TP1 "MODÉLISER L'ALÉA"
CHAÎNES DE MARKOV CACHÉES

PIERRE GIRARDEAU, JEAN-PHILIPPE CHANCELIER, AND JEAN-FRANÇOIS DELMAS

On s'intéresse dans ce TP à deux applications de la théorie des chaînes de Markov cachées : l'estimation d'une loi de mélange de populations et la recherche de zones homogènes dans l'ADN. Les exercices ainsi que la présentation sont largement inspirés de l'ouvrage de Delmas et Jourdain (2006, Chap. 5), dans lequel on trouvera des explications plus détaillées ainsi que les démonstrations des résultats énoncés ici.

1. Présentation

On rappelle la notation condensée suivante x_m^n pour le vecteur $(x_m, \dots, x_n)$ avec $m \leq n \in \mathbb{Z}$.

On considère $S = (S_n, n \geq 1)$ une chaîne de Markov à valeurs dans $\mathcal{I}$, un espace fini non réduit à un élément, de matrice de transition a et de loi initiale π_0. Soit $(Y_n, n \geq 1)$ une suite de variables à valeurs dans $\mathcal{X}$, un espace d'état fini, telle que conditionnellement à S les variables aléatoires $(Y_n, n \geq 1)$ sont indépendantes et la loi de Y_k sachant S ne dépend que de la valeur de S_k. Plus précisément, pour tout $N \geq 1$, conditionnellement à S_1^N, les variables aléatoires Y_1^N sont indépendantes : pour tous $N \geq 1$, $y_1^N \in \mathcal{X}^N$ et $s_1^N \in \mathcal{I}^N$, on a

$$\mathbb{P}(Y_1^N = y_1^N | S_1^N = s_1^N) = \prod_{n=1}^{N} \mathbb{P}(Y_n = y_n | S_1^N = s_1^N). \tag{1}$$

De plus il existe une matrice $b = (b(i,x); i \in \mathcal{I}, x \in \mathcal{X})$, telle que

$$\mathbb{P}(Y_n = y_n | S_1^N = s_1^N) = \mathbb{P}(Y_n = y_n | S_n = s_n) = b(s_n, y_n). \tag{2}$$

Lemme 1. *La suite $((S_n, Y_n), n \geq 1)$ est une chaîne de Markov. On a pour tous $n \geq 2$, $s_1^n \in \mathcal{I}^n$ et $y_1^n \in \mathcal{X}^n$,*

$$\mathbb{P}(S_n = s_n, Y_n = y_n | S_1^{n-1} = s_1^{n-1}, Y_1^{n-1} = y_1^{n-1}) = a(s_{n-1}, s_n) b(s_n, y_n),$$

et

$$\mathbb{P}(S_n = s_n | S_1^{n-1} = s_1^{n-1}, Y_1^{n-1} = y_1^{n-1}) = a(s_{n-1}, s_n).$$

Dans le modèle de chaîne de Markov cachée, lors d'une réalisation, on observe simplement y_1^N, une réalisation de Y_1^N. Les variables S_1^N sont appelées variables cachées, et leur valeur prise lors d'une réalisation, les valeurs cachées. Dans ce modèle, on cherche à estimer, à partir de l'observation y_1^N, le paramètre $\theta = (a, b, \pi_0)$ puis à calculer, pour $i \in \mathcal{I}$, les probabilités $\mathbb{P}(S_n = i | Y_1^N = y_1^N)$. L'ensemble des paramètres possibles forme un compact Θ de $[0,1]^{\mathcal{I}^2} \times [0,1]^{\mathcal{I} \times \mathcal{X}} \times [0,1]^{\mathcal{I}}$.

1.1. Maximum de vraisemblance.

On écrit $\mathbb{P}_\theta$ et $\mathbb{E}_\theta$ pour les probabilités et espérances calculées quand le vrai paramètre de la chaîne de Markov $((S_n, Y_n), n \geq 1)$ est $\theta = (a, b, \pi_0)$. Pour abréger les notations, on notera $S = S_1^N$, $s = s_1^N \in \mathcal{I}^N$, $Y = Y_1^N$ et $y = y_1^N \in \mathcal{X}$.

Date: 17 mai 2010.

Définition 1. La vraisemblance du modèle incomplet est définie par :

$$p_N(\theta; y) = \mathbb{P}_\theta(Y = y).$$

On a, en utilisant (1) et (2) :

$$\begin{aligned} p_N(\theta; y) &= \sum_{s \in \mathcal{I}^N} \mathbb{P}_\theta(Y = y | S = s) \mathbb{P}_\theta(S = s) \\ &= \sum_{s \in \mathcal{I}^N} \left(\prod_{n=1}^N b(s_n, y_n) \right) \pi_0(s_1) \prod_{n=2}^N a(s_{n-1}, s_n). \end{aligned} \tag{3}$$

La vraisemblance du modèle incomplet est donc la probabilité que la suite de valeurs y soit observée, si le modèle sous-jacent est donné par le paramètre θ. On définit également la log-vraisemblance par :

$$L_N(\theta; y) = \log p_N(\theta; y).$$

Notre objectif étant de trouver le paramètre le plus probable sachant l'observation y, on est naturellement amené à introduire la définition suivante.

Définition 2. L'Estimateur du Maximum de Vraisemblance (EMV) est donné par :

$$\hat{\theta}(y) = \arg \max_\theta p_N(\theta, y).$$

Pour déterminer l'EMV de θ, il faut donc maximiser $p_N(\cdot; y)$ en $\theta = (a, b, \pi)$ ou, la fonction log étant croissante, maximixer la fonction $L_N(\cdot; y)$. Bien sûr, il faut tenir compte des contraintes suivantes : $\sum_{j \in \mathcal{I}} a(i, j) = 1$ pour tout $i \in \mathcal{I}$ (a est la matrice de transition d'une chaîne de Markov), $\sum_{x \in \mathcal{X}} b(i, x) = 1$ pour tout $i \in \mathcal{I}$ ($b(i, \cdot)$ est une probabilité) et $\sum_{i \in \mathcal{I}} \pi_0(i) = 1$ (π_0 est la loi de S_1).

Pour calculer numériquement l'EMV, remarquons qu'il faut maximiser $p_N(\theta; y)$, un polynôme de degré $2N$ à $\mathrm{Card}(\mathcal{I}^2 \times (\mathcal{I} \times \mathcal{X}) \times \mathcal{I})$ variables sous $2\mathrm{Card}(\mathcal{I}) + 1$ contraintes linéaires libres. Pour des applications courantes, on ne peut pas espérer calculer numériquement l'EMV par des algorithmes classiques d'optimisation. On peut, en revanche, utiliser des algorithmes de recuit simulé.

1.2. **Algorithme EM.** Nous considérons ici une autre approche : l'algorithme EM (Espérance Maximisation). Soit la vraisemblance du modèle complet définie par $p_N^{complet}(\theta; s, y) = \mathbb{P}_\theta(S = s, Y = y)$. On a

$$p_N^{complet}(\theta; s, y) = \pi_0(s_1) b(s_1, y_1) \prod_{n=2}^N a(s_{n-1}, s_n) b(s_n, y_n).$$

Soit, de plus, la loi conditionnelle de S sachant Y donnée par :

$$\pi_N(\theta; s|y) = \mathbb{P}_\theta(S = s | Y = y) = \frac{\mathbb{P}_\theta(S = s, Y = y)}{\mathbb{P}_\theta(Y = y)} = \frac{p_N^{complet}(\theta; s, y)}{p_N(\theta; y)}. \tag{4}$$

Il est alors facile de vérifier que l'on peut écrire la log-vraisemblance sous la forme :

$$L_N(\theta; y) = Q(\theta, \theta') - \mathcal{H}_{\theta'}(\theta),$$

avec, pour $y \in \mathcal{X}^N$:

$$Q(\theta, \theta') = \sum_{s \in \mathcal{I}^N} \pi_N(\theta'; s|y) \log p_N^{complet}(\theta; s, y), \tag{5}$$

et

$$\mathcal{H}_{\theta'}(\theta) = \sum_{s \in \mathcal{I}^N} \pi_N(\theta'; s|y) \log \pi_N(\theta; s|y).$$

On a alors le lemme suivant.

Lemme 2. *Soit θ' fixé. Soit θ^* le (ou un) paramètre qui maximise la fonction $\theta \mapsto Q(\theta, \theta')$. Alors $L_N(\theta^*; y) \geq L_N(\theta'; y)$.*

L'algorithme EM consiste à construire par récurrence une suite de paramètres $(\theta^{(r)}, r \in \mathbb{N})$ de la manière suivante :
- $\theta^{(0)}$ est choisi de manière quelconque.
- On suppose $\theta^{(r)}$ construit. On calcule $Q(\theta, \theta^{(r)})$. Il s'agit d'un calcul d'espérance (étape E).
- Puis, on choisit $\theta^{(r+1)}$ tel que la fonction $\theta \mapsto Q(\theta, \theta^{(r)})$ atteigne son maximum en la valeur $\theta^{(r+1)}$. Il s'agit d'une maximisation (étape M).

D'après le lemme précédent, la suite $(L_N(\theta^{(r)}; y), r \in \mathbb{N})$ est donc croissante.

2. Crabes de Weldon

2.1. **Introduction.** À la fin du *XIX*ème siècle, Weldon mesure le rapport entre la largeur du front et la longueur du corps de 1 000 crabes de la baie de Naples. Le tableau 1 donne le nombre d'individus observés sur 29 intervalles pour le rapport des deux mesures (les mesures sont faites avec une précision du dixième de millimètre, et la longueur moyenne d'un animal est de 35 millimètres).

Intervalle	Nombre	Intervalle	Nombre
$[0.580, 0.584[$	1	$[0.640, 0.644[$	74
$[0.584, 0.588[$	3	$[0.644, 0.648[$	84
$[0.588, 0.592[$	5	$[0.648, 0.652[$	86
$[0.592, 0.596[$	2	$[0.652, 0.656[$	96
$[0.596, 0.600[$	7	$[0.656, 0.660[$	85
$[0.600, 0.604[$	10	$[0.660, 0.664[$	75
$[0.604, 0.608[$	13	$[0.664, 0.668[$	47
$[0.608, 0.612[$	19	$[0.668, 0.672[$	43
$[0.612, 0.616[$	20	$[0.672, 0.676[$	24
$[0.616, 0.620[$	25	$[0.676, 0.680[$	19
$[0.620, 0.624[$	40	$[0.680, 0.684[$	9
$[0.624, 0.628[$	31	$[0.684, 0.688[$	5
$[0.628, 0.632[$	60	$[0.688, 0.692[$	0
$[0.632, 0.636[$	62	$[0.692, 0.696[$	1
$[0.636, 0.640[$	54		

Table 1. Nombre de crabes de la baie de Naples (sur un total de 1 000 crabes) dont le ratio de la largeur du front par la longueur du corps sont dans les intervalles (Weldon, 1893).

Question 1. À l'aide des données du fichier `crabe.txt`, représenter la loi empirique des données. Représenter également la loi gaussienne la plus proche de la loi empirique des données. Est-il satisfaisant d'expliquer les données à l'aide d'une loi normale ?

Question subsidiaire 1. À l'aide d'un test du χ^2, tester l'hypothèse de normalité des données.

2.2. **Estimation d'une loi de mélange.** Nous allons tenter de modéliser la répartition des données à l'aide d'une loi de mélange. En d'autres termes, nous supposons que les crabes proviennent de plusieurs populations différentes dont la répartition des données au sein de chacune d'entre elles suit une loi gaussienne.

Ainsi, soit $I \geq 2$ fixé le nombre de populations différentes supposé. Le modèle de population est alors donné par une suite $(Z_n, Y_n)_{n\geq 1}$ de variables aléatoires indépendantes de même loi. La variable Z_n, à valeurs dans $\mathcal{I} = \{1, \dots, I\}$ et de loi $\pi = (\pi(i), i \in \mathcal{I})$, indique à quelle population appartient l'échantillon n (c'est la variable cachée). La variable Y_n, à valeurs dans $\mathbb{R}$ est la variable observée. On note f_{μ_i,σ_i} la densité associée à la loi de Y_n sachant $Z_n = i$. Il s'agira d'une densité gaussienne, de moyenne μ_i et de variance σ_i^2. L'objectif de l'exercice est d'estimer les meilleurs paramètres μ et σ au sens du maximum de vraisemblance.

Pour déterminer la loi de Y_n, remarquons que pour tous $a < b$, on a, en utilisant la loi de Y_n sachant Z_n,

$$\begin{aligned}\mathbb{P}(Y_n \in [a,b]) &= \sum_{i\in\mathcal{I}} \mathbb{P}(Y_n \in [a,b] | Z_n = i)\mathbb{P}(Z_n = i) \\ &= \sum_{i\in\mathcal{I}} \pi_i \int_{[a,b]} f_{\mu_i,\sigma_i}(y)\, dy = \int_{[a,b]} f_\theta(y)\, dy,\end{aligned}$$

avec $f_\theta = \sum_{i\in\mathcal{I}} \pi_i f_{\mu_i,\sigma_i}$. Ainsi, Y_n est une variable continue de densité f_θ. Comme les variables $(Y_n, n \geq 1)$ sont indépendantes, la vraisemblance du modèle associé à l'échantillon Y_1^N est pour $y = y_1^N \in \mathbb{R}^N$:

$$p_N(\theta; y) = \prod_{k=1}^N f_\theta(y_k),$$

et la log-vraisemblance

$$L_N(\theta; y) = \sum_{k=1}^N \log f_\theta(y_k).$$

La vraisemblance du modèle complet associé à l'échantillon (Z_1^N, Y_1^N) est pour $z = z_1^N \in \mathcal{I}^N$, $y = y_1^N \in \mathbb{R}^N$:

$$p_N^{complet}(\theta; z, y) = \prod_{k=1}^N \pi_{z_k} f_{\mu_{z_k},\sigma_{z_k}}(y_k).$$

2.3. **Étape E.** L'étape E consiste à expliciter la fonction Q définie dans l'équation (5). Remarquons d'abord que :

$$\log p_N^{complet}(\theta; z, y) = \sum_{k=1}^N \left[\log(\pi_{z_k}) + \log(f_{\mu_{z_k},\sigma_{z_k}}(y_k))\right].$$

De plus :

$$\pi_N(\theta'; z|y) = \frac{p_N^{complet}(\theta'; z, y)}{p_N(\theta'; y)} = \prod_{k=1}^N \rho'_{z_k,k},$$

avec, pour tout $i \in \mathcal{I}, k \in \{1, \dots, N\}$:

$$\rho'_{i,k} = \frac{\pi'_i f_{\mu'_i,\sigma'_i}(y_k)}{f_{\theta'}(y_k)}. \tag{6}$$

La quantité $\rho'_{i,k}$ s'interprète comme la probabilité que $Z_k = i$ sachant $Y_k = y_k$, θ' étant le paramètre du modèle. La quantité $\pi_N(\theta'; z|y)$ s'interprète comme la loi conditionnelle des variables cachées Z_1^N sachant les variables observées Y_1^N.

Comme pour tout $l \in \{1, \dots, N\}$, on a $\sum_{j\in\mathcal{I}} \rho'_{j,l} = 1$, il vient

$$\sum_{z\in\mathcal{I}^N; z_k=i} \pi_N(\theta'; z|y) = \rho'_{i,k}.$$

Cette égalité représente le calcul de la loi marginale de Z_k sachant Y_k. On en déduit donc que

$$Q(\theta, \theta') = \sum_{k=1}^{N} \sum_{i \in \mathcal{I}} \rho'_{i,k} [\log(\pi_i) + \log(f_{\mu_i, \sigma_i}(y_k))].$$

2.4. **Étape M.** L’étape M consiste à maximiser $Q(\theta, \theta')$ en θ. Pour ce faire, on écrit Q sous la forme suivante.

$$Q(\theta, \theta') = N \underbrace{\sum_{i \in \mathcal{I}} \pi_i^* \log \pi_i}_{A_0} + \sum_{j \in \mathcal{I}} \underbrace{\sum_{k=1}^{N} \rho'_{j,k} \log(f_{\mu_j, \sigma_j}(y_k))}_{A_j},$$

avec :

$$\pi_i^* = \frac{1}{N} \sum_{k=1}^{N} \rho'_{i,k}, \tag{7}$$

Remarquons que maximiser $Q(\theta, \theta')$ en $\theta \in \Theta'$ revient à maximiser séparément A_0, sous la contrainte que $\pi \in \mathcal{P}_{\mathcal{I}}$, et A_j pour $j \in \mathcal{I}$.

Question 2. Montrer que A_0 est maximal pour $\pi = \pi^*$ et que A_j est maximal pour :

$$\mu_j^* = \frac{\sum_{k=1}^{N} \rho'_{j,k} y_k}{\sum_{k=1}^{N} \rho'_{j,k}} \quad \text{et} \quad (\sigma_j^*)^2 = \frac{\sum_{k=1}^{N} \rho'_{j,k} (y_k - \mu_j^*)^2}{\sum_{k=1}^{N} \rho'_{j,k}}. \tag{8}$$

2.5. **Algorithme EM.** On peut maintenant appliquer l’algorithme à l’aide des relations (7) et (8).

Question 3. Appliquer l’algorithme EM (le squelette de l’algorithme est donné dans le fichier `crabe-tp.sce`) afin d’obtenir la loi de mélange en supposant l’existence de deux populations. Comparer la loi de mélange obtenue à la loi empirique.

Question subsidiaire 2. Appliquer l’algorithme EM en supposant que l’on a affaire à trois populations de crabes.

3. Recherche de zones homogènes dans l’ADN

3.1. **Introduction.** Le bactériophage lambda est un parasite de la bactérie Escherichia coli. Son ADN (acide désoxyribonucléique) circulaire comporte $N_0 = 48\,502$ paires de nucléotides, et il est essentiellement constitué de régions codantes, i.e. de régions lues et traduites en protéines. La transcription, c’est-à-dire la lecture de l’ADN, s’effectue sur des parties de chacun des deux brins qui forment la double hélice de l’ADN. Ainsi sur la séquence d’ADN d’un seul brin on peut distinguer deux types de zones : celles où la transcription a lieu sur le brin et celles où la transcription a lieu sur le brin apparié. On observe sur les parties codantes une certaine fréquence d’apparition des différents nucléotides Adénine (`A`), Cytosine (`C`), Guanine (`G`) et Thymine (`T`). Le nucléotide `A` (resp. `C`) d’un brin est apparié avec le nucléotide `T` (resp. `G`) du brin apparié et vice versa. Les deux types de zones d’un brin décrites plus haut correspondent en fait à des fréquences d’apparitions différentes des quatre nucléotides. Les biologistes ont d’abord analysé l’ADN du bactériophage lambda en identifiant les gènes de l’ADN, c’est-à-dire les parties codantes de l’ADN, et les protéines correspondantes. Et ils ont ainsi constaté que les deux brins de l’ADN comportaient des parties codantes. Il est naturel de vouloir détecter a priori les parties codantes, ou susceptibles d’être codantes, à partir d’une analyse statistique de l’ADN. Cela peut permettre aux biologistes d’identifier plus rapidement les parties codantes pour les organismes dont la séquence d’ADN est

connue. Dans l'exemple, on se sait pas si le k-ième nucléotide observé appartient à une zone transcrite ou à une zone appariée à une zone transcrite. Le brin transcrit au niveau du k-ième nucléotide est donc une variable cachée que l'on désire retrouver.

Nous présentons brièvement le modèle mathématique pour la séquence d'un brin d'ADN, $y_1 \dots y_{N_0}$ du bactériophage lambda. À la séquence d'ADN, on peut associer la séquence non observée, dite séquence cachée, $s_1 \dots s_{N_0}$, où si $s_k = +1$, alors y_k est la réalisation d'une variable aléatoire, Y_k, de loi p_+ sur $\mathcal{X} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, et si $s_k = -1$ alors la loi de Y_k est p_-. Les probabilités p_+ et p_- sont distinctes mais inconnues. On modélise la suite $s_1, \dots, s_{N_0}$ comme la réalisation d'une chaîne de Markov, $(S_n, n \geq 1)$, sur $\mathcal{I} = \{+1, -1\}$ de matrice de transition, a, également inconnue.

3.2. **Étape E.** Avec les mêmes notations que dans l'exemple précédent, on calcule $Q(\theta, \theta')$ pour $y \in \mathcal{X}^N$, dont on peut montrer qu'il s'exprime ici :

$$\begin{aligned} Q(\theta, \theta') = &\sum_{i \in \mathcal{I}} \mathbb{P}_{\theta'}(S_1 = i | Y = y) \log\, \pi_0(i) \\ &+ \sum_{n=1}^{N} \sum_{i \in \mathcal{I}} \mathbb{P}_{\theta'}(S_n = i | Y = y) \log\, b(i, y_n) \\ &+ \sum_{n=2}^{N} \sum_{i,j \in \mathcal{I}} \mathbb{P}_{\theta'}(S_{n-1} = i, S_n = j | Y = y) \log\, a(i, j). \end{aligned}$$

Nous devons donc calculer, pour la chaîne de Markov de paramètre θ', les probabilités $\mathbb{P}_{\theta'}(S_{n-1} = i, S_n = j | Y = y)$ pour $2 \leq n \leq N$ et $\mathbb{P}_{\theta'}(S_n = i | Y = y)$ pour $1 \leq n \leq N$. Pour résoudre ce problème, appelé problème de **filtrage**, on effectue les étapes suivantes :

(1) Prédire la valeur de S_n connaissant les observations partielles jusqu'à l'instant $n-1$. Il s'agit de la prévision.

(2) Estimer la valeur de S_n connaissant les observations partielles jusqu'à l'instant n. Il s'agit du filtrage.

(3) Estimer la valeur de S_n connaissant les observations partielles jusqu'à l'instant final N. Il s'agit du lissage.

Lemme 3 (Prévision)**.** *On a, pour $n \geq 2$, $y_1^{n-1} \in \mathcal{X}^{n-1}$,*

$$\mathbb{P}_{\theta'}(S_n = i | Y_1^{n-1} = y_1^{n-1}) = \sum_{j \in \mathcal{I}} a'(j, i) \mathbb{P}_{\theta'}(S_{n-1} = j | Y_1^{n-1} = y_1^{n-1}).$$

Lemme 4 (Filtrage)**.** *On a, pour $n \geq 1$, $y_1^n \in \mathcal{X}^n$,*

$$\mathbb{P}_{\theta'}(S_n = i | Y_1^n = y_1^n) = \frac{b'(i, y_n) \mathbb{P}_{\theta'}(S_n = i | Y_1^{n-1} = y_1^{n-1})}{\sum_{j \in \mathcal{I}} b'(j, y_n) \mathbb{P}_{\theta'}(S_n = j | Y_1^{n-1} = y_1^{n-1})}.$$

Remarquons que les termes de prévision à l'instant n s'écrivent en fonction des termes de filtrage à l'instant $n-1$. Ces derniers s'écrivent en fonction des termes de prévision à l'instant $n-1$. On en déduit que l'on peut donc calculer les termes de prévision et de filtrage à l'instant n en fonction de a', b' et $\mathbb{P}_{\theta'}(S_1 = i | Y_1 = y_1)$. Or d'après la formule de Bayes, on a

$$\mathbb{P}_{\theta'}(S_1 = i | Y_1 = y_1) = \frac{\mathbb{P}_{\theta'}(S_1 = i, Y_1 = y_1)}{\sum_{j \in \mathcal{I}} \mathbb{P}_{\theta'}(S_1 = j, Y_1 = y_1)} = \frac{b'(i, y_1) \pi_0'(i)}{\sum_{j \in \mathcal{I}} b'(j, y_1) \pi_0'(j)}.$$

On en déduit donc que l'on peut exprimer les termes de prévision et de filtrage en fonction de $\theta' = (a', b', \pi_0')$.

Lemme 5 (Lissage). *On a, pour* $2 \leq n \leq N$, $y_1^N \in \mathcal{X}^N$,

$$\mathbb{P}_{\theta'}(S_{n-1} = i, S_n = j | Y_1^N = y_1^N) = a'(i,j) \frac{\mathbb{P}_{\theta'}(S_{n-1} = i | Y_1^{n-1} = y_1^{n-1})}{\mathbb{P}_{\theta'}(S_n = j | Y_1^{n-1} = y_1^{n-1})} \mathbb{P}_{\theta'}(S_n = j | Y_1^N = y_1^N),$$

et, pour $1 \leq n \leq N-1$, $y_1^N \in \mathcal{X}^N$,

$$\mathbb{P}_{\theta'}(S_n = j | Y_1^N = y_1^N) = \sum_{l \in \mathcal{I}} a'(j,l) \frac{\mathbb{P}_{\theta'}(S_n = j | Y_1^n = y_1^n)}{\mathbb{P}_{\theta'}(S_{n+1} = l | Y_1^n = y_1^n)} \mathbb{P}_{\theta'}(S_{n+1} = l | Y_1^N = y_1^N).$$

Remarquons que le calcul de $\mathbb{P}_{\theta'}(S_N = j | Y_1^N = y_1^N)$ provient des équations de filtrage et de prévision. Son calcul nécessite le parcours complet de la suite $y = y_1^N$. À partir de cette quantité, on déduit des équations de lissage que l'on peut calculer par récurrence descendante $\mathbb{P}_{\theta'}(S_n = j | Y_1^N = y_1^N)$ (on part donc de $n = N$). Et parallèlement, on peut calculer les quantités $\mathbb{P}_{\theta'}(S_{n-1} = i, S_n = j | Y_1^N = y_1^N)$. Ces calculs nécessitent le parcours complet de la suite $y = y_1^N$ de 1 à N puis de N à 1. On fait référence à ces calculs sous le nom d'algorithme "forward-backward". On a ainsi calculé les coefficients de $Q(\theta, \theta')$ qui sont fonction de $\theta' = (a', b', \pi_0')$.

3.3. **Étape M.** On peut montrer que $Q(\theta, \theta')$ est maximal pour $\theta = (a, b, \pi_0)$ définis pour $i, j \in \mathcal{I}, x \in \mathcal{X}$ par

$$\begin{aligned} b(i,x) &= \frac{\sum_{n=1}^N \mathbf{1}_{\{y_n = x\}} \mathbb{P}_{\theta'}(S_n = i | Y = y)}{\sum_{n=1}^N \mathbb{P}_{\theta'}(S_n = i | Y = y)}, \\ a(i,j) &= \frac{\sum_{n=2}^N \mathbb{P}_{\theta'}(S_{n-1} = i, S_n = j | Y = y)}{\sum_{l \in \mathcal{I}} \sum_{n=2}^N \mathbb{P}_{\theta'}(S_{n-1} = i, S_n = l | Y = y)} \\ &= \frac{\sum_{n=2}^N \mathbb{P}_{\theta'}(S_{n-1} = i, S_n = j | Y = y)}{\sum_{n=1}^{N-1} \mathbb{P}_{\theta'}(S_{n-1} = i | Y = y)}, \\ \pi_0(i) &= \mathbb{P}_{\theta'}(S_1 = i | Y = y). \end{aligned}$$

3.4. **Algorithme EM.** On peut maintenant appliquer l'algorithme à l'aide des relations obtenues ci-dessus.

Question 4. Appliquer l'algorithme EM (le squelette de l'algorithme est donné dans le fichier `hmc-tp.sce`) pour estimer les matrices π_0, a et b. On pourra restreindre, dans un premier temps, la taille de la chaîne d'ADN afin de limiter le temps de calcul. Observer en sortie l'estimation des zones homogènes, ainsi que l'évolution des termes des matrices en jeu au cours de l'algorithme.

Références

Delmas, J.-F. et Jourdain, B. (2006). Modèles Aléatoires : Applications aux Sciences de L'ingénieur et du Vivant. *Springer-Verlag*, Coll. Mathématiques et Applications.

E-mail address: `pierre.girardeau@cermics.enpc.fr`

E-mail address: `jpc@cermics.enpc.fr`

E-mail address: `delmas@cermics.enpc.fr`