Reconnaissance de mots isolés (utilisation des modèles HMM)

Maurice Charbit

October 25, 2002

Contents

1 Modèle de Markov caché

Une chaîne de Markov homogène est un processus aléatoire $Q(n)$, $n \geq 1$, à valeurs dans un alphabet fini $(1, \cdots, S)$, tel que $P(Q(n) = j | Q(n-1) = i, Q(n-2) = q_{n-2}, \cdots, Q(1) = q_1) = P(Q(n) = j | Q(n-1) = i) = a_{ij}$ où $1 \leq i, j \leq S$. Le terme homogène fait référence au fait que a_{ij} est indépendant de n. Une chaîne de Markov est donc décrite par la donnée :

- d'un nombre S d'états numérotés de 1 à S,
- de la loi de probabilité initiale Π de la variable aléatoire $Q(1)$, $\Pi = (\pi_1, \cdots, \pi_S)$ avec $\pi_i \geq 0$ et $\sum_i \pi_i = 1$,
- de la matrice A de transition d'états, dont l'élément a_{ij} représente la probabilité de passer de l'état i, à un instant quelconque, à l'état j à l'instant suivant. On a donc $\sum_j a_{ij} = 1$.

Elle dépend donc du paramètre :

$$\nu = (\Pi, A) \tag{1}$$

Un *modèle de Markov caché* est obtenu en associant alors à la chaîne précédente un processus de sortie $U(n)$ de la façon suivante :

- on considère S processus aléatoires $U^s(n)$ ($s \in \{1, \cdots, S\}$), vectoriels de dimension D, indépendants, de lois respectives $P(du; \rho_s)$. En reconnaissance de la parole, on considère souvent pour $P(du; \rho_s)$ un mélange de P_s composantes gaussiennes, de densité :

$$p_s(u; \rho_s) = \sum_{c=1}^{P_s} \lambda_{cs} (2\pi)^{-D/2} (\det(\Sigma_{cs}))^{-1/2} \exp\left(-\frac{1}{2} (u - \mu_{cs})^T \Sigma_{cs}^{-1} (u - \mu_{cs}) \right) \quad \text{où} \quad u \in \mathbb{R}^D$$

où $\lambda_{cs} \geq 0$, avec $\sum_c \lambda_{cs} = 1$, désignent les proportions respectives de chaque mélange. μ_{cs} est une suite de vecteurs de dimension D et Σ_{cs} une suite de matrices de covariance de dimension $D \times D$. Dans ce cas, ρ_s désigne le paramètre :

$$\rho_s = (\{\lambda_{cs}\}_{c=1:P_s}, \{\mu_{cs}\}_{c=1:P_s}, \{\Sigma_{cs}\}_{c=1:P_s}) \tag{2}$$

- le processus observé est alors défini par $U(n) = U^s(n)$ si $Q(n) = s$. Ce qui s'écrit :

$$U(n) = \sum_{s=1}^{S} U^s(n)\mathbf{1}(Q(n) = s)$$

Un modèle de Markov est dit *caché* (en anglais, *Hidden Markov Model*, en abrégé *HMM*) si on n'a pas accès à la suite des états mais que l'on peut uniquement observer le processus $U(n)$.

Dans la suite θ désigne l'ensemble des paramètres du modèle HMM soit :

$$\theta = \{\{\rho_s\}_{s=1:S}, A, \Pi\}$$

avec les contraintes $a_{ij} \geq 0$, $\pi_i \geq 0$, $\sum_j a_{ij} = 1$ et $\sum_i \pi_i = 1$.

Dans le cas où les $U^s(n)$ sont des mélanges de P_s composantes gaussiennes, ρ_s est donné par (2). ρ_s se simplifie alors si on suppose que les lois sont purement gaussiennes ($P_s = 1$) et que les matrices Σ_s sont diagonales :

$$\rho_s = (\mu_{s1}, \cdots, \mu_{sD}, \sigma_{s1}, \cdots, \sigma_{sD})$$

Loi des états

Notons que l'on peut écrire $P(Q(n) = q_n | Q(n-1) = q_{n-1}) = \sum_{i=1}^{S} \sum_{j=1}^{S} a_{ij}\mathbf{1}(q_n = i, q_{n-1} = j)$ et que $P(Q(1) = q_1) = \sum_{i=1}^{S} \pi(i)\mathbf{1}(q_1 = i)$. En utilisant l'identité $f(\sum_i x_i \mathbf{1}(q = i)) = \sum_i f(x_i)\mathbf{1}(q = i)$ à la fonction logarithme, on obtient :

$$\log(P(Q(n) = q_n | Q(n-1) = q_{n-1}) = \sum_{i=1}^{S} \sum_{j=1}^{S} \log(a_{ij})\mathbf{1}(q_n = i, q_{n-1} = j)$$

et que :

$$\log(P(Q(1) = q_1)) = \sum_{i=1}^{S} \log(\pi(i))\mathbf{1}(q_1 = i)$$

En utilisant la propriété markovienne de la suite $Q(n)$, on obtient pour la loi conjointe d'une suite de T états l'expression[1] :

$$P(Q_{1:T} = q_{1:T}) = \prod_{t=1}^{T-1} P(Q(t+1) = q_{t+1} | Q(t) = q_t) P(Q(1) = q_1)$$

qui s'écrit encore :

$$P(Q_{1:T} = q_{1:T}) = \prod_{t=1}^{T-1} \sum_{i=1}^{S} \sum_{j=1}^{S} a_{ij}\mathbf{1}(q_{t+1} = i, q_t = j) \sum_{i=1}^{S} \pi(i)\mathbf{1}(q_1 = i)$$

En passant au logarithme, il vient :

$$\log(P(Q_{1:T} = q_{1:T})) = \sum_{t=1}^{T-1} \sum_{i=1}^{S} \sum_{j=1}^{S} \log(a_{ij})\mathbf{1}(q_{t+1} = i, q_t = j) + \sum_{i=1}^{S} \log(\pi(i))\mathbf{1}(q_1 = i) \tag{3}$$

[1] la notation $Q_{1:T} = q_{1:T}$ signifie $Q(1) = q(1), \cdots, Q(T) = q(T)$.

Loi des observations conditionnellement aux états

La densité de probabilité de la loi de $U(n)$ conditionnellement à $Q(n) = q_n$ a pour expression :

$$p(u|Q_n = q_n) = \sum_{s=1}^{S} p_s(u;\rho_s)\mathbf{1}(q_n = s)$$

En passant au logarithme, il vient :

$$\log(p(u|Q_n = q_n)) = \sum_{s=1}^{S} \log(p_s(u;\rho_s))\mathbf{1}(q_n = s)$$

Comme on a supposé les lois conditionnelles indépendantes, le logarithme de la densité de probabilité de la loi des T observations $U_{(1:T)}$ conditionnellement à la suite des états $Q_{(1:T)}$ a pour expression :

$$\log(p(u_{1:T}|Q_{1:T} = q_{1:T})) = \sum_{t=1}^{T}\sum_{i=1}^{S} \log(p_i(u_t;\rho_i))\mathbf{1}(q_t = i) \tag{4}$$

Loi conjointe

Partant des expressions (3) et (4), on en déduit par addition le logarithme de la loi conjointe des observations et des états.

On rappelle que, si $p(x,\zeta)$ désigne la densité de probabilité de la loi d'une variable aléatoire X pour le paramètre ζ, la log-vraisemblance de ζ est définie par $\log(p(X,\zeta))$ vue comme une fonction de la variable aléatoire X. Partant de là, la log-vraisemblance de la suite conjointe $(U_{1:T}, Q_{1:T})$, pour le paramètre θ, s'écrit :

$$\begin{aligned}
\ell(U,Q;\theta) &= \log(p(U_{1:T},Q_{1:T};\theta)) \\
&= \sum_{t=1}^{T}\sum_{i=1}^{S} \log(p_i(U_t;\rho_i))\mathbf{1}(Q_t = i) + \sum_{t=1}^{T-1}\sum_{i=1}^{S}\sum_{j=1}^{S} \log(a_{ij})\mathbf{1}(Q_{t+1} = i, Q_t = j) \\
&\quad + \sum_{i=1}^{S} \log(\pi(i))\mathbf{1}(Q_1 = i)
\end{aligned} \tag{5}$$

Considérons à présent, pour la même valeur de θ, R suites d'observations $\{\{U^1_{1:T_1}\},\cdots,\{U^R_{1:T_R}\}\}$, que nous notons $U^{1:R}_{1:T_R}$. En supposant que les R suites sont indépendantes et de durées respectives $(T_1,\cdots,T_R)$, la log-vraisemblance a pour expression :

$$\begin{aligned}
\ell_R(U^{1:R}_{1:T_R}, Q^{1:R}_{1:T_R};\theta) &= \sum_{r=1}^{R} \log(p(U^r_{1:T_r}, Q^r_{1:T_r};\theta)) \\
&= \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{S} \log(p_i(U^r_t;\rho_i))\mathbf{1}(Q^r_t = i) \\
&\quad + \sum_{r=1}^{R}\sum_{t=1}^{T_r-1}\sum_{i=1}^{S}\sum_{j=1}^{S} \log(a_{ij})\mathbf{1}(Q^r_{t+1} = i, Q^r_t = j) \\
&\quad + \sum_{r=1}^{R}\sum_{i=1}^{S} \log(\pi(i))\mathbf{1}(Q^r_1 = i)
\end{aligned} \tag{6}$$

Problèmes dans les modèles de Markov cachés

1. Etant donnée une observation $U_{1:T}$ et un ensemble $\Theta = \{\theta_1, \cdots, \theta_M\}$ de M valeurs de θ, quelle est la valeur de Θ qui rend maximale la vraisemblance. Ce problème est celui de la *reconnaissance* d'un mot dans un dictionnaire de M mots.

2. Etant donnée une suite de R observations $U_{1:T}^{1:R}$ déterminer la valeur de θ qui rend maximale la vraisemblance. Ce problème est celui de l'*apprentissage*.

3. Etant donnée une observation $U_{1:T}$ de durée T, quelle est, pour un valeur de θ donnée, la suite des états qui a la vraisemblance maximale. Ce problème se rencontre, par exemple, lors de la reconnaissance de mots en parole continue.

2 Coefficients cepstraux

2.1 HMM en reconnaissance de la parole

Le signal associé à un mot isolé peut être considéré comme une suite de sons de base agissant comme un alphabet. Un mot est alors caractérisé par une suite caractéristique d'éléments de cet alphabet. En absence de mémoire, il suffirait de comparer élément par élément les constituants d'un mot. Toutefois en pratique on constate que la probabilité d'un élément, dans la suite considérée, dépend des éléments qui ont précédé. D'où l'idée d'une modélisation markovienne de la suite de ces éléments.

Un premier traitement, dont la valeur est essentielle, est donc d'extraire du signal un nombre faible d'éléments pertinents qui sont supposés suivre un modèle HMM. Deux approches sont fréquemment utilisées :

- approche temporelle: on effectue une prédiction linéaire (LPC: linear prediction coding) sur le signal. Cela consiste à approximer le signal par une combinaison linéaire des p valeurs précédentes. On peut donc écrire que $s(s) = a_1 s(n-1) + \cdots + a_p s(n-p) + b(n)$ où $b(n)$ est un processus aléatoire qui prend en compte les erreurs du modèle. On le modélise par un bruit blanc de puissance σ^2. Le vecteur $\theta = (\sigma^2, a_1, \cdots, a_p)$ est choisi de façon à minimiser σ^2. La solution est donnée par les équations de Yule-Walker : $R(1, a_1, \cdots, a_p)^T = (\sigma^2, 0, \cdots, 0)^T$ qui lie θ à la matrice de covariance R du signal $s(n)$. Une estimation de R permet d'estimer le paramètre θ caractérisant le bloc.

- approche fréquentielle: le signal est passé par un banc de filtres. En général les filtres couvrent la bande de Nyquist, avec un chevauchement adéquate, suivant une échelle non linéaire. Les plus fréquemment utilisées sont des échelles logarithmiques analogues à celle de l'oreille humaine.

  Les points de FFT de l'ensemble des sous-bandes représentent le vecteur de paramètres du bloc. Il est parfois intéressant de prendre le cepstre plutôt que le spectre dans la mesure où les coefficients cepstraux sont généralement décorrélées. Cela justifie l'utilisation de matrice diagonale dans le traitement HMM.

Nous nous limiterons ici à la deuxième approche.

2.2 Prétraitement du signal

D'après le théorème d'échantillonnage, la parole est tout d'abord échantillonnée à une fréquence F_e double de la fréquence maximale utile. Pour les essais $F_e = 48\,000$Hz. On effectue ensuite une quantification sur $b = 8$ bits, ce qui donne un rapport signal sur bruit de quantification légèrement inférieur à 48 dB.

Une fois numérisé, le signal subit une opération de *préaccentuation*, qui consiste en un filtrage de type passe-haut qui relève le niveau des aigus. En pratique, on utilise simplement un filtre de réponse impulsionnelle finie $(1, a)$ avec $a = -0.95$. Si $s(n)$ désigne le signal de parole et $s_p(n)$ le signal pré-accentué on a :

$$s_p(n) = s(n) - 0.95s(n-1)$$

D'autre part on est conduit, dans la suite, à traiter les données en travaillant sur des trames de valeurs consécutives. Cette façon de procéder revient à appliquer une fenêtre rectangulaire de durée finie sur l'ensemble du signal. Pour réduire les effets dus aux discontinuités aux bords de la fenêtre, il est fréquent de pondérer une trame de longueur N par la *fenêtre de Hamming*. Cette opération donne la trame fenêtrée :

$$s_w(n) = s_p(n)w(n) \quad \text{où} \quad w(n) = 0.54 - 0.46\cos(2\pi n/(N-1)) \quad \text{avec } 0 \leq n \leq N-1$$

La longueur N d'une trame est choisie de façon à avoir des trames dont la durée est de l'ordre de 20ms. Enfin l'opération de découpage en trames de longueur N comporte un recouvrement de 50% entre trames successifs. En conséquence, pour un mot de durée de 0.5s (comme par exemple la liste des chiffres : zéro, un, ..., ,neuf), cela donne entre 30 et 40 trames à traiter.

2.3 Calcul des coeffients cepstraux

On effectue ensuite sur chaque trame une transformée de Fourier discrète (TFD) :

$$Y(k) = \sum_{n=0}^{L-1} s_w(n)e^{-2j\pi kn/L}$$

où la suite $s_w(n)$ est éventuellement complétée par des zéros. On a pris pour longueur de la TFD $L = 1024$.

On partitionne ensuite l'échelle des fréquences entre 0 et $F_e/2$ en 24 bandes correspondant à l'échelle logarithmique *mel*[2] donnée par :

$$mel(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right) \tag{7}$$

où f est la fréquence en Hz. Pour une fréquence d'échantillonnage $F_e = 48\,000$H, on obtient les 24 bandes représentées à la figure 1.

On note que les bandes se chevauchent. Les valeurs des fréquences sont données dans le tableau 1.

On somme les valeurs de la transformée de Fourier discrète sur chacune des $B = 24$ bandes et on en prend le logarithme. Ce qui donne :

$$Y_m(p) = \log(\sum_{k \in B_p} |Y(k)|) \quad \text{pour} \quad p \in \{0, \cdots, B-1\} \tag{8}$$

On effectue enfin un transformée inverse en cosinus sur ces B valeurs, dont on rappelle l'expression :

$$z_m = C_B Y_m$$

où Y_m désigne le vecteur de composantes $Y_m(p)$ (équation (8)) et C_B la matrice d'éléments :

$$[C_B]_{k,n} = \sqrt{\frac{2}{B}}c(k)\cos\left(\frac{\pi k(n+0.5)}{B}\right) \quad \text{pour} \quad 0 \leq k, n \leq B-1$$

[2]L'échelle *mel* (comme melody) est une échelle psychoacoustique. L'équation (7) correspond assez bien aux valeurs expérimentales. Par définition, l'échelle *mel* associe à la fréquence $f = 1000$ Hz la valeur 1000 mel.

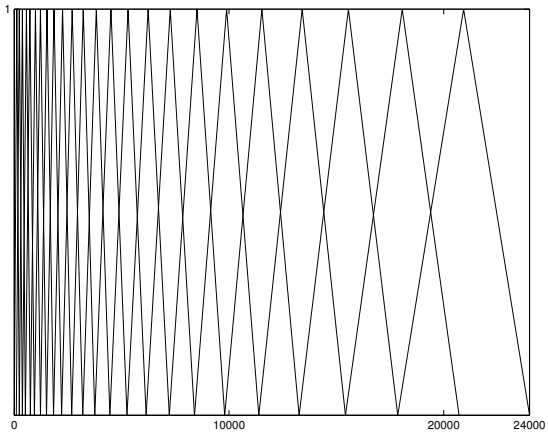

Figure 1: Bandes de fréquences en échelle *mel*

no	*bande (Hz)*	*no*	*bande (Hz)*
1	$0-234$	13	$3188-4453$
2	$94-375$	14	$3750-5250$
3	$234-516$	15	$4453-6141$
4	$375-750$	16	$5250-7219$
5	$516-938$	17	$6141-8391$
6	$750-1219$	18	$7219-9797$
7	$938-1500$	19	$8391-11391$
8	$1219-1828$	20	$9797-13266$
9	$1500-2203$	21	$11391-15422$
10	$1828-2672$	22	$13266-17859$
11	$2203-3188$	23	$15422-20719$
12	$2672-3750$	24	$17859-24000$

Table 1: *Bandes de partition mel*

avec $c(0) = 1/\sqrt{2}$ et $c(k) = 1$ pour $k \neq 0$.

On conserve, pour chaque trame, les $(D-1)$ premières valeurs de $z_m(1:D-1)$, après les avoir centrées, ainsi que l'énergie dans chaque trame :

$$e_r = \log(\sum_n |s_w(n)|^2)$$

Typiquement on prend $D = 13$. On ajoute, souvent, les différences $\Delta_{\{j=1:J\}} = C_j(t) - C_j(t-1)$ ainsi que les différences secondes $\Delta^2_{\{j=1:J\}} = \Delta_j(t) - \Delta_j(t-1)$ par rapport à la trame précédente.

En considérant des mots d'une durée de l'ordre de 0.5s, on obtient une suite de 30 à 40 vecteurs de dimension $D = 13$ (voire $D = 37$), suite qui est supposée suivre un modèle HMM à S états. L'ordre S du modèle HMM dépend du mot et est choisi de façon empirique. Toutefois pour éviter une trop grande complexité, dans l'estimation des paramètres de la chaîne, la valeur de S considérée est inférieure à 6. Enfin, pour le problème de la reconnaissance de mots isolés, on suppose que la topologie de la chaîne est de type *gauche-droite*, cad $a_{ij} = 0$ pour $i < j$ et que on démarre toujours dans le même état noté 1 ce qui conduit à prendre $\Pi = (1, 0, \cdots, 0)$.

3 Estimation pour un HMM : algorithme EM

Partant de la loi conjointe de $p(u, Q; \theta)$, on peut déterminer la loi marginale de l'observation U et en déduire un estimateur du maximum de vraisemblance de θ. En pratique ce calcul est infaisable. C'est pourquoi on adopte une estimation à partir de l'algorithme EM qui, dans le cas des modèles HMM, a une forme plus simple.

3.1 Principes

On pose :

$$L(\theta, \theta^{(p)}) = E_{\theta_p}(\ell_R(U_{1:T_R}^{1:R}, Q_{1:T_R}^{1:R}; \theta)|U_{1:T_R}^{1:R})$$

L'exposant (p) indique que l'espérance conditionnelle $L(\theta, \theta^{(p)})$ est calculée en considérant que la variable aléatoire conjointe $(U_{1:T_R}^{1:R}, Q_{1:T_R}^{1:R})$ suit la loi, pour la valeur du paramètre θ obtenu à l'étape p de l'algorithme.

L'algorithme EM pour estimer de θ, au pas d'itération $(p+1)$, la valeur :

$$\theta^{(p+1)} = \arg\max_{\theta} L(\theta, \theta^{(p)}) \tag{9}$$

Comme l'algorithme comporte le calcul d'une espérance suivi d'une maximisation, il est dit *EM* pour *Expectation-Maximization*. La propriété fondamentale de l'algorithme EM est d'assurer une augmentation de la vraisemblance de l'observation $U_{1:T_R}^{1:R}$ à chaque étape d'itération.

En utilisant la formule (6) et le fait que $E(f(X)g(Y)|X) = f(X)E(g(Y)|X)$, on obtient :

$$\begin{aligned} L(\theta, \theta^{(p)}) &= \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{s=1}^{S} \log(p_s(U_t^r; \rho_s)) P_{\theta^{(p)}}(Q_t^r = s|U_{1:T_r}^r) \\ &\quad + \sum_{r=1}^{R}\sum_{t=1}^{T_r-1}\sum_{i=1}^{S}\sum_{j=1}^{S} \log(a_{ij}) P_{\theta^{(p)}}(Q_{t+1}^r = i, Q_t^r = j|U_{1:T_r}^r) \\ &\quad + \sum_{r=1}^{R}\sum_{i=1}^{S} \log(\pi(i)) P_{\theta^{(p)}}(Q_1^r = i|U_{1:T_r}^r) \end{aligned} \tag{10}$$

Supposons que l'on ait déjà calculé les quantités :

$$\gamma_{rts}^{(p)} = P_{\theta^{(p)}}(Q_t^r = s|U_{1:T_r}^r) \tag{11}$$

pour $r \in \{1, \cdots, R\}$, $t \in \{1, \cdots, T_r\}$ et $s \in \{1, \cdots, S\}$, et

$$\xi_{rtij}^{(p)} = P_{\theta^{(p)}}(Q_{t+1}^r = i, Q_t^r = j|U_{1:T_r}^r) \tag{12}$$

pour $r \in \{1, \cdots, R\}$, $t \in \{1, \cdots, T_r - 1\}$ et $i, j \in \{1, \cdots, S\}$.

On peut alors re-estimer la valeur de θ, qui maximise $L(\theta, \theta^{(p)})$, en annulant la dérivée de $L(\theta, \theta^{(p)})$ par rapport à θ.

3.2 Formules de re-estimation

Re-estimation de π_i

Pour déterminer la formule de re-estimation de π_i, rappelons tout d'abord que l'on doit satisfaire les contraintes $\pi_i \geq 0$ et $\sum_i \pi_i = 1$. Pour résoudre le problème, nous utilisons la méthode

des multiplicateurs de Lagrange en ne considérant que la contrainte $\sum_i \pi_i = 1$. Il faut alors maximiser $L(\theta, \theta^{(p)}) - \lambda(\sum_i \pi_i - 1)$. En annulant la dérivée par rapport à π_i, il vient :

$$\sum_{r=1}^{R} \frac{1}{\pi(i)} \gamma_{r1i} - \lambda = 0$$

ce qui donne $\pi(i) = \sum_{r=1}^{R} \gamma_{r1i}/\lambda$. La constante λ se calcule en vérifiant que $\sum_i \pi_i = 1$, ce qui donne :

$$\pi_i^{(p+1)} = \frac{\sum_{r=1}^{R} \gamma_{r1i}}{\sum_{i=1}^{S} \sum_{r=1}^{R} \gamma_{r1i}^{(p)}} \tag{13}$$

Notons ici que la contrainte $\pi_i \geq 0$ est vérifée car on verra que $\gamma_{r1i} \geq 0$ à toutes les étapes.

Re-estimation de a_{ij}

Un calcul en tout point analogue au calcul précédent donne pour formule de re-estimation de a_{ij} l'expression :

$$a_{ij}^{(p+1)} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r - 1} \xi_{rtij}^{(p)}}{\sum_{j=1}^{S} \sum_{r=1}^{R} \sum_{t=1}^{T_r - 1} \xi_{rtij}^{(p)}} \tag{14}$$

Notons ici que la contrainte $a_{ij} \geq 0$ est vérifiée car on verra que $\xi_{rtij} \geq 0$ à toutes les étapes. De même on verra que l'expression récurrente donnant ξ_{rtij} est telle que, si à l'initialisation $a_{ij} = 0$, alors a_{ij} reste nulle après chaque itération. Cela permet de satisfaire la contrainte de topologie gauche-droite du modèle HMM choisi.

Re-estimation de ρ_s

Limitons nous au cas où les lois de $U_s(n)$ sont gaussiennes pures (sans mélange) et où les matrices de covariance sont supposées *diagonales*. Dans ce cas on a :

$$\log(p_s(U_t^r; \rho_s)) = -\frac{D}{2} \log(2\pi) - \frac{1}{2} \sum_{d=1}^{D} \log(\sigma_{sd}) - \frac{1}{2} \sum_{d=1}^{D} \frac{(u_{td}^r - \mu_{sd})^2}{\sigma_{sd}} \tag{15}$$

où μ_{sd} et σ_{sd} désignent respectivement la moyenne et la variance de la d-ème composante de la loi caractérisant l'état s.

En annulant la dérivée de $L(\theta, \theta^{(p)})$ par rapport à μ_{sd} il vient :

$$\sum_{r=1}^{R} \sum_{t=1}^{T_r} (u_{td}^r - \mu_{sd}) \gamma_{rts^{(p)}} = 0$$

qui donne :

$$\mu_{sd}^{(p+1)} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} u_{td}^r \gamma_{rts}^{(p)}}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{rts}^{(p)}} \tag{16}$$

avec $1 \leq s \leq S$ et $1 \leq d \leq D$.

De même en annulant la dérivée de $L(\theta, \theta^{(p)})$ par rapport à σ_{sd} il vient :

$$\sigma_{sd}^{(p+1)} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} (u_{td}^r - \mu_{sd}^{(p+1)})^2 \gamma_{rts}^{(p)}}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{rts}^{(p)}} \tag{17}$$

avec $1 \leq s \leq S$ et $1 \leq d \leq D$.

3.3 Algorithme *Forward-Backward*

Partant de la loi conjointe de $(U_{1:T_R}^{1:R}, Q_{1:T_R}^{1:R})$ pour la valeur du paramètre égale à θ_p, on peut évidemment déduire les valeurs de γ_{rts} et ξ_{rtij} données par les expressions (11) et (12). Toutefois, un algorithme itératif a été trouvé qui simplifie beaucoup le calcul. Pour cela on introduit les deux quantités auxiliaires α_{irt} et β_{irt} définies respectivement par :

$$\alpha_{rts} = p(u_1^r, \cdots, u_t^r, Q_t^r = s) \tag{18}$$

avec $1 \leq r \leq R$, $1 \leq t \leq T_r$ et $1 \leq s \leq S$.

$$\beta_{rts} = p(u_{t+1}^r, \cdots, u_{T_r}^r | Q_t^r = s) \tag{19}$$

avec $1 \leq r \leq R$, $1 \leq t \leq T_r$ et $1 \leq s \leq S$.

On montre que, pour r et s fixés, α_{rts} et β_{rts} se calculent de façon récurrente sur t. La récurrence pour α_{rts} se fait pour t allant de 1 à T_r et la récurrence pour β_{rts} se fait pour t allant de T_r à 1. D'où le nom d'algorithme *forward-backward* qui lui est donné.

Pour cela introduisons tout d'abord la notation :

$$b_{rts} = p_s(u_t; \rho_{sp}) \tag{20}$$

cad la valeur de la vraisemblance de ρ_{sp} pour l'observation effectuée au temps t, la valeur de ρ_{sp} étant celle calculée à l'étape précédente pour l'état s. Dans le cas gaussien pur, b_{rts} est égal à l'exponentiel de la quantité donnée par (15) soit :

$$b_{rts}^{(p)} = \frac{1}{(2\pi\sigma_{sd}^{(p)})^{D/2}} \exp\left(-\frac{1}{2} \sum_{d=1}^{D} \frac{(u_{td}^r - \mu_{sd}^{(p)})^2}{\sigma_{sd}^{(p)}} \right) \tag{21}$$

Nous donnons ci-dessous sans démonstration les formules de récurrence.

Calcul de α_{rts} à l'étape p

Pour $1 \leq r \leq R$ et $1 \leq s \leq S$,

- valeurs initiales : $\alpha_{r1s}^{(p)} = \pi_s^{(p)} b_{r,1,s}^{(p)}$,
- récurrence $t \in \{1, \cdots, T_r - 1\}$:

$$\alpha_{r,t+1,s}^{(p)} = \left(\sum_{i=1}^{S} \alpha_{r,t,i}^{(p)} a_{is}^{(p)} \right) b_{r,t+1,s}^{(p)} \tag{22}$$

Calcul de β_{rts} à l'étape p

Pour $1 \leq r \leq R$ et $1 \leq s \leq S$,

- valeurs initiales : $\beta_{r,T_r,s}^{(p)} = 1$,
- récurrence $t \in \{T_r - 1, \cdots, 1\}$:

$$\beta_{r,t,s}^{(p)} = \sum_{i=1}^{S} \beta_{r,t+1,i}^{(p)} a_{is}^{(p)} b_{r,t+1,i}^{(p)} \tag{23}$$

Calculs de γ_{rts} et ξ_{rtij} à l'étape p

On montre que :

$$\gamma_{rts}^{(p)} = P_{\theta^{(p)}}(Q_t^r = s|U_{1:T_r}^r) = \frac{\alpha_{rts}^{(p)}\beta_{rts}^{(p)}}{\sum_{i=1}^S \alpha_{rti}^{(p)}\beta_{rti}^{(p)}} \tag{24}$$

$$\xi_{rtij}^{(p)} = P_{\theta^{(p)}}(Q_{t+1}^r = i, Q_t^r = j|U_{1:T_r}^r) = \frac{\alpha_{rti}^{(p)}\beta_{r,t+1,j}^{(p)}b_{r,t+1,j}^{(p)}a_{ij}^{(p)}}{\sum_{s=1}^S \alpha_{rts}^{(p)}\beta_{rts}^{(p)}} \tag{25}$$

En utilisant la formule de récurrence (23), on pourra vérifier que l'expression donnée par (25) est telle que $\sum_{i,j=1}^S \xi_{rtij} = 1$. En pratique on pourra encore utiliser cette propriété de normalisation, pour calculer ξ_{rtij}, en calculant uniquement le numérateur.

En se reportant à l'expression (14) on voit que, si $a_{ij} = 0$ à une étape de l'algorithme, alors il reste égal à 0 dans la suite.

3.4 Facteur d'échelle

En pratique les valeurs numériques de α_{rts} et β_{rts} sont *très* petites, ce qui conduit à des problèmes d'*underflows*. On remarque toutefois que, dans le calcul de γ_{rts} et de ξ_{rtij}, les quantités α_{rts} et β_{rts} peuvent être calculées avec un *facteur d'échelle*. On prend, pour α_{rts}, le facteur d'échelle $\sum_{k=1}^S \alpha_{rtk}$, ce qui conduit à poser :

$$\tilde{\alpha}_{rts} = \frac{\alpha_{rts}}{\sum_{k=1}^S \alpha_{rtk}}$$

L'algorithme par récurrence porte alors sur les quantités $\tilde{\alpha}_{rts}$ dont les valeurs sont mieux adaptées au calcul, dû au fait que $\sum_s \tilde{\alpha}_{rts} = 1$. Partant de là, on a d'après (22) :

$$\begin{aligned}\tilde{\alpha}_{r,t+1,s} &= \frac{\alpha_{r,t+1,s}}{\sum_{k=1}^S \alpha_{r,t+1,k}} = b_{r,t+1,s}\frac{\sum_{i=1}^S \alpha_{r,t,i}a_{is}}{\sum_{k=1}^S \alpha_{r,t+1,k}} = \underbrace{b_{r,t+1,s}\sum_{i=1}^S \tilde{\alpha}_{r,t,i}a_{is}}_{\bar{\alpha}_{r,t+1,s}}\underbrace{\frac{\sum_{j=1}^S \alpha_{r,t,j}}{\sum_{k=1}^S \alpha_{r,t+1,k}}}_{1/c(t+1,r)} \\ &= \bar{\alpha}_{r,t+1,s}/c(t+1,r)\end{aligned}$$

Comme le terme $c(t+1,r)$ ne dépend pas de s, il suffit de calculer tout d'abord, pour toutes les valeurs de s, $\bar{\alpha}_{r,t+1,s}$ à partir de $\tilde{\alpha}_{r,t,i}$ puis d'en déduire alors puisque $\sum_s \tilde{\alpha}_{r,t+1,s} = 1$:

$$c(t+1,r) = \frac{\sum_{j=1}^S \alpha_{r,t+1,j}}{\sum_{k=1}^S \alpha_{r,t,k}} = \sum_{s=1}^S \bar{\alpha}_{r,t+1,s} \tag{26}$$

On peut enfin calculer $\tilde{\alpha}_{r,t+1,i}$ en divisant $\bar{\alpha}_{r,t+1,s}$ par $c(t+1,r)$. De même pour β_{rts}, on prend :

$$\tilde{\beta}_{r,t,s} = \beta_{rts}\sum_{k=1}^S \alpha_{r,t-1,k} = \underbrace{\sum_{i=1}^S \tilde{\beta}_{r,t+1,i}a_{is}b_{r,t+1,i}}_{\bar{\beta}_{r,t,s}}\underbrace{\frac{\sum_{k=1}^S \alpha_{r,t-1,k}}{\sum_{k=1}^S \alpha_{r,t,k}}}_{1/c(t,r)}$$

On peut donc calculer $\bar{\beta}_{r,t,s}$ à partir de $\tilde{\beta}_{r,t+1,s}$ puis, pour obtenir $\tilde{\beta}_{r,t,s}$ diviser le résultat par $c(t,r)$ calculé précédemment. L'initialisation de $\tilde{\beta}_{r,T_r,s}$ peut se faire avec une valeur quelconque, dans la mesure où les valeurs de $\xi_{rtij}^{(p)}$, $\gamma_{rts}^{(p)}$ n'en dépendent pas. On prend $\tilde{\beta}_{r,T_r,s} = 1$. Ce qui

conduit à l'algorithme suivant :

- initialiser : $\alpha_{r1s} = \pi_s b_{r,1,s}$
- calculer : $c(1,r) = \sum_s \alpha_{r1s}$ puis $\tilde{\alpha}_{r1s} = \alpha_{r1s}/c(1,r)$
- calculer pour $s = (1:S),\ t = (2:T_r)$: $\bar{\alpha}_{r,t+1,s} = \left(\sum_{i=1}^{S} \tilde{\alpha}_{r,t,i} a_{is}\right) b_{r,t+1,s}$
- calculer : $c(t+1,r) = \sum_{s=1}^{S} \bar{\alpha}_{r,t+1,s}$
- calculer : $\tilde{\alpha}_{r,t+1,s} = \dfrac{\bar{\alpha}_{r,t+1,s}}{c(t+1,r)}$
- initialiser : $\tilde{\beta}_{r,T_r,s} = 1$
- calculer pour $s = (1:S),\ t = (T_r - 1 : -1 : 1)$: $\bar{\beta}_{r,t,s} = \sum_{i=1}^{S} \tilde{\beta}_{r,t+1,i} a_{is} b_{r,t+1,i}$
- calculer : $\tilde{\beta}_{r,t,s} = \bar{\beta}_{r,t,s}/c(t,r)$

On peut alors calculer $\gamma_{rts}^{(p)}$ et $\xi_{rtij}^{(p)}$ en remplaçant simplement, dans les formules (24) et (25), $\alpha_{rts}^{(p)}$ et $\beta_{rts}^{(p)}$ par $\tilde{\alpha}_{rts}^{(p)}$ et $\tilde{\beta}_{rts}^{(p)}$. Ce qui donne :

$$\gamma_{rts}^{(p)} = \frac{\tilde{\alpha}_{rts}^{(p)} \tilde{\beta}_{rts}^{(p)}}{\sum_{i=1}^{S} \tilde{\alpha}_{rti}^{(p)} \tilde{\beta}_{rti}^{(p)}}$$
$$\xi_{rtij}^{(p)} = \frac{\tilde{\alpha}_{rti}^{(p)} \tilde{\beta}_{r,t+1,j}^{(p)} b_{r,t+1,j}^{(p)} a_{ij}^{(p)}}{\sum_{s=1}^{S} \tilde{\alpha}_{rts}^{(p)} \tilde{\beta}_{rts}^{(p)}}$$

D'après (18), on a :

$$\ell(t,r) = p(u_1^r, \cdots, u_t^r) = \sum_{s=1}^{S} p(u_1^r, \cdots, u_t^r, Q_t^r = s) = \sum_{s=1}^{S} \alpha_{rts}$$

D'après (26), on a :

$$c(t+1,r) = \sum_{s=1}^{S} \bar{\alpha}_{r,t+1,s} = \frac{\sum_{s=1}^{S} \alpha_{r,t+1,s}}{\sum_{k=1}^{S} \alpha_{r,t,k}} = \frac{\ell(t+1,r)}{\ell(t,r)}$$

qui donne par récurrence :

$$\log(\ell(t,r)) = \sum_{u=1}^{t} \log(c(u,r))$$

où on a posé $c(1,r) = \ell(1,r) = \pi_s b_{r,1,s}$. On en déduit la log-vraisemblance de l'ensemble de la séquence d'apprentissage :

$$\Lambda(U_{1:T_R}^{1:R}) = \sum_{r=1}^{R} \log p(u_1^r, \cdots, u_{T_r}^r) = \sum_{r=1}^{R} \log(\ell(T_r,r)) = \sum_{r=1}^{R} \sum_{u=1}^{T_r} \log(c(u,r))$$

3.5 Initialisation

Comme dans tout algorithme de maximisation, le problème du choix de la valeur initiale est cruciale. Ici une façon simple d'opérer consiste à découper la suite des T vecteurs, correspondant au mot à apprendre, en S portions égales. Ce qui consiste à supposer que les S états sont d'égales durées (!). On estime alors, dans chaque portion, les moyennes et les variances respectives de chaque composante. Cela donne les S vecteurs moyennes initiaux $\mu_{1:S}$ et les S matrices diagonales initiales $\Sigma_{1:S}$. En ce qui concerne la matrice de transition, on choisit, comme valeur initiale, une matrice de dimension $S \times S$, avec $a_{ij} = 0$ pour $i > j$ (HMM gauche-droite) et $a_{i,i-1} = a_{ii} = a_{i,i+1} = 1/3$ (équirépartition des 2 états à droite). Rappelons que le vecteur de probabilité de l'état initial est supposé connu et égal à $\Pi = (1, 0, \cdots, 0)$.