrélé à *Cancer*, mais var. inutile pour prédire la maladie
- Plus généralement, quelles sont les variables nécessaires par rapport à une variable cible ?

Couverture de Markov

- MB = ensemble des parents, enfants et autres parents des enfants
- C est indépendant des X_j (non MB) conditionnellement à MB

Couverture de Markov



- $MB = X_4, X_5, X_6, X_7, X_8, X_9$
- C est indépendant de $X_1, X_2, X_3, X_{10}, X_{11}, X_{12}$ conditionnellement à MB
- $P(C|X_1...X_{12}) = P(C|MB)$

Causalité et sélection de variables

Algorithmes de sélection de variables

- IAMB (Aliferis 2002)
- MMPC (Tsamardinos et al. 2003), ...

Algorithmes d'apprentissage de structure

- MMHC (Tsamardinos et al. 2006)

Références



- **Les Réseaux Bayésiens** - P. Naïm, P.H. Willemin, 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 (Prentice 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