Modèles de Markov Cachés

François Le Gland
INRIA Rennes et IRMAR

http://www.irisa.fr/aspi/legland/rennes-1/

classification bayésienne

modèles de Markov cachés

équations forward / backward de Baum

algorithme de Viterbi

re–estimation de Baum–Welch

détection de changement

objectif : *classifier*, i.e. décider parmi un nombre *fini* d'hypothèses, décrites par un ensemble *fini* noté E, au vu d'observations généralement bruitées, dont la distribution dépend de l'hypothèse vraie

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* noté 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 $(Y_0, \cdots, Y_n)$ généralement bruitées

chaque observation Y_k est reliée à l'hypothèse X_k par une relation probabiliste (supposée indépendante de l'instant considéré) de la forme

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

par exemple

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

avec un bruit additif V_k indépendant de X_k

il est utile d'introduire un modèle *a priori* qui donne une description probabiliste de la suite d'hypothèses $(X_0, \cdots, X_n)$

cadre *statique* : étant donné deux variables aléatoires X et Y, dont la loi jointe est connue, qu'apporte le fait d'observer la réalisation $Y = y$ sur la connaissance que l'on a de X ?

on suppose ici 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 noté $I(Y)$ appartenant à E, c'est–à–dire une règle de décision, ou un *classifieur*, qui fait pour toute observation le choix d'un élément de E

une mesure de l'écart entre l'estimateur et la vraie valeur est fournie par la probabilité d'erreur

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

et on définit l'estimateur du minimum de la probabilité d'erreur (MPE, pour *minimum probability of error*) comme l'estimateur $X_*(Y)$ tel que

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

pour tout autre estimateur $I(Y)$

cet estimateur défini de manière implicite peut être obtenu de manière plus explicite, à l'aide de la distribution de probabilité conditionnelle de X sachant $Y = y$, définie elle–même à partir de la distribution de probabilité jointe de (X, Y) par la décomposition

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

Proposition 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

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

pour tout $y \in F$

de sorte que

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

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

$$= \sum_{i \in E} [\, 1(I(y) = i) - 1(X_*(y) = i) \,] \, \mathbb{P}[X = i \mid Y = y]$$

$$= \sum_{i \in E} [\, 1(I(y) = i) - 1(X_*(y) = i) \,] \, [\, \mathbb{P}[X = i \mid Y = y] - p_*(y) \,]$$

où

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

en effet, on a juste retranché le terme

$$\sum_{i \in E} [\, 1(I(y) = i) - 1(X_*(y) = i) \,] \, p_*(y) = 0$$

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

$$1(I(y) = i) - 1(X_*(y) = i) = 1(I(y) = i) \geq 0$$

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

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

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

retour au cadre *dynamique* : on rappelle qu'il s'agit d'estimer l'hypothèse X_n (ou l'hypothèse X_k à un instant intermédiaire) au vu des observations $(Y_0, \cdots, Y_n)$

un des objectifs de ce cours est de fournir des algorithmes efficaces pour le calcul des probabilités conditionnelles

$$p_n^i = \mathbb{P}[X_n = i \mid Y_0, \cdots, Y_n] \qquad \text{ou} \qquad q_k^i = \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n]$$

dans le cas particulier où la suite des hypothèses et des observations forme un modèle de Markov caché

la première étape consiste à introduire et à étudier la notion de modèle de Markov caché

classification bayésienne

modèles de Markov cachés

- chaînes de Markov à espace d'état fini
- modèles de Markov cachés

équations forward / backward de Baum

algorithme de Viterbi

re–estimation de Baum–Welch

détection de changement

chaînes de Markov à espace d'état fini

une suite $\{X_k\}$ de v.a. à valeurs dans un ensemble fini 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 $i_0, \cdots, i_k \in E$
Remarque 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}

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 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 toute suite $i_0, \cdots, i_k \in E$

Preuve on conditionne par l'évènement $\{X_0 = i_0, \cdots, X_{k-1} = i_{k-1}\}$, on utilise la formule de Bayes 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 □

modèles de Markov cachés

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 $\mathbb{R}^d$, recueillies à travers un canal *sans mémoire*, c'est–à–dire que conditionnellement aux états $\{X_k\}$, (i) les observations $\{Y_k\}$ sont mutuellement indépendantes, et (ii) 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 toute suite $i_0, \cdots, i_n \in E$, et toute suite $\ell_0, \cdots, \ell_n \in O$
- et dans le cas *numérique*

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

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

Exemple 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

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

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 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* $g = (g_i)$

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

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 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, toute suite $i_0, \cdots, i_k \in E$, et toute suite $\ell_0, \cdots, \ell_k \in O$

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, toute suite $i_0, \cdots, i_k \in E$, et toute suite $y_0, \cdots, y_k \in \mathbb{R}^d$

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

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

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

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

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 en utilisant la caractérisation des chaînes de Markov 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 :

- Évaluer le modèle : 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 $\boldsymbol{M}$
  $\longrightarrow$ la réponse à ce problème est fournie par l'équation *forward* de Baum
- Identifier le modèle : étant 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 $\boldsymbol{M}$
  $\longrightarrow$ 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

- Estimer l'état de la chaîne : étant 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 $\boldsymbol{M}$
  $\longrightarrow$ 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 $(Y_0, \cdots, Y_n)$
  $\longrightarrow$ 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 $(X_0, \cdots, X_n)$ sachant les observations $(Y_0, \cdots, Y_n)$

classification bayésienne

modèles de Markov cachés

équations forward / backward de Baum

- équation forward
- équation forward
- équation backward

algorithme de Viterbi

re–estimation de Baum–Welch

détection de changement

méthode élémentaire

on commence par présenter une première méthode (élémentaire mais inefficace) pour calculer la distribution de probabilité des observations $(Y_0, \cdots, Y_n)$

Proposition 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 toute suite $\ell_0, \cdots, \ell_n \in O$

- 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 toute suite $y_0, \cdots, y_n \in \mathbb{R}^d$

Preuve on se ramène à la distribution de probabilité jointe des états cachés et des observations, puis on marginalise : on considère d'abord le cas *symbolique*

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

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

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

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 cette méthode élémentaire 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 identités suivantes : 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 le nombre d'opérations nécessaires pour calculer la distribution de probabilité des observations $(Y_0, \cdots, Y_n)$ à partir de cette méthode élémentaire 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

équation forward

on définit la variable *forward* $p_k = (p_k^i)$ — vue comme un *vecteur–ligne* — 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 la variable forward permet de calculer la distribution de probabilité conditionnelle de l'état présent X_k sachant les observations $(Y_0, \cdots, Y_k)$:

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

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

$$L_k = \sum_{i \in E} p_k^i$$

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

Théorème la suite $\{p_k\}$ vérifie l'équation récurrente suivante :

- dans le cas *symbolique*

$$p_k^j = \Big[\sum_{i \in E} p_{k-1}^i \, \pi_{i,j} \Big] \, b_j^{Y_k} \qquad \text{pour tout } j \in E \qquad \text{(F-sym)}$$

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

- et dans le cas *numérique*

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

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

Remarque ce résultat énoncé composante–par–composante peut être aussi formulé pour la variable forward vue comme un *vecteur–ligne*

- 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*

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

avec les matrices *diagonales* définies par

$$B^\ell = \mathrm{diag}(b_i^\ell) \qquad \text{pour tout } \ell \in O$$

et

$$G(y) = \mathrm{diag}(g_i(y)) \qquad \text{pour tout } y \in \mathbb{R}^d$$

respectivement

Preuve on considère uniquement le cas *symbolique* : on rappelle l'identité suivante, obtenue avec la méthode élémentaire

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

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

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

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

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

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

pour tout $i, j \in E$, et il suffit de sommer par rapport à $i \in E$ $\square$

Remarque 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

mise en œuvre numérique

au lieu de résoudre d'abord l'équation forward pour la version non–normalisée de la distribution conditionnelle, définie à tout instant k comme

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

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

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

il est plus efficace, d'un point de vue *numérique*, de propager directement la log–vraisemblance et le filtre

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

- dans le cas *symbolique*

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

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

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

- et dans le cas *numérique*

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

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

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

Remarque ce résultat énoncé composante–par–composante peut être aussi formulé pour la variable forward normalisée vue comme un *vecteur–ligne*

- 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

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

Remarque la suite $\{\log L_k\}$ vérifie l'équation récurrente suivante : 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

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

Preuve on considère uniquement le cas *symbolique* : en utilisant l'équation forward (F-sym), on obtient

$$\bar{p}^j_k = \frac{1}{L_k}\, p^j_k \;=\; \frac{1}{L_k}\;\Big[\sum_{i \in E} p^i_{k-1}\,\pi_{i,j}\,\Big]\; b_j^{Y_k} = \frac{L_{k-1}}{L_k}\;\Big[\sum_{i \in E} \bar{p}^i_{k-1}\,\pi_{i,j}\,\Big]\; b_j^{Y_k}$$

pour tout $j \in E$, et nécessairement

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

en utilisant la condition initiale de l'équation forward (F-sym)

$$\bar{p}^i_0 = \frac{1}{L_0}\, p^i_0 = \frac{1}{L_0}\,\nu_i\; b_i^{Y_0}$$

pour tout $i \in E$, et nécessairement

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

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

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

avec la constante de normalisation

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

Remarque 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*

$$q_k^i = \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n]\, L_n$$

$$= \sum_{\substack{i_0, \cdots, i_{k-1} \in E \\ i_{k+1}, \cdots, i_n \in E}} \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i, X_{k+1} = i_{k+1}, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n]\, L_n$$

$$= \sum_{\substack{i_0, \cdots, i_{k-1} \in E \\ i_{k+1}, \cdots, i_n \in E}} \nu_{i_0}\, \pi_{i_0, i_1} \cdots \pi_{i_{k-1}, i}\, \pi_{i, i_{k+1}} \cdots \pi_{i_{n-1}, i_n}\, b_{i_0}^{Y_0} \cdots b_{i_{k-1}}^{Y_{k-1}}\, b_i^{Y_k}\, b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n}$$

$$= \sum_{i_{k+1}, \cdots, i_n \in E} \Big[\sum_{i_0, \cdots, i_{k-1} \in E} \nu_{i_0}\, \pi_{i_0, i_1} \cdots \pi_{i_{k-1}, i}\, b_{i_0}^{Y_0} \cdots b_{i_{k-1}}^{Y_{k-1}}\, b_i^{Y_k} \Big] \quad \longleftarrow \quad p_k^i$$

$$\pi_{i, i_{k+1}} \cdots \pi_{i_{n-1}, i_n}\, b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n}$$

$$= p_k^i \Big[\sum_{i_{k+1}, \cdots, i_n \in E} \pi_{i, i_{k+1}} \cdots \pi_{i_{n-1}, i_n}\, b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n} \Big] \quad \longleftarrow \quad v_k^i$$

une expression similaire peut être obtenue dans le cas *numérique*
ceci 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 $i \in E$

avec cette définition, on obtient $q_k^i = p_k^i \, v_k^i$ pour tout $i \in E$

Remarque conditionnellement à $(X_k = i)$, la suite $X_{k+1}, X_{k+2}, \cdots$ des états cachés à venir 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 en déduit que la variable *backward* peut s'interpréter comme la vraisemblance du modèle issu de l'état $X_k = i$ à l'instant k, sachant les observations $(Y_{k+1}, \cdots, Y_n)$

Théorème la suite $\{v_k\}$ vérifie l'équation récurrente rétrograde suivante :

- dans le cas *symbolique*

$$v_{k-1}^i = \sum_{j \in E} \pi_{i,j} \; b_j^{Y_k} \; v_k^j \qquad \text{pour tout } i \in E \qquad \text{(B-sym)}$$

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

- et dans le cas *numérique*

$$v_{k-1}^i = \sum_{j \in E} \pi_{i,j} \; g_j(Y_k) \; v_k^j \qquad \text{pour tout } i \in E \qquad \text{(B-num)}$$

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

Remarque ce résultat énoncé composante–par–composante peut être aussi formulé pour la variable backward vue comme un *vecteur–colonne*

- 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 (B-sym) 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 les équations forward et backward sont *duales* l'une de l'autre :

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

ne dépend pas de l'instant considéré

Preuve on considère uniquement le cas *symbolique* : en utilisant successivement l'équation forward (F-sym) et l'équation backward (B-sym), 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 □

Proposition la distribution de probabilité conditionnelle 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

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

  pour tout $i, j \in E$

- et dans le cas *numérique* par

$$\mathbb{P}[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n] = \frac{1}{L_n} \, p_{k-1}^i \, \pi_{i,j} \, g_j(Y_k) \, v_k^j$$

  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 la distribution de probabilité conditionnelle de l'état présent X_k sachant toutes les observations $(Y_0, \cdots, Y_n)$ est donnée par :

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

avec la définition

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

Remarque 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 considéré, et s'interprètent comme la vraisemblance du modèle sachant les observations $(Y_0, \cdots, Y_n)$

Preuve de la Proposition 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 $(k+1)$

on se ramène à la distribution de probabilité conditionnelle des états cachés successifs sachant les observations, puis on marginalise et on utilise l'identité

$$\mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i,$$

$$X_k = j, X_{k+1} = i_{k+1}, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n]\, L_n$$

$$= \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-2},i}\ \pi_{i,j}\ \pi_{j,i_{k+1}} \cdots \pi_{i_{n-1},i_n}$$

$$b_{i_0}^{Y_0} \cdots b_{i_{k-2}}^{Y_{k-2}}\ b_i^{Y_{k-1}}\ b_j^{Y_k}\ b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n}$$

obtenue avec la méthode élémentaire, de sorte que

$$\mathbb{P}[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n]\; L_n =$$

$$= \sum_{\substack{i_0, \cdots, i_{k-2} \in E \\ i_{k+1}, \cdots, i_n \in E}} \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i, X_k = j, X_{k+1} = i_{k+1}, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n]\; L_n$$

$$= \sum_{\substack{i_0, \cdots, i_{k-2} \in E \\ i_{k+1}, \cdots, i_n \in E}} \nu_{i_0}\; \pi_{i_0, i_1} \cdots \pi_{i_{k-2}, i}\; \pi_{i,j}\; \pi_{j, i_{k+1}} \cdots \pi_{i_{n-1}, i_n}\; b_{i_0}^{Y_0} \cdots b_{i_{k-2}}^{Y_{k-2}}\; b_i^{Y_{k-1}}\; b_j^{Y_k}\; b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n}$$

$$= \sum_{i_0, \cdots, i_{k-2} \in E} \nu_{i_0}\; \pi_{i_0, i_1} \cdots \pi_{i_{k-2}, i}\; b_{i_0}^{Y_0} \cdots b_{i_{k-2}}^{Y_{k-2}}\; b_i^{Y_{k-1}}\; \pi_{i,j}\; b_j^{Y_k} \Big[\sum_{i_{k+1}, \cdots, i_n \in E} \pi_{j, i_{k+1}} \cdots \pi_{i_{n-1}, i_n}\; b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n} \Big]$$

$$= \Big[\sum_{i_0, \cdots, i_{k-2} \in E} \nu_{i_0}\; \pi_{i_0, i_1} \cdots \pi_{i_{k-2}, i}\; b_{i_0}^{Y_0} \cdots b_{i_{k-2}}^{Y_{k-2}}\; b_i^{Y_{k-1}} \Big]\; \pi_{i,j}\; b_j^{Y_k}\; v_k^j \qquad \square$$

mise en œuvre numérique (suite)

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 à tout instant k comme

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

puis la version normalisée de la distribution conditionnelle (lisseur)

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

il est plus efficace, d'un point de vue *numérique*, de propager directement la log–vraisemblance et le filtre, puis de propager la variable définie à tout instant k comme

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

Remarque 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 la suite $\{\bar{v}_k\}$ vérifie l'équation récurrente rétrograde suivante :

- dans le cas *symbolique*

$$\bar{v}^i_{k-1} = \frac{1}{c_k} \sum_{j \in E} \pi_{i,j} \, b_j^{Y_k} \, \bar{v}^j_k \qquad \text{pour tout } i \in E$$

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

- et dans le cas *numérique*

$$\bar{v}^i_{k-1} = \frac{1}{c_k} \sum_{j \in E} \pi_{i,j} \, g_j(Y_k) \, \bar{v}^j_k \qquad \text{pour tout } i \in E$$

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

où les constantes de normalisation sont celles déjà définies pour la normalisation de la variable forward

Remarque ce résultat énoncé composante–par–composante peut être aussi formulé pour la variable backward normalisée vue comme un *vecteur–colonne*

- 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 déjà définies pour la normalisation de la variable forward

Preuve on considère uniquement le cas *symbolique* : on remarque que

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

et en utilisant l'équation backward (B-sym), on obtient

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

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

Remarque 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 plus haut, 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 $i,j \in E$

classification bayésienne

modèles de Markov cachés

équations forward / backward de Baum

algorithme de Viterbi

re–estimation de Baum–Welch

détection de changement

les variables forward et backward 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 considéré, et s'interprète comme la vraisemblance du modèle sachant les observations $(Y_0, \cdots, Y_n)$

il n'y a pas lieu de calculer des moyennes conditionnelles, mais on peut utiliser en revanche l'estimateur du *maximum a posteriori*, qui minimise la probabilité de l'erreur d'estimation sachant les observations $(Y_0, \cdots, Y_n)$, et 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

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

il peut arriver que la suite $(X_0^{\rm LMAP}, \cdots, X_n^{\rm LMAP})$ ainsi générée soit incohérente avec le modèle, dans le sens suivant : il peut arriver que l'on obtienne $X_{k-1}^{\rm LMAP} = i$ et $X_k^{\rm LMAP} = j$ 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 juste impossible pour le modèle

pour cette raison, on utilise plutôt un autre estimateur, appelé estimateur *trajectoriel* du *maximum a posteriori*, défini par

$$(X_0^{\rm MAP}, \cdots, X_n^{\rm MAP}) = \mathop{\rm argmax}_{i_0,\cdots,i_n \in E} \mathbb{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n]$$

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 bien sûr pas possible d'effectuer cette maximisation de manière exhaustive, en énumérant toutes les $|E|^{n+1}$ trajectoires possibles : le calcul efficace de cet estimateur est fourni par un algorithme de programmation dynamique, appelé *algorithme de Viterbi*

Remarque 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 le maximum des fonctions

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

définies sur les ensembles E_1 et E_2 respectivement : 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

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

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 supremum, c'est–à–dire que

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

pour tout $x_2 \in E_2$

programmation dynamique

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 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*

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

pour tout $i \in E$

Théorème la suite $\{V_k\}$ vérifie la récurrence suivante :

- dans le cas *symbolique*

$$V_k^j = \max_{i \in E} [V_{k-1}^i \, \pi_{i,j}] \, b_j^{Y_k} \qquad \text{pour tout } j \in E \qquad \text{(V-sym)}$$

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

- et dans le cas *numérique*

$$V_k^j = \max_{i \in E} [V_{k-1}^i \, \pi_{i,j}] \, g_j(Y_k) \qquad \text{pour tout } j \in E \qquad \text{(V-num)}$$

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 et pour tout état $j \in E$, l'indice

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

qui s'interprète comme un pointeur vers un état à l'instant précédent

Preuve on considère uniquement le cas *symbolique* : on rappelle l'identité suivante, obtenue avec la méthode élémentaire

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

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

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

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

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

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

pour tout $i, j \in E$, et il suffit de maximiser par rapport à $i \in E$ $\square$

Remarque soit $j \in E$ un état 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 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

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

en d'autres termes, parmi toutes les trajectoires qui aboutissent dans l'état j à l'instant k, la trajectoire 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 $(k-1)$

Théorème la suite $\{X_k^{\text{MAP}}\}$ vérifie l'équation récurrente rétrograde suivante :

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

avec la condition initiale

$$X_n^{\text{MAP}} = \operatorname*{argmax}_{i \in E} \; V_n^i$$

Preuve si la suite $(i_0^*, \cdots, i_{n-1}^*, i_n^*)$ atteint le maximum de la fonction

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

alors nécessairement 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

$$i_n^* = \operatorname*{argmax}_{i \in E} V_n^i$$

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

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

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

on rappelle l'identité suivante, obtenue avec la méthode élémentaire

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

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

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

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

on en déduit 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 i_{k-1}^* atteint le maximum de la fonction

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

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

$$i \longmapsto V^i_{k-1}\, \pi_{i,i^*_k}$$

en d'autres termes

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

jusqu'ici, 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 *donné*

l'objectif ici est d'*identifier* le modèle, c'est–à–dire 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*

cas symbolique

dans le cas *symbolique*, la fonction de vraisemblance du modèle $\boldsymbol{M} = (\nu, \pi, b)$ admet l'expression

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

obtenue avec la méthode élémentaire, 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 la fonction de vraisemblance prend la valeur

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

le rapport de vraisemblance entre le modèle $\boldsymbol{M}$ et le modèle $\boldsymbol{M'}$ peut s'écrire

$$\frac{L_n}{L'_n} = \frac{\displaystyle\sum_{i_0,\cdots,i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}}{\displaystyle\sum_{j_0,\cdots,j_n \in E} \nu'_{j_0}\ \pi'_{j_0,j_1} \cdots \pi'_{j_{n-1},j_n}\ {b'}_{j_0}^{Y_0} \cdots {b'}_{j_n}^{Y_n}}$$

$$= \frac{\displaystyle\sum_{i_0,\cdots,i_n \in E} \nu'_{i_0}\ \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\ {b'}_{i_0}^{Y_0} \cdots {b'}_{i_n}^{Y_n}\ \Big[\, \frac{\nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}}{\nu'_{i_0}\ \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\ {b'}_{i_0}^{Y_0} \cdots {b'}_{i_n}^{Y_n}} \,\Big]}{\displaystyle\sum_{j_0,\cdots,j_n \in E} \nu'_{j_0}\ \pi'_{j_0,j_1} \cdots \pi'_{j_{n-1},j_n}\ {b'}_{j_0}^{Y_0} \cdots {b'}_{j_n}^{Y_n}}$$

$$= \mathbb{E}'\Big[\, \frac{\nu_{X_0}\ \pi_{X_0,X_1} \cdots \pi_{X_{n-1},X_n}\ b_{X_0}^{Y_0} \cdots b_{X_n}^{Y_n}}{\nu'_{X_0}\ \pi'_{X_0,X_1} \cdots \pi'_{X_{n-1},X_n}\ {b'}_{X_0}^{Y_0} \cdots {b'}_{X_n}^{Y_n}} \mid Y_0, \cdots, Y_n \Big]$$

et compte tenu que la fonction $x \longmapsto \log x$ est concave, on a

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

le (logarithme du) rapport de vraisemblance entre le modèle $\boldsymbol{M}$ et le modèle $\boldsymbol{M}'$ est donc minoré par

$$Q_n = \mathbb{E}'[\log \frac{\nu_{X_0}\ \pi_{X_0,X_1} \cdots \pi_{X_{n-1},X_n}\ b_{X_0}^{Y_0} \cdots b_{X_n}^{Y_n}}{\nu'_{X_0}\ \pi'_{X_0,X_1} \cdots \pi'_{X_{n-1},X_n}\ b'^{Y_0}_{X_0} \cdots b'^{Y_n}_{X_n}} \mid Y_0, \cdots, Y_n]$$

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 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

$$\nu_i = \bar{p}'^{\,i}_0 \; \bar{v}'^{\,i}_0 \qquad \text{et} \qquad \pi_{i,j} = \pi'_{i,j} \, \frac{\displaystyle\sum_{k=1}^{n} \frac{1}{c'_k} \; \bar{p}'^{\,i}_{k-1} \; b'^{\,Y_k}_j \; \bar{v}'^{\,j}_k}{\displaystyle\sum_{k=1}^{n} \bar{p}'^{\,i}_{k-1} \; \bar{v}'^{\,i}_{k-1}}$$

et

$$b^\ell_i = \frac{\displaystyle\sum_{k=0}^{n} 1(Y_k = \ell) \; \bar{p}'^{\,i}_k \; \bar{v}'^{\,i}_k}{\displaystyle\sum_{k=0}^{n} \bar{p}'^{\,i}_k \; \bar{v}'^{\,i}_k}$$

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 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 respectivement
- on calcule les paramètres $(\nu_s, \pi_s, b_s) = (\nu, \pi, b)$ grâce aux formules de re–estimation

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

Preuve on remarque que

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

on rappelle également les expressions obtenues pour les distributions conditionnelles de l'état X_k ou de la transition (X_{k-1}, X_k) sachant les observations $(Y_0, \cdots, Y_n)$

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

et

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

pour tout $i, j \in E$, et on en déduit que

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

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

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

et

$$b_i^\ell = \frac{\displaystyle\sum_{k=0}^{n} 1(Y_k = \ell) \; \bar{p}_k'^{\,i} \; \bar{v}_k'^{\,i}}{\displaystyle\sum_{k=0}^{n} \bar{p}_k'^{\,i} \; \bar{v}_k'^{\,i}}$$

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

cas numérique

dans le cas *numérique*, on s'intéresse au cas des densités d'émission 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

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

la fonction de vraisemblance du modèle $\boldsymbol{M} = (\nu, \pi, h, R)$ admet l'expression

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

obtenue avec la méthode élémentaire, 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 la fonction de vraisemblance prend la valeur

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

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

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

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 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

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

et

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

pour tout $i, j \in E$, 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', h', R') des paramètres

Remarque 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 normalisées $\{\bar{p}'_k\}$ et $\{\bar{v}'_k\}$ des équations forward et backward respectivement
- on calcule les paramètres $(\nu_s, \pi_s, h_s, R_s) = (\nu, \pi, h, R)$ grâce aux formules de re–estimation

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

Preuve on remarque que

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

$$= \mathbb{E}'[\log \nu_{X_0} + \sum_{k=1}^{n} \log \pi_{X_{k-1},X_k} + \sum_{k=0}^{n} \log g_{X_k}(Y_k) \mid Y_0, \cdots , Y_n] + \text{cste}$$

$$= \sum_{i \in E} \mathbb{P}'[X_0 = i \mid Y_0, \cdots , Y_n]\ \log \nu_i$$

$$+ \sum_{i,j \in E} \{ \sum_{k=1}^{n} \mathbb{P}'[X_{k-1} = i, X_k = j \mid Y_0, \cdots , Y_n] \}\ \log \pi_{i,j}$$

$$+ \sum_{i \in E} \{ \sum_{k=0}^{n} \mathbb{P}'[X_k = i \mid Y_0, \cdots , Y_n]\ \log g_i(Y_k) \} + \text{cste}$$

on remarque que

$$\log g_i(y) = -\tfrac{1}{2} \log\det R_i - \tfrac{1}{2} (y - h_i)^* R_i^{-1} (y - h_i) + \mathrm{cste}$$

$$= \tfrac{1}{2} \log\det M_i - \tfrac{1}{2} \mathrm{trace}[(y - h_i)(y - h_i)^* M_i] + \mathrm{cste}$$

avec $M_i = R_i^{-1}$ pour tout $i \in E$ et tout $y \in \mathbb{R}^d$
on rappelle également les expressions obtenues pour les distributions conditionnelles de l'état X_k ou de la transition (X_{k-1}, X_k) sachant les observations $(Y_0, \cdots, Y_n)$

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

et

$$\mathbb{P}'[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n] = \frac{1}{c'_k} \, \bar{p}'^{\,i}_{k-1} \, \pi'_{i,j} \, g'_j(Y_k) \, \bar{v}'^{\,j}_k$$

pour tout $i, j \in E$

on en déduit que

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

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

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

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

$$a\,\mathrm{trace}(R\,D) - \mathrm{trace}(A\,D) = \mathrm{trace}[\,(a\,R - A)\,D] \qquad \text{où} \qquad R = M^{-1}$$

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

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

et

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

pour tout $i, j \in E$

classification bayésienne

modèles de Markov cachés

équations forward / backward de Baum

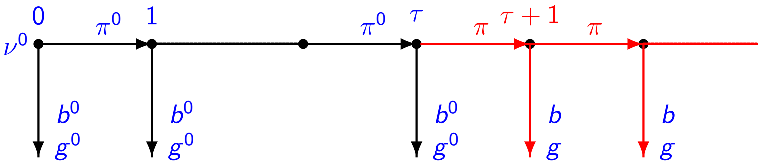
algorithme de Viterbi

re–estimation de Baum–Welch

détection de changement

on considère le problème de détecter au vu des observations $(Y_0, \cdots, Y_n)$ un (éventuel) changement dans le modèle $\boldsymbol{M} = (\nu, \pi, b)$ dans le cas *symbolique*, ou dans le modèle $\boldsymbol{M} = (\nu, \pi, g)$ dans le cas *numérique*, à un instant τ inconnu compris entre l'instant initial 0 et l'instant final n (ou bien au–delà de l'instant final n)

- à l'instant 0, la loi initiale est $\nu^0 = (\nu_i^0)$
- jusqu'à l'instant τ, i.e. entre l'instant initial 0 et l'instant τ, le modèle est *nominal*, c'est–à–dire que la matrice de transition est $\pi^0 = (\pi_{ij}^0)$ et les probabilités d'émission sont $b^0 = (b_i^{l,0})$, ou les densités d'émission sont $g^0 = (g_i^0)$
- au–delà de l'instant τ, i.e. entre l'instant τ et l'instant final n, le modèle est *modifié*, c'est–à–dire que la matrice de transition est $\pi = (\pi_{ij})$ et les probabilités d'émission sont $b = (b_i^l)$, ou les densités d'émission sont $g = (g_i)$


0
ν^0
π^0
1
π^0
τ
π
$\tau + 1$
π
b^0
g^0
b^0
g^0
b^0
g^0
b
g
b
g

on rappelle que la vraisemblance pour le modèle sans changement où $\tau = n$, i.e. la vraisemblance pour le modèle *nominal*, peut être calculée comme

$$L_n^0 = \sum_{i \in E} p_n^{i,0}$$

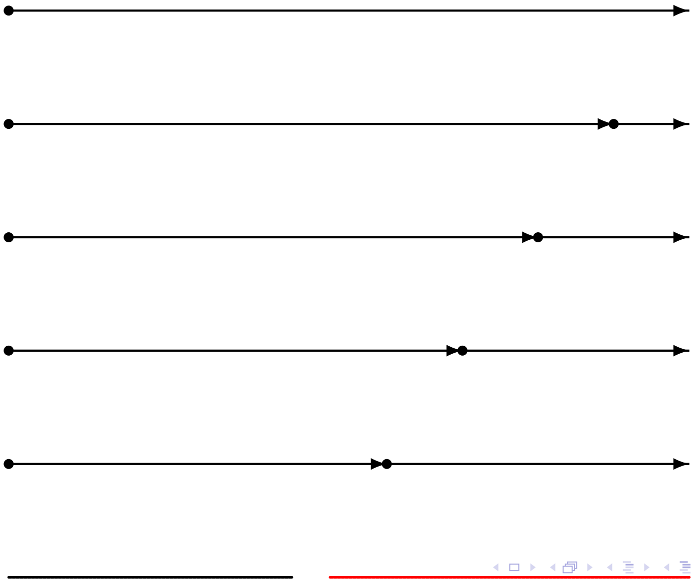
où $p_n^0 = (p_n^{i,0})$ désigne la variable forward à l'instant final n, calculée en utilisant le modèle *nominal* uniquement

pour tout $\tau = 0 \cdots (n-1)$, la vraisemblance pour le modèle avec changement à l'instant τ, peut être calculée comme

$$L_n(\tau) = \sum_{i \in E} p_n^i(\tau)$$

où $p_n(\tau) = (p_n^i(\tau))$ désigne la variable forward à l'instant final n, calculée en utilisant

- le modèle *nominal* depuis l'instant initial 0 et jusqu'à l'instant τ
- le modèle *modifié* à partir de l'instant τ et jusqu'à l'instant final n



propriété de dualité des variables forward et backward :
pour tout $\tau = 0 \cdots (n-1)$, la vraisemblance pour le modèle avec changement à l'instant τ, peut être calculée comme

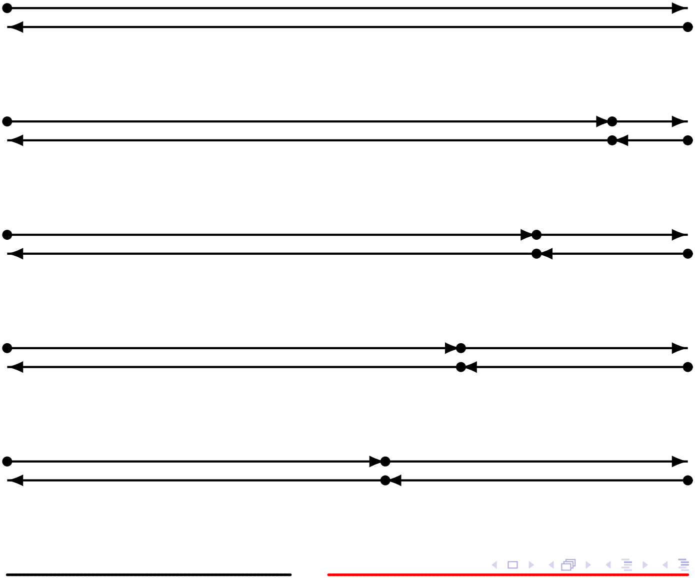
$$L_n(\tau) = \sum_{i \in E} p_n^i(\tau) = \sum_{i \in E} p_k^i(\tau)\, v_k^i(\tau)$$

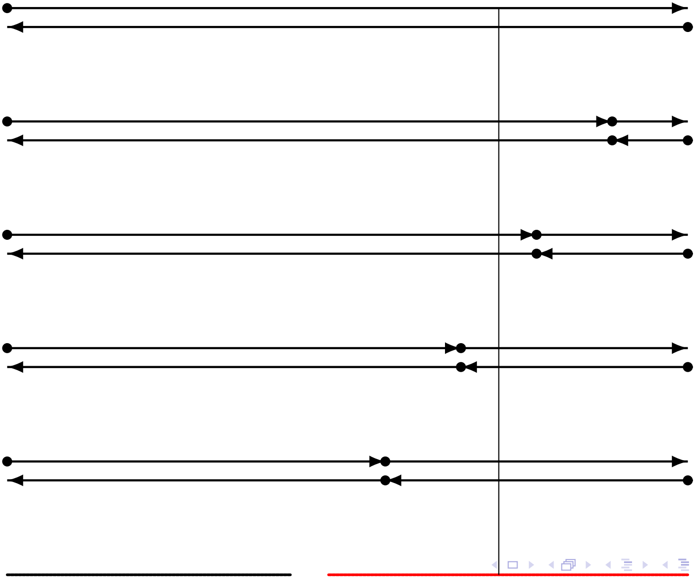
où $p_k(\tau) = (p_k^i(\tau))$ désigne la variable forward à un instant intermédiaire k, calculée en utilisant

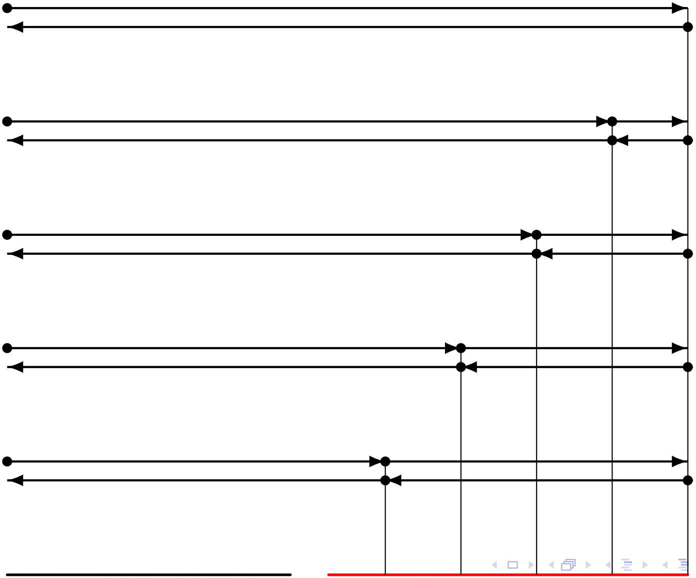
- le modèle *nominal* depuis l'instant initial 0 et jusqu'à l'instant τ
- le modèle *modifié* à partir de l'instant τ et jusqu'à l'instant final n

et où $v_k(\tau) = (v_k^i(\tau))$ désigne la variable backward à un instant intermédiaire k, calculée en utilisant

- le modèle *modifié* depuis l'instant initial n et jusqu'à l'instant τ
- le modèle *nominal* à partir de l'instant τ et jusqu'à l'instant final 0







ce dernier schéma adapté permet d'évaluer la vraisemblance $L_n(\tau)$ pour tous les instants de changement possibles $\tau = 0, 1 \cdots n$ en propageant deux équations seulement

- une équation forward utilisant le modèle *nominal* uniquement
- une équation backward utilisant le modèle *modifié* uniquement