INSTITUT FRANCAIS
DES SCIENCES
ET TECHNOLOGIES
DES TRANSPORTS,
DE L'AMENAGEMENT
ET DES RESEAUX

Introduction à l'apprentissage de structure de *Réseaux Bayésiens*.

UEC1 : Apprentissage artificiel avancé

Olivier FRANÇOIS

IFSTTAR

GRETTIA – 21 novembre 2011

Plan du cours : Partie I Introduction et Objectifs

- Exemple introductif
- Motivation des modèles graphiques probablistes
- Problématique du cours

Plan du cours : Partie II Apprentissage et généralisation

- Erreur théorique et erreur empirique
- Sur-apprentissage
- Généralisation
  - Ensemble de validation
  - Ensemble de test
- Rasoir d'Occam
- Dilemme Biais/Variance
- Régularisation
- Validation Croisée

Plan du cours : Partie III Rappel de Probabilités

- Indépendance conditionnelle
- Interprétation fréquentiste
- Dilemme du Prisonnier
- Probabilités des causes et Théorème de Bayes

Plan du cours : Partie IV Réseaux Bayésiens

- Définitions des Modèles Graphiques Probabilistes
- Raisonnement probabiliste
- Réseaux bayésiens : Définition
- Indépendance conditionnelle et Théorème de Bayes
- La d-séparation

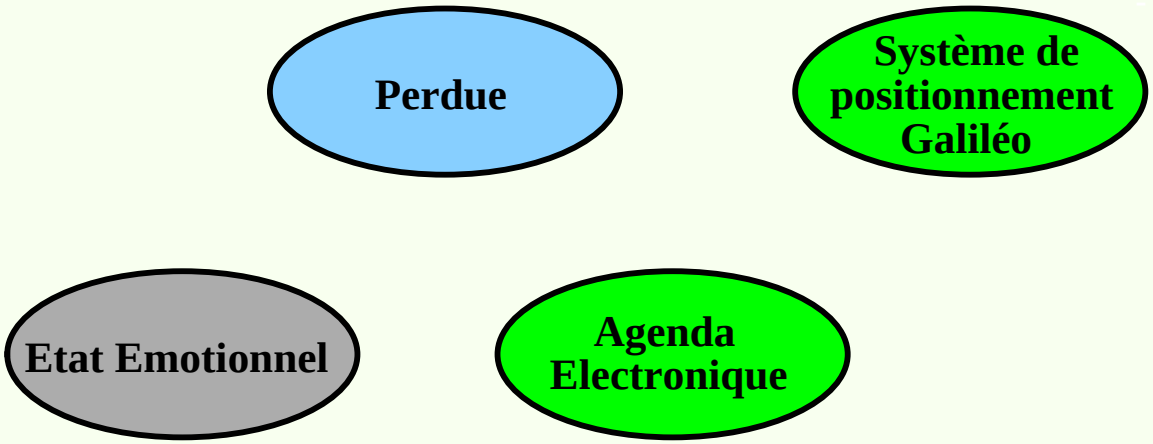
Plan du cours : Partie V Inférence dans les RB

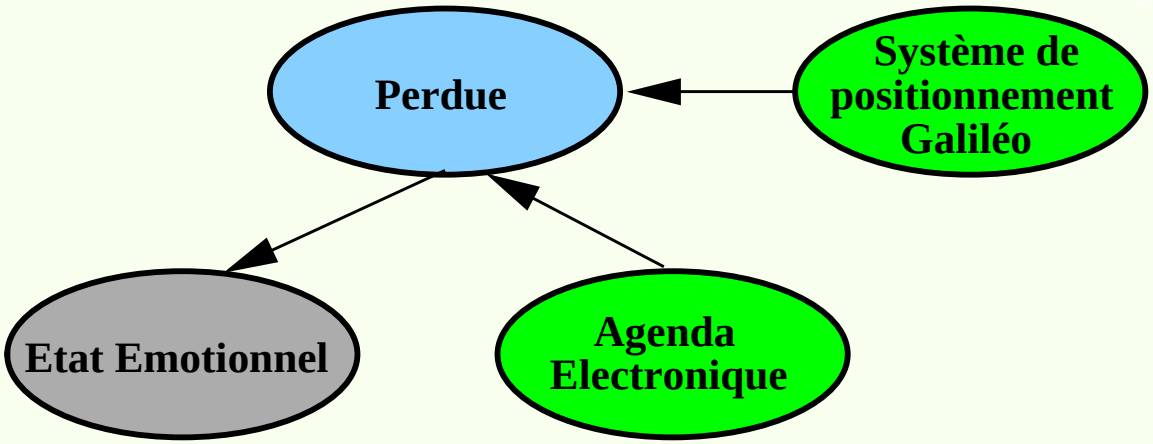
- Marginalisation : *Bucket Elimination*
- *Message Passing* de Pearl
- *Junction tree* de Jensen
- Exemple d'application

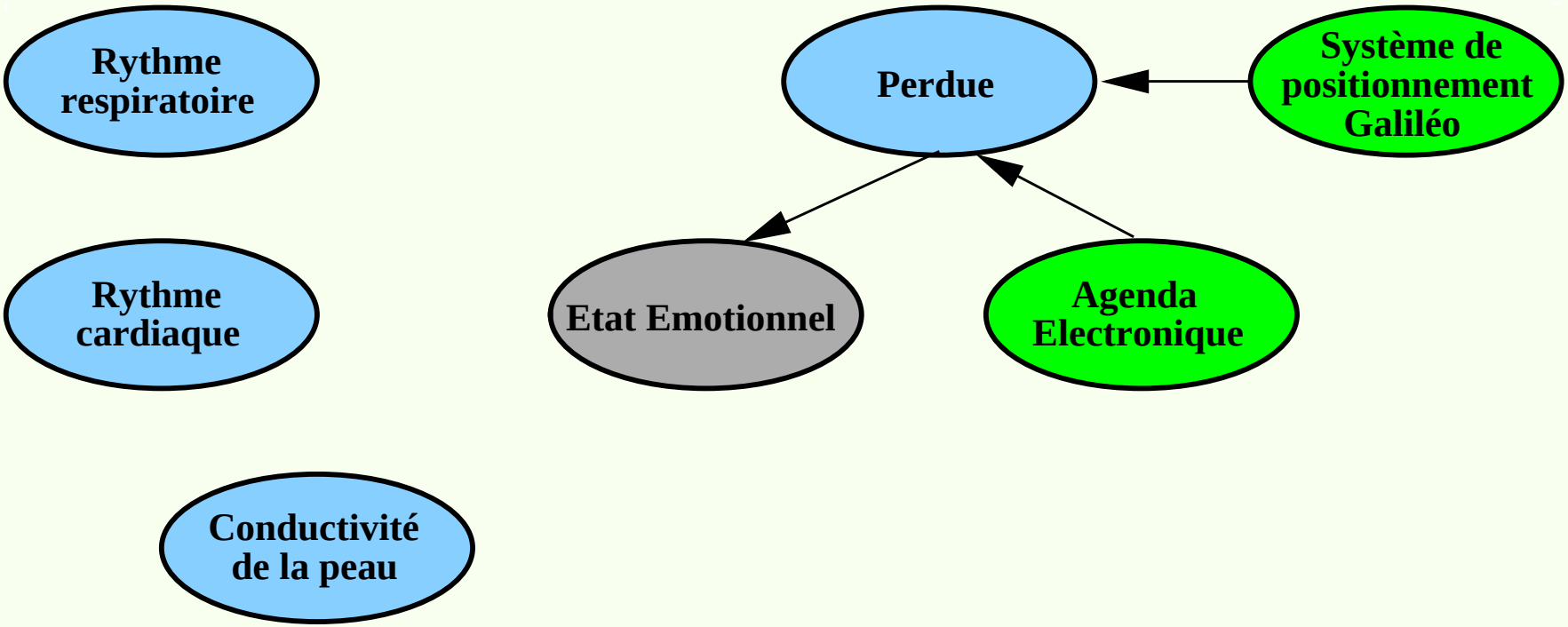
Plan du cours : Partie VI Apprentissage de RB

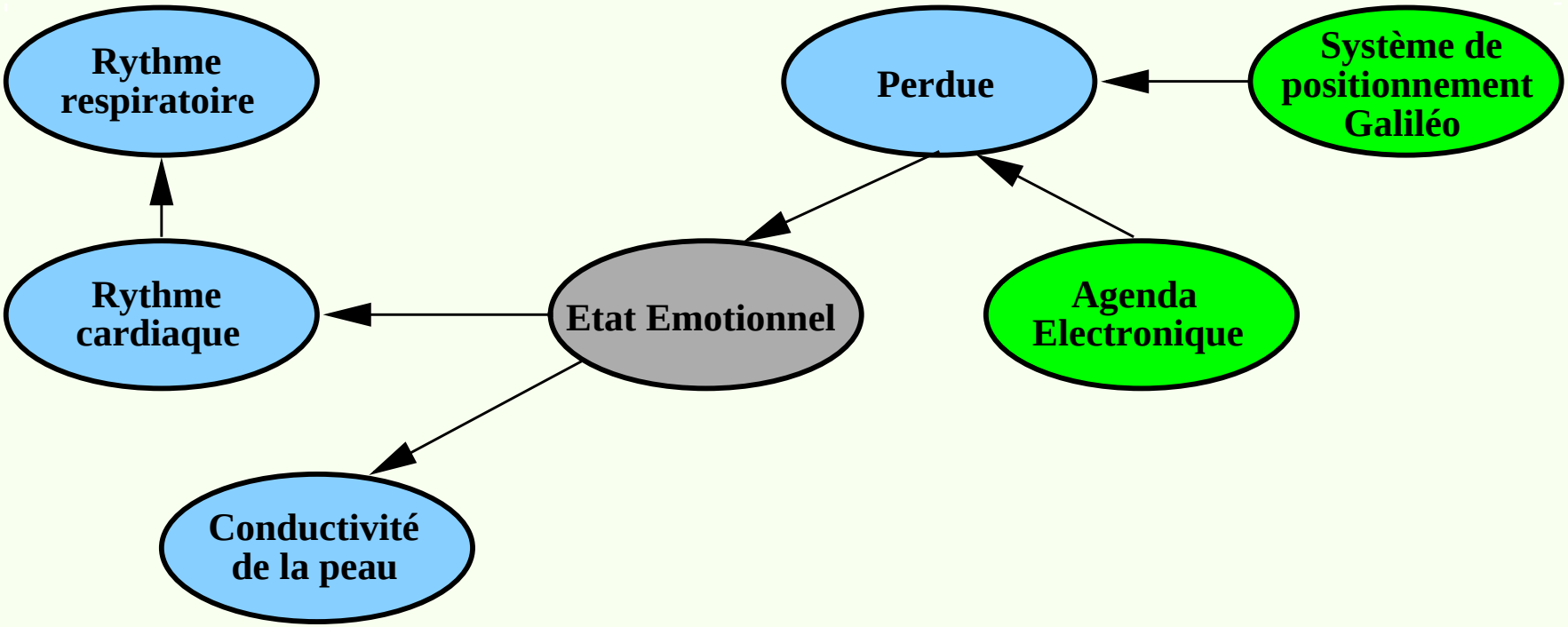
- Apprentissage des paramètres
  - Approche Statistique
  - Approche Bayésienne
- Algorithme EM
- Apprentissage de Structure
  - Espace de recherche
  - Notion de Score
  - Recherche de causalité
  - Algorithme K2
  - Recherche Gloutonne
- Variables latentes

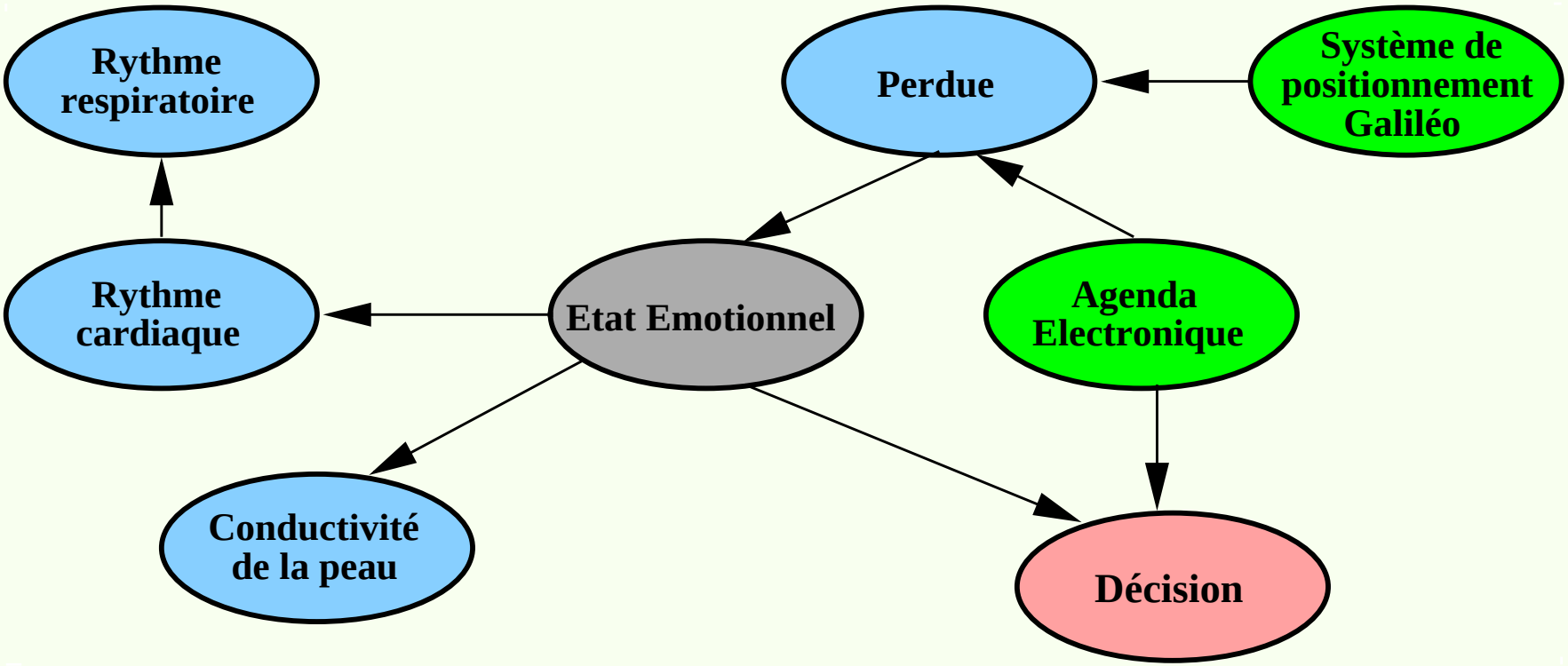

Etat Emotionnel


Perdue
Système de positionnement Galiléo
Etat Emotionnel
Agenda Electronique

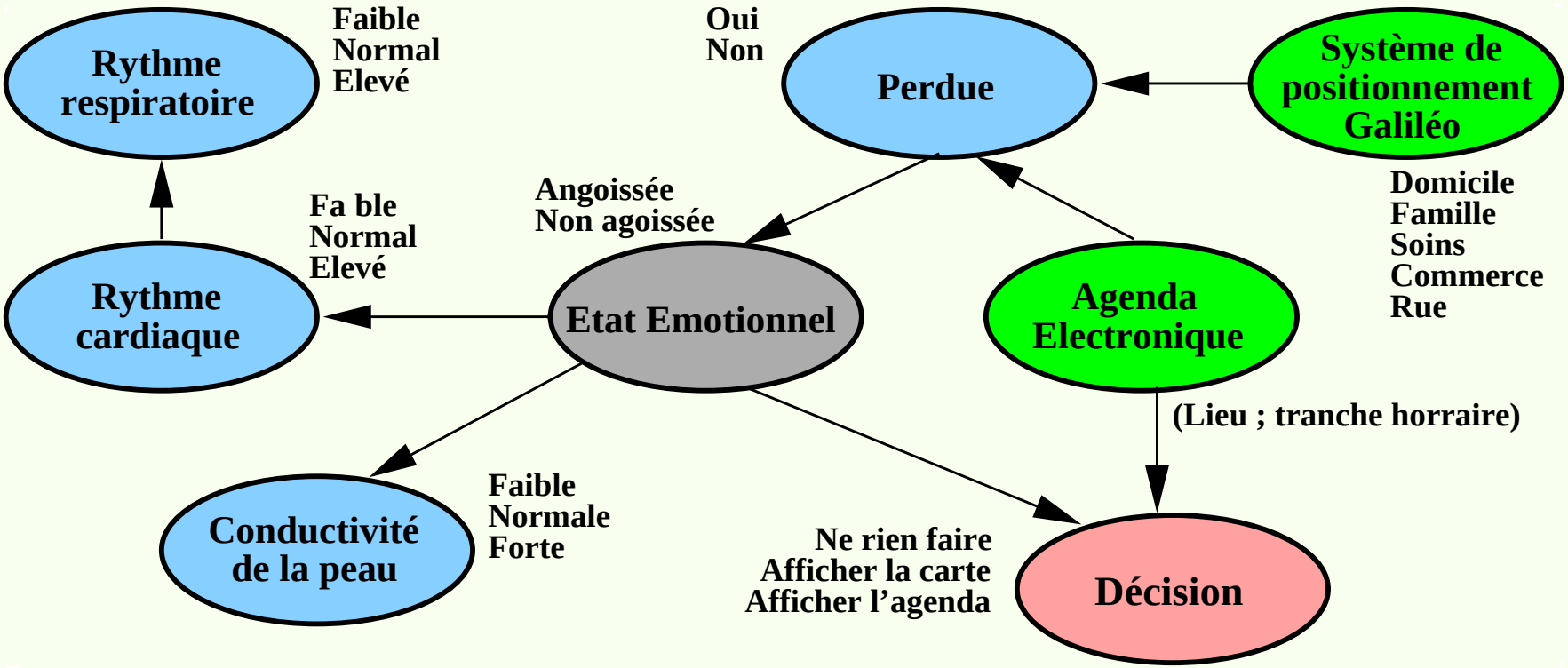

Perdue
Système de positionnement Galiléo
Etat Emotionnel
Agenda Electronique


Rythme respiratoire
Rythme cardiaque
Conductivité de la peau
Perdue
Système de positionnement Galiléo
Etat Emotionnel
Agenda Electronique

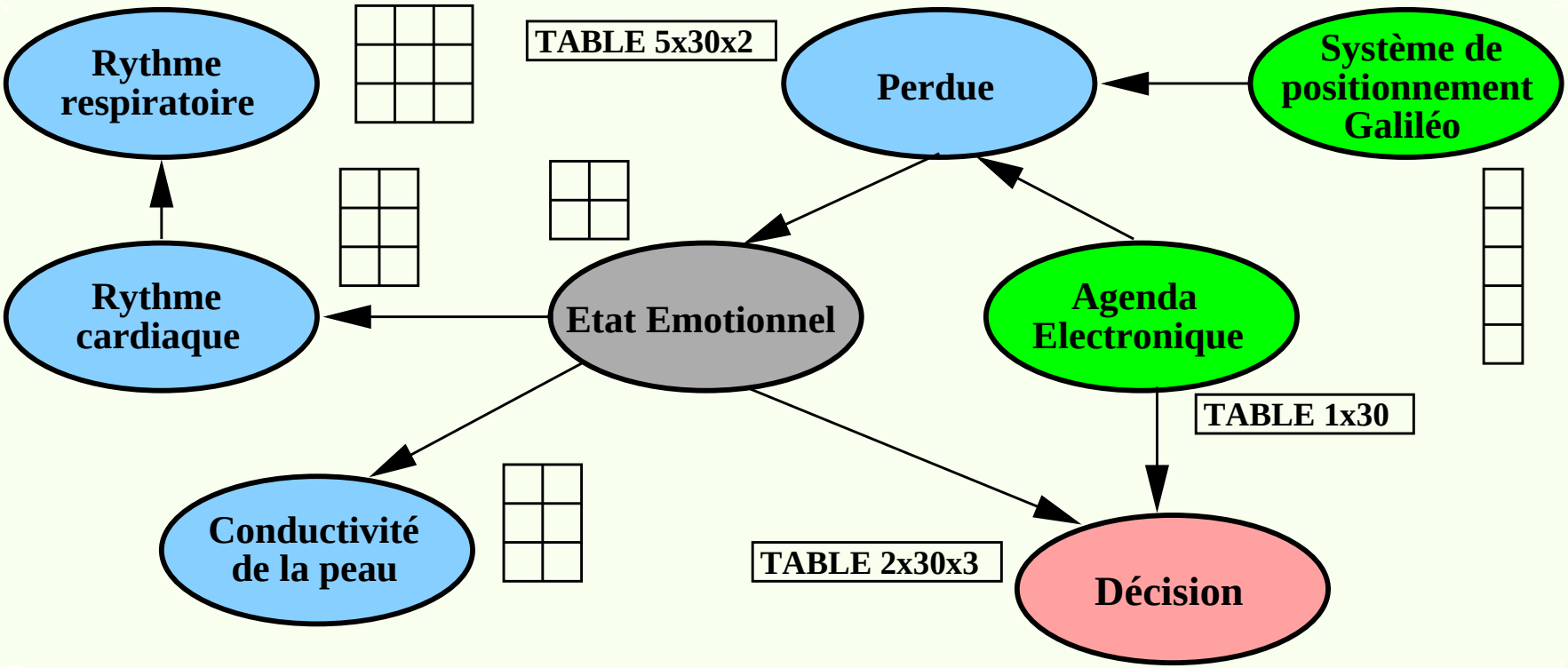

Rythme respiratoire
Perdue
Système de positionnement Galiléo
Rythme cardiaque
Etat Emotionnel
Agenda Electronique
Conductivité de la peau


Rythme respiratoire
Perdue
Système de positionnement Galiléo
Rythme cardiaque
Etat Emotionnel
Agenda Electronique
Conductivité de la peau
Décision

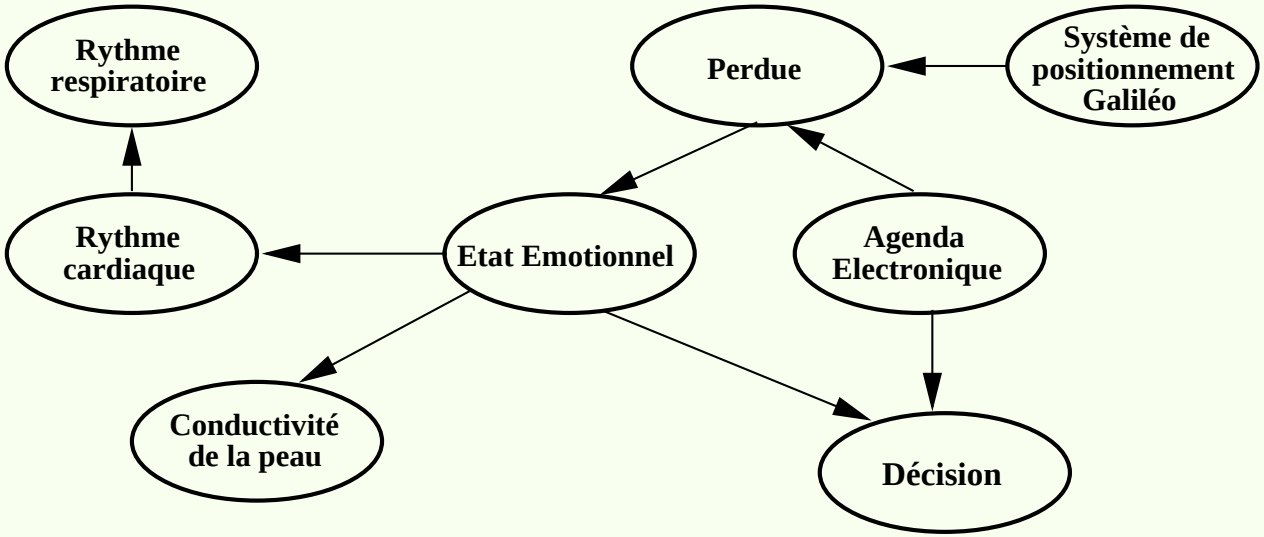
Définition d'un Réseau Bayésien



Définition d'un Réseau Bayésien

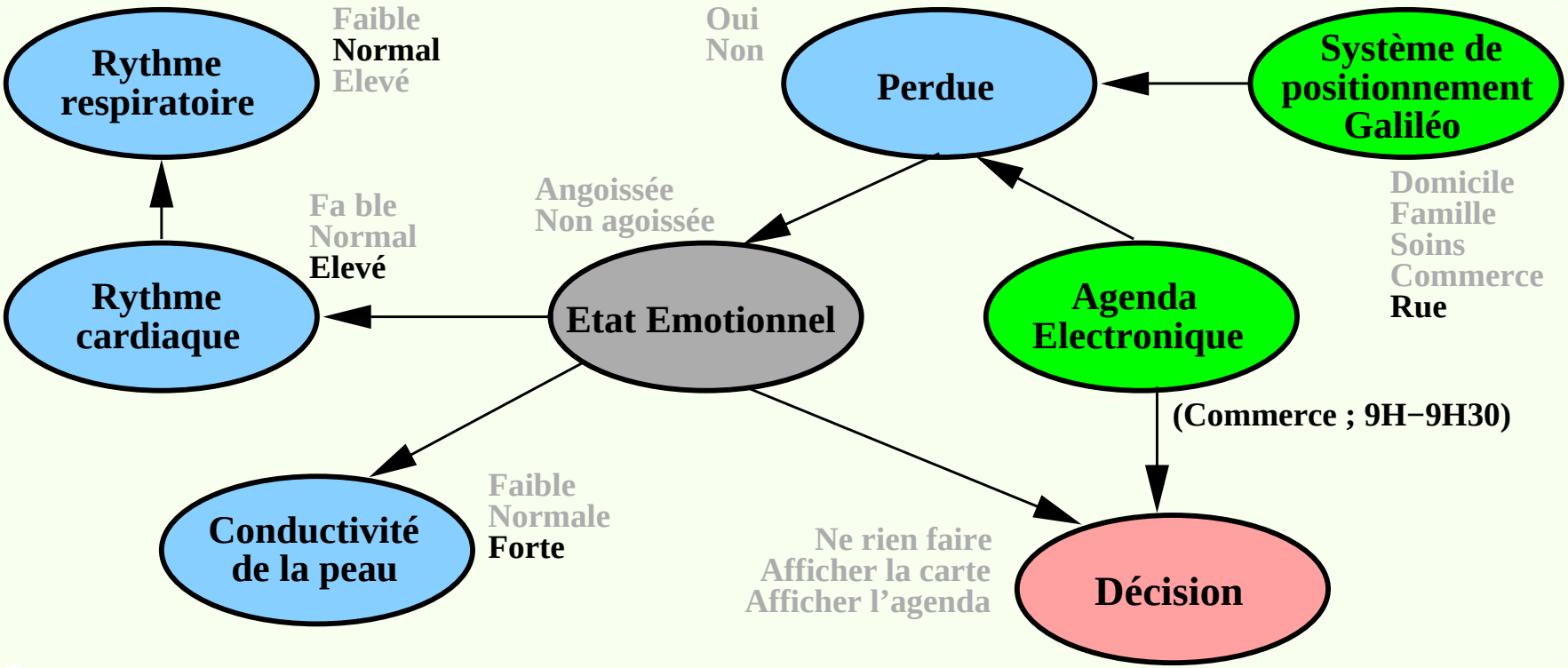


Apprentissage de structure : Quoi ?

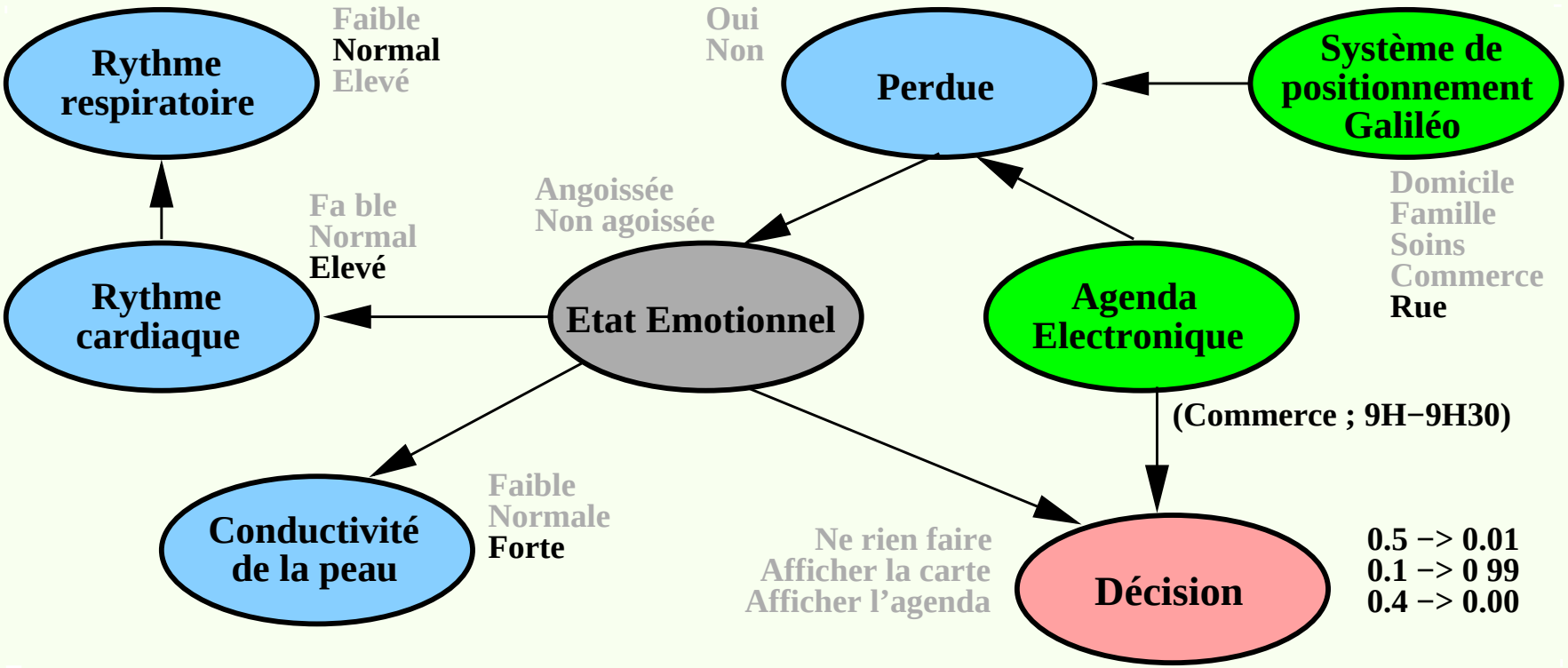


respi.	card.	peau	EE	Agenda	GPS	Perdue	Décision
normal	élevé	normale	?	RDV med.	maison	non	?
élevé	normal	forte	?	?	rue	?	?
normal	normal	faible	?	?	famille	?	?
...	...	...	...	...	...	...	

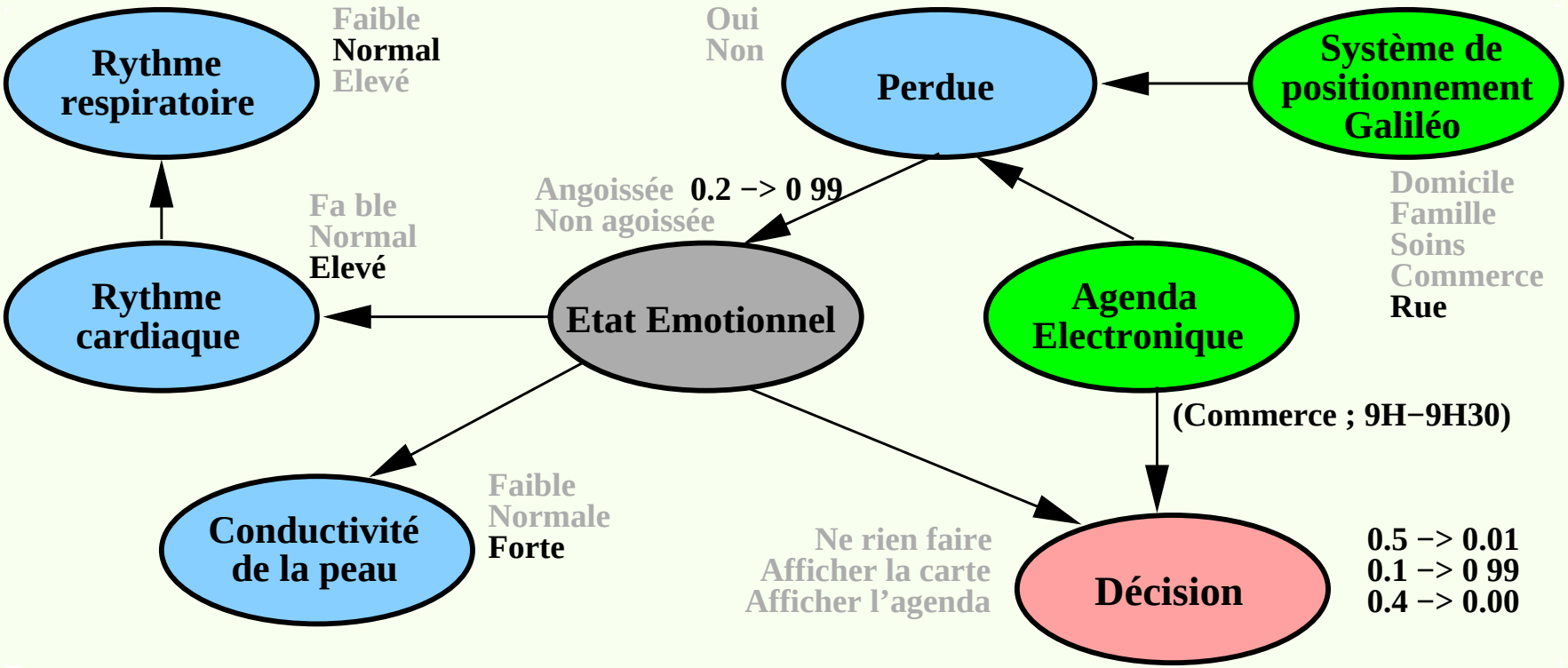
Apprentissage de structure : Pourquoi ?



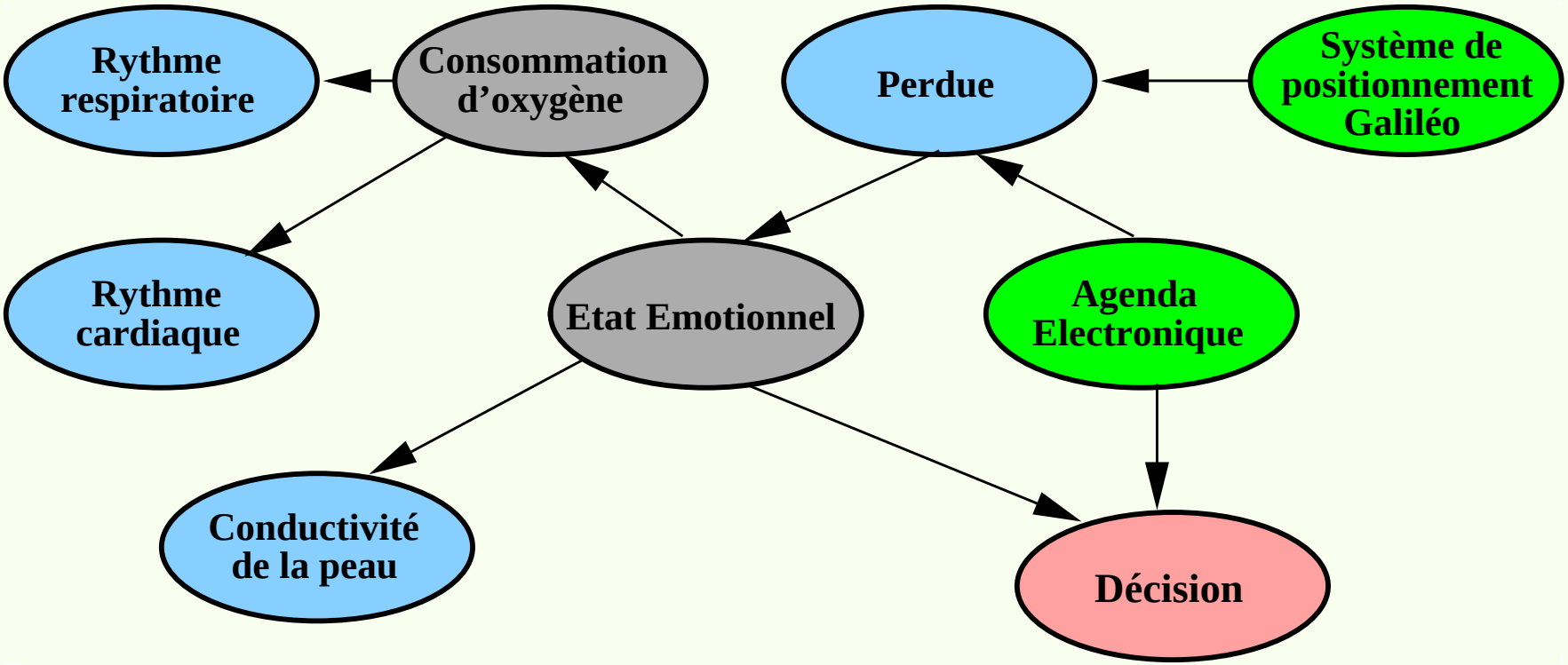
Apprentissage de structure : Pourquoi ?



Apprentissage de structure : Pourquoi ?



Apprentissage de structure : Pourquoi ?



Activités de recherche en quelques mots

Mots clés

- Apprentissage Statistique
- Modélisation probabiliste
- Modèles Graphiques Probabilistes
  - Réseaux Bayésiens (Dynamiques)
  - Modèles de Markov Cachées
  - Diagrammes d'Influence et Processus de décision semi-markoviens
  - Champs de Markov (*random fields*)
  - Filtres de Kalman
  - Chain graphs
  - Réseaux de préférences (*GAI-nets*)
- Simulation stochastique

Activités de recherche en quelques mots

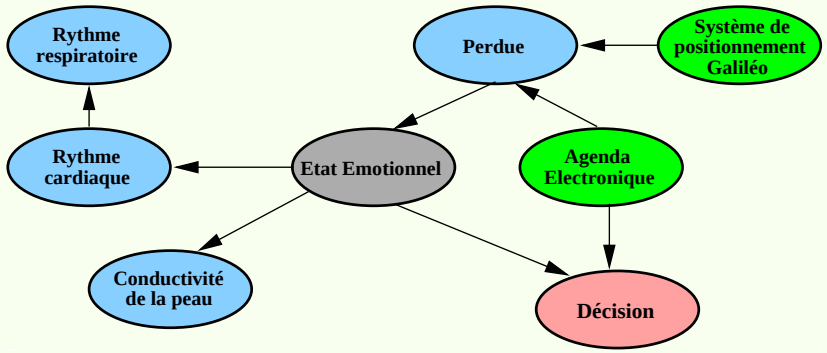
Problématiques

- Traitement des masses de données
  - Sélection de modèles, d'attributs
  - Classement, classification — *ex : Firebird*
  - Aide à la décision — *ex : MS-Office*
  - Fusion de données, de sources hétérogènes
- Ingénierie des connaissances
  - Modélisation des connaissances
  - Prise en compte d'experts et retours d'expériences
  - Découverte de connaissances
- Sciences de l'information
  - Extraction d'information
  - Modélisation *économique* de l'information
  - Compression — *ex : Djvu*

Intérêts et motivation

Intérêts des RB

- Outil de représentation *graphique* des connaissances
- Représentation de l'incertain et de l'imprécis
- Raisonnement à partir de donneés incomplètes



Motivation

- Comment déterminer automatiquement la structure ?

Intérêts et motivation

Des domaines d'application variés

- Aide au diagnostic
- Fiabilité
- Contrôle de production
- Maintenance
- Contrôle-commande
- Sécurité informatique
- Psychologie
- Sciences cognitives

Motivation

- Fournir des outils génériques pour **modéliser** et **simuler** des systèmes complexes.

Plan du cours : Partie II Apprentissage et généralisation

- Erreur théorique et erreur empirique
- Sur-apprentissage
- Généralisation
  - Ensemble de validation
  - Ensemble de test
- Rasoir d'Occam
- Dilemme Biais/Variance
- Régularisation
- Validation Croisée

Introduction

- **On a un modèle *f***
- **Erreur commise par un modèle = erreur théorique**

$$EP(w) = \iint (y - f(x,w))^2 P(x,y)\,dxdy$$

- **L'apprentissage se fait sur une base d'exemples (x_i,y_i) en minimisant l'erreur empirique (ici avec une fonction de coût quadratique)**

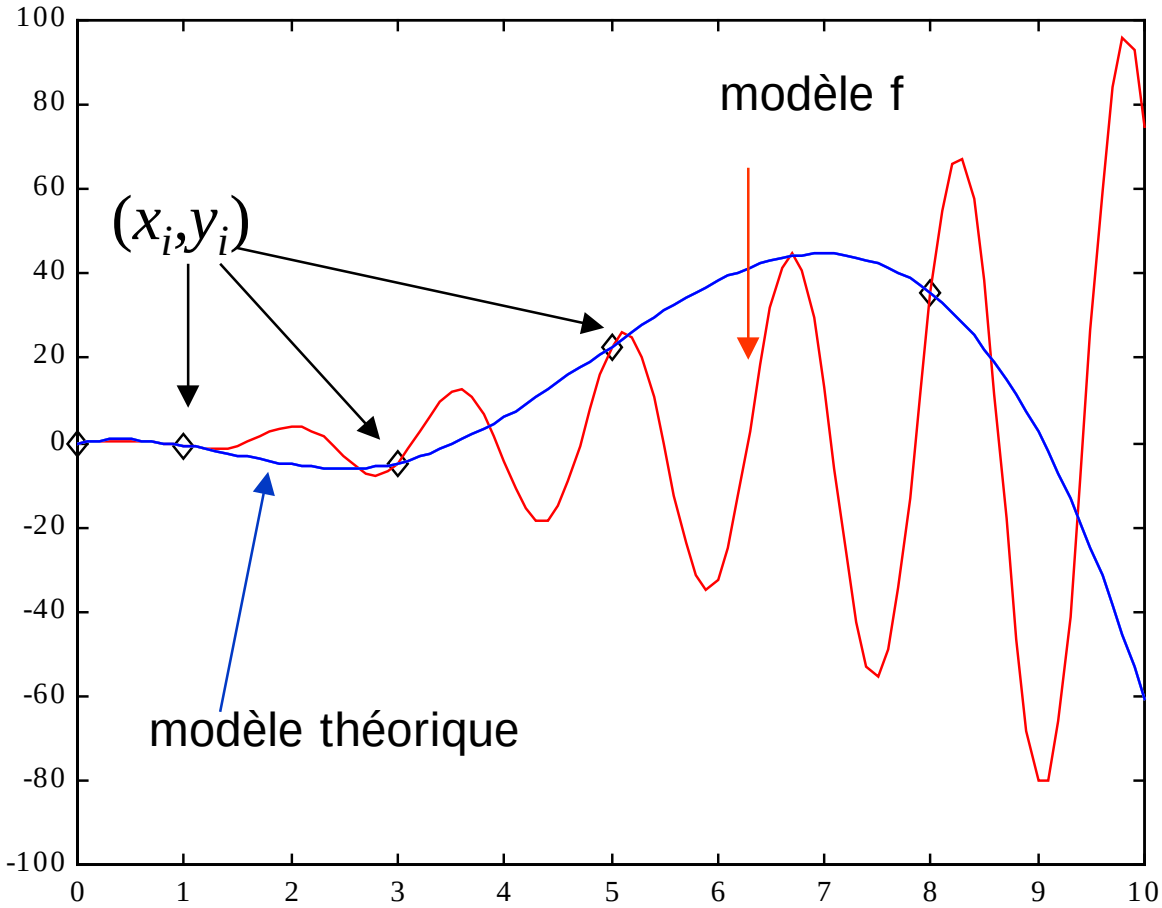
$$J(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i,w))^2$$

- **L'erreur en apprentissage est-elle un bon indicateur de la qualité du modèle *f* ?**

Introduction (fin)

- **L'erreur en apprentissage est-elle un bon indicateur de la qualité du modèle *f* ?**

- ***J(f)=0***



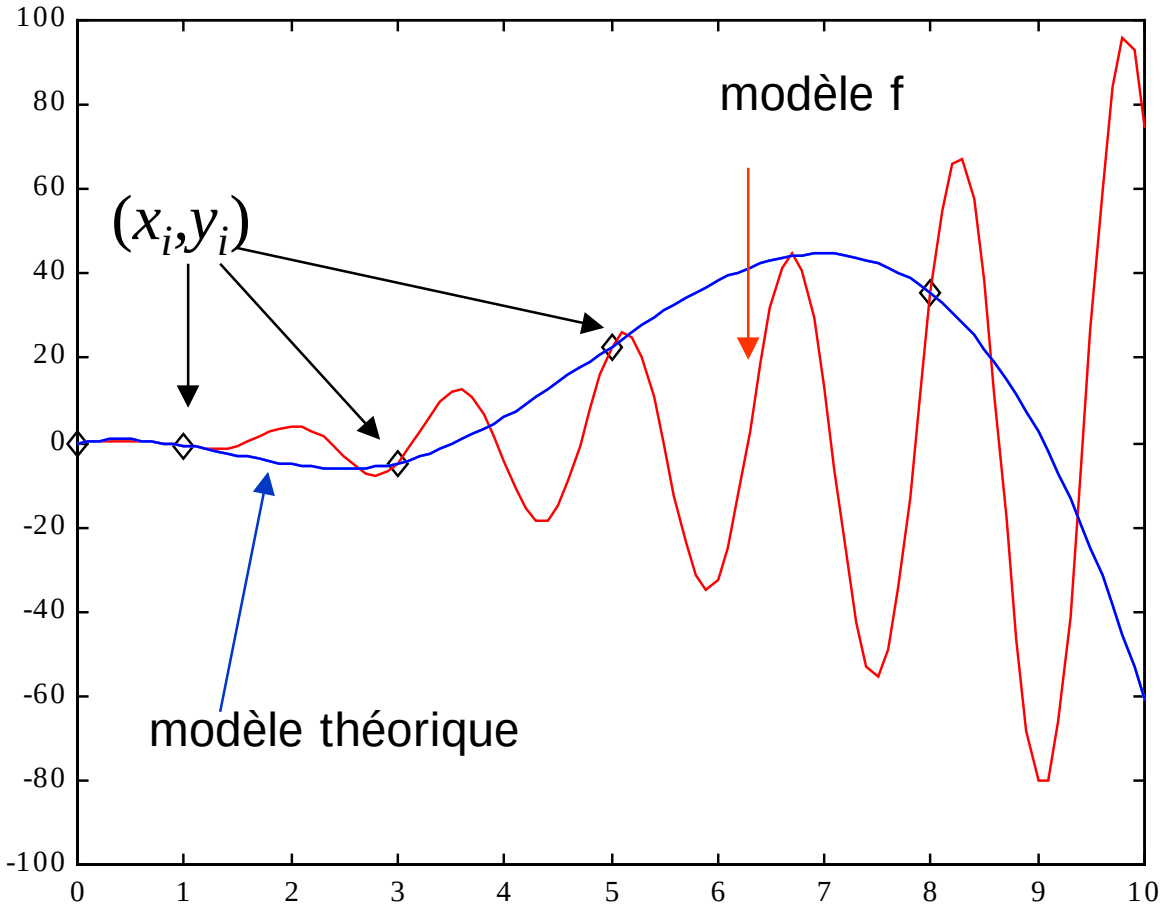
Réponse : Non !

Plan

- **Notion de sur-apprentissage**
- **Une méthode pour éviter le sur-apprentissage :**
  - **Ensemble de validation, early stopping**
- **Comment "bien apprendre" ?**
  - **Complexité d'un modèle : Le rasoir d'Occam**
  - **Dilemme biais/variance**
  - **Comment diminuer la variance ?**
    - » **Régularisation**
    - » **Bootstrap, "committees"**
- **Comment estimer autrement l'erreur en généralisation**

Le sur-apprentissage

- **Apprentissage "par cœur" : le modèle ne "connaît" que les points utilisés pour l'apprentissage et fait n'importe quoi ailleurs.**



Notion de généralisation

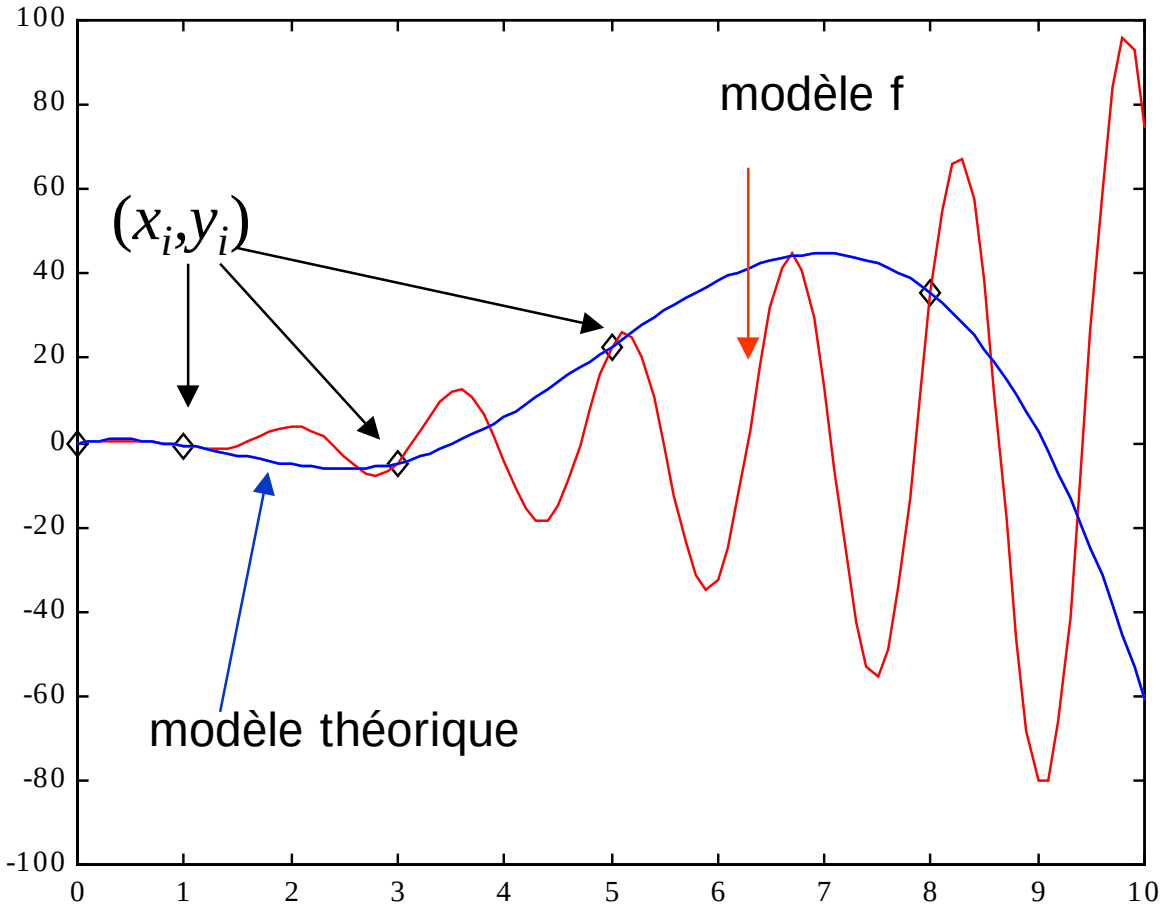
- **Bien apprendre, ce n'est pas apprendre par cœur, mais être capable de bien se comporter devant des points quelconques**

$\Rightarrow$ **C'est la généralisation**

- **Comment éviter le sur-apprentissage ?**
- **Comment "bien" apprendre ?**
- **Comment estimer l'erreur en généralisation autrement ?**

Sur-apprentissage : le retour

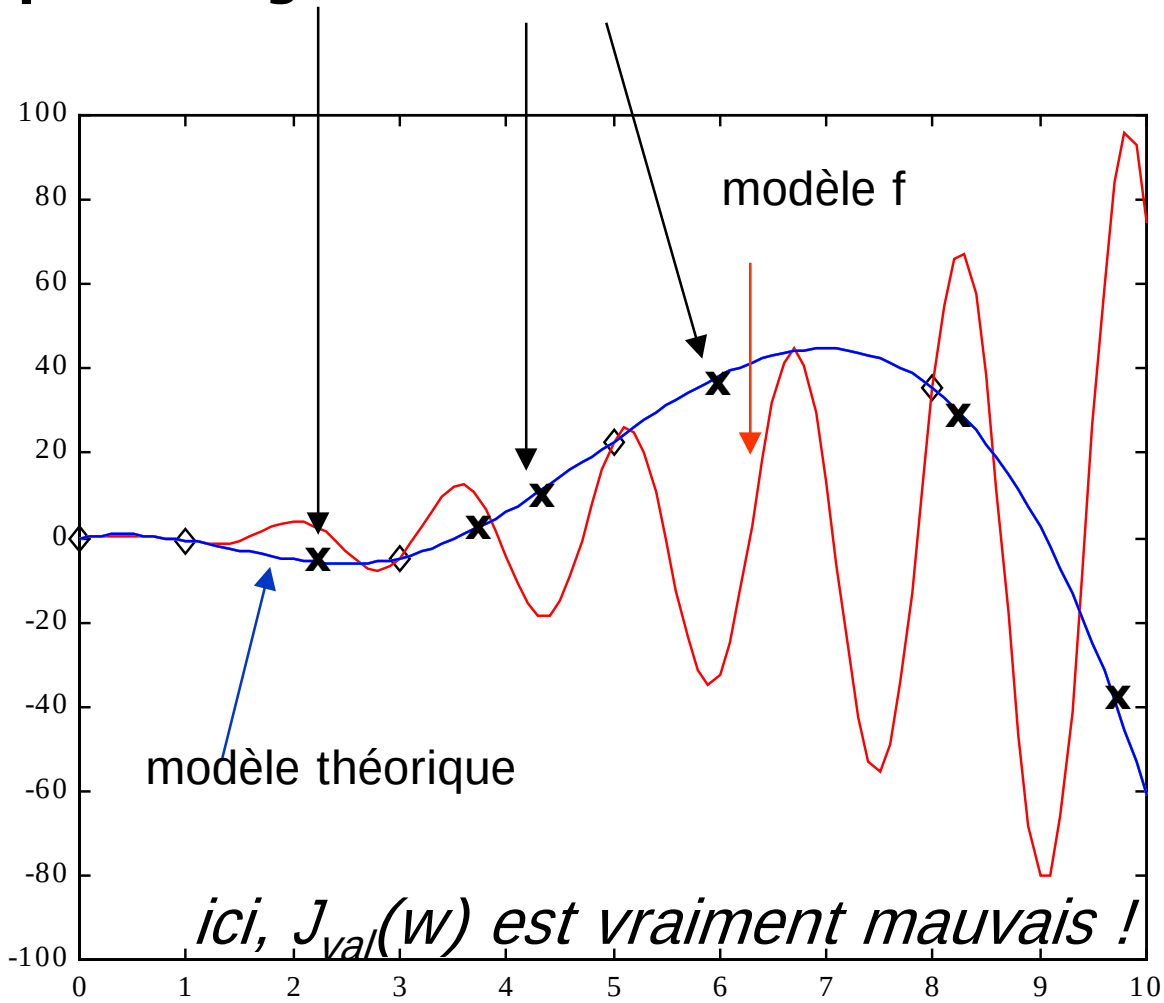
- **Comment déterminer que le choix de *f* n'est pas bon ?**
- ***J(f)=0 : l'erreur en apprentissage ne convient pas !***



Ensemble de Validation

- **On va prendre des points différents de ceux de l'ensemble d'apprentissage : *l'ensemble de validation***

Qualité du modèle : $J_{val}(w)$



Ensemble de Validation

- **Algorithme d'apprentissage itératif**

  Quand arrêter l'apprentissage pour éviter le sur-apprentissage ?

- **On sépare les exemples disponibles en 2 (ou 3) bases**

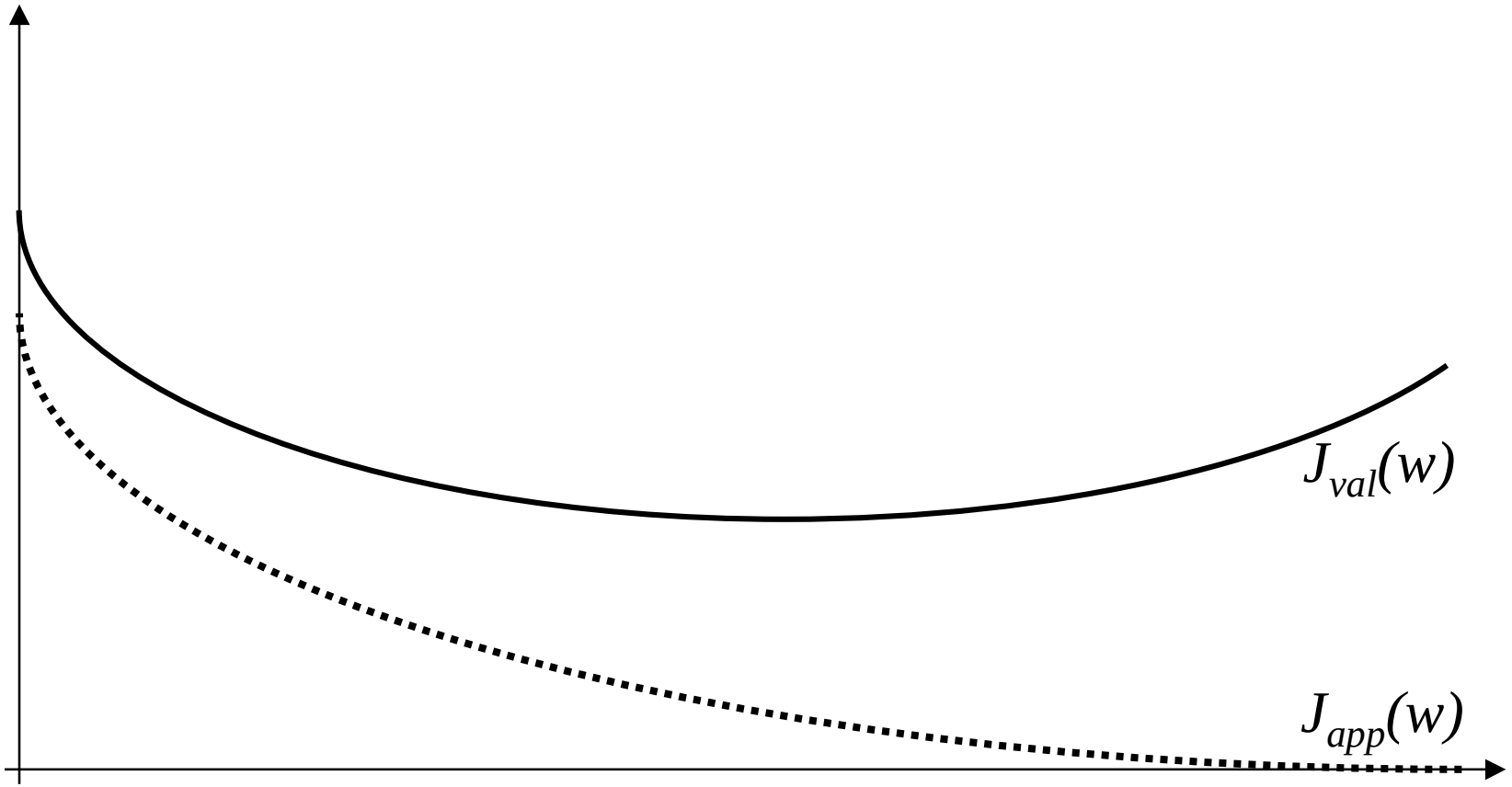
  – **apprentissage** $$J_{app}(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i, w))^2$$

  – **validation** $$J_{val}(w) = \frac{1}{2n'} \sum_{i=1}^{n'} (y'_i - f(x'_i, w))^2$$

  – **test : pour calculer une erreur indépendante des données qui ont servi à faire (et à arrêter) l'apprentissage**

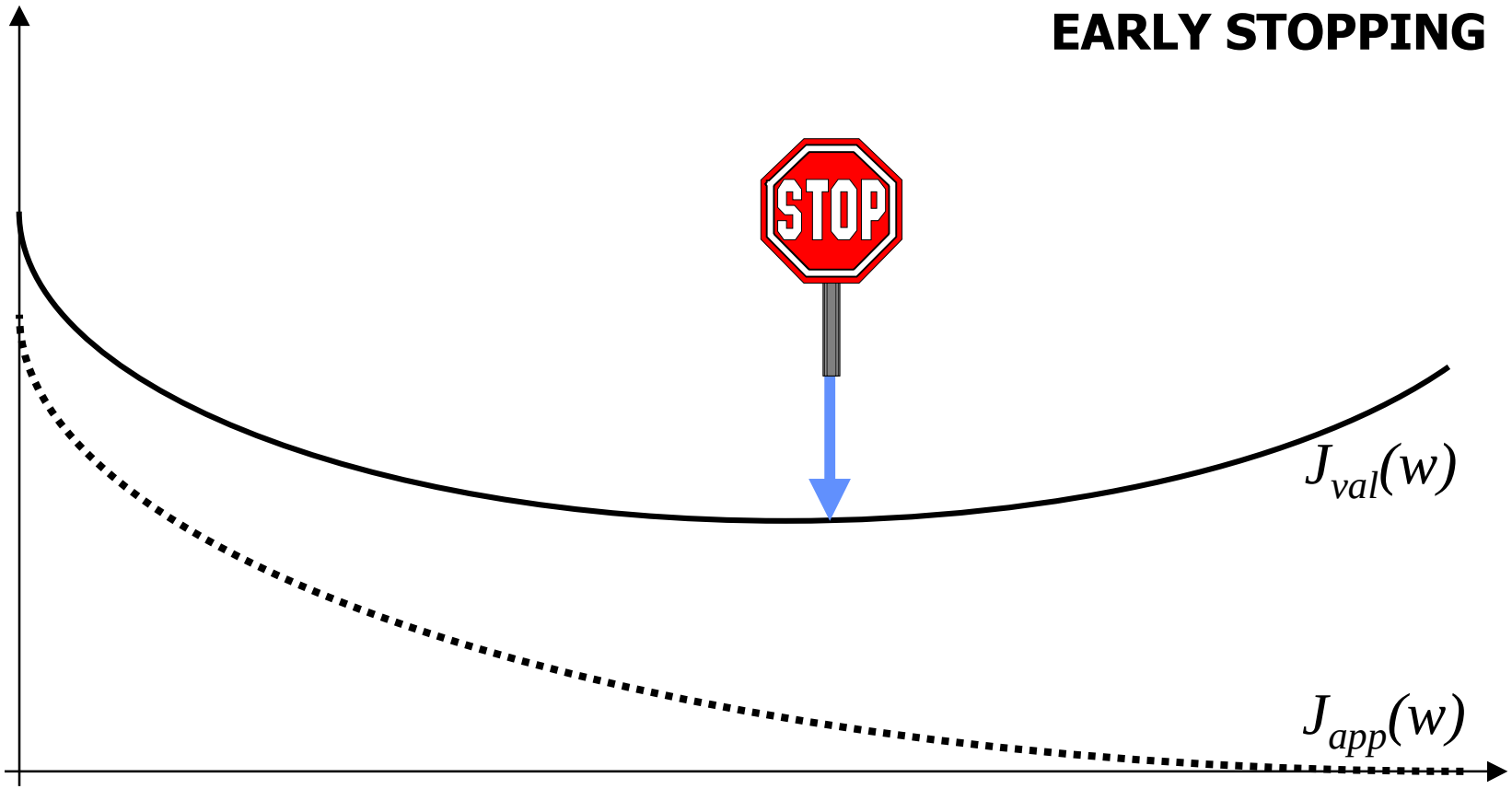
Ensemble de Validation

- $J_{val}(w)$ **est une meilleure estimation de EP**



Ensemble de Validation

- **On peut s'en servir pour arrêter l'apprentissage**



Transition

⇒ **On vient de voir un moyen d'arrêter l'apprentissage avant le "sur-apprentissage"**

- **Comment "bien" apprendre ?**
- **Comment calculer l'erreur en généralisation sans ensemble de validation ?**

Complexité d'un modèle

- **Principe général : le rasoir d'Occam *(14ème siècle)***
  Shave away all that is unnecessary

⇒ **Il faut toujours préférer un modèle simple à un modèle complexe**

- **Comment déterminer la complexité d'un modèle ?**
  - **Nb de paramètres (degré pour un polynôme, etc…)**
  - **Bonne approximation, mais pas suffisante pour des modèles compliqués (Réseaux de Neurones, etc…)**
  - **Travaux de Vapnik et Chervonenkis : la VC dimension**

Le dilemme biais/variance

- **Notations**
  - **(*x*,*y*) un point**
  - ***f* un modèle.**
  - **E(y|x) meilleur modèle possible**
  - ***C* une fonction de coût locale** $C(f,x,y)=\|f(x)-E(y|x)\|^2$
    - » **coût quadratique**
  - ***EP(f)* : erreur de modélisation**

$$EP(f)=E_{xy}[C(f,x,y)]=E_{xy}\left\{\|f(x)-E(y|x)\|^2\right\}$$

  - ***D* : Échantillon de taille N**
  - **Δ : ensemble de tous les échantillons**
  - **le modèle *f* est appris sur *D* : f_D**
  - **erreur de modélisation "moyenne"** $EP=E_{\Delta}[EP(f_D)]$

Le dilemme biais/variance (suite)

- **Ré-écriture de C :**

$$C(f,x,y) = \left\| f(x) - E(y|x) \right\|^2 = \left\| f(x) - E_\Delta(f(x)) + E_\Delta(f(x)) - E(y|x) \right\|^2$$

$$C(f,x,y) = \left\| f(x) - E_\Delta(f(x)) \right\|^2 + \left\| E_\Delta(f(x)) - E(y|x) \right\|^2 + 2\left\{ E_\Delta(f(x)) - E(y|x) \right\}\left\{ f(x) - E_\Delta(f(x)) \right\}$$

$$A = \left\| f(x) - E_\Delta(f(x)) \right\|^2$$

$$B = \left\| E_\Delta(f(x)) - E(y|x) \right\|^2$$

$$C = 2\left\{ E_\Delta(f(x)) - E(y|x) \right\}\left\{ f(x) - E_\Delta(f(x)) \right\}$$

indépendant de Δ

Le dilemme biais/variance (suite)

- **"Moyennage" sur Δ :** $E_\Delta[C(f,x,y)] = E_\Delta[A] + E_\Delta[B] + E_\Delta[C]$

$$E_\Delta[C] = E_\Delta[\{E_\Delta(f(x)) - E(y|x)\}\{f(x) - E_\Delta(f(x))\}]$$

indépendant de Δ

$$E_\Delta[C] = \{E_\Delta(f(x)) - E(y|x)\} \times E_\Delta[f(x) - E_\Delta(f(x))]$$

$$E_\Delta[C] = \{E_\Delta(f(x)) - E(y|x)\} \times \{E_\Delta(f(x)) - E_\Delta(f(x))\}$$ **= 0 !!!!!**

Le dilemme biais/variance (suite)

- **Retour à EP**

$$EP = E_{\Delta}[EP(f_D)] = E_{\Delta}[E_{xy}(C(f_D, x, y))]$$

$$EP = E_{xy}[E_{\Delta}(C(f_D, x, y))]$$

$$= E_{xy}[E_{\Delta}(A)] + E_{xy}[E_{\Delta}(B)]$$

$$E_{\Delta}(A) = E_{\Delta}\left(\| f(x) - E_{\Delta}(f(x)) \|^2\right)$$

VARIANCE

$$E_{\Delta}(B) = E_{\Delta}\left(\| E_{\Delta}(f(x)) - E(y|x) \|^2\right)$$

BIAIS²

Le dilemme biais/variance (suite)

- **EP =Biais² + Variance**

  Biais = écart entre le modèle "moyen" et le modèle idéal

  Variance = variance de tous les modèles *f* possibles

⇒ On cherche à minimiser EP, donc le biais ET la variance …

- **En pratique, biais et variance sont antagonistes …**
  - **diminution de l'un = augmentation de l'autre**
    - » **ex :**
      - **on choisit *f* dans une grande famille de fonctions (polynômes de degré <100)**
      - **on est sur de trouver de bonnes approximations : biais faible**
      - **la famille de fonctions est tellement grande que la variance est énorme**

⇒ il faut trouver un compromis …

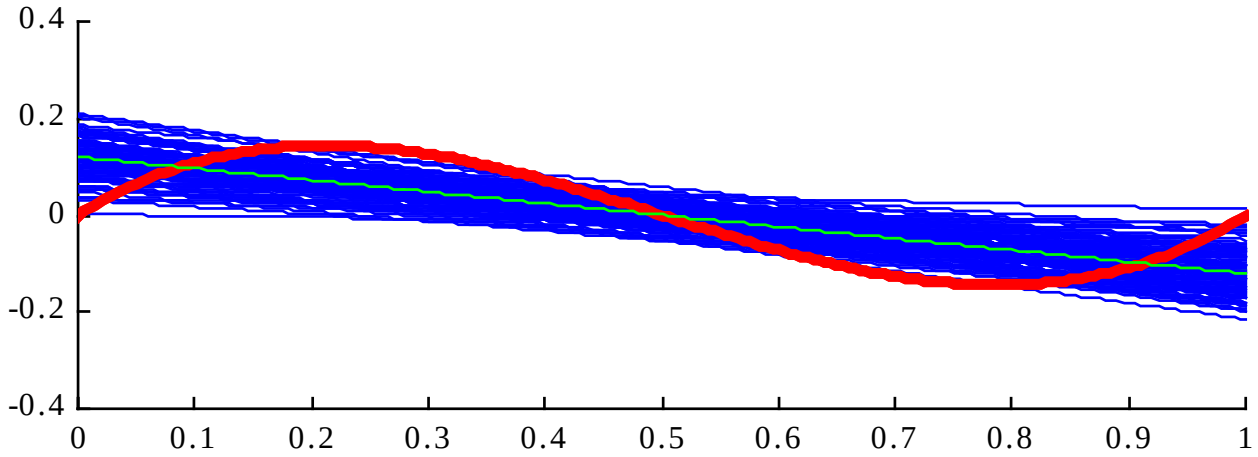
C'est le dilemme biais/variance

Exemple

50 interpolations polynomiales à partir de 15 points de la fonction bruitée

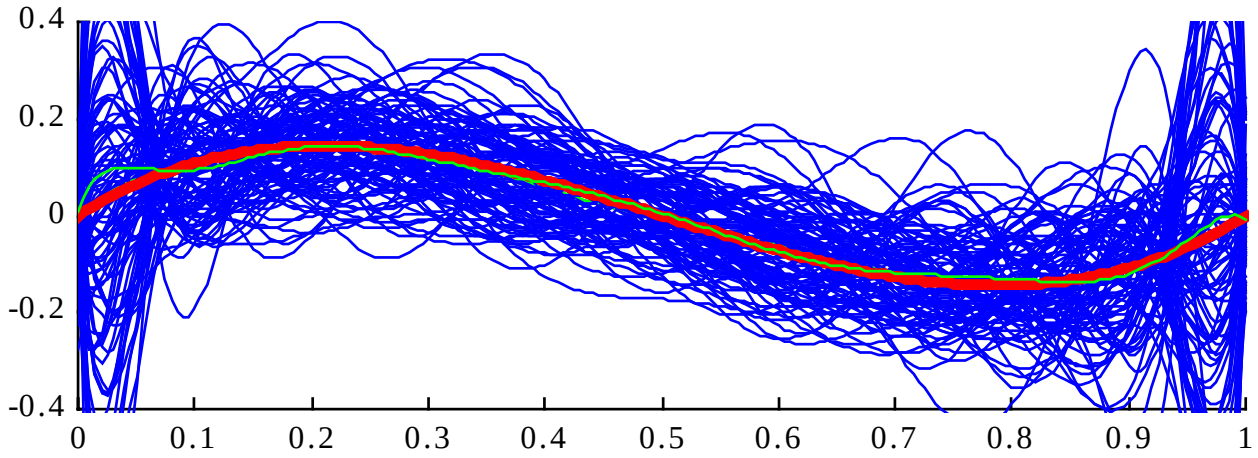
polynômes de degré 1

biais important
variance faible



polynômes de degré 12

biais faible
variance importante



Régularisation

- **Supposons que le biais soit faible …**
- **Comment arriver à "limiter" la variance ?**

- **On peut "brider" les paramètres des fonctions pour les empêcher de s'écarter de la solution …**
  - **ex : les polynômes de degré < 100**
    MAIS avec le + possible de coefficients nuls …
    (Occam : on favorise les polynômes simples)

  - **comment faire ?**
    on change la fonction d'erreur empirique :

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i, \boldsymbol{w})\right)^2 + \lambda\|\boldsymbol{w}\|^2$$

erreur "classique"

pénalisation des coefficients

équilibrage

Committees / Bootstrap

- **Autre façon de diminuer la variance**

- **On prend L modèles différents :**

  – **même base d'apprentissage**
    nb de paramètres (ou initialisation) différent

  → **"committees" de modèles**

  – **même nb de paramètres**
    base d'apprentissage différente

  → **bootstrap, validation croisée, bagging**

Validation croisée

- La base d'exemple est séparée en L segments B_1 ... B_L
- On peut créer L ensembles d'apprentissage (+validation si nécessaire) + ensemble de test associé et utiliser chaque ensemble pour apprendre un modèle :
  - $App+Val_1 = \{B_1 \dots B_{L-1}\}$ $Test_1 = B_L$ $\Rightarrow$ f_1
  - $App+Val_2 = \{B_1 \dots B_{L-2}\, B_L\}$ $Test_2 = B_{L-1}$ $\Rightarrow$ f_2
  - ...
  - $App+Val_L = \{B_2 \dots B_L\}$ $Test_L = B_1$ $\Rightarrow$ f_L
- on utilise le modèle moyen

$$f(x) = \frac{1}{L}\sum_{j=1}^{L} f_j(x)$$

Validation croisée

- **Avantage :**
  - **la variance est diminuée**

- **Inconvénients :**
  - **il faut faire L apprentissages au lieu d'un (bon compromis : L=10)**

- **Cas particulier :**
  - **séparation des exemples en N segments de 1 exemple**

  $\rightarrow$ **méthodes leave-one-out ou Jack-Knife**

Bootstrap

- **On crée L ensembles d'apprentissage(+validation) en tirant N exemples au hasard AVEC REMISE**

- **En pratique L>30**

Estimer EP autrement

- **Idée : estimer l'erreur de prédiction**
  - **sans données de validation**
  - **avec**
    - » **l'erreur en apprentissage**
    - » **une idée de la complexité du modèle**

- **Il existe un certain nombre de méthodes dérivées des statistiques ou de la théorie de l'information**

$$EP = J_{app} + F\left(ComplexitéModèle\right)$$

Plan du cours : Partie III Rappel de Probabilités

- Indépendance conditionnelle
- Interprétation fréquentiste
- Dilemme du Prisonnier
- Probabilités des causes et Théorème de Bayes

Notions de Probabilité Conditionnelle

- **Pourquoi ?**

- **Les Réseaux de Neurones et les Réseaux Bayesiens utilisent tous les deux la notion de probabilité conditionnelle :**

  - **X = [hauteur, largeur, poids] d'un objet**
  - **C = {voiture, camion, vélo}**
  - **On mesure X=[15 5 500],**
    quelle est la probabilité que C soit une voiture sachant X ?

  - **A : "La bourse de New-York est en hausse"**
  - **B : "Il pleut à Rouen"**
  - **Quelle est la probabilité que la bourse de New-York soit en hausse quand il pleut à Rouen ?**

Définition

- **Soient 2 événements A et M, on suppose que <u>M s'est produit</u>**

- **{M s'est produit} = information a priori** **(** $P(M) \neq 0$ **)**

  S'il existe un lien entre A et M, cette information va modifier la probabilité de A

⇒ **Probabilité de A conditionnellement à M**
Probabilité de A sachant M

$$P(A|M) = \frac{P(A \& M)}{P(M)}$$

- $P(A|M)$ **est une probabilité**

Interprétation "fréquentiste"

- **On répète N fois une expérience E sur laquelle on a défini 2 événements A et M**
  - *n(A)* **= nb de réalisations de A**
  - *n(M)* **= nb de réalisations de M**
  - *n(A&M)* **= nb de réalisations simultanées de A et M**

$$\hat{P}(A) = \frac{n(A)}{N} \qquad \hat{P}(M) = \frac{n(M)}{N} \qquad \hat{P}(A\&M) = \frac{n(A\&M)}{N}$$

$$\Rightarrow \quad \hat{P}(A|M) = \frac{n(A\&M)}{n(M)}$$

Rappel : $P = \lim_{N \to +\infty} \hat{P}$

théorique / empirique

Indépendance de 2 événements

- **A et B sont 2 événements indépendants ssi**

$$P(A\&B) = P(A) \times P(B)$$

- **en utilisant la définition de la probabilité conditionnelle :**

$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

⇒ **la connaissance de B n'apporte rien sur celle de A et réciproquement**

Exemple

- **La législation de la marijuana aux USA**

	Pour	Contre
Est des USA	7.8 %	22.2 %
Autres régions	18.2 %	51.8 %

$\hat{P}(Est) = 30\ \%$

$\hat{P}(Autres) = 70\ \%$

$\hat{P}(Pour) = 26\ \%$

° $\hat{P}(Pour \mid Est) = \hat{P}(Pour \& Est) \ / \ \hat{P}(Est) = 7.8 \ / \ 30 = 26\ \%$

$\Rightarrow$ Pour et Est sont 2 événements indépendants

° En pratique : que dire si $\hat{P}(Pour \mid Est) =$ 27 % 35 % 50 %

$\Rightarrow$ Il faudra fixer une frontière de décision

Probabilités Totales

- **Soient N événements** $M_1, \ldots M_N$
  - **complets**
  - **mutuellement exclusifs**

- **Théorème des Probabilités Totales**

$$P(A) = \sum_{i=1}^{N} P(A|M_i)P(M_i)$$

Exemple

- **Soit un lot de 4500 pièces**
  - **3700 bonnes**
  - **800 fausses**

- **réparties dans 4 boites**
  - **B1 : 2000 pièces (1600 bonnes, 400 fausses)**
  - **B2 : 500 pièces (300 bonnes, 200 fausses)**
  - **B3 : 1000 pièces (900 bonnes, 100 fausses)**
  - **B4 : 1000 pièces (900 bonnes, 100 fausses)**

- **F = {tirer une pièce fausse}**

P(F) ?

Exemple

- **Il faut déjà choisir la boite dans laquelle on va prendre une pièce …**
  - **4 boites, choix équiprobable :** *P(Bi) = 1/4*
- **Pour chaque boite, on peut déterminer la probabilité d'avoir une pièce fausse**
  - *P(F|B1) = 400/2000 = 0.2*
  - *P(F|B2) = 200/500 = 0.4*
  - *P(F|B3) = 100/1000 = 0.1*
  - *P(F|B4) = 100/1000 = 0.1*

$$P(F) = \sum_{i=1}^{4} P(F|B_i)P(B_i) = \frac{1}{4}(0.2 + 0.4 + 0.1 + 0.1) = 0.2$$

Exercice : le jeu des $$$

les dangers du conditionnel !

- **Jeu TV :**
  - **3 portes, des $$$ derrière une des 3 portes A, B et C**
  - **Le présentateur SAIT où sont les $$$**
  - **Vous êtes le candidat …**
  - **1ère question du présentateur :**
    - » **quelle porte choisissez-vous (sans l'ouvrir) ?**
  - **Le présentateur ouvre une des 2 portes restantes (et qui ne cache pas les $$$, il ne veut pas perdre)**
  - **2ème question du présentateur :**
    - » **choisissez-vous toujours la même porte ?**

Exercice : le jeu des $$$

- **1ère question du présentateur :**
  - » **quelle porte choisissez-vous (sans l'ouvrir) ?**
  - » **Aucun a priori :**
    P(A=$$$) = P(B=$$$) = P(C=$$$) = 1/3
  - » **on peut choisir au hasard A**
- **Le présentateur ouvre une des 2 portes restantes (et qui ne cache pas les $$$) de manière équiprobable**

Exercice : le jeu des $$$

- **résolution**

	A=$$$	B=$$$	C=$$$	
présentateur ouvre B	*1/6*	*0*	*1/3*	*1/2*
ouvre C	*1/6*	*1/3*	*0*	*1/2*
	1/3	*1/3*	*1/3*	

Si le présentateur ouvre B, la porte la plus probable est C
Si le présentateur ouvre C, la porte la plus probable est B
Þ ***dans les 2 cas, notre premier choix (A) n'est plus le plus probable !***

Probabilités des Causes

- **Question :**
  Un événement A peut résulter de plusieurs causes Mi (s'excluant mutuellement),
  quelle est la probabilité de A connaissant
  - **les probabilités élémentaires** $P(Mi)$ ***probabilités à priori***
  - **les probabilités conditionnelles de A à chaque cause Mi.**

- **Réponse = Théorème des probabilités totales**

- **Mais comment répondre à la question inverse :**
  Un événement A s'est produit,
  quelle est la probabilité que ce soit la cause Mi qui l'ait produit ?
- **comment calculer** $P(Mi|A)$ **?** ***probabilité à posteriori***

- **Réponse = Théorème de Bayes ...**

Théorème de Bayes

- **Soient plusieurs événements Mi (s'excluant mutuellement),**
  - *P(Mi)* **probabilité à priori de Mi**
  - *P(A|Mi)* **probabilité de A conditionnellement à Mi**
  - *P(Mi|A)* **probabilité à posteriori de Mi (condit. à A)**

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

  - *P(A)* **s'obtient par le Théorème des probabilités totales**

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

Théorème de Bayes

Exemple

Statistiques - Wonnacott - p.104

- **Comment acheter une voiture d'occasion ?**
  - **je suis intéressé par le modèle X**
  - **les revues spécialisées indiquent que 30 % de ces voitures ont une transmission défectueuse**
  - **je trouve ce modèle chez un garagiste ...**
- **idée = obtenir des informations complémentaires en demandant à un ami mécanicien d'essayer la voiture et d'effectuer un diagnostic.**
  - **le diagnostic du mécanicien est généralement bon :**
    il reconnaît 90 % des voitures X défectueuses
    il reconnaît 80 % des voitures X non défectueuses

Exemple

- **Formulation mathématique**
  - **état réel de la voiture :** **DEF / OK**
  - **état diagnostiqué par le mécanicien :** **def / ok**

  - ***P(DEF) = 30 %*** **probabilité à priori**

  - ***P(def|DEF) = 90 %*** ***P(ok|DEF) = 10 %***
  - ***P(ok|OK) = 80 %*** ***P(def|OK) = 20 %***

  - **Quelles sont les probabilités à posteriori après le test du mécanicien ?**

    P(DEF|def) = ?
    P(DEF|ok) = ?

  - **on espère P(DEF|def) > P(DEF) > P(DEF|ok)**

Exemple

- **Règle de Bayes :**

$$P(DEF|def) = \frac{P(def|DEF)P(DEF)}{P(def|DEF)P(DEF) + P(def|OK)P(OK)}$$

$$\boxed{P(DEF|def) = \frac{90 \times 30}{90 \times 30 + 20 \times 70} = \frac{27}{41} = 66\,\%}$$

$$P(DEF|ok) = \frac{P(ok|DEF)P(DEF)}{P(ok|DEF)P(DEF) + P(ok|OK)P(OK)}$$

$$\boxed{P(DEF|ok) = \frac{10 \times 30}{10 \times 30 + 80 \times 70} = \frac{3}{59} = 5\,\%}$$

Exemple

- **Plus compliqué, mais plus réaliste …**

$$85\% < P(def|DEF) < 95\ \%$$

- **Quelles sont les probabilités à posteriori après le test du mécanicien ?**

$$P(DEF|def) = ?$$
$$P(DEF|ok) = ?$$

Et en continu ?

- **X est une variable aléatoire**
- **fonction de densité $f_X(x)$**

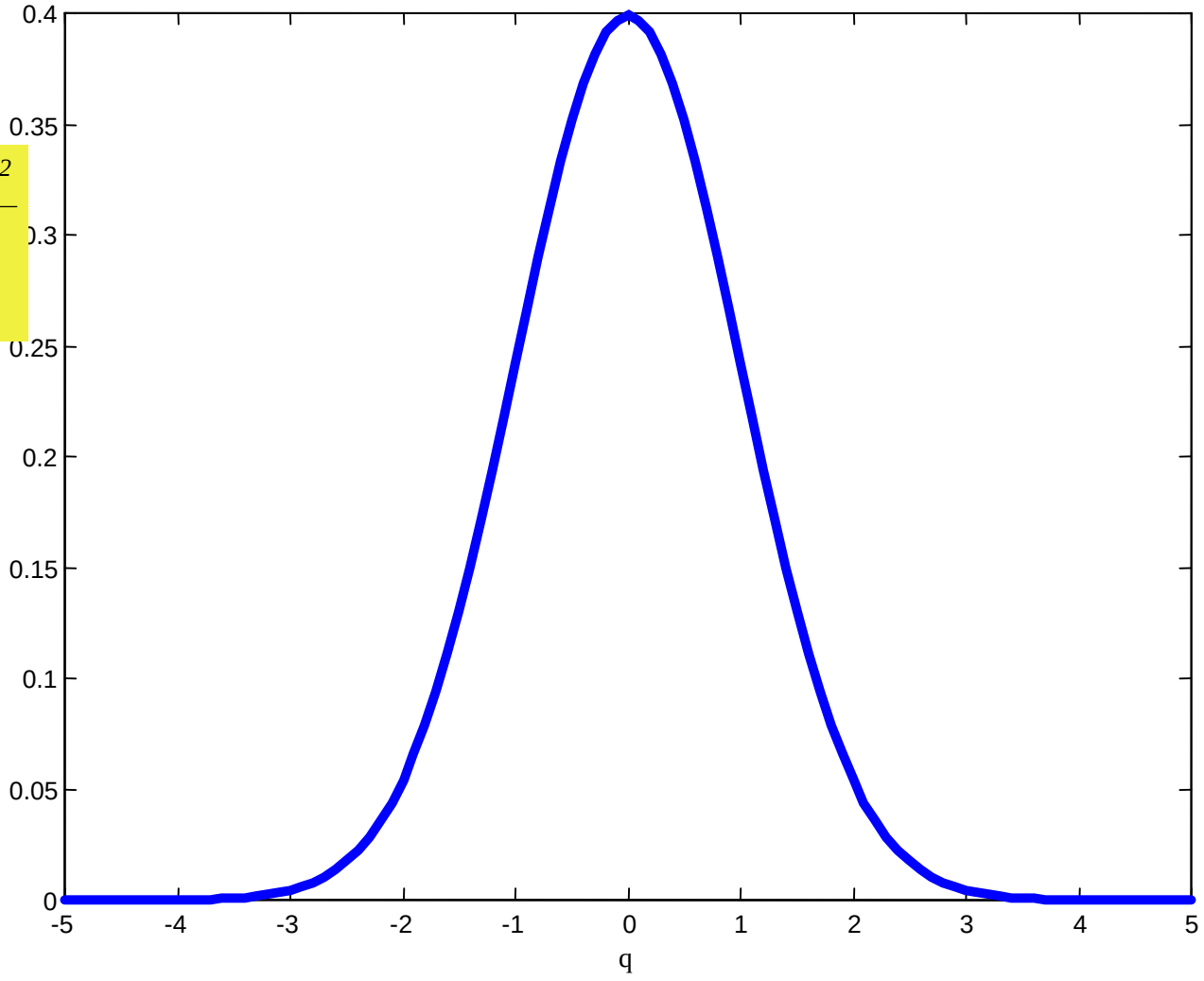
$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f_X(t)\,dt$$

- **Espérance :** $E(X) = \int_{-\infty}^{+\infty} x\ f_X(x)\,dx$
- **Variance :** $VAR(X) = E\left[(X - E(X))^2\right] = \int_{-\infty}^{+\infty} (x - E(X))^2\ f_X(x)\,dx$
- **Ecart-type :** $s_X = \sqrt{VAR(X)}$ $\qquad VAR(X) = s_X^2$

Exemple

- **La loi normale :**

$$f_X(x) = \frac{1}{\sqrt{2p\,s^{\,2}}}\,exp^{-\frac{(x-m)^2}{2s^{\,2}}}$$



Et en multi-dimensionnel ?

- **X et Y deux variables aléatoires**
- **fonction de densité $f_{X,Y}(x,y)$**

- **densités marginales :**

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy \qquad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx$$

- **densités conditionnelles :**

$$f_X(x|Y=y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \qquad f_Y(y|X=x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

rappel du cas continu

$$P(A|M) = \frac{P(A\&M)}{P(M)}$$

Exemple "mixte"

- X = taux de globule rouge — variable continue
- Y = état du patient — M = malade ¬M

- De manière générale :
  - $f_X(X|M) \sim N(30, s^2=5)$ — P(M) = 0.2
  - $f_X(X|\neg M) \sim N(27, s^2=5)$ — P(¬M) = 0.8

- Un patient se présente avec X=24 ...
  - P(M|X=24) ?

Exemple "mixte"

- **Règle de Bayes**

$$P(M|X=24) = \frac{f_X(X=24|M)P(M)}{f_X(X=24|M)P(M) + f_X(X=24|\neg M)P(\neg M)}$$

$f_X(X=24|M)$: $N(30, s^2=5)$ → ***0.0049***

$P(M)$: ***0.2***

$f_X(X=24|\neg M)$: $N(27, \sigma^2=5)$ → ***0.0725***

$P(\neg M)$: ***0.8***

P(M|X=24) = 0.0166

Variables aléatoires indépendantes

- **X et Y deux variables aléatoires indépendantes ssi**

$$f_X(x|Y=y) = f_X(x) \qquad \forall y$$
$$f_Y(y|X=x) = f_Y(y) \qquad \forall x$$

- **en utilisant la formule donnant les densités conditionnelles :**

$$\boxed{f_{X,Y}(x,y) = f_X(x) \times f_Y(y) \qquad \forall x,y}$$

rappel du cas continu
P(A&B) = P(A) x P(B)

Et on augmente le nb de variables

- **densités conditionnelles**

$$f\left(x_1, x_2, \Lambda\ , x_k \middle| x_{k+1}, x_{k+2}, \Lambda\ , x_n\right) = \frac{f\left(x_1, x_2, \Lambda\ , x_n\right)}{f\left(x_{k+1}, x_{k+2}, \Lambda\ , x_n\right)}$$

- **indépendance**
  - **X_1 ... X_n sont indépendantes ssi** $\boxed{f\left(x_1, \Lambda\ , x_n\right) = f\left(x_1\right) \Lambda\ f\left(x_n\right)}$
  - **Attention :**
    - » **si *n* variables sont indépendantes, alors elles sont indépendantes *k* à *k* *(k < n)*** **VRAI**
    - » **la réciproque est FAUSSE**
  - **exemple**
    X_1 X_2 indépendantes
    X_1 X_3 indépendantes ⇏ **X_1 X_2 X_3 indépendantes**
    X_2 X_3 indépendantes

Exemple (en discret)

- 4 événements X1, X2, X3, X4 équiprobables $P(Xi)=1/4$

- $A=\{X1 \text{ ou } X2\}$ $P(A)=P(X1)+P(X2)=1/2$
- $B=\{X1 \text{ ou } X3\}$ $P(B)=1/2$
- $C=\{X1 \text{ ou } X4\}$ $P(C)=1/2$

- A, B, C indépendants 2 à 2 :
  - $A\&B = \{X1 \text{ ou } X2\} \&\{X1 \text{ ou } X3\} = X1$ $P(A\&B) = P(X1) = 1/4$
    $P(A)P(B)=1/2*1/2=1/4$
  - (idem avec A&C et B&C)

- A,B,C indépendantes ?
  - $A\&B\&C = X1$ $P(A\&B\&C)= P(X1) = 1/4$
    $P(A)P(B)P(C)= 1/8$

Non !

Plan du cours : Partie IV Réseaux Bayésiens

- Définitions des Modèles Graphiques Probabilistes
- Raisonnement probabiliste
- Réseaux bayésiens : Définition
- Indépendance conditionnelle et Théorème de Bayes
- La d-séparation

Estimation de densité

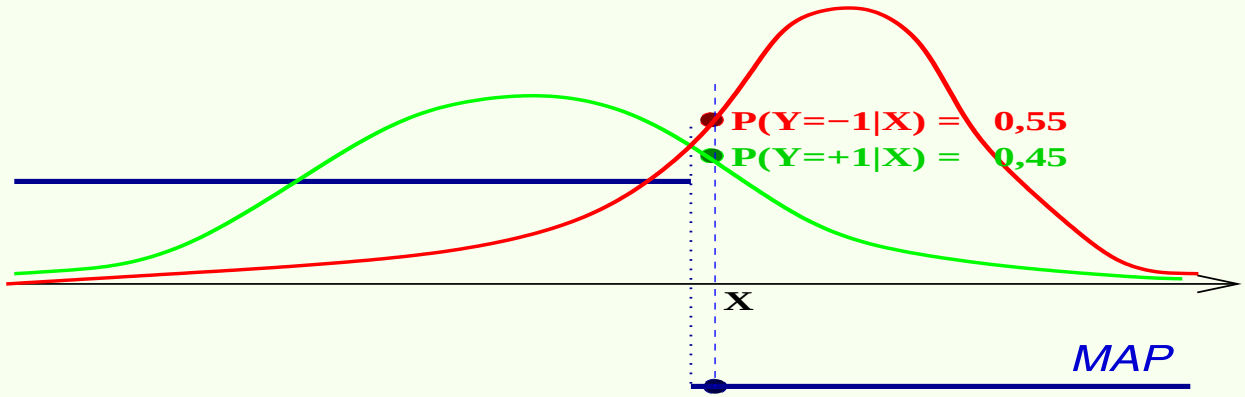
Remarque

Si $\mathbb{P}(X, Y)$ est connue,

1) classification (inférence $\Rightarrow \mathbb{P}(Y|X) \Rightarrow$ **MAP**) et

2) régression ($\mathbb{E}(Y|X)$ minimise le risque quadratique)

deviennent triviales.



Principe : **Maximum de vraisemblance**

Estimation de densité

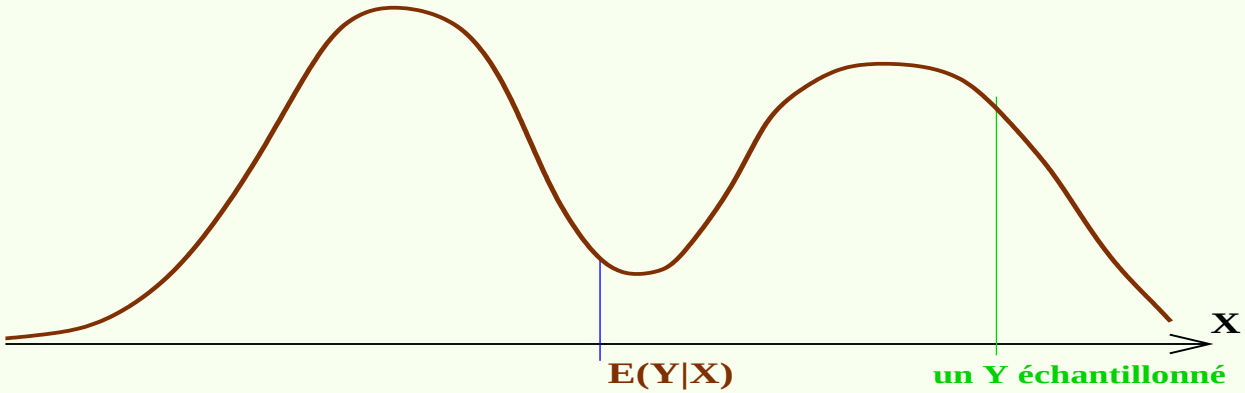
Remarque

Si $\mathbb{P}(X, Y)$ est connue,

1) classification (inférence $\Rightarrow \mathbb{P}(Y|X) \Rightarrow$ **MAP**) et

2) régression (échantillonnage à partir de $\mathbb{P}(Y|X)$)

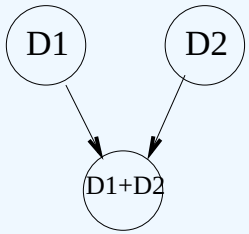
deviennent triviales.



Principe : **Maximum de vraisemblance**

Dirigé ou non-orienté ?

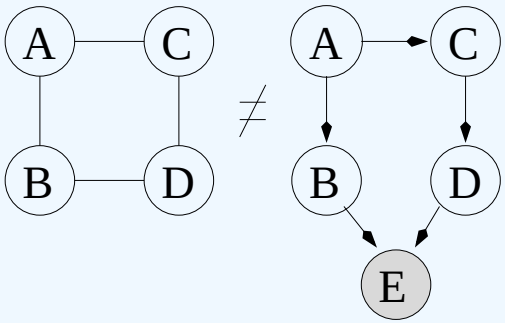
Deux avantages de l'aspect dirigé :



peut représenter simultanément
l'indépendance $D_1 \perp\!\!\!\perp D_2$ et
la dépendance $D_1 \not\perp\!\!\!\perp D_2 | Somme$

- Le sens des flèches "peut" représenter des relations de **causalité**, des relations **temporelles**, *etc.*

Deux inconvénients de l'aspect dirigé :



un modèle non dirigé peut représenter
$A \perp\!\!\!\perp D | \{B, C\}$ et $B \perp\!\!\!\perp C | \{A, D\}$,
un graphe dirigé doit se limiter à
$A \perp\!\!\!\perp D | \{B, C\}$ et $B \perp\!\!\!\perp C | \{A, D\}$.

- Taille de l'espace de recherche $\left(2^{\frac{n(n-1)}{2}} \text{ vs. } n^{2^{\mathcal{O}(n)}}\right)$

Objectif

Trouver comment les $\neq$ attributs interagissent ?

$\rightarrow$ *Utiliser une loi de probabilité*

- bon moyen de modéliser les systèmes *chaotiques* ou trop *complexes* (médecine, prévisions météo...)
- une même configuration des attributs *observés* peut provenir de *différents états*.

Avantage d'une telle modélisation :

- distribution de probabilité sur l'ensemble des configurations possibles.

Le décideur peut alors évaluer les risques.

Apprentissage de structure

Deux classes d'algos. :

Méthodes à base de détection de contraintes

Utiliser des **tests statistiques** pour tester les **indépendances conditionnelles** des attributs et en déduire une structure.

- → test du χ^2
- → test du rapport de vraisemblance
- → **information mutuelle**
- → étude des corrélations. . .

Méthodes à base de score (dites bayésiennes)

Maximiser une mesure/approximation de **la vraisemblance** bayésienne sur un espace de structures.

- → critère *AIC*
- → **critère** *BIC*
- → Minimum Description Length
- → approximation de Laplace au n-ième ordre
- → *AICc*, *ICL* . . .

Et la causalité dans tout ça ?

Une définition possible de la causalité :

$$\text{Si } t_1 < t_2$$
$$\text{et } \mathbb{P}(B_{t_2}|A_{t_1}) > \mathbb{P}(B_{t_2})$$
Alors A produit à t_1 est dit **cause** de B produit à t_2

Un exemple

Soit M la variable représentant **être malade**
et soit D celle représentant **des microbes se développent**
Laquelle est la cause ?

Et la causalité dans tout ça ?

Une définition possible de la causalité :

$$\text{Si } t_1 < t_2$$
$$\text{et } \mathbb{P}(B_{t_2}|A_{t_1}) > \mathbb{P}(B_{t_2})$$
Alors A produit à t_1 est dit **cause** de B produit à t_2

Un exemple

Soit M la variable représentant **être malade**
et soit D celle représentant **des microbes se développent**
Laquelle est la cause ?

Un exemple plus simple

Soit O la variable représentant **il va y avoir un orage**
et soit B celle représentant **le baromètre indique orageux**
Quelle est la cause ?

Et la causalité dans tout ça ?

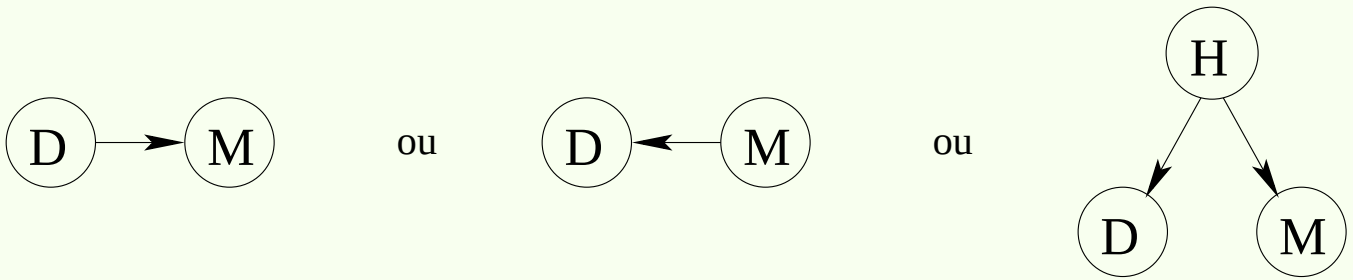
Si $t_1 > t_2$
et $\mathbb{P}(B_{t_2}|A_{t_1}) > \mathbb{P}(B_{t_2})$
Alors A produit à t_1 est dit **cause** de B produit à t_2

n'était donc pas une définition de la causalité...

Et la causalité dans tout ça ?

Si $t_1 > t_2$
et $\mathbb{P}(B_{t_2}|A_{t_1}) > \mathbb{P}(B_{t_2})$
Alors A produit à t_1 est dit **cause** de B produit à t_2

n'était donc pas une définition de la causalité...



Les modèles à variables latentes (ou facteurs) postulent l'existence de variables inobservables directement (intelligence, engagement religieux) mais dont on peut mesurer ou observer les effets (fréquentation des lieux de culte, réussite à certains tests).

$\longrightarrow$ Nécécite des expériences/données supplémentaires

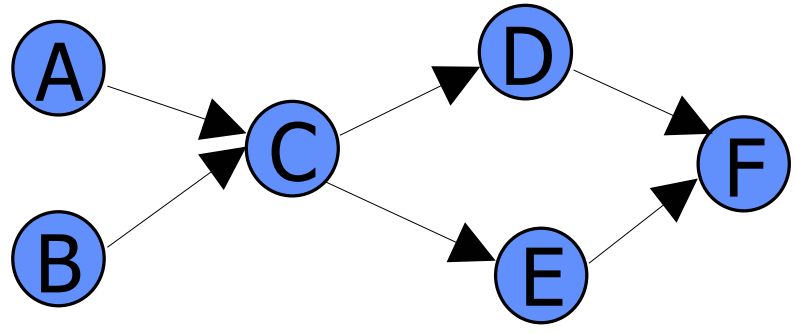
Utiliser au mieux les données

Pouvoir prendre en compte le maximum d'information :

- information imprécise (taille moyenne :
  20%|{<165} & 70%|(165;175) & 10%|{>175})
- information incertaine (taille petite ou moyenne :
  50%|{<165} & 45%|(165;175) & 5%|{>175})
- information manquante (taille inconnue :
  distribution *a posteriori*)
- information censurée (taille <170 :
  65%|{<165} & 35%|(165;175) & 0%|{>175})

Introduction

- Modèle graphique (MG) = mariage entre la théorie des graphes et la théorie des probabilités
  - graphe d'états
  - probabilités de transition



- De nombreux modèles utilisés en Machine Learning peuvent être vus comme des cas précis de MG
  - Modèles de Markov Cachés (HMM)
  - Réseaux Bayésiens (RB)
  - ...

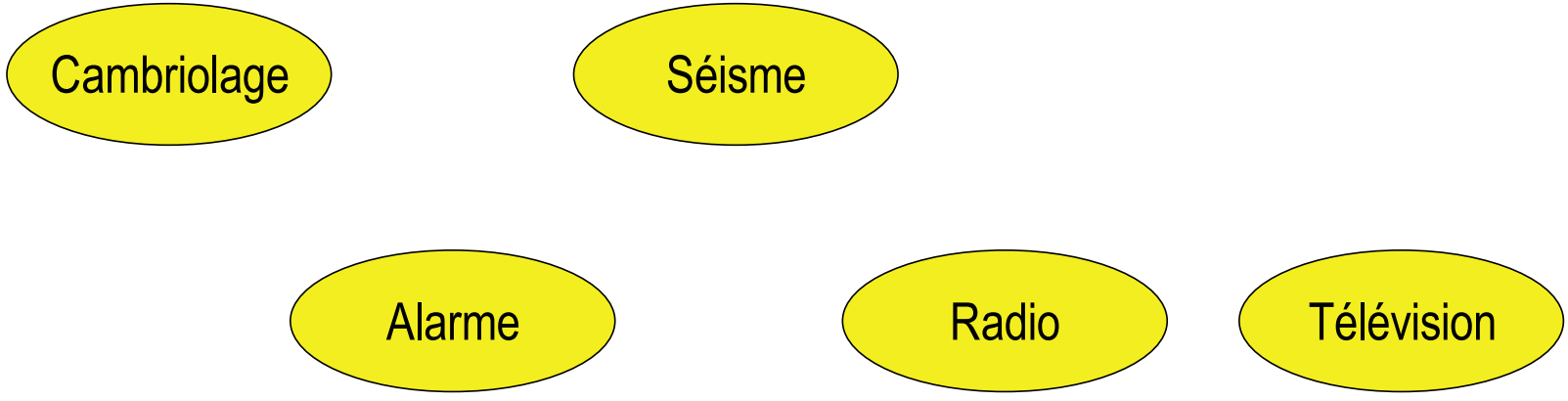
Origine

- Inspiration des systèmes experts :
  - règles :
    - si soleil=VRAI et arrosage=FAUX, alors sol=sec
  - comment rajouter des informations probabilistes ?
    - $P(soleil) = 0.3 \qquad P(arrosage) = 0.5 \qquad P(sec) = ?$
  - comment inverser l'inférence d'un SE ?
    - $P(sec) = 0.9 \qquad P(arrosage) = 0.5 \qquad P(soleil) ?$
- (Pearl 1988) Raisonnement probabiliste
- De nombreuses appellations :
  - SE bayesien, SE probabiliste, réseau de croyance, réseau causal
  - belief network, bayesian network, probabilistic independence networks

Réseau Bayésien

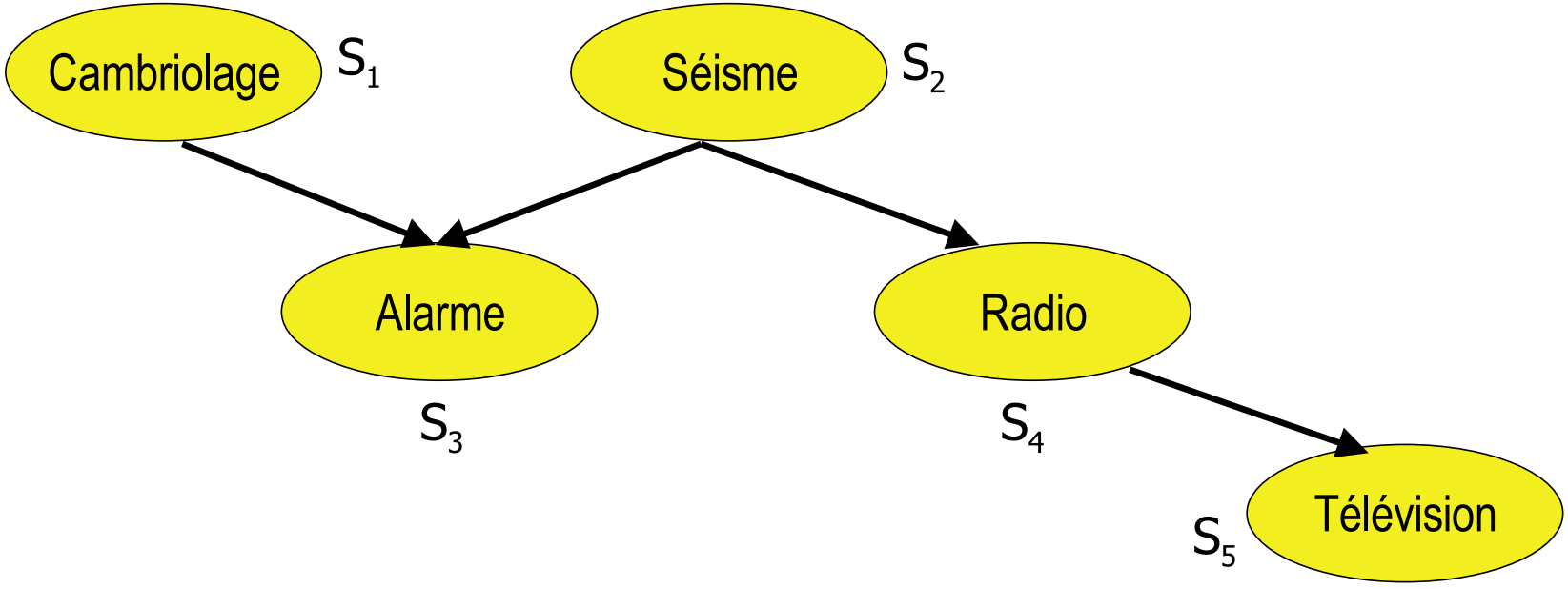
- Réseau bayésien =
  - description qualitative des dépendances entre des variables (graphe causal)
  - description quantitative de ces dépendances

- Exemple



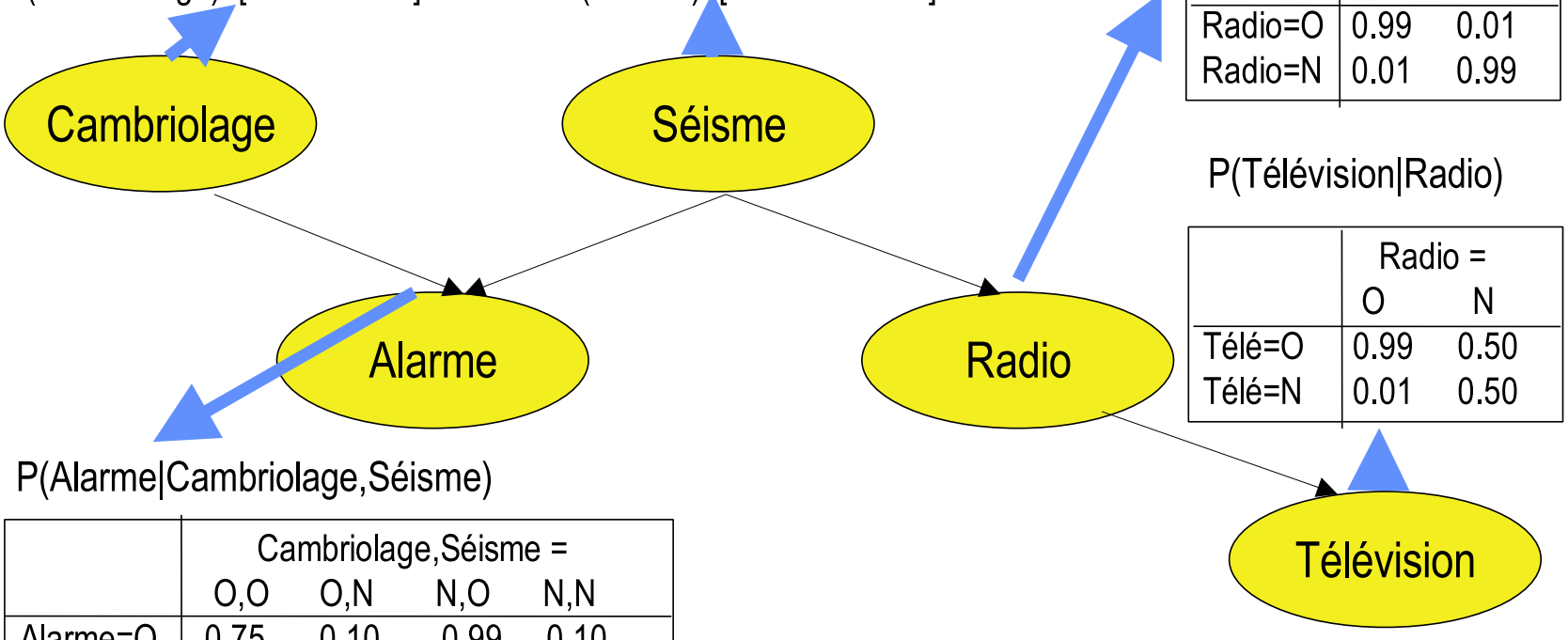
Exemple

- Dépendance entre les variables (graphe)
  - DAG = Directed Acyclic Graph
  - Numérotation des variables dans l'ordre topologique



Exemple

- Description quantitative des dépendances (probabilités conditionnelles



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

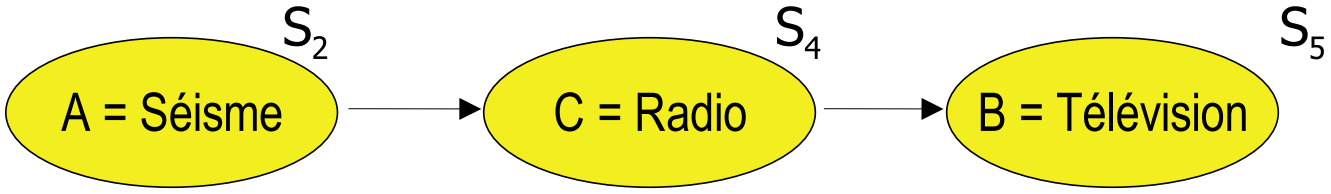
P(Alarme|Cambriolage,Séisme)

	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

Indépendance Conditionnelle

- Indépendance conditionnelle :
  - A et B sont indépendants conditionnellement à C ssi :
    - lorsque l'état de C est connu, toute connaissance sur B n'altère pas A
    - $$P(A|B,C) = P(A|C)$$
- Les RB vont servir à représenter graphiquement les indépendances conditionnelles
- Exemple sur 3 nœuds
  - 3 types de relations possibles entre A, B et C ...

Indépendance Conditionnelle

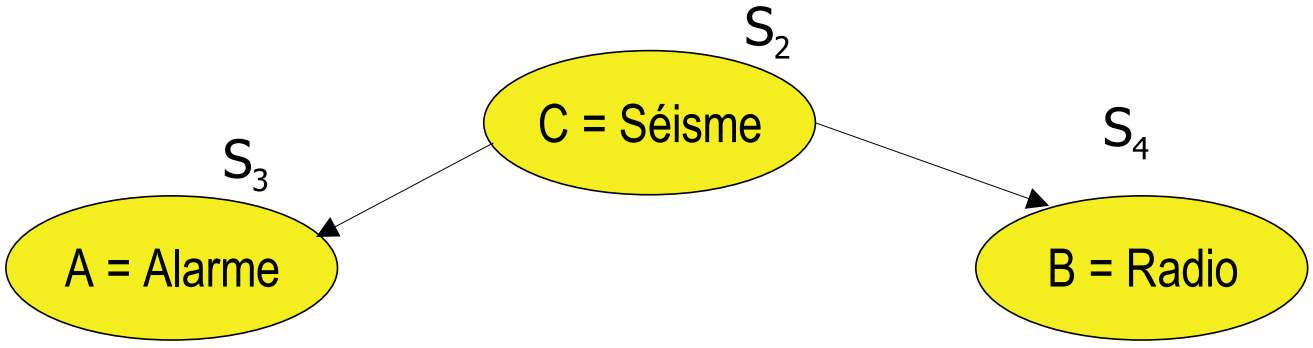


- Connexion série
- A et B sont dépendants

- A et B sont indépendants conditionnellement à C
  - si $P(C)$ est connue,
    A n'intervient pas dans le calcul de $P(B)$

  - $P(S_5|S_4, S_2) = P(S_5|S_4) = P(S_5|parents(S_5))$

Indépendance Conditionnelle

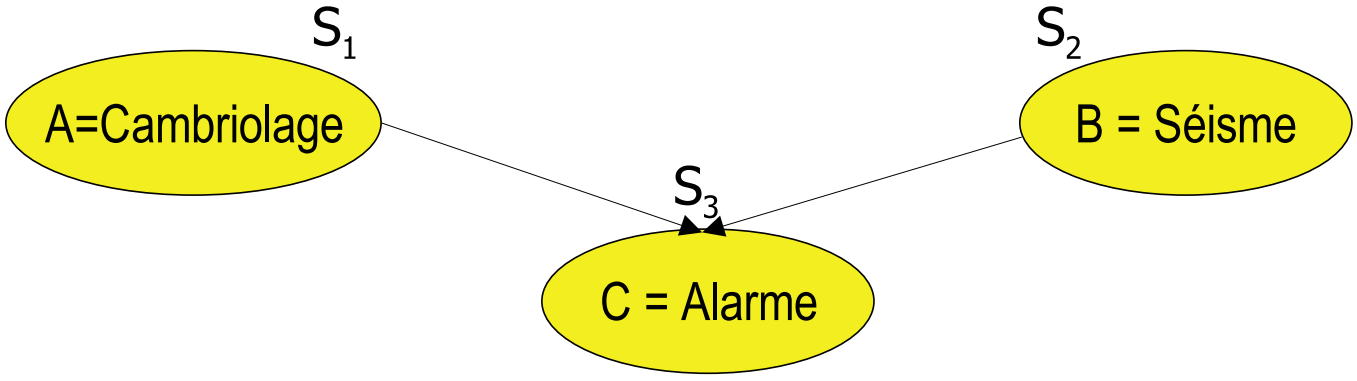


- Connexion divergente
- A et B sont dépendants

- A et B sont indépendants conditionnellement à C
  - si $P(C)$ est connue,
    A n'intervient pas dans le calcul de $P(B)$

  - $P(S_4|S_2, S_3) = P(S_4|S_2) = P(S_4|parents(S_4))$

Indépendance Conditionnelle



- Connexion convergente (A, B, C est une V-structure)
- A et B sont indépendants

- A et B sont dépendants conditionnellement à C
  - si $P(C)$ est connue,
    $P(A)$ intervient pas dans le calcul de $P(B)$

  - $P(S_3|S_1, S_2) = P(S_3|parents(S_3))$

Conséquence

- RB = représentation compacte de la loi jointe $P(S)$
- Théorème de Bayes :

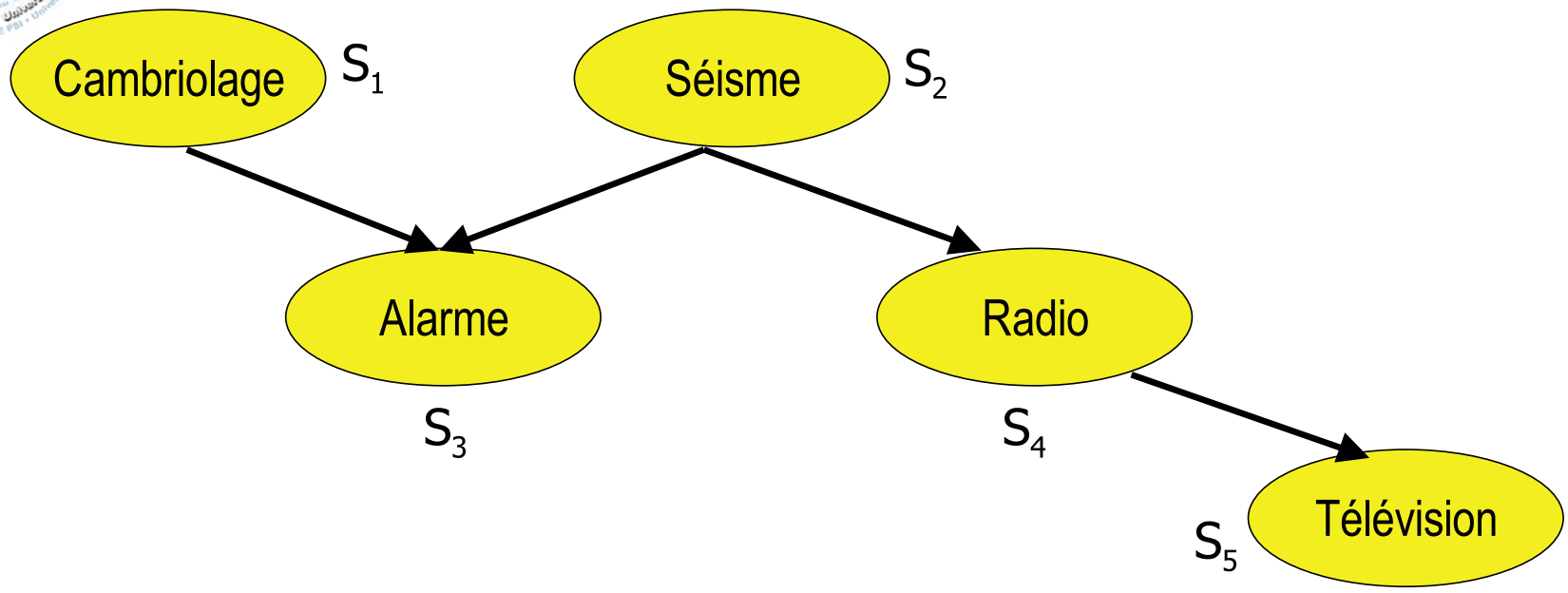
$$P(S) = P(S_1) \times P(S_2|S_1) \times P(S_3|S_1, S_2) \times \cdots \times P(S_n|S_1 \ldots S_{n-1})$$

- Mais 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 locales

Exemple

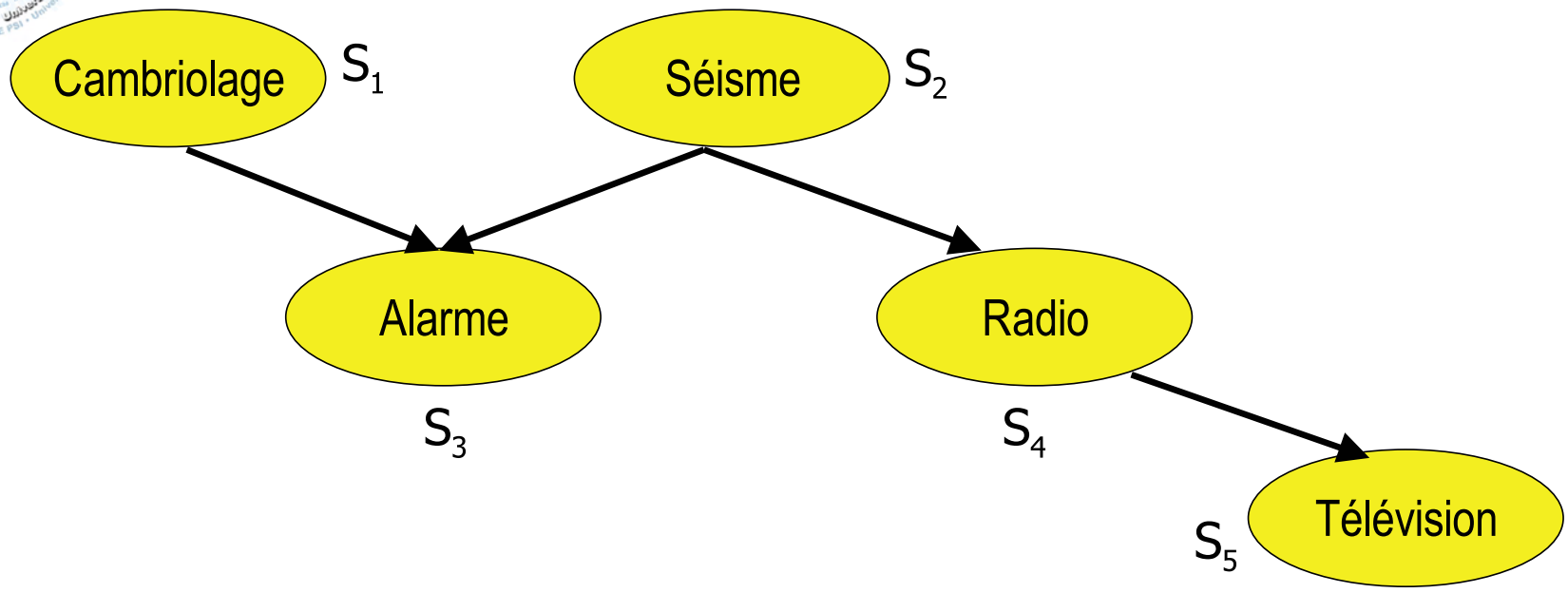


$$P(Cambriolage, Seisme, Alarme, Radio, Tele) =$$
$$P(S_1)P(S_2|S_1)P(S_3|S_1,S_2)P(S_4|S_1,S_2,S_3)P(S_5|S_1,S_2,S_3,S_4)$$
$$P(S_1) \quad P(S_2) \quad P(S_3|S_1,S_2) \quad P(S_4|S_2) \quad P(S_5|S_4)$$

La d-séparation

- Principe
  - déterminer si 2 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
    - 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

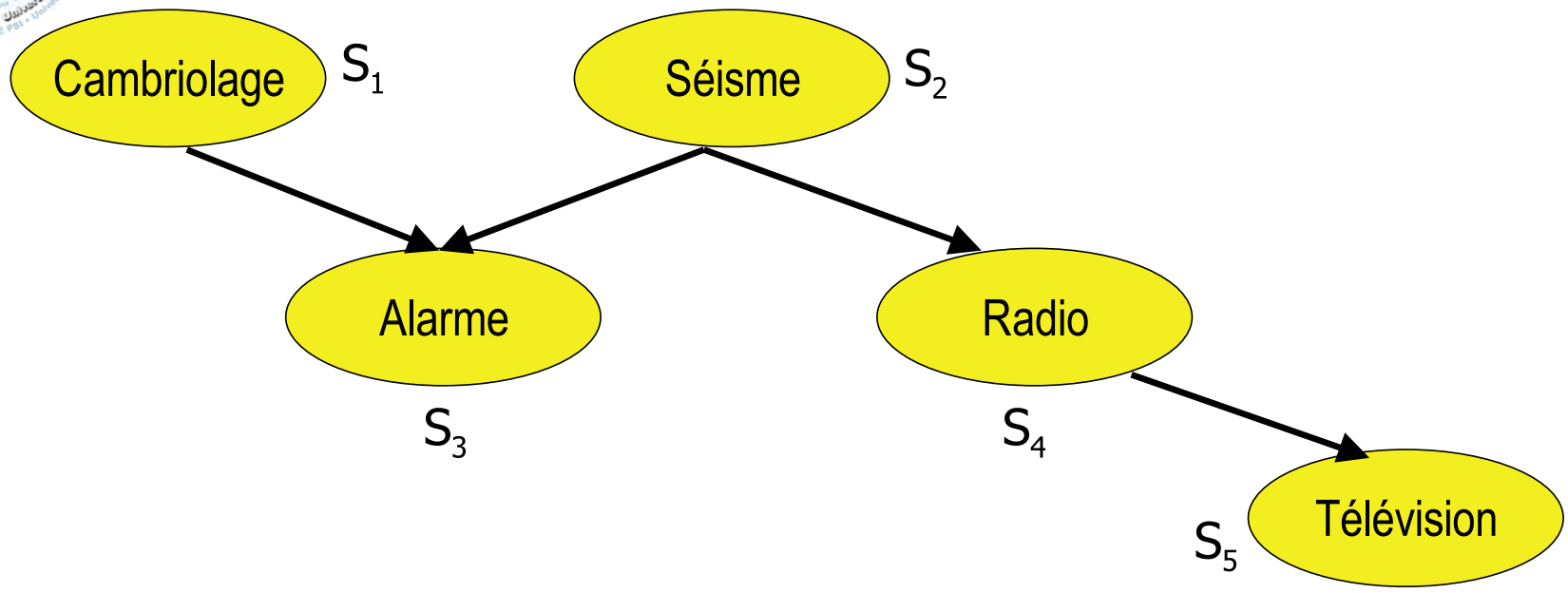
Exemple



- exemple de d-séparation : $S_1 \ldots S_4$?
  - $V = S_3$ sur le chemin entre S_1 et S_4.
  - la connexion est convergente en V
  - V n'est pas instancié

  $\rightarrow$ S_1 et S_4 sont d-séparés
  (si S_3 était mesuré, S_1 et S_4 seraient d-connectés)

Exemple



- autre exemple de d-séparation : $S_2 \ldots S_5$?
  - $V = S_4$ sur le chemin entre S_2 et S_5.
  - la connexion est série en V
  - V n'est pas instancié

  $\rightarrow$ S_2 et S_5 sont d-connectés
  (si S_4 était mesuré, S_2 et S_5 seraient d-séparés)

Plan du cours : Partie V Inférence dans les RB

- Marginalisation : *Bucket Elimination*
- *Message Passing* de Pearl
- *Junction tree* de Jensen
- Exemple d'application

Inférence

- Inférence = calcul de n'importe quelle $P(S_i|S_j = x)$
  NB : l'observation $\{S_j = x\}$ est appellée l'évidence
- A quoi sert la loi jointe $P(S) = P(S_1, \ldots, S_n)$?
- Rappel : marginalisation

$$P(S_i) = \sum_{s_1, s_2 \ldots s_n} P(S_1 = s_1, S_2 = s_2, \ldots, S_i, \ldots S_n = s_n)$$

- d'où

$$P(S_i|S_j = x) = \frac{P(S_i, S_j = x)}{P(S_j = x)} = \frac{\sum_{\{s_k\}_{k \neq i,j}} P(S_1{=}s_1, \ldots, S_i, S_j{=}x, \ldots S_n{=}s_n)}{\sum_{\{s_k\}_{k \neq j}} P(S_1{=}s_1, \ldots, S_j{=}x, \ldots S_n{=}s_n)}$$

- Un RB décompose cette loi jointe, ce qui permet de simplifier les calculs

Quelques algorithmes d'inférence

- Algorithmes exacts
  - Bucket Elimination
  - Message Passing (Pearl 88) pour les arbres
  - Junction Tree (Jensen 90)

Problème = explosion combinatoire de ces méthodes pour des graphes fortement connectés, etc . . .
(inférence = problème NP-complet)

- Algorithmes approchés
  - Echantillonnage : Markov Chain Monte Carlo, ...
  - Méthodes variationnelles

Bucket Elimination

- Principe
  - grâce à la décomposition de la loi jointe, certaines étapes de la marginalisation de $P(S_i, S_j = x)$ se simplifient

- Exemple
  - évidence $E = \{S_4 = O\}$, on cherche $P(S_2|E)$

$$P(S, E) = P(S_1)P(S_2)P(S_3|S_1S_2)P(S_4 = 0|S_2)P(S_5|S_4 = O)$$

$$P(S_2, E) = \sum_{S_1,S_3,S_5} P(S, E)$$

  - et si on choisit l'ordre des variables pour la marginalisation ?

Bucket Elimination

- Commençons par S_5

$$\sum_{S_5} P(S_1, S_2, S_3, S_4 = O, S_5) = P(S_1)P(S_2)P(S_3|S_1, S_2)\dots$$

$$\dots \quad P(S_4 = O|S_2) \sum_{S_5} P(S_5|S_4 = O)$$

- Cette dernière somme vaut 1 ! On a éliminé S_5

$$P(S_1, S_2, S_3, S_4 = O) = P(S_1)P(S_2)P(S_3|S_1, S_2)P(S_4 = O|S_2)$$

- Au tour de S_1

Bucket Elimination

$$\sum_{S_1} P(S_1, S_2, S_3, S_4 = O) \quad = \quad P(S_2)P(S_4 = O|S_2) \ldots$$

$$\ldots \quad \sum_{S_1} P(S_1)P(S_3|S_1, S_2)$$

- Cette dernière somme nous rend une table dépendant de S_2 et S_3 : $T(S_2, S_3)$

$$P(S_2, S_3, S_4 = O) \quad = \quad P(S_2)P(S_4 = O|S_2)T(S_2, S_3)$$

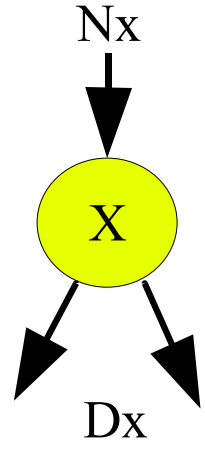
- Idem avec S_3 pour obtenir $P(S_2, S_4 = O)$

Marginalisation = série de produits locaux de matrices et de marginalisations locales

Message Passing (Pearl 1988)

- 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 ?
  - Si X instancié, $\lambda(X) = [001 \ldots 0]$
    (la position du 1 correspond à la valeur donnée à X
  - sinon
    - si X est une feuille, $\lambda(X) = [1 \ldots 1]$
    - sinon

$$\lambda(X = x) = \Pi_{Y \in Enf(X)} \lambda_Y(X = x)$$

Message Passing

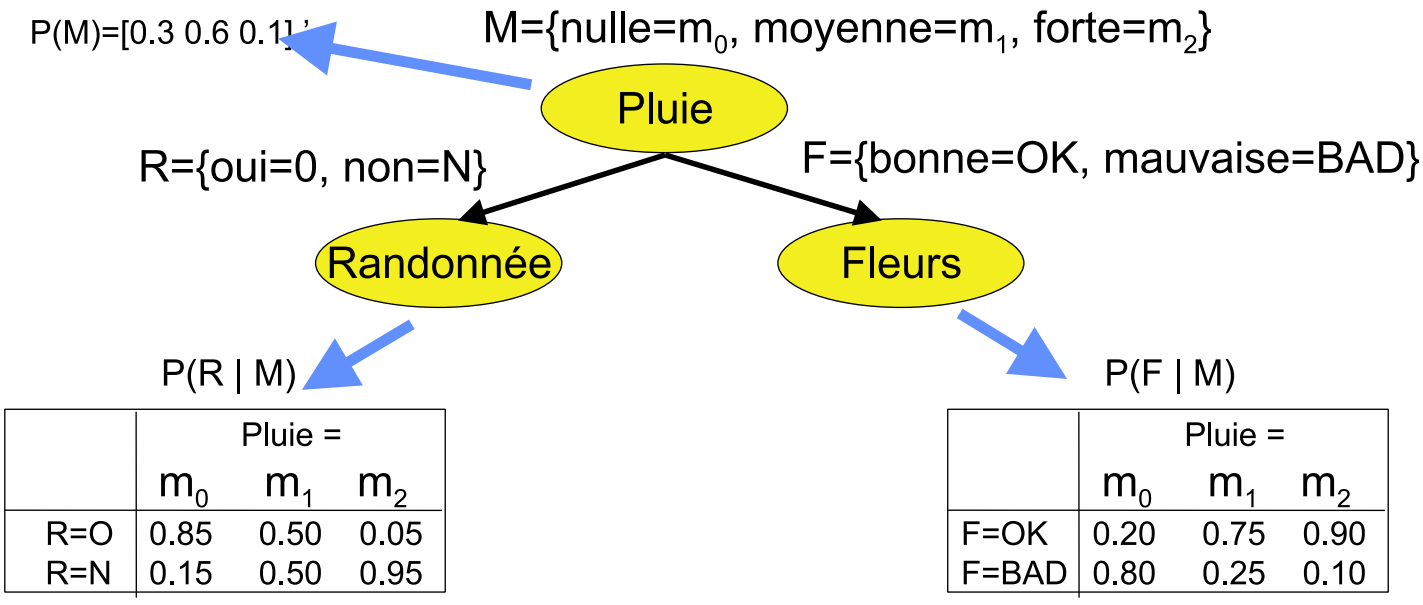
- Les messages π
  - Pour Z l'unique parent de X,

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

- Comment calculer π en chaque nœud ?
  - Si X instancié, $\lambda(X) = [001 \ldots 0]$
    (la position du 1 correspond à la valeur donnée à X
  - sinon
    - si X est la racine, $\pi(X) = P(X)$
    - sinon

$$\pi(X = x) = \sum_z P(X = x|Z = z)\pi_X(Z = z)$$

Exemple



	Pluie = m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

	Pluie = m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

- $E = \emptyset$

$P(F) = ?$

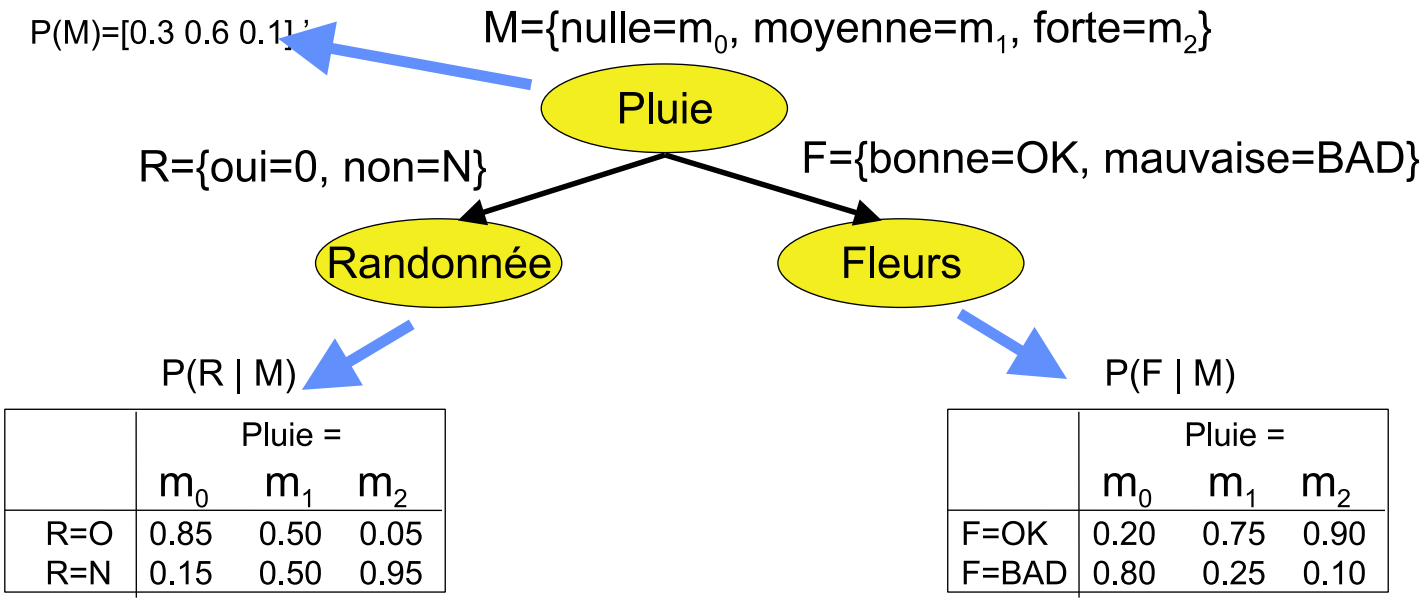
$P(R) = ?$

$$P(F) = \sum_m P(F|M=m)P(M=m)$$

			0.3
			0.6
			0.1
0.20	0.75	0.90	0.6
0.80	0.25	0.10	0.4

			0.3
			0.6
			0.1
0.85	0.50	0.05	0.56
0.15	0.50	0.95	0.44

Exemple

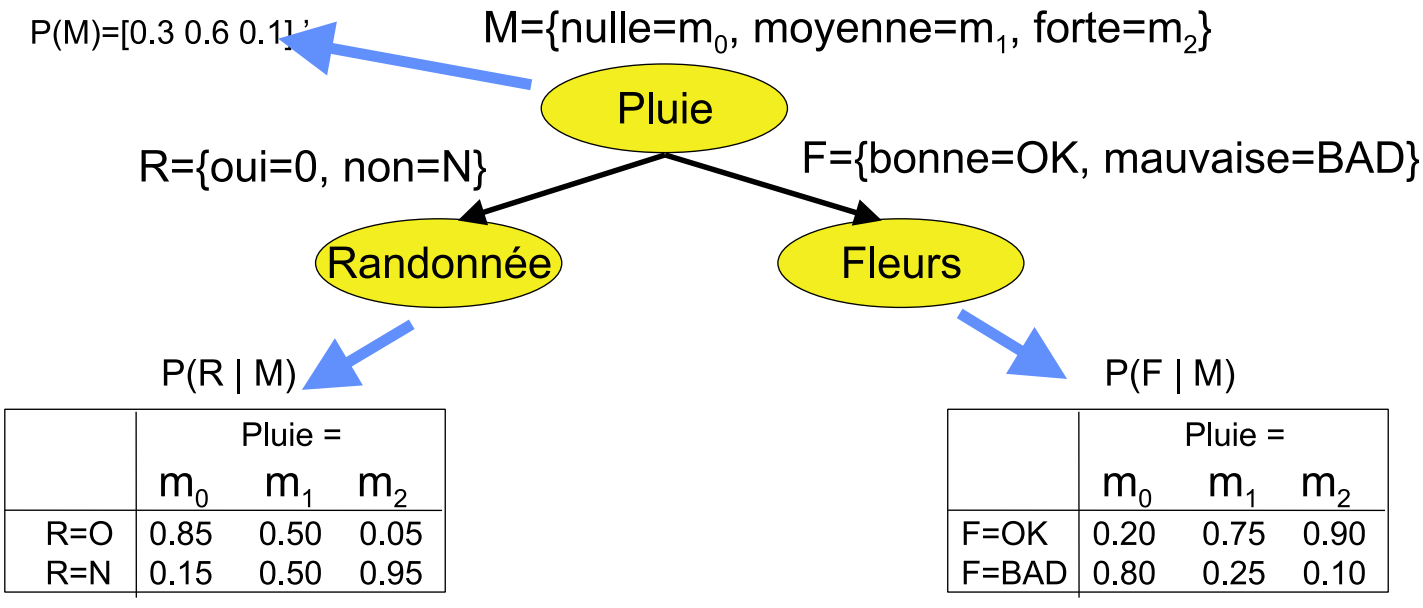


	Pluie = m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

	Pluie = m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

- $E = \{M = m_2\}$
  - $\lambda(M) = \pi(M) = [0 \;\; 0 \;\; 1]$ (nœud instancié)
  - $P(M|E) \propto \lambda(M)\pi(M) = [0 \;\; 0 \;\; 1]$ (logique)
  - messages envoyés aux enfants :
    - $\pi_R(M) = \pi(M)\lambda_F(M) = [0 \;\; 0 \;\; 1]$
    - $\pi_F(M) = \pi(M)\lambda_R(M) = [0 \;\; 0 \;\; 1]$

Exemple



	Pluie = m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

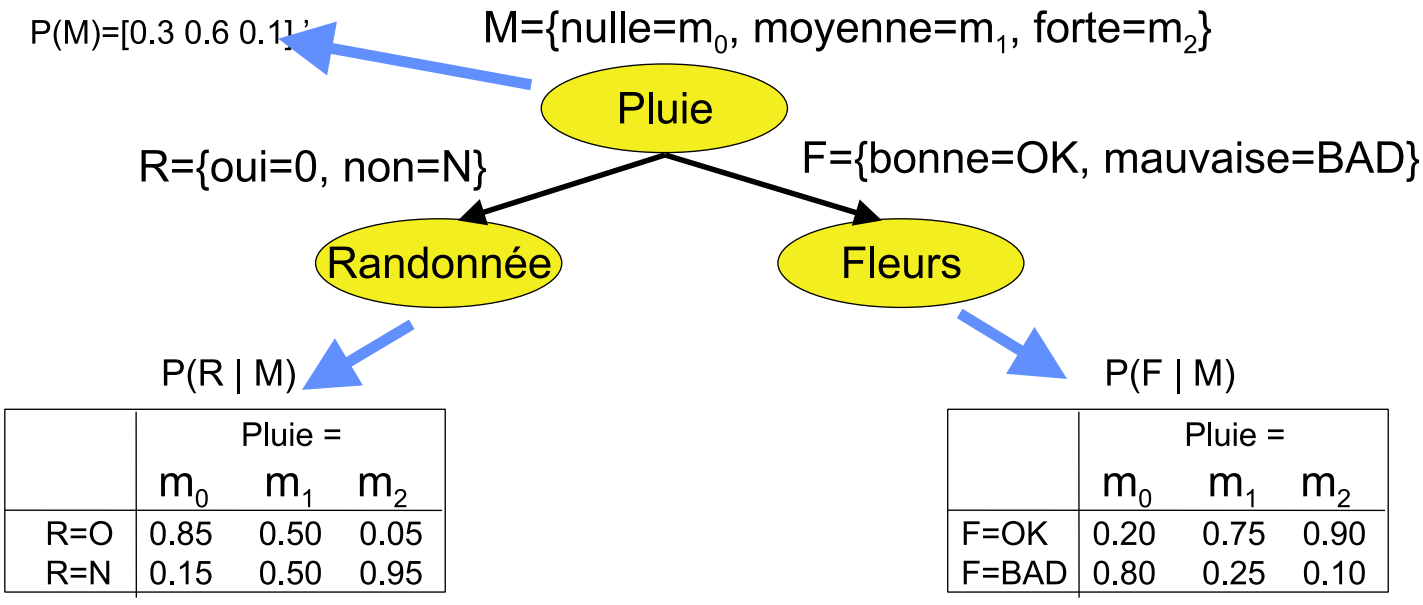
	Pluie = m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

- $E = \{M = m_2\}$ suite...
  - en R :
    - $\pi(R) = P(R|M)\pi_R(M) = [0.05 \;\; 0.95]$
    - $\lambda(R) = [1 \;\; 1]$ (feuille)
    - $P(R|E) \propto \lambda(R)\pi(R) = [0.05 \;\; 0.95]$

$$\begin{bmatrix} 0.85 & 0.50 & 0.05 \\ 0.15 & 0.50 & 0.95 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.05 \\ 0.95 \end{bmatrix}$$

Pluie=Forte $\Rightarrow$ Randonnée=Non

Exemple



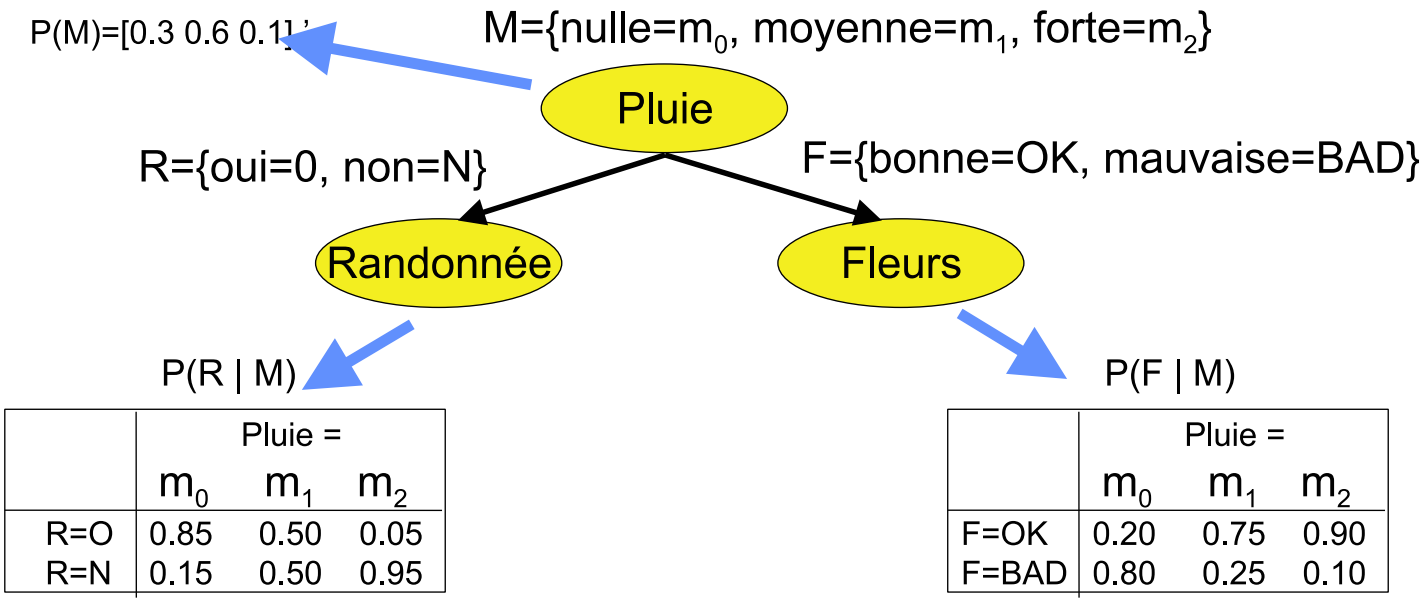
	Pluie = m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

	Pluie = m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

- $E = \{M = m_2\}$ fin
  - en F :
    - $\pi(F) = P(F|M)\pi_F(M) = [0.9 \;\; 0.1]$
    - $\lambda(F) = [1 \;\; 1]$ (feuille)
    - $P(F|E) \propto \lambda(F)\pi(F) = [0.9 \;\; 0.1]$

Pluie=Forte $\Rightarrow$ Fleurs arrosées

Exemple



	Pluie = m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

	Pluie = m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

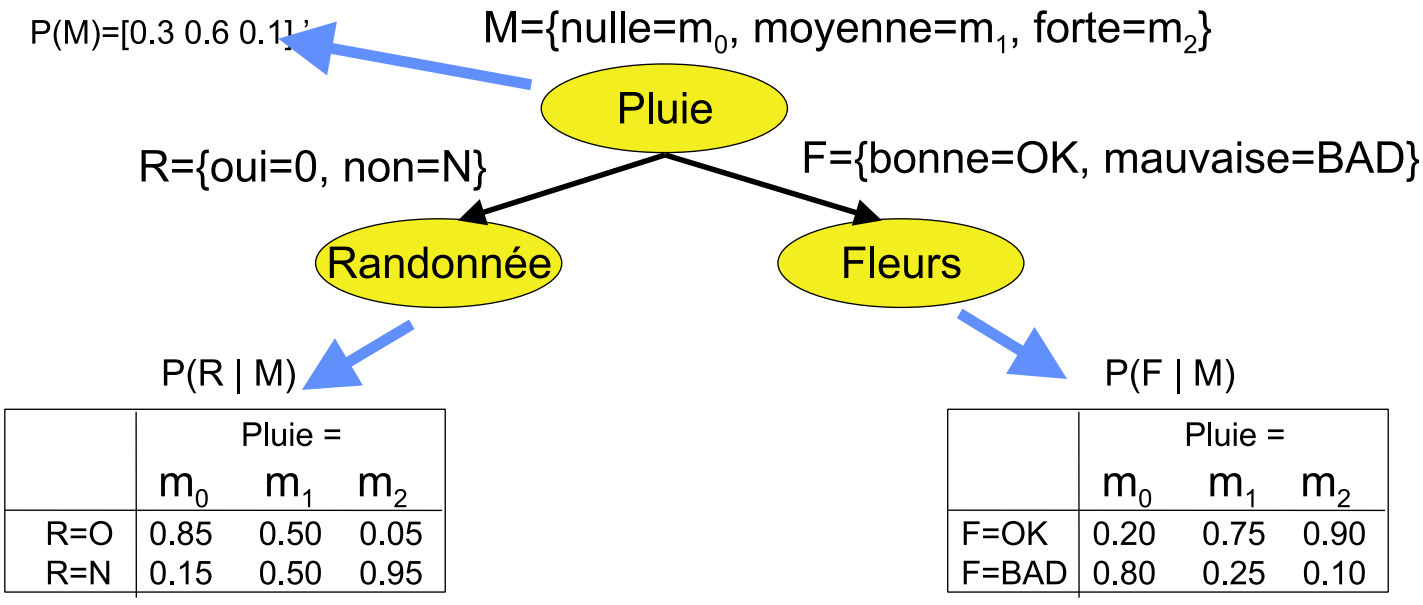
- $E = \{F = OK(0.2) | BAD(0.8)\}$ (soft evidence)

En F :

- $\lambda(F) = \pi(F) = [0.2 \;\; 0.8]$
- $P(F|E) \propto \lambda(F)\pi(F) = [0.2 \;\; 0.8]$
- message envoyé au parent :
  - $\lambda_F(M) = \lambda(F)P(F|M) = [0.68 \;\; 0.35 \;\; 0.26]$

	0.20	0.75	0.90
	0.80	0.25	0.10
0.2 0.8	0.68	0.35	0.26

Exemple



	Pluie = m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

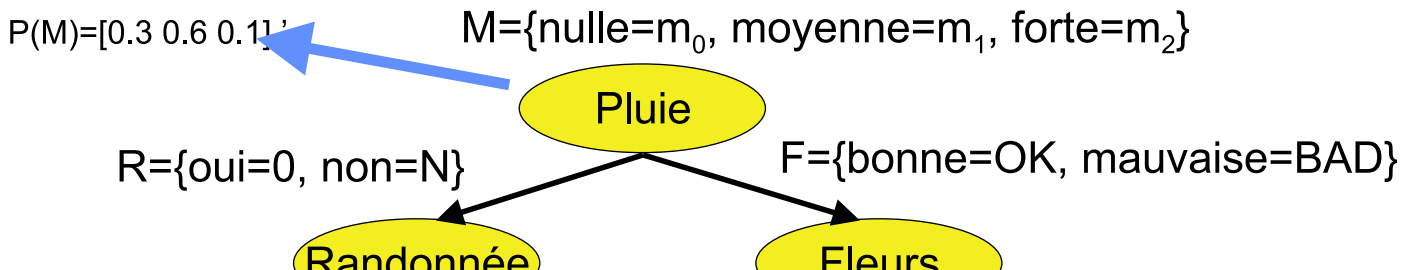
	Pluie = m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

- $E = \{F = OK(0.2) | BAD(0.8)\}$ (soft evidence)

En R :

- $\lambda(R) = [1 \;\; 1]$ (feuille)
- $\pi(R) = ?$
- message envoyé au parent :
  - $\lambda_R(M) = \lambda(R)P(R|M) = [1 \;\; 1 \;\; 1]$

Exemple



P(R | M)

	Pluie = m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

P(F | M)

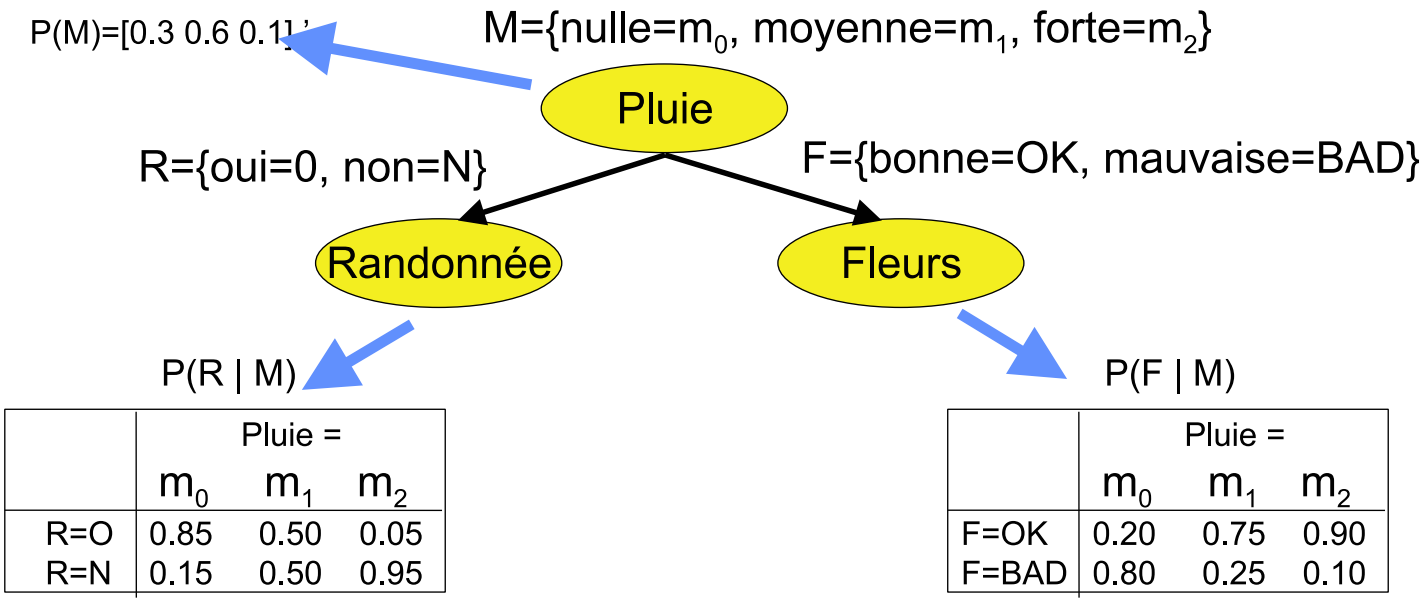
	Pluie = m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

- $E = \{F = OK(0.2) | BAD(0.8)\}$ (soft evidence)

En M :

- $\lambda(M) = \lambda_R(M)\lambda_F(M) = [0.68 \;\; 0.35 \;\; 0.26]$
- $\pi(M) = P(M) = [0.3 \;\; 0.6 \;\; 0.1]$ (racine)
- $P(M|E) \propto \lambda(M)\pi(M) = [0.463 \;\; 0.477 \;\; 0.060]$
- message : $\pi_R(M) = \pi(M)\lambda_F(M) = [0.204 \; 0.216 \; 0.026]$

Exemple



	Pluie = m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

	Pluie = m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

- $E = \{F = OK(0.2) | BAD(0.8)\}$ (soft evidence)

Retour en R :

- $\lambda(R) = [1 \;\; 1]$ (feuille)
- $\pi(R) = P(R|M)\pi_R(M) = [0.283 \;\; 0.229]$
- $P(R|E) \propto \lambda(R)\pi(R) =$ $[0.553 \;\; 0.447]$

Fleurs plutôt en mauvais état $\Rightarrow$ Randonnée = plutôt oui

Junction Tree (Jensen 1990)

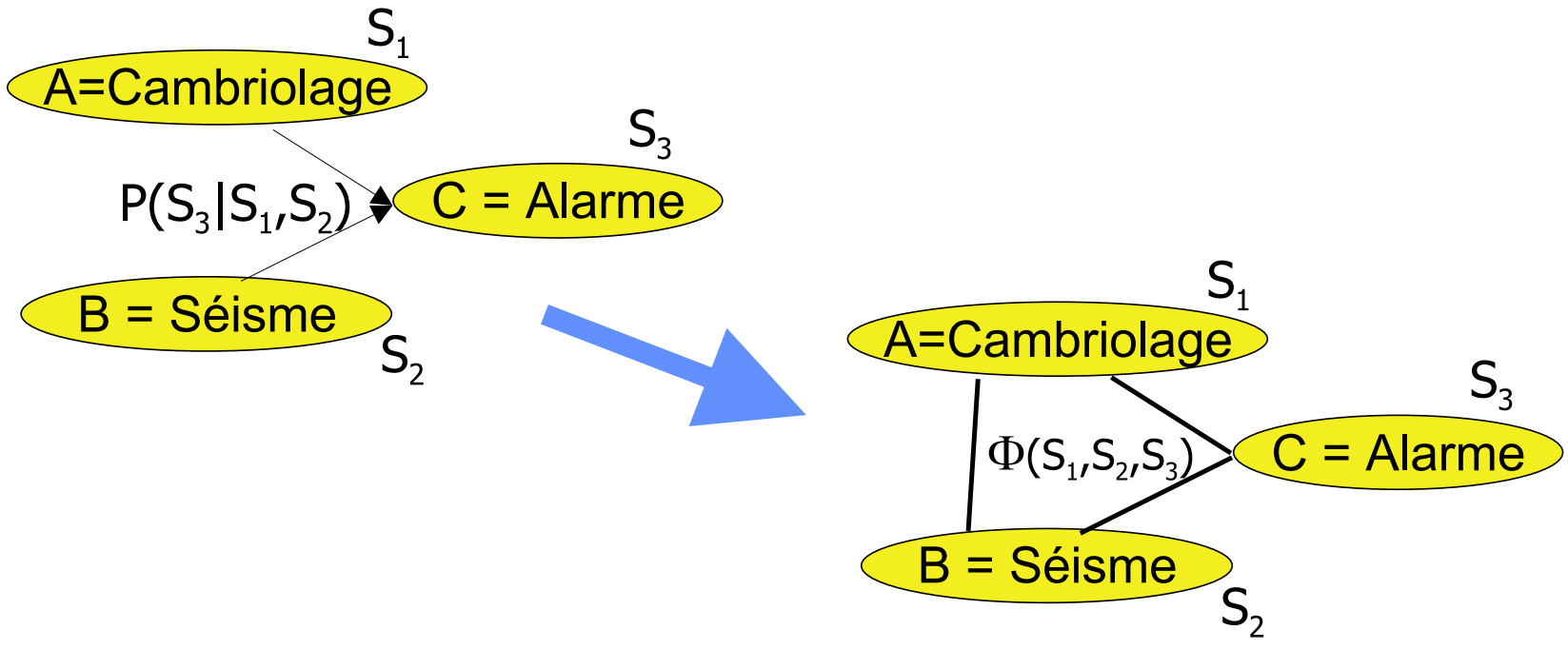
- Message Passing ne s'applique bien qu'aux arbres
- Besoin d'un algorithme plus général

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

- Moralisation = marier les parents et "désorienter" le graphe
- Triangulation = éviter les cycles dans le graphe non orienté.

Junction Tree

- Moralisation : marier les parents de chaque nœud



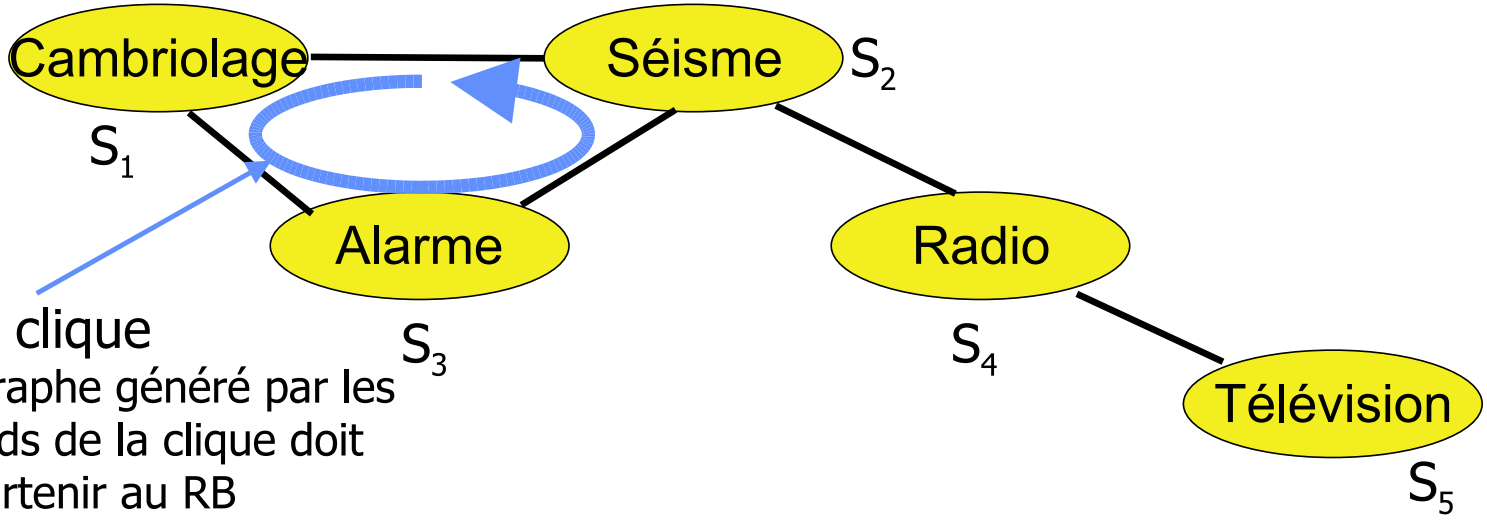
Junction Tree

- 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 > 3).

- Triangulation optimale pour des graphes non-dirigés = NP-difficile (comment choisir les meilleures cordes ?)

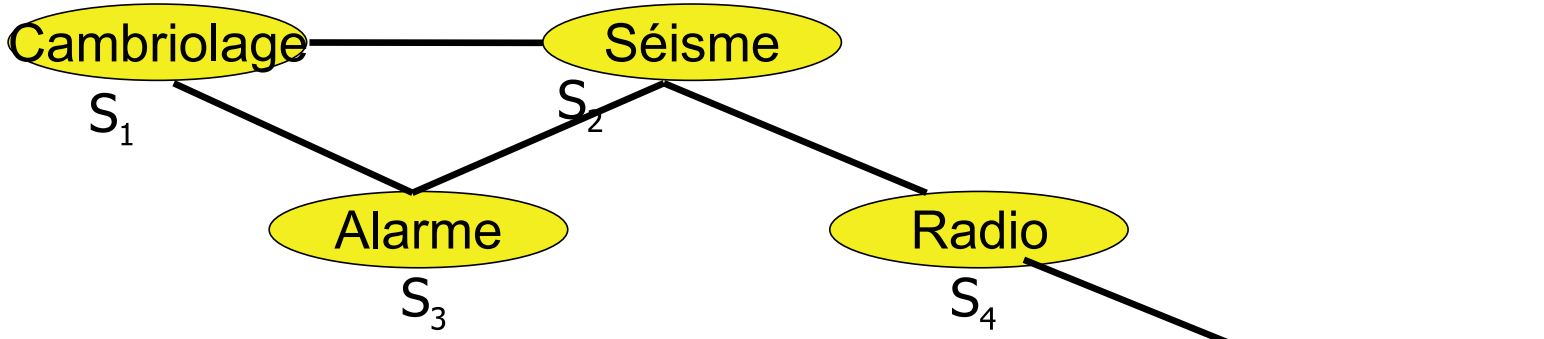
Junction Tree

- Clique = sous-graphe du RB 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

Applications

- Diagnostic et raisonnement dans des systèmes complexes
- Marketing/Finance (modélisation de risques) :
  - ATT : détection de fraudes (mauvais payeurs) pour les factures de téléphone
  - Altaprofit : optimation de portefeuilles (contrats d'assurance vie)
- Informatique :
  - Microsoft : printer troubleshooting, assistant Office
  - MODIST : évaluation de la qualité pour des développements logiciels

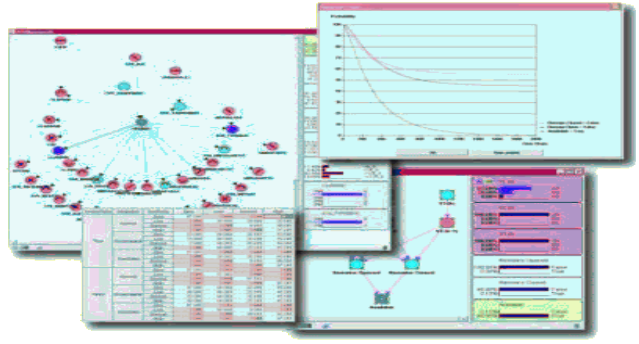
Applications

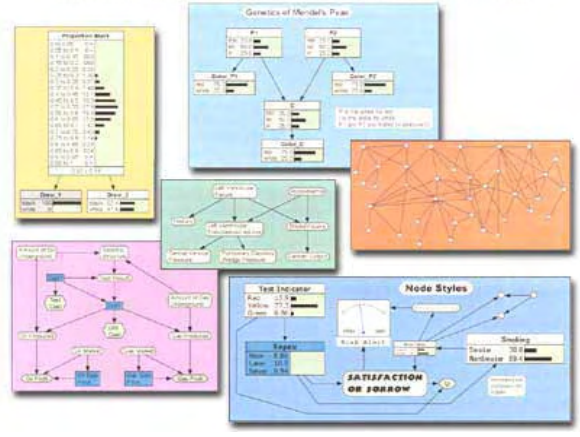
- Médecine :
  - Aide au diagnostic de problèmes cardio-vasculaires
  - Surveillance transfusionnelle, ...

- Industrie :
  - NASA : aide au diagnostic de pannes en temps réel pour les systèmes de propulsion de la navette spatiale
  - Lockheed Martin : système de contrôle d'un véhicule sous-marin autonome
  - Ricoh : aide au télédiagnostic
  - EDF : modélisation de groupes electrogènes

Offre logicielle

- Toolbox
  - Bayes Net Toolbox (BNT) pour Matlab
  - gR, GRAPPA, ... pour $\mathcal{R}$
  - BNJ, JavaBayes, ... pour Java
- Logiciels non commerciaux
  - Microsoft Belief Network [US]
  - Genie 2/Smile [US]
- Logiciels commerciaux
  - Bayesia [FR]
  - ProBT (inférence probabiliste) [FR]
  - Hugin [DK]
  - Netica [CA]





Plan du cours : Partie VI Apprentissage de RB

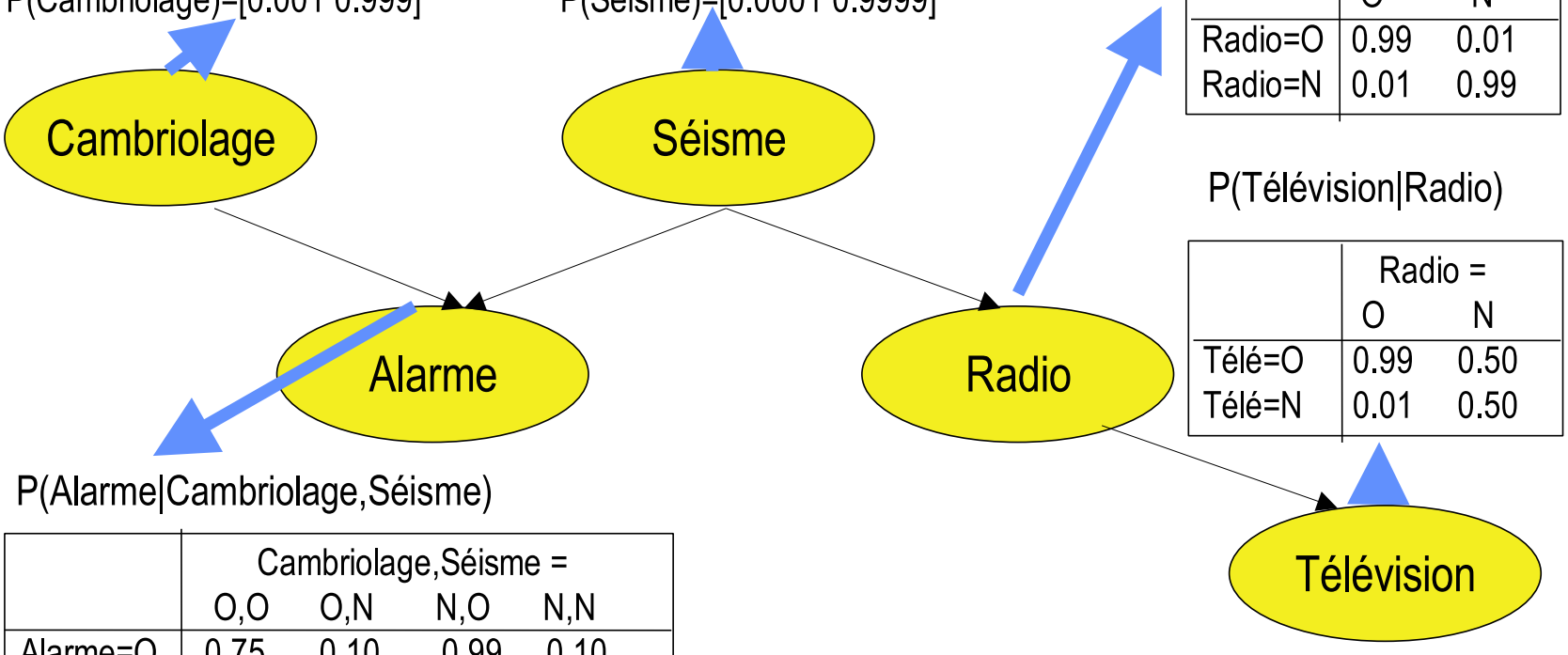
- Apprentissage des paramètres
  - Approche Statistique
  - Approche Bayésienne
- Algorithme EM
- Apprentissage de Structure
  - Espace de recherche
  - Notion de Score
  - Recherche de causalité
  - Algorithme K2
  - Recherche Gloutonne
- Variables latentes

Références

- **Les Réseaux Bayésiens** - P. Naïm, P.H. Wuillemin, Ph. Leray, O. Pourret, A. Becker (Eyrolles)
- **Probabilistic reasoning in Intelligent Systems : Networks of plausible inference** - J. Pearl (Morgan Kaufman)
- **An introduction to Bayesian Networks** - F. Jensen (Springer Verlag)
- **Probabilistic Networks and Expert Systems** - R.G. Cowell & al. (Springer Verlag)
- **Learning Bayesian Networks** - R. Neapolitan (Prenctice Hall)
- **Learning in Graphical Models** - Jordan M.I. ed. (Kluwer)

Réseau Bayésien

- Réseau bayésien =
  - des variables
  - un graphe entre ces variables
  - des probabilités conditionnelles



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

Plan

- Définition
- Algorithmes d'inférence
  - Bucket Elimination
  - Message Passing (Pearl)
  - Junction Tree (Jensen)
- Applications et Offre logicielle

- Apprentissage
  - des paramètres
  - de la structure

- Modèles étendus
  - variables continues : modèles conditionnels gaussiens
  - problèmes temporels : modèles dynamiques
  - théorie de la décision : diagrammes d'influence

Apprentissage des paramètres

- Structure fixée
- Il faut déterminer chaque $P(X_i = x_k | pa(X_i) = x_j) = \theta_{i,j,k}$
- Notons $\theta = \{\theta_{i,j,k}\}_{i=1:n,\ j=1:q_i,\ k=1:r_i}$

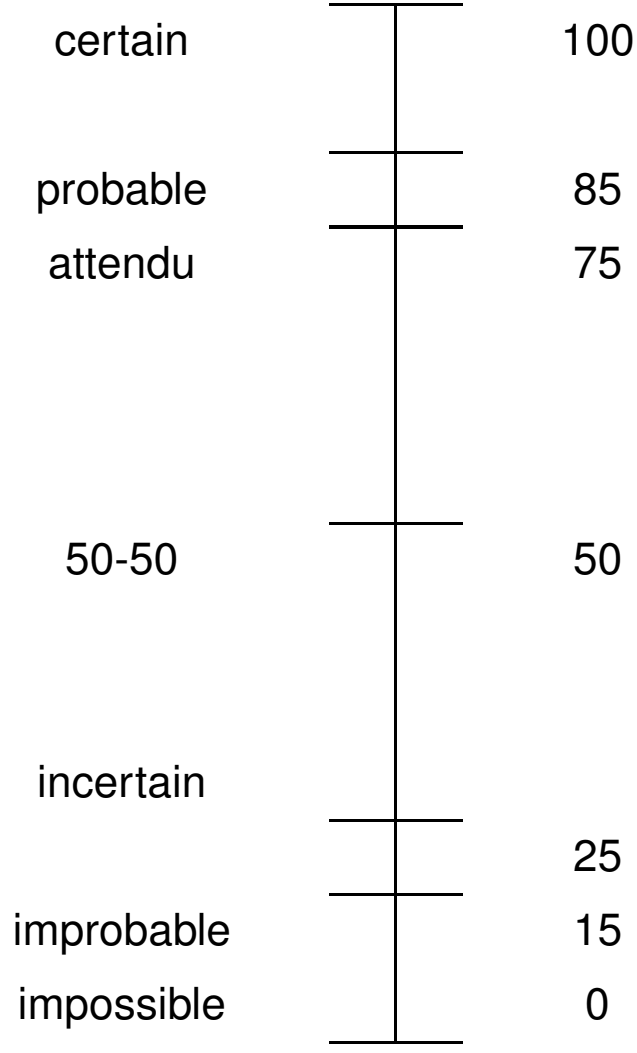
- 2 approches

  - **acquisition de connaissances**
    - avec un expert du domaine.

  - **apprentissage à partir de données**,
    - complètes ou non
    - approche statistique ou bayésienne

Acquisition de connaissances

- Détermination des probabilités conditionnelles sans données
  - Trouver un expert fiable et coopératif,
  - Le familiariser à la notion de probabilité,
  - Tenir compte des biais éventuels parfois subconscients (un expert va souvent sur-estimer la probabilité de réussite d'un projet le concernant, etc ...).
  - Lui fournir un outil pour déterminer les probabilités : échelle de probabilité (Drusdel 2001)



Acquisition de connaissances

- Simplification d'un problème complexe :
  - $P(Y|X_1 \dots X_n)$ (variables binaires) = 2^n valeurs !
  - Supposons
    - que l'on puisse estimer $p_i = P(y|\bar{x}_1, \bar{x}_2, ..., x_i, ..., \bar{x}_n)$
    - qu'il n'y a pas d'effet mutuel des variables X_i sur Y.
  - Alors (modèle OU bruité, Pearl 1986)

    - si un X_i est vrai, alors Y est presque toujours vrai (avec la probabilité p_i)
    - si plusieurs X_i sont vrais, alors

$$P(y|\mathcal{X}) = 1 - \prod_{i/X_i \in \mathcal{X}_p} (1 - p_i)$$

    où $\mathcal{X}_p$ est l'ensemble des X_i vrais.

Acquisition de connaissances

- Modèle OU bruité
  - étendu au cas où Y peut être vrai sans qu'une seule des causes soit vraie (*leaky noisy-OR gate*) (Henrion 1989)
  - aux variables multivaluées (*generalized noisy-OR gate*)

  - Ce type de CPD s'intègre très facilement aux algorithmes d'inférence tels que message passing ou junction tree.
  - Il peut aussi être utilisé pour de l'apprentissage à partir de données...

Apprentissage (données complètes)

- A partir de données complètes / incomplètes
- Avec des données complètes $\mathcal{D}$,
  - Approche statistique = *max. de vraisemblance (MV)*

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

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

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

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

Apprentissage (données complètes)

- Avec des données complètes $\mathcal{D}$,
  - 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)

- Approche bayésienne, *max. à 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.

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

- mais avec des données incomplètes ? . . .

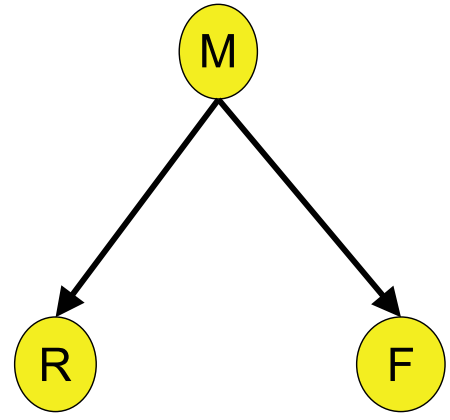
Exemple

- Données complètes (MAP)

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

$$\hat{P}(F = OK|M = m_0) = 1/6 = 0.17$$
$$\hat{P}(F = BAD|M = m_0) = 5/6 = 0.83$$

etc . . .



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)

- Algorithme EM (Expectation Maximisation)
  - (Dempster 1977)
  - Algo général d'estimation de paramètres avec des données incomplètes

  - Algorithme itératif
    - initialiser les paramètres $\theta^{(0)}$
    - 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
    - M ré-estimer les paramètres $\theta^{(t+1)}$ à partir des données complétées
      – en utilisant MV, MAP, ou EAP

Exemple

- Données manquantes (EM+MAP)
  - 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] :

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

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

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

Apprentissage de la structure

- Détermination de la structure d'un réseau bayésien
- Pas d'expert pour obtenir de structure (complète) du RB
- Apprentissage à partir de données
  - données complètes
  - données incomplètes
- Variables latentes ?

Point de départ

Recherche d'un bon réseau bayésien

- Recherche :
  - exhaustive = impossible / taille de l'espace.
    le nombre de structures possibles à partir de n nœuds est super-exponentiel (Robinson 77)

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

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

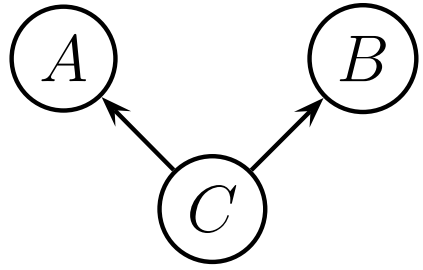
  - dans quel espace ?
    - $\mathcal{B}$ = espace des RB
    - $\mathcal{E}$ = espace des représentants des classes d'équivalence de Markov

Point de départ

Recherche d'un bon réseau bayésien

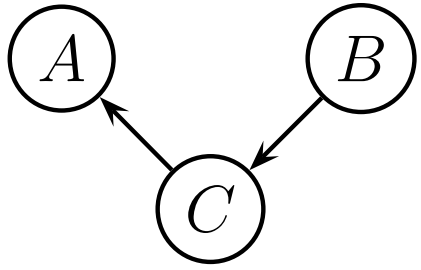
- Notion d'équivalence de Markov :

Deux structures B_1 et B_2 sont équivalentes si elles représentent la même distribution de probabilité jointe.



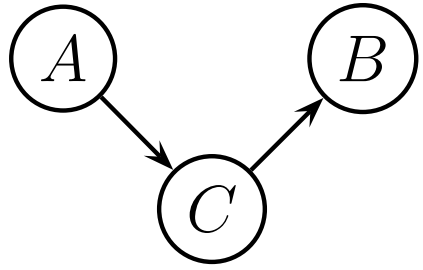
$P(A|C)P(B|C)P(C)$

$\propto$



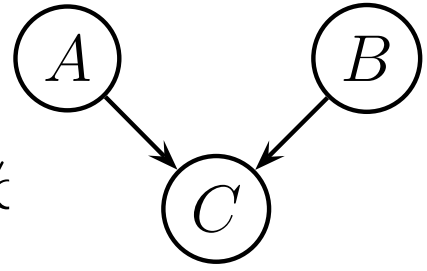
$= P(A|C)P(B)P(C|B)$

$\propto$



$= P(A)P(B|C)P(C|A)$

$\not\propto$

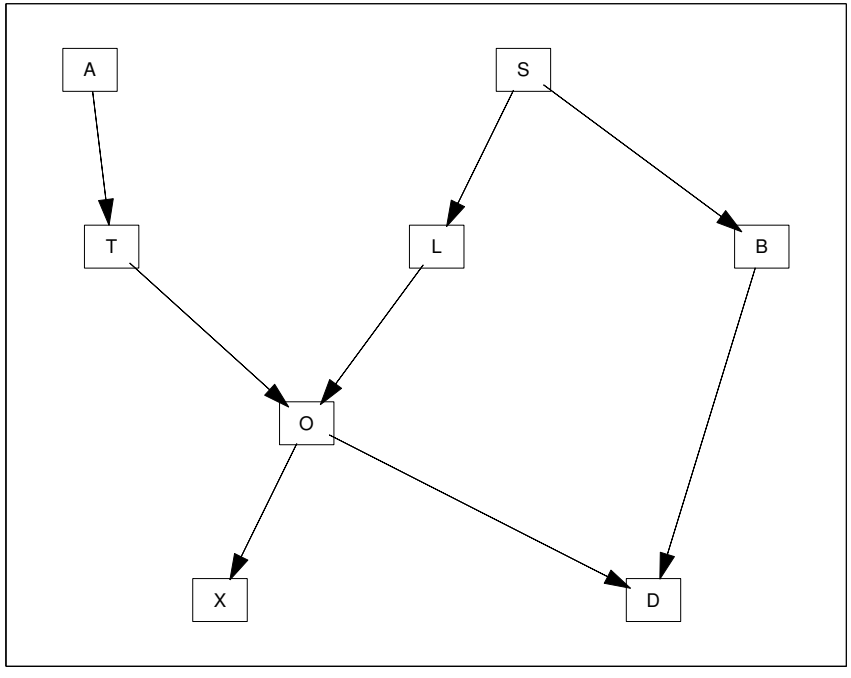


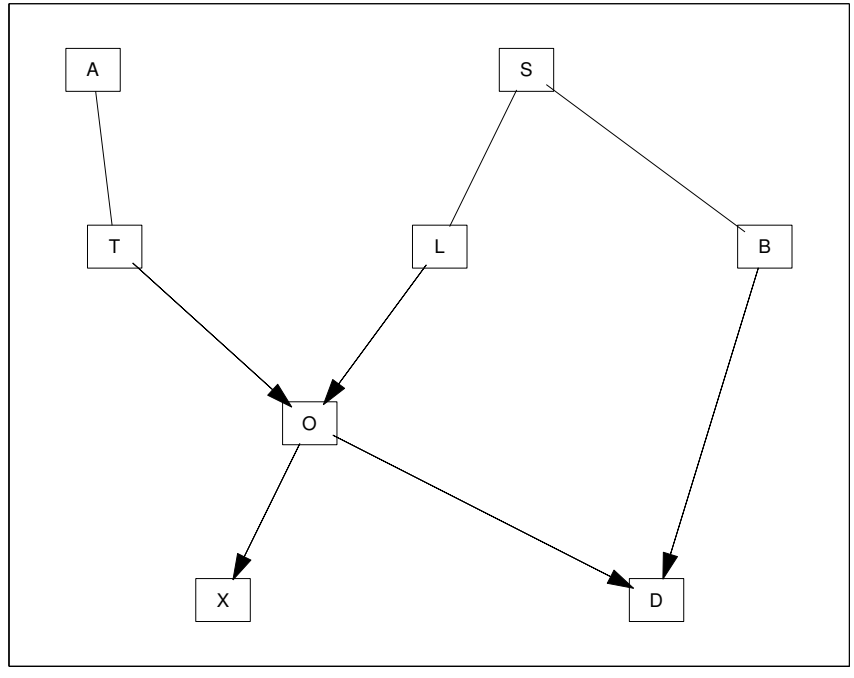
$P(A)P(B)P(C|A,B)$

Point de départ

- Notion d'équivalence de Markov (fin) :

On peut représenter les différentes structures équivalents par un graphe acyclique partiellement dirigé (PDAG).





Point de départ

Recherche d'un bon réseau bayésien

- Besoin d'un score pour "noter" chaque structure
  - calculable "rapidement"
  - décomposable localement
- Notion de *score équivalence*
  - Un score S est dit *score equivalent* ssi pour deux structures B_1 et B_2 équivalentes on a $S(B_1) = S(B_2)$.

Survol de différentes familles de méthodes

- Recherche de causalité (IC/**PC**, IC*/FCI)

- Recherche par calcul de score
  - quelques scores :
    - entropie, AIC, BIC, MDL
    - BD, BDe, BDeu

  - parcours de recherche :
    - espace $\mathcal{B}$
      - restriction aux arbres : Chow&Liu, **MWST**
      - ordonnancement des nœuds : **K2**
      - heuristique : **Greedy Search**
      - données incomplètes : **Structural EM**

    - espace $\mathcal{E}$
      - **Greedy Equivalence Search**

Survol de différentes familles de méthodes

- Recherche de causalité (IC/**PC**, IC*/FCI)

- Recherche par calcul de score
  - quelques scores :
    - entropie, AIC, BIC, MDL
    - BD, BDe, BDeu
  - parcours de recherche :
    - espace $\mathcal{B}$
      - restriction aux arbres : Chow&Liu, **MWST**
      - ordonnancement des nœuds : **K2**
      - heuristique : **Greedy Search**
      - données incomplètes : **Structural EM**
    - espace $\mathcal{E}$
      - **Greedy Equivalence Search**

Causalité

- Séries d'algos proposés en "parallèle" :
  - Pearl et Verma : IC et IC*
  - Spirtes, Glymour et Scheines : SGS, PC, CI, FCI

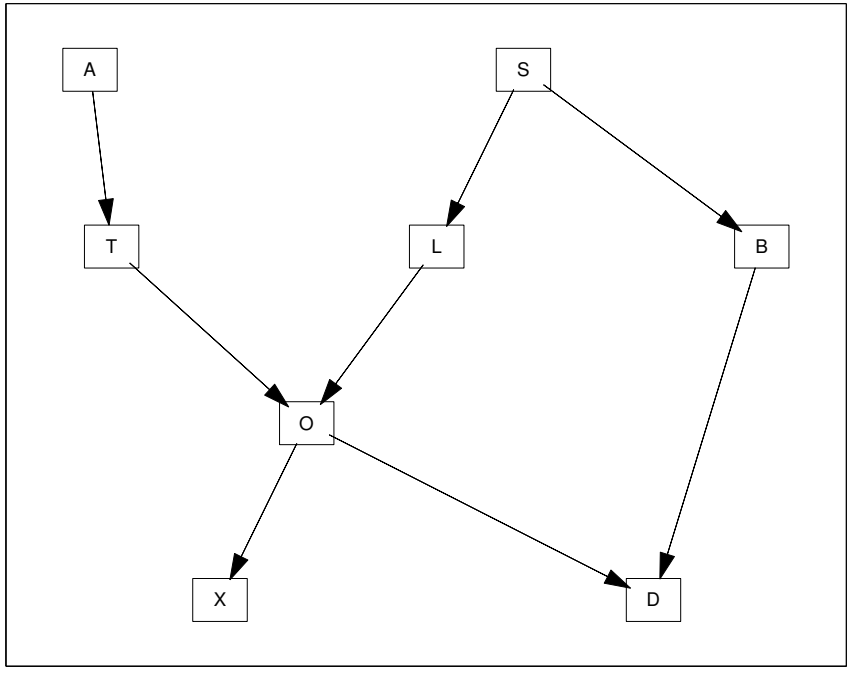
- Principe commun :
  - construire un graphe non dirigé contenant les relations entre les variables (tests d'indépendance conditionnelle du χ^2)
    - par ajout d'arêtes (Pearl et Verma)
    - par suppression d'arêtes (SGS)
  - détecter les V-structures (idem)
  - "propager" les orientations de certains arcs, etc ...
  - détecter des variables latentes (IC*, CI, FCI)

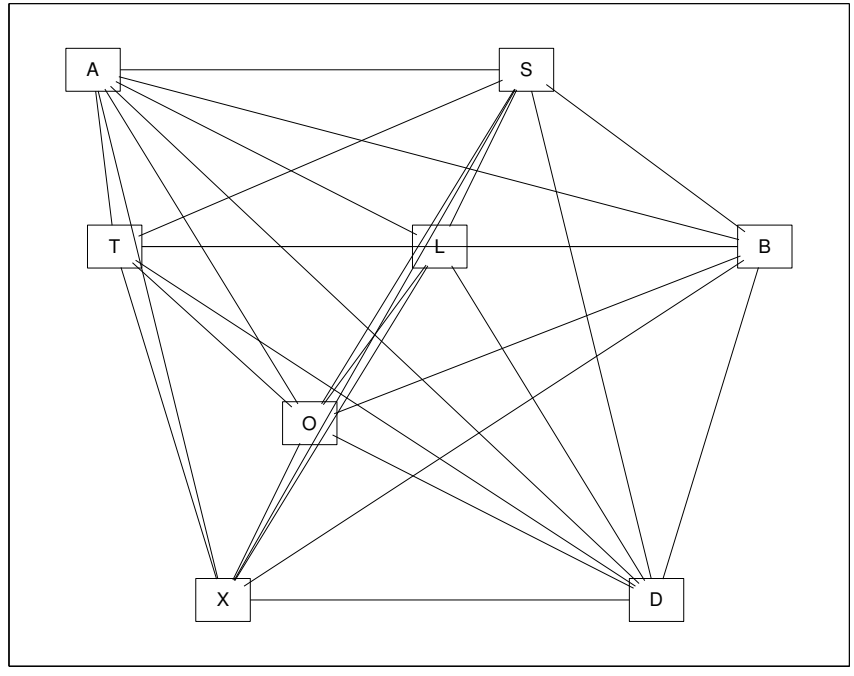
Causalité

- 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.

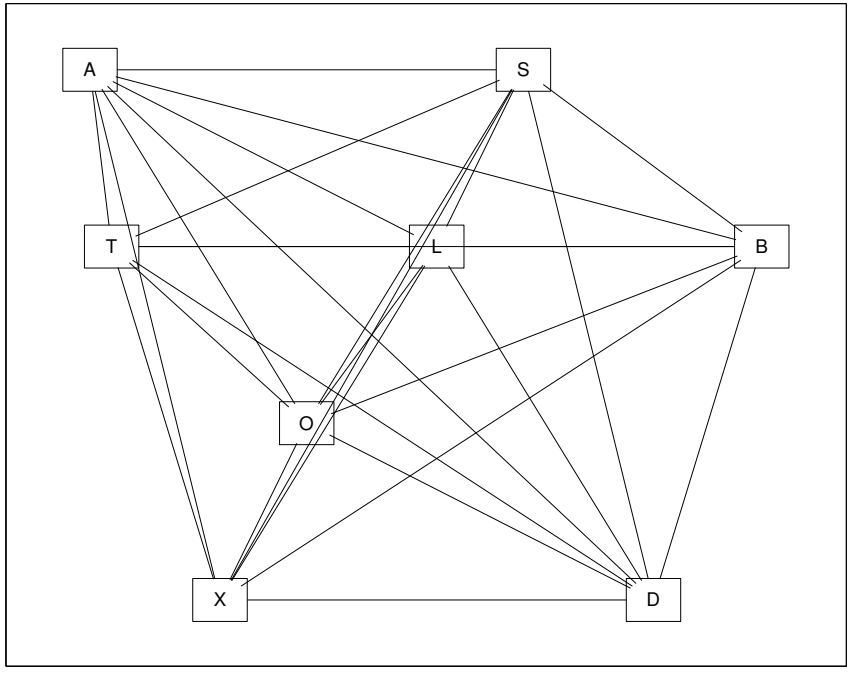


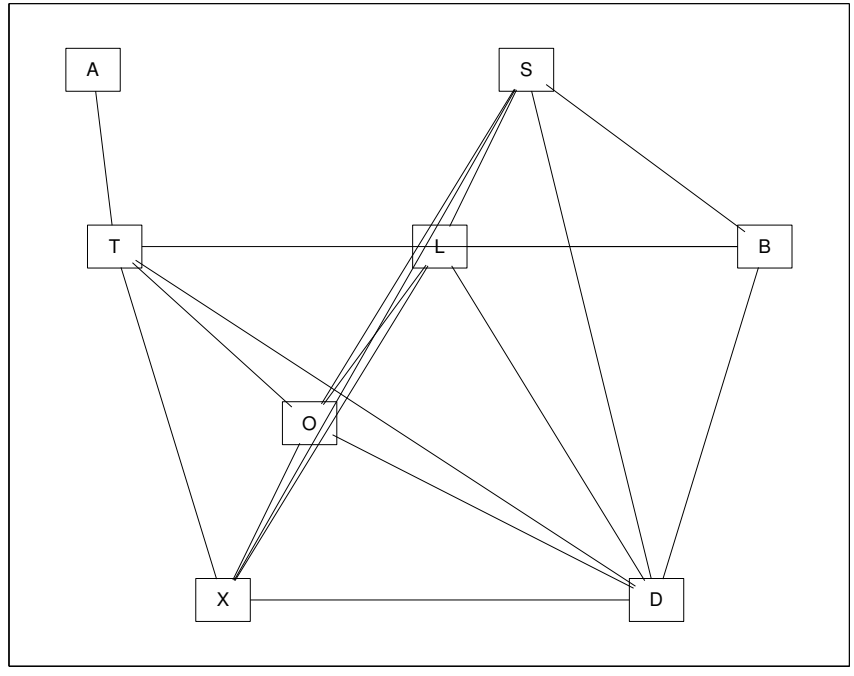


A gauche, le réseau "théorique" utilisé pour générer 5000 exemples.

Algorithme PC

Etape 1a : Suppression des ind. conditionnelles d'ordre 0



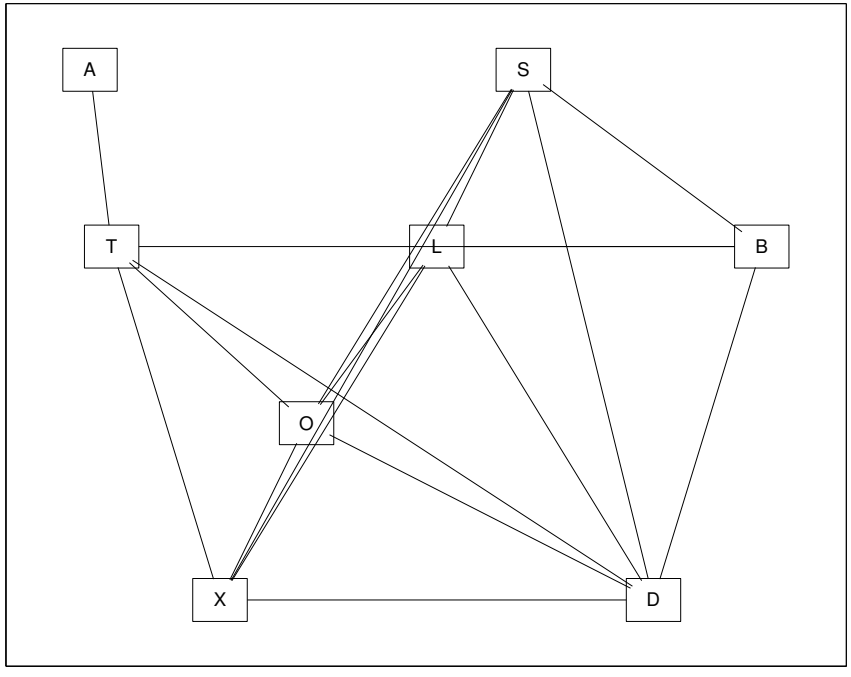


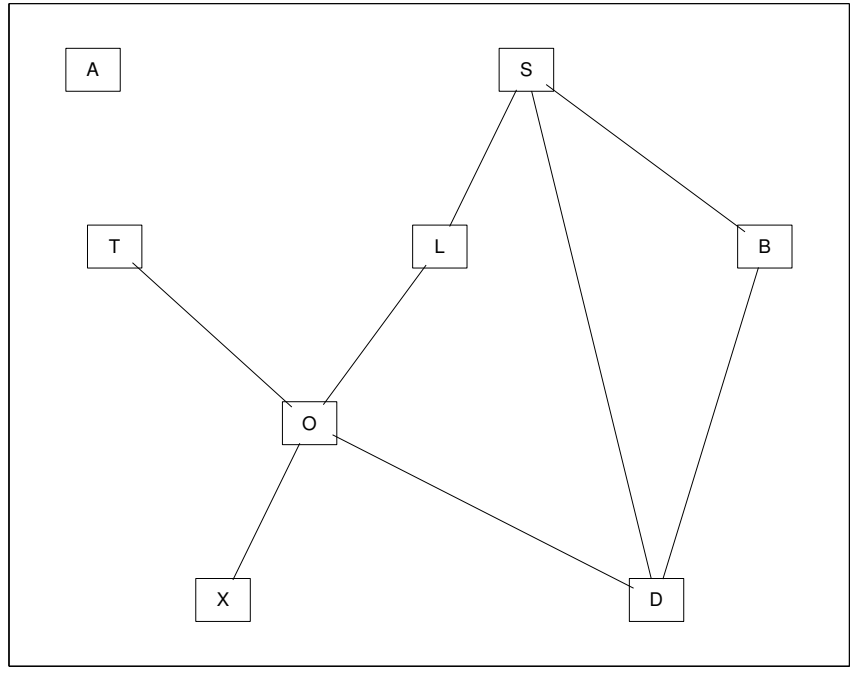
Test du χ^2 sur nos données :

$S \perp A$	$L \perp A$	$B \perp A$	$O \perp A$	$X \perp A$	$D \perp A$
$T \perp S$	$L \perp T$	$O \perp B$	$X \perp B$		

Algorithme PC

Etape 1b : Suppression des ind. conditionnelles d'ordre 1



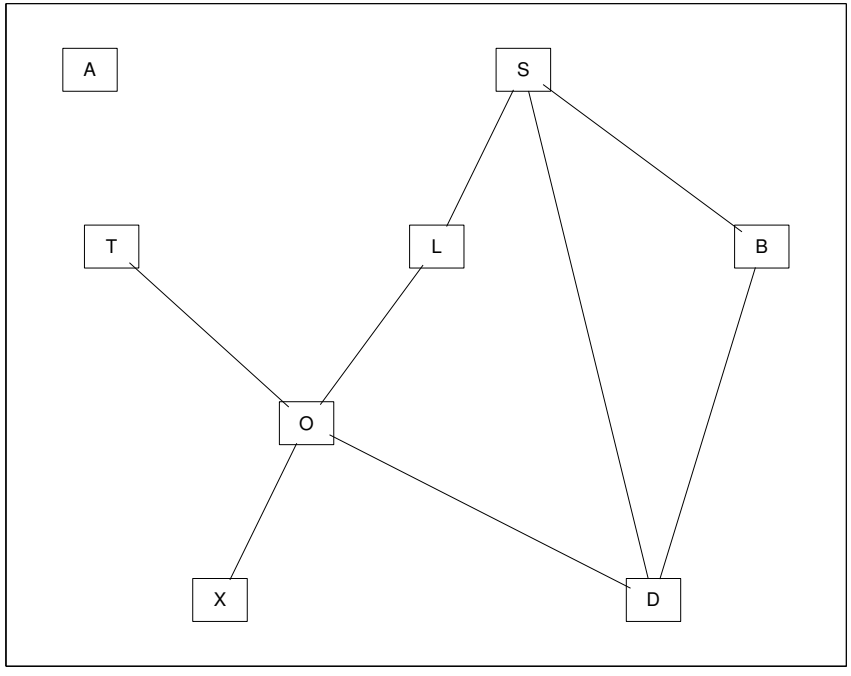


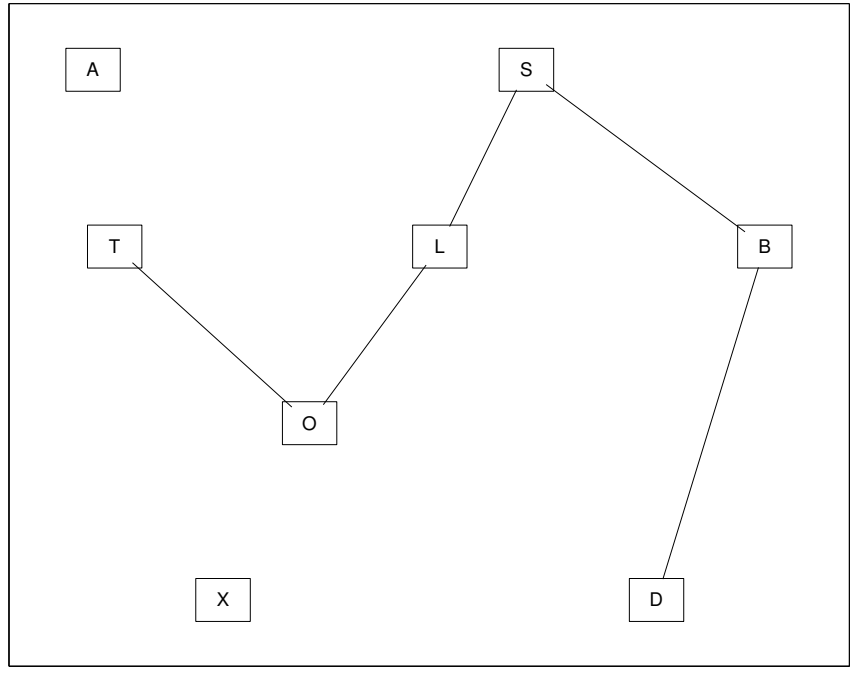
Test du χ^2 sur nos données :

$T \perp A \mid O$	$O \perp S \mid L$	$X \perp S \mid L$	$B \perp T \mid S$
$X \perp T \mid O$	$D \perp T \mid O$	$B \perp L \mid S$	$X \perp L \mid O$
$D \perp L \mid O$	$D \perp X \mid O$		

Algorithme PC

Etape 1c : Suppression des ind. conditionnelles d'ordre 2



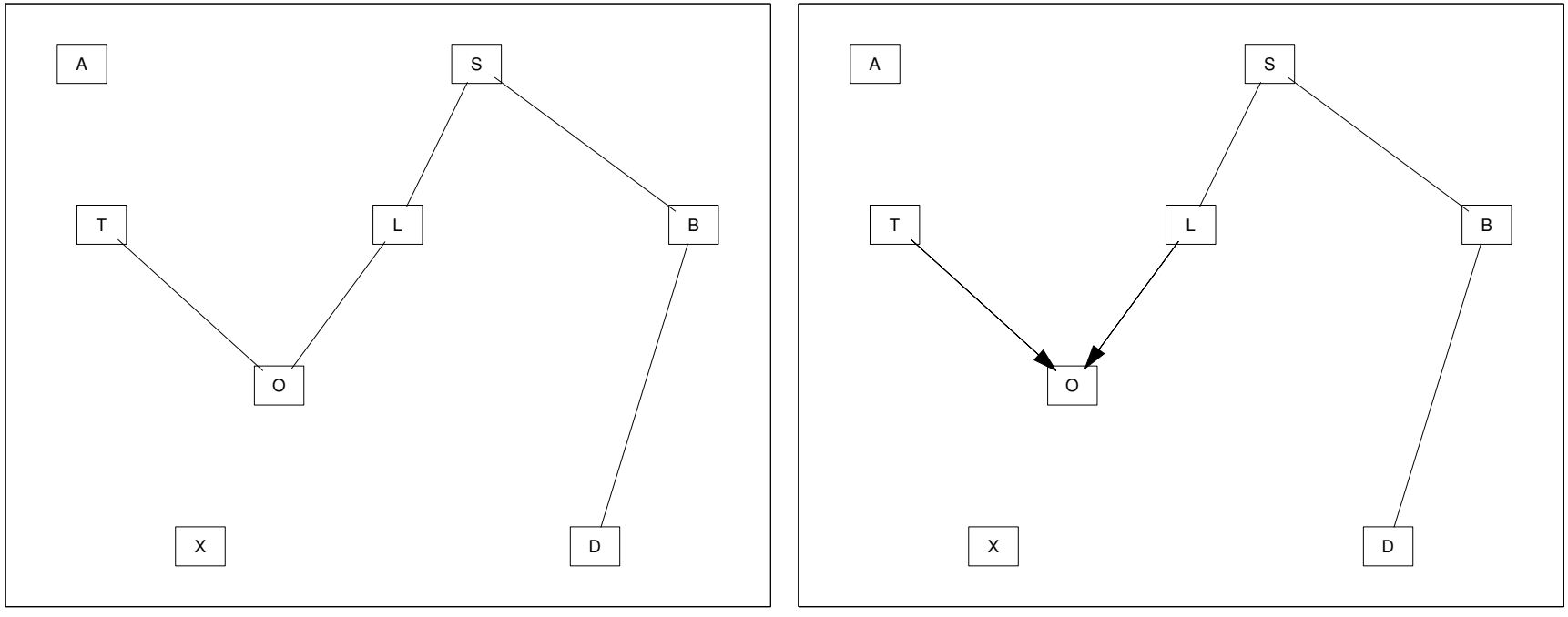


Test du χ^2 sur nos données :

$$D \perp S | \{L, B\} \quad X \perp O | \{T, L\} \quad D \perp O | \{T, L\}$$

Algorithme PC

Etape 2 : Recherche des V-structures



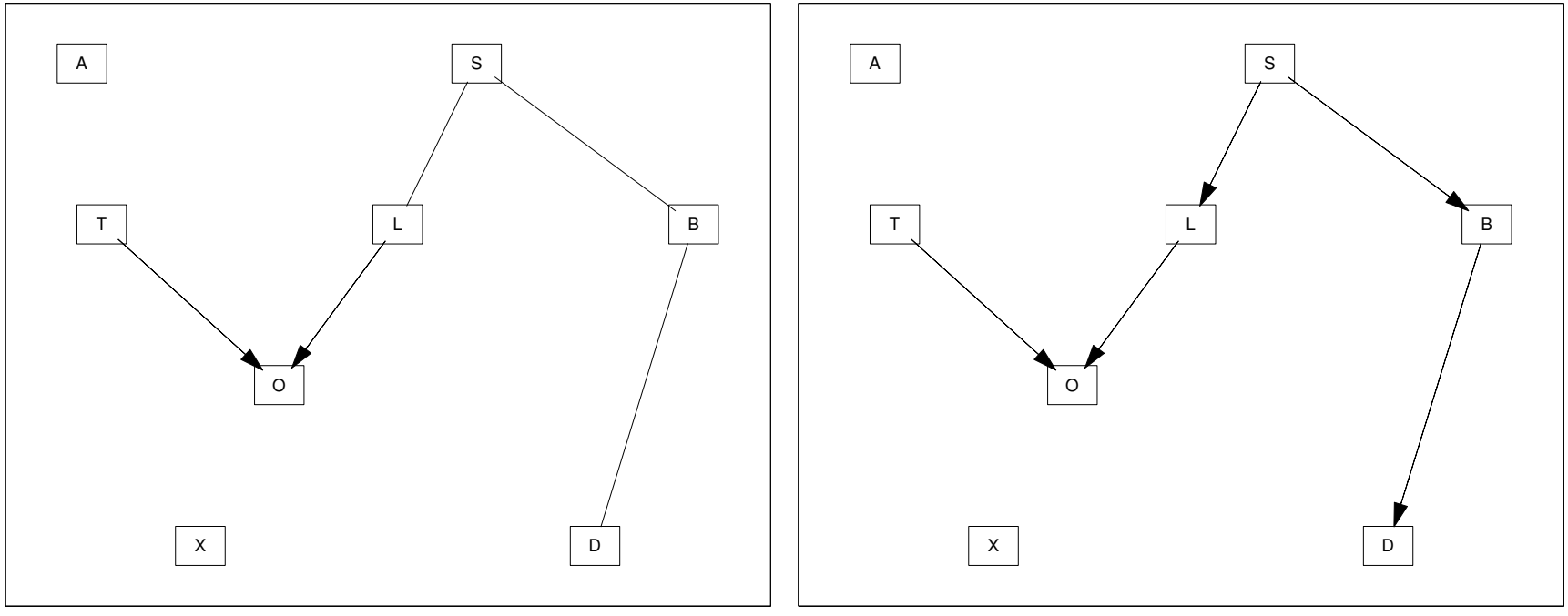
Test du χ^2 sur nos données : découverte de la V-structure

$$T \rightarrow O \leftarrow L$$

\+ Etape 3 : Orientation récursive de certaines arêtes (aucune ici)

Algorithme PC

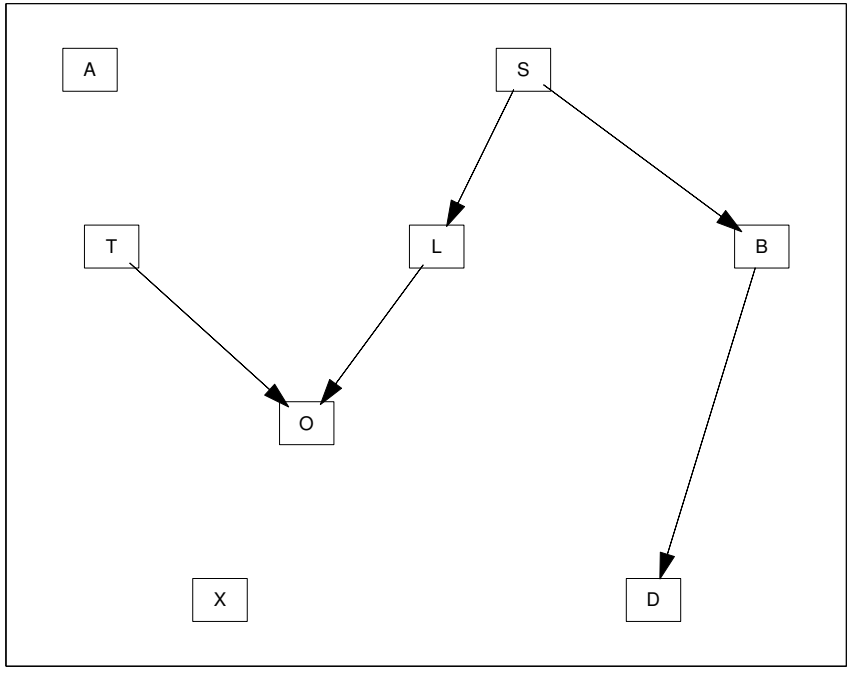
Etape 4 : Instanciation du PDAG

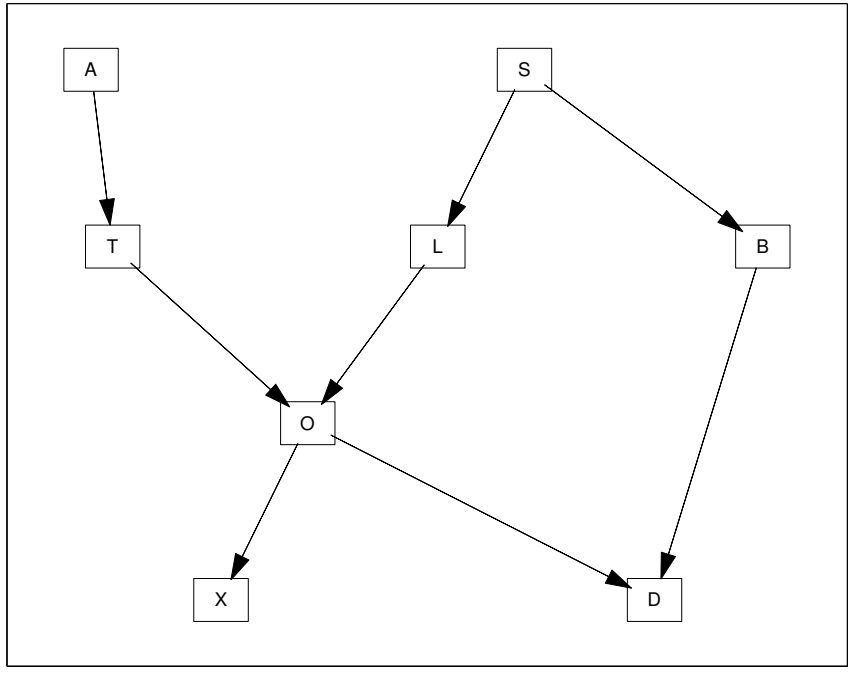


Orientation des arcs restants
(seule condition : ne pas introduire de nouvelle V-structure)

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 $O \rightarrow D$.

Survol de différentes familles de méthodes

- Recherche de causalité (IC/**PC**, IC*/FCI)

- Recherche par calcul de score
  - quelques scores :
    - entropie, AIC, BIC, MDL
    - BD, BDe, BDeu

- parcours de recherche :
  - espace $\mathcal{B}$
    - restriction aux arbres : Chow&Liu, **MWST**
    - ordonnancement des nœuds : **K2**
    - heuristique : **Greedy Search**
    - données incomplètes : **Structural EM**
  - espace $\mathcal{E}$
    - **Greedy Equivalence Search**

Score

- Principe du rasoir d'Occam : trouver le modèle
  - qui colle le mieux aux données $\mathcal{D}$: vraisemblance $L(\mathcal{D}|\theta, B)$
  - et qui soit le plus simple possible : $Dim(B)$ = nb de paramètres pour décrire B

- Problèmes non traités ici
  - Estimation de L avec des données manquantes
  - Calcul de $Dim(B)$ avec des variables latentes

Quelques scores : Entropie conditionnelle

$$H(B,\mathcal{D}) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} -\frac{N_{i,j,k}}{N}\ \log(\frac{N_{i,j,k}}{N_{i,j}})$$

- lien entre l'entropie et le maximum de la log-vraisemblance :

$$\log L(\mathcal{D}|\theta^{MV}, B) = -N \times H(B,\mathcal{D})$$

Problème avec la vraisemblance (ou l'entropie) = pas de contrôle sur la complexité de la structure ...

Quelques scores : application de AIC et BIC

- Tenir compte de la vraisemblance maximale et de la complexité du modèle
- Application aux critères AIC (Akaïke 70) et BIC (Schwartz 78)

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

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

Quelques scores : MDL

- Application du principe MDL
  (Minimum Description Length, Rissanen 78)
- Minimiser la somme de :
  (1) la longueur de codage du modèle
  (2) la longueur de codage des données lorsqu'on utilise le modèle pour les représenter.
- Plusieurs approches proposées (Bouckaert 93, **Lam et Bacchus 93**, Suzuki 99)

$$S_{MDL}(B, \mathcal{D}) = \log L(\mathcal{D}|\theta^{MV}, B) - |A_B| \log N - c.Dim(B)$$

- $|A_B|$ est le nombre d'arcs dans le graphe B
- c est le nombre de bits utilisés pour stocker chaque paramètre numérique

Quelques scores : Bayesian Dirichlet

- (Cooper et Herskovits 92) proposent une approche bayésienne :

$$\begin{aligned} S_{BD}(B,\mathcal{D}) &= P(B,\mathcal{D}) = \int_\theta L(\mathcal{D}|\theta,B)P(\theta|B)p(B)\,d\theta \\ &= P(B)\int_\theta L(\mathcal{D}|\theta,B)P(\theta|B)\,d\theta \end{aligned}$$

En faisant les hypothèses classiques d'indépendance des exemples, en prenant une distribution a priori de Dirichlet sur les paramètres :

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

Quelques scores : BDe

- (Heckerman 94) : BD + *score équivalence*
- Distribution a priori des paramètres :

$$\alpha_{ijk} = N' \times P(X_i = x_k, pa(X_i) = x_j | B_c)$$

où B_c est un graphe complétement relié (n'encodant aucune indépendance conditionnelle) et N' est un nombre d'exemples "équivalent" défini par l'utilisateur.

- si chaque instance de X_i est équiprobable (cond. à B_c), on obtient :

$$\alpha_{ijk} = \frac{N'}{r_i q_i}$$

- le score BDe utilisant ces α_{ijk} est souvent appelé score BDeu.

Survol de différentes familles de méthodes

- Recherche de causalité (IC/**PC**, IC*/FCI)

- Recherche par calcul de score
  - quelques scores :
    - entropie, AIC, BIC, MDL
    - BD, BDe, BDeu

  - parcours de recherche :
    - espace $\mathcal{B}$
      - restriction aux arbres : Chow&Liu, **MWST**
      - ordonnancement des nœuds : **K2**
      - heuristique : **Greedy Search**
      - données incomplètes : **Structural EM**

    - espace $\mathcal{E}$
      - **Greedy Equivalence Search**

Recherche dans l'espace des RB

- Décomposabilité du score :

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

- Avantages :
  - utilisation de méthodes incrémentales
  - diminution du nb de calculs dans l'évaluation des scores

Restriction du parcours de recherche

- Restriction à l'espace des arbres :
  - quel est le meilleur arbre passant par tous les nœuds (i.e. maximisant un score défini pour chaque arc possible ?

- 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 :

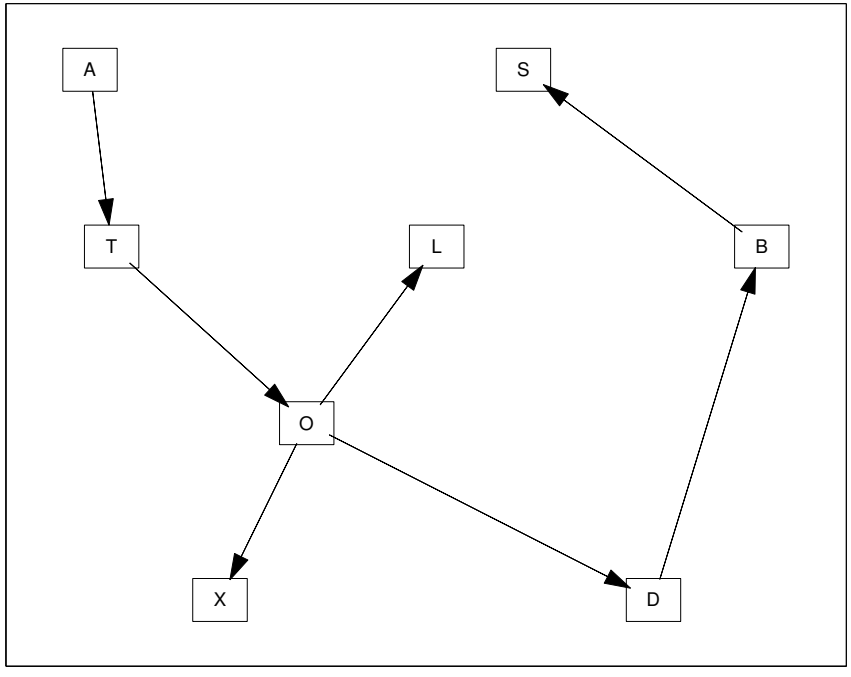
$$W(X_A, X_B) = score(X_A, Pa(X_A) = X_B) - score(X_A, \emptyset)$$

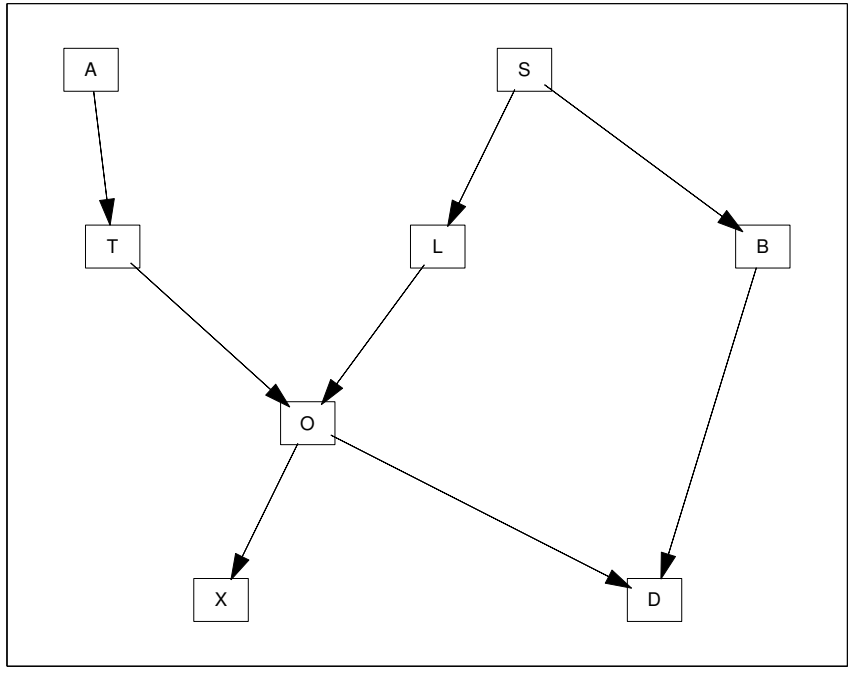
Restriction du parcours de recherche

- Restriction à l'espace des arbres :
  - MWST donne un arbre non orienté reliant toutes les variables.
  - arbre non orienté = 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.

Algorithme MWST

Réseau obtenu vs. théorique





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

Restriction du parcours de recherche

- Ordonnancement des nœuds :
  - Hypothèse : si X_i est avant X_j alors il ne pourra y avoir d'arc de X_j vers X_i.

$\Rightarrow$ Réduction du nb de structures possibles de $NS(n)$ à $NS'(n) = 2^{n(n-1)/2}$.

  - Par exemple, $NS'(5) = 1024$ contre $NS(5) = 29281$ et $NS'(10) = 3.5 \times 10^{13}$ contre $NS(10) = 4.2 \times 10^{18}$.

Restriction du parcours de recherche

- Ordonnancement des nœuds :
  - Algorithme K2 (Cooper et Herskovits 92) : score *BD* avec un apriori uniforme sur les structures :

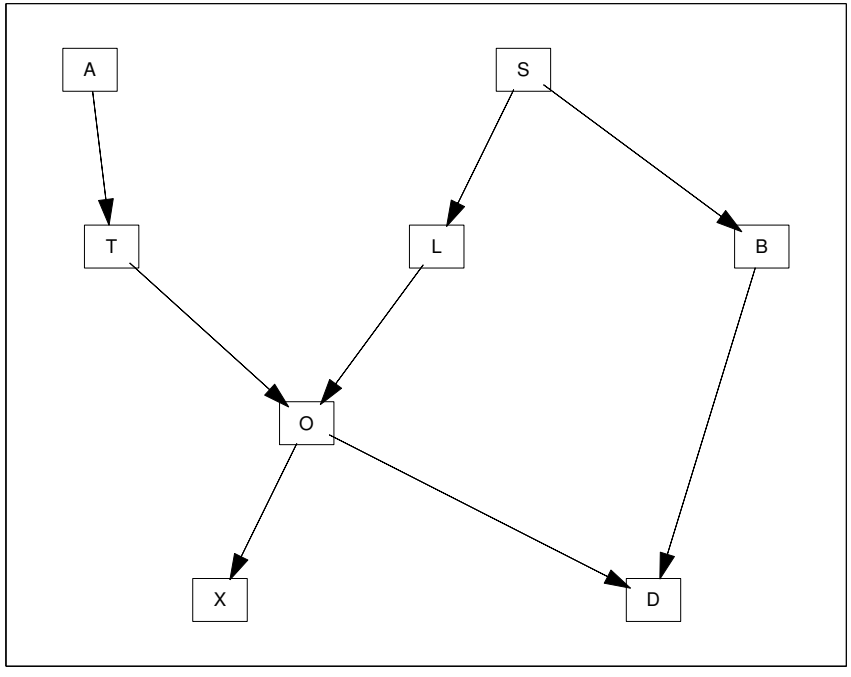
$$S_{BD}(B,\mathcal{D}) \propto \prod_{i=1}^{n}\left(\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij}+\alpha_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N_{ijk}+\alpha_{ijk})}{\Gamma(\alpha_{ijk})}\right)$$

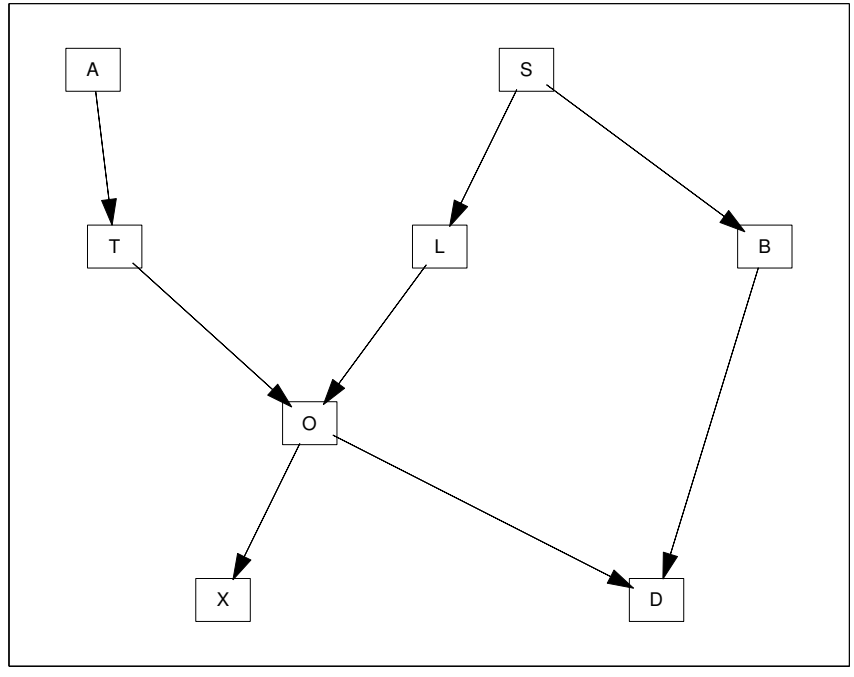
$$\propto \prod_{i=1}^{n} g(X_i, pa_i)$$

  - Maximisation de S_{BD} par recherche gloutonne en cherchant les parents pa_i du nœud X_i qui vont maximiser $g(X_i, pa_i)$, et ainsi de suite...
  - Ajout d'une borne supérieure u au nombre de parents possibles pour un nœud.

Algorithme K2

Réseau obtenu vs. théorique

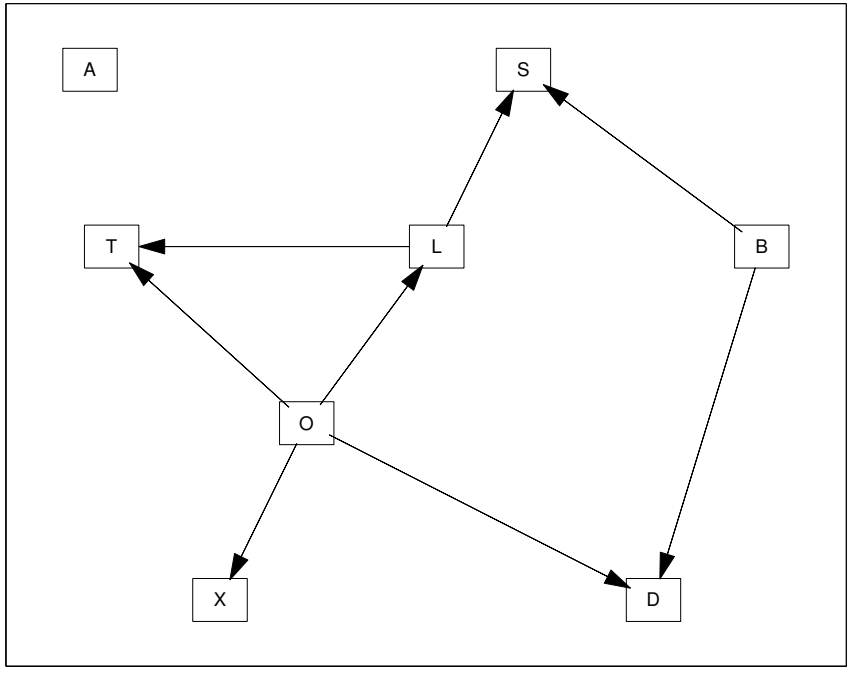


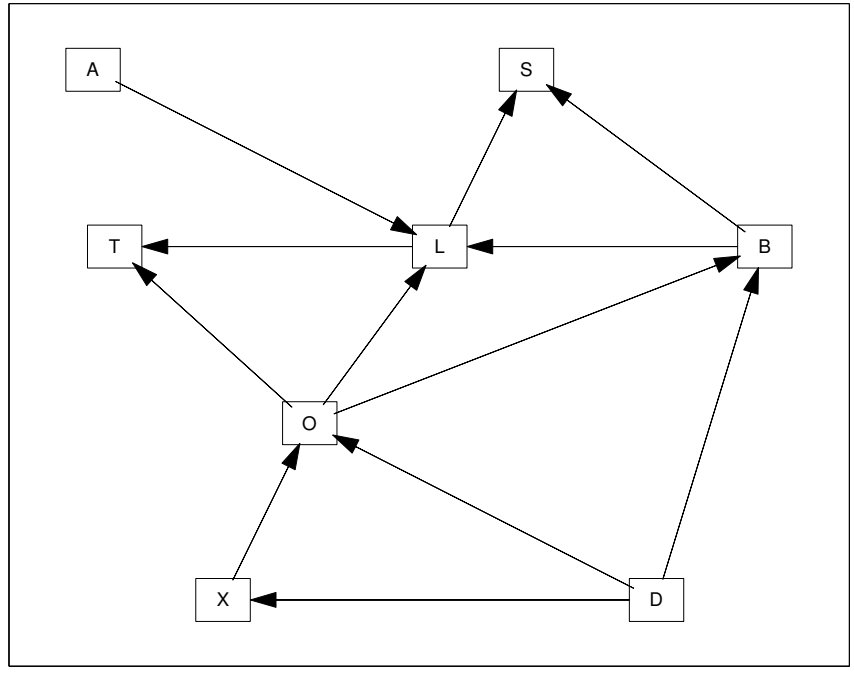


En donnant l'ordre "théorique" et les 5000 exemples, ca converge bien ... mais ...

Algorithme K2

Situation réaliste : initial^{o} aléatoire de l'ordre de parcours





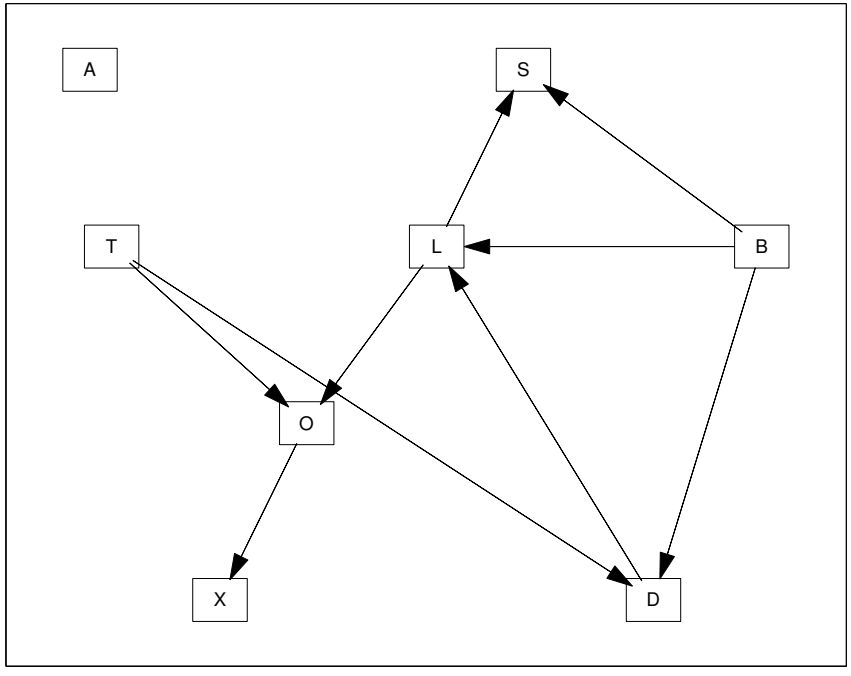
convergence vers des minima locaux ...

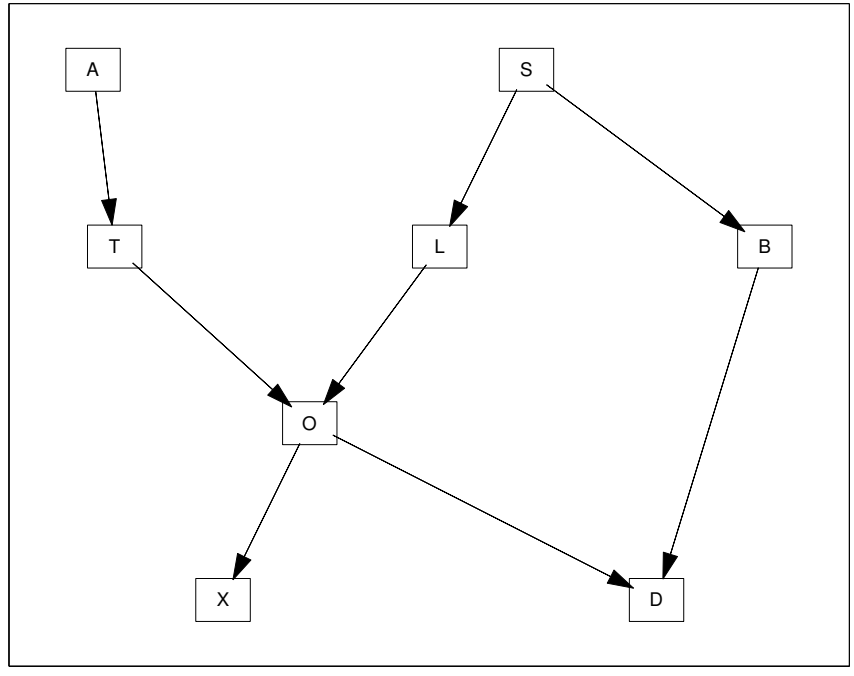
Restriction du parcours de recherche

- Parcours de type *greedy search* :
  - opérateurs classiques :
    - ajout d'arc
    - inversion d'arc
    - suppression d'arc
  - sous reserve que le graphe obtenu soit toujours un DAG (pas de cycle)
  - possibilité de commencer à partir d'un graphe précis

Algorithme Greedy Search

Réseau obtenu vs. théorique

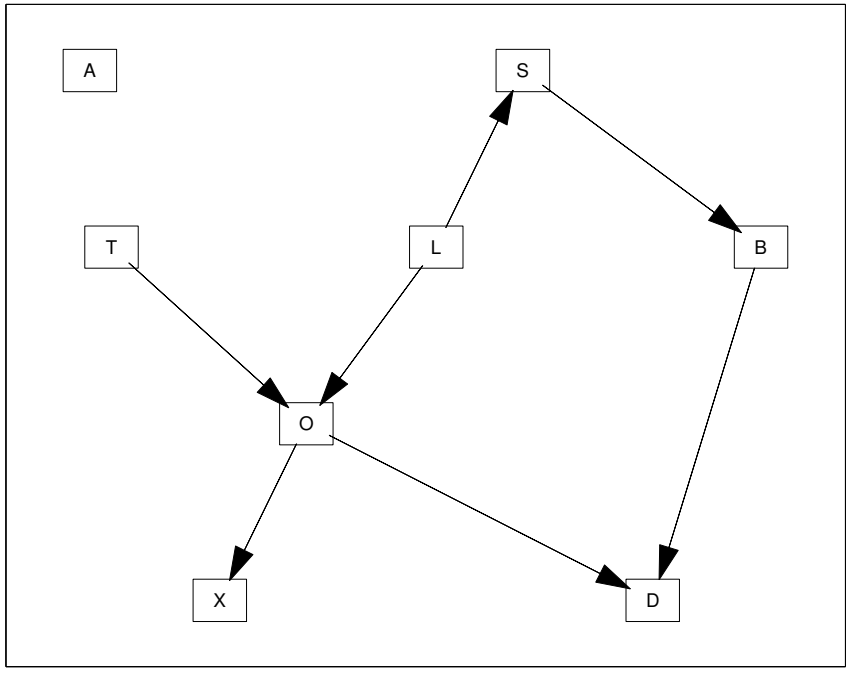


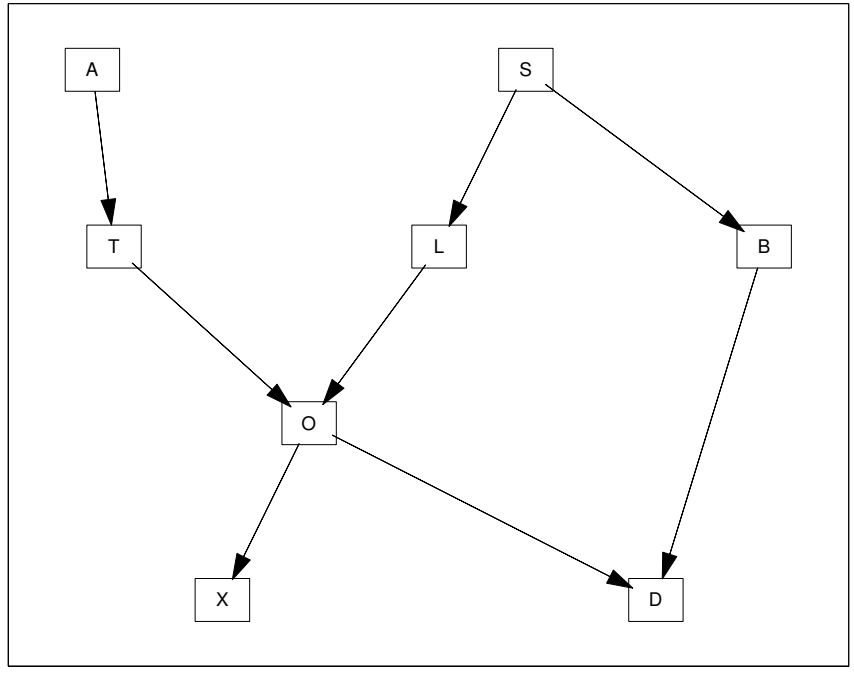


On tombe surement dans un optimum local

Algorithme Greedy Search

Réseau obtenu vs. théorique





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

Prise en compte des données incomplètes

- Problème = calcul du score lorsque les données sont incomplètes $\mathcal{X} = \{\mathcal{D}, \mathcal{H}\}$
- Algorithme Structural EM (Friedman 97)
  - Score classique = $\log L(\mathcal{D}|\theta, B) - \text{pen}(\mathcal{B}, \theta, D)$
  - Pour une distribution $P(\mathcal{H}|\mathcal{D})$ estimée à partir de θ', B'

$$S_{EM}(\theta, B|\theta', B') = E_{\mathcal{H}}[\log L(\mathcal{D}, \mathcal{H}|\theta, B) - \text{pen}(B, \theta, \mathcal{D})]$$

  - Algorithme itératif : détail d'une itération :
    - estimer $P(\mathcal{H}|\mathcal{D})$ à partir de $\{\theta^{(i)}, B^{(i)}\}$
    - trouver $B^{(i+1)}$ qui maximise $S_{EM}(\theta, B|\theta^{(i)}, B^{(i)})$
    - trouver $\theta^{(i+1)}$ qui maximise $S_{EM}(\theta, B^{(i+1)}|\theta^{(i)}, B^{(i)})$

Algorithme Structural EM

- En pratique :
  - inutile de maximiser S_{EM}, il suffit de trouver un meilleur $B^{(i+1)}$ et pas forcément le meilleur dans $\mathcal{B}$.

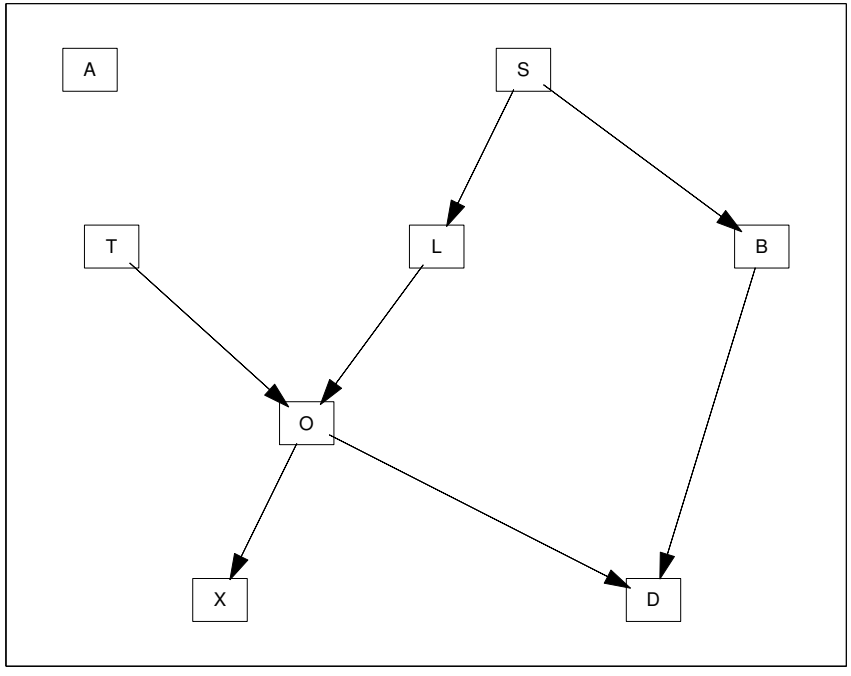
  $\Rightarrow$ recherche de $B^{(i+1)}$ parmi les voisins de $B^{(i)}$

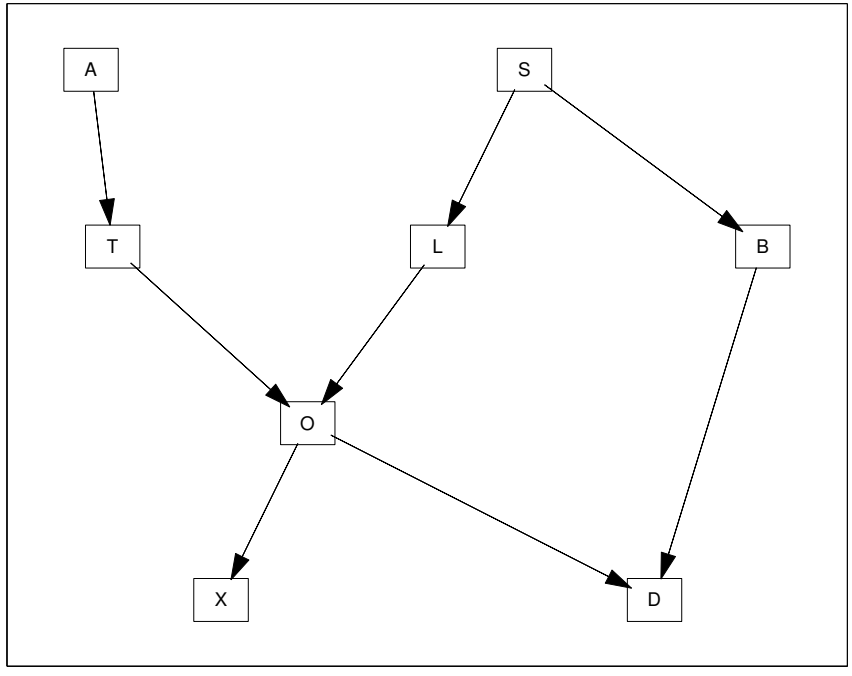
  Structural EM $\simeq$ *Greedy Search* + *EM* sur les paramètres

  - variante "alternante"
    - Faire plusieurs itérations de l'algo EM paramètrique pour améliorer $\theta^{(i)}$ avant de chercher $B^{(i+1)}$
  - variante avec un score bayésien (*Bayesian Structural EM*)

Algorithme Structural EM

Réseau obtenu vs. théorique



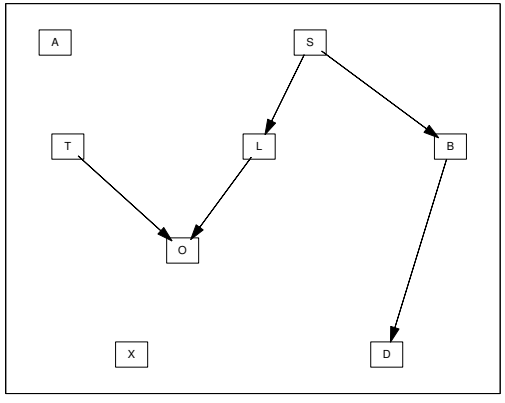


avec nos 5000 exemples (en enlevant aléatoirement 20% des données), ca converge bien (sauf $A \rightarrow T$) ...
mais c'est lent...

Récapitulatif

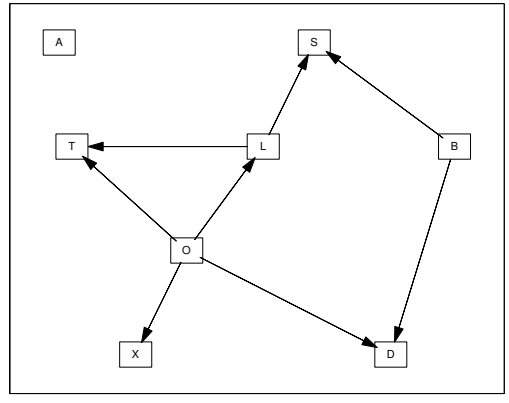
PC

pb test statistique



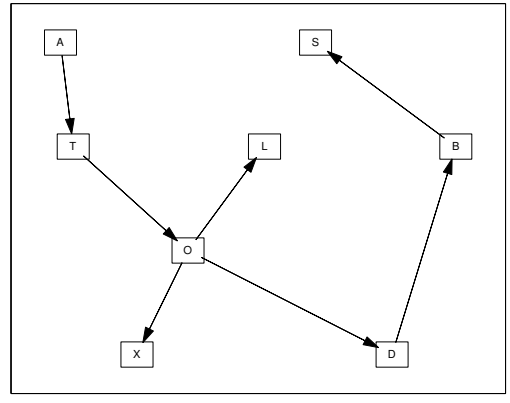
K2 (ordre aléatoire)

pb initialisation



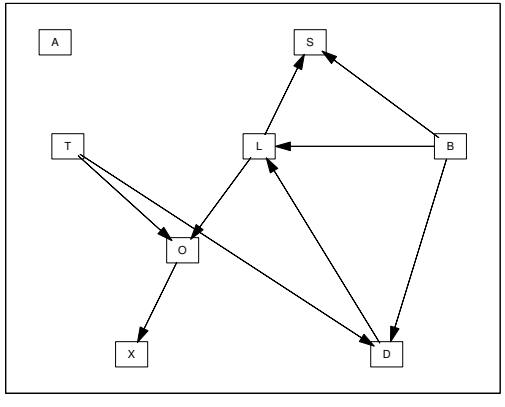
MSWT

rapide mais arbre



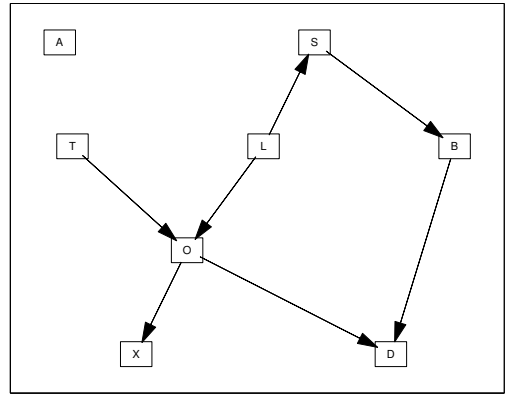
Greedy Search

lent et optimum local



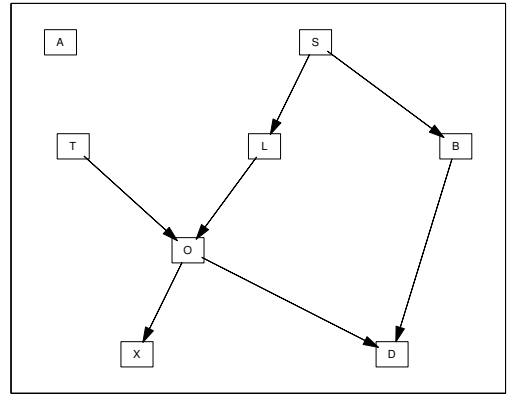
Greedy Search+MWST

lent



SEM

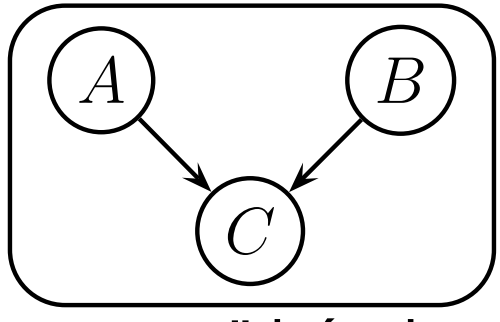
pour données manquantes



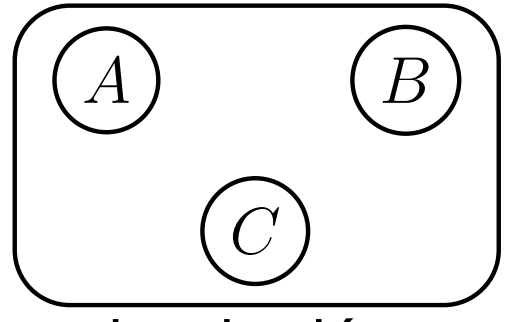
Et si on changeait d'espace de recherche ...

- 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

- Exemple (Munteanu 2001) : au lieu de retrouver la V-structure initiale, un algorithme de type Greedy Search peut converger vers un optimum local.



structure "théorique"

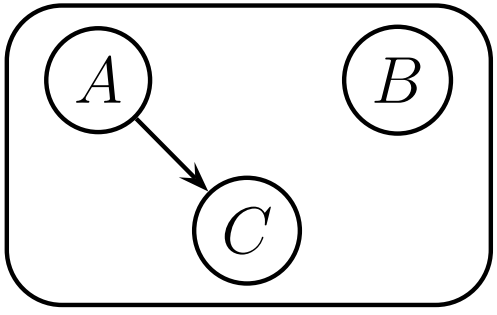


point de départ

Et si on changeait d'espace de recherche ...

- 2 scores maximums (structures équivalentes) :

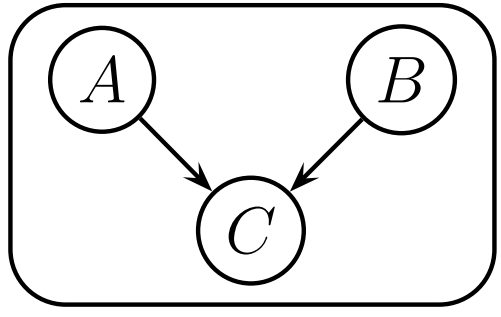
$$score(A \rightarrow C) = score(C \leftarrow A)$$



score maximal :

$$score(B \rightarrow C)$$

$\Downarrow$

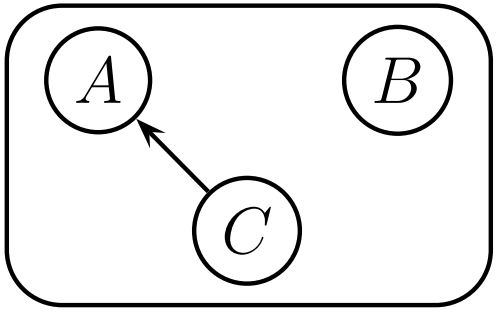


optimum global

Et si on changeait d'espace de recherche ...

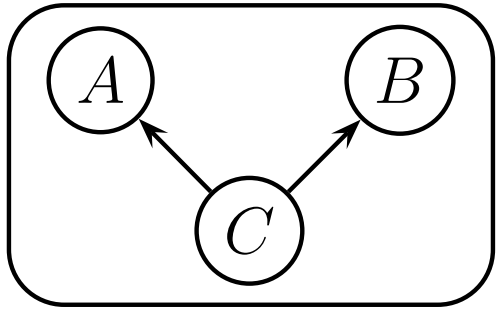
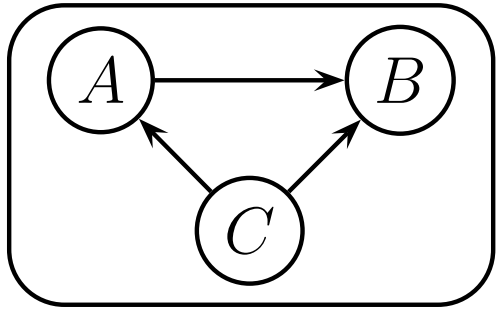
- 2 scores maximums (structures équivalentes) :

$$score(A \rightarrow C) = \boxed{score(C \leftarrow A)}$$



2 scores maximums
(structures équivalentes) :

$$score(B \rightarrow C) = \boxed{score(C \rightarrow B)}$$

$\Longrightarrow$  $\cdots$ etc $\cdots \Longrightarrow$ 

optimum local

Recherche dans $\mathcal{E}$

- $\mathcal{E}$ = espace des représentants des classes d'équivalence de Markov
- Meilleures propriétés : OUI
  - 2 structures équivalentes (de même score) = 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)

Recherche dans $\mathcal{E}$

- Deux situations :
  - travailler comme avant dans $\mathcal{B}$
    - on "bride" les opérateurs pour qu'ils ne proposent que des "bons" RB (non équivalents) (Castello et Kocka 2002)
  - travailler directement dans $\mathcal{E}$
    - (Chickering 2000) : algos lourds, trop d'opérateurs. . .
    - (Auvray 2002)
    - (Chickering 2002) : **Greedy Equivalence Search** algo optimal avec une série d'ajout d'arcs puis une série de retraits
- et pourquoi ne pas imaginer un S2EM (Structural Equivalent EM) ? ? ?

Variables latentes

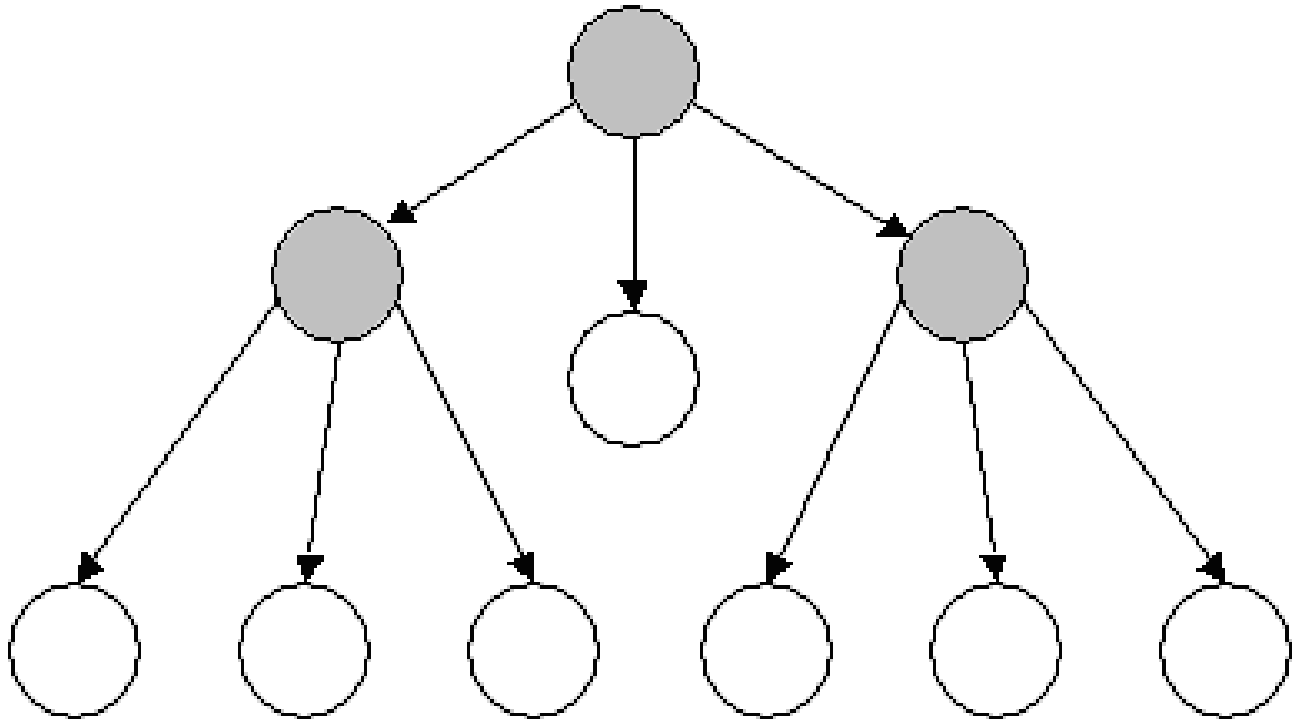
- Toutes les méthodes présentées précédemment font la même hypothèse :
  les variables du problème sont toutes à notre disposition

- Cette hypothèse est souvent fausse
- L'incorporation de variables latentes dans le modèle donne souvent de meilleurs résultats

- Comment trouver des variables latentes ?

Modèles avec variables latentes

- La connaissance du problème à résoudre peut déterminer la position des variables latentes
- Exemple pour des problèmes de clustering : Hierarchical Latent Class Models



Découverte de variables latentes

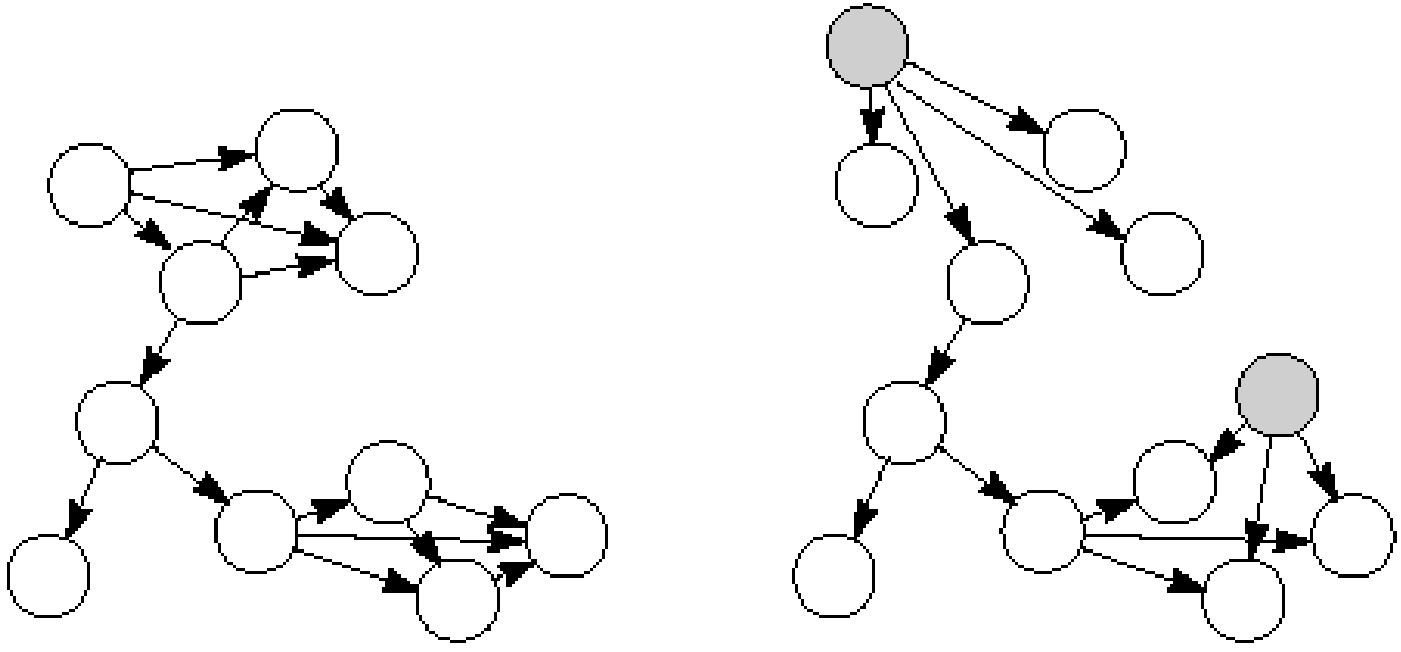
- On peut aussi essayer de découvrir la présence de variables latentes à partir des données
  - Extension des méthodes de recherche de causalité
  - Heuristiques

Causalité et variables latentes

- Algorithmes IC*, CI, FCI
- Principe :
  - Notations :
    - $A \leftrightarrow B$: il existe une variable latente $A \leftarrow L \rightarrow B$
    - $A \mapsto B$: $A \rightarrow B$ ou $A \leftrightarrow B$
    - $A - B$: $A \rightarrow B$ ou $A \leftarrow B$ ou $A \leftrightarrow B$
  - Algorithme :
    - les V structures découvertes en (2) sont marquées par des $\mapsto$ au lieu de $\rightarrow$
    - lors de l'étape (3), certaines règles permettent de rajouter des flèches et de lever des incertitudes (cf. Pearl 2000)

Recherche heuristique de variables latentes

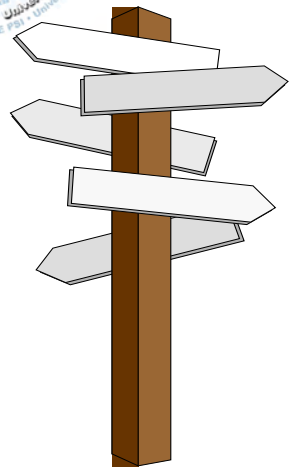
- (Martin et Vanlehn 1995)
  si un ensemble de variables sont toutes mutuellement dépendantes, alors cela peut signifier que ces variables possèdent en commun une unique cause cachée qui les rend mutuellement indépendantes



Cardinalité des variables latentes

- Déterminer la position des variables latentes dans le graphe n'est pas suffisant
- Il faut aussi déterminer la cardinalité de ces variables !
  - On peut faire varier la cardinalité et choisir la structure qui donne le meilleur score
  - Quelques problèmes pour calculer la vraisemblance marginale des RB avec variables latentes (Chickering 96)

Références



- **Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference** - J. Pearl (Morgan Kaufman)
- **An introduction to Bayesian Networks** - F. Jensen (Springer Verlag)
- **Probabilistic Networks and Expert Systems** - R.G. Cowell & al. (Springer Verlag)
- **Learning Bayesian Networks** - R. Neapolitan (Prenctice Hall)
- **Learning in Graphical Models** - Jordan M.I. ed. (Kluwer)
- **Les réseaux bayésiens : modèles graphiques de connaissance** - A. Becker & P. Naïm (Eyrolles)

Pour conclure

Les modèles graphiques sont

- adaptés pour la modélisation de connaissances,
- un outil d'inférence/simulation puissant lorsque les données sont incomplètes,
- utilisables dans de nombreux domaines d'application lorsque le cadre est incertain.

Leurs méthodes d'apprentissage permettent

- d'extraire automatiquement des connaissances,
- d'effectuer une sélection de variables.

Merci pour votre attention.

IFSTTAR - GRETTIA
Cité Descartes
2, rue de la butte verte
93166 Noisy-Le-Grand Cedex

Mél. : olivier.francois@ifsttar.fr
Tél. : +33 (0)1 45 92 56 67
Fax : +33 (0)1 45 92 56 40
Site : `http://www.inrets.fr/ur/gretia/poles/diag`

Mél. : francois.olivier.c.h@gmail.com
Tél. : +33 (0)6 28 34 03 96
Site : `http://ofrancois.tuxfamily.org`