

# Coic Quantin Modèles graphiques : DM 2

NAAMANE

Chamsdin

Balme 1

1) Par la définition de l'indépendance conditionnelle,

$$X \perp\!\!\!\perp Y | Z \Leftrightarrow p(x, y | z) = p(x | z) p(y | z) \quad \forall x, y, z \text{ tels que } p(z) > 0$$

$$\Leftrightarrow p(x | y, z) p(y | z) = p(x | z) p(y | z) \quad \forall x, y, z \text{ tels que } p(z) > 0$$

comme d'après l'axiome (b), pour  $x, y, z$  tels que  $p(z) > 0$ ,

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

$$\Leftrightarrow p(x | y, z) \frac{p(y | z)}{\frac{p(y)}{p(z)} > 0} = p(x | z) \frac{p(y | z)}{\frac{p(y)}{p(z)} > 0} \quad \forall x, y, z \text{ tels que } p(z) > 0$$

$$\Leftrightarrow p(x | y, z) = p(x | z) \quad \forall x, y, z \text{ tels que } p(y | z) > 0.$$

Remarque:  $p(y | z) > 0 \Rightarrow p(z) > 0$  car  $p(z) = \sum_{y'} p(y' | z)$  qui est  $> 0$  car

c'est une somme de termes  $\geq 0$  avec un terme  $(p(y | z)) > 0$ .

2) Soit  $G$ :

```

graph TD
    X((X)) --> Z((Z))
    Y((Y)) --> Z((Z))
    Z((Z)) --> T((T))
  
```

,  $p \in \mathcal{L}(G) \Leftrightarrow p(x, y, z, t) = p(x)p(y)p(z|x, y)p(t|z)$

$\forall x, y, z, t$ .

Il n'est pas vrai que  $X \perp\!\!\!\perp Y | T \wedge p \in L(G)$ :

Tout d'abord on utilise le Boyer ball algorithm:

Si on veut faire arriver une balle du noeud  $X$  au noeud  $Y$ :

- On ne peut pas aller en  $Y$  depuis  $X$  en passant par  $Z$  car c'est une structure

de type "explaining array"  et la balle est bloquée en  $Z$

lorsque on ne conditionne pas sur  $Z$ .

- On peut aller en  $T$  depuis  $X$  en passant par  $Z$  car c'est une structure de type

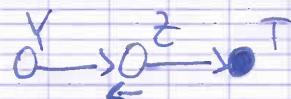
"chaîne de Markov" et la balle peut passer lorsque qu'on ne conditionne pas sur  $Z$ . 

- On peut rebondir de  $Z$  sur  $Z$  en passant par  $T$  car c'est une structure

de type "explaining array" et la balle peut passer lorsque qu'on ne conditionne sur  $T$ .

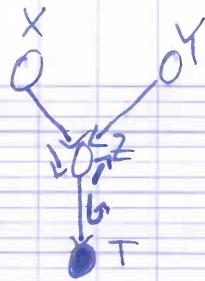
- Enfin la balle peut aller de  $T$  en  $Y$  en passant par  $Z$  car c'est une structure de type

"chaîne de Markov" et on peut passer lorsque qu'on ne conditionne pas sur  $Z$



On a donc pu trouver un moyen du Boyer ball algorithm un chemin allant de  $X$  à  $Y$

sous cette bloqué avant d'arriver en Y :



Ainsi, l'algorithme nous dit que  $\exists p \in \mathcal{L}(G)$  telle que  $X \perp\!\!\!\perp Y \mid T$

Vérifions cette conclusion par un contre-exemple :

Revois l'exemple du cours :



où G est une variable binaire représentant le fait que les personnes sont rentrées ou non

R " " " "

" "

" "

S " " " "

" "

" "

P " " " "

" "

" "

On va montrer que  $P(R=1, S=1 \mid P=1) \neq P(R=1 \mid P=1) \cdot P(S=1 \mid P=1)$

On dispose des probabilités suivantes :  $P(S=1)=0,5$   $P(R=1)=0,2$

$P(G=1 \mid S, R)$	$R=0$	$R=1$
$S=0$	0,01	0,8
$S=1$	0,8	0,95

$P(P=1 \mid G)$	$G=0$	$G=1$
	0,2	0,7

Les calculs seront donnés à une précision de  $10^{-4}$  pour les résultats approximatifs

Calcul de  $P(R=1, S=1 | P=1)$ :

$$P(R=1, S=1 | P=1) = \frac{P(P=1 | R=1, S=1)}{P(P=1)} P(R=1, S=1)$$

$$\begin{aligned} P(P=1 | R=1, S=1) &= P(P=1 | G=1, R=1, S=1) P(G=1 | R=1, S=1) \\ &\quad + P(P=1 | G=0, R=1, S=1) P(G=0 | R=1, S=1) \end{aligned}$$

( $\lambda_1, \rho \perp\!\!\!\perp (R, S) | G$  car  $\rho \in \mathcal{L}(G) \Leftrightarrow p(\lambda_1, \rho, g, p) = p(\lambda_1)p(\rho)p(g|\lambda_1, \rho)p(p|g) \forall \lambda_1, \rho, g, p$ )

$$\begin{aligned} \text{et donc } p(\lambda_1, \rho, g) &= \sum_{p} p(\lambda_1)p(\rho)p(g|\lambda_1, \rho)p(p|g) = p(\lambda_1)p(\rho)p(g|\lambda_1, \rho) \underbrace{\sum_{p} p(p|g)}_{=1} \\ &= p(\lambda_1)p(\rho)p(g|\lambda_1, \rho) \end{aligned}$$

$$\text{d'où } p(p|\lambda_1, \rho, g) = \frac{p(p|\lambda_1, \rho, g)}{p(\lambda_1, \rho, g)} = p(p|g)$$

$$\begin{aligned} \text{Ainsi, } P(P=1 | R=1, S=1) &= P(P=1 | G=1) P(G=1 | R=1, S=1) \\ &\quad + P(P=1 | G=0) P(G=0 | R=1, S=1) \\ &= 0,7 \times 0,95 + 0,2 \times 0,05 = 0,675. \end{aligned}$$

$$P(P=1) = P(P=1 | G=1) P(G=1) + P(P=1 | G=0) P(G=0)$$

$$\begin{aligned} P(G=1) &= P(G=1 | R=1, S=1) P(R=1, S=1) + P(G=1 | R=0, S=0) P(R=0, S=0) \\ &\quad + P(G=1 | R=0, S=1) P(R=0, S=1) + P(G=1 | R=1, S=0) P(R=1, S=0) \end{aligned}$$

Par indépendance de  $R$  et  $S$ :  $P(R=1, S=1) = P(R=1)P(S=1) = 0,2 \times 0,5 = 0,1$

$$P(R=0, S=0) = 0,8 \times 0,5 = 0,4, \quad P(R=0, S=1) = 0,8 \times 0,5 = 0,4$$

$$P(R=1, S=0) = 0,2 \times 0,5 = 0,1$$

$$\text{Dès } P(G=1) = 0,95 \times 0,1 + 0,01 \times 0,4 + 0,8 \times 0,4 + 0,8 \times 0,1 \\ = 0,499$$

$$\text{et } P(G=0) = 1 - P(G=1) = 0,501$$

$$\text{Dès } P(P=1) = 0,7 \times 0,499 + 0,2 \times 0,501 = 0,4495$$

$$\text{Ainsi } P(R=1, S=1 | P=1) = \frac{0,675 \times 0,1}{0,4495} \approx 0,1502$$

Calcul de  $P(R=1 | P=1)$ :

$$P(R=1 | P=1) = \frac{P(P=1 | R=1) P(R=1)}{P(P=1)}$$

$$P(P=1 | R=1) = P(P=1 | G=1, R=1) P(G=1 | R=1) + P(P=1 | G=0, R=1) P(G=0 | R=1)$$

$$\text{Or, } P \perp\!\!\!\perp R | G \text{ car taylor d'après la factorisation du graphe: } p(x_1, g, p) \\ = p(x_1) p(g) p(p|x_1, g)$$

$$p(x_1, g) = \sum_s \sum_p p(x_1) p(s) p(g|x_1, s) p(p|g) = p(x_1) \sum_s p(s) p(g|x_1, s) \sum_p p(p|g) \\ = p(x_1) \sum_s p(s) p(g|x_1, s)$$

$$\text{et } p(p|x_1, g) = \sum_s p(x_1) p(s) p(g|x_1, s) p(p|g) = p(x_1) p(p|g) \sum_s p(s) p(g|x_1, s)$$

$$\text{dès } p(p|x_1, g) = \frac{p(p|x_1, g)}{p(x_1, g)} = p(p|g)$$

$$\text{Ainsi, } P(P=1 | R=1) = P(P=1 | G=1) P(G=1 | R=1) + P(P=1 | G=0) P(G=0 | R=1)$$

$$P(G=1|R=1) = P(G=1|R=1, S=1) P(S=1|R=1) + P(G=1|R=1, S=0) P(S=0|R=1)$$

Puis indépendance de R et S :  $P(S=1|R=1) = P(S=1)$  et  $P(S=0|R=1) = P(S=0)$

$$\text{Donc } P(G=1|R=1) = 0,95 \times 0,5 + 0,8 \times 0,5 = 0,875$$

$$\text{et } P(G=0|R=1) = 1 - 0,875 = 0,125$$

$$\text{Donc } P(P=1|R=1) = 0,7 \times 0,875 + 0,2 \times 0,125 = 0,6375$$

$$\text{Ainsi, } P(R=1|P=1) = \frac{0,6375 \times 0,2}{0,4495} \approx 0,2836$$

Calcul de  $P(S=1|P=1)$ :

$$P(S=1|P=1) = \frac{P(P=1|S=1) P(S=1)}{P(P=1)}$$

$$\begin{aligned} P(P=1|S=1) &= P(P=1|G=1, S=1) P(G=1|S=1) + P(P=1|G=0, S=1) P(G=0|S=1) \\ &= P(P=1|G=1) P(G=1|S=1) + P(P=1|G=0) P(G=0|S=1) \end{aligned}$$

car  $P \perp\!\!\!\perp S | G$  (démonstration similaire à  $P \perp\!\!\!\perp R | G$ )

$$P(G=1|S=1) = P(G=1|R=1, S=1) P(R=1|S=1) + P(G=1|R=0, S=1) P(R=0|S=1)$$

$$= P(G=1|R=1, S=1) P(R=1) + P(G=1|R=0, S=1) P(R=0) \text{ car } R \perp\!\!\!\perp S.$$

$$= 0,95 \times 0,2 + 0,8 \times 0,8 = 0,83$$

$$\text{et } P(G=0|S=1) = 1 - P(G=1|S=1) = 1 - 0,83 = 0,17$$

$$\text{Donc } P(R=1 | S=1) = 0,7 \times 0,83 + 0,2 \times 0,17 = 0,615$$

$$\text{Ainsi, } P(S=1 | R=1) = \frac{0,615 \times 0,15}{0,4495} \approx 0,6841$$

$$\text{Et donc } P(R=1 | P=1) \times P(S=1 | R=1) \approx 0,2836 \times 0,6841 \approx 0,1940$$

Donc  $P(R=1, S=1 | P=1) \neq P(R=1 | P=1) \cdot P(S=1 | P=1)$  et ainsi  $R \perp\!\!\!\perp S | P$ .

PS: On montre juste l'indépendance de R et S qu'on a utilisée tout au long des calculs:

$$p \in \mathcal{P}(G) \Leftrightarrow p(\lambda_1, \alpha, g, \rho) = p(\lambda)p(\alpha)p(g|\lambda, \alpha)p(\rho|g) \quad \forall \lambda, \alpha, g, \rho$$

$$p(\lambda, \alpha) = \sum_g \sum_p p(\lambda, \alpha, g, \rho) = p(\lambda)p(\alpha) \sum_g p(g|\lambda, \alpha) \sum_p p(\rho|g) = p(\lambda)p(\alpha) \underbrace{\sum_g p(g|\lambda, \alpha)}_{=1} \underbrace{\sum_p p(\rho|g)}_{=1} = p(\lambda)p(\alpha)$$

3) (a) Suppose que  $X \perp\!\!\!\perp Y | Z$  et  $X \perp\!\!\!\perp Y$  avec  $Z$  une variable aléatoire bininaire.

$$\text{Soit } x, y, z, \text{ mai: } p(x, y, z) = p(x, y|z)p(z) = p(x|z)p(y|z)p(z)$$

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

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

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

$$\Leftrightarrow \left( \sum_z p(x|z)p(z) \right) \left( \sum_z p(y|z)p(z) \right) = \sum_z p(x|z)p(y|z)p(z)$$

$$\Leftrightarrow (p(x|0)(1-p) + p(x|1)p) (p(y|0)(1-p) + p(y|1)p) = p(x|0)p(y|0)(1-p) + p(x|1)p(y|1)p$$

où  $z \in \{0, 1\}$  et on note  $p(z=1) = p = 1 - p(z=0)$ .

$$\Leftrightarrow p(x|0)p(y|0)(1-p)^2 + p(x|0)p(y|1)p(1-p) + p(x|1)p(y|0)p(1-p) + p(x|1)p(y|1)p^2$$

$$= p(x|0)p(y|0)(1-p) + p(x|1)p(y|1)p$$

$$\Leftrightarrow (p(x|1)p(y|1) + p(x|0)p(y|0) - p(x|0)p(y|1) - p(x|1)p(y|0))p^2$$

$$+ (p(x|0)p(y|1) + p(x|1)p(y|0) - 2p(x|0)p(y|0))p + p(x|0)p(y|0)$$

$$= (p(x|1)p(y|1) - p(x|0)p(y|0))p + p(x|0)p(y|0) \quad (*)$$

On a donc une égalité de polynômes en  $p$ ,  $\forall p \in [0,1]$ . Les deux polynômes sont égaux si et seulement si leurs coefficients sont égaux :

$$(*) \Leftrightarrow \begin{cases} p(x|1)p(y|1) + p(x|0)p(y|0) - p(x|0)p(y|1) - p(x|1)p(y|0) = 0 \\ p(x|0)p(y|1) + p(x|1)p(y|0) - 2p(x|0)p(y|0) = p(x|1)p(y|1) - p(x|0)p(y|0) \end{cases} \quad \begin{matrix} \text{mêmes} \\ \text{conditions} \end{matrix}$$

$$\Leftrightarrow p(x|0)p(y|1) + p(x|1)p(y|0) = p(x|1)p(y|1) + p(x|0)p(y|0)$$

$$\Leftrightarrow p(x|0)(p(y|1) - p(y|0)) + p(x|1)(p(y|0) - p(y|1)) = 0$$

$$\Leftrightarrow (p(y|1) - p(y|0))(p(x|0) - p(x|1)) = 0$$

$$\Leftrightarrow p(y|1) = p(y|0) \quad \text{ou} \quad p(x|0) = p(x|1) \quad (*)$$

$$\Leftrightarrow Y \perp\!\!\!\perp Z \quad \text{ou} \quad X \perp\!\!\!\perp Z.$$

$$(*) \text{ car } p(x|z) = p(x) \text{ pour } z=0,1 \Leftrightarrow \begin{cases} p(x|0) = p(x|0)(1-p) + p(x|1)p \\ p(x|1) = p(x|0)(1-p) + p(x|1)p \end{cases}$$

$$\Leftrightarrow \begin{cases} p(x|0) - p(x|1) = 0 \\ p(x|1) = p(x|0)(1-p) + p(x|1)p \end{cases} \Leftrightarrow \begin{cases} p(x|0) = p(x|1) \\ p(x|1) = p(x|1)(1-p) + p(x|1)p \end{cases} \Leftrightarrow \begin{cases} p(x|0) = p(x|1) \\ p(x|1) = p(x|0) \end{cases}$$

même chose pour  $Y$ .

Ainsi, on a montré que si  $X \perp\!\!\!\perp Y|Z$  et  $X \perp\!\!\!\perp Y$  over  $Z$  une variable aléatoire binaire

alors  $X \perp\!\!\!\perp Z$  et  $Y \perp\!\!\!\perp Z$ .

## Balise 2

$$1) p \in \mathcal{L}(G) \Leftrightarrow p(x) = \prod_{a \in V} p(x_a | x_{\pi_a}) = \prod_{\substack{a \in V \\ i, j \in V}} p(x_a | x_{\pi_a}) \times p(x_i | x_{\pi_i}) \times p(x_j | x_{\pi_j}) \quad \forall x$$

$$p \in \mathcal{L}(G') \Leftrightarrow p(x) = \prod_{a \in V} p(x_a | x_{\pi'_a}) = \prod_{\substack{a \in V \\ i, j \in V}} p(x_a | x_{\pi'_a}) \times p(x_i | x_{\pi'_i}) \times p(x_j | x_{\pi'_j}) \quad \forall x$$

On va montrer que  $p \in \mathcal{L}(G) \Leftrightarrow p \in \mathcal{L}(G')$ .

Tout d'abord, les noeuds différents de  $i, j$  ne sont pas affectés par la transformation entre les deux graphes dans le sens que  $\pi_a = \pi'_a \quad \forall a \neq i, j$

$$\text{Donc } \prod_{\substack{a \in V \\ i, j \in V}} p(x_a | x_{\pi_a}) = \prod_{\substack{a \in V \\ i, j \in V}} p(x_a | x_{\pi'_a}) \quad \forall x.$$

Il nous reste donc à montrer que  $p(x_i | x_{\pi_i}) p(x_j | x_{\pi_j}) = p(x_i | x_{\pi'_i}) p(x_j | x_{\pi'_j}) \quad \forall x$

On a que dans  $G$ ,  $\pi_j = \pi_i \cup \{j\}$  puis après la transformation on a dans  $G'$ ,

$$\pi'_j = \pi_j \setminus \{i\} \quad \text{donc } \pi'_j = \pi_i \cup \{j\} \setminus \{i\} = \pi_i \text{ et } \pi'_i = \pi_i \cup \{j\}$$

$$\begin{aligned} \text{Ainsi, } p(x_i | x_{\pi_i}) p(x_j | x_{\pi_j}) &= p(x_i | x_{\pi_i}) p(x_j | x_{\pi_i \cup \{j\}}) \\ &= \frac{p(x_{\pi_i} | x_i)}{p(x_{\pi_i})} \times \frac{p(x_{\pi_i \cup \{j\}} | x_i, x_j)}{p(x_{\pi_i \cup \{j\}})} = \frac{p(x_{\pi_i} | x_i, x_j)}{p(x_{\pi_i})} = p(x_i, x_j | x_{\pi_i}) \quad \forall x \end{aligned}$$

$$et p(x_i | x_{\pi'_i}) p(x_j | x_{\pi'_j}) = p(x_i | x_{\pi'_i \cup \{j\}}) p(x_j | x_{\pi'_i})$$

$$= \frac{p(x_{\pi'_i} | x_i, x_j)}{p(x_{\pi'_i})} \times \frac{p(x_{\pi'_i \cup \{j\}} | x_i)}{p(x_{\pi'_i \cup \{j\}})} = \frac{p(x_{\pi'_i} | x_i, x_j)}{p(x_{\pi'_i})} = p(x_i, x_j | x_{\pi'_i}) \quad \forall x$$

$$\text{Dès } p(x_i | x_{\pi_i}) p(x_j | x_{\pi_j}) = p(x_i | x_{\pi_i}) p(x_j | x_{\pi_j}) \quad \forall x$$

$$\text{et donc } \prod_{a \in V} p(x_a | x_{\pi_a}) = \prod_{a \in V} p(x_a | x_{\pi_a}) \quad \forall x$$

$$\text{i.e. } p \in \mathcal{L}(G) \Leftrightarrow p \in \mathcal{L}(G').$$

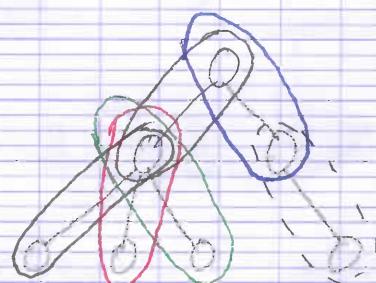
2) Tout d'abord, comme  $G$  est un DAG, on peut supposer sans perte de généralité que les noeuds sont disposés dans un ordre topologique (i.e. si  $x_i$  est un ancêtre de  $x_j$  on a strict order  $i < j$ ). ( $V = \{0, \dots, m\}$ )

$$\text{Si } p \in \mathcal{L}(G) \text{ alors } p \text{ s'écrit: } p(x) = \prod_{i \in V} p(x_i | x_{\pi_i}) \quad \forall x$$

Montrons que  $p \in \mathcal{L}(G')$ .

Comme  $G$  est un ordre orienté, chaque noeud a un unique parent (sauf la racine qui n'en a pas). Ainsi, en enlevant l'orientation des flèches, on a que les cliques maximales de  $G'$  sont respectivement constituées d'un noeud (sauf la racine)

et de son unique parent. Exemple:



la racine

Dès lors on appelle l'ensemble des cliques maximales  $\mathcal{E}_{max} = \{\pi_i; U_i; f_i\} ; i \in V \setminus \{0\}$

$$\text{On pose } \Psi_1(x_1, x_{\pi_1}) = \Psi_1(x_1, x_0) = p(x_0) p(x_1 | x_0)$$

$$\text{et } \psi_i(x_i, x_{\pi_i}) = p(x_i | x_{\pi_i}) \quad \forall i \in V \setminus \{0, 1\}$$

$$\begin{aligned} \text{On pose aussi } z &:= \sum_{x_i} \prod_{i \in V \setminus \{0, 1\}} \psi_i(x_i, x_{\pi_i}) \\ &= \sum_{x_0, x_1, \dots, x_m \text{ tels que}} \underbrace{\prod_{i \in V \setminus \{0, 1\}} \psi_i(x_i, x_{\pi_i})}_{\sum_{x_m} \psi_m(x_m, x_{\pi_m})} \\ &= \sum_{x_m} \underbrace{p(x_m | x_{\pi_m})}_{=1} \end{aligned}$$

$$\begin{aligned} &\stackrel{\text{réécriture}}{=} \dots \\ (\text{grâce à l'ordre topologique}) &= \sum_{x_0} \sum_{x_1} p(x_0) p(x_1 | x_0) = \sum_{x_0} p(x_0) \underbrace{\sum_{x_1} p(x_1 | x_0)}_{=1} \\ &= \sum_{x_0} p(x_0) = 1 \end{aligned}$$

$$\text{Ainsi, } p(x) = \prod_{i \in V} p(x_i | x_{\pi_i}) = \frac{1}{z} \prod_{i \in V \setminus \{0, 1\}} \psi_i(x_i, x_{\pi_i}) \quad \forall x$$

i.e le produit de fractions dépendant respectivement de chaque clique positives

maximal multiplié par une constante de normalisation (ici 1) donc

$$p \in L(G')$$

$$\text{Réciproquement, si } p \in L(G') \text{ alors } p(x) = \frac{1}{z} \prod_{C \in \mathcal{E}} \beta_C(x_C) \quad \forall x$$

$$\text{où } z = \sum_{x_C} \prod_{C \in \mathcal{E}} \beta_C(x_C) \quad \text{où } \mathcal{E} \text{ est l'ensemble des cliques de } G'$$

Remarque: On peut toujours garder l'ordre topologique induit par l'ordre dirigé dans le sens où  $x_0$  est toujours la racine,  $x_m$  une feuille ...

Car dans l'ordre orienté  $G$  chaque noeud a un unique parent

(excepté la racine qui n'a pas), mais donc que alors  $G'$ ,

$$\mathcal{L} = \{j \in V \mid j \in V \setminus \{\pi_i \cup j \mid i \in V \setminus \{\pi_i\}\} \}. Comme dans l'affirmation de p:$$

$$p(x) = \frac{1}{2} \prod_{i \in V \setminus \{\pi_i\}} \psi_0(x_0) \times \prod_{i \in V \setminus \{\pi_i\}} \psi_i(x_i) \times \gamma_i(x_i, x_{\pi_i}) \quad \forall x$$

$$\text{et } Z = \sum_x \psi_0(x_0) \times \prod_{i \in V \setminus \{\pi_i\}} \psi_i(x_i) \times \gamma_i(x_i, x_{\pi_i})$$

$$\text{On va poser } \Lambda_0(x_0) := \psi_0(x_0) \text{ et } \Lambda_i(x_i, x_{\pi_i}) := \psi_i(x_i) \times \gamma_i(x_i, x_{\pi_i}) \quad \forall i \in V \setminus \{\pi_i\}$$

$$\text{alors } p(x) = \frac{1}{2} \prod_{i \in V} \Lambda_i(x_i, x_{\pi_i}) \quad \forall x \quad (\text{puisque } \pi_0 = \emptyset)$$

$$\text{et } Z = \sum_{x_0 \dots x_m} \prod_{i \in V} \Lambda_i(x_i, x_{\pi_i}) = \sum_{x_0 \dots x_m \in V \setminus \{\pi_i\}} \prod_{i \in V} \Lambda_i(x_i, x_{\pi_i}) \sum_{x_m} \Lambda_m(x_m, x_{\pi_m})$$

Réécrivons grâce à l'ordre topologique de  $G$ . (l'autre direction)

$= \dots$

$$= \prod_{i \in V} \sum_{x_i} \Lambda_i(x_i, x_{\pi_i})$$

$$\text{alors } p(x) = \prod_{i \in V} \left( \frac{\Lambda_i(x_i, x_{\pi_i})}{\sum_{x_i} \Lambda_i(x_i, x_{\pi_i})} \right) \quad \forall x$$

$$\text{On pose alors } f_i(x_i, x_{\pi_i}) = \frac{\Lambda_i(x_i, x_{\pi_i})}{\sum_{x_i} \Lambda_i(x_i, x_{\pi_i})}$$

$\forall i \in V \quad f_i > 0 \quad \text{car} \quad \Lambda_i > 0 \quad \text{d'après la définition de la factorisation de polynômes } G'$ .

$$\forall i \in V, \quad \sum_{x_i} f_i(x_i, x_{\pi_i}) = \frac{\sum_{x_i} \Lambda_i(x_i, x_{\pi_i})}{\sum_{x_i} \Lambda_i(x_i, x_{\pi_i})} = 1$$

$$\text{et } p(x) = \prod_{i \in V} f_i(x_i, x_{\pi_i}) \quad \forall x \quad \text{alors } p \in \mathcal{L}(G).$$

Ainsi, on a montré que  $p \in \mathcal{L}(G) \Rightarrow p \in \mathcal{L}(G')$  et  $p \in \mathcal{L}(G') \Rightarrow p \in \mathcal{L}(G)$  donc  $\mathcal{L}(G) = \mathcal{L}(G')$

### Bonnie 3

$$D(p||q) = - \sum_{x \in X} p_x(x) \log p_x(x)$$

Comme  $0 \leq p_x(x) \leq 1$  (et que l'ima  $0 \log 0 = 0$ )  $\forall x \in X$ ,

$$\underbrace{p_x(x)}_{\geq 0} \underbrace{\log p_x(x)}_{\leq 0} \leq 0 \text{ donc } -p_x(x) \log p_x(x) \geq 0 \quad \forall x \in X$$

On a donc une somme de  $|X|=n$  termes positifs donc  $H(X) \geq 0$  et il

$$\text{y a égalité si et seulement si } -p_x(x) \log p_x(x) = 0 \quad \forall x \in X$$

$$\Leftrightarrow p_x(x) \log p_x(x) = 0 \quad \forall x \in X$$

$$\text{Or, } p_x(x) \log(p_x(x)) = 0 \Leftrightarrow p_x(x) = 0 \text{ ou } p_x(x) = 1 \quad (x \in X)$$

$$\text{et comme } \sum_{x \in X} p_x(x) = 1 \text{ alors } H(X) = 0 \Leftrightarrow \exists x^* \in X \text{ tel que } p_x(x^*) = 1$$

$$\text{et } p_x(x) = 0 \quad \forall x \in X \setminus \{x^*\}$$

$$\Leftrightarrow \exists x^* \in X \text{ tel que } x = x^*$$

$\Leftrightarrow X$  est constante presque-sûrement

$$(b) D(p||q) = \sum_{x \in X} p_x(x) \log \left( \frac{p_x(x)}{q(x)} \right) = - \sum_{x \in X} p_x(x) \log (q(x))$$

$$+ \sum_{x \in X} p_x(x) \log (p_x(x))$$

$$= - \sum_{x \in X} p_x(x) \log (q(x)) - H(X)$$

$$\text{Or } q(x) = \frac{1}{|X|} = \frac{1}{n} \quad \forall x \in X$$

$$\text{Dmc } D(p_X || q) = - \sum_{\substack{x \in X \\ \geq 1}} p_X(x) \log\left(\frac{1}{n}\right) - H(X)$$

$$= - \log\left(\frac{1}{n}\right) - H(X) = \log(n) - H(X)$$

Ainsi  $D(p_X || q) = \log(n) - H(X)$ .

$$(c) D(p_X || q) = \log(n) - H(X) \Leftrightarrow H(X) = \log(n) - D(p_X || q) \geq 0$$

donc  $H(X) \leq \log(n)$

$$2) (a) I(X_1, X_2) = \sum_{(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) \log\left(\frac{p_{1,2}(x_1, x_2)}{p_1(x_1)p_2(x_2)}\right)$$

$$= D(p_{1,2} || p_1 p_2) \geq 0$$

$I$  est la divergence de Kullback-Leibler entre les distributions sur  $\mathcal{X}_1 \times \mathcal{X}_2$ ,

$p_{1,2}$  (la distribution jointe de  $(X_1, X_2)$ ) et  $p_1 p_2$  (la distribution formée du produit des distributions marginales de  $X_1$  et  $X_2$ ).

$$(b) I(X_1, X_2) = \sum_{(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) \log\left(\frac{p_{1,2}(x_1, x_2)}{p_1(x_1)p_2(x_2)}\right)$$

$$= \sum_{(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) \log(p_{1,2}(x_1, x_2)) - \sum_{(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) \log(p_1(x_1))$$

$$- \sum_{(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) \log(p_2(x_2))$$

$$-\text{H}(X_1, X_2) = \sum_{x_1 \in \mathcal{X}_1} \log(p_1(x_1)) \underbrace{\sum_{x_2 \in \mathcal{X}_2} p_{1,2}(x_1, x_2)}_{= p_1(x_1)} - \sum_{x_2 \in \mathcal{X}_2} \log(p_2(x_2)) \underbrace{\sum_{x_1 \in \mathcal{X}_1} p_{1,2}(x_1, x_2)}_{= p_2(x_2)}$$

$$= -\text{H}(X_1, X_2) - \sum_{x_1 \in \mathcal{X}_1} p_1(x_1) \log(p_1(x_1)) - \sum_{x_2 \in \mathcal{X}_2} p_2(x_2) \log(p_2(x_2))$$

$$= \text{H}(X_1) + \text{H}(X_2) - \text{H}(X_1, X_2)$$

D'ac,  $\boxed{\text{I}(X_1, X_2) = \text{H}(X_1) + \text{H}(X_2) - \text{H}(X_1, X_2)}$

(c) On a  $p_1$  et  $p_2$  données (donc fixé) donc  $\text{H}(X_1)$  et  $\text{H}(X_2)$  sont fixé.

A  $\text{H}(X_1)$  et  $\text{H}(X_2)$  fixé, comme  $\text{I}(X_1, X_2) \geq 0$ ,  $\text{H}(X_1, X_2) = \text{H}(X_1) + \text{H}(X_2) - \text{I}(X_1, X_2)$

est maximale pour  $\text{I}(X_1, X_2) = 0 \Leftrightarrow D(p_{1,2} \parallel p_1 p_2) = 0$

$\Leftrightarrow p_{1,2} = p_1 p_2$  d'après ce qu'on a vu en cours

$\Leftrightarrow X_1$  et  $X_2$  sont indépendantes

D'ac, avec des marginales  $p_1$  et  $p_2$  données, la distribution  $p_{1,2}$  maximale

en terme d'entropie est la distribution telle que  $p_{1,2} = p_1 p_2$  i.e. telle que  $X_1$  et  $X_2$  sont indépendantes.

## Partie 4

Question (a) après.

(b) On se place dans le cadre où  $X_i | Z_i = k \sim \mathcal{N}(\mu_k, \sigma_k^2 I_d) \quad \forall i=1, \dots, n$   
 $\forall k=1, \dots, K$

L'étape M de l'algorithme correspond à maximiser l'espérance de la log-vraisemblance  
 sous la loi de  $Z|X$ .

$$f(\pi, \mu, \sigma) :=$$

$$\text{à maximiser}, \sum_{i=1}^n \sum_{j=1}^K z_i^j \log(\pi_{j|i}) + \sum_{i=1}^n \sum_{j=1}^K z_i^j \left( \log\left(\frac{1}{(2\pi)^{\frac{d}{2}}}\right) + \log\left(\frac{1}{\sigma_{j|i}^2}\right)^{\frac{1}{2}} \right) - \frac{1}{2} \sum_{j=1}^K (x_i - \mu_{j|i})^T (x_i - \mu_{j|i})$$

$$\text{ou } z_i^j = p_{j|i} (\mathbb{P}_{j|i}(x_i)) = \frac{\pi_{j|i} \mathcal{N}(x_i | \mu_{j|i}, \sigma_{j|i}^2 I_d)}{\sum_{j=1}^K \pi_{j|i} \mathcal{N}(x_i | \mu_{j|i}, \sigma_{j|i}^2 I_d)}$$

où  $\mathcal{N}(x_i | \mu_{j|i}, \sigma_{j|i}^2 I_d)$  est la densité de  $x_i$  d'une  $\mathcal{N}(\mu_{j|i}, \sigma_{j|i}^2 I_d)$

On maximise d'abord par rapport à  $\pi$  sous la contrainte :  $\sum_{j=1}^K \pi_{j|i} = 1$

On écrit le Lagrangien  $F(\pi, \lambda) = f(\pi, \mu, \sigma) + \lambda \left( \sum_{j=1}^K \pi_{j|i} - 1 \right)$  pour  $\lambda \in \mathbb{R}$

$$\frac{\partial F}{\partial \pi_{j|i}} = \sum_{i=1}^n \frac{z_i^j}{\pi_{j|i}} + \lambda = 0 \Leftrightarrow \pi_{j|i} = -\frac{1}{\lambda} \sum_{i=1}^n z_i^j \text{ pour } j=1, \dots, K$$

$$\text{Avec la contrainte, } \sum_{j=1}^K \pi_{j|i} = 1 \Leftrightarrow \sum_{j=1}^K \left( -\frac{1}{\lambda} \sum_{i=1}^n z_i^j \right) = 1 \Leftrightarrow -\frac{1}{\lambda} \sum_{i=1}^n \sum_{j=1}^K z_i^j = 1 \Leftrightarrow -\frac{n}{\lambda} = 1 \Leftrightarrow \lambda = -n$$

$$\text{donc } \pi_{j|i} = \frac{1}{n} \sum_{i=1}^n z_i^j$$

$$\text{Par rapport à } p: \frac{\partial \ell}{\partial p_{j|t}} (\pi_T, p_T, \sigma_T) = 0 \Leftrightarrow \sum_{i=1}^m \frac{z_i^j}{\sigma_{j|t}^2} (x_i - p_{j|t}) = 0$$

$$\Leftrightarrow \sum_{i=1}^m z_i^j (x_i - p_{j|t}) = 0 \Leftrightarrow \sum_{i=1}^m z_i^j x_i = p_{j|t} \sum_{i=1}^m z_i^j$$

$$\Leftrightarrow p_{j|t} = \frac{\sum_{i=1}^m z_i^j x_i}{\sum_{i=1}^m z_i^j} \quad \text{pour } j=1, \dots, K$$

dmc

$$p_{j|t+1} = \frac{\sum_{i=1}^m z_i^j x_i}{\sum_{i=1}^m z_i^j}$$

$$\text{Par rapport à } \sigma: \frac{\partial \ell}{\partial \sigma_{j|t}^2} (\pi_T, p_T, \sigma_T) = 0 \Leftrightarrow -\frac{d}{2} \sum_{i=1}^m \frac{z_i^j}{\sigma_{j|t}^2} + \frac{\sum_{i=1}^m z_i^j}{2\sigma_{j|t}^4} (x_i - p_{j|t})^T (x_i - p_{j|t}) = 0$$

(Remarque:  $\left(\sigma_{j|t}^{-2} \mathbb{I}_d\right)^{-1} = \sigma_{j|t}^{-2} \mathbb{I}_d$ )

$$\Leftrightarrow -d \sum_{i=1}^m \frac{z_i^j}{\sigma_{j|t}^2} + \frac{\sum_{i=1}^m z_i^j}{\sigma_{j|t}^2} (x_i - p_{j|t})^T (x_i - p_{j|t}) = 0$$

$$\Leftrightarrow \sigma_{j|t}^{-2} = \frac{\sum_{i=1}^m z_i^j (x_i - p_{j|t})^T (x_i - p_{j|t})}{d \sum_{i=1}^m z_i^j} \quad \text{pour } j=1, \dots, K$$

dmc

$$\sigma_{j|t+1}^{-2} = \frac{\sum_{i=1}^m z_i^j (x_i - p_{j|t+1})^T (x_i - p_{j|t+1})}{d \sum_{i=1}^m z_i^j}$$

Exercice 4 Les graphiques utilisés pour l'interprétation sont dans l'annexe.

(a) On a testé cinq initialisations aléatoirement. Une fois que l'algorithme a convergé, on obtient toujours un centroide en haut à gauche vers  $(-2.2; 4)$ , un en bas à gauche vers  $(-3.7; -4.2)$ , un en bas à droite vers  $(3.8; 5)$  et un en bas à droite vers  $(3.6; -2.5)$ .

En ce qui concerne la distortion, l'algorithme converge à chaque fois vers une valeur de 1104 environ en moyenne (valeur la plus basse : 1102,54 et valeur la plus haute : 1108,62). Cependant, la convergence n'est pas identique dans chaque cas (cf. graphes en annexe). Soit la convergence est exponentielle et vers 10 itérations, soit la convergence présente une perturbation caractérisée par une déstabilisation de la convergence vers une faible remontée de la distortion avant de reprendre la convergence vers 1104 environ et cela sur une petite vingtaine d'itérations.

Pour finir, on remarque que les groupes de points sont toujours les mêmes mais ils ne sont pas toujours labellisés pareil (i.e. la classe 1 pour une initialisation peut devenir la classe 2 pour une autre initialisation par exemple). Cela est simplement dû au fait de la position du départ de chaque centroide qui est choisie aléatoirement dans tout le

nuage de points.

Par fin, on observe que les points à la frontière des quatre classes (en centre du nuage de points) ne sont pas toujours attribués au même groupe de points.

(d) Tout d'abord, on observe une différence de performance entre l'EM isotropique et l'EM général. En effet, que ce soit sur l'échantillon train ou test la log-vraisemblance finale est toujours meilleure avec l'EM général (EM général: train: -2327,75 test: -2409,05 ; EM isotropique: train: -2645,54 , test: -2692,81). Cela nous semble cohérent. En effet, on peut observer sur le nuage de points qu'il y a quatre groupes de points. Évidemment, les groupes en haut à droite et en bas respectivement à gauche ne présentent clairement pas une dispersion similaire en haut et en bas. Ainsi, une hypothèse de matrice de variance-covariance proportionnelle à  $I_2$  dans chaque classe ne nous semble pas raisonnable.

Ainsi, une hypothèse de matrice de variance-covariance proportionnelle à  $I_2$  dans chaque classe ne nous semble pas raisonnable.

On peut également observer que la classification réalisée par l'EM isotropique est moins bonne que celle réalisée par l'EM général. Autant, l'EM général classe les points dans les groupes qui nous semblent "naturellement" visuellement au vu du nuage de points. Autant, l'EM isotropique ne le fait pas. Par exemple, le groupe

"naturel" en bas à gauche qui forme une bande est coupé par l'EM isotropique qui donne des points de ce groupe ou grappe en bas à droite. (Elle peut s'expliquer par le fait que l'hypothèse qu'on impose à l'EM isotropique la force à classer les points respectivement de telle sorte que chaque grappe ait une dispersion similaire en hauteur et en largeur ce qui le fait appeler la bande en bas à gauche pour rendre le groupe de points d'une telle forme.

Pour finir, on remarque que la performance est légèrement meilleure sur l'échantillon train que sur l'échantillon test pour chaque méthode respectivement. En effet, la log-vraisemblance finale est légèrement meilleure sur l'échantillon train (cf valeurs données à la page précédente). (Elle nous montre tout à fait cohérent du fait que l'on a appris les paramètres avec l'échantillon train. Mais, cela nous montre surtout que l'apprentissage des paramètres a été bon vu que la vraisemblance calculée sur l'échantillon test avec les paramètres appris sur l'échantillon train est proche de celle calculée sur l'échantillon train pour les deux méthodes EM respectivement (chacune à leur niveau de performance). On peut d'ailleurs aussi voir sur les graphiques que sur l'échantillon test, la classification est presque aussi bonne que sur l'échantillon train pour les deux méthodes EM respectivement, mis à part quelques points grisésément mal classés comme par exemple les 2 points en jaune qui devraient naturellement plutôt faire partie du groupe de points rouge dans le cas EM général (sur test set).