Introduction aux processus stochastiques
Leçon 3

Louis Wehenkel

Département EEI
Université de Liège

Montefiore - Liège - 11/3/2008

Find slides: *http://montefiore.ulg.ac.be/∼lwh/ProcStoch/*

Objectifs de cette leçon

Réseaux bayesiens et notion de d-séparation

- Formalisation de la notion de réseau bayesien
- Propriétés graphiques et d-séparation

D-séparation dans les processus Markoviens

- Chaînes de Markov
- Chaînes de Markov cachées

Inférence et d-séparation

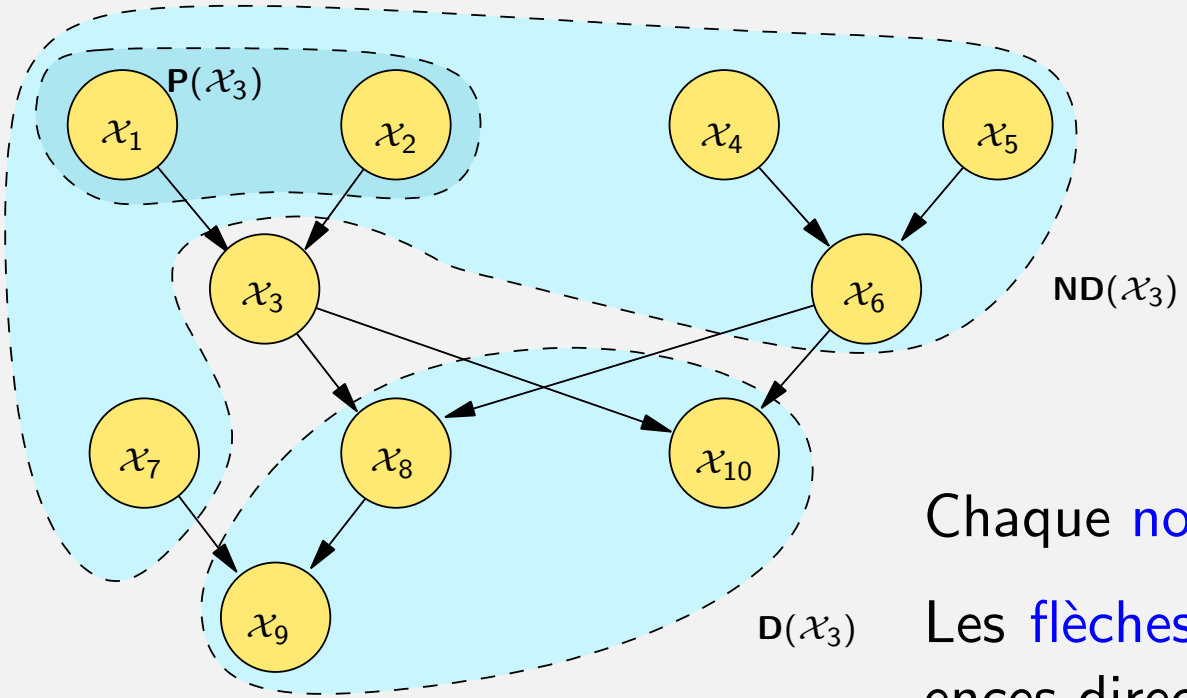
- Principe général
- Application aux chaînes de Markov cachées

Objectifs de cette leçon

- *Expliquer la notion de d-séparation dans les réseaux bayesiens*
- *Montrer comment cette notion peut être exploitée pour inférer des indépendances conditionnelles dans les chaînes de Markov et dans les chaînes de Markov à états cachés.*
- *Développer les algorithmes d'inférence sur chaînes de Markov cachées et illustrer leur intérêt en pratique.*
- *Faire quelques remarques sur les généralisations possibles.*

Réseau bayesien

Une réseau bayesien est un graphe dirigé acyclique qui représente une distribution de probabilités conjointe d'un certain nombre de variables aléatoires discrètes.



Chaque noeud représente une v.a.

Les flèches représentent des influences directes entre v.a.

Acyclique veut dire qu'en suivant les flèches à partir d'un noeud quelconque on ne reviendra jamais à ce noeud.

Terminologie et notations

Désignons par $\mathbf{D} = (\mathbf{V}, \mathbf{E})$ un graphe dirigé acyclique, où $\mathbf{V}$ désigne un ensemble fini de noeuds et $\mathbf{E} \subset \mathbf{V} \times \mathbf{V}$ l'ensemble des flèches.

$\forall \mathcal{X} \in \mathbf{V}$ nous définissons les ensembles de noeuds suivants relatifs à la struture de graphe $\mathbf{D}$:

- $\mathbf{P}(\mathcal{X})$: les parents de $\mathcal{X}$ dans $\mathbf{D}$, i.e. $\{\mathcal{Y} \in \mathbf{V} : (\mathcal{Y}, \mathcal{X}) \in \mathbf{E}\}$.
- $\mathbf{F}(\mathcal{X})$: les fils de $\mathcal{X}$, i.e. l'ensemble $\{\mathcal{Y} \in \mathbf{V} | \mathcal{X} \in \mathbf{P}(\mathcal{Y})\}$.
- $\mathbf{D}(\mathcal{X})$: les descendants de $\mathcal{X}$, i.e. l'union de $\mathbf{F}(\mathcal{X})$ et des descendants de ceux-ci.
- $\mathbf{ND}(\mathcal{X})$: les nondescendants de $\mathcal{X}$, i.e. $\mathbf{V} \setminus (\{\mathcal{X}\} \cup \mathbf{D}(\mathcal{X}))$.

La figure qui précède illustre ces notions pour le noeud $\mathcal{X}_3$.

Pour faciliter la lecture nous utilisons des lettres droites grasses pour désigner des ensembles de noeuds ou de variables, et réservons les lettres rondes pour désigner des noeuds ou variables.

Propriété fondamentale

La structure graphique du réseau bayesien exprime le fait que

> Conditionnellement aux parents d'une variable, celle-ci est indépendante de tout sous-ensemble de ses nondescendants.

Autrement dit:

$$\forall \mathcal{X} \in \mathbf{V}, \forall\, \mathbf{W} \subset \mathbf{ND}(\mathcal{X}) : \mathcal{X} \perp \mathbf{W} | \mathbf{P}(\mathcal{X}).$$

De cela découle la propriété de factorisation:

$$P(\mathbf{V}) = \prod_{\mathcal{X} \in \mathbf{V}} P(\mathcal{X} | \mathbf{P}(\mathcal{X})).$$

Démonstration: supposons que le variables de $\mathbf{V}$ sont numérotées selon un ordre ancestral $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$, c'est-à-dire tel que pour toute variable les numéros d'ordre de ses parents sont inférieurs à son propre numéro d'ordre. Cela implique que les descendants d'une variable sont de numéro d'ordre supérieur à celle-ci. (On montre plus loin qu'il est toujours possible d'ordonner les variables d'un graphe acyclique fini selon un ordre ancestral.)

Factorisons alors $P(\mathbf{V})$ dans cet ordre : en toute généralité on a $P(\mathcal{X}_1, \ldots, \mathcal{X}_n) = \prod_{i=1}^{n} P(\mathcal{X}_i | \mathcal{X}_{i-1} \ldots, \mathcal{X}_1)$.

Mais, $\forall i$ l'ensemble de variables $\{\mathcal{X}_{i-1} \ldots, \mathcal{X}_1\}$ contient par hypothèse les parents de $\mathcal{X}_i$ et aucun de ses descendants. On a donc $\forall i : P(\mathcal{X}_i | \mathcal{X}_{i-1} \ldots, \mathcal{X}_1) = P(\mathcal{X}_i | \mathbf{P}(\mathcal{X}_i))$. CQFD

Algorithme de construction d'un ordre ancestral

Remarques préliminaires:

- Dans tout graphe acyclique fini (disons à n noeuds) il existe au moins un noeud qui ne possède pas de père.

  Sinon, on pourrait construire un chemin de longueur n en remontant de père en père n fois à partir d'un noeud quelconque, ce qui impliquerait l'existence d'un cycle.

- Tout sous-graphe d'un graphe acyclique est acyclique.

  Un sous-graphe $\mathbf{D}'$ est obtenu à partir d'un graphe $\mathbf{D}$ en enlevant certains noeuds et toutes les flèches qui touchent ces noeuds. I.e. $\mathbf{D}' = (\mathbf{V}', \mathbf{E}')$, avec $\mathbf{V}' \subset \mathbf{V}$ et $\mathbf{E}' = \mathbf{E} \cap (\mathbf{V}' \times \mathbf{V}')$.

Algorithme d'ordonnancement ancestral d'un graphe acyclique $\mathbf{D}$:

- $i = 1$, $\mathbf{D}_1 = \mathbf{D}$;
- tant que $i \leq n$: appeler $\mathcal{X}_i$ un des noeuds sans père de $\mathbf{D}_i$; appeler $\mathbf{D}_{i+1}$ le sous-graphe de $\mathbf{D}_i$ obtenu en enlevant $\mathcal{X}_i$; incrémenter i.

  Le fait que le noeud $\mathcal{X}_i$ soit sans père dans $\mathbf{D}_i$ implique que tous ses pères dans $\mathbf{D}$ ont été enlevés lors des étapes précédentes et sont donc de numéro d'ordre inférieur.

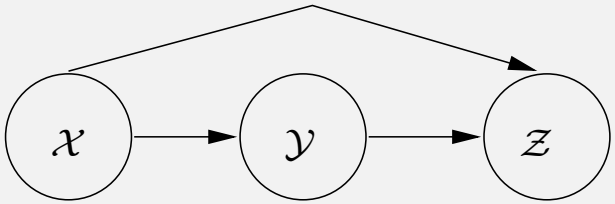
Réseau bayesien

Etant étant donné un ensemble de variables aléatoires $\mathbf{V}$, un réseau bayesien sur $\mathbf{V}$ est défini par un graphe dirigé acyclique $\mathbf{D}$ ayant pour ensemble de noeuds $\mathbf{V}$ et un ensemble de distributions conditionnelles $P(\mathcal{X}|\mathbf{P}(\mathcal{X})), \forall \mathcal{X} \in \mathbf{V}$.

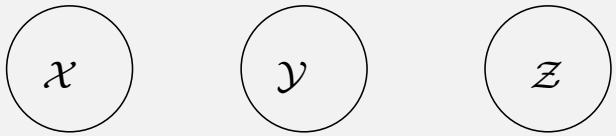
Le réseau bayesien exprime le fait que la distrubution conjointe des variables de $\mathbf{V}$ est égale au produit des distributions conditionnelles $P(\mathcal{X}|\mathbf{P}(\mathcal{X}))$, ou de manière équivalente que la propriété d'indépendance fondamentale est respectée par cette distribution.

Cette propriété d'indépendance fondamentale implique d'autres relations d'indépendance condionnelle qui peuvent être déduites de la structure du graphe par le biais de la notion de d-séparation.

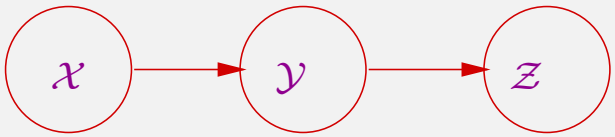
Quelques réseaux bayesiens à trois variables



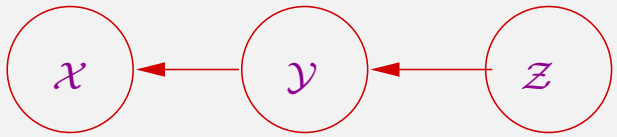
$P(\mathcal{X},\mathcal{Y},\mathcal{Z}) = P(\mathcal{X})P(\mathcal{Y}|\mathcal{X})P(\mathcal{Z}|\mathcal{X},\mathcal{Y})$



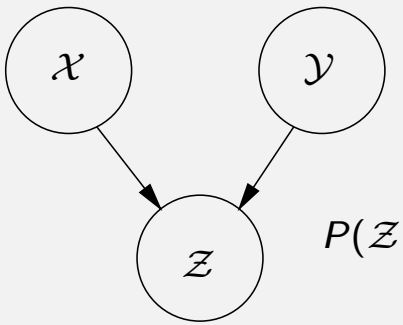
$P(\mathcal{X},\mathcal{Y},\mathcal{Z}) = P(\mathcal{X})P(\mathcal{Y})P(\mathcal{Z})$



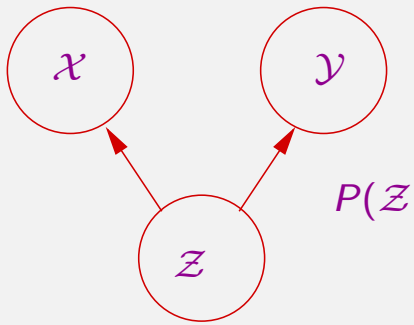
$P(\mathcal{X},\mathcal{Y},\mathcal{Z}) = P(\mathcal{X})P(\mathcal{Y}|\mathcal{X})P(\mathcal{Z}|\mathcal{Y})$



$P(\mathcal{X},\mathcal{Y},\mathcal{Z}) = P(\mathcal{Z})P(\mathcal{Y}|\mathcal{Z})P(\mathcal{X}|\mathcal{Y})$



$P(\mathcal{Z}|\mathcal{X},\mathcal{Y})P(\mathcal{X})P(\mathcal{Y})$



$P(\mathcal{Z})P(\mathcal{X}|\mathcal{Z})P(\mathcal{Y}|\mathcal{Z})$

Communications entre noeuds

Soit un graphe $\mathbf{D} = (\mathbf{V}, \mathbf{E})$ et deux noeuds $\mathcal{X}$ et $\mathcal{Y}$ de $\mathbf{V}$.

On dit que $\mathcal{X}$ et $\mathcal{Y}$ sont directement reliés si $\mathcal{X} \in \mathbf{P}(\mathcal{Y}) \vee \mathcal{Y} \in \mathbf{P}(\mathcal{X})$.

On dit que $\mathcal{X}$ et $\mathcal{Y}$ communiquent s'il existe une suite de noeuds *distincts* $\mathcal{X}_1, \ldots, \mathcal{X}_k$ telle que $\forall i$, $\mathcal{X}_i$ est directement relié à $\mathcal{X}_{i+1}$ et telle que $\mathcal{X}_1 = \mathcal{X}$ et $\mathcal{X}_k = \mathcal{Y}$.

Un telle suite est appelée communication entre $\mathcal{X}$ et $\mathcal{Y}$.

En d'autres mots, deux noeuds de $\mathbf{D}$ communiquent si dans la version nondirigée de $\mathbf{D}$ il existe un "chemin sans cycle" qui les relie.

NB: il est facile de voir que s'il existe un chemin avec cycle qui relie deux noeuds, alors il existe forcément aussi un chemin sans cycle qui les relie.

Communication bloquée

Soit un graphe $\mathbf{D} = (\mathbf{V}, \mathbf{E})$ et $\mathbf{Z}$ un sous-ensemble de noeuds de $\mathbf{V}$.

On dit qu'une communication $\mathcal{X}_1, \ldots, \mathcal{X}_k$ est bloquée par l'ensemble de noeuds $\mathbf{Z}$ s'il existe au moins un $j \in \{2, \ldots, k-1\}$ tel que l'une des trois conditions suivantes soit satisfaite:

1. $\mathcal{X}_j$ joue à la fois le rôle de père et de fils dans le chemin et $\mathcal{X}_j \in \mathbf{Z}$.
   Structure: $\mathcal{X}_{j-1} \rightarrow \mathcal{X}_j \rightarrow \mathcal{X}_{j+1}$ ou $\mathcal{X}_{j-1} \leftarrow \mathcal{X}_j \leftarrow \mathcal{X}_{j+1}$.
2. $\mathcal{X}_j$ joue le rôle de père de deux variables dans le chemin et $\mathcal{X}_j \in \mathbf{Z}$.
   Structure: $\mathcal{X}_{j-1} \leftarrow \mathcal{X}_j \rightarrow \mathcal{X}_{j+1}$.
3. $\mathcal{X}_j$ joue le rôle le fils de deux variables dans le chemin alors que $\mathbf{Z}$ ne contient ni $\mathcal{X}_j$ ni aucun de ces descendants.
   Structure: $\mathcal{X}_{j-1} \rightarrow \mathcal{X}_j \leftarrow \mathcal{X}_{j+1}$, avec $\mathbf{Z} \cap (\{\mathcal{X}_j\} \cup \mathbf{D}(\mathcal{X}_j)) = \emptyset$.

D-séparation

Soit un réseau bayesien de structure $\mathbf{D} = (\mathbf{V}, \mathbf{E})$ et $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ trois sous-ensembles disjoints de variables (et donc de noeuds) de $\mathbf{D}$.

D-séparation: définition

On dit que les ensembles $\mathbf{X}$ et $\mathbf{Y}$ sont d-séparés par l'ensemble $\mathbf{Z}$ si toutes les communications entre un élément quelconque de $\mathbf{X}$ et un élément quelconque de $\mathbf{Y}$ sont bloquées par $\mathbf{Z}$.

D-séparation: propriété fondamentale

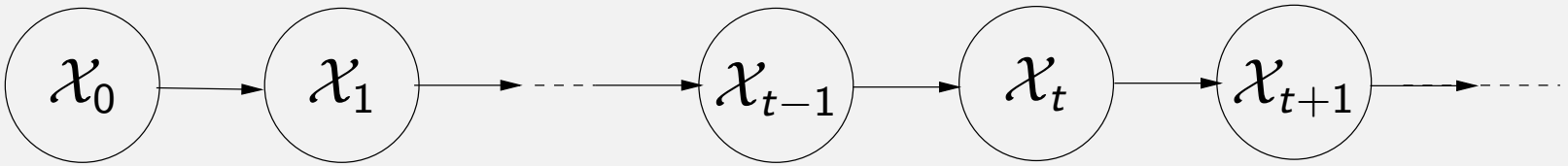
Si $\mathbf{X}$ et $\mathbf{Y}$ sont d-séparés par $\mathbf{Z}$ alors $\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$.

Attention: la réciproque n'est pas vraie.

Notons que $\mathbf{X} \perp \mathbf{Y} | \mathbf{Z} \Rightarrow \mathbf{X}' \perp \mathbf{Y}' | \mathbf{Z}$, pour tout $\mathbf{X}' \subset \mathbf{X}$ et $\mathbf{Y}' \subset \mathbf{Y}$.
Cependant $\mathbf{X} \perp \mathbf{Y} | \mathbf{Z} \not\Rightarrow \mathbf{X} \perp \mathbf{Y} | \mathbf{Z}'$ pour $\mathbf{Z}' \subset \mathbf{Z}$.

Le réseau bayesien d'une chaîne de Markov

La chaîne de Markov correspond au réseau bayesien suivant:



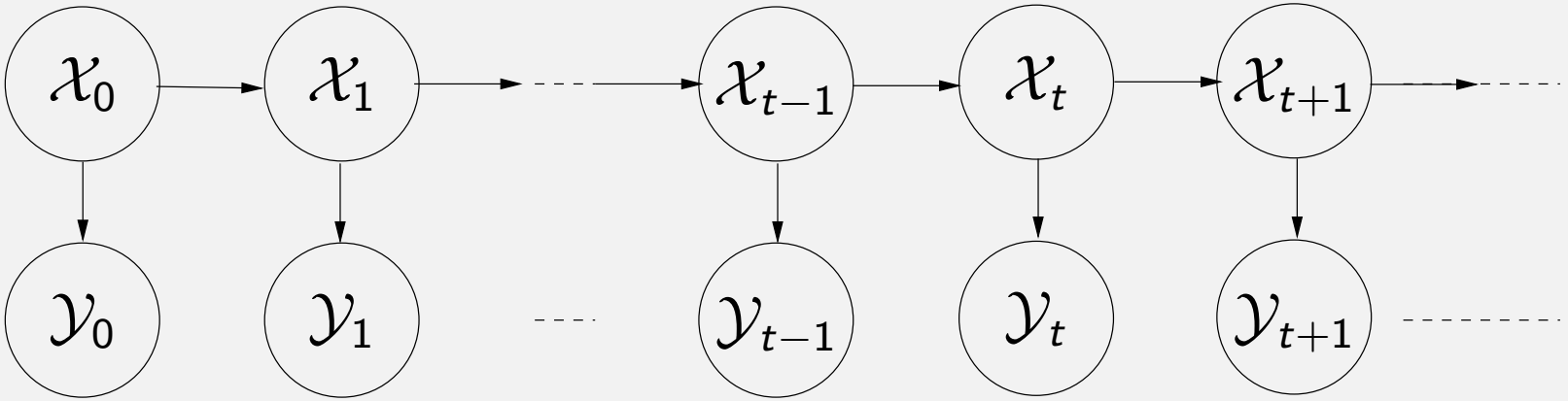
On a $\forall t: P(\mathcal{X}_0, \ldots, \mathcal{X}_t) = P(\mathcal{X}_0) \prod_{i=1}^{t} P(\mathcal{X}_i|\mathcal{X}_{i-1})$.

On a, $\forall i < j < k$: $\{\mathcal{X}_j\}$ d-sépare $\{\mathcal{X}_i\}$ et $\{\mathcal{X}_k\}$, donc $\mathcal{X}_i \perp \mathcal{X}_k|\mathcal{X}_j$.
et $\{\mathcal{X}_{i_1}, \ldots, \mathcal{X}_{i_n}\} \perp \{\mathcal{X}_{k_1}, \ldots, \mathcal{X}_{k_m}\}|\mathcal{X}_j$, si $\max\{i_j\} < j < \min\{k_j\}$.

On en déduit que si $i_1, \ldots, i_n$ forment une suite croissante ou décroissante d'indices, la distribution conjointe $P(\mathcal{X}_{i_1}, \ldots, \mathcal{X}_{i_n})$ se factorise en $P(\mathcal{X}_{i_1}) \prod_{j=2}^{n} P(\mathcal{X}_{i_j}|\mathcal{X}_{i_j-1})$.

$\Rightarrow$ toute sous-suite d'une chaîne de Markov est une chaîne de Markov!

Le réseau bayesien d'une chaîne de Markov cachée



On a $\forall t$: $P(\mathcal{X}_0, \ldots, \mathcal{X}_t, \mathcal{Y}_0, \ldots, \mathcal{Y}_t) = \prod_{i=0}^{t} P(\mathcal{X}_i|\mathcal{X}_{i-1})P(\mathcal{Y}_i|\mathcal{X}_i)$

On a, $\forall i < j < k$: $\{\mathcal{X}_j\}$ d-sépare $\{\mathcal{X}_i\}$ et $\{\mathcal{X}_k\}$, donc $\mathcal{X}_i \perp \mathcal{X}_k|\mathcal{X}_j$.
et $\{\mathcal{X}_j\}$ d-sépare $\{\mathcal{X}_i, \mathcal{Y}_i\}$ et $\{\mathcal{X}_k, \mathcal{Y}_k\}$, donc $\{\mathcal{X}_i, \mathcal{Y}_j\} \perp \{\mathcal{X}_k, \mathcal{Y}_k\}|\mathcal{X}_j$.

On a aussi $\{\mathcal{X}_i, \mathcal{Y}_j\} \perp \{\mathcal{X}_k, \mathcal{Y}_k\}|\{\mathcal{X}_j, \mathcal{Y}_j\}$ $\forall i < j < k$.

Mais $\{\mathcal{Y}_j\}$ ne d-sépare pas $\{\mathcal{Y}_i\}$ et $\{\mathcal{Y}_k\}$, donc $\mathcal{Y}_i \not\perp \mathcal{Y}_k|\mathcal{Y}_j$.

Méthode générale d'inférence

Inférence: calculer une distribution conditionnelle $P(\mathcal{Y}|\mathcal{X} = X)$ à partir d'une représentation de la loi conjointe $P(\mathcal{X}, \mathcal{Y}, \mathcal{Z}_1, \ldots, \mathcal{Z}_n)$.

Méthode générale:

- On calcule d'abord $P(\mathcal{X} = X, \mathcal{Y})$ en intégrant sur les $\mathcal{Z}_i$:

$$P(\mathcal{X} = X, \mathcal{Y}) = \sum_{i_1} \cdots \sum_{i_n} P(\mathcal{X} = X, \mathcal{Y}, \mathcal{Z}_1 = Z_{i_1}, \ldots, \mathcal{Z}_n = Z_{i_n})$$

- On calcule ensuite $P(\mathcal{Y}|\mathcal{X} = X)$ par normalisation:

$$P(\mathcal{Y}|\mathcal{X} = X) = \frac{P(\mathcal{X} = X, \mathcal{Y})}{P(\mathcal{X} = X)} = \frac{P(\mathcal{X} = X, \mathcal{Y})}{\sum_i P(\mathcal{X} = X, \mathcal{Y} = Y_i)}$$

La méthode générale demande de l'ordre de $\prod_{i=1}^{n} |\mathcal{Z}_i|$ opérations.

Exploitation de la d-séparation

Supposons que parmi les variables $\mathcal{Z}_i$ il en existe une, disons $\mathcal{Z}_k$ qui soit telle que $\{\mathcal{X}, \mathcal{Z}_1, \ldots, \mathcal{Z}_{k-1}\} \perp \{\mathcal{Y}, \mathcal{Z}_{k+1}, \ldots, \mathcal{Z}_n\} | \mathcal{Z}_k$.

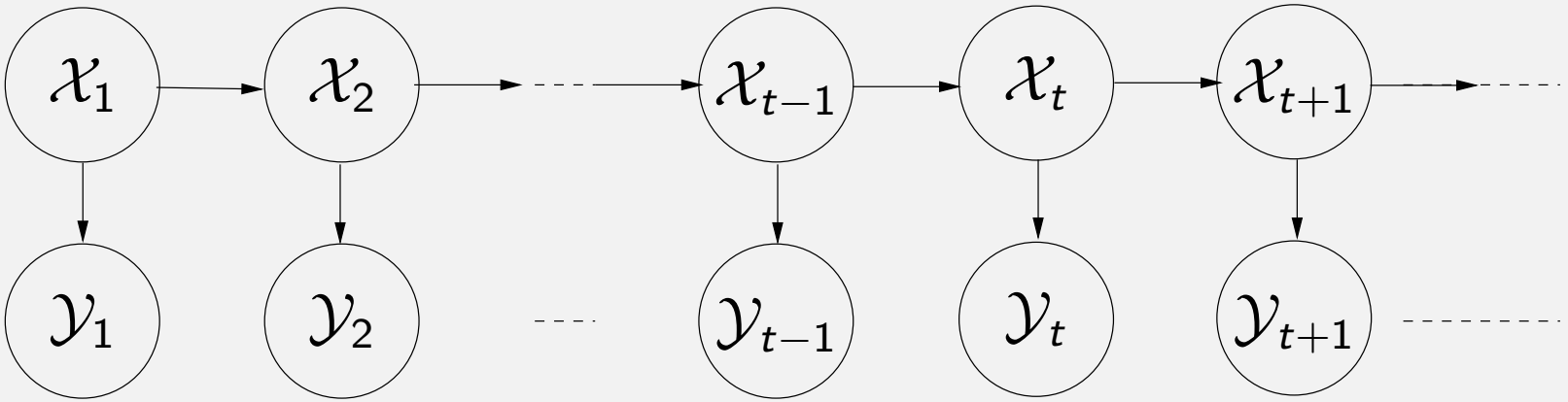
Cela implique que $P(\mathcal{X}, \mathcal{Y}, \mathcal{Z}_k) = P(\mathcal{X}, \mathcal{Z}_k) P(\mathcal{Y} | \mathcal{Z}_k[\mathcal{X}])$ et donc que $P(\mathcal{Y}, \mathcal{X} = X) = \sum_{i_k} P(\mathcal{Z}_k = Z_{i_k}, \mathcal{X} = X) P(\mathcal{Y} | \mathcal{Z}_k = Z_{i_k})$.

On peut donc calculer $P(\mathcal{Y} | \mathcal{X} = X)$ en quatre temps:

1. on calcule d'abord $P(\mathcal{Z}_k, \mathcal{X} = X)$;
2. et en parallèle on calcule $P(\mathcal{Y} | \mathcal{Z}_k = Z_{i_k}), \forall i_k$;
3. $P(\mathcal{Y}, \mathcal{X} = X) = \sum_{i_k} P(\mathcal{Z}_k = Z_{i_k}, \mathcal{X} = X) P(\mathcal{Y} | \mathcal{Z}_k = Z_{i_k})$;
4. et enfin $P(\mathcal{Y} | \mathcal{X} = X)$ par normalisation de $P(\mathcal{Y}, \mathcal{X} = X)$.

Dans un RB où $\mathcal{Z}_k$ d-sépare $\{\mathcal{X}, \mathcal{Z}_1, \ldots, \mathcal{Z}_{k-1}\}$ et $\{\mathcal{Y}, \mathcal{Z}_{k+1}, \ldots, \mathcal{Z}_n\}$ cela permet de réduire la complexité du calcul de $P(\mathcal{Y} | \mathcal{X} = X)$.

Inférence dans les CMC



Problèmes d'inférence:

- Connaissant $Y_1, \ldots, Y_T$ déterminer $P(\mathcal{X}_T | Y_1, \ldots, Y_T)$.
  I.e. le problème d'estimation d'état.
- Connaissant $Y_1, \ldots, Y_T$ déterminer $P(\mathcal{X}_{T+k} | Y_1, \ldots, Y_T)$.
  I.e. le problème de prédiction.
- Connaissant $Y_1, \ldots, Y_T$ déterminer $P(\mathcal{X}_t | Y_1, \ldots, Y_T)$, quelque soit $t \in \{1, \ldots, T\}$.
  I.e. le problème de filtrage.

Notations

Sans restriction, nous pouvons supposer que $\forall t$, $\mathcal{X}_t \in \{1, \ldots, p\}$ et $\mathcal{Y}_t \in \{1, \ldots, q\}$. Nous utiliserons les notations suivantes:

- π_1 pour désigner le vecteur ligne de dimension p dont la i-ème composante vaut $P(\mathcal{X}_1 = i)$.
- Π_t pour désigner la matrice $p \times p$ dont l'élément i, j vaut $P(\mathcal{X}_{t+1} = j | \mathcal{X}_t = i)$.
- Σ_t pour désigner la matrice $p \times q$ dont l'élément i, j vaut $P(\mathcal{Y}_t = j | \mathcal{X}_t = i)$.
- c_i pour désigner le vecteur colonne de dimension q dont la i-ème composante vaut 1 et dont toutes les autres composantes sont nulles.

NB: si la CMC est invariante dans les temps alors $\forall t$ on a $\Pi_t = \Pi$ et $\Sigma_t = \Sigma$.

Propagation gauche droite: α

La procédure "gauche $\rightarrow$ droite" calcule pour $t = 1, \ldots, T$:

$$\hat{\alpha}_t(i) = P(\mathcal{Y}_1 = Y_1, \cdots, \mathcal{Y}_{t-1} = Y_{t-1}, \mathcal{X}_t = i). \quad (1)$$

$$\alpha_t(i) = P(\mathcal{Y}_1 = Y_1, \cdots, \mathcal{Y}_{t-1} = Y_{t-1}, \mathcal{Y}_t = Y_t, \mathcal{X}_t = i). \quad (2)$$

La procédure de calcul est la suivante

1. Initialisation : $\hat{\alpha}_1(i) = P(\mathcal{X}_1 = i) = \pi_{1i}, \forall i = 1, \ldots, p$
   En notation vectorielle: $\hat{\alpha}_1 = \pi_1$
2. Mise à jour (pour $t = 1, \ldots, T$) :
   2.1 Correction: $\alpha_t = \hat{\alpha}_t \times (\Sigma_t c_{Y_t})$ ($\times$ = produit terme à terme)
   2.2 Prédiction: $\hat{\alpha}_{t+1} = \alpha_t \Pi_t$.
   En d'autres termes: $\alpha_{t+1} = (\alpha_t \Pi_t) \times (\Sigma_{t+1} c_{Y_{t+1}})$.

A partir de là on récupère:

- $P(\mathcal{X}_T | \mathcal{Y}_1 = Y_1, \cdots, \mathcal{Y}_T = Y_T)$ par normalisation de α_T.
- $P(\mathcal{X}_{T+1} | \mathcal{Y}_1 = Y_1, \cdots, \mathcal{Y}_T = Y_T)$ par normalisation de $\hat{\alpha}_{T+1}$.

Démonstration de la propagation α

La justification de l'induction est la suivante :

$$\begin{aligned}
\alpha_{t+1}(j) &= P(Y_1 \cdots Y_t, \mathcal{X}_{t+1} = j, Y_{t+1}) && (3)\\
&= P(Y_1 \cdots Y_t, \mathcal{X}_{t+1} = j) P(Y_{t+1} | Y_1 \cdots Y_t, \mathcal{X}_{t+1} = j) && (4)\\
&= P(Y_1 \cdots Y_t, \mathcal{X}_{t+1} = j) P(Y_{t+1} | \mathcal{X}_{t+1} = j) && (5)\\
&= P(Y_1 \cdots Y_t, \mathcal{X}_{t+1} = j) [\Sigma_{t+1}]_{j, Y_{t+1}}, && (6)
\end{aligned}$$

où le passage de (4) vers (5) est justifié car $\mathcal{X}_{t+1}$ bloque la communication avec $\mathcal{Y}_{t+1}$.

$$\begin{aligned}
\hat{\alpha}_{t+1}(j) = P(Y_1 \cdots Y_t, \mathcal{X}_{t+1} = j) &= \sum_{i=1}^{N} P(Y_1 \cdots Y_t, \mathcal{X}_t = i, \mathcal{X}_{t+1} = j) && (7)\\
&= \sum_{i=1}^{N} P(Y_1 \cdots Y_t, \mathcal{X}_t = i) P(\mathcal{X}_{t+1} = j | Y_1 \cdots Y_t, \mathcal{X}_t = i) && (8)\\
&= \sum_{i=1}^{N} P(Y_1 \cdots Y_t, \mathcal{X}_t = i) P(\mathcal{X}_{t+1} = j | \mathcal{X}_t = i) && (9)\\
&= \sum_{i=1}^{N} \alpha_t(i) [\Pi_t]_{i,j}, && (10)
\end{aligned}$$

où le passage de (8) à (9) est justifié par le fait que $\mathcal{X}_t$ est le (seul) père de $\mathcal{X}_{t+1}$ et que les variables $\mathcal{Y}_1, \mathcal{Y}_2 \ldots \mathcal{Y}_t$ sont des non-descendants de $\mathcal{Y}_{t+1}$.

Propagation α: remarques

Aspects numériques:
En pratique les valeurs de α_t et $\hat{\alpha}_t$ on tendance à devenir de plus en plus petites au fil des itérations. Il est donc préférable de les renormaliser à chaque itération. On peut se convaincre que le même algorithme peut encore s'appliquer, en ajoutant après l'étape (2.1) une étape de normalisation

$$\alpha_t(i) := \frac{\alpha_t(i)}{\sum_j \alpha_t(j)}. \tag{11}$$

NB: l'étape (2.2) préserve le caractère normalisé automatiquement.

Prédiction à plusieurs pas de temps:
$P(\mathcal{X}_{T+k}|Y_1,\ldots,Y_T)$ peut-être calculée à partir de (la version normalisée de) α_T en lui appliquant les k matrices de transition $\Pi_T,\ldots\Pi_{T+k-1}$:

$$P(\mathcal{X}_{T+k}|Y_1,\ldots,Y_T) = \alpha_T \Pi_T \cdots \Pi_{T+k-1}. \tag{12}$$

Propagation droite gauche: β

La procédure "droite $\rightarrow$ gauche" calcule les coefficients suivants

$$\beta_t(i) \quad = \quad P(\mathcal{Y}_{t+1} = Y_{t+1}, \cdots, \mathcal{Y}_T = Y_T | \mathcal{X}_t = i). \qquad (13)$$

La procédure de calcul est la suivante

1. Initialisation : $\beta_T(i) = 1, \forall i = 1, \ldots, p$
   En notation vectorielle: $\beta_T = \mathbf{1}$ (vecteur colonne.)
2. Mise à jour (pour $t = T, \ldots, 2$) :
   2.1 Correction: $\hat{\beta}_t = ((\Sigma_t c_{Y_t})^T \times \beta_t)^T$ ($\times$ = produit terme à terme)
   2.2 "Prédiction": $\beta_{t-1} = \Pi_{t-1} \hat{\beta}_t$.

A partir de là on récupère:

- $P(\mathcal{X}_t | \mathcal{Y}_1 = Y_1, \cdots, \mathcal{Y}_T = Y_T)$ par normalisation de $\alpha_t \times \beta_t$.

NB: Démonstrations $\rightarrow$ homework!

Applications

- Codage d'information pour transmission ou stockage à l'aide de dispositifs imparfaits (correction d'erreurs)
- Modélisation de comportements d'utilisateurs (réseaux, marketing...)
- Economie, finance
- Intelligence artificielle (processus de décision intelligents)
- Robotique (planification récative de tâches)
- Génomique, protéomique
- ...

Généralisation des algorithmes d'inférence

- Extension au cas où certaines observations sont manquantes
- Structures de réseaux bayesiens plus générales
- Apprentissage des paramètres

Inférence lorsque certaines observations manquent

- Supposons que la valeur de la variable $\mathcal{Y}_i$ ne soit pas connue.
- But: calculer $P(\mathcal{X}_t|\mathcal{Y}_1 = Y_1, \cdots, [\mathcal{Y}_i] \cdots, \mathcal{Y}_T = Y_T)$
- Revient à calculer $\sum_j P(\mathcal{X}_t, \mathcal{Y}_1 = Y_1, \cdots, \mathcal{Y}_i = j, \cdots, \mathcal{Y}_T = Y_T)$, puis à normaliser.
- Notez que $\alpha_t \beta_t = P(\mathcal{X}_t, \mathcal{Y}_1 = Y_1, \cdots, \mathcal{Y}_i = Y_i, \cdots, \mathcal{Y}_T = Y_T)$
- Calculer $P(\mathcal{X}_t|\mathcal{Y}_1 = Y_1, \cdots, [\mathcal{Y}_i] \cdots, \mathcal{Y}_T = Y_T)$, revient donc à superposer les inférences qui seraient faites pour chacune des valeurs possibles de $\mathcal{Y}_i$, puis normaliser.
- Cela revient donc à propager successivement les vecteurs c_j $\forall j = 1, \ldots, q$ à partir de la variable $\mathcal{Y}_i$
- Ce qui revient à propager "une seule fois" le vecteur $\sum_j c_j = 1$
- (Revient aussi à ne rien propager à partir de $\mathcal{Y}_i$, puisque $\Sigma_i 1 = 1$.)

Inférence lorsque certaines observations manquent

En général

- Initialisation:
  - on associe le vecteur c_{Y_k} à chacune des variables $\mathcal{Y}$ dont la valeur est connue
  - on associe le vecteur 1 à chacune des variables qui ne sont pas connues
  - on associe le vecteur l_{X_j} à chacune des variables $\mathcal{X}$ dont la valeur est connue
  - on associe le vecteur π_1 à $\mathcal{X}_1$
- Propagation avant et arrière pour calculer les probabilités conditionneles des autres variables étant données les observations.

Généralisation aux réseaux bayesiens

- Différentes familles de structures de RB
- Graphes de factorisation
- Transformations de graphes
- Autres problèmes similaires
  - Meilleure explication
  - ...