Télécom Bretagne
Module F4B101A : Traitement Statistique de l'Information

Modèles de Markov Cachés

François Le Gland
INRIA Rennes et IRMAR
http://www.irisa.fr/aspi/legland/

2014–15

Table des matières

Chapitre 1

Introduction

La classification consiste à décider parmi un nombre *fini* d'hypothèses, décrites par un ensemble *fini* E, au vu d'observations généralement bruitées, dont la distribution dépend de l'hypothèse vraie. Ainsi, si l'hypothèse $X = i$ est vraie pour $i \in E$, alors l'observation Y recueillie a pour distribution

$$\mathbb{P}[Y \in dy \mid X = i] = g_i(y)\, dy \qquad \text{ou bien} \qquad \mathbb{P}[Y = \ell \mid X = i] = b_i^\ell \ ,$$

selon qu'il s'agit d'une observation *numérique* à valeurs dans $\mathbb{R}^d$, ou bien d'une observation *symbolique* à valeurs dans un autre ensemble *fini* O.

On considère ici le problème où l'hypothèse vraie varie au cours du temps, et on souhaite décider, de manière récursive si possible, parmi un nombre *fini* d'hypothèses, au vu d'une suite d'observations généralement bruitées.

Typiquement, on dispose d'une suite $(Y_0, Y_1, \cdots, Y_n)$ d'observations, où chaque observation Y_n est reliée à l'hypothèse X_n par une relation probabiliste (supposée indépendante de l'instant considéré) de la forme

$$\mathbb{P}[Y_n \in dy \mid X_n = i] = g_i(y)\, dy \qquad \text{ou bien} \qquad \mathbb{P}[Y_n = \ell \mid X_n = i] = b_i^\ell \ ,$$

par exemple

$$Y_n = h(X_n) + V_n \ ,$$

avec un bruit additif V_n indépendant de X_n.

Tel qu'il est formulé, le problème de décision, vu aussi comme le problème d'estimation de l'état caché X_n, à partir des observations $(Y_0, Y_1, \cdots, Y_n)$ est en général mal–posé, et il est utile d'introduire un modèle *a priori* qui donne une description probabiliste de la suite $(X_0, X_1, \cdots, X_n)$. On considérera en particulier le cas où les états cachés et la suite des observations forment un modèle de Markov caché.

Dans de nombreux cas, la prise en compte de l'information a priori peut se ramener au problème statique suivant : étant donnés deux variables aléatoires X et Y, qu'apporte le fait d'observer la réalisation $Y = y$ sur la connaissance que l'on a de X ?

On suppose que la variable cachée X prend ses valeurs dans un ensemble *fini* E et que la variable observée Y prend ses valeurs dans un ensemble quelconque F. Par définition, un *estimateur* de X à partir de l'observation de Y est un élément aléatoire $I(Y)$ dans E, où I est une application mesurable définie sur F à valeurs dans E, c'est–à–dire une règle de décision, ou un classifieur, qui fait le choix d'un élément de E pour toute observation. Naturellement $I(Y)$ n'est pas égal à X : une mesure de l'écart entre l'estimateur et la vraie valeur est fournie par la probabilité d'erreur

$$\mathbb{P}[I(Y) \neq X] \ . \tag{1.1}$$

L'estimateur du minimum de la probabilité d'erreur (MPE, pour *minimum probability of error*) de X sachant Y est un estimateur $X_*(Y)$ tel que

$$\mathbb{P}[X_*(Y) \neq X] \leq \mathbb{P}[I(Y) \neq X] \ ,$$

pour tout autre estimateur $I(Y)$. La Proposition 1.1 ci–dessous montre que cet estimateur est obtenu à l'aide de la distribution de probabilité conditionnelle de X sachant $Y = y$, définie à partir de la distribution de probabilité jointe de (X, Y) par la décomposition

$$\mathbb{P}[X = i, Y \in dy] = \mathbb{P}[X = i \mid Y = y] \, \mathbb{P}[Y \in dy] \ . \tag{1.2}$$

Proposition 1.1 *Soit X et Y deux variables aléatoires à valeurs dans l'ensemble fini E et dans F respectivement. L'estimateur MPE de X sachant Y est le maximum a posteriori, i.e.*

$$X_*(y) = \operatorname*{argmax}_{i \in E} \mathbb{P}[X = i \mid Y = y] \ .$$

Preuve. Pour tout classifieur I, on a

$$\begin{aligned}
\mathbb{P}[I(Y) = X \mid Y = y] &= \sum_{i \in E} \mathbb{P}[I(Y) = X, X = i \mid Y = y] \\
&= \sum_{i \in E} \mathbb{P}[I(Y) = X \mid X = i, Y = y] \, \mathbb{P}[X = i \mid Y = y] \\
&= \sum_{i \in E} \mathbf{1}_{(I(y) = i)} \, \mathbb{P}[X = i \mid Y = y] \ ,
\end{aligned}$$

pour tout $y \in F$, de sorte que

$$\begin{aligned}
&\mathbb{P}[X_*(Y) \neq X \mid Y = y] - \mathbb{P}[I(Y) \neq X \mid Y = y] \\
&\quad = \sum_{i \in E} [\, \mathbf{1}_{(I(y) = i)} - \mathbf{1}_{(X_*(y) = i)} \,] \, \mathbb{P}[X = i \mid Y = y] \\
&\quad = \sum_{i \in E} [\, \mathbf{1}_{(I(y) = i)} - \mathbf{1}_{(X_*(y) = i)} \,] \, [\, \mathbb{P}[X = i \mid Y = y] - p_*(y) \,] \ ,
\end{aligned}$$

où

$$p_*(y) = \max_{i \in E} \mathbb{P}[X = i \mid Y = y] \ .$$

Par définition

$$\mathbb{P}[X = i \mid Y = y] - p_*(y) \leq 0 \ ,$$

pour tout $i \in E$, avec égalité pour $i = X_*(y)$, tandis que

$$\mathbf{1}(I(y) = i) - \mathbf{1}(X_*(y) = i) = \mathbf{1}(I(y) = i) \geq 0 \ ,$$

pour tout $i \neq X_*(y)$. On en déduit que

$$\mathbb{P}[X_*(Y) \neq X \mid Y = y] - \mathbb{P}[I(Y) \neq X \mid Y = y] \leq 0 \ ,$$

pour tout $y \in F$, de sorte que

$$\mathbb{P}[X_*(Y) \neq X] - \mathbb{P}[I(Y) \neq X]$$

$$= \int_F [\, \mathbb{P}[X_*(Y) \neq X \mid Y = y] - \mathbb{P}[I(Y) \neq X \mid Y = y] \,] \, \mathbb{P}[Y \in dy] \leq 0 \ ,$$

avec égalité pour $I = X_*$. □

L'objectif de ce cours est de fournir des algorithmes efficaces de calcul des probabilités conditionnelles

$$p_n^i = \mathbb{P}[X_n = i \mid Y_1, \cdots, Y_n] \ ,$$

dans le cas particulier où les états cachés et la suite des observations forment un modèle de Markov caché.

Chapitre 2

Modèles de Markov cachés

On se propose d'étudier le problème de filtrage, c'est–à–dire le problème de l'estimation d'un état caché au vu d'observations bruitées, dans le cas où l'état caché est modélisé par une chaîne de Markov à temps *discret* et espace d'état *fini*.

2.1 Chaînes de Markov à état fini

On considère un espace d'état *fini* E. Une suite $\{X_k\}$ de v.a. à valeurs dans E est une chaîne de Markov si la propriété suivante est vérifiée (propriété de Markov)

$$\mathbb{P}[X_k = i_k \mid X_0 = i_0, \cdots, X_{k-1} = i_{k-1}] = \mathbb{P}[X_k = i_k \mid X_{k-1} = i_{k-1}] \ ,$$

pour tout instant k et toute suite $i_0, \cdots, i_k \in E$.

Cette notion généralise la notion de système dynamique déterministe (machine à état fini, suite récurrente, ou équation différentielle ordinaire) : la distribution de probabilité de l'état présent X_k ne dépend que de l'état immédiatement passé X_{k-1}.

Il résulte de la Proposition 2.1 ci–dessous qu'une chaîne de Markov $\{X_k\}$ est entièrement caractérisée par la donnée

- de la *loi initiale* $\nu = (\nu_i)$

$$\nu_i = \mathbb{P}[X_0 = i] \qquad \text{pour tout } i \in E,$$

- et de la *matrice de transition* $\pi = (\pi_{i,j})$

$$\pi_{i,j} = \mathbb{P}[X_k = j \mid X_{k-1} = i] \qquad \text{pour tout } i, j \in E,$$

  qu'on suppose indépendante de l'instant k (chaîne de Markov *homogène*).

Il suffit donc d'une donnée locale (les probabilités de transition entre deux instants successifs) pour caractériser de façon globale une chaîne de Markov.

Proposition 2.1 *Soit ν une probabilité sur E, et π une matrice markovienne sur E. La* distribution de probabilité *de la chaîne de Markov $\{X_k\}$, de loi initiale ν et de matrice de transition π, est donnée par*

$$\mathbb{P}[X_0 = i_0, \cdots, X_k = i_k] = \nu_{i_0}\, \pi_{i_0,i_1} \cdots \pi_{i_{k-1},i_k} \ ,$$

pour tout instant k, et tout $i_0, \cdots, i_k \in E$.

PREUVE. On conditionne par l'évènement $\{X_0 = i_0, \cdots, X_{k-1} = i_{k-1}\}$ et on applique la propriété de Markov

$$\mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i_k] =$$

$$= \mathbb{P}[X_k = i_k \mid X_0 = i_0, \cdots, X_{k-1} = i_{k-1}]\ \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}]$$

$$= \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}]\ \pi_{i_{k-1},i_k} \ .$$

En itérant cette relation, on obtient le résultat annoncé. □

2.2 Modèles de Markov cachés

On considère ensuite le cas des modèles de Markov *cachés*, ou chaînes de Markov partiellement observées. Dans ce modèle, on n'observe pas directement la suite $\{X_k\}$, mais on dispose d'observations $\{Y_k\}$ à valeurs dans un espace *fini* O, ou dans $\mathbb{R}^d$. On suppose que les observations sont recueillies à travers un canal *sans mémoire*, c'est–à–dire que conditionnellement aux états $\{X_k\}$, les observations $\{Y_k\}$ sont mutuellement indépendantes, et que chaque observation Y_k ne dépend que de l'état X_k au même instant. Cette propriété s'exprime de la façon suivante :

- dans le cas *symbolique*

$$\mathbb{P}[Y_0 = \ell_0, \cdots, Y_n = \ell_n \mid X_0 = i_0, \cdots, X_n = i_n] = \prod_{k=0}^{n} \mathbb{P}[Y_k = \ell_k \mid X_k = i_k] \ ,$$

  pour tout $i_0, \cdots, i_n \in E$, et tout $\ell_0, \cdots, \ell_n \in O$,

- et dans le cas *numérique*

$$\mathbb{P}[Y_0 \in dy_0, \cdots, Y_n \in dy_n \mid X_0 = i_0, \cdots, X_n = i_n] = \prod_{k=0}^{n} \mathbb{P}[Y_k \in dy_k \mid X_k = i_k] \ ,$$

  pour tout $i_0, \cdots, i_n \in E$, et tout $y_0, \cdots, y_n \in \mathbb{R}^d$.

Exemple 2.2 Supposons que les observations $\{Y_k\}$ soient reliées aux états $\{X_k\}$ de la façon suivante

$$Y_k = h(X_k) + V_k \ ,$$

où la suite $\{V_k\}$ est un bruit blanc gaussien de dimension d, de moyenne nulle et de matrice de covariance R inversible, indépendant de la chaîne de Markov $\{X_k\}$.

La fonction h définie sur E à valeurs dans $\mathbb{R}^d$ est caractérisée par la donnée d'une famille *finie* $h = (h_i)$ de vecteurs de $\mathbb{R}^d$, et on a

$$\mathbb{P}[Y_k \in dy \mid X_k = i] = \frac{1}{\sqrt{\det(2\pi\, R)}}\ \exp\big\{\ -\tfrac{1}{2}\,(y-h_i)^*\,R^{-1}\,(y-h_i)\ \big\}\ dy\ .$$

Conditionnellement à $\{X_0 = i_0, \cdots, X_n = i_n\}$, les vecteurs aléatoires $Y_0, \cdots, Y_n$ sont mutuellement indépendants, et chaque Y_k est un vecteur aléatoire gaussien de dimension d, de moyenne h_{i_k} et de matrice de covariance R, de sorte que la propriété de canal sans mémoire est vérifiée.

Il résulte de la Proposition 2.3 ci–dessous qu'un modèle de Markov caché $\{(X_k, Y_k)\}$ est entièrement caractérisé par la donnée

- de la *loi initiale* $\nu = (\nu_i)$

$$\nu_i = \mathbb{P}[X_0 = i] \qquad \text{pour tout } i \in E,$$

- de la *matrice de transition* $\pi = (\pi_{i,j})$

$$\pi_{i,j} = \mathbb{P}[X_k = j \mid X_{k-1} = i] \qquad \text{pour tout } i, j \in E,$$

- et dans le cas *symbolique*, des *probabilités d'émission* $b = (b_i^\ell)$

$$b_i^\ell = \mathbb{P}[Y_k = \ell \mid X_k = i] \qquad \text{pour tout } i \in E, \text{ et tout } \ell \in O,$$

- ou dans le cas *numérique*, des *densités d'émission* $g = (g_i)$

$$g_i(y)\, dy = \mathbb{P}[Y_k \in dy \mid X_k = i] \qquad \text{pour tout } i \in E, \text{ et tout } y \in \mathbb{R}^d.$$

Les probabilités / densités d'émissions sont rangées dans les matrices diagonales

$$B^\ell = \mathrm{diag}(b_i^\ell) \quad \text{pour tout } \ell \in O \qquad \text{et} \qquad G(y) = \mathrm{diag}(g_i(y)) \quad \text{pour tout } y \in \mathbb{R}^d.$$

Il suffit donc d'une donnée locale (les probabilités de transition entre deux instants successifs, et les probabilités / densités d'émission à un instant donné) pour caractériser de façon globale un modèle de Markov caché.

Proposition 2.3 *Dans le cas* symbolique, *la* distribution de probabilité *du modèle de Markov caché* $\{(X_k, Y_k)\}$, *de loi initiale* ν, *de matrice de transition* π, *et de probabilités d'émission* b, *est donnée par*

$$\mathbb{P}[X_0 = i_0, \cdots, X_k = i_k, Y_0 = \ell_0, \cdots, Y_k = \ell_k] =$$

$$= \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-1},i_k}\ b_{i_0}^{\ell_0} \cdots b_{i_k}^{\ell_k}\ ,$$

pour tout instant k, tout $i_0, \cdots, i_k \in E$, et tout $\ell_0, \cdots, \ell_k \in O$.

Dans le cas numérique*, la* distribution de probabilité *du modèle de Markov caché $\{(X_k, Y_k)\}$, de loi initiale ν, de matrice de transition π, et de densités d'émission g, est donnée par*

$$\mathbb{P}[X_0 = i_0, \cdots, X_k = i_k, Y_0 \in dy_0, \cdots, Y_k \in dy_k] =$$

$$= \nu_{i_0}\, \pi_{i_0,i_1} \cdots \pi_{i_{k-1},i_k}\, g_{i_0}(y_0) \cdots g_{i_k}(y_k)\, dy_0 \cdots dy_k \ ,$$

pour tout instant k, tout $i_0, \cdots, i_k \in E$, et tout $y_0, \cdots, y_k \in \mathbb{R}^d$.

Preuve. On considère d'abord le cas *symbolique*. On utilise la formule de Bayes, et la propriété de canal sans mémoire

$$\mathbb{P}[X_0 = i_0, \cdots, X_k = i_k, Y_0 = \ell_0, \cdots, Y_k = \ell_k] =$$

$$= \mathbb{P}[Y_0 = \ell_0, \cdots, Y_k = \ell_k \mid X_0 = i_0, \cdots, X_k = i_k]\, \mathbb{P}[X_0 = i_0, \cdots, X_k = i_k]$$

$$= \mathbb{P}[X_0 = i_0, \cdots, X_k = i_k]\, b_{i_0}^{\ell_0} \cdots b_{i_k}^{\ell_k} \ ,$$

et on conclut en utilisant la Proposition 2.1.

Dans le cas *numérique*, on procède de la même manière

$$\mathbb{P}[X_0 = i_0, \cdots, X_k = i_k, Y_0 \in dy_0, \cdots, Y_k \in dy_k] =$$

$$= \mathbb{P}[Y_0 \in dy_0, \cdots, Y_k \in dy_k \mid X_0 = i_0, \cdots, X_k = i_k]\, \mathbb{P}[X_0 = i_0, \cdots, X_k = i_k]$$

$$= \mathbb{P}[X_0 = i_0, \cdots, X_k = i_k]\, g_{i_0}(y_0) \cdots g_{i_k}(y_k)\, dy_0 \cdots dy_k \ ,$$

et on conclut de la même manière, en utilisant la Proposition 2.1. □

On désigne par $\boldsymbol{M} = (\nu, \pi, b)$ dans le cas *symbolique*, et par $\boldsymbol{M} = (\nu, \pi, g)$ dans le cas *numérique*, les paramètres caractéristiques du modèle, et on s'intéresse aux trois problèmes suivants :

- **Evaluer** le modèle $\boldsymbol{M}$: Il s'agit de calculer *efficacement* la distribution de probabilité de la suite d'observations $(Y_0, \cdots, Y_n)$ (ou *fonction de vraisemblance*) en fonction des paramètres du modèle. La réponse à ce problème est fournie par l'équation *forward* de Baum.
- **Estimer** l'état de la chaîne : Etant donnée une suite d'observations $(Y_0, \cdots, Y_n)$, il s'agit d'estimer de façon récursive l'etat présent X_n (problème de *filtrage*), ou bien d'estimer un état intermédiaire X_k pour $k = 0, \cdots, n$ (problème de *lissage*), ou encore d'estimer globalement la suite d'états $(X_0, \cdots, X_n)$, pour un modèle donné $\boldsymbol{M}$. La réponse aux deux premiers problèmes est fournie par les équations *forward* et *backward* de Baum, qui permettent de calculer la distribution de probabilité conditionnelle de l'état X_k sachant les observations $(Y_0, \cdots, Y_n)$. La réponse au dernier problème est fournie par un algorithme de *programmation dynamique*, l'algorithme de Viterbi, qui permet de maximiser la distribution de probabilité conditionnelle de la suite d'états $(X_0, X_1, \cdots, X_n)$.

- **Identifier** le modèle $\boldsymbol{M}$: Etant donnée une suite d'observations $(Y_0, \cdots, Y_n)$, il s'agit de calculer l'estimateur du *maximum de vraisemblance* pour les paramètres inconnus du modèle. La réponse à ce problème est fournie par les *formules de re–estimation* de Baum–Welch, qui définissent un algorithme itératif pour maximiser la fonction de vraisemblance.

Chapitre 3

Equations forward / backward de Baum

On commence par présenter une première méthode pour calculer la distribution de probabilité des observations $(Y_0, \cdots, Y_n)$.

Proposition 3.1 *La distribution de probabilité des observations* $(Y_0, \cdots, Y_n)$ *est donnée :*

- *dans le cas* symbolique *par*

$$\mathbb{P}[Y_0 = \ell_0, \cdots, Y_n = \ell_n] = \sum_{i_0, \cdots, i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{\ell_0} \cdots b_{i_n}^{\ell_n}\ ,$$

  pour tout $\ell_0, \cdots, \ell_n \in O$,

- *et dans le cas* numérique *par*

$$\mathbb{P}[Y_0 \in dy_0, \cdots, Y_n \in dy_n] =$$

$$= \sum_{i_0, \cdots, i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ g_{i_0}(y_0) \cdots g_{i_n}(y_n)\ dy_0 \cdots dy_n\ ,$$

  pour tout $y_0, \cdots, y_n \in \mathbb{R}^d$.

Preuve. On considère d'abord le cas *symbolique*. On utilise la Proposition 2.3 pour calculer la distribution de probabilité marginale

$$\mathbb{P}[Y_0 = \ell_0, \cdots, Y_n = \ell_n] =$$

$$= \sum_{i_0, \cdots, i_n \in E} \mathbb{P}[X_0 = i_0, \cdots, X_n = i_n, Y_0 = \ell_0, \cdots, Y_n = \ell_n]$$

$$= \sum_{i_0, \cdots, i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{\ell_0} \cdots b_{i_n}^{\ell_n}\ .$$

Dans le cas *numérique*, on procède de la même manière

$$\mathbb{P}[Y_0 \in dy_0, \cdots, Y_n \in dy_n] =$$

$$= \sum_{i_0,\cdots,i_n \in E} \mathbb{P}[X_0 = i_0, \cdots, X_n = i_n, Y_0 \in dy_0, \cdots, Y_n \in dy_n]$$

$$= \sum_{i_0,\cdots,i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ g_{i_0}(y_0) \cdots g_{i_n}(y_n)\ dy_0 \cdots dy_n\ . \qquad \square$$

Remarque 3.2 Cette méthode fournit une première expression pour la distribution de probabilité conditionnelle de la suite des états $(X_0, \cdots, X_n)$ sachant les observations $(Y_0, \cdots, Y_n)$:

- dans le cas *symbolique*

$$\mathbb{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n] = \frac{\nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}}{\displaystyle\sum_{j_0,\cdots,j_n \in E} \nu_{j_0}\ \pi_{j_0,j_1} \cdots \pi_{j_{n-1},j_n}\ b_{j_0}^{Y_0} \cdots b_{j_n}^{Y_n}}\ ,$$

- et dans le cas *numérique*

$$\mathbb{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n] =$$

$$= \frac{\nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ g_{i_0}(Y_0) \cdots g_{i_n}(Y_n)}{\displaystyle\sum_{j_0,\cdots,j_n \in E} \nu_{j_0}\ \pi_{j_0,j_1} \cdots \pi_{j_{n-1},j_n}\ g_{j_0}(Y_0) \cdots g_{j_n}(Y_n)}\ ,$$

et pour la vraisemblance du modèle (obtenue en utilisant la suite des observations $(Y_0, \cdots, Y_n)$ à la place des variables muettes) :

- dans le cas *symbolique*

$$L_n = \sum_{i_0,\cdots,i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}\ ,$$

- et dans le cas *numérique*

$$L_n = \sum_{i_0,\cdots,i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ g_{i_0}(Y_0) \cdots g_{i_n}(Y_n)\ .$$

On en déduit les expressions suivantes pour les distributions non–normalisées :

- dans le cas *symbolique*

$$\mathbb{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n]\ L_n = \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}\ ,$$

- et dans le cas *numérique*

$$\mathbb{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n]\ L_n = \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ g_{i_0}(Y_0) \cdots g_{i_n}(Y_n)\ .$$

Remarque 3.3 Le nombre d'opérations nécessaires pour calculer la distribution de probabilité des observations $(Y_0, \cdots, Y_n)$ à partir des formules données dans la Proposition 3.1 est considérable : pour chaque trajectoire possible $(i_0, \cdots, i_n)$ de la chaîne de Markov, il faut effectuer le produit de $2(n+1)$ termes, et il y a $|E|^{n+1}$ trajectoires possibles différentes. Le nombre total d'opérations élémentaires (additions et multiplications) à effectuer est donc de l'ordre de : $2(n+1)\ |E|^{n+1}$. Ce nombre croît *exponentiellement* avec le nombre n d'observations.

3.1 Equation forward

On introduit la variable *forward* $p_k = (p_k^i)$ vue comme un vecteur–ligne, et définie par

$$p_k^i = \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_k]\ L_k \qquad \text{pour tout } i \in E.$$

Remarque 3.4 La variable forward permet de calculer la distribution de probabilité conditionnelle de l'état présent X_n sachant les observations $(Y_0, \cdots, Y_n)$:

$$\mathbb{P}[X_n = i \mid Y_0, \cdots, Y_n] = \frac{1}{L_n}\ p_n^i \qquad \text{pour tout } i \in E,$$

(en ce sens, p_n est une distribution de probabilité non–normalisée), et la constante de normalisation

$$L_n = \sum_{i \in E} p_n^i\ ,$$

s'interprète comme la vraisemblance du modèle sachant les observations $(Y_0, \cdots, Y_n)$.

Théorème 3.5 *La suite* $\{p_k\}$ *vérifie l'équation récurrente suivante :*

- *dans le cas* symbolique

$$p_k^j = \Big[\sum_{i \in E} p_{k-1}^i\ \pi_{i,j}\ \Big]\ b_j^{Y_k} \qquad \textit{pour tout } j \in E, \tag{3.1}$$

  avec la condition initiale : $p_0^i = \nu_i\ b_i^{Y_0}$ *pour tout* $i \in E$,

- *et dans le cas* numérique

$$p_k^j = \Big[\sum_{i \in E} p_{k-1}^i\ \pi_{i,j}\ \Big]\ g_j(Y_k) \qquad \textit{pour tout } j \in E, \tag{3.2}$$

  avec la condition initiale : $p_0^i = \nu_i\ g_i(Y_0)$ *pour tout* $i \in E$.

Remarque 3.6 Ce résultat énoncé composante–par–composante peut être aussi formulé pour la variable forward vue comme un vecteur–ligne, ce qui donne

- dans le cas *symbolique*

$$p_k = p_{k-1}\ \pi\ B^{Y_k} \qquad \text{et} \qquad p_0 = \nu\ B^{Y_0}\ ,$$

- et dans le cas *numérique*

$$p_k = p_{k-1}\ \pi\ G(Y_k) \qquad \text{et} \qquad p_0 = \nu\ G(Y_0)\ .$$

PREUVE. On considère uniquement le cas *symbolique*. Il résulte de la Remarque 3.2 que

$$\begin{aligned}
&\mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_k]\ L_k = \\
&= \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-2},i}\ \pi_{i,j}\ b_{i_0}^{Y_0} \cdots b_i^{Y_{k-1}}\ b_j^{Y_k} \\
&= \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i \mid Y_0, \cdots, Y_{k-1}]\ L_{k-1}\ \pi_{i,j}\ b_j^{Y_k}\ ,
\end{aligned}$$

pour tout $i, j \in E$ et tout $i_0, \cdots, i_{k-2} \in E$. En sommant par rapport à $i_0, \cdots, i_{k-2} \in E$, on obtient

$$\begin{aligned}
\mathbb{P}[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_k]\ L_k\ &=\ \mathbb{P}[X_{k-1} = i \mid Y_0, \cdots, Y_{k-1}]\ L_{k-1}\ \pi_{i,j}\ b_j^{Y_k} \\
&=\ p_{k-1}^i\ \pi_{i,j}\ b_j^{Y_k}\ ,
\end{aligned}$$

pour tout $i, j \in E$. En sommant ensuite par rapport à $i \in E$, on obtient

$$p_k^j = \sum_{i \in E} [p_{k-1}^i\ \pi_{i,j}]\ b_j^{Y_k}\ ,$$

pour tout $j \in E$, d'où le résultat. □

Remarque 3.7 Le calcul récursif de la variable forward p_n fait seulement intervenir des produits matrice / vecteur, et permet de calculer plus efficacement la distribution de probabilité des observations $(Y_0, \cdots, Y_n)$. Il suffit de $|E|\,(2|E| + 1)$ opérations élémentaires (additions et multiplications) pour passer de l'instant k à l'instant $(k + 1)$. Le nombre total d'opérations élémentaires à effectuer est donc de l'ordre de : $n\ |E|\,(2|E| + 1) + (|E| - 1)$. Ce nombre croît seulement *linéairement* avec le nombre n d'observations.

Au lieu de résoudre d'abord l'équation forward pour la version non–normalisée de la distribution conditionnelle, définie par

$$p_k^i = \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_k]\ L_k \qquad \text{pour tout } i \in E,$$

et d'en déduire ensuite la constante de normalisation (vraisemblance) et la version normalisée de la distribution conditionnelle (filtre), définies par

$$L_k = \sum_{i \in E} p_k^i \qquad \text{et} \qquad \bar{p}_k^i = \frac{p_k^i}{\sum_{j \in E} p_k^j} = \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_k] \qquad \text{pour tout } i \in E,$$

respectivement, il est plus efficace, d'un point de vue *numérique*, de propager directement le filtre.

Proposition 3.8 *La suite* $\{\bar{p}_k\}$ *vérifie l'équation récurrente suivante :*

- *dans le cas* symbolique

$$\bar{p}_k^j = \frac{1}{c_k} \left[\sum_{i \in E} \bar{p}_{k-1}^i \, \pi_{i,j} \right] b_j^{Y_k} \qquad \textit{pour tout } j \in E,$$

avec la condition initiale : $\bar{p}_0^i = \dfrac{1}{c_0} \, \nu_i \, b_i^{Y_0}$ *pour tout* $i \in E$*, où les constantes de normalisation sont définies par*

$$c_k = \sum_{i,j \in E} \bar{p}_{k-1}^i \, \pi_{i,j} \, b_j^{Y_k} \qquad \textit{et} \qquad c_0 = \sum_{i \in E} \nu_i \, b_i^{Y_0} \ ,$$

- *et dans le cas* numérique

$$\bar{p}_k^j = \frac{1}{c_k} \left[\sum_{i \in E} \bar{p}_{k-1}^i \, \pi_{i,j} \right] g_j(Y_k) \qquad \textit{pour tout } j \in E,$$

avec la condition initiale : $\bar{p}_0^i = \dfrac{1}{c_0} \, \nu_i \, g_i(Y_0)$ *pour tout* $i \in E$*, où les constantes de normalisation sont définies par*

$$c_k = \sum_{i,j \in E} \bar{p}_{k-1}^i \, \pi_{i,j} \, g_j(Y_k) \qquad \textit{et} \qquad c_0 = \sum_{i \in E} \nu_i \, g_i(Y_0) \ .$$

Remarque 3.9 Ce résultat énoncé composante–par–composante peut être aussi formulé pour la variable forward normalisée vue comme un vecteur–ligne, ce qui donne

- dans le cas *symbolique*

$$\bar{p}_k = \frac{1}{c_k} \, \bar{p}_{k-1} \, \pi \, B^{Y_k} \qquad \text{et} \qquad \bar{p}_0 = \frac{1}{c_0} \, \nu \, B^{Y_0} \ ,$$

où les constantes de normalisation sont définies par

$$c_k = \bar{p}_{k-1} \, \pi \, b^{Y_k} \qquad \text{et} \qquad c_0 = \nu \, b^{Y_0} \ ,$$

- et dans le cas *numérique*

$$\bar{p}_k = \frac{1}{c_k}\, \bar{p}_{k-1}\, \pi\, G^{Y_k} \qquad \text{et} \qquad \bar{p}_0 = \frac{1}{c_0}\, \nu\, G^{Y_0} \ ,$$

où les constantes de normalisation sont définies par

$$c_k = \bar{p}_{k-1}\, \pi\, g^{Y_k} \qquad \text{et} \qquad c_0 = \nu\, g^{Y_0} \ .$$

PREUVE. On considère uniquement le cas *symbolique* : en utilisant l'équation forward (3.1), on obtient

$$\bar{p}_k^j = \frac{1}{L_k}\, p_k^j \ = \ \frac{1}{L_k}\ \Big[\, \sum_{i\in E} p_{k-1}^i\, \pi_{i,j}\, \Big]\ b_j^{Y_k} = \frac{L_{k-1}}{L_k}\ \Big[\, \sum_{i\in E} \bar{p}_{k-1}^i\, \pi_{i,j}\, \Big]\ b_j^{Y_k} \ ,$$

pour tout $j \in E$, et nécessairement

$$\frac{L_k}{L_{k-1}} = \sum_{i,j\in E} \bar{p}_{k-1}^i\, \pi_{i,j}\, b_j^{Y_k} = c_k \ ,$$

et en utilisant la condition initiale de l'équation forward (3.1), on obtient

$$\bar{p}_0^i = \frac{1}{L_0}\, p_0^i = \frac{1}{L_0}\, \nu_i\, b_i^{Y_0} \ ,$$

pour tout $i \in E$, et nécessairement

$$L_0 = \sum_{i\in E} \nu_i\, b_i^{Y_0} = c_0 \ . \qquad \square$$

Remarque 3.10 La suite $\{\log L_k\}$ vérifie l'équation récurrente suivante, valide dans le cas *symbolique* et dans le cas *numérique*

$$\log L_k = \log L_{k-1} + \log c_k \ ,$$

avec la condition initiale

$$\log L_0 = \log c_0 \ ,$$

et en itérant cette relation, on obtient

$$\log L_n = \sum_{k=0}^{n} \log c_k \ .$$

3.2 Equation backward

Pour tout instant intermédiaire k, antérieur à l'instant final n, on définit

$$q_k^i = \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n]\ L_n \qquad \text{pour tout } i \in E.$$

Remarque 3.11 Cette variable permet de calculer la distribution de probabilité conditionnelle de l'état présent X_k sachant toutes les observations $(Y_0, \cdots, Y_n)$:

$$\mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n] = \frac{1}{L_n}\ q_k^i \qquad \text{pour tout } i \in E,$$

et la constante de normalisation

$$L_n = \sum_{i \in E} q_k^i\ ,$$

s'interprète comme (une autre expression de) la vraisemblance du modèle sachant les observations $(Y_0, \cdots, Y_n)$.

Fixer l'état à l'instant k permet d'effectuer une coupure entre le passé jusqu'à l'instant $(k-1)$ et le futur à partir de l'instant $(k+1)$: dans le cas *symbolique*, il résulte en effet de la Remarque 3.2 que

$$\begin{aligned}
q_k^i \ &= \ \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n]\ L_n \\
&= \sum_{\substack{i_0, \cdots, i_{k-1} \in E \\ i_{k+1}, \cdots, i_n \in E}} \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i, \\
&\qquad\qquad X_{k+1} = i_{k+1}, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n]\ L_n \\
&= \sum_{\substack{i_0, \cdots, i_{k-1} \in E \\ i_{k+1}, \cdots, i_n \in E}} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-1},i}\ \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_{k-1}}^{Y_{k-1}}\ b_i^{Y_k}\ b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n} \\
&= \sum_{i_{k+1}, \cdots, i_n \in E} \Big[\sum_{i_0, \cdots, i_{k-1} \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-1},i}\ b_{i_0}^{Y_0} \cdots b_{i_{k-1}}^{Y_{k-1}}\ b_i^{Y_k}\ \Big] \\
&\qquad\qquad \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n} \\
&= \sum_{i_{k+1}, \cdots, i_n \in E} p_k^i\ \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n} \\
&= p_k^i\ \Big[\sum_{i_{k+1}, \cdots, i_n \in E} \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n}\ \Big]\ ,
\end{aligned}$$

et une expression similaire peut être obtenue dans le cas *numérique*, ce qui justifie d'introduire la variable *backward* $v_k = (v_k^i)$ vue comme un vecteur–colonne, et définie :

- dans le cas *symbolique* par

$$v_k^i = \sum_{i_{k+1},\cdots,i_n \in E} \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n} \qquad \text{pour tout } i \in E,$$

  et en particulier : $v_{n-1}^i = \sum_{j \in E} \pi_{i,j}\ b_j^{Y_n}$ pour tout $i \in E$,

- et dans le cas *numérique* par

$$v_k^i = \sum_{i_{k+1},\cdots,i_n \in E} \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ g_{i_{k+1}}(Y_{k+1}) \cdots g_{i_n}(Y_n) \qquad \text{pour tout } i \in E,$$

  et en particulier : $v_{n-1}^i = \sum_{j \in E} \pi_{i,j}\ g_j(Y_n)$ pour tout $i \in E$.

Remarque 3.12 Conditionnellement à $(X_k = i)$, la suite $X_{k+1}, X_{k+2}, \cdots$ est une chaîne de Markov, de loi initiale $\pi_{i,\bullet}$ (ligne i de la matrice π) — c'est–à–dire que

$$\mathbb{P}[X_{k+1} = j \mid X_k = i] = \pi_{i,j} \qquad \text{pour tout } j \in E,$$

et de matrice de transition π. On déduit alors de la Proposition 3.1 que la distribution de probabilité des observations $(Y_{k+1}, \cdots, Y_n)$ sachant $(X_k = i)$ est donnée :

- dans le cas *symbolique* par

$$\mathbb{P}[Y_{k+1} = \ell_{k+1}, \cdots, Y_n = \ell_n \mid X_k = i] = \sum_{i_{k+1},\cdots,i_n \in E} \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{\ell_{k+1}} \cdots b_{i_n}^{\ell_n}\ ,$$

  pour tout $\ell_{k+1}, \cdots, \ell_n \in O$,

- et dans le cas *numérique* par

$$\mathbb{P}[Y_{k+1} \in dy_{k+1}, \cdots, Y_n \in dy_n \mid X_k = i] =$$

$$= \sum_{i_{k+1},\cdots,i_n \in E} \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ g_{i_{k+1}}(y_{k+1}) \cdots g_{i_n}(y_n)\ dy_{k+1} \cdots dy_n\ ,$$

  pour tout $y_{k+1}, \cdots, y_n \in \mathbb{R}^d$,

ce qui permet d'interpréter la variable *backward* comme la vraisemblance du modèle issu de l'état $X_k = i$ à l'instant k (obtenue en utilisant la suite des observations $(Y_{k+1}, \cdots, Y_n)$ à la place des variables muettes).

Théorème 3.13 *La suite* $\{v_k\}$ *vérifie l'équation récurrente rétrograde suivante :*

- *dans le cas* symbolique

$$v_{k-1}^i = \sum_{j\in E} \pi_{i,j}\ b_j^{Y_k}\ v_k^j \qquad \textit{pour tout } i \in E, \tag{3.3}$$

  avec la condition initiale : $v_n^i = 1$ *pour tout* $i \in E$,

- *et dans le cas* numérique

$$v_{k-1}^i = \sum_{j\in E} \pi_{i,j}\ g_j(Y_k)\ v_k^j \qquad \textit{pour tout } i \in E, \tag{3.4}$$

  avec la condition initiale : $v_n^i = 1$ *pour tout* $i \in E$.

Remarque 3.14 Ce résultat énoncé composante–par–composante peut être aussi formulé pour la variable backward vue comme un vecteur–colonne, ce qui donne

- dans le cas *symbolique*

$$v_{k-1} = \pi\ B^{Y_k}\ v_k \qquad \text{et} \qquad v_n \equiv 1\ ,$$

- et dans le cas *numérique*

$$v_{k-1} = \pi\ G(Y_k)\ v_k \qquad \text{et} \qquad v_n \equiv 1\ .$$

Preuve. On considère uniquement le cas *symbolique*. Avec l'initialisation proposée à l'instant n, l'équation (3.3) permet de retrouver à l'instant $(n-1)$

$$v_{n-1}^i = \sum_{j\in E} \pi_{i,j}\ b_j^{Y_n} \qquad \text{pour tout } i \in E.$$

Par définition

$$\begin{aligned}
v_{k-1}^i &= \sum_{i_k,\cdots,i_n\in E} \pi_{i,i_k}\cdots\pi_{i_{n-1},i_n}\ b_{i_k}^{Y_k}\cdots b_{i_n}^{Y_n} \\
&= \sum_{j\in E}\ \sum_{i_{k+1},\cdots,i_n\in E} \pi_{i,j}\ \pi_{j,i_{k+1}}\cdots\pi_{i_{n-1},i_n}\ b_j^{Y_k}\ b_{i_{k+1}}^{Y_{k+1}}\cdots b_{i_n}^{Y_n} \\
&= \sum_{j\in E} \pi_{i,j}\ b_j^{Y_k}\ \Big[\ \sum_{i_{k+1},\cdots,i_n\in E} \pi_{j,i_{k+1}}\cdots\pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{Y_{k+1}}\cdots b_{i_n}^{Y_n}\ \Big] \\
&= \sum_{j\in E} \pi_{i,j}\ b_j^{Y_k}\ v_k^j\ ,
\end{aligned}$$

pour tout $i \in E$, d'où le résultat. □

Proposition 3.15 *Les équations forward et backward sont* duales *l'une de l'autre :*

$$\sum_{i\in E} p_0^i \; v_0^i = \sum_{i\in E} p_k^i \; v_k^i = \sum_{i\in E} p_n^i = L_n \;, \tag{3.5}$$

pour tout instant k.

PREUVE. On considère uniquement le cas *symbolique*. En utilisant successivement l'équation forward (3.1) et l'équation backward (3.3), on obtient

$$\begin{aligned}\sum_{j\in E} p_k^j \; v_k^j \; &= \; \sum_{j\in E} \Big[\sum_{i\in E} p_{k-1}^i \; \pi_{i,j} \Big] \; b_j^{Y_k} \; v_k^j \\ &= \sum_{i\in E} p_{k-1}^i \; \Big[\sum_{j\in E} \pi_{i,j} \; b_j^{Y_k} \; v_k^j \Big] = \sum_{i\in E} p_{k-1}^i \; v_{k-1}^i \;,\end{aligned}$$

d'où le résultat. □

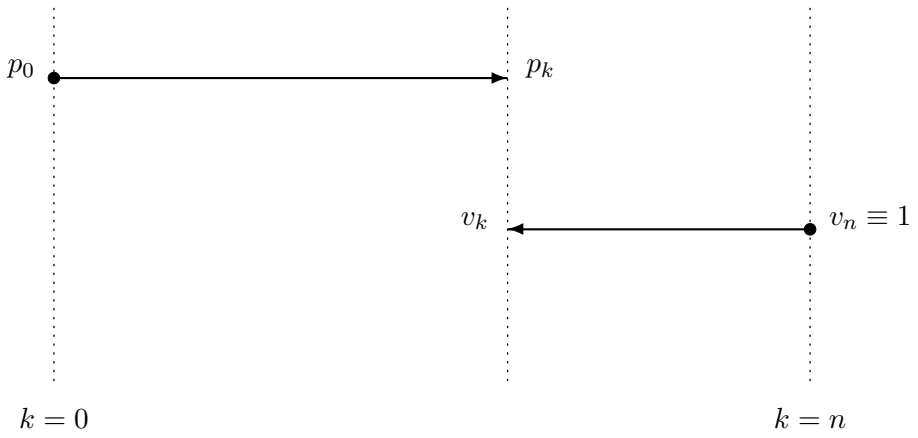

FIGURE 3.1 – Equations forward / backward

Proposition 3.16 *La distribution de probabilité de la* transition (X_{k-1}, X_k) *à un instant intermédiaire sachant les observations* $(Y_0, \cdots, Y_n)$ *jusqu'à l'instant final est donnée :*

- *dans le cas* symbolique *par*

$$\mathbb{P}[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n] = \frac{1}{L_n} \; p_{k-1}^i \; \pi_{i,j} \; b_j^{Y_k} \; v_k^j \qquad \textit{pour tout } i, j \in E,$$

- *et dans le cas* numérique *par*

$$\mathbb{P}[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n] = \frac{1}{L_n} \; p_{k-1}^i \; \pi_{i,j} \; g_j(Y_k) \; v_k^j \qquad \textit{pour tout } i, j \in E.$$

En sommant pour tout $j \in E$ et en utilisant l'équation backward, ou bien en sommant pour tout $i \in E$ et en utilisant l'équation forward, on retrouve le résultat suivant, en terme du produit composante–par–composante des variables forward et backward.

Corollaire 3.17 *La distribution de probabilité de l'état X_k à un instant intermédiaire sachant les observations $(Y_0, \cdots, Y_n)$ jusqu'à l'instant final est donnée par :*

$$\mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n] = \frac{1}{L_n}\, q_k^i \qquad \textit{pour tout } i \in E,$$

avec la définition

$$q_k^i = p_k^i\, v_k^i \qquad \textit{pour tout } i \in E.$$

Remarque 3.18 On vérifie que les constantes de normalisation

$$\sum_{i,j \in E} p_{k-1}^i\, \pi_{i,j}\, b_j^{Y_k}\, v_k^j = \sum_{j \in E} \Big[\sum_{i \in E} p_{k-1}^i\, \pi_{i,j} \Big]\; b_j^{Y_k}\, v_k^j = \sum_{j \in E} p_k^j\, v_k^j = L_n \ ,$$

et

$$\sum_{i \in E} q_k^i = \sum_{i \in E} p_k^i\, v_k^i = L_n \ ,$$

ne dépendent pas de l'instant k considéré, et s'interprètent comme la vraisemblance du modèle sachant les observations $(Y_0, \cdots, Y_n)$.

Preuve de la Proposition 3.16. On considère uniquement le cas *symbolique*. Fixer la transition entre les instants $(k-1)$ et k permet d'effectuer une coupure entre le passé jusqu'à l'instant $(k-2)$ et le futur à partir de l'instant $(k+1)$: il résulte en effet de la Remarque 3.2 que

$$\begin{aligned}
&\mathbb{P}[X_{k-1}=i, X_k=j \mid Y_0,\cdots,Y_n]\; L_n = \\
&= \sum_{\substack{i_0,\cdots,i_{k-2}\in E\\ i_{k+1},\cdots,i_n\in E}} \mathbb{P}[X_0=i_0,\cdots,X_{k-2}=i_{k-2},X_{k-1}=i,\\
&\qquad\qquad X_k=j,X_{k+1}=i_{k+1},\cdots,X_n=i_n \mid Y_0,\cdots,Y_n]\; L_n \\
&= \sum_{\substack{i_0,\cdots,i_{k-2}\in E\\ i_{k+1},\cdots,i_n\in E}} \nu_{i_0}\; \pi_{i_0,i_1}\cdots\pi_{i_{k-2},i}\; \pi_{i,j}\; \pi_{j,i_{k+1}}\cdots\pi_{i_{n-1},i_n}\; b_{i_0}^{Y_0}\cdots b_{i_{k-2}}^{Y_{k-2}}\; b_i^{Y_{k-1}}\; b_j^{Y_k}\; b_{i_{k+1}}^{Y_{k+1}}\cdots b_{i_n}^{Y_n} \\
&= \sum_{i_0,\cdots,i_{k-2}\in E} \nu_{i_0}\; \pi_{i_0,i_1}\cdots\pi_{i_{k-2},i}\; b_{i_0}^{Y_0}\cdots b_{i_{k-2}}^{Y_{k-2}}\; b_i^{Y_{k-1}}\; \pi_{i,j}\; b_j^{Y_k} \\
&\qquad\qquad \Big[\sum_{i_{k+1},\cdots,i_n\in E} \pi_{j,i_{k+1}}\cdots\pi_{i_{n-1},i_n}\; b_{i_{k+1}}^{Y_{k+1}}\cdots b_{i_n}^{Y_n}\Big] \\
&= \sum_{i_0,\cdots,i_{k-2}\in E} \nu_{i_0}\; \pi_{i_0,i_1}\cdots\pi_{i_{k-2},i}\; b_{i_0}^{Y_0}\cdots b_{i_{k-2}}^{Y_{k-2}}\; b_i^{Y_{k-1}}\; \pi_{i,j}\; b_j^{Y_k}\; v_k^j \\
&= \Big[\sum_{i_0,\cdots,i_{k-2}\in E} \nu_{i_0}\; \pi_{i_0,i_1}\cdots\pi_{i_{k-2},i}\; b_{i_0}^{Y_0}\cdots b_{i_{k-2}}^{Y_{k-2}}\; b_i^{Y_{k-1}}\Big]\; \pi_{i,j}\; b_j^{Y_k}\; v_k^j \\
&= p_{k-1}^i\; \pi_{i,j}\; b_j^{Y_k}\; v_k^j\;,
\end{aligned}$$

d'où le résultat. □

Au lieu de résoudre d'abord l'équation forward et l'équation backward séparément, et d'en déduire successivement la version non–normalisée de la distribution conditionnelle, définie par

$$q_k^i = p_k^i\; v_k^i = \mathbb{P}[X_k=i \mid Y_0,\cdots,Y_n]\; L_n \qquad \text{pour tout } i\in E,$$

puis la version normalisée de la distribution conditionnelle (lisseur), définie par

$$\bar{q}_k^i = \frac{q_k^i}{\sum_{j\in E} q_k^j} = \frac{p_k^i\; v_k^i}{\sum_{j\in E} p_k^j\; v_k^j} = \frac{\bar{p}_k^i\; v_k^i}{\sum_{j\in E} \bar{p}_k^j\; v_k^j} = \mathbb{P}[X_k=i \mid Y_0,\cdots,Y_n] \qquad \text{pour tout } i\in E,$$

il est plus efficace, d'un point de vue *numérique*, de propager directement le filtre, comme dans la Proposition 3.8, puis de propager la variable $\bar{v}_k = (\bar{v}_k^i)$ définie par

$$\bar{v}_k^i = \frac{v_k^i}{\sum_{j\in E} \bar{p}_k^j\; v_k^j} \qquad \text{pour tout } i\in E.$$

Remarque 3.19 Avec cette normalisation de la variable backward, la distribution de probabilité conditionnelle de l'état X_k sachant les observations $(Y_0, \cdots, Y_n)$ s'exprime comme

$$\mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n] = \bar{p}_k^i \; \bar{v}_k^i = \bar{q}_k^i \qquad \text{pour tout } i \in E.$$

Proposition 3.20 *La suite $\{\bar{v}_k\}$ vérifie l'équation récurrente rétrograde suivante :*

- *dans le cas* symbolique

$$\bar{v}_{k-1}^i = \frac{1}{c_k} \sum_{j \in E} \pi_{i,j} \; b_j^{Y_k} \; \bar{v}_k^j \qquad \textit{pour tout } i \in E,$$

  avec la condition initiale : $\bar{v}_n^i = 1$ pour tout $i \in E$,

- *et dans le cas* numérique

$$\bar{v}_{k-1}^i = \frac{1}{c_k} \sum_{j \in E} \pi_{i,j} \; g_j(Y_k) \; \bar{v}_k^j \qquad \textit{pour tout } i \in E,$$

  avec la condition initiale : $\bar{v}_n^i = 1$ pour tout $i \in E$,

où les constantes de normalisation sont celles introduites dans l'énoncé de la Proposition 3.8.

Remarque 3.21 Ce résultat énoncé composante–par–composante peut être aussi formulé pour la variable backward normalisée vue comme un vecteur–colonne, ce qui donne

- dans le cas *symbolique*

$$\bar{v}_{k-1} = \frac{1}{c_k} \; \pi \; B^{Y_k} \; \bar{v}_k \qquad \text{et} \qquad \bar{v}_n \equiv 1 \; ,$$

- et dans le cas *numérique*

$$\bar{v}_{k-1} = \frac{1}{c_k} \; \pi \; G(Y_k) \; \bar{v}_k \qquad \text{et} \qquad \bar{v}_n \equiv 1 \; ,$$

où les constantes de normalisation sont celles introduites à la Remarque 3.9.

Preuve. On considère uniquement le cas *symbolique* : en utilisant la relation (3.5), on remarque que

$$\bar{v}_k^i = \frac{v_k^i}{\sum_{j \in E} \bar{p}_k^j \; v_k^j} = \sum_{j \in E} p_k^j \; \frac{v_k^i}{\sum_{j \in E} p_k^j \; v_k^j} = \frac{L_k}{L_n} \; v_k^i \; ,$$

et en utilisant l'équation backward (3.13), on obtient

$$\bar{v}_{k-1}^i = \frac{L_{k-1}}{L_n} \sum_{j \in E} \pi_{i,j} \; b_j^{Y_k} \; v_k^j = \frac{L_{k-1}}{L_n} \; \frac{L_n}{L_k} \sum_{j \in E} \pi_{i,j} \; b_j^{Y_k} \; \bar{v}_k^j \; ,$$

pour tout $i \in E$, d'où le résultat compte tenu que $\dfrac{L_k}{L_{k-1}} = c_k$. □

Remarque 3.22 On remarque que

$$\frac{1}{L_n}\, p_{k-1}^i\, v_k^j = \frac{L_{k-1}}{L_n}\, \bar{p}_{k-1}^i\, \frac{L_n}{L_k}\, \bar{v}_k^j = \frac{1}{c_k}\, \bar{p}_{k-1}^i\, \bar{v}_k^j \qquad \text{pour tout } i,j \in E,$$

et en reportant cette identité dans les expressions obtenues à la Proposition 3.16, on vérifie que la distribution de probabilité conditionnelle de la transition (X_{k-1}, X_k) sachant les observations $(Y_0, \cdots, Y_n)$ s'exprime

- dans le cas *symbolique*, comme

$$\mathbb{P}[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n] = \frac{1}{c_k}\, \bar{p}_{k-1}^i\, \pi_{i,j}\, b_j^{Y_k}\, \bar{v}_k^j\ ,$$

  pour tout $i, j \in E$,

- et dans le cas *numérique*, comme

$$\mathbb{P}[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n] = \frac{1}{c_k}\, \bar{p}_{k-1}^i\, \pi_{i,j}\, g_j(Y_k)\, \bar{v}_k^j\ ,$$

  pour tout $i, j \in E$,

et les constantes de normalisation sont celles introduites dans l'énoncé de la Proposition 3.8.

Chapitre 4

Algorithme de Viterbi

Il résulte de la Remarque 3.4 et du Corollaire 3.17 que les variables forward et backward étudiées au Chapitre 3 permettent de calculer la distribution de probabilité conditionnelle de l'état présent X_n, ou de l'état X_k à un instant intermédiaire, sachant les observations $(Y_0, \cdots, Y_n)$, définies par

$$\mathbb{P}[X_n = i \mid Y_0, \cdots, Y_n] = \frac{1}{L_n}\, p_n^i \qquad \text{pour tout } i \in E,$$

et

$$\mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n] = \frac{1}{L_n}\, q_k^i \qquad \text{pour tout } i \in E,$$

respectivement, où la constante de normalisation

$$L_n = \sum_{i \in E} p_n^i = \sum_{i \in E} p_k^i \; v_k^i = \sum_{i \in E} q_k^i \;,$$

ne dépend pas de l'instant k considéré, et s'interprète comme la vraisemblance du modèle sachant les observations $(Y_0, \cdots, Y_n)$.

Compte tenu que les états possibles pour la chaîne de Markov ne se prêtent pas en général aux opérations *algébriques*, il n'y aurait aucun sens à utiliser ces distributions de probabilités conditionnelles pour calculer des moyennes conditionnelles. D'après la Proposition 1.1, on peut proposer en revanche l'estimateur du *maximum a posteriori*, qui minimise la probabilité de l'erreur d'estimation sachant les observations $(Y_0, \cdots, Y_n)$, défini pour l'état présent par

$$X_n^{\mathrm{LMAP}} = \operatorname*{argmax}_{i \in E} \; \mathbb{P}[X_n = i \mid Y_0, \cdots, Y_n] = \operatorname*{argmax}_{i \in E} \; p_n^i \;,$$

et pour l'état à un instant intermédiaire par

$$X_k^{\mathrm{LMAP}} = \operatorname*{argmax}_{i \in E} \; \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n] = \operatorname*{argmax}_{i \in E} \; q_k^i \;,$$

(en supposant que dans chacun des cas le maximum est atteint en un point unique).

Cependant, il peut arriver que la suite $(X_0^{\mathrm{LMAP}}, \cdots, X_n^{\mathrm{LMAP}})$ ainsi générée soit incohérente avec le modèle, dans le sens suivant : il peut arriver que l'on obtienne $X_{k-1}^{\mathrm{LMAP}} = i$ et $X_k^{\mathrm{LMAP}} = j$ pour deux instants successifs, alors que $\pi_{i,j} = 0$ pour cette même paire (i, j), ce qui signifie que

la transition de l'état i vers l'état j est *impossible* pour le modèle. Pour cette raison, on utilise plutôt un autre estimateur, appelé estimateur *trajectoriel* du *maximum a posteriori*, défini par

$$(X_0^{\mathrm{MAP}}, \cdots, X_n^{\mathrm{MAP}}) = \underset{i_0,\cdots,i_n \in E}{\operatorname{argmax}} \mathbb{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n] \ .$$

et qui minimise la probabilité de l'erreur d'estimation de la suite des états cachés sachant les observations $(Y_0, \cdots, Y_n)$. Il n'est pas possible de calculer cette probabilité pour chacune des $|E|^{n+1}$ suites possibles, ni d'effectuer la maximisation de manière exhaustive. Le calcul efficace de cet estimateur est fourni par un algorithme de programmation dynamique, appelé *algorithme de Viterbi*, qui exploite la remarque suivante.

Remarque 4.1 Si (x_1^*, x_2^*) atteint le maximum de la fonction $f(x_1, x_2)$ définie sur l'ensemble produit $E_1 \times E_2$, alors nécessairement x_1^* et x_2^* atteignent respectivement le maximum des fonctions

$$h(x_1) = f(x_1, x_2^*) \qquad \text{et} \qquad g(x_2) = \max_{x_1 \in E_1} f(x_1, x_2) \ ,$$

définies sur les ensembles E_1 et E_2. Clairement

$$h(x_1^*) = f(x_1^*, x_2^*) \geq f(x_1, x_2^*) = h(x_1) \ ,$$

pour tout $x_1 \in E_1$, et d'autre part

$$g(x_2^*) = \max_{x_1 \in E_1} f(x_1, x_2^*) \geq f(x_1^*, x_2^*) \geq f(x_1, x_2) \ ,$$

pour tout $(x_1, x_2) \in E_1 \times E_2$, et comme la majoration est valide pour tout $x_1 \in E_1$, alors elle reste valide pour le maximum, c'est–à–dire que

$$g(x_2^*) \geq \max_{x_1 \in E_1} f(x_1, x_2) = g(x_2) \ ,$$

pour tout $x_2 \in E_2$.

Si la suite $(i_0^*, \cdots, i_k^*)$ atteint le maximum de la fonction

$$(i_0, \cdots, i_k) \longmapsto \mathbb{P}[X_0 = i_0, \cdots, X_k = i_k \mid Y_0, \cdots, Y_k] \ ,$$

alors nécessairement, d'après la Remarque 4.1, i_k^* atteint le maximum de la fonction

$$i \longmapsto \max_{i_0,\cdots,i_{k-1} \in E} \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i \mid Y_0, \cdots, Y_k] \ ,$$

ce qui justifie d'introduire la fonction *valeur* $V_k = (V_k^i)$ définie par

$$V_k^i = \max_{i_0,\cdots,i_{k-1} \in E} \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i \mid Y_0, \cdots, Y_k] \, L_k \ ,$$

pour tout $i \in E$.

Théorème 4.2 *La suite* $\{V_k\}$ *vérifie l'équation récurrente suivante :*

- *dans le cas* symbolique

$$V_k^j = \max_{i \in E} \left[V_{k-1}^i \, \pi_{i,j} \right] b_j^{Y_k} \qquad \textit{pour tout } j \in E, \tag{4.1}$$

  avec la condition initiale : $V_0^i = \nu_i \, b_i^{Y_0}$ *pour tout* $i \in E$,

- *et dans le cas* numérique

$$V_k^j = \max_{i \in E} \left[V_{k-1}^i \, \pi_{i,j} \right] g_j(Y_k) \qquad \textit{pour tout } j \in E, \tag{4.2}$$

  avec la condition initiale : $V_0^i = \nu_i \, g_i(Y_0)$ *pour tout* $i \in E$.

La suite $\{V_k\}$ *est instrumentale, et permet de définir à chaque instant* k *l'indice*

$$I_{k-1}(j) = \operatorname*{argmax}_{i \in E} \left[V_{k-1}^i \, \pi_{i,j} \right] \qquad \textit{pour tout } j \in E,$$

qui peut s'interpréter comme un pointeur vers un état à l'instant précédent $(k-1)$ *(en supposant que le maximum est atteint en un point unique).*

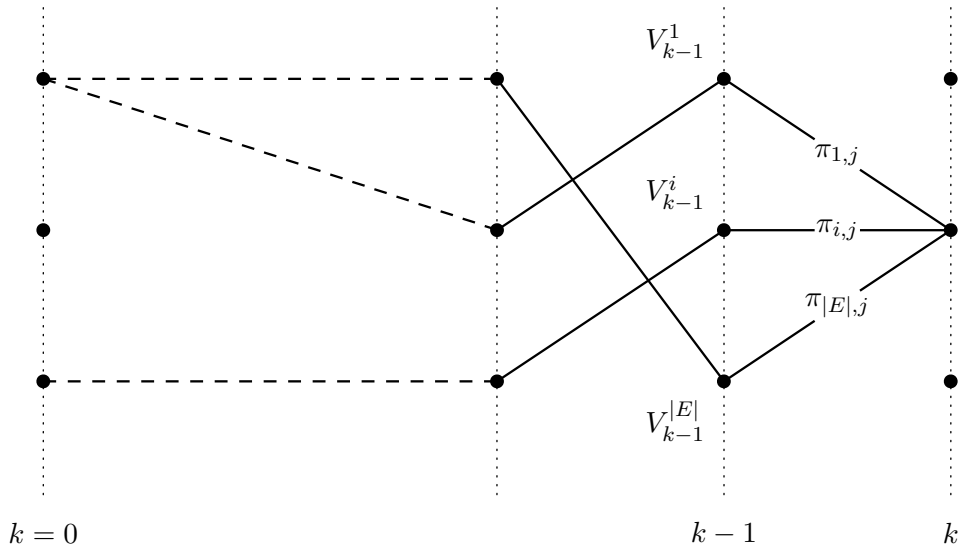

FIGURE 4.1 – Algorithme de Viterbi (programmation dynamique)

PREUVE. On considère uniquement le cas *symbolique*. Il résulte de la Remarque 3.2 que

$$\mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_k]\ L_k =$$

$$= \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-2},i}\ \pi_{i,j}\ b_{i_0}^{Y_0} \cdots b_i^{Y_{k-1}}\ b_j^{Y_k}$$

$$= \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i \mid Y_0, \cdots, Y_{k-1}]\ L_{k-1}\ \pi_{i,j}\ b_j^{Y_k}\ ,$$

pour tout $i, j \in E$ et tout $i_0, \cdots, i_{k-2} \in E$. En maximisant par rapport à $i_0, \cdots, i_{k-2} \in E$, on obtient

$$\max_{i_0,\cdots,i_{k-2} \in E} \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_k]\ L_k =$$

$$= \max_{i_0,\cdots,i_{k-2} \in E} \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i \mid Y_0, \cdots, Y_{k-1}]\ L_{k-1}\ \pi_{i,j}\ b_j^{Y_k}$$

$$= V_{k-1}^i\ \pi_{i,j}\ b_j^{Y_k}\ ,$$

pour tout $i, j \in E$. En maximisant ensuite par rapport à $i \in E$, on obtient

$$V_k^j\ =\ \max_{i \in E} [V_{k-1}^i\ \pi_{i,j}]\ b_j^{Y_k}\ ,$$

pour tout $j \in E$, d'où le résultat. □

Remarque 4.3 Soit $j \in E$ fixé. Si la suite $(i_0^*, \cdots, i_{k-1}^*)$ atteint le maximum de la fonction

$$(i_0, \cdots, i_{k-1}) \longmapsto \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = j \mid Y_0, \cdots, Y_k]\ ,$$

alors nécessairement, d'après la Remarque 4.1, i_{k-1}^* atteint le maximum de la fonction

$$i \longmapsto \max_{i_0,\cdots,i_{k-2} \in E} \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_k]\ ,$$

c'est–à–dire que i_{k-1}^* atteint le maximum de la fonction

$$i \longmapsto V_{k-1}^i\ \pi_{i,j}\ .$$

En d'autres termes, parmi toutes les suites qui aboutissent dans l'état $j \in E$ à l'instant k, la suite de plus grande probabilité, conditionnellement aux observations $(Y_0, \cdots, Y_k)$, est nécessairement passée dans l'état

$$I_{k-1}(j) = \operatorname*{argmax}_{i \in E} [V_{k-1}^i\ \pi_{i,j}]\ ,$$

à l'instant précédent $(k-1)$ (en supposant que le maximum est atteint en un point unique). Ce calcul permet de pré–positionner à tout instant k et pour tout état $j \in E$ un pointeur vers un état $I_{k-1}(j)$ à l'instant précédent $(k-1)$. En outre, on a nécessairement

$$\pi_{I_{k-1}(j),j} > 0\ ,$$

ce qui garantit que la transition de l'état $I_{k-1}(j)$ vers l'état j est possible pour le modèle.

On obtient alors un algorithme efficace pour le calcul de la suite optimale, c'est–à–dire de l'estimateur *trajectoriel* du *maximum a posteriori*.

Théorème 4.4 *La suite* $\{X_k^{\mathrm{MAP}}\}$ *vérifie l'équation récurrente rétrograde suivante :*

$$X_{k-1}^{\mathrm{MAP}} = I_{k-1}(X_k^{\mathrm{MAP}}) \ ,$$

avec la condition initiale

$$X_n^{\mathrm{MAP}} = \underset{i\in E}{\mathrm{argmax}}\ V_n^i \ .$$

Preuve. Si la suite $(i_0^*, \cdots, i_n^*)$ atteint le maximum de la fonction

$$(i_0, \cdots, i_n) \longmapsto \mathbb{P}[X_0 = i_0, \cdots, X_{n-1} = i_{n-1}, X_n = i_n \mid Y_0, \cdots, Y_n] \ ,$$

alors nécessairement, d'après la Remarque 4.1, i_n^* atteint le maximum de la fonction

$$i \longmapsto \max_{i_0,\cdots,i_{n-1}\in E} \mathbb{P}[X_0 = i_0, \cdots, X_{n-1} = i_{n-1}, X_n = i \mid Y_0, \cdots, Y_n] \ ,$$

c'est–à–dire que

$$i_n^* = \underset{i\in E}{\mathrm{argmax}}\, V_n^i \ ,$$

en supposant que le maximum est atteint en un point unique.

Si la suite $(i_0^*, \cdots, i_{k-1}^*, i_k^*, \cdots, i_n^*)$ atteint le maximum de la fonction

$$(i_0, \cdots, i_{k-1}, i_k, \cdots, i_n)$$

$$\longmapsto \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i_k, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n] \ ,$$

alors nécessairement, d'après la Remarque 4.1, la suite $(i_0^*, \cdots, i_{k-1}^*)$ atteint le maximum de la fonction

$$(i_0, \cdots, i_{k-1}) \longmapsto \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i_k^*, \cdots, X_n = i_n^* \mid Y_0, \cdots, Y_n] \ .$$

Dans le cas *symbolique*, il résulte de la Remarque 3.2 que

$$\mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i_k^*, \cdots, X_n = i_n^* \mid Y_0, \cdots, Y_n]\ L_n =$$

$$= \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-2},i_{k-1}}\ \pi_{i_{k-1},i_k^*} \cdots \pi_{i_{n-1}^*,i_n^*}\ b_{i_0}^{Y_0} \cdots b_{i_{k-1}}^{Y_{k-1}}\ b_{i_k^*}^{Y_k} \cdots b_{i_n^*}^{Y_n}$$

$$= \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i_{k-1} \mid Y_0, \cdots, Y_{k-1}]\ L_{k-1}$$

$$\pi_{i_{k-1},i_k^*} \cdots \pi_{i_{n-1}^*,i_n^*}\ b_{i_k^*}^{Y_k} \cdots b_{i_n^*}^{Y_n} \ ,$$

c'est–à–dire que la suite $(i_0^*, \cdots, i_{k-1}^*)$ atteint le maximum de la fonction

$$(i_0, \cdots, i_{k-1}) \longmapsto \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1} \mid Y_0, \cdots, Y_{k-1}]\ \pi_{i_{k-1},i_k^*} \ ,$$

et nécessairement, d'après la Remarque 4.1, i^*_{k-1} atteint le maximum de la fonction

$$i \longmapsto \max_{i_0,\cdots,i_{k-2}\in E} \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i \mid Y_0, \cdots, Y_{k-1}]\ \pi_{i,i^*_k}\ ,$$

c'est–à–dire que

$$i^*_{k-1} = \operatorname*{argmax}_{i\in E} [V^i_{k-1}\ \pi_{i,i^*_k}] = I_{k-1}(i^*_k)\ ,$$

en supposant que le maximum est atteint en un point unique. □

Chapitre 5

Formules de re–estimation de Baum–Welch

Dans les chapitres précédent, l'accent a porté sur l'estimation d'un état caché ou de la suite des états cachés successifs, à partir d'une suite d'observations et pour un modèle connu a priori et caractérisé par la donnée

- de la *loi initiale* $\nu = (\nu_i)$

$$\nu_i = \mathbb{P}[X_0 = i] \qquad \text{pour tout } i \in E,$$

- de la *matrice de transition* $\pi = (\pi_{i,j})$

$$\pi_{i,j} = \mathbb{P}[X_{k+1} = j \mid X_k = i] \qquad \text{pour tout } i, j \in E,$$

- et dans le cas *symbolique*, des *probabilités d'émission* $b = (b_i^\ell)$

$$b_i^\ell = \mathbb{P}[Y_k = \ell \mid X_k = i] \qquad \text{pour tout } i \in E, \text{ et tout } \ell \in O,$$

- ou dans le cas *numérique*, des *densités d'émission* $g = (g_i)$

$$g_i(y)\, dy = \mathbb{P}[Y_k \in dy \mid X_k = i] \qquad \text{pour tout } i \in E, \text{ et tout } y \in \mathbb{R}^d,$$

  par exemple des densités gaussiennes caractérisées par la donnée d'une famille *finie* $h = (h_i)$ de vecteurs de $\mathbb{R}^d$, et d'une famille *finie* $R = (R_i)$ de matrices de covariance inversibles, c'est–à–dire

$$g_i(y) = g(h_i, R_i, y) = \frac{1}{\sqrt{\det(2\pi\, R_i)}}\ \exp\{-\tfrac{1}{2}\,(y - h_i)^*\, R_i^{-1}\,(y - h_i)\}\ ,$$

  pour tout $i \in E$, et tout $y \in \mathbb{R}^d$.

L'objectif ici est d'estimer les paramètres caractéristiques du modèle, à partir d'une suite d'observations, et le point de vue adopté est celui de l'estimation par maximum de vraisemblance.

Dans le cas *symbolique*, il résulte de la Remarque 3.2 que la fonction de vraisemblance du modèle $\boldsymbol{M} = (\nu, \pi, b)$ admet l'expression suivante

$$L_n = \sum_{i_0,\cdots,i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}\ ,$$

et on se propose d'étudier un algorithme itératif pour maximiser la fonction de vraisemblance L_n par rapport aux paramètres (ν, π, b) du modèle. Soit $\boldsymbol{M'} = (\nu', \pi', b')$ un autre modèle, pour lequel on a déja évalué la fonction de vraisemblance

$$L'_n = \sum_{i_0,\cdots,i_n \in E} \nu'_{i_0}\ \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\ b_{i_0}'^{Y_0} \cdots b_{i_n}'^{Y_n}\ ,$$

par exemple en terme des solutions $\{p'_k\}$ et $\{v'_k\}$ des équations forward / backward de Baum pour le modèle $\boldsymbol{M'}$. D'après la Remarque 3.2, le rapport de vraisemblance entre le modèle $\boldsymbol{M}$ et le modèle $\boldsymbol{M'}$ peut s'écrire

$$\begin{aligned}
\frac{L_n}{L'_n} &= \frac{\displaystyle\sum_{i_0,\cdots,i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}}{\displaystyle\sum_{i_0,\cdots,i_n \in E} \nu'_{i_0}\ \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\ b_{i_0}'^{Y_0} \cdots b_{i_n}'^{Y_n}} \\
&= \frac{\displaystyle\sum_{i_0,\cdots,i_n \in E} \nu'_{i_0}\ \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\ b_{i_0}'^{Y_0} \cdots b_{i_n}'^{Y_n}\ [\ \frac{\nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}}{\nu'_{i_0}\ \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\ b_{i_0}'^{Y_0} \cdots b_{i_n}'^{Y_n}}\ ]}{\displaystyle\sum_{i_0,\cdots,i_n \in E} \nu'_{i_0}\ \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\ b_{i_0}'^{Y_0} \cdots b_{i_n}'^{Y_n}} \\
&= \mathbb{E}'[\ \frac{\nu_{X_0}\ \pi_{X_0,X_1} \cdots \pi_{X_{n-1},X_n}\ b_{X_0}^{Y_0} \cdots b_{X_n}^{Y_n}}{\nu'_{X_0}\ \pi'_{X_0,X_1} \cdots \pi'_{X_{n-1},X_n}\ b_{X_0}'^{Y_0} \cdots b_{X_n}'^{Y_n}}\ |\ Y_0, \cdots, Y_n]\ ,
\end{aligned}$$

et compte tenu que la fonction $x \mapsto \log x$ est concave, le logarithme du rapport de vraisemblance est minoré par

$$\begin{aligned}
\log \frac{L_n}{L'_n} &= \log \mathbb{E}'[\ \frac{\nu_{X_0}\ \pi_{X_0,X_1} \cdots \pi_{X_{n-1},X_n}\ b_{X_0}^{Y_0} \cdots b_{X_n}^{Y_n}}{\nu'_{X_0}\ \pi'_{X_0,X_1} \cdots \pi'_{X_{n-1},X_n}\ b_{X_0}'^{Y_0} \cdots b_{X_n}'^{Y_n}}\ |\ Y_0, \cdots, Y_n] \\
&\geq \mathbb{E}'[\log \frac{\nu_{X_0}\ \pi_{X_0,X_1} \cdots \pi_{X_{n-1},X_n}\ b_{X_0}^{Y_0} \cdots b_{X_n}^{Y_n}}{\nu'_{X_0}\ \pi'_{X_0,X_1} \cdots \pi'_{X_{n-1},X_n}\ b_{X_0}'^{Y_0} \cdots b_{X_n}'^{Y_n}}\ |\ Y_0, \cdots, Y_n] = Q_n\ ,
\end{aligned}$$

qui s'annule quand le modèle $\boldsymbol{M}$ coïncide avec le modèle $\boldsymbol{M'}$. Maximiser Q_n par rapport aux paramètres (ν, π, b) du modèle $\boldsymbol{M}$ garantit donc que la vraisemblance du modèle qui atteint le maximum de Q_n sera supérieure à la vraisemblance L'_n du modèle courant $\boldsymbol{M'}$. Les formules de re–estimation de Baum–Welch permettent de trouver explicitement les paramètres du nouveau modèle en fonction des paramètres (ν', π', b') du modèle courant $\boldsymbol{M'}$. En répétant cette procédure, on construit une suite de modèles de vraisemblance croissante, et idéalement cette suite converge vers un modèle qui atteint le maximum de la fonction de vraisemblance.

Théorème 5.1 *Dans le cas* symbolique*, l'algorithme itératif pour l'estimation par maximum de vraisemblance des paramètres du modèle au vu des observations* $(Y_0, \cdots, Y_n)$*, est donné par les formules explicites de re–estimation*

$$\nu_i = \bar{p}_0'^i \; \bar{v}_0'^i \;, \qquad \pi_{i,j} = \pi'_{i,j} \; \frac{\displaystyle\sum_{k=1}^{n} \frac{1}{c'_k} \; \bar{p}_{k-1}'^i \; b_j'^{Y_k} \; \bar{v}_k'^j}{\displaystyle\sum_{k=1}^{n} \bar{p}_{k-1}'^i \; \bar{v}_{k-1}'^i} \qquad \textit{et} \qquad b_i^\ell = \frac{\displaystyle\sum_{k=0}^{n} \mathbf{1}(Y_k = \ell) \; \bar{p}_k'^i \; \bar{v}_k'^i}{\displaystyle\sum_{k=0}^{n} \bar{p}_k'^i \; \bar{v}_k'^i} \;,$$

pour tout $i, j \in E$*, et tout* $\ell \in O$*, où les deux suites* $\{\bar{p}'_k\}$ *et* $\{\bar{v}'_k\}$ *sont les solutions normalisées des équations forward et backward respectivement pour les valeurs* (ν', π', b') *des paramètres.*

Remarque 5.2 Concrètement, si $\boldsymbol{M}_{s-1} = (\nu_{s-1}, \pi_{s-1}, b_{s-1})$ désigne le modèle courant à l'étape $(s-1)$ de l'algorithme, alors

- pour les valeurs $(\nu', \pi', b') = (\nu_{s-1}, \pi_{s-1}, b_{s-1})$ des paramètres, on calcule les solutions normalisées $\{\bar{p}'_k\}$ et $\{\bar{v}'_k\}$ des équations forward et backward définies aux Propositions 3.8 et 3.20 respectivement,
- on calcule les paramètres $(\nu_s, \pi_s, b_s) = (\nu, \pi, b)$ grâce aux formules de re–estimation du Théorème 5.1,

ce qui définit le nouveau modèle $\boldsymbol{M}_s = (\nu_s, \pi_s, b_s)$ à l'étape s de l'algorithme.

Preuve. On remarque que

$$\begin{aligned}
Q_n \;&=\; \mathbb{E}'[\log \frac{\nu_{X_0} \; \pi_{X_0,X_1} \cdots \pi_{X_{n-1},X_n} \; b_{X_0}^{Y_0} \cdots b_{X_n}^{Y_n}}{\nu'_{X_0} \; \pi'_{X_0,X_1} \cdots \pi'_{X_{n-1},X_n} \; b_{X_0}'^{Y_0} \cdots b_{X_n}'^{Y_n}} \mid Y_0, \cdots, Y_n] \\
&=\; \mathbb{E}'[\log \nu_{X_0} + \sum_{k=1}^{n} \log \pi_{X_{k-1},X_k} + \sum_{k=0}^{n} \log b_{X_k}^{Y_k} \mid Y_0, \cdots, Y_n] + \text{cste} \\
&=\; \sum_{i \in E} \mathbb{P}'[X_0 = i \mid Y_0, \cdots, Y_n] \; \log \nu_i \\
&\quad + \sum_{i,j \in E} \Big\{ \sum_{k=1}^{n} \mathbb{P}'[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n] \Big\} \; \log \pi_{i,j} \\
&\quad + \sum_{i \in E} \sum_{\ell \in O} \Big\{ \sum_{k=0}^{n} \mathbf{1}(Y_k = \ell) \; \mathbb{P}'[X_k = i \mid Y_0, \cdots, Y_n] \Big\} \; \log b_i^\ell + \text{cste} \;,
\end{aligned}$$

et en utilisant les expressions obtenues aux Remarques 3.19 et 3.22, on obtient

$$\begin{aligned} Q_n \;=\; & \sum_{i\in E} \bar{p}_0'^{\,i}\; \bar{v}_0'^{\,i}\; \log \nu_i \\ & + \sum_{i,j\in E} \Big\{ \sum_{k=1}^{n} \frac{1}{c'_k}\; \bar{p}_{k-1}'^{\,i}\; \pi'_{i,j}\; b_j'^{\,Y_k}\; \bar{v}_k'^{\,j} \Big\}\; \log \pi_{i,j} \\ & + \sum_{i\in E} \sum_{\ell\in O} \Big\{ \sum_{k=0}^{n} \mathbf{1}(Y_k=\ell)\; \bar{p}_k'^{\,i}\; \bar{v}_k'^{\,i} \Big\}\; \log b_i^{\ell} + \text{cste}\ . \end{aligned}$$

La maximisation par rapport aux paramètres (ν,π,b) sous les contraintes d'égalité

$$\sum_{i\in E} \nu_i = 1\ , \qquad \sum_{j\in E} \pi_{i,j} = 1 \qquad \text{et} \qquad \sum_{\ell\in O} b_i^{\ell} = 1 \qquad \text{pour tout } i\in E,$$

est explicite, et on obtient les formules de re–estimation

$$\nu_i = \bar{p}_0'^{\,i}\; \bar{v}_0'^{\,i}\ , \qquad \pi_{i,j} = \pi'_{i,j}\; \frac{\displaystyle\sum_{k=1}^{n} \frac{1}{c'_k}\; \bar{p}_{k-1}'^{\,i}\; b_j'^{\,Y_k}\; \bar{v}_k'^{\,j}}{\displaystyle\sum_{k=1}^{n} \bar{p}_{k-1}'^{\,i}\; \bar{v}_{k-1}'^{\,i}} \qquad \text{et} \qquad b_i^{\ell} = \frac{\displaystyle\sum_{k=0}^{n} \mathbf{1}(Y_k=\ell)\; \bar{p}_k'^{\,i}\; \bar{v}_k'^{\,i}}{\displaystyle\sum_{k=0}^{n} \bar{p}_k'^{\,i}\; \bar{v}_k'^{\,i}}\ ,$$

pour tout $i,j\in E$, et tout $\ell\in O$. □

Dans le cas *numérique*, il résulte de la Remarque 3.2 que la fonction de vraisemblance du modèle $\boldsymbol{M}=(\nu,\pi,h,R)$ avec des densités d'émission gaussiennes, admet l'expression suivante

$$L_n = \sum_{i_0,\cdots,i_n\in E} \nu_{i_0}\; \pi_{i_0,i_1}\cdots\pi_{i_{n-1},i_n}\; g_{i_0}(Y_0)\cdots g_{i_n}(Y_n)\ ,$$

et on se propose d'étudier un algorithme itératif pour maximiser la fonction de vraisemblance L_n par rapport aux paramètres (ν,π,h,R) du modèle. Soit $\boldsymbol{M'}=(\nu',\pi',h',R')$ un autre modèle, pour lequel on a déja évalué la fonction de vraisemblance

$$L'_n = \sum_{i_0,\cdots,i_n\in E} \nu'_{i_0}\; \pi'_{i_0,i_1}\cdots\pi'_{i_{n-1},i_n}\; g'_{i_0}(Y_0)\cdots g'_{i_n}(Y_n)\ ,$$

par exemple en terme des solutions $\{p'_k\}$ et $\{v'_k\}$ des équations forward / backward de Baum pour le modèle $\boldsymbol{M'}$. En procédant comme dans le cas *symbolique*, le (logarithme du) rapport de vraisemblance entre le modèle $\boldsymbol{M}$ et le modèle $\boldsymbol{M'}$ est minoré par

$$\log\frac{L_n}{L'_n} \geq \mathbb{E}'[\log \frac{\nu_{X_0}\; \pi_{X_0,X_1}\cdots\pi_{X_{n-1},X_n}\; g_{X_0}(Y_0)\cdots g_{X_n}(Y_n)}{\nu'_{X_0}\; \pi'_{X_0,X_1}\cdots\pi'_{X_{n-1},X_n}\; g'_{X_0}(Y_0)\cdots g'_{X_n}(Y_n)} \mid Y_0,\cdots,Y_n] = Q_n\ ,$$

qui s'annule quand le modèle $\boldsymbol{M}$ coïncide avec le modèle $\boldsymbol{M'}$. Maximiser Q_n par rapport aux paramètres (ν,π,h,R) du modèle $\boldsymbol{M}$ garantit donc que la vraisemblance du modèle qui atteint le maximum de Q_n sera supérieure à la vraisemblance L'_n du modèle courant $\boldsymbol{M'}$. Les formules de

re–estimation de Baum–Welch permettent de trouver explicitement les paramètres du nouveau modèle en fonction des paramètres (ν', π', h', R') du modèle courant $\boldsymbol{M}'$. En répétant cette procédure, on construit une suite de modèles de vraisemblance croissante, et idéalement cette suite converge vers un modèle qui atteint le maximum de la fonction de vraisemblance.

Théorème 5.3 *Dans le cas* numérique *avec des densités d'émission gaussiennes, l'algorithme itératif pour l'estimation par maximum de vraisemblance des paramètres du modèle au vu des observations* $(Y_0, \cdots, Y_n)$*, est donné par les formules explicites de re–estimation*

$$\nu_i = \vec{p}_0'^{\,i}\ \bar{v}_0'^{\,i}\ , \qquad \pi_{i,j} = \pi'_{i,j}\ \frac{\displaystyle\sum_{k=1}^{n} \frac{1}{c'_k}\ \vec{p}_{k-1}'^{\,i}\ g'_j(Y_k)\ \bar{v}_k'^{\,j}}{\displaystyle\sum_{k=1}^{n} \vec{p}_{k-1}'^{\,i}\ \bar{v}_{k-1}'^{\,i}}\ ,$$

$$h_i = \frac{\displaystyle\sum_{k=0}^{n} Y_k\ \vec{p}_k'^{\,i}\ \bar{v}_k'^{\,i}}{\displaystyle\sum_{k=0}^{n} \vec{p}_k'^{\,i}\ \bar{v}_k'^{\,i}} \qquad \textit{et} \qquad R_i = \frac{\displaystyle\sum_{k=0}^{n} (Y_k - h_i)\,(Y_k - h_i)^*\ \vec{p}_k'^{\,i}\ \bar{v}_k'^{\,i}}{\displaystyle\sum_{k=0}^{n} \vec{p}_k'^{\,i}\ \bar{v}_k'^{\,i}}\ ,$$

pour tout $i, j \in E$*, où les deux suites* $\{\vec{p}'_k\}$ *et* $\{\bar{v}'_k\}$ *sont les solutions normalisées des équations forward et backward respectivement pour les valeurs* (ν', π', h', R') *des paramètres.*

Remarque 5.4 Concrètement, si $\boldsymbol{M}_{s-1} = (\nu_{s-1}, \pi_{s-1}, h_{s-1}, R_{s-1})$ désigne le modèle courant à l'étape $(s-1)$ de l'algorithme, alors

- pour les valeurs $(\nu', \pi', h', R') = (\nu_{s-1}, \pi_{s-1}, h_{s-1}, R_{s-1})$ des paramètres, on calcule les solutions normalisée $\{\vec{p}'_k\}$ et $\{\bar{v}'_k\}$ des équations forward et backward définies aux Propositions 3.8 et 3.20 respectivement,
- on calcule les paramètres $(\nu_s, \pi_s, h_s, R_s) = (\nu, \pi, h, R)$ grâce aux formules de re–estimation du Théorème 5.3,

ce qui définit le nouveau modèle $\boldsymbol{M}_s = (\nu_s, \pi_s, h_s, R_s)$ à l'étape s de l'algorithme.

PREUVE. On remarque que

$$\begin{aligned} Q_n &= \mathbb{E}'[\log \frac{\nu_{X_0}\ \pi_{X_0,X_1}\cdots\pi_{X_{n-1},X_n}\ g_{X_0}(Y_0)\cdots g_{X_n}(Y_n)}{\nu'_{X_0}\ \pi'_{X_0,X_1}\cdots\pi'_{X_{n-1},X_n}\ g'_{X_0}(Y_0)\cdots g'_{X_n}(Y_n)} \mid Y_0,\cdots,Y_n] \\ &= \mathbb{E}'[\log \nu_{X_0} + \sum_{k=1}^{n} \log \pi_{X_{k-1},X_k} + \sum_{k=0}^{n} \log g_{X_k}(Y_k) \mid Y_0,\cdots,Y_n] + \text{cste} \\ &= \sum_{i\in E} \mathbb{P}'[X_0 = i \mid Y_0,\cdots,Y_n]\ \log \nu_i \\ &\quad + \sum_{i,j\in E} \Big\{ \sum_{k=1}^{n} \mathbb{P}'[X_{k-1} = i, X_k = j \mid Y_0,\cdots,Y_n] \Big\}\ \log \pi_{i,j} \\ &\quad + \sum_{i\in E} \Big\{ \sum_{k=0}^{n} \mathbb{P}'[X_k = i \mid Y_0,\cdots,Y_n]\ \log g_i(Y_k) \Big\} + \text{cste}\ , \end{aligned}$$

et aussi que

$$\begin{aligned} \log g_i(y) &= -\tfrac{1}{2}\ \log\det R_i - \tfrac{1}{2}\ (y-h_i)^*\ R_i^{-1}\ (y-h_i) + \text{cste} \\ &= \tfrac{1}{2}\ \log\det M_i - \tfrac{1}{2}\ \text{trace}[\,(y-h_i)\,(y-h_i)^*\ M_i\,] + \text{cste}\ , \end{aligned}$$

avec $M_i = R_i^{-1}$ pour tout $i \in E$, et tout $y \in \mathbb{R}^d$, et en utilisant les expressions obtenues aux Remarques 3.19 et 3.22, on obtient

$$\begin{aligned} Q_n &= \sum_{i\in E} \bar{p}_0'^{\,i}\ \bar{v}_0'^{\,i}\ \log \nu_i \\ &\quad + \sum_{i,j\in E} \Big\{ \sum_{k=1}^{n} \frac{1}{c'_k}\ \bar{p}_{k-1}'^{\,i}\ \pi'_{i,j}\ g'_j(Y_k)\ \bar{v}_k'^{\,j}\ \Big\}\ \log \pi_{i,j} \\ &\quad + \tfrac{1}{2}\ \sum_{i\in E} \Big\{ \sum_{k=0}^{n} \bar{p}_k'^{\,i}\ \bar{v}_k'^{\,i}\ \Big\}\ \log\det M_i \\ &\quad - \tfrac{1}{2}\ \sum_{i\in E} \text{trace}[\ \Big\{ \sum_{k=0}^{n} \bar{p}_k'^{\,i}\ \bar{v}_k'^{\,i}\ (Y_k - h_i)\ (Y_k - h_i)^*\ \Big\}\ M_i\,] + \text{cste}\ . \end{aligned}$$

On rappelle que la dérivée dans la direction D de l'application

$$M \longmapsto a\ \log\det M - \text{trace}(A\,M)\ ,$$

définie sur l'ensemble des matrices inversibles, est égale à

$$a\,\text{trace}(R\,D) - \text{trace}(A\,D) = \text{trace}[\,(a\,R - A)\,D]\ ,$$

où $R = M^{-1}$ par définition. La maximisation par rapport aux paramètres (ν, π, h, R) sous les contraintes d'égalité

$$\sum_{i\in E} \nu_i = 1 \qquad \text{et} \qquad \sum_{j\in E} \pi_{i,j} = 1 \qquad \text{pour tout } i \in E,$$

est explicite, et on obtient les formules de re–estimation

$$\nu_i = \vec{p}_0'^{\,i}\ \vec{v}_0'^{\,i}\ , \qquad \pi_{i,j} = \pi'_{i,j}\ \frac{\displaystyle\sum_{k=1}^{n} \frac{1}{c'_k}\ \vec{p}_{k-1}'^{\,i}\ g'_j(Y_k)\ \vec{v}_k'^{\,j}}{\displaystyle\sum_{k=1}^{n} \vec{p}_{k-1}'^{\,i}\ \vec{v}_{k-1}'^{\,i}}\ ,$$

$$h_i = \frac{\displaystyle\sum_{k=0}^{n} Y_k\ \vec{p}_k'^{\,i}\ \vec{v}_k'^{\,i}}{\displaystyle\sum_{k=0}^{n} \vec{p}_k'^{\,i}\ \vec{v}_k'^{\,i}} \qquad \text{et} \qquad R_i = \frac{\displaystyle\sum_{k=0}^{n} (Y_k - h_i)\,(Y_k - h_i)^*\ \vec{p}_k'^{\,i}\ \vec{v}_k'^{\,i}}{\displaystyle\sum_{k=0}^{n} \vec{p}_k'^{\,i}\ \vec{v}_k'^{\,i}}\ ,$$

pour tout $i, j \in E$. □