

## 4.1 Rappels de Notation

- $(X_1, \dots, X_n)$  v. a. discrètes
- $p(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$  Loi jointe
- $p(x_A) = \sum_{x_{A^c}} p(x_A, x_{A^c})$  Loi marginale
- $p(x_A | x_{A^c}) = \frac{p(x_A, x_{A^c})}{p(x_{A^c})}$  Loi conditionnelle
- $p(x_1, \dots, x_n) = p(x_1) p(x_2 | x_1) p(x_3 | x_2, x_1) \cdots p(x_n | x_{n-1}, \dots, x_1)$
- $X \perp\!\!\!\perp Y \mid Z \Leftrightarrow p(x, y | z) = p(x | z) p(y | z) \Leftrightarrow p(x | y, z) = p(x | z) = \frac{p(x, y | z)}{p(y | z)}$

## 4.2 Modèles graphiques orientés

(Directed Graphical Models, Belief Networks, Bayesian Networks, Réseaux Bayésiens)



ATTENTION : l'indépendance par paires n'implique pas forcément l'indépendance jointe.

### 4.2.1 Définitions et premières propriétés

Soient  $G = (V, E)$  un DAG avec  $V = \{1, \dots, n\}$ , et  $X_1, \dots, X_n$   $n$  variables aléatoires discrètes. Soit :

$$p(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_i(x_i, x_{\pi_i})$$

avec :

1.  $\pi_i$  est l'ensemble des parents de  $i$
2.  $\forall i, f_i \geq 0$
3.  $\forall i, \sum_{x_i} f_i(x_i, x_{\pi_i}) = 1$

**Proposition 4.1**  $p(x_1, \dots, x_n)$  est une distribution sur  $(X_1, \dots, X_n)$ , et :

$$\forall i, p(x_i | x_{\pi_i}) = f_i(x_i, x_{\pi_i})$$

**Démonstration** Elle se fait par récurrence sur le nombre de sommets. Le cas  $n = 1$  est trivial. On passe de  $n - 1$  à  $n$  en considérant que comme  $G$  est acyclique, il contient une feuille. Sans perte de généralité, on peut supposer que cette feuille est le nœud  $n$ .

**Hypothèse de récurrence**  $(x_1, \dots, x_{n-1})$  : Si  $p(x_1, \dots, x_{n-1}) = \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i})$  alors  $p(x_1, \dots, x_n)$  est une distribution et  $\forall i \in \{1, \dots, n\}$ ,  $p(x_i | x_{\pi_i}) = f_i(x_i, x_{\pi_i})$

On a :

$$\begin{aligned} p(x_1, \dots, x_{n-1}) &= \sum_{x_n} p(x) \\ &= \sum_{x_n} \left( \prod_{i=1}^n f_i(x_i, x_{\pi_i}) \right) \\ &= \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) \sum_{x_n} f_n(x_n, x_{\pi_n}) \\ &= \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) \\ &= g(x_1, \dots, x_{n-1}) \end{aligned}$$

Par hypothèse de récurrence, on a pour tout  $i \leq n - 1$ ,  $f(x_i, x_{\pi_i}) = p(x_i | x_{\pi_i})$ , donc :

$$\begin{aligned} p(x_n, x_{\pi_n}) &= \sum_{x_i, i \notin \{n\} \cup \pi_n} p(x_1, \dots, x_n) \\ &= \sum_{x_i, i \notin \{n\} \cup \pi_n} g(x) f_n(x_n, x_{\pi_n}) \\ &= f_n(x_n, x_{\pi_n}) \underbrace{\sum_{x_i, i \notin \{n\} \cup \pi_n} g(x)}_{h(x_{\pi_n})} \end{aligned}$$

avec  $h(x_{\pi_n})$  qui ne dépend que de  $x_{\pi_n}$ .

On en déduit que :

$$p(x_{\pi_n}) = \underbrace{\sum_{x_n} f_n(x_n, x_{\pi_n})}_{=1} h(x_{\pi_n})$$

Ainsi :

$$p(x_n | x_{\pi_n}) = \frac{p(x_n, x_{\pi_n})}{p(x_{\pi_n})} = \frac{h(x_{\pi_n}) f_n(x_n, x_{\pi_n})}{h(x_{\pi_n})} = f_n(x_n, x_{\pi_n})$$

Et finalement, on montre que  $\sum_x p(x) = 1$  :

$$p(x_1, \dots, x_n) = f_n(x_n, x_{\pi_n}) p(x_1, \dots, x_{n-1})$$

$$\sum_x p(x) = \underbrace{\sum_{x_1} \cdots \sum_{x_{n-1}} p(x_1, \dots, x_{n-1})}_{=1} \underbrace{\sum_{x_n} f_n(x_n, x_{\pi_n})}_{=1} = 1$$

■

**Définition 4.2** Soit  $G$  un graphe orienté. On dit que  $p$  se factorise dans  $G$  ( $p \in \mathcal{L}(G)$ ) si et seulement si il existe  $(f_1, \dots, f_n)$  positives et de somme égale à un, telles que :

$$p(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_i, x_{\pi_i}).$$

## 4.2.2 Exemples

**Graphe sans arêtes**  $G = (V, E), V = \{1, 2, \dots, n\}$  avec  $E = \emptyset$

$$p(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_i)$$

Alors  $p(x) \in \mathcal{L}(G) \iff p(x) = \prod_{i=1}^n p(x_i) \iff (X_1, \dots, X_n)$  sont indépendantes. Donc  $\mathcal{L}(G)$  est l'ensemble des lois indépendantes sur  $(X_1, \dots, X_n)$ .

**Graphe complet**  $G = (V, E), V = \{1, 2, \dots, n\}$  avec  $\forall i \leq j, (i, j) \in E$

$$p(x_1, \dots, x_n) = f_1(x_1) f_2(x_2, x_1) f_3(x_3, x_2, x_1) \cdots f_n(x_n, x_{n-1}, \dots, x_1)$$

Toute distribution sur  $(X_1, \dots, X_n)$  peut s'écrire sous cette forme, donc  $\mathcal{L}(G)$  est l'ensemble de toutes les distributions possibles sur  $(X_1, \dots, X_n)$ .

## 4.2.3 Exemples canoniques de graphes à trois noeuds

« Chaîne de Markov » Montrons que pour toute loi se factorisant dans  $G$  on a :  $X \perp\!\!\!\perp Z \mid Y$ , en montrant que  $p(z|x, y) = p(z|y)$ .

On a :

$$p(z|x, y) = \frac{p(x, y, z)}{p(x, y)} = \frac{p(x)p(y|x)p(z|y)}{p(x)p(y|x)}$$

donc,  $p(z|x, y) = p(z|y) \quad \square$

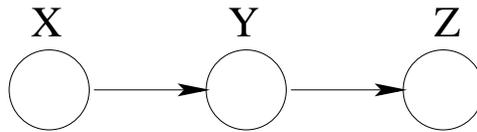


Figure 4.1. « Chaîne de Markov »

## « Cause commune »



ATTENTION : le terme de "cause" est à bannir en statistiques car celles-ci décrivent seulement des corrélations entre différentes variables aléatoires, à ne pas confondre avec des potentielles relations de causalité (Ici corrélation n'est pas à prendre dans le sens mathématiques de corrélation linéaire, mais dans le sens de relation entre deux quantités).

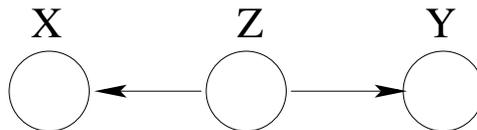


Figure 4.2. « Cause commune »

Montrons que pour toute loi se factorisant dans  $G$ , on a  $X \perp\!\!\!\perp Y \mid Z$ , en montrant que  $p(x, y|z) = p(x|z)p(y|z)$ .

On a :

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(z)p(x|z)p(y|z)}{p(z)}$$

$$\text{donc, } p(x, y|z) = p(x|z)p(y|z) \quad \square$$

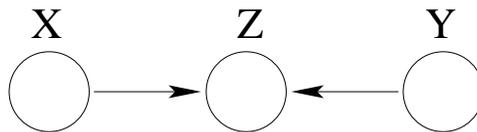


Figure 4.3. « Explaining away »

« Explaining away » Montrons que pour toute loi se factorisant dans  $G$ , on a  $X \perp\!\!\!\perp Y$ , en montrant que  $p(x, y) = p(x)p(y)$ .

On a :

$$p(x, y) = \sum_z p(x, y, z) = \sum_z p(x)p(y)p(z|x, y) = p(x)p(y)$$

$$\text{donc, } p(x, y) = p(x)p(y) \quad \square$$

– Remarque  $X \perp\!\!\!\perp Y \mid Z$ .

**Proposition 4.3** Soient  $G = (V, E)$  et  $\tilde{G} = (V, \tilde{E})$  deux DAG tels que  $E \subset \tilde{E}$ , alors  $\mathcal{L}(G) \subset \mathcal{L}(\tilde{G})$ .

**Définition 4.4 Arête couverte**  $X_i \leftarrow X_j$  est couverte si et seulement si  $\pi_i = \pi_j \cup \{j\}$

**Proposition 4.5** Soit  $G$  et  $(i, j)$  une arête couverte.  $G' = (V, E')$  avec  $E' = E \setminus \{(i, j)\} \cup \{(j, i)\}$ . Si  $G'$  est un DAG, alors  $\mathcal{L}(G) = \mathcal{L}(G')$ .

**Proposition 4.6** Si  $G$  est un DAG muni d'un ordre  $\{1, \dots, n\}$  topologique, alors

$$\forall i : p(x_i | x_{i-1}, \dots, x_1) = p(x_i | x_{\pi_i})$$

**Démonstration** On démontre ce résultat par récurrence sur le nombre de nœuds, en considérant une feuille ( $G$  étant un DAG, il contient forcément une feuille). ■

**Corollaire 4.7** Si  $G$  est un DAG muni d'un ordre topologique,

$$p(x) \in \mathcal{L}(G) \Rightarrow \forall i, X_i \perp\!\!\!\perp X_{\{1, \dots, i-1\} \setminus \pi_i} \mid X_{\pi_i}$$

#### 4.2.4 Marginalisation dans un modèle graphique orienté (MGO)

**Proposition 4.8** Soit  $p(x) \in \mathcal{L}(G)$ . Si  $n$  est une feuille alors  $p(x_1, \dots, x_{n-1})$  se factorise dans le sous-graphe  $\tilde{G} = (\tilde{V}, \tilde{E})$ ,  $\tilde{V} = V \setminus \{n\}$  et  $\tilde{E} = E \cap (\tilde{V} \times \tilde{V})$ .

En d'autres termes, en marginalisant par rapport à une feuille, on obtient un nouveau modèle graphique à partir de l'ancien modèle dont on a enlevé la feuille.

**Démonstration**

$$\begin{aligned} p(x_1, \dots, x_{n-1}) &= \sum_{x_n} p(x_1, \dots, x_n) \\ &= \sum_{x_n} \left( \prod_{i=1}^{n-1} p(x_i | x_{\pi_i}) p(x_n | x_{\pi_n}) \right) \\ &= \prod_{i=1}^{n-1} p(x_i | x_{\pi_i}) \sum_{x_n} p(x_n | x_{\pi_n}) \\ &= \prod_{i=1}^{n-1} p(x_i | x_{\pi_i}) \end{aligned}$$

■

**Remarque** Dans le cas d'un nœud  $n$  qui n'est pas une feuille, on n'obtient pas forcément une distribution qui se factorise dans un modèle graphique en marginalisant par rapport à  $X_n$ . On verra dans la suite, que la marginalisation par rapport à un nœud quelconque se comporte beaucoup mieux dans le cas des modèles non orientés.

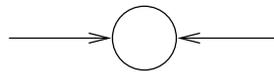
### 4.2.5 Notion de d-separation (*directed separation*)

La séparation classique n'est pas suffisante dans le cadre de graphes orientés. Une séparation plus complexe a donc dû être introduite pour répondre à la question suivante :

**Pour  $p(x) \in \mathcal{L}(G)$ , étant donné  $A, B$  et  $C$  trois nœuds, sous quelles conditions a-t-on  $X_A \perp\!\!\!\perp X_B \mid X_C$  ?**

Cette séparation s'appelle la **d-separation**.

**Définition 4.9** *Structure en  $v$  (**v-structure**)*. On dit qu'il y a une structure en  $v$  au nœud  $d$  si  $d$  a au moins deux parents.



**Figure 4.4.** v-structure

**Définition 4.10** Soit  $a, b \in V$ . Une **chaîne** de  $a$  vers  $b$  est une séquence de nœuds  $a = v_1, \dots, v_k = b$  telle que

$$\forall k, (v_{k-1}, v_k) \in E \text{ ou } (v_k, v_{k-1}) \in E$$

Une **chaîne** de  $a$  vers  $b$  est un **chemin** de  $a$  vers  $b$  dans le graphe symétrisé (c'est-à-dire en supprimant le sens des flèches).

**Définition 4.11** On dit que la chaîne de  $a$  vers  $b$  est **bloquée en  $d$**  par un ensemble de nœuds  $C \subset V$  si et seulement si il existe  $i$  dans  $\{1, 2, \dots, k\}$  tels que  $d = v_i$  et :

- Si  $d \in C$ , alors  $v_{i-1}, v_i, v_{i+1}$  ne forment pas une structure en  $v$ ;
- Si  $d \notin C$ , alors si  $v_{i-1}, v_i, v_{i+1}$  forment une structure en  $v$ , alors ni  $d$  ni aucun descendant de  $d$  dans  $G$  n'appartient à  $C$ .

**Définition 4.12** Une chaîne  $H$  est **bloquée** par  $C$  si et seulement si  $\exists d \in H$  tel que  $H$  soit bloqué en  $d$  par  $C$ .

**Définition 4.13** Soient  $A, B, C \subset V$  disjoints. On dit que  $A$  et  $B$  sont **d-séparés** par  $C$  si et seulement si  $\forall a \in A, \forall b \in B$ , toutes les chaînes de  $a$  vers  $b$  sont bloquées par  $C$ .

**Théorème 4.14** Si  $A$  et  $B$  sont d-séparés par  $C$  alors  $X_A \perp\!\!\!\perp X_B \mid X_C$

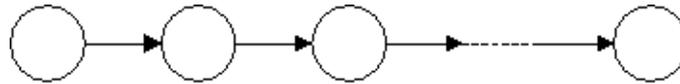


Figure 4.5. Modèle de chaîne de Markov

## 4.2.6 Modèles classiques simples

### Chaîne de Markov

$$p(x_1, \dots, x_n) = p(x_1) \prod_{i=2}^n p(x_i | x_{i-1})$$

Avec la notion de d-séparation, il devient plus facile de démontrer que le futur est indépendant du passé sachant le présent.

### Modèle de Markov caché ou HMM (Hidden Markov Model)

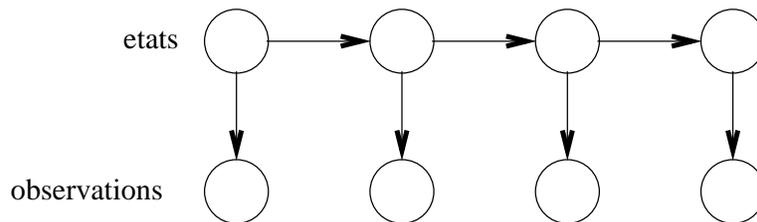


Figure 4.6. Modèle de Markov caché

Dans le cas d'un modèle de Markov caché, la chaîne de Markov n'est pas directement observable. On observe seulement des variables aléatoires reliées à la chaîne. Par exemple, la chaîne de Markov peut être la position d'un mobile au cours du temps et on observe sa position sur un écran radar (donc avec des erreurs et des imprécisions).

## 4.3 Modèles Graphiques non-orientés

(Markov Random Fields, Markov Network, Undirected Graphical Model)

### 4.3.1 Définition

**Définition 4.15** Soient  $G = (V, E)$  un graphe non orienté,  $\mathcal{C}$  l'ensemble des cliques de  $G$ , alors la famille des lois associées à  $G$ , est :

$$\mathcal{L}(G) = \left\{ p \mid p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C), \text{ avec } \forall C \in \mathcal{C}, \psi_C \geq 0 \right\}$$

où  $Z$  est une constante de normalisation (encore appelée fonction de partition) :

$$Z = \sum_x \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

**Remarque** Les éléments de  $\mathcal{L}(G)$  sont clairement des distributions de probabilité.

**Remarque** Les éléments de  $\mathcal{L}(G)$  sont appelées « loi de Gibbs » sur  $G$ . Les fonctions  $\psi_C$  sont appelées « potentiels ».



Dans cette définition de la factorisation dans les modèles graphiques non orientés, on peut prendre pour  $\mathcal{C}$ , soit l'ensemble des cliques, soit l'ensemble des cliques maximales. En effet, chaque clique  $C$  est incluse dans une clique maximale et le potentiel  $\psi_C$  associé peut être combiné avec celui de cette clique maximale.



Les  $\psi_C$  sont définies à une constante près.

Quand  $p \in \mathcal{L}(G)$ , on dit que  $p$  se factorise dans  $G$ . On peut étendre cette terminologie en disant que la fonction  $u$  se factorise dans  $G$  si

$$u(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C).$$

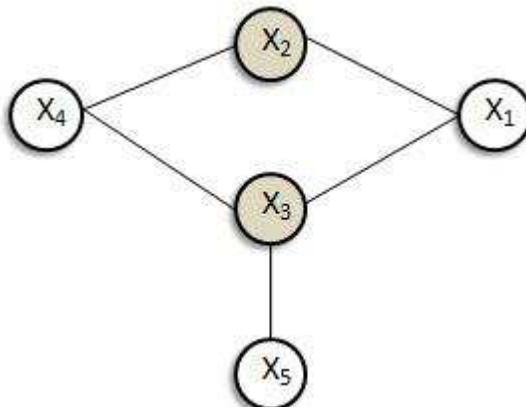
**Exemple** Dans le cas où  $(X_1, \dots, X_n)$  sont indépendantes ( $E = \emptyset$ ), on a bien :

$$p(x) = \prod_{i=1}^n \psi_i(x_i).$$

### 4.3.2 Séparation et indépendance conditionnelle

**Définition 4.16** Soit  $G = (V, E)$  un graphe non orienté avec  $V = \{1, \dots, n\}$  et  $p$  une distribution de probabilité quelconque sur  $(X_1, \dots, X_n)$ .

On dit que  $p$  a la **propriété de Markov globale** sur  $G$  si (et seulement si) pour tous  $A, B, S \subset V$ , disjoints, tels que  $S$  sépare  $A$  et  $B$  (au sens usuel de la théorie des graphes), on a  $X_A \perp\!\!\!\perp X_B \mid X_S$ .



**Figure 4.7. Exemple de modèle graphique non orienté** Dans ce graphe,  $p(x)$  se factorise sous la forme :  $\psi_{12}(x_{12})\psi_{24}(x_{24})\psi_{43}(x_{43})\psi_{31}(x_{31})\psi_{35}(x_{35})$ .

**Proposition 4.17** Si  $p \in \mathcal{L}(G)$ , alors  $p$  à la propriété de Markov globale sur  $G$ .

**Démonstration**  $A$ ,  $B$  et  $S$  sont disjoint, donc  $V$  peut être partitionné par  $\tilde{A}$ ,  $\tilde{B}$  et  $S$ , où  $\tilde{A} = A \cup \{a \in V \mid a \text{ et } A \text{ ne sont pas séparés par } S\}$ ,  $\tilde{B} = V \setminus (S \cup \tilde{A})$ . Remarquons que comme  $A$  et  $B$  étaient séparés par  $S$ ,  $\tilde{A}$  et  $\tilde{B}$  le sont aussi par construction. Soit  $C$  une clique du graphe, alors il est impossible que  $C \cap \tilde{A} \neq \emptyset$  et  $C \cap \tilde{B} \neq \emptyset$ ; car sinon, s'il existe  $c_1 \in C \cap \tilde{A}$  et  $c_2 \in C \cap \tilde{B}$ , alors, comme  $C$  est une clique,  $\{c_1, c_2\}$  est une arête du graphe et  $c_2$  n'est donc pas séparé de  $\tilde{A}$  par  $S$  ce qui est une contradiction. Donc  $C \cap \tilde{A} = \emptyset$  ou  $C \cap \tilde{B} = \emptyset$ , ce qui est équivalent à :

$$C \subset \tilde{A} \cup S \quad \text{ou} \quad C \subset \tilde{B} \cup S$$

Alors  $p$  peut être décomposée comme :

$$p(x) = \prod_{C \subset \tilde{A} \cup S} \psi_C(x_C) \prod_{C \subset \tilde{B} \cup S} \psi_C(x_C) := g(x_{\tilde{A} \cup S}) \cdot h(x_{\tilde{B} \cup S}).$$

Or  $A \subset \tilde{A}$  et  $B \subset \tilde{B}$ , on en déduit que

$$\begin{aligned} p(x_A | x_B, x_S) &= \frac{p(x_A, x_B, x_S)}{p(x_B, x_S)} = \frac{g(x_A, x_S)h(x_B, x_S)}{g(x_S)h(x_B, x_S)} \\ &= \frac{g(x_A, x_S)h(x_S)}{g(x_S)h(x_S)} = \frac{p(x_A, x_S)}{p(x_S)} = p(x_A | x_S). \end{aligned}$$

On peut ainsi conclure que  $X_A \perp\!\!\!\perp X_B \mid X_S$ . ■

Une propriété qui est presque la réciproque est admise :

**Théorème 4.18** (Hammersley-Cliford) Soit  $p$  une distribution telle que  $\forall x, p(x) > 0$  alors :

$$p \text{ a la propriété de Markov globale sur } G \Rightarrow p \in \mathcal{L}(G).$$

### 4.3.3 Marginalisation

**Définition 4.19** On appelle « couverture de Markov » d'un noeud  $i$  de  $G$ , le plus petit ensemble  $\mathcal{N}(i)$  de nœuds de  $G$ , tel que  $i$  soit séparé de  $V \setminus \mathcal{N}(i) \setminus \{i\}$  par  $\mathcal{N}(i)$ .

**Remarque** Il s'agit de l'ensemble des voisins de  $i$ .

**Remarque** On peut également définir une notion de couverture de Markov dans les graphes orientés en remplaçant séparé dans la définition ci-dessus par  $d$ -séparé. Dans ce cas, la couverture de Markov de  $i$  ne coïncide généralement plus avec l'ensemble des voisins de  $i$ .

**Proposition 4.20** (Marginalisation) Soit  $G = (V, E)$  un graphe non orienté et  $p \in \mathcal{L}(G)$ , alors  $p(x_1, \dots, x_{n-1})$  se factorise dans le graphe  $G' = (V', E')$  avec :

- $V' = V \setminus \{n\}$
- $E' = (E \cup \mathcal{N}(n) \times \mathcal{N}(n)) \cap (V' \times V')$ .

**Démonstration** Il suffit de séparer les cliques qui contiennent  $n$  :

$$p(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C) = \prod_{n \in C} \psi_C(x_C) \prod_{n \notin C} \psi_C(x_C),$$

$$p(x_1, \dots, x_{n-1}) = \sum_{x_n} p(x) = g(x_{\text{voisins de } n}) \prod_{n \notin C} \psi_C(x_C).$$

■

## 4.4 Comparaisons et relations entre modèles graphiques orientés et non orientés

	Modèles graphiques orientés	Modèles graphiques non-orientés
Définition	$p(x) = \prod_{i=1}^n p(x_i   x_{\pi_i})$	$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$
Séparation	$d$ -séparation	séparation
Marginalisation	non close	close

Soit  $G$  un DAG. Est-il possible de trouver  $G'$  non orienté tel  $\mathcal{L}(G) = \mathcal{L}(G')$ ?  $\mathcal{L}(G) \subset \mathcal{L}(G')$ ?

**Définition 4.21** Soit  $G = (V, E)$  un DAG. Le graphe symétrisé de  $G$  est  $\tilde{G} = (V, \tilde{E})$ , avec  $(u, v) \in \tilde{E}$  si et seulement si  $(u, v) \in E$  ou  $(v, u) \in E$ .

**Définition 4.22** Soit  $G = (V, E)$  un DAG. Le graphe moralisé  $\bar{G}$  de  $G$  est le graphe symétrisé  $\tilde{G}$ , auquel on a rajouté des arêtes de manière à ce que tout couple de noeuds de  $G$  qui sont parents d'un même troisième (dans  $G$  orienté) soit connecté dans  $\bar{G}$ .

On admet les propositions suivantes :

**Proposition 4.23** Soit  $G$  un DAG qui n'admet pas de  $v$ -structure, alors  $\bar{G} = \tilde{G}$  et  $\mathcal{L}(G) = \mathcal{L}(\tilde{G}) = \mathcal{L}(\bar{G})$ .

**Proposition 4.24** Soit  $G$  un DAG, alors  $\mathcal{L}(G) \subset \mathcal{L}(\bar{G})$ .

On peut aussi montrer que  $\bar{G}$  est un élément minimal en nombre d'arêtes dans l'ensemble des graphes  $H$  non orientés tels que  $\mathcal{L}(G) \subset \mathcal{L}(H)$ .



Il existe des structures d'indépendance conditionnelle entre variables aléatoires qui ne peuvent être entièrement spécifiées ni par des modèles graphiques orientés ni par des modèles graphiques non orientés.