5

Recherche de zones homogènes dans l'ADN

Le bactériophage lambda est un parasite de la bactérie Escherichia coli. Son ADN (acide désoxyribonucléique) circulaire comporte $N_0 = 48\,502$ paires de nucléotides (voir [15]), 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.

Les paragraphes qui suivent montrent comment, en modélisant la séquence d'ADN comme une réalisation partielle d'une chaîne de Markov, on peut détecter les zones où les fréquences d'apparitions des quatre nucléotides sont significativement différentes. L'algorithme EM (Espérance Maximisation) que nous présentons et son utilisation pour l'analyse de l'ADN ont été étudiés en détail et dans un cadre plus général par Muri [12]. De nombreux travaux récents permettent d'améliorer l'algorithme EM pour la détection de zones intéressantes de l'ADN en tenant compte d'informations biologiques connues a priori, voir par exemple les travaux du Laboratoire Statistique et Génome (`http://stat.genopole.cnrs.fr`).

Les algorithmes EM ont été initialement introduits en 1977 par Dempster, Laid et Rubin [8]. Ils sont utilisés pour l'estimation de paramètres dans des modèles où des variables sont cachées, c'est-à-dire non observées (voir par exemple [5] p. 213). Dans l'exemple ci-dessus, avec l'interprétation biologique que l'on espère retrouver, on ne 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. Il existe de nombreuses applications des algorithmes EM, voir par exemple [10]. Signalons, sans être exhaustif, que ces algorithmes sont utilisés dans les domaines suivants :

- Classification ou étude de données mélangées dont les sources sont inconnues : pour les données mélangées voir par exemple [11], pour l'analyse d'image voir par exemple [7 et 9], voir aussi l'exemple du paragraphe 5.6.1.
- Analyse de données censurées ou tronquées, voir le problème du paragraphe 5.6.2.
- Estimation de matrice de covariance avec des données incomplètes, etc.

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.

Remarque. La matrice de transition a est de la forme

$$a = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon' & 1-\varepsilon' \end{pmatrix},$$

où ε et ε' sont intuitivement inversement proportionnels à la longueur moyenne des zones homogènes où les fréquences d'apparitions des quatre nucléotides sont constantes. En effet, si par exemple $S_1 = +1$, la loi du premier instant où la chaîne de Markov change d'état, $T = \inf\{k \geq 1\,;\ S_{k+1} \neq +1\}$, suit une loi géométrique de paramètre ε car

$$\mathbb{P}(T = k | S_1 = +1) = \mathbb{P}(S_{k+1} = -1, S_k = 1, \dots, S_2 = 1 | S_1 = 1) = (1-\varepsilon)^{k-1}\varepsilon,$$

et son espérance vaut $1/\varepsilon$. ◊

Le modèle utilisé comporte une chaîne de Markov, S, qui n'est pas directement observée. Ce type de modèle, dit modèle de chaînes de Markov cachées, est présenté de manière détaillée au paragraphe 5.1. Pour identifier les zones homogènes, il faut estimer les paramètres inconnus a, p_+ et p_-. Pour cela, on utilisera les estimateurs du maximum de vraisemblance (EMV) qui possèdent de bonnes propriétés. La construction de ces estimateurs ainsi que leurs propriétés asymptotiques sont présentées dans le paragraphe 5.2 au

travers d'un exemple élémentaire et dans un cadre simple. La convergence de l'EMV vers les paramètres inconnus du modèle dans le cadre des chaînes de Markov cachées est plus complexe à établir. Ce résultat et sa démonstration technique sont reportés au paragraphe 5.5. Nous verrons au paragraphe 5.3, qu'il est impossible de calculer explicitement l'EMV dans le cas particulier des chaînes de Markov cachées. Mais nous exhiberons une méthode, l'algorithme EM, pour en donner une bonne approximation. Le paragraphe 5.4, qui est le cœur de ce chapitre, présente la mise en œuvre explicite de l'algorithme EM. En particulier, on calcule la loi des états cachés $S_1, \ldots, S_{N_0}$ sachant les observations $y_1, \ldots, y_{N_0}$, voir la Fig. 5.7 pour les valeurs de $\mathbb{P}(S_n = +1 | y_1, \ldots, y_{N_0})$ concernant l'ADN du bactériophage lambda. Dans les modèles de mélanges ou de données censurées, l'algorithme EM s'exprime simplement. Ces applications importantes, en marge des modèles de chaînes de Markov, sont abordées au paragraphe 5.6. Pour les modèles de mélanges, on utilise les données historiques des crabes de Weldon, analysées par Pearson en 1894, première approche statistique d'un modèle de mélange. Les modèles de données censurées sont évoqués au travers d'un problème. Enfin, dans la conclusion, paragraphe 5.7, nous présentons les résultats numériques obtenus pour le bactériophage lambda ainsi que quelques commentaires sur la méthode utilisée.

5.1 Chaînes de Markov cachées

On rappelle la notation condensée suivante x_m^n pour le vecteur $(x_m, \ldots, x_n)$ 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

$$\mathbb{P}(Y_1^N = y_1^N | S_1^N = s_1^N) = \prod_{n=1}^{N} \mathbb{P}(Y_n = y_n | S_1^N = s_1^N), \tag{5.1}$$

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

$$\mathbb{P}(Y_n = y_n | S_1^N = s_1^N) = \mathbb{P}(Y_n = y_n | S_n = s_n) = b(s_n, y_n). \tag{5.2}$$

Lemme 5.1.1. *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

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

Démonstration. En utilisant les égalités (5.1) et (5.2) ainsi que la propriété de Markov pour $(S_n, n \geq 1)$, on a

$$\begin{aligned}
&\mathbb{P}(S_1^n = s_1^n, Y_1^n = y_1^n) \\
&\qquad = \mathbb{P}(Y_1^n = y_1^n | S_1^n = s_1^n)\mathbb{P}(S_1^n = s_1^n) \\
&\qquad = \left(\prod_{k=1}^{n} b(s_k, y_k)\right) \mathbb{P}(S_n = s_n | S_1^{n-1} = s_1^{n-1})\mathbb{P}(S_1^{n-1} = s_1^{n-1}) \\
&\qquad = \left(\prod_{k=1}^{n} b(s_k, y_k)\right) \mathbb{P}(S_n = s_n | S_{n-1} = s_{n-1})\mathbb{P}(S_1^{n-1} = s_1^{n-1}) \\
&\qquad = \left(\prod_{k=1}^{n} b(s_k, y_k)\right) a(s_{n-1}, s_n)\mathbb{P}(S_1^{n-1} = s_1^{n-1}).
\end{aligned}$$

D'autre part, en sommant sur $y_n \in \mathcal{X}$ et en utilisant $\sum_{x\in\mathcal{X}} b(s_n, x) = 1$, puis en sommant sur $s_n \in \mathcal{I}$ et en utilisant $\sum_{s_n\in\mathcal{I}} a(s_{n-1}, s_n) = 1$, il vient

$$\mathbb{P}(S_1^{n-1} = s_1^{n-1}, Y_1^{n-1} = y_1^{n-1}) = \left(\prod_{k=1}^{n-1} b(s_k, y_k)\right) \mathbb{P}(S_1^{n-1} = s_1^{n-1}).$$

On en déduit donc la première égalité du lemme. Ainsi la suite $((S_n, Y_n), n \geq 1)$ est une chaîne de Markov.

La deuxième égalité du lemme se déduit de la première en sommant sur $y_n \in \mathcal{X}$. □

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 $[0,1]^{\mathcal{I}^2} \times [0,1]^{\mathcal{I}\times\mathcal{X}} \times [0,1]^{\mathcal{I}}$.

Remarquons que la loi du vecteur des observations Y_1^N ne détermine pas complètement le paramètre $\theta = (a, b, \pi_0)$. En effet, soit σ une permutation de $\mathcal{I}$. On note $\theta_\sigma = \Big(a(\sigma_i, \sigma_j), i \in \mathcal{I}, j \in \mathcal{I}), (b(\sigma_i, x), i \in \mathcal{I}, x \in \mathcal{X}), (\pi_0(\sigma_i), i \in \mathcal{I})\Big)$. Les paramètres θ et θ_σ génèrent la même loi pour le processus des observations Y_1^N (mais pas pour (S_1^N, Y_1^N) en général). On ne peut pas espérer distinguer θ de θ_σ à la seule vue des observations. On dit que le modèle n'est pas identifiable.

Remarquons également que s'il existe une probabilité p sur $\mathcal{X}$ telle que pour tous $i \in \mathcal{I}$, $x \in \mathcal{X}$, $b(i, x) = p(x)$ alors la suite Y est une suite de variables aléatoires indépendantes et de même loi p. En particulier la loi de Y_1^N ne dépend pas des valeurs de a et π_0.

Définition 5.1.2. *Soit X une variable aléatoire dont on peut observer les réalisations. On suppose que la loi, a priori inconnue, de X appartient à une famille de lois indicée par un paramètre : $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$, où Θ est un ensemble de paramètres. Il s'agit d'un* ***modèle paramétrique****. On dit que le modèle est* ***identifiable*** *si pour $\theta \neq \theta' \in \Theta$, on a $\mathrm{P}_\theta \neq \mathrm{P}_{\theta'}$.*

Nous reviendrons sur la notion de modèle identifiable dans le paragraphe suivant lors de la construction d'un estimateur de θ, paramètre a priori inconnu de la loi de X.

Pour que le modèle de chaîne de Markov cachée soit identifiable, il suffit de restreindre l'ensemble des paramètres possibles à un sous-ensemble Θ' de Θ. Pour cela, on vérifie qu'il est possible de choisir Θ' un ouvert de Θ, tel que si $\theta = (a, b, \pi_0) \in \Theta$, alors

- soit il existe $i \neq i'$ et pour tout $x \in \mathcal{X}$, $b(i, x) = b(i', x)$, alors on a $\theta \notin \Theta'$,
- soit pour tous $i \neq i'$, il existe $x \in \mathcal{X}$ tel que $b(i, x) \neq b(i', x)$, et alors il existe σ une permutation unique de $\mathcal{I}$ telle que $\theta_\sigma \in \Theta'$.

Ainsi si θ et θ' appartiennent à Θ' et sont distincts, alors les lois des observations Y_1^N sont différentes. Le modèle, où l'ensemble des paramètres est Θ', est alors identifiable.

5.2 L'estimateur du maximum de vraisemblance (EMV)

5.2.1 Définitions et exemples

Nous illustrons l'estimation par maximum de vraisemblance dans l'exemple qui suit.

Exemple 5.2.1. Vous désirez jouer à un jeu de pile ou face avec un adversaire. Vous savez qu'il dispose en fait de deux pièces biaisées. On note θ_i la probabilité d'obtenir pile pour la pièce $i \in \{1, 2\}$, avec $\theta_1 = 0.3$ et $\theta_2 = 0.8$. La pièce avec laquelle votre adversaire vous propose de jouer, a déjà été utilisée dans le jeu de pile ou face précédent, où vous avez observé $k_0 = 4$ piles sur $n = 10$ lancers.

Le nombre de pile obtenu lors de n lancers suit une loi binomiale de paramètre (n, θ), θ étant la probabilité d'obtenir pile. On note $p(\theta, k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$ la probabilité qu'une variable de loi binomiale de paramètre (n, θ) prenne la valeur k. On visualise ces probabilités pour $n = 10$, et $\theta \in \{\theta_1, \theta_2\}$ sur la Fig. 5.1. Il est raisonnable de supposer que la pièce utilisée est la pièce 1 car la probabilité d'observer $k_0 = 4$ est plus grande pour $\theta = \theta_1$, $p(\theta_1; k_0) \simeq 0.2$, que pour $\theta = \theta_2$, $p(\theta_2\,; k_0) \simeq 0.006$. On choisit ainsi le paramètre $\theta \in \{\theta_1, \theta_2\}$ qui maximise la fonction

$$\theta \to p(\theta\,;\ k_0) \quad \text{où } \theta \in \Theta = \{\theta_1, \theta_2\}.$$

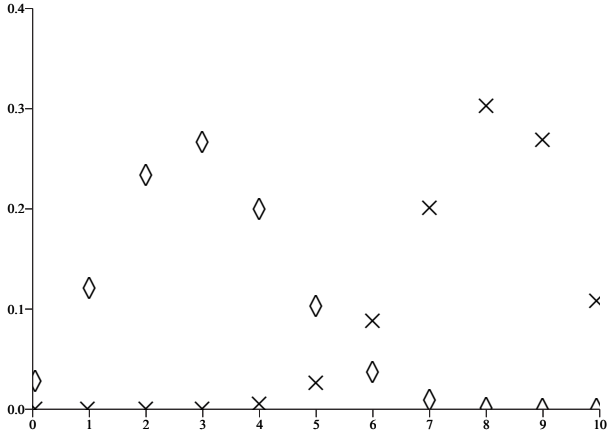

Fig. 5.1. Lois binomiales de paramètres $(10, \theta)$ avec $\theta = 0.3$ (losanges) et $\theta = 0.8$ (croix)

La fonction ci-dessus s'appelle la vraisemblance, et le paramètre qui la maximise s'appelle l'estimateur du maximum de vraisemblance. ◊

Définition 5.2.2. *On considère un modèle paramétrique : $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ une famille de lois (resp. de lois à densité) sur un espace E discret (resp. sur $E = \mathbb{R}^n$), où Θ est un ensemble de paramètres. On note $p(\theta\,;x)$ la probabilité qu'une variable de loi P_θ prenne la valeur $x \in E$ (resp. la densité en $x \in E$ d'une variable de loi P_θ). La fonction définie sur Θ, à $x \in E$ fixé, par $\theta \to p(\theta\,;x)$ s'appelle la* ***vraisemblance****.*

*Supposons que pour tout $x \in E$, il existe une unique valeur de θ, notée $\hat{\theta}(x)$, telle que la vraisemblance soit maximale en $\hat{\theta}(x)$: $p(\hat{\theta}(x)\,;x) > p(\theta, x)$ pour tous $x \in E$, $\theta \in \Theta$ et $\theta \neq \hat{\theta}(x)$. La fonction $x \to \hat{\theta}(x)$, pour $x \in E$, s'appelle l'****Estimateur du Maximum de Vraisemblance*** *(EMV) de θ.*

Comme la fonction log est strictement croissante, on peut aussi rechercher l'EMV comme la valeur de θ qui maximise la **log-vraisemblance** $\theta \to \log p(\theta\,;x)$. Cette approche est souvent techniquement plus simple. Dans le cas où la vraisemblance atteint son maximum en plusieurs points, l'EMV est mal défini, voir la remarque 5.2.10 à ce sujet.

Soit X est une variable aléatoire de loi P_{θ_0}, avec $\theta_0 \in \Theta$ inconnu.

Définition 5.2.3. *La variable aléatoire $\hat{\theta} = \hat{\theta}(X)$ est également appelée l'EMV de θ.*

Remarque 5.2.4. Si g est une bijection définie sur Θ, alors on vérifie facilement que l'EMV de $r = g(\theta)$, qui est l'EMV associé au modèle paramétrique

$\mathcal{Q} = \{\mathrm{Q}_r = \mathrm{P}_{g^{-1}(r)}, r \in g(\Theta)\}$ est $g(\hat{\theta})$, où $\hat{\theta}$ est l'EMV de θ. Par convention si g est une fonction définie sur Θ, alors l'EMV de $g(\theta)$ est $g(\hat{\theta})$. $\Diamond$

L'estimateur du maximum de vraisemblance possède de bonnes propriétés statistiques : convergence, normalité asymptotique. En revanche, cet estimateur est souvent biaisé. Ce handicap est généralement compensé par le fait qu'il permet de construire un intervalle de confiance étroit comparativement à d'autres estimateurs. Après avoir donné une définition précise aux termes précédents, nous illustrerons ces propriétés sur l'estimation de la probabilité d'avoir un garçon à la naissance.

Définition 5.2.5. *On considère un modèle paramétrique. Soit* $X = (X_n, n \geq 1)$ *une suite de variables aléatoires à valeurs dans* E*, dont la loi* P_{θ_0} *appartient à une famille de lois* $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$*, où* Θ *est un ensemble de paramètres. Le paramètre* θ_0 *est inconnu a priori.*

Un ***estimateur*** *de* θ_0 *construit à partir de* X_1^n *est une fonction explicite de* X_1^n*. En particulier, elle ne fait pas intervenir* θ_0*.*

Un estimateur $h(X_1^n)$ *de* θ_0 *est dit* ***sans biais*** *s'il est intégrable et si* $\mathbb{E}[h(X_1^n)] = \theta_0$ *pour tout* $\theta_0 \in \Theta$*. Sinon, on dit que l'estimateur est biaisé.*

Soit $(\delta_n, n \geq 1)$ *une suite d'estimateurs de* θ_0*, où* δ_n *est construit à partir de* X_1^n*. On dit que la suite* $(\delta_n, n \geq 1)$ *est un estimateur* ***convergent*** *(on dit aussi fortement convergent) de* θ_0*, si pour tout* $\theta_0 \in \Theta$*, on a p.s.*

$$\lim_{n\to\infty} \delta_n = \theta_0.$$

Si la convergence a lieu en probabilité seulement, on parle d'estimateur faiblement convergent.

On dit que la suite $(\delta_n, n \geq 1)$ *est un estimateur* ***asymptotiquement normal*** *de* θ_0*, si pour tout* $\theta_0 \in \Theta$*, on a*

$$\sqrt{n}(\delta_n - \theta_0) \xrightarrow[n\to\infty]{\textbf{Loi}} \mathcal{N}(0, \Sigma(\theta_0)),$$

où $\Sigma(\theta_0)$ *est la* ***variance asymptotique*** *(ou matrice de covariance asymptotique si le paramètre* θ *est multidimensionnel).*

Exemple 5.2.6. On suppose qu'à la naissance chaque bébé a une probabilité θ_0 d'être un garçon et une probabilité $1-\theta_0$ d'être une fille. On considère une population de n bébés et on désire donner une estimation de θ_0 afin de savoir si à la naissance il naît significativement plus de garçons que de filles ou plus de filles que de garçons ou autant de garçons que de filles.

On précise le modèle paramétrique. On note $X_i = 1$ si le $i^{\text{ème}}$ bébé est un garçon et $X_i = 0$ sinon. Il est naturel de supposer que les variables aléatoires $X = (X_i, i \geq 1)$ sont indépendantes et de même loi de Bernoulli de paramètre $\theta_0 \in \Theta = [0, 1]$.

Pour calculer l'EMV $\hat{\theta}_n$, à partir de l'échantillon X_1^n, on remarque que $\mathbb{P}(X_i = x_i) = \theta_0^{x_i}(1-\theta_0)^{1-x_i}$. Par indépendance, on en déduit que la vraisemblance est la fonction de θ définie sur Θ, pour $x = x_1^n \in \{0, 1\}^n$, par

$$p_n(\theta\,;x) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{n_1}(1-\theta)^{n-n_1},$$

où $n_1 = \sum_{i=1}^n x_i$ représente le nombre de garçons. La log-vraisemblance est définie par

$$L_n(\theta\,;x) = \log p_n(\theta\,;x) = n_1 \log(\theta) + (n-n_1)\log(1-\theta),$$

avec la convention que $0 \log 0 = 0$. Dans un premier temps on cherche θ qui annule sa dérivée :

$$\frac{\partial L_n}{\partial \theta}(\theta\,;x) = \frac{n_1}{\theta} - \frac{n-n_1}{1-\theta} = 0,$$

soit $\theta = n_1/n$. Comme $\partial L_n/\partial\theta$ est une fonction strictement décroissante, on en déduit dans un deuxième temps, que la log-vraisemblance est maximale pour $\theta = \dfrac{n_1}{n} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$. L'EMV est donc $\hat\theta_n = \frac{1}{n}\sum_{i=1}^n X_i$. On peut remarquer que dans ce cas précis l'EMV est un estimateur sans biais de θ_0.

On déduit de la loi forte des grands nombres que $(\hat\theta_n, n \geq 1)$ converge presque sûrement vers le vrai paramètre inconnu θ_0, i.e. l'EMV est convergent. On déduit du théorème central limit que $(\sqrt{n}(\hat\theta_n - \theta_0), n \geq 1)$ converge en loi vers une variable de loi gaussienne $\mathcal{N}(0, \Sigma(\theta_0))$, où $\Sigma(\theta_0) = \mathrm{Var}(X_1) = \theta_0(1-\theta_0)$. L'EMV est donc asymptotiquement normal de variance asymptotique $\Sigma(\theta_0)$.

On peut alors, en remplaçant $\Sigma(\theta_0)$ par l'estimation $\hat\theta_n(1-\hat\theta_n)$, en déduire un intervalle de confiance de θ_0 de niveau asymptotique 95 % (cf. le paragraphe A.3.4). Par exemple, aux U.S.A. en 1996 on compte $n_1 = 1\,990\,480$ naissances de garçons et $n - n_1 = 1\,901\,014$ naissances de filles. On en déduit que $\hat\theta_n = n_1/n \simeq 0.511495$. On calcule l'intervalle de confiance de niveau asymptotique 95 % pour θ : $\left[\hat\theta_n \pm 1.96\sqrt{\hat\theta_n(1-\hat\theta_n)}/\sqrt{n}\right] \simeq [0.511, 0.512]$. Il naît significativement plus de garçons que de filles. ◊

Exercice 5.2.7. On considère le modèle paramétrique gaussien. Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes et de même loi gaussienne de moyenne $\mu \in \mathbb{R}$ et de variance $\sigma^2 > 0$. Le paramètre est $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R}\times]0, \infty[$. La vraisemblance associée au vecteur X_1^n est sa densité.

1. Expliciter la vraisemblance et la log-vraisemblance du modèle gaussien.
2. Montrer que l'EMV, $\hat\theta_n = (\hat\mu_n, \hat\sigma_n^2)$, de θ associé à X_1^n est défini par la moyenne empirique et la variance empirique :

$$\hat\mu_n = \frac{1}{n}\sum_{k=1}^{n} X_k, \quad \hat\sigma_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \hat\mu_n)^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \hat\mu_n^2. \tag{5.3}$$

3. Vérifier que la moyenne empirique est un estimateur sans biais de μ, mais que la variance empirique est un estimateur biaisé de σ^2. Construire à partir de $\hat{\sigma}_n^2$ un estimateur sans biais de σ^2.
4. Vérifier que l'EMV, $\hat{\theta}_n$, est un estimateur de θ convergent et asymptotiquement normal.

♦

Sous des hypothèses assez générales sur la vraisemblance, on peut montrer que la suite d'estimateurs $(\hat{\theta}_n, n \geq 1)$, où $\hat{\theta}_n$ est l'EMV de θ construit à partir de X_1^n, est convergente et asymptotiquement normale. On pourra consulter [4], paragraphe 16, pour une démonstration précise de ces résultats quand les variables $(X_n, n \geq 1)$ sont indépendantes et de même loi.

Dans le paragraphe qui suit, nous démontrons la convergence de l'EMV, dans le cas où les variables $(X_n, n \geq 1)$, indépendantes et de même loi, dépendant d'un paramètre $\theta \in \Theta$, sont à valeurs dans un espace discret E. Des arguments similaires permettront de montrer la convergence de l'EMV pour les chaînes de Markov cachées (voir le paragraphe 5.5). Au chapitre 6, la démonstration du lemme 6.2.1 établit directement la convergence de l'EMV de la matrice de transition pour un modèle de chaîne de Markov à l'aide du théorème ergodique pour les chaînes de Markov et la normalité asymptotique à l'aide du TCL ergodique.

5.2.2 Convergence de l'EMV dans un modèle simple

On considère $\mathcal{P} = \{\mathrm{P}_\theta\,;\theta \in \Theta\}$, une famille de lois sur un espace discret E, indicée par un paramètre $\theta \in \Theta$. Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes et de même loi P_{θ_0}, θ_0 étant inconnu. Pour $x_1 \in E$ on pose $p(\theta_0\,;x_1) = \mathbb{P}(X_1 = x_1)$. La vraisemblance associée à X_1 est donc $\theta \to p(\theta, x_1)$. Par indépendance, la vraisemblance associée à l'échantillon X_1^N est, pour $x_1^n \in E$, $p_n(\theta\,;x_1^n) = \prod_{i=1}^n p(\theta\,;x_i)$. La log-vraisemblance est

$$L_n(\theta\,;x_1^n) = \sum_{i=1}^n \log p(\theta\,;x_i). \tag{5.4}$$

On suppose que l'EMV de θ, $\hat{\theta}_n$ est bien défini : la variable aléatoire $\hat{\theta}_n$ est l'unique valeur de Θ en laquelle $L_n(\theta\,;X_1^n)$ et donc $\frac{1}{n}L_n(\theta\,;X_1^n)$ sont maximaux :

$$\hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmax}}\ \frac{1}{n}\sum_{i=1}^n \log p(\theta\,;X_i). \tag{5.5}$$

On considère le lemme suivant dont la démonstration est reportée à la fin de ce paragraphe.

Lemme 5.2.8. *Soit un espace E discret, $\mathcal{P}_E$ l'ensemble des probabilités sur E, et $p = (p(x), x \in E) \in \mathcal{P}_E$. On considère la fonction $\mathcal{H}_p$ à valeurs dans $[-\infty, 0]$ définie sur $\mathcal{P}_E$ par*

$$\mathcal{H}_p : \ p' \to \mathcal{H}_p(p') = \sum_{x \in E} p(x) \log p'(x),$$

avec la convention $0 \log 0 = 0$. *On suppose que* $\mathcal{H}_p(p) > -\infty$. *Alors la fonction* $\mathcal{H}_p$ *atteint son unique maximum pour* $p' = p$.

Remarquons que la quantité $\mathcal{H}_p(p)$ est au signe près l'**entropie** de p.

Par simplicité d'écriture, on note pour $\theta, \theta_0 \in \Theta$, $\mathcal{H}_{\theta_0}(\theta) = \mathcal{H}_{p_0}(p)$, où $p = p(\theta\,;\cdot)$ et $p_0 = p(\theta_0\,;\cdot)$. Comme $p(\theta\,; X_1) \leq 1$, on a

$$\mathcal{H}_{\theta_0}(\theta) = \sum_{x \in E} p(\theta_0\,; x) \log p(\theta, x) = \mathbb{E}[\log p(\theta\,; X_1)] \in [-\infty, 0],$$

et par la loi forte des grands nombres, cf. corollaire A.3.14,

$$\frac{1}{n} L_n(\theta\,; X_1^n) = \frac{1}{n} \sum_{i=1}^{n} \log p(\theta\,; X_i) \xrightarrow[n\to\infty]{\textbf{p.s.}} \mathcal{H}_{\theta_0}(\theta). \tag{5.6}$$

Ceci suggère que $\hat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} \frac{1}{n} L_n(\theta\,; X_1^n)$ converge presque sûrement vers $\operatorname*{argmax}_{\theta \in \Theta} \mathcal{H}_{\theta_0}(\theta)$ (i.e. vers θ_0 d'après le lemme 5.2.8) quand n tend vers l'infini ; et donc que l'EMV est convergent. Plus précisément, on a le théorème suivant.

Théorème 5.2.9. *On suppose les conditions suivantes :*

1. Θ *est compact.*
2. *Le modèle est identifiable.*
3. *La vraisemblance définie sur* Θ, $\theta \to p(\theta\,; x)$, *est continue pour tout* $x \in E$.
4. *P.s. pour* n *assez grand, (5.5) définit uniquement l'EMV* $\hat{\theta}_n$.
5. *La quantité* $\mathcal{H}_\theta(\theta)$ *est finie pour tout* $\theta \in \Theta$.

Alors l'EMV de θ, *défini par (5.5), est un estimateur convergent.*

Démonstration. On pose pour tout $x \in E$, $f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i = x\}}$, et on remarque que

$$\frac{1}{n} L_n(\theta\,; X_1^n) = \sum_{x \in E} \log p(\theta\,; x) f_n(x).$$

La loi forte des grands nombres assure que p.s. pour tout $x \in E$,

$$f_n(x) \xrightarrow[n\to\infty]{\textbf{p.s.}} p(\theta_0\,; x).$$

Comme Θ est compact, et que p.s. l'EMV est bien défini pour n assez grand, la suite des EMV admet au moins un point d'accumulation $\theta_* \in \Theta$. Et il existe p.s. une fonction strictement croissante (aléatoire), σ, de $\mathbb{N}^*$ dans $\mathbb{N}^*$, telle que

la suite $(\hat{\theta}_{\sigma(n)}, n \geq 1)$ converge vers θ_*. Par continuité de la vraisemblance, on a pour tout $x \in E$,

$$\log p(\hat{\theta}_{\sigma(n)}\,;x) \xrightarrow[n\to\infty]{\textbf{p.s.}} \log p(\theta_*\,;x).$$

Comme les fonctions $f_{\sigma(n)}$ et $-\log p(\hat{\theta}_{\sigma(n)}\,;\cdot)$ sont positives, on déduit du lemme de Fatou que p.s.

$$\begin{aligned}\liminf_{n\to\infty} -\frac{1}{\sigma(n)}\, L_{\sigma(n)}(\hat{\theta}_{\sigma(n)}\,;X_1^{\sigma(n)}) &= \liminf_{n\to\infty} \sum_{x\in E} -\log p(\hat{\theta}_{\sigma(n)}\,;x) f_{\sigma(n)}(x)\\ &\geq \sum_{x\in E} \liminf_{n\to\infty} -\log p(\hat{\theta}_{\sigma(n)}\,;x) f_{\sigma(n)}(x)\\ &= -\mathcal{H}_{\theta_0}(\theta_*).\end{aligned}$$

Comme $\hat{\theta}_n$ est l'EMV, on a $L_n(\theta_0\,;X_1^n) \leq L_n(\hat{\theta}_n\,;X_1^n)$, et grâce à (5.6), p.s.

$$\begin{aligned}\liminf_{n\to\infty} -\frac{1}{\sigma(n)}\, L_{\sigma(n)}(\hat{\theta}_{\sigma(n)}\,;X_1^{\sigma(n)}) &\leq \liminf_{n\to\infty} -\frac{1}{\sigma(n)}\, L_{\sigma(n)}(\theta_0\,;X_1^{\sigma(n)})\\ &= -\mathcal{H}_{\theta_0}(\theta_0).\end{aligned}$$

On déduit de ces inégalités, que p.s. $\mathcal{H}_{\theta_0}(\theta_*) \geq \mathcal{H}_{\theta_0}(\theta_0)$. Le modèle étant identifiable et $\mathcal{H}_{\theta_0}(\theta_0)$ fini, on déduit du lemme 5.2.8 que p.s. $\theta_* = \theta_0$. Ceci implique que la suite des EMV admet p.s. un seul point d'accumulation θ_0. Elle est donc p.s. convergente et sa limite est θ_0. Ceci démontre donc le théorème. □

Remarque 5.2.10. Si la vraisemblance $p_n(\cdot, X_1^n)$ atteint son maximum en plusieurs points l'EMV n'est pas défini (condition 4 du théorème 5.2.9 non vérifiée). Les arguments de la démonstration ci-dessus assurent en fait que toute suite $(\hat{\theta}_n, n \geq 1)$, telle que la vraisemblance $p_n(\cdot, X_1^n)$ atteint son maximum en $\hat{\theta}_n$, admet p.s. un seul point d'accumulation qui est θ_0. La suite converge donc p.s. vers la vraie valeur θ_0. Ainsi, on peut étendre la définition de l'EMV à tout point de Θ tel que la vraisemblance soit maximale en ce point, et conserver les propriétés de convergence. ♢

Démonstration du lemme 5.2.8. Remarquons que pour $r \geq 0$, on a $\log r \leq r - 1$ avec égalité si et seulement si $r = 1$. Pour $y > 0$, $z \geq 0$, on a $y \log z - y \log y = y \log(z/y) \leq y\Big(\frac{z}{y} - 1\Big) = z - y$, avec égalité si et seulement si $z = y$. Avec la convention $0 \log 0 = 0$, on obtient que pour $y \geq 0$, $z \geq 0$

$$y \log z - y \log y \leq z - y, \tag{5.7}$$

avec égalité si et seulement si $y = z$. On a

$$\begin{aligned}\mathcal{H}_p(p') - \mathcal{H}_p(p) &= \sum_{x\in E} p(x)\log p'(x) - \sum_{x\in E} p(x)\log p(x)\\ &= \sum_{x\in E} p(x)[\log p'(x) - \log p(x)]\\ &\le \sum_{x\in E} p'(x) - p(x)\\ &= 0,\end{aligned}$$

où l'on a utilisé le fait que $\sum_{x\in E} p(x)\,|\log p(x)| < \infty$ pour la deuxième égalité et (5.7) pour l'inégalité. Ainsi on a $\mathcal{H}_p(p') \le \mathcal{H}_p(p)$. Enfin comme (5.7) n'est une égalité que si $y = z$, on en déduit que $\mathcal{H}_p(p') = \mathcal{H}_p(p)$ si et seulement si $p' = p$. □

5.3 Présentation générale de l'algorithme EM

On écrit $\mathbb{P}_\theta$ et $\mathbb{E}_\theta$ pour les probabilités et espérances calculées quand le vrai paramètre de la chaîne de Markov $((S_n, Y_n), n \ge 1)$ 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 $y = y_1^N \in \mathcal{X}$. La vraisemblance du modèle incomplet est définie par $p_N(\theta\,; y) = \mathbb{P}_\theta(Y = y)$. On a, en utilisant (5.1) et (5.2),

$$\begin{aligned}p_N(\theta\,; y) &= \sum_{s\in\mathcal{I}^N} \mathbb{P}_\theta(Y = y|S = s)\mathbb{P}_\theta(S = s)\\ &= \sum_{s\in\mathcal{I}^N} \left(\prod_{n=1}^N b(s_n, y_n)\right)\pi_0(s_1)\prod_{n=2}^N a(s_{n-1}, s_n). \qquad (5.8)\end{aligned}$$

La log-vraisemblance du modèle incomplet est $L_N(\theta\,; y) = \log p_N(\theta\,; y)$. Pour déterminer l'EMV de θ, il faut maximiser $p_N(\cdot\,; y)$ en $\theta = (a, b, \pi)$. 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). On choisit Θ', défini à la fin du paragraphe 5.1, pour l'ensemble des paramètres possibles de sorte que le modèle paramétrique soit identifiable.

L'existence et la convergence de l'EMV sont présentées au paragraphe 5.5, voir le théorème 5.5.2. Pour calculer numériquement l'EMV, remarquons qu'il faut maximiser $p_N(\theta\,; y)$, un polynôme de degré $2N$ à Card $(\mathcal{I}^2 \times (\mathcal{I} \times \mathcal{X}) \times \mathcal{I})$ variables sous 2Card $(\mathcal{I}) + 1$ 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 (cf. [16] pour la détection de zones homogènes de l'ADN).

Pour une autre approche, on considère la vraisemblance du modèle complet définie par $p_N^{complet}(\theta\,;s,y) = \mathbb{P}_\theta(S=s, Y=y)$. On a

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

Nous verrons au paragraphe 5.4.2 que le calcul de l'EMV pour le modèle complet est élémentaire.

La loi conditionnelle de S sachant Y est donnée par

$$\pi_N(\theta\,;s|y) = \mathbb{P}_\theta(S=s|Y=y) = \frac{\mathbb{P}_\theta(S=s,Y=y)}{\mathbb{P}_\theta(Y=y)} = \frac{p_N^{complet}(\theta\,;s,y)}{p_N(\theta\,;y)}. \tag{5.9}$$

On écrit artificiellement la log-vraisemblance du modèle incomplet, calculée pour θ, avec la log-vraisemblance du modèle complet calculée pour θ' distinct de θ a priori. Comme $\sum_{s\in\mathcal{I}^N}\pi_N(\theta'\,;s|y)=1$, on a

$$\begin{aligned}
&L_N(\theta\,;y)\\
&\quad= \log p_N(\theta\,;y)\\
&\quad= \sum_{s\in\mathcal{I}^N}\pi_N(\theta'\,;s|y)\log p_N(\theta\,;y)\\
&\quad= \sum_{s\in\mathcal{I}^N}\pi_N(\theta'\,;s|y)\log p_N^{complet}(\theta\,;s,y) - \sum_{s\in\mathcal{I}^N}\pi_N(\theta'\,;s|y)\log \pi_N(\theta\,;s|y),
\end{aligned}$$

où l'on a utilisé la définition (5.9) de $\pi_N(\theta\,;s|y)$ pour la dernière égalité. On pose, pour $y\in\mathcal{X}^N$,

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

et

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

On a donc

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

On établit le lemme suivant.

Lemme 5.3.1. *Soit θ' fixé. Soit θ^* le (ou un) paramètre qui maximise la fonction $\theta\to Q(\theta,\theta')$. Alors $L_N(\theta^*\,;y)\geq L_N(\theta'\,;y)$.*

Démonstration. On déduit du lemme 5.2.8 que $\mathcal{H}_{\theta'}(\theta^*)\leq\mathcal{H}_{\theta'}(\theta')$. Comme $Q(\theta^*,\theta')\geq Q(\theta',\theta')$, cela implique que $L_N(\theta^*\,;y)\geq L_N(\theta'\,;y)$. □

L'algorithme EM (Espérance Maximisation) consiste à construire par récurrence une suite de paramètres $(\theta^{(r)}, r \in \mathbb{N})$ de la manière suivante. $\theta^{(0)} \in \Theta'$ 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 $\theta \to Q(\theta, \theta^{(r)})$ 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.

Soit $\delta > 0$. On considère l'hypothèse suivante notée (H_δ) : *Pour tous* $i, j \in \mathcal{I}$ *et* $x \in \mathcal{X}$, *on a* $a(i,j) > \delta$, $b(i,x) > \delta$ *et* $\pi_0(i) > \delta$.

Soit Θ_δ l'ensemble des paramètres $\theta \in \Theta'$ qui vérifient la condition (H_δ). On suppose que l'on peut choisir δ assez petit pour que le vrai paramètre θ_0 soit dans Θ_δ. On admet le théorème suivant qui découle des résultats de [14].

Théorème 5.3.2. *La suite construite par l'algorithme EM dans* Θ_δ, $(\theta^{(r)}, r \in \mathbb{N})$, *converge vers l'EMV de* θ, $\hat{\theta}_N$, *dès que* $\theta^{(0)}$ *est assez proche de* $\hat{\theta}_N$.

Comme le souligne le théorème, la difficulté de l'algorithme EM réside dans le choix du point d'initialisation $\theta^{(0)}$. On peut démontrer, sous des hypothèses assez générales, que la suite générée par l'algorithme EM converge vers un point en lequel la dérivée de la log-vraisemblance s'annule. Il peut très bien s'agir d'un point selle ou d'un maximum local et non du maximum global $\hat{\theta}_N$. De plus l'algorithme EM converge mal si le point initial se trouve dans une région où la log-vraisemblance ne varie pas beaucoup. Il existe des procédures pour introduire de l'aléatoire dans les premières itérations de l'algorithme (variante stochastique (SEM) de l'algorithme EM) afin de s'affranchir de ces problèmes. On peut aussi utiliser l'algorithme EM avec plusieurs points d'initialisation. On pourra consulter [10] pour des résultats précis concernant ces questions.

Exemple 5.3.3. Les calculs explicites du paragraphe suivant permettent d'implémenter facilement l'algorithme EM pour l'estimation des paramètres et des variables cachées pour les chaînes de Markov cachées. On peut ainsi vérifier la pertinence de cet algorithme sur des simulations. On choisit un exemple avec des paramètres proches de ceux estimés dans l'exemple de la séquence d'ADN du bactériophage lambda (voir le paragraphe 5.7 pour les résultats numériques et plus particulièrement (5.22) pour la valeur des paramètres estimés). On considère une simulation $(s_1^{N_0}, y_1^{N_0})$ de la chaîne de Markov cachée $(S_1^{N_0}, Y_1^{N_0})$, avec $N_0 = 48\ 502$, $\mathcal{I} = \{+1, -1\}$, $\mathcal{X} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, et les paramètres suivants :

$$a = \begin{pmatrix} 0.9999 & 0.0001 \\ 0.0002 & 0.9998 \end{pmatrix}, \quad b = \begin{pmatrix} 0.246 & 0.248 & 0.298 & 0.208 \\ 0.270 & 0.208 & 0.198 & 0.324 \end{pmatrix} \quad \text{et} \quad \pi_0 = (1, 0).$$

Après 1 000 itérations de l'algorithme EM, initialisé avec

$$a^{(0)} = \begin{pmatrix} 0.28 & 0.72 \\ 0.19 & 0.81 \end{pmatrix}, \quad b^{(0)} = \begin{pmatrix} 0.21 & 0.36 & 0.37 & 0.06 \\ 0.27 & 0.27 & 0.26 & 0.20 \end{pmatrix} \quad \text{et} \quad \pi_0^{(0)} = (0.5, 0.5), \tag{5.10}$$

on obtient l'estimation suivante des paramètres a et b :

$$a \simeq \begin{pmatrix} 0.99988 & 0.00012 \\ 0.00015 & 0.99985 \end{pmatrix}, \quad b \simeq \begin{pmatrix} 0.2456 & 0.2505 & 0.2946 & 0.2096 \\ 0.2723 & 0.2081 & 0.1952 & 0.3244 \end{pmatrix}.$$

Cette estimation dépend assez peu du point de départ pourvu que les termes diagonaux de $a^{(0)}$ ne soient pas simultanément trop petits. Dans la Fig. 5.2, on présente les valeurs de la simulation $s_1^{N_0}$ et les valeurs restaurées ($\mathbb{P}(S_n = +1 | Y_1^{N_0} = y_1^{N_0}), n \in \{1, \ldots, N_0\}$). La figure 5.3 (resp. 5.4) présente l'évolution au cours des itérations de l'algorithme EM, des coefficients diagonaux de a (resp. des coefficients de b). On constate que les estimations sont constantes après un petit nombre (devant N_0) d'itérations, et que les valeurs numériques estimées sont proches des vraies valeurs des paramètres. ◊

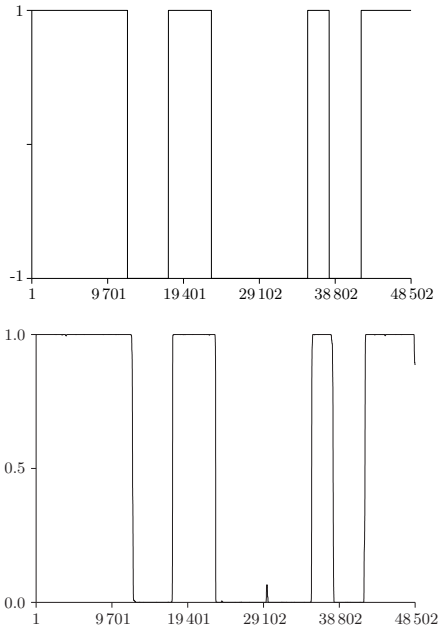

Fig. 5.2. Valeur des états cachés simulés ($n \to S_n$) en haut et valeur des probabilités estimées de l'état caché $+1$ ($n \to \mathbb{P}(S_n = +1 | Y_1^{N_0} = y_1^{N_0})$) en bas

5.4 Mise en œuvre de l'algorithme EM

5.4.1 L'étape espérance : étape E

On suppose construit $\theta^{(r)}$. On désire construire $\theta^{(r+1)}$. Pour cela, il faut calculer $Q(\theta\,;\theta^{(r)})$. Soit $\theta, \theta' \in \Theta_\delta$. On note $\theta = (a, b, \pi_0)$ et $\theta' = (a', b', \pi_0')$. On rappelle que $\mathbb{P}_{\theta'}$ et $\mathbb{E}_{\theta'}$ désignent les probabilités et espérances quand le vrai

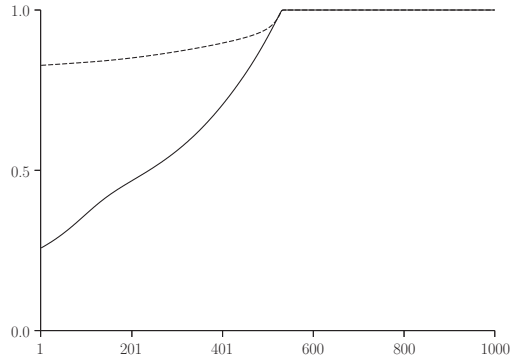

Fig. 5.3. Évolution de l'estimation des termes diagonaux de la matrice de transition des états cachés en fonction des itérations, obtenue pour 1 000 itérations

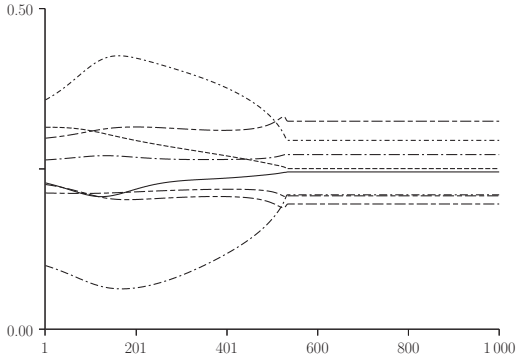

Fig. 5.4. Évolution de l'estimation des termes de la matrice b en fonction des itérations, obtenue pour 1 000 itérations

paramètre de la chaîne de Markov $((S_n, Y_n), n \geq 1)$ est θ'. Comme $\theta' \in \Theta_\delta$, remarquons que toutes les probabilités de transition sont strictement positives. En particulier, pour tous $n \geq 1$, $s_1^n \in \mathcal{I}^n$, $y_1^n \in \mathcal{X}^n$, $\mathbb{P}_{\theta'}(S_1^n = s_1^n, Y_1^n = y_1^n) > 0$.

Avec les notations du paragraphe précédent, on calcule $Q(\theta, \theta')$ pour $y \in \mathcal{X}^N$:

$$\begin{aligned} Q(\theta, \theta') &= \sum_{s \in \mathcal{I}^N} \pi_N(\theta' \,; s|y) \log p_N^{complet}(\theta \,; s, y) \\ &= \mathbb{E}_{\theta'}[\log p_N^{complet}(\theta \,; S, y)|Y = y] \\ &= \mathbb{E}_{\theta'}\left[\log \Big(\pi_0(S_1) b(S_1, y_1) \prod_{n=2}^{N} a(S_{n-1}, S_n) b(S_n, y_n) \Big) \Big| Y = y \right] \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\theta'}[\log \pi_0(S_1)|Y=y] + \mathbb{E}_{\theta'}[\log b(S_1,y_1)|Y=y] \\
&\qquad + \sum_{n=2}^{N} \mathbb{E}_{\theta'}[\log a(S_{n-1},S_n)|Y=y] + \sum_{n=2}^{N} \mathbb{E}_{\theta'}[\log b(S_n,y_n)|Y=y] \\
&= \sum_{i\in\mathcal{I}} \mathbb{P}_{\theta'}(S_1=i|Y=y)\log \pi_0(i) \\
&\qquad\qquad + \sum_{n=1}^{N}\sum_{i\in\mathcal{I}} \mathbb{P}_{\theta'}(S_n=i|Y=y)\log b(i,y_n) \\
&\qquad\qquad + \sum_{n=2}^{N}\sum_{i,j\in\mathcal{I}} \mathbb{P}_{\theta'}(S_{n-1}=i,S_n=j|Y=y)\log a(i,j).
\end{aligned}
$$

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 $2\le n\le N$ et $\mathbb{P}_{\theta'}(S_n=i|Y=y)$ pour $1\le n\le N$. 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 5.4.1 (Prévision). *On a, pour $n\ge 2$, $y_1^{n-1}\in\mathcal{X}^{n-1}$,*

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

Démonstration. On décompose suivant les valeurs de S_1^{n-1} :

$$
\begin{aligned}
&\mathbb{P}_{\theta'}(S_n=i|Y_1^{n-1}=y_1^{n-1}) \\
&\qquad = \sum_{s_1^{n-1}\in\mathcal{I}^{n-1}} \mathbb{P}_{\theta'}(S_n=i,S_1^{n-1}=s_1^{n-1}|Y_1^{n-1}=y_1^{n-1}) \\
&\qquad = \sum_{s_1^{n-1}\in\mathcal{I}^{n-1}} \mathbb{P}_{\theta'}(S_n=i|S_1^{n-1}=s_1^{n-1},Y_1^{n-1}=y_1^{n-1}) \\
&\qquad\qquad\qquad \mathbb{P}_{\theta'}(S_1^{n-1}=s_1^{n-1}|Y_1^{n-1}=y_1^{n-1}) \\
&\qquad = \sum_{s_1^{n-1}\in\mathcal{I}^{n-1}} a'(s_{n-1},i)\mathbb{P}_{\theta'}(S_1^{n-1}=s_1^{n-1}|Y_1^{n-1}=y_1^{n-1}) \\
&\qquad = \sum_{s_{n-1}\in\mathcal{I}} a'(s_{n-1},i)\mathbb{P}_{\theta'}(S_{n-1}=s_{n-1}|Y_1^{n-1}=y_1^{n-1}),
\end{aligned}
$$

où l'on a utilisé le lemme 5.1.1 pour la troisième égalité. □

Lemme 5.4.2 (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 $\theta' = (a', b', \pi'_0)$.

Avant de détailler la démonstration du lemme 5.4.2, nous démontrons un résultat technique intermédiaire.

Lemme 5.4.3. *Soit $\theta \in \Theta$. Soit $m \geq 0$, $n \geq 2$, $y_n^{n+m+1} \in \mathcal{X}^{m+1}$, $s_k \in \mathcal{I}$ et $J_n = \{(S_1^{n-1}, Y_1^{n-1}) \in B\}$, où $B \subset (\mathcal{I} \times \mathcal{X})^{n-1}$. Si $\mathbb{P}_\theta(S_n = s_n, J_n) > 0$, alors on a*

$$\mathbb{P}_\theta(Y_n^{n+m} = y_n^{n+m} | S_n = s_n, J_n) = \mathbb{P}_\theta(Y_n^{n+m} = y_n^{n+m} | S_n = s_n),$$

et si $\mathbb{P}_\theta(Y_n^{n+m} = y_n^{n+m}, S_n = s_n, J_n) > 0$

$$\begin{aligned}\mathbb{P}_\theta(Y_{n+m+1} = y_{n+m+1} | Y_n^{n+m} = y_n^{n+m}, S_n = s_n, J_n) \\ = \mathbb{P}_\theta(Y_{n+m+1} = y_{n+m+1} | Y_n^{n+m} = y_n^{n+m}, S_n = s_n).\end{aligned}$$

Démonstration. On calcule dans un premier temps $\mathbb{P}_\theta(Y_n = y_n | S_n = s_n, J_n)$. En décomposant suivant les valeurs possibles de (S_1^{n-1}, Y_1^{n-1}), il vient

$$\begin{aligned}\mathbb{P}_\theta(Y_n = y_n, S_n = s_n, J_n) &= \sum_{(s_1^{n-1}, y_1^{n-1}) \in B} \mathbb{P}_\theta(S_1^n = s_1^n, Y_1^n = y_1^n) \\ &= \sum_{(s_1^{n-1}, y_1^{n-1}) \in B} \pi_0(s_1) b(s_1, y_1) \prod_{k=2}^{n} a(s_{k-1}, s_k) b(s_k, y_k),\end{aligned}$$

où l'on a utilisé pour la dernière égalité le fait que $((S_k, Y_k), k \geq 1)$ est une chaîne de Markov, cf. le lemme 5.1.1, de loi initiale $\mathbb{P}_\theta(S_1 = s_1, Y_1 = y_1) = \mathbb{P}_\theta(Y_1 = y_1 | S_1 = s_1)\mathbb{P}_\theta(S_1 = s_1) = \pi_0(s_1)b(s_1, y_1)$. En sommant sur $y_n \in \mathcal{X}$, on en déduit que

$$\begin{aligned}&\mathbb{P}_\theta(S_n = s_n, J_n) \\ &\quad = \sum_{(s_1^{n-1}, y_1^{n-1}) \in B} \pi_0(s_1) b(s_1, y_1) \left[\prod_{k=2}^{n-1} a(s_{k-1}, s_k) b(s_k, y_k)\right] a(s_{n-1}, s_n),\end{aligned}$$

et donc si $\mathbb{P}_\theta(S_n = s_n, J_n) > 0$,

$$\mathbb{P}_\theta(Y_n = y_n | S_n = s_n, J_n) = b(s_n, y_n) = \mathbb{P}_\theta(Y_n = y_n | S_n = s_n).$$

On suppose $m \geq 1$. On a alors

$$\begin{aligned}
&\mathbb{P}_\theta(Y_n^{n+m} = y_n^{n+m} | S_n = s_n, J_n) \\
&\quad = \mathbb{P}_\theta(Y_{n+1}^{n+m} = y_{n+1}^{n+m} | S_n = s_n, Y_n = y_n, J_n) \mathbb{P}_\theta(Y_n = y_n | S_n = s_n, J_n) \\
&\quad = \mathbb{P}_\theta(Y_{n+1}^{n+m} = y_{n+1}^{n+m} | S_n = s_n, Y_n = y_n) \mathbb{P}_\theta(Y_n = y_n | S_n = s_n, J_n) \\
&\quad = \mathbb{P}_\theta(Y_{n+1}^{n+m} = y_{n+1}^{n+m} | S_n = s_n, Y_n = y_n) \mathbb{P}_\theta(Y_n = y_n | S_n = s_n) \\
&\quad = \mathbb{P}_\theta(Y_n^{n+m} = y_n^{n+m} | S_n = s_n),
\end{aligned}$$

où on utilise la formule de décomposition des probabilités conditionnelles pour les première et dernière égalités, et la proposition 1.1.7 pour la chaîne de Markov $((S_k, Y_k), k \geq 1)$ avec $I_n = \{(S_{n+1}^{n+m}, Y_{n+1}^{n+m}) \in \mathcal{I}^m \times \{y_{n+1}^{n+m}\}\}$ pour la deuxième égalité. Ceci démontre la première égalité du lemme.

Remarquons, que grâce à la définition des probabilités conditionnelles, on a si $\mathbb{P}_\theta(Y_n^{n+m} = y_n^{n+m}, S_n = s_n, J_n) > 0$,

$$\begin{aligned}
&\mathbb{P}_\theta(Y_{n+m+1} = y_{n+m+1} | Y_n^{n+m} = y_n^{n+m}, S_n = s_n, J_n) \\
&\qquad = \frac{\mathbb{P}_\theta(Y_n^{n+m+1} = y_n^{n+m+1} | S_n = s_n, J_n)}{\mathbb{P}_\theta(Y_n^{n+m} = y_n^{n+m} | S_n = s_n, J_n)} \\
&\qquad = \frac{\mathbb{P}_\theta(Y_n^{n+m+1} = y_n^{n+m+1} | S_n = s_n)}{\mathbb{P}_\theta(Y_n^{n+m} = y_n^{n+m} | S_n = s_n)} \\
&\qquad = \mathbb{P}_\theta(Y_{n+m+1} = y_{n+m+1} | Y_n^{n+m} = y_n^{n+m}, S_n = s_n),
\end{aligned}$$

où l'on a utilisé la première égalité du lemme deux fois pour obtenir l'avant-dernière égalité. Ceci termine la démonstration du lemme. □

Démonstration du lemme 5.4.2. On a

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

où l'on a utilisé la première égalité du lemme 5.4.3 (avec $m = 0$) pour la quatrième égalité. □

Lemme 5.4.4 (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 $1 \leq n \leq N-1$, $y_1^N \in \mathcal{X}^N$,

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

Démonstration. Soit $2 \leq n \leq N$. En utilisant la première égalité du lemme 5.4.3 avec $n + m = N$, il vient

$$\begin{aligned}
&\mathbb{P}_{\theta'}(S_{n-1} = i, S_n = j, Y_1^N = y_1^N) \\
&\quad = \mathbb{P}_{\theta'}(Y_n^N = y_n^N | S_{n-1} = i, S_n = j, Y_1^{n-1} = y_1^{n-1}) \\
&\qquad\qquad \mathbb{P}_{\theta'}(S_{n-1} = i, S_n = j, Y_1^{n-1} = y_1^{n-1}) \\
&\quad = \mathbb{P}_{\theta'}(Y_n^N = y_n^N | S_n = j) \mathbb{P}_{\theta'}(S_{n-1} = i, S_n = j, Y_1^{n-1} = y_1^{n-1}).
\end{aligned}$$

On a aussi

$$\mathbb{P}_{\theta'}(S_n = j, Y_1^N = y_1^N) = \mathbb{P}_{\theta'}(Y_n^N = y_n^N | S_n = j) \mathbb{P}_{\theta'}(S_n = j, Y_1^{n-1} = y_1^{n-1}).$$

En particulier, en faisant le rapport de ces deux égalités, on obtient

$$\begin{aligned}
&\mathbb{P}_{\theta'}(S_{n-1} = i | S_n = j, Y_1^N = y_1^N) \\
&\quad = \frac{\mathbb{P}_{\theta'}(S_n = j, S_{n-1} = i, Y_1^{n-1} = y_1^{n-1})}{\mathbb{P}_{\theta'}(S_n = j, Y_1^{n-1} = y_1^{n-1})} \\
&\quad = \frac{\mathbb{P}_{\theta'}(S_n = j | S_{n-1} = i, Y_1^{n-1} = y_1^{n-1}) \mathbb{P}_{\theta'}(S_{n-1} = i | Y_1^{n-1} = y_1^{n-1})}{\mathbb{P}_{\theta'}(S_n = j | Y_1^{n-1} = y_1^{n-1})}.
\end{aligned}$$

En utilisant la proposition 1.1.7 pour la chaîne de Markov $((S_n, Y_n), n \geq 1)$ et le lemme 5.1.1, il vient

$$\begin{aligned}
&\mathbb{P}_{\theta'}(S_n = j | S_{n-1} = i, Y_1^{n-1} = y_1^{n-1}) \\
&\qquad = \sum_{x \in \mathcal{X}} \mathbb{P}_{\theta'}(S_n = j, Y_n = x | S_{n-1} = i, Y_1^{n-1} = y_1^{n-1}) \\
&\qquad = \sum_{x \in \mathcal{X}} \mathbb{P}_{\theta'}(S_n = j, Y_n = x | S_{n-1} = i, Y_{n-1} = y_{n-1}) \\
&\qquad = \sum_{x \in \mathcal{X}} a'(i,j) b'(j,x) \\
&\qquad = a'(i,j).
\end{aligned}$$

On en déduit

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

On calcule maintenant $\mathbb{P}_{\theta'}(S_{n-1} = i, S_n = j | Y_1^N = y_1^N)$. On déduit de l'égalité précédente que

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

Il reste à calculer $\mathbb{P}_{\theta'}(S_n = j | Y_1^N = y_1^N)$. On déduit de l'égalité précédente, en remplaçant n par $n+1$, que, pour $1 \leq n \leq N-1$,

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

□

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 $\theta' = (a', b', \pi'_0)$.

5.4.2 L'étape maximisation : étape M

On recherche $\theta = (a, b, \pi_0)$ qui maximise $Q(\theta, \theta')$ à y et θ' fixés. On maximise la quantité $Q(\theta, \theta')$ définie par

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

sous les contraintes que $(\pi_0(j), j \in \mathcal{I})$ est une probabilité, de même que $(b(i, x), x \in \mathcal{X})$ et $(a(i, j), j \in \mathcal{I})$ pour tout $i \in \mathcal{I}$. On remarque que l'on peut maximiser séparément les sommes correspondant à chacune des contraintes précédentes. Considérons par exemple la somme intervenant dans $Q(\theta, \theta')$ qui fait intervenir les termes $b(i, x)$ pour $x \in \mathcal{X}$ et i fixé :

$$\sum_{n=1}^{N} \mathbb{P}_{\theta'}(S_n = i|Y = y) \log b(i, y_n).$$

On peut récrire cette somme de la manière suivante $\sum_{x\in\mathcal{X}} q(x) \log b(i, x)$ où

$$q(x) = \sum_{n=1}^{N} \mathbf{1}_{\{y_n=x\}} \mathbb{P}_{\theta'}(S_n = i|Y = y).$$

Remarquons que la probabilité $(b(i, x), x \in \mathcal{X})$ maximise $\sum_{x\in\mathcal{X}} q(x) \log b(i, x)$ si et seulement si elle maximise la somme $\sum_{x\in\mathcal{X}} p(x) \log b(i, x)$, où $p(x) = q(x)/\sum_{z\in\mathcal{X}} q(z)$. La suite $p = (p(x), x \in \mathcal{X})$ définit une probabilité sur $\mathcal{X}$. On déduit du lemme 5.2.8 que la somme est maximale pour $b(i, \cdot) = p$. On peut utiliser des arguments similaires pour calculer a et π_0. En définitive, on en déduit que $Q(\theta, \theta')$ est maximal pour $\theta = (a, b, \pi_0)$ défini pour $i, j \in \mathcal{I}, x \in \mathcal{X}$ par

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

Ceci termine l'étape M.

5.5 Convergence de l'EMV pour les chaînes de Markov cachées

Le théorème 5.3.2 assure, sous certaines hypothèses, la convergence de la suite construite avec l'algorithme EM vers l'EMV de θ. Dans ce paragraphe nous indiquons les arguments qui permettent de montrer que l'EMV est un estimateur convergent pour le modèle de chaîne de Markov cachée présenté au paragraphe 5.1. Ces arguments sont similaires à ceux employés dans le paragraphe 5.2, pour la convergence de l'EMV dans le modèle paramétrique où les variables $(X_n, n \geq 1)$ sont indépendantes et de même loi.

On reprend les notations du paragraphe 5.3. La log-vraisemblance des données observées pour un échantillon de taille N est, pour $y \in \mathcal{X}^N$,

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

On introduit les probabilités conditionnelles suivantes pour $k \geq 2$:

$$p(\theta\,;y_k|y_1^{k-1}) = \mathbb{P}_\theta(Y_k = y_k|Y_1^{k-1} = y_1^{k-1}) = \frac{p_k(\theta\,;y_1^k)}{p_{k-1}(\theta\,;y_1^{k-1})}.$$

Comme $p_N(\theta\,;y) = \prod_{k=1}^n p(\theta\,;y_k|y_1^{k-1})$, avec la convention $p(\theta\,;y_k|y_1^{k-1}) = p(\theta\,;y_1)$ si $k = 1$, on peut alors récrire la log-vraisemblance de la manière suivante :

$$L_N(\theta\,;y) = \sum_{k=1}^N \log p(\theta\,;y_k|y_1^{k-1}).$$

(Comparer cette expression avec celle de (5.4), concernant un modèle paramétrique de variables aléatoires indépendantes et de même loi.) Remarquons que par construction, voir (5.8), la log-vraisemblance est une fonction continue sur Θ_δ, l'espace des paramètres décrit page 134.

Avant de démontrer la proposition suivante (cf. théorème 3.1 dans [1]), nous indiquons comment elle implique la convergence de l'EMV.

Proposition 5.5.1. *Soit $\theta_0 \in \Theta_\delta$ le vrai paramètre. Alors pour tout $\theta \in \Theta_\delta$, on a*

$$\frac{1}{N}\, L_N(\theta\,;Y_1^N) \xrightarrow[N\to+\infty]{\textbf{\textit{p.s.}}} \mathcal{H}_{\theta_0}(\theta),$$

où la fonction $\mathcal{H}_{\theta_0}$ est continue. De plus si $\theta \neq \theta_0$, alors on a $\mathcal{H}_{\theta_0}(\theta) < \mathcal{H}_{\theta_0}(\theta_0)$.

Quand l'EMV existe, il est défini par

$$\hat{\theta}_N = \operatorname*{argmax}_{\theta\in\Theta_\delta}\ \frac{1}{N}\, L_N(\theta\,;Y_1^N). \tag{5.11}$$

D'après la proposition précédente,

$$\theta_0 = \underset{\theta\in\Theta_\delta}{\operatorname{argmax}}\ \mathcal{H}_{\theta_0}(\theta),$$

et $\mathcal{H}_{\theta_0}(\theta)$ est la limite p.s. de $\frac{1}{N} L_N(\theta\,; Y_1^N)$. Vu le paragraphe 5.2.2, et plus particulièrement le théorème 5.2.9, il est naturel de penser que l'EMV est un estimateur convergent.

Le résultat qui suit est démontré par exemple dans [1] (théorème 3.4) :

Théorème 5.5.2. *Soit $\theta_0 \in \Theta_\delta$ le vrai paramètre. Pour N assez grand, l'EMV, $\hat{\theta}_N$, de $\theta \in \Theta_\delta$ est bien défini. De plus l'EMV est convergent :*

$$\hat{\theta}_N \xrightarrow[N\to+\infty]{\textbf{\textit{p.s.}}} \theta_0.$$

Sous des hypothèses supplémentaires sur le modèle, difficiles à vérifier en pratique, on peut montrer que l'EMV est également asymptotiquement normal (voir [1]).

Le reste du paragraphe est consacré aux éléments de démonstration de la proposition 5.5.1, qui incluent une cascade de lemmes.

Démonstration de la proposition 5.5.1. Soit θ le paramètre de la chaîne de Markov $((S_n, Y_n), n \in \mathbb{N}^*)$. Grâce à (H_δ), la chaîne de Markov est irréductible. Comme l'espace d'état est fini, elle possède une unique probabilité invariante, μ_θ, d'après la remarque 1.5.7.

On continue la démonstration sous l'hypothèse que (S_1, Y_1) a pour loi μ_θ. Pour tout $n \geq 1$, (S_n, Y_n) a pour loi μ_θ, et plus généralement, pour tout $k \in \mathbb{N}^*$, les suites $((S_{n+k}, Y_{n+k}), n \geq -k)$ ont même loi, c'est-à-dire que pour tous $m, k \in \mathbb{N}^*$, les suites $((S_{n+k}, Y_{n+k}), m \geq n \geq -k)$ ont même loi. Le théorème d'extension de Kolmogorov (voir [3], appendice II) permet d'une certaine manière de passer à la limite $k \to \infty$, et plus précisément de construire une suite $Z = (Z_n, n \in \mathbb{Z})$ telle que pour tout $k \in \mathbb{N}^*$, la suite $(Z_{n+k}, n \geq 0)$ est une chaîne de Markov issue de μ_θ et de même loi que $((S_n, Y_n), n \in \mathbb{N}^*)$ (i.e. même loi initiale et même matrice de transition). En particulier, la loi de Z_n est la loi invariante μ_θ. Par abus de notation, on écrit $Z_n = (S_n, Y_n)$ pour $n \in \mathbb{Z}$. Soit $y = (\ldots, y_{-1}, y_0) \in \mathcal{X}^{-\mathbb{N}}$. Pour tout $n \in \mathbb{N}^*$, on définit la fonction

$$g_n(\theta\,; y) = \mathbb{P}_\theta(Y_0 = y_0 | Y_{-n}^{-1} = y_{-n}^{-1}),$$

où $\theta \in \Theta_\delta$ est le paramètre de la loi de Z. Le lemme suivant, dont la démonstration est reportée à la suite de celle-ci, assure que pour $n > 0$ grand, l'information supplémentaire donnée par $Y_{-(n+1)} = y_{-(n+1)}$ a peu d'influence sur la connaissance de Y_0 quand on connaît déjà Y_{-n}^{-1}.

Lemme 5.5.3. *Il existe $\rho \in [0, 1[$ tel que pour tous $\theta \in \Theta_\delta$, $y \in \mathcal{X}^{-\mathbb{N}}$, $n \in \mathbb{N}^*$, on a*

$$|g_n(\theta\,; y) - g_{n+1}(\theta\,; y)| \leq \rho^{n-1}.$$

De plus les fonctions g_n sont uniformément minorées par une constante $c > 0$.

On déduit de ce lemme que la suite de fonctions $(g_n, n \geq 1)$ converge uniformément en $y \in \mathcal{X}^{-\mathbb{N}}$, $\theta \in \Theta_\delta$ vers une limite g. Les fonctions g_n étant continues en θ à valeurs dans $[c,1]$, on en déduit que la fonction g est continue en θ, à valeurs dans $[c,1]$.

La fin de la démonstration de la proposition est scindée en trois étapes : dans la première on construit la fonction $\mathcal{H}_{\theta_0}$, dans la deuxième on vérifie la convergence énoncée dans la proposition, enfin dans la dernière étape on montre que la fonction $\mathcal{H}_{\theta_0}$ atteint son maximum en θ_0.

Première étape. Remarquons que $g_n(\theta\,;(Y_r, r \leq k))$ est en fait une fonction de Y_{k-n}^k et donc une fonction de Z_{k-n}^k que l'on note $\exp f_n$:

$$\log g_n(\theta\,;(Y_r, r \leq k)) = f_n(Z_{k-n}^k). \tag{5.12}$$

Comme $g_n \in [c,1]$, on en déduit que les fonctions f_n sont bornées et négatives. Comme $(Z_{k-n}, k \geq 0)$ est une chaîne de Markov irréductible de probabilité invariante μ_{θ_0}, on déduit du lemme 1.5.10 que $(Z_{k-n}^k, k \geq 0)$ est une chaîne de Markov irréductible de probabilité invariante la loi de Z_{-n}^0 (rappelons que la loi de Z_{-n} est la probabilité invariante μ_{θ_0}). Comme la fonction f_n est bornée, on déduit de (5.12) et du corollaire 1.5.11 que

$$\frac{1}{N}\sum_{k=1}^{N} \log g_n(\theta\,;(Y_r, r \leq k)) \xrightarrow[n\to\infty]{\textbf{p.s.}} \mathbb{E}_{\theta_0}[f_n(Z_{-n}^0)] = \mathbb{E}_{\theta_0}[\log g_n(\theta\,;(Y_r, r \leq 0))]. \tag{5.13}$$

Par convergence dominée, on a

$$\lim_{n\to\infty} \mathbb{E}_{\theta_0}[\log g_n(\theta\,;(Y_r, r \leq 0))] = \mathcal{H}_{\theta_0}(\theta), \tag{5.14}$$

où $\mathcal{H}_{\theta_0}(\theta) = \mathbb{E}_{\theta_0}[\log g(\theta\,;(Y_r, r \leq 0))]$. La fonction g étant continue en θ, on en déduit, par convergence dominée, que la fonction $\theta \to \mathcal{H}_{\theta_0}(\theta)$ est continue sur Θ_δ.

Deuxième étape. Soit $N \geq n \geq 1$. On pose

$$A_N^n = \left| \frac{1}{N} \log p_N(\theta\,; Y_1^N) - \frac{1}{N}\sum_{k=1}^{N} \log g_n(\theta\,;(Y_r, r \leq k)) \right|.$$

Il existe C tel que pour tous $a, b \geq c$, on a $|\log a - \log b| \leq C\,|a-b|$. Il vient en utilisant le lemme 5.5.3, ainsi que $\log p_N(\theta\,; Y_1^N) = \sum_{k=1}^N \log g_k(\theta\,;(Y_r, r \leq k))$,

$$\begin{aligned} A_N^n &\leq \frac{1}{N}\sum_{k=1}^{N} |\log g_k(\theta\,;(Y_r, r \leq k)) - \log g_n(\theta\,;(Y_r, r \leq k))| \\ &\leq C\,\frac{1}{N}\sum_{k=1}^{N} |g_k(\theta\,;(Y_r, r \leq k)) - g_n(\theta\,;(Y_r, r \leq k))| \end{aligned}$$

$$
\begin{aligned}
&\leq C\,\frac{1}{N}\sum_{k=1}^{n}|g_k(\theta\,;(Y_r,r\leq k))-g_n(\theta\,;(Y_r,r\leq k))|\\
&\qquad\qquad +C\,\frac{1}{N}\sum_{k=n+1}^{N}\sum_{l=n}^{k-1}|g_{l+1}(\theta\,;(Y_r,r\leq k))-g_l(\theta\,;(Y_r,r\leq k))|\\
&\leq 2Cc\frac{n}{N}+C\,\frac{1}{N}\sum_{k=n+1}^{N}\sum_{l=n}^{k-1}\rho^{l-1}\\
&\leq 2Cc\frac{n}{N}+C\frac{\rho^{n-1}}{1-\rho}.
\end{aligned}
$$

On en déduit que

$$\lim_{n\to\infty}\lim_{N\to\infty}A_N^n=0.$$

Comme d'après (5.13) et (5.14), p.s.

$$\lim_{n\to\infty}\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\log g_n(\theta\,;(Y_r,r\leq k))=\mathcal{H}_{\theta_0}(\theta),$$

on en déduit que

$$\frac{1}{N}\,L_N(\theta\,;Y_1^N)\xrightarrow[N\to+\infty]{\textbf{p.s.}}\mathcal{H}_{\theta_0}(\theta).$$

On admet que le résultat reste vrai, même si la loi de (S_1,Y_1) n'est pas la probabilité invariante μ_{θ_0}.

Troisième étape. On remarque que

$$
\begin{aligned}
\mathbb{E}_{\theta_0}[\log g_n(\theta\,;(Y_r,r\leq 0))] &= \sum_{y_0^n\in\mathcal{X}^{n+1}}p_{n+1}(\theta_0\,;y_0^n)\log p(\theta\,;y_n|y_0^{n-1})\\
&= \sum_{y_0^{n-1}\in\mathcal{X}^n}p_n(\theta_0\,;y_0^{n-1})h_{\theta_0}(\theta),
\end{aligned}
$$

où la fonction h dépend de y_0^{n-1} :

$$h_{\theta_0}(\theta)=\sum_{y\in\mathcal{X}}p(\theta_0\,;y|y_0^{n-1})\log p(\theta\,;y|y_0^{n-1}).$$

Remarquons que $h_{\theta_0}(\theta)$ peut s'écrire $\sum_{x\in\mathcal{X}}p(x)\log p'(x)$ avec les probabilités $p=p(\theta_0\,;\cdot|y_0^{n-1})$ et $p'=p(\theta\,;\cdot|y_0^{n-1})$. D'après le lemme 5.2.8, on a $h_{\theta_0}(\theta)\leq h_{\theta_0}(\theta_0)$ pour tout $\theta\in\Theta_\delta$. En particulier, on en déduit que pour tout $\theta\in\Theta_\delta$,

$$\mathbb{E}_{\theta_0}[\log g_n(\theta\,;(Y_r,r\leq 0))]\leq\mathbb{E}_{\theta_0}[\log g_n(\theta_0\,;(Y_r,r\leq 0))].$$

Par passage à la limite, on obtient que $\mathcal{H}_{\theta_0}(\theta)\leq\mathcal{H}_{\theta_0}(\theta_0)$ pour tous $\theta,\theta_0\in\Theta_\delta$. On admet que le modèle étant identifiable, l'inégalité est stricte si $\theta\neq\theta_0$. □

Démonstration du lemme 5.5.3. Pour $n \geq 1$, on considère les quantités

$$M_n^+(y) = \max_{i \in \mathcal{I}} \mathbb{P}_\theta(Y_0 = y_0 | Y_{-n}^{-1} = y_{-n}^{-1}, S_{-n} = i)$$

et

$$M_n^-(y) = \min_{i \in \mathcal{I}} \mathbb{P}_\theta(Y_0 = y_0 | Y_{-n}^{-1} = y_{-n}^{-1}, S_{-n} = i).$$

On a

$$\begin{aligned} g_n(\theta\,; y) &= \frac{\mathbb{P}_\theta(Y_0 = y_0, Y_{-n}^{-1} = y_{-n}^{-1})}{\mathbb{P}_\theta(Y_{-n}^{-1} = y_{-n}^{-1})} \\ &= \frac{\sum_{i \in \mathcal{I}} \mathbb{P}_\theta(Y_0 = y_0, Y_{-n}^{-1} = y_{-n}^{-1}, S_{-n} = i)}{\sum_{i \in \mathcal{I}} \mathbb{P}_\theta(Y_{-n}^{-1} = y_{-n}^{-1}, S_{-n} = i)}. \end{aligned}$$

Soit $a_r > 0$, $b_r > 0$ pour $1 \leq r \leq k$. On a les inégalités $b_u \min_{1 \leq r \leq k} \frac{a_r}{b_r} \leq a_u \leq b_u \max_{1 \leq r \leq k} \frac{a_r}{b_r}$. En sommant sur $u \in \{1, \ldots, k\}$, il vient aisément

$$\min_{1 \leq r \leq k} \frac{a_r}{b_r} \leq \frac{\sum_{1 \leq u \leq k} a_u}{\sum_{1 \leq u \leq k} b_u} \leq \max_{1 \leq r \leq k} \frac{a_r}{b_r}. \tag{5.15}$$

On en déduit que

$$M_n^-(y) \leq g_n(\theta\,; y) \leq M_n^+(y). \tag{5.16}$$

De plus, on a

$$\begin{aligned} g_{n+1}(y) &= \mathbb{P}_\theta(Y_0 = y_0 | Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1}) \\ &= \sum_{i \in \mathcal{I}} \mathbb{P}_\theta(Y_0 = y_0, S_{-n} = i | Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1}) \\ &= \sum_{i \in \mathcal{I}} \mathbb{P}_\theta(Y_0 = y_0 | Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1}, S_{-n} = i) \\ &\qquad\qquad \mathbb{P}_\theta(S_{-n} = i | Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1}) \\ &= \sum_{i \in \mathcal{I}} \mathbb{P}_\theta(Y_0 = y_0 | Y_{-n}^{-1} = y_{-n}^{-1}, S_{-n} = i) \\ &\qquad\qquad \mathbb{P}_\theta(S_{-n} = i | Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1}), \end{aligned}$$

où on a utilisé la deuxième égalité du lemme 5.4.3 pour la dernière égalité. Comme $\sum_{i \in \mathcal{I}} \mathbb{P}_\theta(S_{-n} = i | Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1}) = 1$, on en déduit que

$$M_n^-(y) \leq g_{n+1}(\theta\,; y) \leq M_n^+(y). \tag{5.17}$$

Grâce à (5.16) et (5.17), on obtient $|g_n(\theta\,; y) - g_{n+1}(\theta\,; y)| \leq M_n^+(y) - M_n^-(y)$. La démonstration du lemme sera complète dès que le lemme suivant sera démontré. □

Lemme 5.5.4. *Il existe $\rho \in [0,1[$ tel que, pour tous $\theta \in \Theta_\delta$, $y \in \mathcal{X}^{-\mathbb{N}}$ et $n \in \mathbb{N}^*$, on a*

$$M_n^+(y) - M_n^-(y) \leq \rho^{n-1}.$$

De plus les fonctions M_n^- sont uniformément minorées par une constante $c > 0$.

Pour cela on démontre d'abord le lemme technique suivant.

Lemme 5.5.5. *Il existe $\eta_\delta > 0$ tel que pour tous $\theta \in \Theta_\delta$, $i, h \in \mathcal{I}$, $y \in \mathcal{X}^{-\mathbb{N}}$ et $n \geq 1$, on a*

$$\mathbb{P}_\theta(S_{-n} = i | Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1}, S_{-(n+1)} = h) \geq \eta_\delta.$$

Démonstration du lemme 5.5.5. Soit $i, h \in \mathcal{I}$. On a pour $n \geq 2$,

$$\begin{aligned}
&\mathbb{P}_\theta(S_{-n} = i, S_{-(n+1)} = h, Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1}) \\
&\quad = \sum_{l \in \mathcal{I}} \mathbb{P}_\theta(Y_{-(n-1)}^{-1} = y_{-(n-1)}^{-1}, S_{-(n+1)}^{-(n-1)} = (h,i,l), Y_{-(n+1)}^{-n} = y_{-(n+1)}^{-n}) \\
&\quad = \sum_{l \in \mathcal{I}} \mathbb{P}_\theta(Y_{-(n-1)}^{-1} = y_{-(n-1)}^{-1} | S_{-(n+1)}^{-(n-1)} = (h,i,l), Y_{-(n+1)}^{-n} = y_{-(n+1)}^{-n}) \\
&\qquad\qquad \mathbb{P}_\theta(S_{-(n+1)}^{-(n-1)} = (h,i,l), Y_{-(n+1)}^{-n} = y_{-(n+1)}^{-n}) \\
&\quad = \sum_{l \in \mathcal{I}} \mathbb{P}_\theta(Y_{-(n-1)}^{-1} = y_{-(n-1)}^{-1} | S_{-(n-1)} = l) \\
&\qquad\qquad \mu_\theta(h) b(h, y_{-(n+1)}) a(h,i) b(i, y_{-n}) a(i,l),
\end{aligned}$$

où l'on a utilisé la première égalité du lemme 5.4.3 pour la troisième égalité. On en déduit donc que pour $i, j, h \in \mathcal{I}$, on a

$$\begin{aligned}
&\frac{\mathbb{P}_\theta(S_{-n} = j | S_{-(n+1)} = h, Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1})}{\mathbb{P}_\theta(S_{-n} = i | S_{-(n+1)} = h, Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1})} \\
&\qquad = \frac{\mathbb{P}_\theta(S_{-n} = j, S_{-(n+1)} = h, Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1})}{\mathbb{P}_\theta(S_{-n} = i, S_{-(n+1)} = h, Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1})} \\
&\qquad = \frac{\sum_{l \in \mathcal{I}} \mathbb{P}_\theta(Y_{-(n-1)}^{-1} = y_{-(n-1)}^{-1} | S_{-(n-1)} = l)}{\sum_{l' \in \mathcal{I}} \mathbb{P}_\theta(Y_{-(n-1)}^{-1} = y_{-(n-1)}^{-1} | S_{-(n-1)} = l')} \cdots \\
&\qquad\qquad \cdots \frac{\mu_\theta(h) b(h, y_{-(n+1)}) a(h,j) b(j, y_{-n}) a(j,l)}{\mu_\theta(h) b(h, y_{-(n+1)}) a(h,i) b(i, y_{-n}) a(i,l')} \\
&\qquad \leq \frac{\sum_{l \in \mathcal{I}} \mathbb{P}_\theta(Y_{-(n-1)}^{-1} = y_{-(n-1)}^{-1} | S_{-(n-1)} = l)}{\sum_{l' \in \mathcal{I}} \mathbb{P}_\theta(Y_{-(n-1)}^{-1} = y_{-(n-1)}^{-1} | S_{-(n-1)} = l')} \frac{1}{\delta^3} \\
&\qquad = \frac{1}{\delta^3},
\end{aligned}$$

où l'on a utilisé pour l'inégalité les inégalités $\delta \leq a(i',j') \leq 1$ et $\delta \leq b(i',x') \leq 1$ pour $i',j' \in \mathcal{I}$, $x' \in \mathcal{X}$. Pour $n=1$, des calculs similaires donnent

$$\frac{\mathbb{P}_\theta(S_{-1}=j|S_{-2}=h,Y_{-2}^{-1}=y_{-2}^{-1})}{\mathbb{P}_\theta(S_{-1}=i|S_{-2}=h,Y_{-2}^{-1}=y_{-2}^{-1})} \leq \frac{1}{\delta^2} \leq \frac{1}{\delta^3}\,.$$

On déduit de l'égalité

$$\sum_{j\in\mathcal{I}} \mathbb{P}_\theta(S_{-n}=j|Y_{-(n+1)}^{-1}=y_{-(n+1)}^{-1},S_{-(n+1)}=h)=1,$$

que

$$\begin{aligned}
&\frac{1}{\mathbb{P}_\theta(S_{-n}=i|Y_{-(n+1)}^{-1}=y_{-(n+1)}^{-1},S_{-(n+1)}=h)} \\
&\qquad = 1+\sum_{j\neq i\in\mathcal{I}} \frac{\mathbb{P}_\theta(S_{-n}=j|Y_{-(n+1)}^{-1}=y_{-(n+1)}^{-1},S_{-(n+1)}=h)}{\mathbb{P}_\theta(S_{-n}=i|Y_{-(n+1)}^{-1}=y_{-(n+1)}^{-1},S_{-(n+1)}=h)} \\
&\qquad \leq 1+\frac{\text{Card }(\mathcal{I})-1}{\delta^3}.
\end{aligned}$$

Donc pour $n \geq 1$, $\mathbb{P}_\theta(S_{-n}=i|Y_{-(n+1)}^{-1}=y_{-(n+1)}^{-1},S_{-(n+1)}=h)$ est uniformément minoré par une constante strictement positive. □

Démonstration du lemme 5.5.4. Remarquons dans un premier temps que

$$\begin{aligned}
&\mathbb{P}_\theta(Y_0=y_0|Y_{-(n+1)}^{-1}=y_{-(n+1)}^{-1},S_{-(n+1)}=h) \\
&=\sum_{i\in\mathcal{I}} \mathbb{P}_\theta(Y_0=y_0,S_{-n}=i|Y_{-(n+1)}^{-1}=y_{-(n+1)}^{-1},S_{-(n+1)}=h) \\
&=\sum_{i\in\mathcal{I}} \mathbb{P}_\theta(Y_0=y_0|Y_{-n}^{-1}=y_{-n}^{-1},S_{-n}=i,Y_{-(n+1)}=y_{-(n+1)},S_{-(n+1)}=h) \\
&\qquad\qquad \mathbb{P}_\theta(S_{-n}=i|Y_{-(n+1)}^{-1}=y_{-(n+1)}^{-1},S_{-(n+1)}=h) \\
&=\sum_{i\in\mathcal{I}} \mathbb{P}_\theta(Y_0=y_0|Y_{-n}^{-1}=y_{-n}^{-1},S_{-n}=i) \\
&\qquad\qquad \mathbb{P}_\theta(S_{-n}=i|Y_{-(n+1)}^{-1}=y_{-(n+1)}^{-1},S_{-(n+1)}=h),
\end{aligned} \tag{5.18}$$

où l'on a utilisé la deuxième égalité du lemme 5.4.3 pour la dernière égalité. On en déduit donc que

$$\begin{aligned}
M_{n+1}^-(y) &\geq \min_{h\in\mathcal{I}} \sum_{i\in\mathcal{I}} M_n^-(y)\mathbb{P}_\theta(S_{-n}=i|Y_{-(n+1)}^{-1}=y_{-(n+1)}^{-1},S_{-(n+1)}=h) \\
&\geq M_n^-(y).
\end{aligned}$$

En particulier la suite $(M_n^-, n \geq 1)$ est uniformément minorée par M_1^- qui est strictement positif grâce à l'hypothèse (H_δ).

De plus, on a en utilisant (5.18) à nouveau,

$$\begin{aligned}
&M_{n+1}^+(y) - M_{n+1}^-(y) \\
&\quad = \max_{h,j\in\mathcal{I}} \Big\{ \mathbb{P}_\theta(Y_0 = y_0 | Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1}, S_{-(n+1)} = h) \\
&\qquad\qquad - \mathbb{P}_\theta(Y_0 = y_0 | Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1}, S_{-(n+1)} = j) \Big\} \\
&\quad = \max_{h,j\in\mathcal{I}} \Big\{ \sum_{i\in\mathcal{I}} [A(h,i) - A(j,i)] \mathbb{P}_\theta(Y_0 = y_0 | Y_{-n}^{-1} = y_{-n}^{-1}, S_{-n} = i) \Big\},
\end{aligned}$$

où $A(l,i) = \mathbb{P}_\theta(S_{-n} = i | Y_{-(n+1)}^{-1} = y_{-(n+1)}^{-1}, S_{-(n+1)} = l)$. On considère les ensembles suivants qui dépendent de h et j :

$$\mathcal{I}^+ = \{i \in \mathcal{I} \,;\, A(h,i) - A(j,i) \geq 0\} \text{ et } \mathcal{I}^- = \{i \in \mathcal{I} \,;\, A(h,i) - A(j,i) < 0\}.$$

On a

$$\begin{aligned}
&M_{n+1}^+(y) - M_{n+1}^-(y) \\
&\quad \leq \max_{h,j\in\mathcal{I}} \Big\{ \sum_{i\in\mathcal{I}^+} [A(h,i) - A(j,i)] M_n^+(y) + \sum_{i\in\mathcal{I}^-} [A(h,i) - A(j,i)] M_n^-(y) \Big\} \\
&\quad = \max_{h,j\in\mathcal{I}} \Big\{ \sum_{i\in\mathcal{I}^+} [A(h,i) - A(j,i)] (M_n^+(y) - M_n^-(y)) \Big\},
\end{aligned}$$

en ayant remarqué pour la dernière égalité que $\sum_{i\in\mathcal{I}} A(h,i) = \sum_{i\in\mathcal{I}} A(j,i) = 1$ implique $\sum_{i\in\mathcal{I}^-} [A(h,i) - A(j,i)] = -\sum_{i\in\mathcal{I}^+} [A(h,i) - A(j,i)]$. Remarquons enfin, grâce au lemme 5.5.5, que

$$\sum_{i\in\mathcal{I}^+} [A(h,i) - A(j,i)] = 1 - \sum_{i\in\mathcal{I}^-} A(h,i) - \sum_{i\in\mathcal{I}^+} A(j,i) \leq 1 - \text{Card}\,(\mathcal{I})\eta_\delta \leq 1 - 2\eta_\delta.$$

On pose $\rho = 1 - 2\eta_\delta \in [0,1[$, et on obtient

$$M_{n+1}^+(y) - M_{n+1}^-(y) \leq \rho (M_n^+(y) - M_n^-(y)).$$

Par définition de $M_1^+(y)$ et $M_1^-(y)$, on a $0 \leq M_1^+(y) - M_1^-(y) \leq 1$. On en déduit que $M_n^+(y) - M_n^-(y) \leq \rho^{n-1}$. Cela termine la démonstration du lemme 5.5.4. □

5.6 Autres exemples d'application de l'algorithme EM

5.6.1 Le mélange

Un des premiers exemples d'étude de loi de mélange remonte à la fin du $XIX^{\text{ème}}$ siècle. Il s'agit aujourd'hui d'une problématique courante, voir par exemple [11] ou Chap. 9 dans [5].

Les crabes de Weldon

À la fin du $XIX^{\text{ème}}$ 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 5.1 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).

Si l'on suppose un modèle gaussien pour les données de ratio, on calcule à partir de (5.3) la moyenne empirique $\mu_0 \simeq 0.645$ et l'écart type empirique, racine carrée de la variance empirique, $\sigma_0 \simeq 0.019$.

L'asymétrie des données, voir les histogrammes de la figure 5.5, indique que les données ne proviennent pas de réalisations de variables gaussiennes indépendantes et de même loi. Effectivement, un test d'adéquation de loi (χ^2, Shapiro-Wilk, ..., cf. [2]) permet de rejeter l'hypothèse de normalité pour les données. Ceci induit Weldon à postuler l'existence de deux sous-populations. À partir de ces données, Pearson [13] estime les paramètres d'un

Tableau 5.1. 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)

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		

modèle à $I = 2$ populations différentes. Ainsi un crabe pris au hasard a une probabilité π_i d'appartenir à la population i, π_i étant proportionnel à la taille de la population i. Et, au sein de la population i, les mesures du ratio sont distribuées suivant une loi gaussienne réelle de moyenne μ_i, de variance σ_i^2 et de densité f_{μ_i,σ_i}. On suppose de plus que les mesures sont des réalisations de variables indépendantes $(Y_n, n \geq 1)$. L'objectif est d'estimer les probabilités π_i et les paramètres $(\mu_i, \sigma_i), i \in \mathcal{I}$. Le groupe, Z_n, du n-ième crabe mesuré est une variable cachée, que l'on essaie également de restaurer.

Dans ce qui suit, nous nous proposons d'estimer les paramètres avec leur EMV, que nous approchons à l'aide de l'algorithme EM. Plusieurs autres méthodes existent pour l'approximation de ces EMV, cf. [11, 5]. Historiquement, Pearson a estimé les paramètres de sorte que les cinq premiers moments de la loi de Y_n égalent les moments empiriques. Cette méthode conduit à rechercher les racines d'un polynôme de degré neuf.

Le modèle de mélange.

Soit $I \geq 2$ fixé. Le modèle complet est donné par une suite de variables aléatoires indépendantes de même loi, $((Z_n, Y_n), n \geq 1)$, où Z_n est à valeurs dans $\mathcal{I} = \{1, \ldots, I\}$, de loi $\pi = (\pi(i), i \in \mathcal{I})$, et la loi de Y_n sachant $Z_n = i$ a pour densité f_{μ_i,σ_i}. On observe les réalisations des variables $(Y_n, n \geq 1)$ et les variables $(Z_n, n \geq 1)$ sont cachées. Il s'agit d'un **modèle de mélange** de lois gaussiennes. Le modèle est paramétrique, de paramètre inconnu $\theta = (\pi, ((\mu_i, \sigma_i), i \in \mathcal{I})) \in \Theta = \mathcal{P}_{\mathcal{I}} \times (\mathbb{R} \times]0, \infty[)^I$, où $\mathcal{P}_{\mathcal{I}}$ est l'ensemble des probabilités sur $\mathcal{I}$.

Remarquons que le nombre I est fixé a priori. L'estimation du nombre I de populations différentes est un problème délicat en général, voir [11], Chap. 6.

Comme à la fin du paragraphe 5.1, il est facile de vérifier que le modèle est identifiable si l'on restreint l'ensemble des paramètres à $\Theta' = \{\theta \in \Theta$ tels que si $i < j \in \mathcal{I}$, alors soit $\mu_i < \mu_j$ soit $\mu_i = \mu_j$ et $\sigma_i < \sigma_j\}$. On admet alors que l'EMV de θ, $\hat{\theta}_n$, construit à partir de (Z_1^n, Y_1^n), est un estimateur convergent.

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 $z = z_1^N \in \mathcal{I}^N$, $y = y_1^N \in \mathbb{R}^N$:

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

Il est difficile de calculer numériquement l'EMV de θ du modèle incomplet. L'algorithme EM, que nous explicitons, est rapide à mettre en œuvre dans ce cadre. Les mêmes arguments permettent de démontrer le lemme 5.3.1, avec Q défini ici par

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

où $\theta, \theta' = (\pi', ((\mu_i', \sigma_i'), i \in \mathcal{I})) \in \Theta'$, et par définition

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

où pour tous $i \in \mathcal{I}, k \in \{1,\dots,N\}$

$$\rho'_{i,k} = \frac{\pi_i' f_{\mu_i',\sigma_i'}(y_k)}{f_{\theta'}(y_k)}. \tag{5.19}$$

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.

L'étape E

On explicite la fonction $Q(\theta,\theta')$. On remarque que

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

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

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

L'étape M

On écrit $Q(\theta,\theta') = NA_0 + \sum_{j\in\mathcal{I}} A_j$ avec $A_0 = \sum_{i\in\mathcal{I}} \pi_i^* \log \pi$,

$$\pi_i^* = \frac{1}{N}\sum_{k=1}^{N} \rho'_{i,k}, \tag{5.20}$$

et $A_j = \sum_{k=1}^{N} \rho'_{j,k} \log(f_{\mu_j,\sigma_j}(y_k))$ pour $j \in \mathcal{I}$. 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 $j \in \mathcal{I}$.

Comme $\pi^* = (\pi_i^*, i \in \mathcal{I})$ est une probabilité sur $\mathcal{I}$, on déduit du lemme 5.2.8 que, sous la contrainte $\pi \in \mathcal{P}_{\mathcal{I}}$, A_0 est maximal pour $\pi = \pi^*$. On cherche ensuite les zéros des dérivées de A_j par rapport à μ_j et σ_j. Comme $\log f_{\mu,\sigma}(v) = -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{(v-\mu)^2}{2\sigma^2}$, on a

$$\frac{\partial A_j}{\partial \mu_j} = \sum_{k=1}^{N} \rho'_{j,k} \frac{(y_k - \mu_j)}{\sigma_j^2},$$

et

$$\frac{\partial A_j}{\partial \sigma_j} = -\sum_{k=1}^{N} \rho'_{j,k} \frac{1}{\sigma_j}\left[1 - \frac{(y_k-\mu_j)^2}{\sigma_j^2}\right].$$

Les deux dérivées ci-dessus s'annulent en

$$\mu_j^* = \frac{\sum_{k=1}^{N} \rho'_{j,k} y_k}{\sum_{k=1}^{N} \rho'_{j,k}} \quad \text{et} \quad (\sigma_j^*)^2 = \frac{\sum_{k=1}^{N} \rho'_{j,k}(y_k - \mu_j^*)^2}{\sum_{k=1}^{N} \rho'_{j,k}}. \tag{5.21}$$

On vérifie aisément que A_j possède un unique maximum pour $(\mu_j,\sigma_j) \in \mathbb{R}\times]0,\infty[$, et qu'il est atteint en (μ_j^*,σ_j^*). On en déduit donc que $\theta \to Q(\theta,\theta')$ atteint son unique maximum en $\theta^* = (\pi^*, ((\mu_i^*,\sigma_i^*), i \in \mathcal{I}))$.

L'algorithme EM consiste donc, à partir d'un point initial $\theta^{(0)} \in \Theta'$, à itérer les opérations définies par (5.19), (5.20) et (5.21). On pourra remarquer que les actualisations (5.20) et (5.21) s'interprètent comme le calcul de la moyenne empirique et de la variance empirique pondérées par $\rho'_{i,\cdot}$, qui est la probabilité, calculée avec les anciens paramètres, d'être dans la population i.

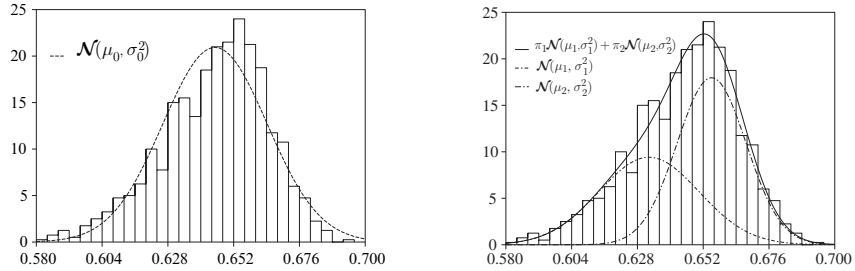

Fig. 5.5. Histogrammes des mesures pour les crabes de Weldon avec, à droite, la densité de la loi gaussienne $\mathcal{N}(\mu_0, \sigma_0^2)$ et, à gauche, les densités des lois gaussiennes $\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)$ et la densité de la loi mélange $f_\theta = \pi_1\mathcal{N}(\mu_1, \sigma_1^2)+\pi_2\mathcal{N}(\mu_2, \sigma_2^2)$. On remarque une meilleure adéquation de la densité f_θ (figure de droite) aux données par rapport à la densité gaussienne $\mathcal{N}(\mu_0, \sigma_0^2)$ (figure de gauche)

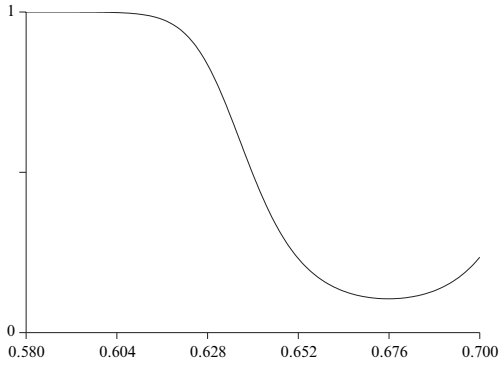

Fig. 5.6. Restauration des données manquantes pour les crabes de Weldon : probabilité d'appartenir à la population 1, sachant la valeur du ratio

Résultats

La limite et la convergence de l'algorithme dépendent peu du point de départ $\theta^{(0)}$. Dans l'exemple des crabes de Weldon, on obtient les valeurs numériques suivantes estimées par l'algorithme EM pour les paramètres du mélange :

$$\begin{aligned} \pi_1 &\simeq 0.434, \quad \mu_1 \simeq 0.632, \quad \sigma_1 \simeq 0.018, \\ \pi_2 &\simeq 0.566, \quad \mu_2 \simeq 0.655, \quad \sigma_2 \simeq 0.013. \end{aligned}$$

La figure 5.5 permet de visualiser l'adéquation des données à la densité pour les paramètres estimés. Enfin, dans la Fig. 5.6 on trace la probabilité d'appartenir à la population 1, sachant la valeur y du ratio :

$$y \to \pi(\theta\,;1|y) = \frac{\pi_1 f_{\mu_1,\sigma_1}(y)}{\pi_1 f_{\mu_1,\sigma_1}(y) + \pi_2 f_{\mu_2,\sigma_2}(y)}.$$

Pour confirmer l'adéquation des données à la loi du mélange, on peut faire un test du χ^2, voir [2]. On peut également proposer un modèle paramétrique à partir de la loi de Weibull qui est asymétrique. Cette approche fournit également une bonne adéquation aux données et donne une interprétation différente des observations.

5.6.2 Données censurées

Dans de nombreuses études, en particulier les études médicales ou les études de qualité, certaines données sont censurées, par exemple si le temps d'observation est limité par t_0. Ainsi, au lieu d'observer une réalisation de $(X_n, n \geq 1)$, suites de variables à valeurs dans $\mathbb{R}$, on observe une réalisation de $(Y_n = \min(X_n, t_0), n \geq 1)$, où t_0 est connu. Il s'agit à nouveau d'un modèle à variables cachées. Le problème qui suit permet d'expliciter l'algorithme EM dans ce cadre.

Problème 5.6.1. On se place dans le cadre d'un modèle paramétrique. Soit $(X_n, n \geq 1)$ une suite de variables aléatoires réelles indépendantes de même loi de densité f_θ, où $\theta \in \Theta$ est inconnu. On pose pour $v \in \mathbb{R}$

$$\bar{F}_\theta(v) = \mathbb{P}(X_1 \geq v) = \int_v^\infty f_\theta(r)\, dr.$$

On suppose que l'on n'observe que les réalisations de $(Y_n = \min(X_n, t_0), n \geq 1)$. La vraisemblance associée à Y_1 est donnée pour $y_1 \in \mathbb{R}$ par

$$p_1(\theta\,; y_1) = \begin{cases} f_\theta(y_1) & \text{si } y_1 < t_0, \\ \bar{F}_\theta(t_0) & \text{si } y_1 = t_0, \end{cases}$$

et la vraisemblance associée à (X_1, Y_1) est donnée pour $x_1, y_1 \in \mathbb{R}$ par

$$p_1^{complet}(\theta\,; x_1, y_1) = f_\theta(x_1)\mathbf{1}_{\{y_1 = \min(x_1, t_0)\}}.$$

On pose $\mathcal{C}_N = \{k \in \{1, \ldots, N\}\,; y_k < t_0\}$, et $n = N - \operatorname{Card} \mathcal{C}_N$, le nombre de données censurées.

1. Calculer la vraisemblance du modèle incomplet Y_1^N, $p_N(\theta\,; y)$, pour $y = y_1^N \in \mathbb{R}^N$.
2. En déduire la log-vraisemblance :

$$L_N(\theta\,; y) = \sum_{k \in \mathcal{C}_N} \log f_\theta(y_k) + n \log \bar{F}_\theta(t_0).$$

3. Dans le cas particulier du modèle exponentiel, $f_\theta(v) = \theta\, \mathrm{e}^{-\theta v}\, \mathbf{1}_{\{v>0\}}$, $\theta \in \Theta =]0, \infty[$, calculer l'EMV, $\hat{\theta}_N$, de θ construit à partir de Y_1^N. Montrer directement la convergence de l'EMV. On peut également démontrer la normalité asymptotique de l'EMV.

En général, on ne peut pas calculer explicitement l'EMV de θ. On peut alors utiliser l'algorithme EM pour donner une approximation de l'EMV.

4. Vérifier que la vraisemblance du modèle complet (X_1^N, Y_1^N), pour $x = x_1^N, y = y_1^N \in \mathbb{R}^N$ tels que $x_k = y_k$ si $k \in \mathcal{C}_N$, peut s'écrire

$$p_N^{complet}(\theta\,; x, y) = \prod_{k\in\mathcal{C}_N} f_\theta(y_k) \prod_{l\notin\mathcal{C}_N} f_\theta(x_l)\mathbf{1}_{\{x_l\geq t_0\}}.$$

On définit pour $x = x_1^N, y = y_1^N \in \mathbb{R}^N$ tels que $x_k = y_k$ si $k \in \mathcal{C}_N$,

$$\pi_N(\theta\,; x|y) = \frac{p_N^{complet}(\theta\,; x, y)}{p_N(\theta\,; y)}$$

qui s'interprète comme la loi conditionnelle des variables cachées sachant les variables observées. On pose également

$$Q(\theta, \theta') = \int_{[t_0,+\infty[^n} \pi_N(\theta'\,; x|y) \log(p_N^{complet}(\theta\,; x, y)) \prod_{l\notin\mathcal{C}_N} dx_l,$$

où $\theta, \theta' \in \Theta'$, avec la convention que si $n = 0$, alors $Q(\theta, \theta') = \sum_{k=1}^N \log f_\theta(y_k)$.

5. Montrer que

$$Q(\theta, \theta') = \sum_{k\in\mathcal{C}_N} \log f_\theta(y_k) + n\,\frac{1}{\bar{F}_{\theta'}(t_0)} \int_{[t_0,\infty[} f_{\theta'}(v) \log f_\theta(v)\; dv.$$

On peut vérifier que si h et g sont deux densités sur $\mathbb{R}$, bornées continues, et si $\int_{\mathbb{R}} g(v)\,|\log g(v)|\;\; dv < \infty$, alors $h \neq g$ implique $\int_{\mathbb{R}} g(v) \log h(v)\;\; dv < \int_{\mathbb{R}} g(v) \log g(v)\; dv$ (cf. le lemme 5.2.8 pour des variables discrètes).

On suppose que f_θ est la densité de la loi gaussienne de moyenne θ et de variance 1.

6. Vérifier que $\displaystyle\int_{\mathbb{R}} f_\theta(x)\,|\log f_\theta(x)|\;\; dx < \infty$, puis le lemme 5.3.1, pour le modèle de données censurées.
7. Montrer que $\theta \to Q(\theta, \theta')$ est maximal pour

$$\theta = \frac{1}{N}\left[\sum_{k\in\mathcal{C}_N} y_k + n\,\frac{1}{\bar{F}_{\theta'}(t_0)} \int_{[t_0,\infty[} f_{\theta'}(v) v\; dv\right].$$

8. Vérifier que $\int_{[t_0,\infty[} f_{\theta'}(v) v\; dv = \theta' \bar{F}_{\theta'}(t_0) + f_{\theta'}(t_0)$.
9. En déduire que la suite $(\theta^{(r)}, r \geq 0)$ de l'algorithme EM est définie pour $r \geq 0$ par la relation de récurrence

$$\theta^{(r+1)} = \frac{1}{N}\left[\sum_{k\in\mathcal{C}_N} y_k + n\theta^{(r)} + n\frac{f_0(t_0 - \theta^{(r)})}{\bar{F}_0(t_0 - \theta^{(r)})}\right].$$

10. On peut vérifier que pour toute valeur de $\theta^{(0)}$, la suite $(\theta^{(r)}, r \geq 0)$ converge vers une limite θ^*. En déduire l'équation satisfaite par θ^*.
11. Vérifier que la dérivée de la log-vraisemblance s'annule en θ^*.

On peut vérifier que la dérivée de la log-vraisemblance ne s'annule qu'en un seul point, θ^*, et donc la log-vraisemblance est maximale en θ^*. En particulier l'EMV de θ est θ^*. Ainsi la suite issue de l'algorithme EM converge vers l'EMV. On peut également vérifier sur cet exemple que l'EMV est convergent. ⧫

5.7 Conclusion

Pour le bactériophage lambda, on obtient après 1 000 itérations de l'algorithme EM, initialisé avec (5.10), l'approximation suivante de l'EMV des paramètres $a = (a(i,j)\,; i,j \in \{-1,+1\})$ et $b = (b(i,j)\,; i \in \{-1,+1\}, j \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\})$:

$$a \simeq \begin{pmatrix} 0.99988 & 0.00012 \\ 0.00023 & 0.99977 \end{pmatrix}, \quad b \simeq \begin{pmatrix} 0.24635 & 0.24755 & 0.29830 & 0.20780 \\ 0.26974 & 0.20845 & 0.19834 & 0.32347 \end{pmatrix}. \quad (5.22)$$

L'algorithme EM fournit également les valeurs de $\mathbb{P}(S_n = i | Y = y)$ par les équations de lissage, calculées avec l'approximation (5.22) de l'EMV de θ. La figure 5.7 met en évidence la présence de six grandes zones homogènes associées aux valeurs cachées $+1$ ou -1. Ces zones correspondent à des proportions différentes des quatre nucléotides. Ces proportions différentes pourraient provenir du fait que la transcription se fait sur le brin d'ADN analysé ou sur le brin apparié (voir [6]). Enfin les Figs. 5.8 et 5.9 représentent l'évolution des paramètres estimés, termes diagonaux de la matrice a, et termes de la matrice b, en fonction du nombre d'itérations de l'algorithme EM. On observe une convergence très rapide de l'algorithme. Toutefois, pour des initialisations éloignées des valeurs données dans (5.22), on observe une convergence de l'algorithme EM vers une valeur différente, correspondant à un maximum local de la log-vraisemblance.

Si on augmente le nombre de valeurs cachées possibles, certaines des zones précédentes se divisent, mais les résultats deviennent moins nets. On peut également choisir des modèles plus compliqués (et donc avec plus de paramètres) où la loi de Y_n peut dépendre de S_n et aussi de Y_{n-1} ; on peut aussi tenir compte dans les modèles du fait que trois nucléotides codent pour un acide aminé, etc. Nous renvoyons à [12] pour une étude très détaillée des différents modèles et des résultats obtenus pour chacun. On retrouve également dans ces modèles un découpage de l'ADN proche du cas présenté ici, où l'on se limite à deux états cachés. Les résultats sont donc robustes. Ils suggèrent donc vraiment l'existence de six zones homogènes de deux types différents.

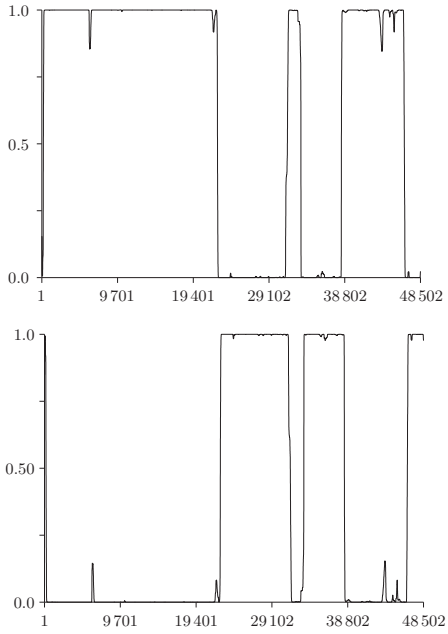

Fig. 5.7. Probabilité des états cachés pour la séquence d'ADN du bactériophage lambda dans un modèle à deux états cachés : $\mathcal{I} = \{-1, 1\}$, obtenu avec 1 000 itérations (de haut en bas : $n \to \mathbb{P}(S_n = i | Y_1^{N_0} = y_1^{N_0})$, pour $i = -1, +1$)

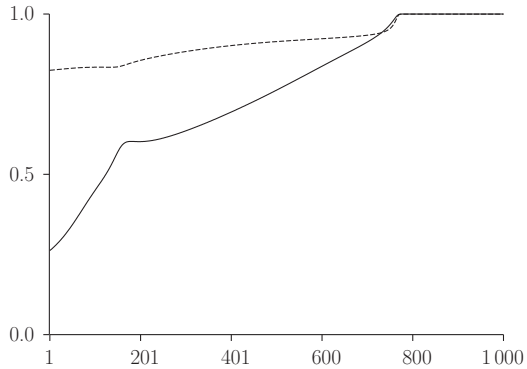

Fig. 5.8. Évolution de l'estimation des termes diagonaux de la matrice de transition des états cachés en fonction des itérations, obtenue pour 1 000 itérations

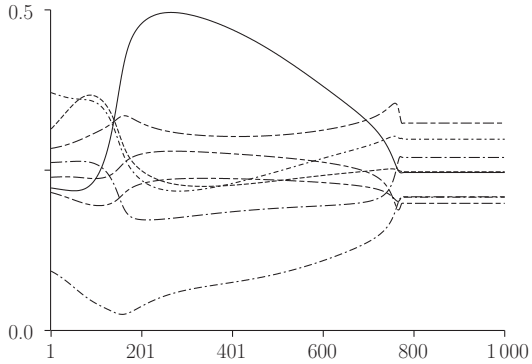

Fig. 5.9. Évolution de l'estimation des termes de la matrice b en fonction des itérations, obtenue pour 1 000 itérations

Références

1. L. Baum et T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, 37 : 1554–1563, 1966.
2. P. Bickel et K. Doksum. *Mathematical statistics. Basic ideas and selected topics.* Holden-Day Series in Probability and Statistics. Holden-Day, San Francisco, 1977.
3. P. Billingsley. *Convergence of probability measures.* John Wiley & Sons Inc., New York, 1968.
4. A.A. Borovkov. *Mathematical statistics.* Gordon and Breach Science Publishers, Amsterdam, 1998.
5. R. Casella et C. Robert. *Monte Carlo statistical methods.* Springer texts in statistics. Springer, 1999.
6. G. Churchill. Hidden Markov chains and the analysis of genome structure. *Comput. Chem.*, 16(2) : 105–115, 1992.
7. F. Dellaert. *Monte Carlo EM for data-association and its application in computer vision.* PhD thesis, Carnegie Mellon (Pittsburgh, U.S.A.), 2001.
8. A.P. Dempster, N.M. Laird et D.B. Rubin. Maximum likelihood from incomplete data via the EM algoritm (with discussion). *J. of the Royal Stat. Soc. B*, 39 : 1–38, 1977.
9. D. Forsyth et J. Ponce. *Computer vision – A modern approach.* Prentice Hall, 2003.
10. G. McLachlan et T. Krishnan. *The EM algorithm and extensions.* Wiley Series in Probability and Mathematical Statistics. Wiley & Sons, 1997.
11. G. McLachlan et D. Peel. *Finite mixture models.* Wiley Series in Probability and Mathematical Statistics. Wiley & Sons, 2001.
12. F. Muri. *Comparaison d'algorithmes d'identification de chaînes de Markov cachées et application à la détection de régions homogènes dans les séquences d'ADN.* Thèse, Université René Descartes (Paris V), 1997.

13. K. Pearson. Contributions to the theory of mathematical evolution. *Phil. Trans. of the Royal Soc. of London A*, 185 : 71–110, 1894.

14. R. Redner et H. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, 26(2) : 195–239, 1984.

15. F. Sanger, A. Coulson, G. Hong, D. Hill et G. Petersen. Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.*, 162(4) : 729–773, 1982.

16. P. Vandekerkhove. *Contribution à l'étude statistique des chaînes de Markov cachées*. Thèse, Université de Montpellier II, 1997.