Dauphine | PSL
UNIVERSITÉ PARIS

Master de Mathématiques appliquées

Mémoire de M1
Méthodes d'inférence pour les chaînes de Markov cachées

Auteurs :
Emma Kopp
Thomas Kermorvant
Emile Kosseim

Encadrant :
Julien Stoehr

2020 - 2021

Remerciements

Nous tenons à remercier sincèrement Monsieur Julien Stoehr pour son écoute, sa disponibilité et ses conseils réguliers qui nous ont permis de comprendre les enjeux du sujet et ainsi nous guider dans la rédaction de ce mémoire.

Lien d'accès aux codes

`https://github.com/emmakopp1/HMM`

Table des matières

Introduction

Depuis toujours l'Homme est soumis à l'incertitude. Il vit dans un monde qu'il ne connaît pas et est tributaire de ses lois qu'il cherche à comprendre. Dans tous les domaines, de la physique à l'économie, en passant par la biologie et la finance, il modélise son environnement et tente d'en tirer les lois qui régissent ce dernier. S'appuyant sur des principes, il émet des hypothèses et teste ses modèles en les confrontant à la réalité. Depuis l'avènement de la statistique, en considérant les observations comme des réalisations de variables aléatoires, l'homme est capable de mieux comprendre ce qui l'entoure et donc de prédire l'évolution future de sa quantité d'intérêt avec un certain degré de confiance.

Parmi les modèles standards en statistique, on retrouve notamment les modèles de Markov cachés (Baum and Petrie [1966]. Ces derniers permettent de modéliser des situations où les observations sont une représentation "bruitée" ou partielle d'un processus non observé. De par leur structure, ces modèles sont fortement utilisés avec des applications dans de multiples domaines comme la reconnaissance vocale (Rabiner [1989]), la bioinformatique (Koski [2001]) ou encore la finance (Bhar and Hamori [2004]). De nombreux résultats concernant les modèles de Markov cachés ont été établis afin de répondre à diverses problématiques. Une question fondamentale autour de ces modèles est l'inférence des paramètres du modèle (Cappé et al. [2006]) à travers l'étude de l'algorithme Expectation Maximisation (EM) (Dempster et al. [1977]) et des méthodes de Monte Carlo par chaînes de Markov (MCMC) (Robert et al. [2011]). On trouve également dans la littérature la question de la recherche du chemin le plus probable qui repose sur l'algorithme de Viterbi (Forney [1973]).

Dans ce mémoire, on introduira dans un premier temps le modèle de Markov caché (1) en présentant les différentes propriétés qui lui sont associées. Dans un deuxième temps, on étudiera les algorithmes *Forward-Backward* (2) (Cappé et al. [2006]) qui nous permettront d'obtenir la loi du processus observable. Par la suite, on parlera d'inférence des paramètres (3) en nous appuyant sur l'article de Rydén [2008]. Enfin, dans la logique d'une procédure d'inférence complète, on s'intéressera à la problématique du chemin optimal (4) grâce à l'algorithme de Viterbi.

1 Modèle de Markov caché

1.1 Rappels sur les chaînes de Markov

Soit (Ω,$\mathcal{F}$,$\mathbb{P}$) un espace de probabilité. Les variables aléatoires considérées dans cette première partie sont à valeurs dans un espace mesurable $(E, \mathcal{F})$ avec E de cardinal fini tel que $E = \{e_1, \ldots, e_n\}$.

Définition 1 : On appelle processus stochastique discret à valeurs dans E une suite de variables aléatoires $(X_t)_{t\in\mathbb{N}}$ définies sur le même espace de probabilité (Ω,$\mathcal{F}$,$\mathbb{P}$) à valeurs dans un ensemble discret E.

Dans ce mémoire on utilisera un type particulier de processus, à savoir les processus markoviens.

Définition 2 : Soit $(X_t)_{t\in\mathbb{N}}$ un processus stochastique à valeurs dans E. On dit que $(X_t)_{t\in\mathbb{N}}$ est markovien si pour tout $t \in \mathbb{N}$ et tout $(x_0, ..., x_{t-1}, x, y) \in E^{t+2}$ tel que $\mathbb{P}(X_0 = x_0, ..., X_t = x) > 0$, on a

$$\mathbb{P}(X_{t+1} = y \mid X_t = x, X_{t-1} = x_{t-1} ..., X_0 = x_0) = \mathbb{P}(X_{t+1} = y \mid X_t = x).$$

Autrement dit, la loi du futur d'un processus markovien sachant la trajectoire passée est la même que la loi du futur du processus sachant l'état présent. Cette propriété, caractéristique des processus markoviens, est également appelée propriété de Markov. Par la suite, sauf mention contraire, on travaillera avec des processus de Markov dits homogènes.

Définition 3 : On dit que le processus markovien est homogène en temps si la probabilité de transition d'un certain état à un autre est indépendante du temps, autrement dit s'il existe une fonction $A : E \times E \to \mathbb{R}$ telle que pour tout t appartenant à $\mathbb{N}$,

$$A(x, y) = \mathbb{P}\left(X_{t+1} = y | X_t = x\right).$$

Dans ce cas, on définit $(a_{ij})_{1\leq i,j\leq n}$ la matrice de transition du processus $(X_t)_{t\in\mathbb{N}}$ par

$$\forall\,(i, j) \in [\![1, n]\!]^2, \quad a_{ij} = \mathbb{P}\left(X_{t+1} = e_j \mid X_t = e_i\right).$$

On note indifféremment l'application A et sa représentation matricielle. Les coefficients a_{ij} de cette matrice représentent la probabilité de passer de l'état e_i à l'état e_j.

Exemple (Chaînes de Markov à 3 états) : On se donne $E = \{1, 2, 3\}$ et soit (Ω,$\mathcal{F}$,$\mathbb{P}$) un espace de probabilité avec $\Omega = E^t$ où $t \in \mathbb{N}^*$ et $\mathcal{F} = \mathcal{P}(\Omega)$. On considère une chaîne de Markov à valeur dans E et soit $A \in \mathcal{M}_3(\mathbb{R})$ sa matrice de transition définie par :

$$A = \begin{pmatrix} 0.3 & 0.5 & 0.2 \\ 0 & 0.3 & 0.7 \\ 0 & 0 & 1 \end{pmatrix}$$

On peut représenter cette chaîne de Markov sous la forme d'un graphe (cf Figure 1).

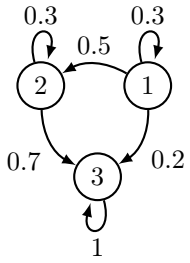

FIGURE 1 – Le graphe correspondant à la matrice A

1.2 Définition du modèle de Markov caché (*Hidden Markov Model*)

Le modèle de Markov caché (HMM) est un modèle statistique qui est utilisé pour répondre à des problématiques qui présentent des structures latentes. Il s'agit d'un processus bivarié en temps discret $(H_k, O_k)_{k\in\mathbb{N}}$ avec (H_k) un processus caché ou latent (n'étant pas observable) et (O_k) un processus observé défini conditionnellement à (H_k).

(H_k) est supposé markovien à état dans E et homogène en temps. On notera $A = (a_{ij})_{1\leq i,j\leq n}$ la matrice de transition associée à ce processus. La distribution initiale du processus est donnée par le vecteur de probabilité initiale $\mu = (\mu_1, \ldots, \mu_n)$, c'est à dire que μ_i correspond à la probabilité que l'état initial de la chaîne de Markov cachée soit l'état e_i :

$$\forall i \in [\![1, n]\!], \quad \mu_i = \mathbb{P}(H_1 = e_i).$$

Pour désigner le processus sur l'intervalle de temps $[\![1, t]\!]$, on utilisera la notation $H_{1:t} = (H_1, ..., H_t)$. De même, une réalisation de ce processus sera notée $h_{1:t} = (h_1, ...h_t)$. Enfin, par soucis de simplicité, nous serons parfois amené dans ce mémoire à adopter la notation $a_{h_i h_j}$ qui correspond à la probabilité de passer de l'état h_i à l'état h_j. Cet élément correspond bien à un coefficient de la matrice de transition. Le modèle est ensuite spécifié via la distribution conditionnelle.

Définition 4 : On appelle loi d'émission la loi conditionnelle $O_k \mid H_k$ telle que $\forall k \in [\![1, t]\!]$,

(i) Les variables aléatoires $(O_1, \ldots O_k)$ sont conditionnellement indépendantes sachant $H_1 \ldots H_k$.
(ii) L'observation O_k et $H_1, \ldots, H_{k-1}$ sont conditionnellement indépendantes sachant H_k.

À travers ce mémoire, nous étudierons le modèle de Markov caché dans le cadre de lois d'émission discrètes et continues. Dans le cas d'une loi d'émission discrète, la suite des états observables est modélisée par un processus aléatoire à état dans $V = \{v_1, \ldots, v_m\}$. On définit alors la matrice de probabilités d'émission $B = (b_j(i))_{j=1,\ldots,n}^{i=1,\ldots,m}$ par :

$$\forall i \in [\![1, m]\!], \quad \forall j \in [\![1, n]\!], \quad b_j(i) = \mathbb{P}(O_k = v_i \mid H_k = e_j),$$

où $b_j(i)$ correspond à la probabilité d'émettre l'élément v_i partant de l'état e_j. On introduit également la notation $b_j(o_k)$ qui correspond à la probabilité d'émettre l'élément o_k partant de h_j.

Dans le cas d'une loi d'émission continue, $O_k \mid H_k$ est à valeurs dans un ensemble V non dénombrable et admet pour densité $w_\theta(\cdot \mid h_k)$ où θ désigne le vecteur des paramètres de la densité considérée. Pour notre étude, on se restreindra au cas où la loi d'émission continue est gaussienne, *i.e.*,

$$O_k \mid H_k = e_i \sim \mathcal{N}(m_i, \sigma_i^2), \qquad \text{avec } B = (m_1, \sigma_1^2, ..., m_n, \sigma_n^2)$$

Dans le cas gaussien, on notera cette densité $\phi(\cdot \mid m_i, \sigma_i^2)$.

A l'instar du processus latent, pour désigner le processus (O_k) sur l'intervalle de temps $[\![1;t]\!]$, on utilisera la notation $O_{1:t} = (O_1, \dots, O_t)$ et une réalisation de ce processus sera notée $o_{1:t} = (o_1, \dots, o_t)$.

On définit alors un modèle de Markov caché par $\mathcal{M} = (E, V, \mu, A, B)$. Par la suite on notera $\theta = (A, B, \mu)$ le paramètre du modèle appartenant à $\Theta \subseteq \mathbb{R}^d$. Dans l'ensemble du mémoire, lorsque que l'on travaillera avec des densités ou des probabilités sachant un modèle paramétré par θ, on utilisera par exemple la notation f_θ ou $\mathbb{P}_\theta$.

1.3 Exemples de modèles de Markov caché

Dans ce mémoire on s'intéresse à un modèle $\mathcal{M} = (E, V, \mu, A, B)$ où le processus latent est le processus markovien à 3 états de la partie 1.1. Pour rappel $E = \{1, 2, 3\}$ et on introduit $(\Omega,\mathcal{F},\mathbb{P})$ un espace de probabilité avec $\Omega = E^t \times V^t$ et $\mathcal{F} = \mathcal{P}(E^t) \otimes \mathcal{V}$ où $\mathcal{V}$ est la tribu associée à l'ensemble V^t. La structure latente du modèle est donnée par la matrice A et le vecteur de probabilité initiale μ :

$$A = \begin{pmatrix} 0.3 & 0.5 & 0.2 \\ 0 & 0.3 & 0.7 \\ 0 & 0 & 1 \end{pmatrix} \quad \mu = \begin{pmatrix} 0.6 \\ 0.4 \\ 0 \end{pmatrix}$$

Exemple 1 (Loi d'émission discrète) : Dans le cas discret, on suppose que chaque état peut émettre un élément dans $V = \{a, b\}$ et on a $\mathcal{V} = \mathcal{P}(V^t)$. La matrice de probabilité d'émission B est donné par

$$B = \begin{pmatrix} 1 & 0 \\ 0.5 & 0.5 \\ 0 & 1 \end{pmatrix}$$

Exemple 2 (Loi d'émission gaussienne) : Dans le cas gaussien, on émet dans $V = \mathbb{R}$ et on suppose que toutes les lois gaussiennes ont la même variance $\sigma^2 = 1$ et que le paramètre inconnu est la moyenne m_i. Ici on a alors $\mathcal{V} = \mathcal{B}(\mathbb{R}^t)$ et le vecteur des paramètres B est donnée par :

$$B = \begin{pmatrix} 2 \\ 1 \\ 1.5 \end{pmatrix}$$

Remarque : On remarque dans le cas gaussien que la loi marginale du processus observé est un modèle de mélange. En effet, en intégrant sur tous les états latents possibles, la loi de O_k s'écrit comme un somme de variables gaussiennes pour tout $k \in [\![1;t]\!]$.

2 Calcul de la loi marginale du processus observable

Une question d'intérêt lorsque l'on étudie les HMM est le calcul de la loi marginale du processus observable. Nous verrons dans cette partie que c'est un problème qui se résout raisonnablement de manière théorique mais qui, en pratique, s'avère coûteux en temps de calcul. Pour palier à ce problème de complexité, nous allons nous intéresser aux algorithmes progressif-rétrograde (*backward-forward*) et nous les utiliserons pour calculer la loi marginale du processus observable.

Considérons un HMM dont la loi du processus observable est f_θ. On peut écrire la densité du processus observable de la façon suivante :

$$f_\theta\left(o_{1:t}\right) = \int_{\mathcal{H}} f_\theta\left(o_{1:t}, h_{1:t}\right)\, \nu(dh).$$

où $\mathcal{H}$ correspond à l'ensemble des séquences des états latents de longueur t et ν représente ici la mesure de comptage. D'après cette expression, on somme sur toutes les suites d'états possibles. Dès lors, ce calcul est de complexité $O(n^t)$. Pour une séquence de longueur 30 et un modèle à trois états, le coût de calcul est de l'ordre de $3^{30} \approx 10^{14}$. Cette méthode de calcul n'est donc pas possible dès lors que l'on a beaucoup de données. Pour optimiser la complexité du problème, nous utilisons un algorithme d'analyse combinatoire appelé algorithme progressif-rétrograde (*forward-backward*).

2.1 Algorithme progressif

Dans cette partie nous allons étudier la partie rétrograde de l'algorithme. Le but de cet algorithme est de calculer la loi du processus observable de façon plus rapide. Notons α_k la loi jointe de $(O_1, \dots O_k, H_k)$. On notera $\alpha_k(j)$ la loi de $(O_1, \dots O_k, H_k = e_j)$. Si l'on intègre cette loi par rapport à H_k, on obtient la loi du processus observé. On peut donc écrire :

$$\forall k \in [\![1; t]\!], \quad f(o_{1:k}) = \sum_{j=1}^{n} \alpha_k(j).$$

Pour calculer α_k, on chercher à obtenir une relation de récurrence entre α_k et α_{k-1} de façon à obtenir un algorithme récursif. Ce sera l'objet de la proposition suivante.

Proposition 1 (Algorithme progressif - relation de récurrence) : On a pour tout $j \in [\![1; n]\!]$, l'expression de la suite $(\alpha_k(j))_{k\in[\![1;t]\!]}$:

$$\alpha_k(j) = \begin{cases} \mu_j w_\theta(o_1 \mid e_j) & \text{si } k = 1 \\ \sum_{i=1}^{n} \alpha_{k-1}(i) a_{ij} w_\theta(o_k \mid e_j) & \text{sinon} \end{cases}$$

Preuve (Cas discret) : Nous effectuerons la preuve uniquement dans le cas discret. On peut écrire la quantité $\alpha_k(j)$ pour tout $j \in [\![1, n]\!]$ de la façon suivante :

$$\alpha_k(j) = \mathbb{P}_\theta(O_{1:k} = o_{1:k}, H_k = e_j).$$

En conditionnant par rapport aux états cachés, on écrit :

$$\begin{aligned} \alpha_k(j) &= \sum_{i=1}^{n} \mathbb{P}_\theta\left(O_{1:k} = o_{1:k}, H_k = e_j, H_{k-1} = e_i\right) \\ &= \sum_{i=1}^{n} \mathbb{P}_\theta\left(O_k = o_k, H_k = e_j \mid O_{1:k-1} = o_{1:k-1}, H_{k-1} = e_i\right) \mathbb{P}_\theta\left(O_{1:k-1} = o_{1:k-1}, H_{k-1} = e_i\right). \end{aligned}$$

Enfin, en utilisant (i) de la propriété 1, *i.e.* l'indépendance de la variable aléatoire O_k sachant $(H_1, \dots, H_{k-1})$ on peut écrire :

$$\alpha_k(j) = \sum_{i=1}^{n} \mathbb{P}_\theta\left(O_k = o_k, H_k = e_j \mid H_{k-1} = e_i\right) \alpha_{k-1}(i) = \sum_{i=1}^{n} \alpha_{k-1}(i)\ a_{ij}\ b_j(o_k).$$

□

On peut alors calculer explicitement la probabilité d'observation d'une séquence de façon itérative. Cette technique est nettement plus efficace. En effet, on est passé d'une complexité $O(n^t)$ à $O(n^2 t)$, ce qui rend le calcul possible rapidement, même pour une séquence très longue.

2.2 Algorithme rétrograde

Il est également possible de calculer la loi marginale du processus observé de façon rétrograde (*backward*). Cela sera l'objet de cette partie. Notons $\beta_k(j)$ la loi jointe de $O_{k+1}, \dots, O_t$ sachant H_k. L'étude de $\beta_k(j)$ est symétrique à celle de $\alpha_k(j)$. Ainsi en sommant sur l'ensemble des valeurs que prend la variable aléatoire H_k, on peut aussi écrire :

$$\forall k \in [\![1, t]\!], \quad f_\theta(o_{1:k}) = \sum_{i=1}^{n} \beta_0(i).$$

La notation β_0 signifie que l'on ne met pas de condition sur un état de la séquence.
Nous allons exhiber une relation de récurrence décroissante entre $\beta_k(j)$ et $\beta_{k+1}(j)$ pour un état j quelconque.

Proposition 2 : Pour tout $j \in [\![1, n]\!]$, l'expression de la suite $(\beta_k(j))_{k \in [\![1,t]\!]}$ s'écrit :

$$\beta_k(j) = \begin{cases} 1 & \text{si } k = t \\ \mu_j \beta_1(j) & \text{si } k = 0 \\ \sum_{i=1}^{n} w_\theta(o_{k+1} \mid e_i) a_{ji} \beta_{k+1}(i) & \text{sinon} \end{cases}$$

Preuve (Cas discret) : Nous effectuons la preuve dans le cas discret. On peut écrire la quantité $\beta_k(j)$ pour tout $j \in [\![1, n]\!]$ de la façon suivante :

$$\beta_k(j) = \mathbb{P}_\theta(O_{k+1:t} = o_{k+1:t} \mid H_k = e_j).$$

Nous allons utiliser la formule des probabilités totales sur les états latents au temps $k+1$. Alors pour tout $k \in [\![1, t]\!]$,

$$\begin{aligned} \beta_k(j) &= \mathbb{P}_\theta(O_{k+1:t} = o_{k+1:t} \mid H_k = e_j) \\ &= \sum_{i=1}^{n} \mathbb{P}_\theta\left(O_{k+1:t} = o_{k+1:t}, H_{k+1} = e_i \mid H_k = e_j\right) \\ &= \sum_{i=1}^{n} \mathbb{P}_\theta\left(H_{k+1} = e_i \mid H_k = e_j, \right) \mathbb{P}_\theta\left(O_{k+1:t} = o_{k+1:t} \mid H_{k+1} = e_i, H_k = e_j\right). \end{aligned}$$

Enfin, en isolant les variables au temps $k+1$ et de par l'indépendance de $O_{k+1} \mid H_{k+1}$ avec H_k ((i) de la propriété 1) , on en déduit :

$$\begin{aligned} \beta_k(j) &= \sum_{i=1}^{n} \mathbb{P}_\theta\left(H_{k+1} = e_i \mid H_k = e_j\right) \mathbb{P}_\theta\left(O_{k+2:t} = o_{k+2:t} \mid H_{k+1} = e_i\right) \mathbb{P}_\theta\left(O_{k+1} = o_{k+1} \mid H_{k+1} = e_i\right) \\ &= \sum_{i=1}^{n} b_i(o_{k+1}) a_{ji} \beta_{k+1}(i). \end{aligned}$$

□

De même que pour l'algorithme progressif, on améliore considérablement la complexité du problème initial. On passe d'une compléxité $O(n^t)$ à $O(n^2 t)$.

Ces deux résultats sont des cas particuliers d'une formule plus générale que nous allons présenter dans la prochaine partie. Elle permet de combiner les deux algorithmes pour obtenir la loi d'émission d'une séquence observée à partir d'un temps $t \in [\![1, t]\!]$ quelconque.

2.3 Algorithme progressif-rétrograde

Il est possible de combiner les algorithmes rétrograde-progressif. Le nouvel algorithme que nous allons présenter nous permet de calculer à partir du couple (α_k, β_k) à n'importe quel temps k dans $[\![1, t]\!]$ la loi d'émission d'une séquence.

Proposition 3 :

$$\forall k \in [\![1, t]\!], \quad f_\theta\left(o_{1:t}\right) = \sum_{i=1}^{n} \alpha_k(i)\beta_k(i).$$

Preuve (Cas discret) : Dans un premier temps, nous utilisons la formule des probabilités totales de sorte à conditionner par rapport aux états latents. En scindant la trajectoire de 1 à k et de $k+1$ à t, on obtient :

$$\begin{aligned}
\mathbb{P}_\theta\left(O_{1:t} = o_{1:t}\right) &= \sum_{i=1}^{n} \mathbb{P}_\theta\left(O_{1:k} = o_{1:k}, O_{k+1:t} = o_{k+1:t}, H_k = e_i\right) \\
&= \sum_{i=1}^{n} \mathbb{P}_\theta\left(O_{1:k} = o_{1:k}, O_{k+1:t} = o_{k+1:t} \mid H_k = e_i\right) \mathbb{P}_\theta\left(H_k = e_i\right)
\end{aligned}$$

D'après (i) de la proposition 1, on peut écrire :

$$\begin{aligned}
\mathbb{P}_\theta\left(O_{1:t} = o_{1:t}\right) &= \sum_{i=1}^{n} \mathbb{P}_\theta\left(O_{1:k} = o_{1:k} \mid H_k = e_i\right) \mathbb{P}_\theta\left(O_{k+1:t} = o_{k+1:t} \mid H_k = e_i\right) \mathbb{P}_\theta\left(H_k = e_i\right) \\
&= \sum_{i=1}^{n} \mathbb{P}_\theta\left(O_{1:k} = o_{1:k}, H_k = e_i\right) \mathbb{P}_\theta\left(O_{k+1:t} = o_{k+1:t} \mid H_k = e_i\right) \\
&= \sum_{i=1}^{n} \alpha_k(i)\beta_k(i).
\end{aligned}$$

□

Désormais, nous pouvons obtenir la probabilité d'observation d'une séquence à partir de la donnée de $(\alpha_k(i), \beta_k(i))$ pour tout $i \in E$. Nous allons illustrer ces algorithmes à partir des modèles exposés dans l'exemple introduit dans la partie 1.

2.4 Exemples sur un Modèle de Markov caché à trois états

On se place dans le cadre de l'exemple de la partie 1 avec $\mathcal{M}$ le modèle introduit précédemment. Nous allons expliciter les matrices $\alpha_{1:5}$ et $\beta_{1:5}$ pour calculer la loi d'émission de la séquence. Nous calculerons aussi cette loi à partir de la formule faisant intervenir β_0.

Exemple 1 (Loi d'émission discrète) :
Si on se place dans un cas discret, où le processus observé prend ses valeurs dans $V = \{a, b\}$, on observe la séquence $o_{1:5} = (a, b, a, a, b)$. Les matrices $\alpha_{1:5}$ et $\beta_{1:5}$ sont :

$$\alpha_{1:5} = \begin{pmatrix} 0.6 & 0 & 0 & 0 & 0 \\ 0.2 & 0.18 & 0.027 & 0.0041 & 6.1 \times 10^{-4} \\ 0 & 0.26 & 0 & 0 & 2.8 \times 10^{-3} \end{pmatrix}$$

$$\beta_{1:5} = \begin{pmatrix} 0.0048 & 0.1361 & 0.3475 & 0.45 & 1 \\ 0.0029 & 0.0191 & 0.1275 & 0.85 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

Enfin on obtient le vecteur $\beta_0 = (0.0029, 5.7 \times 10^{-3}, 0)$. Si l'on utilise l'algorithme progressif, on s'appuie sur la formule suivante :

$$\mathbb{P}_\theta\left(O_{1:t} = o_{1:t}\right) = \sum_{i=1}^{n} \alpha_t(i).$$

On en déduit en sommant les éléments de la dernière colonne le résultat : $\mathbb{P}_\theta\left(abaab\right) = 6.1 \times 10^{-4} + 2.8 \times 10^{-3} = 0.0034$. De même, en utilisant uniquement l'algorithme rétrograde et la formule :

$$\mathbb{P}_\theta\left(O_{1:t} = o_{1:t}\right) = \sum_{i=1}^{n} \beta_0(i).$$

On en déduit $\mathbb{P}_\theta\left(abaab\right) = \sum_{i=1}^{n} \beta_0(i) = 0.0029 + 5.7 \times 10^{-3} = 0.22 = 0.0034$.

Enfin, pour utiliser l'algorithme progressif-rétrograde et en utilisant la proposition 3, la probabilité recherchée est obtenue en effectuant le produit matriciel de la $(k+1)^{\text{ième}}$ colonne de α avec la $(k+1)^{\text{ième}}$ colonne de β. Si on choisit $k = 3$, on obtient alors : $\mathbb{P}_\theta\left(O_{1:t} = o_{1:t}\right) = 0.0048 \times 0.6 + 0.0029 \times 0.2 = 0.0034$. Ce qui conclut l'exemple dans le cas discret.

Exemple 2 (Loi d'émission continue) :
Si on se place dans un cas continu gaussien, où le processus observé prend ses valeurs dans $V = \mathbb{R}$, on observe la séquence $o_{1:5} = (1.73, 2.17, 1.28, 1.44, 1.51)$. Pour éviter la redondance des résultats, nous ne montrerons que le cas général avec l'algorithme progressif-rétrograde. Avec une loi d'émission gaussienne, les formules de récurrence de α_t et β_t sont les suivantes : pour tout $j \in [\![1; n]\!]$ et pour tout $k \in [\![1; t]\!]$,

$$\alpha_k(j) = \begin{cases} \mu_j \ \phi_\theta(o_1; m_j, \sigma^2) & \text{si } k = 1 \\ \sum_{i=1}^{n} \alpha_{k-1}(i) a_{ij} \ \phi_\theta(o_k \mid m_j, \sigma^2) & \text{sinon} \end{cases}$$

$$\beta_k(j) = \begin{cases} 1 & \text{si } k = t \\ \mu_j \beta_1(j) & \text{si } k = 0 \\ \sum_{i=1}^{n} \phi(o_{k+1} \mid m_j, \sigma^2) a_{ji} \beta_{k+1}(i) & \text{sinon} \end{cases}$$

On obtient les matrices $\alpha_{1:5}$ et $\beta_{1:5}$ suivantes :

$$\alpha_{1:5} = \begin{pmatrix} 0.2308 & 0.272 & 0.0025 & 0.0003 & 2.7 \times 10^{-5} \\ 0 & 0.0232 & 0.0079 & 0.0013 & 1.8 \times 10^{-4} \\ 0.1554 & 0.0642 & 0.335 & 0.0157 & 6.6 \times 10^{-3} \end{pmatrix}$$

$$\beta_{1:5} = \begin{pmatrix} 0.0165 & 0.0544 & 0.1383 & 0.3611 & 1 \\ 0.0175 & 0.0609 & 0.1530 & 0.3843 & 1 \\ 0.0197 & 0.0619 & 0.1589 & 0.3989 & 1 \end{pmatrix}.$$

Ces deux matrices nous permettent d'obtenir le résultat recherché. On obtient alors :

$$f_\theta\left(1.73, 2.17, 1.28, 1.44, 1.51\right) = 0.0069.$$

3 Inférence des paramètres

On s'intéresse dans cette partie aux méthodes d'inférence des paramètres dans un modèle de Markov caché. Pour ce faire et face à la particularité de la structure de ce modèle, nous serons amenés à étudier deux algorithmes. Le premier est l'algorithme *Expectation-Maximisation* (EM) (Dempster et al. [1977]), le second est l'algorithme de Gibbs (Gelfand and Smith [1990]) qui est une méthode de Monte-Carlo par Chaine de Markov (MCMC)(Robert et al. [2011]) que nous présenterons dans un contexte bayésien. Enfin, en nous basant sur l'article de Rydén [2008] nous conclurons cette partie par un exemple d'application de l'algorithme de Gibbs dans le cadre des HMM.

3.1 Algorithme Expectation-Maximisation (EM)

3.1.1 Présentation de l'algorithme EM

Afin d'aborder de façon progressive notre problématique d'inférence dans le cadre des HMM, nous allons tout d'abord commencer par présenter la théorie de l'algorithme EM (Dempster et al. [1977]). Cet algorithme, que l'on doit au statisticien Arthur Pentland Dempster, permet de résoudre de façon itérative le problème de maximisation de la vraisemblance dans un contexte où le calcul de cette quantité peut s'avérer complexe. Il est assez courant d'avoir recours à cet algorithme pour déterminer les paramètres d'un modèle qui dépend de variables latentes non observables. Baum et al. [1970] ont présenté une version de cet algorithme dans le cas particulier des HMM.

Pour formaliser le problème, on introduit $(\Omega,\mathcal{F},\mathbb{P}_\theta)$ un espace de probabilité avec θ qui appartient à $\Theta \subseteq \mathbb{R}^d$, l'espace des paramètres. On considère un modèle statistique associé à la loi jointe d'un couple de variables aléatoires $(O_k, H_k)_{k\in\mathbb{N}}$ où (O_k) est un processus observable et (H_k) un processus latent. Nous introduisons $L\left(\cdot \mid o_{1:t}\right)$ la fonction de vraisemblance des observations $o_{1:t}$ qui prend en argument le paramètre θ. Si on note f_θ la densité marginale du processus (O_k), on peut écrire :

$$L\left(\theta \mid o_{1:t}\right) = \int_{\mathcal{H}} f_\theta\left(o_{1:t}, h_{1:t}\right)\, \nu(dh).$$

où $\mathcal{H}$ qui correspond à l'ensemble des séquences des états latents de longueur t et ν la mesure de comptage. Une procédure classique d'inférence consisterait à maximiser la vraisemblance, mais dans notre modèle considéré, ce n'est pas possible en pratique. En effet, comme expliqué durant la partie 2, lorsque les variables latentes prennent leur valeurs dans E avec Card$(E)= n$, l'intégrale sur $\mathcal{H}$ représente la

somme sur toutes les trajectoires possibles qui sont du nombre de n^t. On est face à un problème combinatoire qui rend donc la maximisation de la vraisemblance incalculable en pratique. L'algorithme EM permet de contourner ce problème en introduisant la vraisemblance complète que l'on notera par la suite $L^c(\cdot \mid o_{1:t}, h_{1:t})$ qui contient à la fois les données observées et les données cachées. Si on note $\kappa_\theta(\cdot \mid o_{1:t})$ la densité conditionnelle de la variable $H_k \mid O_k$, on peut alors écrire

$$L^c(\theta \mid o_{1:t}, h_{1:t}) = L(\theta \mid o_{1:t}) \times \kappa_\theta(h_{1:t} \mid o_{1:t}).$$

En passant à la log-vraisemblance et en intégrant par rapport à la mesure $\kappa_{\theta'}(. \mid o_{1:t})$ paramétré par θ' on a

$$\int_{\mathcal{H}} \log(L(\theta \mid o_{1:t}))\, \kappa_{\theta'}(h_{1:t} \mid o_{1:t})\, \nu(dh) = \int_{\mathcal{H}} \{\log(L^c(\theta \mid o_{1:t}, h_{1:t})) - \log(\kappa_\theta(h_{1:t} \mid o_{1:t}))\}\, \kappa_{\theta'}(h_{1:t} \mid o_{1:t})\, \nu(dh)$$

$\kappa_{\theta'}(. \mid o_{1:t})$ étant une densité on a

$$\log(L(\theta \mid o_{1:t})) \underbrace{\int_{\mathcal{H}} \kappa_{\theta'}(h_{1:t} \mid o_{1:t})\, \nu(dh)}_{=1} = \int_{\mathcal{H}} \{\log(L^c(\theta \mid o_{1:t}, h_{1:t})) - \log(\kappa_\theta(h_{1:t} \mid o_{1:t}))\}\, \kappa_{\theta'}(h_{1:t} \mid o_{1:t})\, \nu(dh).$$

Puisque le paramètre de la densité κ caractérise sa loi, on peut finalement écrire la formule précédente sous la forme d'une espérance de la log-vraisemblance complète prise par rapport à la loi de $H_{1:t} \mid O_{1:t}$.

$$\log(L(\theta \mid o_{1:t})) = \underbrace{\mathbb{E}_{\theta'}[\log L^c(\theta \mid O_{1:t}, H_{1:t}) \mid O = o_{1:t}]}_{=:Q(\theta;\theta')} - \underbrace{\mathbb{E}_{\theta'}[\log \kappa_\theta(H_{1:t} \mid O_{1:t}) \mid O = o_{1:t}]}_{=:R(\theta;\theta')} \tag{3.1.1}$$

On souhaite maintenant maximiser la quantité $\log(L(\theta \mid o_{1:t}))$. Pour ce faire, nous allons nous appuyer sur la proposition suivante.

Proposition 4 : Pour toute paire $(\theta, \theta') \in \Theta^2$ on a

$$R(\theta'; \theta) \leq R(\theta; \theta).$$

Preuve : Soient $\theta, \theta' \in \Theta$, alors on a

$$\begin{aligned} R(\theta'; \theta) - R(\theta; \theta) &= \mathbb{E}_\theta[\log \kappa_{\theta'}(H_{1:t} \mid O_{1:t}) \mid O_{1:t} = o_{1:t}] - \mathbb{E}_\theta[\log \kappa_\theta(H_{1:t} \mid O_{1:t}) \mid O = o_{1:t}] \\ &= \mathbb{E}_\theta\left[\log \frac{\kappa_{\theta'}(H_{1:t} \mid O_{1:t})}{\kappa_\theta(H_{1:t} \mid O_{1:t})} \mid O = o_{1:t}\right]. \end{aligned}$$

La fonction $x \longmapsto \log(x)$ étant concave, l'inégalité de Jensen donne

$$\begin{aligned} \mathbb{E}_\theta\left[\log \frac{\kappa_{\theta'}(H_{1:t} \mid O_{1:t})}{\kappa_\theta(H_{1:t} \mid O_{1:t})} \mid O = o_{1:t}\right] &\leq \log\left(\mathbb{E}_\theta\left[\frac{\kappa_{\theta'}(H_{1:t} \mid O_{1:t})}{\kappa_\theta(H_{1:t} \mid O_{1:t})} \mid O_{1:t} = o_{1:t}\right]\right) \\ &\leq \log\left(\int_{\mathcal{H}} \frac{\kappa_{\theta'}(h \mid o_{1:t})}{\kappa_\theta(h_{1:t} \mid o_{1:t})} \kappa_\theta(h_{1:t} \mid o_{1:t})\, \nu(dh)\right) \\ &\leq \log\left(\int_{\mathcal{H}} \kappa_{\theta'}(h_{1:t} \mid o_{1:t})\, \nu(dh)\right) \\ &\leq \log(1) \\ &\leq 0. \end{aligned}$$

On obtient finalement que $R(\theta'; \theta) - R(\theta; \theta) \leq 0$

□

Une conséquence importante de la proposition 4 est qu'elle nous permet de trouver une condition suffisante pour que la maximisation de $\log\left(L\left(\theta \mid o_{1:t}\right)\right)$ ait lieu. Supposons qu'il existe θ^m et $\theta^{m+1} \in \Theta$ avec $m \in \mathbb{N}$ satisfaisant

$$\text{(H1)}: \quad Q\left(\theta^{m+1}; \theta^m\right) \geq Q\left(\theta^m; \theta^m\right).$$

Proposition 5 : Si l'hypothèse (H1) est vérifiée alors $\log\left(L\left(\theta^{m+1} \mid o_{1:t}\right)\right) \geq \log\left(L\left(\theta^m \mid o_{1:t}\right)\right)$.

Preuve : Soit $m \in \mathbb{N}$, il suffit de montrer que $\log\left(L\left(\theta^{m+1} \mid o_{1:t}\right)\right) - \log\left(L\left(\theta^m \mid o_{1:t}\right)\right) \geq 0$

On a $\log\left(L\left(\theta^{m+1} \mid o_{1:t}\right)\right) - \log\left(L\left(\theta^m \mid o_{1:t}\right)\right) = \underbrace{Q\left(\theta^{m+1}; \theta^m\right) - Q\left(\theta^m; \theta^m\right)}_{(1)} - \left[R\left(\theta^{m+1}; \theta^m\right) - R\left(\theta^m; \theta^m\right)\right]$.

Or $R\left(\theta^{m+1}; \theta^m\right) - R\left(\theta^m; \theta^m\right) \leq 0$ d'après la proposition 4. De plus d'après (H1) on a que (1) est ≥ 0. D'où le résultat.

□

Pour faire croître la vraisemblance, il suffit donc de construire une suite de points $\left(\theta^m\right)_{m \in \mathbb{N}}$ dans $\Theta^{\mathbb{N}}$ telle que (H1) est vérifiée. Un candidat naturel pour θ^{m+1} est

$$\theta^{m+1} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, Q\left(\theta; \theta^m\right)$$

qui est celui utilisé habituellement pour l'algorithme EM.

Algorithm 1: Algorithme EM

Input : $\theta^0 \in \Theta$, $\epsilon > 0$, $o_{1:t}$
Set : $m = 0$
while $|Q(\theta^{m+1}; \theta^m) - Q(\theta^m; \theta^m)| > \epsilon$ **do**
  Étape E : $Q\left(\theta; \theta^m\right) = \mathbb{E}_{\theta^m}\left[\log L^c\left(\theta \mid O_{1:t}, H_{1:t}\right) \mid O_{1:t} = o_{1:t}\right]$
  Étape M : $\theta^{m+1} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, Q\left(\theta; \theta^m\right)$;
  m = m+1 ;
end

La maximisation de Q n'est néanmoins pas nécessaire. D'après (H1) il suffit de faire croître Q et on peut utiliser un algorithme de descente de gradient. Cependant, il convient de vérifier que le pas α est choisi intelligemment de manière à ce que (H1) soit vérifiée. Pour ce faire, on peut notamment utiliser le principe de recherche linéaire.

Algorithm 2: Algorithme EM - Descente de Gradient

Input : $\theta^0 \in \Theta$, $\epsilon > 0$, $\alpha > 0$
Set : $m = 0$
while $|Q(\theta^{m+1}; \theta^m) - Q(\theta^m; \theta^m)| > \epsilon$ **do**
  Étape E : $Q\left(\theta; \theta^m\right) = \mathbb{E}_{\theta^m}\left[\log L^c\left(\theta \mid O_{1:t}, H_{1:t}\right) \mid O_{1:t} = o_{1:t}\right]$
  Étape M : $\theta^{m+1} = \theta^m - \alpha \nabla Q(\theta^m; \theta^m)$
  m = m+1
end

.

3.1.2 Application de l'algorithme EM dans le cadre des HMM (Formules de Baum-Welch)

Nous présentons dans cette partie l'application de l'algorithme EM au cas des HMM (Baum et al. [1970]). On rappelle qu'à chaque étape (E), on souhaite calculer :

$$Q\left(\theta^{(m+1)}, \theta^{(m)}\right) = \mathbb{E}_{\theta^m}\left[\log L^c\left(\theta^{(m+1)} \mid O_{1:t}, H_{1:t}\right) \mid O_{1:t} = o_{1:t}\right].$$

Calculons $Q(\theta, \theta')$ pour $(\theta, \theta') \in \Theta^2$ où l'on supposera θ' antérieur à θ. De plus, on notera $Q = Q(\theta, \theta')$. En utilisant le rapport de vraisemblance de la loi jointe des processus (O_k, H_k) sous θ et θ', on obtient :

$$\begin{aligned} Q &= \mathbb{E}_{\theta'}\left[\log \mu_{h_1} + \sum_{k=1}^{t} \log(a_{h_{k-1}h_k}) + \sum_{k=1}^{t} \log(w_\theta(o_k \mid h_k)) \mid O_{1:t} = o_{1:t}\right] + \mathrm{C} \\ &= \sum_{e_i \in E} \kappa_{\theta'}(e_i \mid o_{1:t}) \log(\mu_i) + \sum_{e_i, e_j \in E^2} \sum_{k=1}^{t} \kappa_{\theta'}(e_i, e_j \mid o_{1:t}) \log(a_{ij}) + \sum_{e_i \in E} \sum_{k=1}^{t} \kappa_{\theta'}(e_i \mid o_{1:t}) \log(w_\theta(o_k \mid h_k)) + \mathrm{C}. \end{aligned}$$

Enfin, avant d'énoncer un théorème important dans la résolution de l'étape (M) et pour faciliter les notations, nous allons introduire différentes variables qui interviendront par la suite.

Définition 5 : Soit un paramètre $\theta \in \Theta$. On note $\xi_k(\cdot, \cdot; \theta)$ la densité jointe de H_k et H_{k+1} sachant la séquence $O_{1:t}$ et le paramètre θ. Ainsi pour tout $(i, j) \in [\![1, n]\!]^2$, $\xi_k(i, j; \theta)$ correspond à la densité que l'état e_i ait été émis par o_k et que l'état e_j ait été émis par o_{k+1}. On a donc :

$$\forall k \in [\![1, n]\!], \quad \xi_k(i, j; \theta) = \kappa_\theta\left(e_i, e_j \mid o_{1:t}\right).$$

Par définition des quantités $\alpha_k(i)$ et $\beta_k(j)$ (*Forward/Backward*) introduites dans la partie 2, on a la proposition suivante.

Proposition 6 : Soit $o_{1:t}$ une séquence de longueur t. Alors,

$$\forall k \in [\![1, t]\!], \quad \xi_k(i, j; \theta) = \frac{\alpha_k(i)\ a_{ij}\ w_\theta(o_{k+1} \mid h_j)\ \beta_{k+1}(j)}{f_\theta\left(o_{1:t}\right)}.$$

Preuve (Cas discret) : Nous effectuons la preuve uniquement dans le cas discret. Dans un premier temps, nous allons scinder la trajectoire $o_{1:t}$ en deux trajectoires $o_{1:k}$ et $o_{k+1:t}$ et en utilisant la formule de probabilité conditionnelle, on écrit :

$$\begin{aligned} \mathbb{P}_\theta\left(H_k = e_i, H_{k+1} = e_j, O_{1:t}\right) &= \mathbb{P}_\theta\left(o_{1:k}, o_{k+1:t}, H_k = e_i, H_{k+1} = e_j\right) \\ &= \mathbb{P}_\theta\left(o_{1:k}, o_{k+1:t} \mid H_k = e_i, H_{k+1} = e_j\right) \mathbb{P}_\theta\left(H_k = e_i, H_{k+1} = e_j\right). \end{aligned}$$

D'après (i) de la proposition 1, on a l'indépendance du processus $O_{1:t}$ conditionnellement aux états latents. Par ailleurs, on rappelle les formules suivantes :

$$\beta_k(j) = \mathbb{P}_\theta(O_{k+1:t} = o_{k+1:t} \mid H_k = h_j)$$

$$\alpha_k(j) = \mathbb{P}_\theta(O_{1:k} = o_{1:k}, H_k = e_j)$$

Alors, on peut écrire :

$$\begin{aligned}\mathbb{P}_\theta\left(H_k = e_i, H_{k+1} = e_j, O_{1:t} = o_{1:t}\right) &= \mathbb{P}_\theta\left(O_{1:k} = o_{1:k} \mid H_k = e_i, H_{k+1} = e_j\right) \mathbb{P}_\theta\left(O_{k+1:t} = o_{k+1:t} \mid H_k = e_i, H_{k+1} = e_j\right) \\ &\quad \times\ \mathbb{P}_\theta\left(H_{k+1} = e_j \mid H_k = e_i\right) \mathbb{P}_\theta\left(H_k = e_i\right) \\ &= \alpha_k(i) \times a_{ij} \times \mathbb{P}_\theta\left(O_{k+1} = o_{k+1}, o_{k+2:t} \mid H_{k+1} = e_j, H_k = e_i\right) \\ &= \alpha_k(i) a_{ij} b_j\left(o_{k+1}\right) \beta_{k+1}(j).\end{aligned}$$

On isole l'observation $k+1$ pour faire apparaître les membres voulus. On remarque le fait que l'état latent au temps $k+1$ ne dépend que de l'observation au temps $k+1$. On obtient donc :

$$\begin{aligned}\mathbb{P}_\theta\left(H_k = e_i, H_{k+1} = e_j, O\right) &= \alpha_k(i) \times a_{ij} \times \mathbb{P}_\theta\left(o_{k+1}, o_{k+2:t} \mid H_{k+1} = e_j, H_k = e_i\right) \\ &= \alpha_k(i) \times a_{ij} \times \mathbb{P}_\theta\left(o_{k+1} \mid H_{k+1} = e_j\right) \mathbb{P}_\theta\left(o_{k+2:t} \mid H_{k+1} = e_j\right) \\ &= \alpha_k(i) a_{ij} b_j\left(O_{k+1}\right) \beta_{k+1}(j).\end{aligned}$$

Finalement, on obtient :

$$\xi_k(i, j; \theta) = \frac{\mathbb{P}_\theta\left(H_k = e_i, H_{k+1} = e_j, O_{1:t} = o_{1:t}\right)}{\mathbb{P}_\theta\left(O_{1:t} = o_{1:t}\right)} = \frac{\alpha_k(i)\ a_{ij}\ b_j(O_{k+1})\ \beta_k(j)}{\mathbb{P}_\theta\left(O_{1:t} = o_{1:t}\right)}.$$

□

Définition 6 : Soit un paramètre $\theta \in \Theta$. On note $\gamma_k(.;\theta)$ la loi de l'état latent au temps k sachant le processus $O_{1:t}$ et la paramètre θ. Ainsi, pour tout état i, $\gamma_k(i;\theta)$ correspond à la densité que l'état e_i ait émis o_k la k$^{\text{ième}}$ observation. On écrit :

$$\gamma_k(i;\theta) = \kappa_\theta\left(e_i \mid o_{1:t}\right)$$

Il est ici encore possible de réécrire $\gamma_k(i)$ à l'aide des fonctions *Forward/Backward*.

Proposition 7 : Pour tout i $\in E$

$$\gamma_k(i;\theta) = \sum_{j=1}^{n} \frac{\alpha_k(i)\ a_{ij}\ w_\theta(o_{k+1} \mid h_j)\ \beta_{k+1}(j)}{f_\theta\left(o_{1:t}\right)} = \frac{\alpha_k(i)\beta_k(i)}{f_\theta\left(o_{1:t}\right)}.$$

Preuve (Cas discret) : En utilisant la formule des probabilités totales sur l'événement $\{H_{k+1} = e_j\}$, on peut écrire :

$$\begin{aligned}\gamma_k(i;\theta) &= \sum_{j=1}^{n} \mathbb{P}_\theta(H_k = e_i, H_{k+1} = e_j \mid O_{1:t} = o_{1:t}) \\ &= \sum_{j=1}^{n} \xi_k(i, j; \theta) \\ &= \sum_{j=1}^{n} \frac{\alpha_k(i)\ a_{ij}\ b_j(o_{k+1})\ \beta_{k+1}(j)}{\mathbb{P}_\theta(O_{1:t} = o_{1:t})}.\end{aligned}$$

En remarquant que $a_{ij}\ b_j(o_{k+1})\ \beta_{k+1}(j) = \mathbb{P}_\theta(O_{k+1:t} = o_{k+1:t}, H_{k+1} = e_j \mid H_k = e_i)$, on peut écrire :

$$\begin{aligned}
\sum_{j=1}^{n} \alpha_k(i)\ a_{ij}\ b_j(o_{k+1})\ \beta_{k+1}(j) &= \alpha_k(i)\ \sum_{j=1}^{n} a_{ij}\ b_j(o_{k+1})\ \beta_{k+1}(j) \\
&= \alpha_k(i)\ \sum_{j=1}^{n} \mathbb{P}_\theta(O_{k+1:t} = o_{k+1:t}, H_{k+1} = e_j \mid H_k = e_i) \\
&= \alpha_k(i)\mathbb{P}_\theta(O_{k+1:t} = o_{k+1:t} \mid H_k = e_i) \\
&= \alpha_k(i)\beta_k(i).
\end{aligned}$$

On a finalement la relation :

$$\gamma_k(i;\theta) = \sum_{j=1}^{n} \frac{\alpha_k(i)\ a_{ij}\ b_j(o_{k+1})\ \beta_{k+1}(j)}{\mathbb{P}_\theta(O_{1:t} = o_{1:t})} = \frac{\alpha_k(i)\beta_k(i)}{\mathbb{P}_\theta(O_{1:t} = o_{1:t})}.$$

□

Après avoir introduit ces différentes quantités, nous pouvons maintenant énoncer le théorème qui nous permet d'expliciter la solution au problème de maximisation (étape (M)).

Théorème 1 (Formules de réestimation de Baum Welch) : Dans le cadre des HMM, la maximisation de $Q(\theta, \theta')$ conduit aux formules de réestimation des paramètres du modèle suivantes :

(i) Cas discret :

$$\bar{\mu}_i = \frac{\gamma_1(i;\theta')}{\sum_{j\in E}\gamma_1(j;\theta')} \qquad \bar{a}_{ij} = \frac{\sum_{k=1}^{t}\xi_k(i,j;\theta')}{\sum_{k=1}^{t}\gamma_k(i;\theta')} \qquad \bar{b}_j(v) = \frac{\sum_{k=1}^{t}1_{(o_k=v)}\gamma_k(j;\theta')}{\sum_{k=1}^{t}\gamma_k(j;\theta')}$$

(ii) Cas continu gaussien :

$$\bar{\mu}_i = \frac{\gamma_1(i;\theta')}{\sum_{j\in E}\gamma_1(j;\theta')} \qquad \bar{a}_{ij} = \frac{\sum_{k=1}^{t}\xi_k(i,j;\theta')}{\sum_{k=1}^{t}\gamma_k(i;\theta')} \qquad \bar{m}_i = \frac{\sum_{k=1}^{t}\gamma_k(i;\theta')\ o_k}{\sum_{k=1}^{t}\ \gamma_k(i;\theta')}, \quad i \in \{1,2,3\}$$

Preuve : (i) Cas discret : Soient deux paramètres θ et θ' d'un modèle de HMM où θ' est antérieur à θ. En utilisant les définition 5 et 6, on en déduit :

$$Q = \sum_{e_i\in E}\gamma_t(i;\theta')\log(\mu_i) + \sum_{i,j\in E}\sum_{k=1}^{t}\xi_k(i,j;\theta')\log(a_{ij}) + \sum_{e_i\in E}\sum_{v\in V}\sum_{k=1}^{t}\mathbb{I}_{(o_k=v)}\ \gamma_k(i;\theta')\log(b_i(v)) + \mathrm{C}$$

Par ailleurs, la maximisation par rapport aux paramètres (A, B, μ) du HMM se fait sous les contraintes d'égalités suivantes :

$$.\sum_{i\in E}\mu_i = 1 \quad (1) \qquad .\sum_{i\in E}a_{ij} = 1 \quad (2) \qquad .\sum_{l\in V}b_i(v) = 1 \quad \forall i \in E \quad (3)$$

Nous devons résoudre le problème :

$$\underset{(A,B,\mu)\in K}{\operatorname{argmax}} \; Q(A,B,\mu)$$

où

$$K = \left\{ \sum_{j\in E} \mu_j = 1, \quad \forall i \in E, \quad \sum_{j\in E} a_{ij} = 1, \quad \sum_{v\in V} b_i(v) = 1 \right\}$$

On utilise ensuite le multiplicateur de Lagrange pour résoudre le problème sous contraintes. Rappelons que argmax $f(x) =$ argmin $\{-f(x)\}$. Notre problème de maximisation d'une fonction concave devient donc un problème de minimisation d'une fonction convexe. On compte bien trois contraintes d'égalités. Celle-ci sont qualifiées car chacune des contraintes est une combinaison linéaire des variables (A, B, μ). De plus, Q est une fonction convexe et C^1 comme somme et composée de fonctions linéaires C^1. Ainsi l'unique solution au problème vérifie :

Il existe (x, y, z) trois réels tels que pour tout $i \in E$,

$$\nabla Q(\mu, A, B) + \sum_{j\in E} x\nabla\left(\mu_j - 1\right) \sum_{j\in E} y\nabla\left(a_{ij} - 1\right) + \sum_{v\in V} z\nabla\left(b_i\left(v\right) - 1\right) = 0 \tag{3.1.2}$$

Donc l'équation (3.1.2) se réecrit : $\forall\left(i, j, o\right) \in E^2 \times V$,

$$\begin{cases} x + \frac{\gamma_1(i;\theta')}{\mu_i} = 0, \\ \sum_{k=1}^{t} \frac{\xi_k\left(i,j;\theta'\right)}{a_{ij}} + y = 0 \\ \sum_{k=1}^{t} 1_{(O_k=v)} \frac{\gamma_k(i;\theta')}{b_i(v)} + z = 0 \\ \sum_{j\in E} \mu_j - 1 = 0 \\ \sum_{j\in E} a_{ij} - 1 = 0 \\ \sum_{v\in V} b_i(v) - 1 = 0 \end{cases}$$

On remarque que ce problème peut se résoudre en 3 problèmes indépendants. Dans un premier temps, pour la résolution de μ, on regarde le problème :

$$\begin{cases} \mu_i = -\frac{\gamma_1(i;\theta')}{x} \quad i \in E \\ \sum_{j\in E} \mu_j = 1 \end{cases}$$

Ce qui implique alors $x = -\sum_{j\in E} \gamma_1(j;\theta')$ et ainsi en remplaçant x par sa formule on trouve que $\mu_i = \frac{\gamma_1(i;\theta')}{\sum_{j\in E} \gamma_1(j;\theta')}$ pour $i \in E$.

Dans un second temps, on résout l'équation pour le paramètre a_{ij} :

$$\begin{cases} \sum_{k=1}^{t} \frac{\xi_k\left(i,j;\theta'\right)}{a_{ij}} + y = 0 \quad (i,j) \in E^2 \\ \sum_{j\in E} a_{ij} = 1 \quad i \in E \end{cases}$$

En combinant les deux équations précédantes, on déduit $y = -\sum_{j\in E} \sum_{k=1}^{t} \xi_k\left(i, j; \theta'\right)$ et donc que :

$$a_{ij} = \frac{\sum_{k=1}^{t} \xi_k\left(i, j; \theta'\right)}{\sum_{j\in E} \sum_{k=1}^{t} \xi_k\left(i, j; \theta'\right)} = \frac{\sum_{k=1}^{t} \xi_k\left(i, j; \theta'\right)}{\sum_{k=1}^{t} \gamma_k\left(i; \theta'\right)}$$

Enfin, pour le paramètre $b_j(v)$:

$$\begin{cases} \sum_{k=1}^{t} 1_{(O_k=v)} \frac{\gamma_k(i;\theta')}{b_i(v)} + z = 0 \\ \sum_{v \in V} b_i(v) = 1 \quad i \in E \end{cases}$$

On trouve le résultat suivant :

$$b_i(v) = \frac{\sum_{k=1}^{t} 1_{(O_k=v)} \gamma_k(j;\theta')}{\sum_{k=1}^{t} \gamma_k(j;\theta')}$$

Ce qui conclut la preuve.

(ii) Cas continu gaussien : On doit résoudre le problème :

$$\underset{(A,B,\mu)\in K}{\operatorname{argmax}} \; Q(A,B,\mu)$$

On réécrit Q en adaptant les formules au cas gaussien :

$$Q = \sum_{e_i \in E} \gamma_t(i;\theta') \log(\mu_i) + \sum_{i,j \in E} \sum_{k=1}^{t} \xi_k(i,j;\theta') \log(a_{ij}) + \sum_{e_i \in E} \sum_{k=1}^{t} \gamma_k(i;\theta') \log(\phi(o_k \mid m_k, \sigma^2)) + \mathrm{C}$$

Dans cette preuve, on s'occupera de maximiser uniquement le troisième terme de Q car les autres ont déjà été traités dans la preuve du cas discret. Explicitons ce troisième terme :

$$\sum_{e_i \in E} \sum_{k=1}^{t} \gamma_k(i;\theta') \log(\phi(o_k \mid m_k, \sigma^2)) = \sum_{e_i \in E} \sum_{k=1}^{t} \gamma_k(i;\theta') \frac{1}{2\pi} \exp\left(-\frac{(o_k - m_i)^2}{2\pi\sigma^2}\right)$$

On doit maximiser cette expression par rapport à m_i pour tout i dans E. L'expression étant continue et convexe, il nous suffit d'égaliser le gradient à 0 et de résoudre l'équation. On a donc pour tout i appartenant à $[\![1; n]\!]$,

$$\sum_{k=1}^{t} \gamma_k(i;\theta') \frac{o_k - m_i}{\pi\sigma^2} \exp\left(-\frac{(o_k - m_i)^2}{2\pi\sigma^2}\right) = 0$$

Puisque σ^2 est strictement positif on en déduit :

$$\sum_{k=1}^{t} \gamma_k(i;\theta')(m_i - o_k) = 0 \iff m_i = \frac{\sum_{k=1}^{t} \gamma_k(i;\theta') \; o_k}{\sum_{k=1}^{t} \; \gamma_k(i;\theta')}$$

□

3.1.3 Exemples d'application

Nous allons maintenant appliquer les résultats démontrés sur des exemples concrets.

Exemple 1 : Nous allons ici vérifier la validité de l'algorithme EM en montrant qu'au fur et à mesure des itérations, la probabilité d'observation d'une séquence augmente. On considère un modèle de HMM à trois états et on observe des éléments $V = \{a, b\}$. On suppose que le HMM est paramétré par θ_0 que l'on définit de la façon suivante :

$$A_0 = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.3 & 0.6 \end{pmatrix} \quad B_0 = \begin{pmatrix} 1 & 0 \\ 0.5 & 0.5 \\ 0 & 1 \end{pmatrix} \quad \mu_0 = \begin{pmatrix} 0.2 \\ 0.6 \\ 0.2 \end{pmatrix}.$$

Avec ces paramètres initiaux on obtient $\mathbb{P}_\theta\,(\text{abba}) = 0.045$. Après plusieurs itérations de l'algorithme, on trouve les paramètres optimaux qui sont :

$$A_{10} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad B_{10} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mu_{10} = \begin{pmatrix} 0.6 \\ 0.4 \\ 0 \end{pmatrix}.$$

Dans ce cas, la probabilité d'observation de la séquence est maximale, c'est-à-dire $\mathbb{P}_{\theta_{10}}\,(\text{abba}) = 1$.

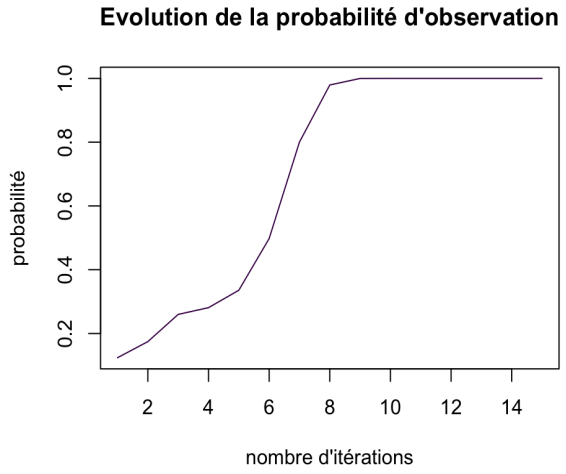

FIGURE 2 –

On remarque ici que l'algorithme converge à partir de 10 itérations pour une précision $\epsilon = 1.0 \times 10^{-7}$.

Exemple 2 : On se place maintenant dans le cadre d'une loi d'émission gaussienne. Dans l'exemple précédent, on a montré que la probabilité d'observation d'une séquence augmentait au fur et à mesure des itérations, ce qui nous garantissait la validité des formules de réestimation. Pour mettre en évidence la pertinence de l'algorithme EM dans le cas continu, nous allons montrer que la log-vraisemblance complète augmente au cours des itérations :

On suppose dans un premier temps que l'on observe une séquence générée par le modèle de HMM gaussien suivant où le paramètre de variance de la gausienne est fixé à $\sigma = 0.5$:

$$A = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.3 & 0.6 \end{pmatrix} \quad B = \begin{pmatrix} -2 \\ 0 \\ 2 \end{pmatrix} \quad \mu = \begin{pmatrix} 0.6 \\ 0.4 \\ 0 \end{pmatrix}.$$

Pour 30 itérations de l'algorithme EM, on obtient les paramètres optimaux suivants pour $\epsilon = 1.0 \times 10^{-7}$:

$$A_{30} = \begin{pmatrix} 0.65 & 0.25 & 0.05 \\ 0.07 & 0.82 & 0.11 \\ 0.08 & 0.24 & 0.67 \end{pmatrix} \quad B_{30} = \begin{pmatrix} -1.95 \\ -0.04 \\ 1.96 \end{pmatrix} \quad \mu_{30} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

On retrouve bien les paramètres du modèle qui ont permis de générer la séquence observée. Pour se convaincre de l'efficacité de l'algorithme, on peut tracer l'évolution des moyennes des lois gaussienne au cours des itérations ainsi que celle de la log-vraissemblance complète.

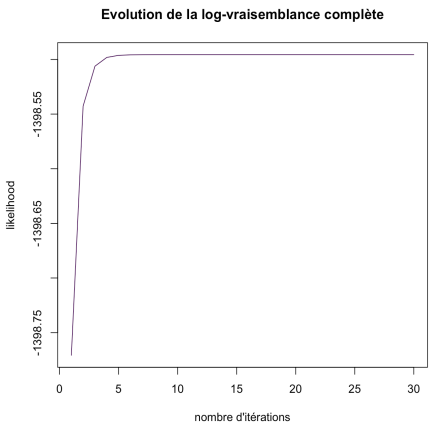

FIGURE 3 – Évolution de la log-vraissemblance complète.

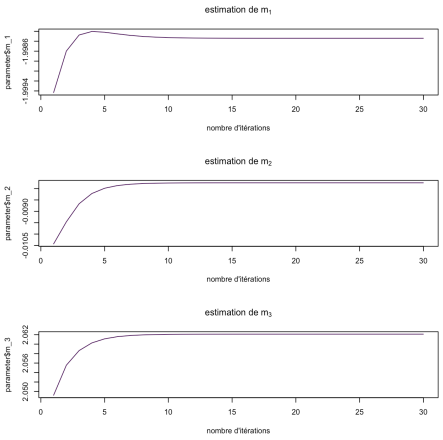

FIGURE 4 – Evolution de l'estimation des paramètres de moyenne au cours des itérations.

3.1.4 Construction d'intervalles de confiance pour les paramètres du HMM

On a vu dans les parties précédentes que l'algorithme EM fournissait un estimateur du maximum de vraisemblance pour les paramètres du HMM. On souhaiterait maintenant pouvoir lui associer une mesure d'incertitude. Dans cette partie, on s'intéresse donc à la construction de plusieurs intervalles de confiance pour les différents paramètres estimés du HMM, à savoir $\theta = (A, B, \mu)$. Nous allons étudier deux méthodes pour y parvenir, la première utilise les propriétés asymptotiques du maximum de vraisemblance et la deuxième repose sur la mise en place d'une procédure bootstrap.

Intervalle de confiance asymptotique

On commence par énoncer quelques rappels concernant l'estimateur du maximum de vraisemblance.

Définition 7 : Soit $L(\theta \mid x_1, \ldots, x_n)$ la fonction de vraisemblance associée à l'échantillon $(x_1, \ldots, x_n)$. Si $\log(L(\theta \mid x_1, \ldots, x_n))$ est deux fois dérivable par rapport à θ, on définit la matrice d'information de Fisher par :

$$J(\theta) = J(\theta_i; \theta_j) = \mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log(L(\theta \mid x_1, \ldots, x_n))\right].$$

Théorème 2 (Théorème fondamentale de la statistique) : Si l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ est solution de $\nabla \log(L(\theta \mid x_1, \ldots, x_n)) = 0$ ($\hat{\theta}_n$ est un M-estimateur) et qu'il vérifie pour tout $x = (x_1, \ldots, x_n)$,

(i) $\theta \mapsto \log(L(\theta \mid x)$ est partout deux fois dérivable.

(ii) $\theta \mapsto \log(L(\theta \mid x)$ est deux fois dérivable par rapport à θ sous le signe somme.

(iii) La fonction $\theta \mapsto \frac{\partial^2 \log(L(\theta|x)}{\partial \theta^2}$ est continue.

(iv) Pour tout $\theta \in \Theta$ l'information de Fisher $J(\theta)$ est finie.

alors on a :

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, J(\theta)^{-1}),$$

où $J(\theta)$ correspond à la matrice d'information de Fisher.

On peut ainsi en déduire une propriété asymptotique pour chacun des paramètres.

Proposition 8 : Si on note $\hat{j}_k^2 = \left[J\left(\hat{\theta}\right)^{-1}\right]_{kk}$, alors

$$\sqrt{n}\,\frac{\hat{\theta}_k - \theta_k}{\hat{j}_k^2} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Preuve : D'après le théorème précédent, on a :

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, J(\theta)^{-1})$$

Donc on en déduit :

$$J(\theta)^{\frac{1}{2}} \sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, I_d)$$

De plus, on a par le théorème de continuité $J\left(\hat{\theta}\right) \xrightarrow[n\to+\infty]{\mathbb{P}} J(\theta)$. Donc d'après le théorème de Slustky, on obtient :

$$J\left(\hat{\theta}_n\right)^{\frac{1}{2}} \sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, I_d)$$

En prenant $\hat{j}_k^2 = \left[J\left(\hat{\theta}\right)^{-1}\right]_{kk}$, on a bien

$$\sqrt{n}\,\frac{\hat{\theta}_k - \theta_k}{\hat{j}_k^2} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

□

On peut ainsi construire un intervalle de confiance pour les différents paramètres de notre HMM. Si l'on considère un vecteur de paramètres :

$$\theta = (a_{11}, \ldots, b_{11}, \ldots, \mu_1, \ldots, \mu_n)^T$$

et la matrice d'information de Fisher correspondante :

$$J(\theta) = J(\theta_i;\theta_j) = \mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log\left(L^c\left(\theta \mid o_1,\ldots,o_t,h_1,\ldots h_t\right)\right)\right]$$

Si l'on souhaite construire un intervalle de confiance pour a_{ij}, le théorème 3 ainsi que la proposition précédente nous assure que :

$$\sqrt{n}\,\frac{(\hat{a_{ij}} - a_{ij})}{\sqrt{J^{-1}_{a_{ij}}}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Avec $J^{-1}_{a_{ij}}$ l'élément diagonal de J^{-1} correspondant au paramètre a_{ij}.

Donc, l'intervalle de confiance de nieau $1-\alpha$ est donné par :

$$IC(1-\alpha) = \left[\hat{a_{ij}} - \frac{q_{1-\frac{\alpha}{2}}}{\sqrt{nJ_{a_{ij}}}}, \hat{a_{ij}} + \frac{q_{1-\frac{\alpha}{2}}}{\sqrt{nJ_{a_{ij}}}}\right]$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$.

Intervalle de confiance via procédure Bootstrap paramétrique

Une autre méthode pour construire des intervalles de confiance pour les paramètres $\theta = (A, B, \mu)$ du HMM est d'utiliser une procédure bootstrap paramétrique (Visser et al. [2000]) qui se décompose en plusieurs étapes.

On suppose dans un premier temps que l'on observe une séquence $o_{1:t} = (o_1, ..., o_t)$ issue d'un HMM dont on ignore les paramètres $\theta = (A, B, \mu)$. Ce dernier peut être alors estimé à l'aide de l'algorithme EM et des formules de Baum-Welch décrites précédemment. On obtient alors une estimation des paramètres que l'on notera $\hat{\theta} = (\hat{A}, \hat{B}, \hat{\mu})$. À partir du paramètre $\hat{\theta}$ on peut simuler K séquences de taille t qui nous serviront d'échantillon Bootstrap. On a alors K séquences notées $o^{(1)}_{1:t}, ..., o^{(K)}_{1:t}$. Pour chaque séquence, on estime à nouveau les paramètres à l'aide de l'algorithme EM. On a donc K estimations de θ que l'on note $\hat{\theta}_1, ..., \hat{\theta}_K$. On peut alors construire un intervalle de confiance pour chacun des paramètres A, B, μ.

Supposons que l'on veuille construire un intervalle de confiance de niveau $1-\alpha$ pour un coefficient a_{ij} de la matrice A. D'après la procédure Bootstrap, on dispose d'un échantillon $a_{ij} = (\hat{a}^*_{ij} - a^1_{ij}, ..., \hat{a}^*_{ij} - a^K_{ij})$. Dès lors, on peut utiliser les quantiles empiriques de notre échantillon bootstrap. Si l'on note $\hat{F}_n$ la fonction de répartition empirique de notre échantillon et $\hat{F}^{(-1)}_n$ son inverse généralisé, alors on a :

$$\mathrm{IC}(1-\alpha) = \left[\hat{F}^{-1}_n\left(\frac{\alpha}{2}\right); \hat{F}^{-1}_n\left(1-\frac{\alpha}{2}\right)\right]$$

En pratique, il nous suffit donc de ranger les valeurs de notre échantillon dans l'ordre croissant et prendre comme borne de l'intervalle la $\left(\left\lfloor\frac{\alpha K}{2}\right\rfloor + 1\right)^{\text{ième}}$ et la $\left(\left\lfloor\left(1-\frac{\alpha}{2}\right)K\right\rfloor\right)^{\text{ième}}$ valeur de l'échantillon. Ainsi un intervalle de confiance de niveau $1-\alpha$ est donné par :

$$\mathrm{IC}(1-\alpha) = \left[\hat{a}^{\left(\left\lfloor\frac{\alpha K}{2}\right\rfloor+1\right)}_{ij}; \hat{a}^{\left(\left\lfloor\left(1-\frac{\alpha}{2}\right)K\right\rfloor\right)}_{ij}\right]$$

On peut procéder de même pour mesurer l'incertitude des coefficients de B et de μ.

Exemple de construction d'intervalle de confiance

On va ici mettre en pratique la construction d'intervalle de confiance par la méthode Bootstrap paramétrique. On se place dans le cadre de l'exemple 2 de la partie 3.1.3 (loi d'émission gaussienne). Pour le paramètre a_{22}, on obtient l'intervalle de confiance $IC(0.95) = [0.79, 0.85]$, alors que la valeur réelle du paramètre est 0.8. De même, pour le paramètre m_3, on obtient l'intervalle de confiance $IC(0.95) = [1.87, 2.04]$, alors que la valeur réelle du paramètre est 2. Les histogrammes des échantillons bootstrap sont données ci-dessous.

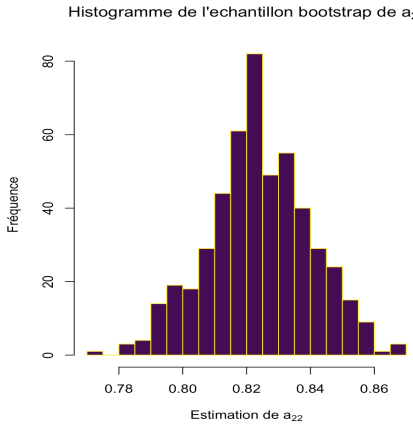

FIGURE 5 –

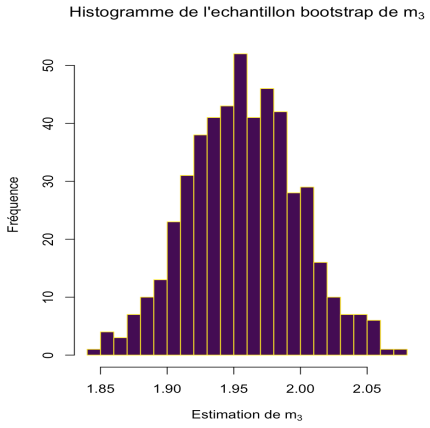

FIGURE 6 –

3.2 Approche bayésienne et algorithme MCMC

Dans cette partie, nous allons à présent nous placer d'un point de vue bayésien. À la différence du point de vue fréquentiste, l'étude bayésienne permet de combiner l'information que l'on a sur les données avec la connaissance *a priori* que l'on a sur les paramètres du modèle. Dans cette partie, nous présenterons dans un premier temps les principes de la modélisation Bayesienne puis on s'intéressera à l'algorithme de Gibbs, un cas particulier de l'algorithme de Metropolis-Hasting, une méthode MCMC (Chib and Greenberg [1995]). Enfin, nous présenterons un exemple d'application de l'algorithme de Gibbs dans un modèle de HMM.

3.2.1 Formalisme bayésien

Le principe de la statistique bayésienne est de probabiliser l'espace des paramètres. Désormais, on considère un modèle statistique paramétré par un vecteur aléatoire $\theta \in \Theta$. On fait le choix d'associer à ce vecteur une loi de probabilité $\pi(\theta)$ qui admet une densité dans $\mathbb{R}^d$ absolument continue par rapport à la mesure de Lebesgues. On appelle cette loi, la loi *a priori* (*prior*) de θ. Notons que ce choix est arbitraire et correspond à l'incertitude que l'on a *a priori* sur le paramètre θ.

L'objectif par la suite est de construire une loi de probabilité pour θ qui tient compte des observations. Nous introduisons alors la loi *a posteriori* (*posterior*), $\pi(\theta \mid o_{1:t})$. De plus, usuellement, on va noter $L(\theta \mid o_{1:t})$ la vraisemblance des observations et de θ. D'après le théorème de Bayes, la loi a posteriori est proportionnel au produit de la loi a priori et de la vraisemblance.

$$\pi(\theta \mid o_{1:t}) \propto \pi(\theta)\ L(\theta \mid o_{1:t})$$

Par normalisation, on peut déterminer le coefficient de proportionnalité.

$$\pi(\theta \mid o_{1:t}) = \frac{\pi(\theta)\ L(\theta \mid o_{1:t})}{\int_{\mathbb{R}^d} \pi(\theta)\ L(\theta \mid o_{1:t}) d\theta} \tag{3.2.1}$$

Dans le cadre bayésien, ce que l'on cherche à déterminer est la loi *a posteriori*. D'une part, à l'instar de l'estimateur du maximum de vraisemblance (EMV) dans le cadre fréquentiste, dans le cadre bayésien, on peut chercher à déterminer le maximum a posteriori (MAP). Il est obtenu en maximisant la densité de la loi a posteriori sur θ.

$$\hat{\theta}_{MAP} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \pi(\theta \mid o_{1:t})$$

Cette recherche du MAP est motivé par le résultat du Théorème de Bernstein - Von Mises qui implique que les estimateurs MAP et EMV ont des propriétés similaires sur de larges échantillons.

Théorème 3 (Bernstein - Von Mises) : Sous certaines hypothèses de régularité et pour un échantillon suffisamment large, la loi a posteriori de θ est "asymptomatiquement proche" d'une loi normale de moyenne la valeur exacte de θ et de matrice de covariance égale à l'inverse d'une matrice d'information de Fisher.

D'autre part, une quantité d'intérêt est le calcul d'espérance selon la loi *a posteriori* de la forme

$$\mathbb{E}_\pi[g] = \int g(\theta)\pi(\theta \mid o_{1:t})\ d\theta.$$

Cependant, même si l'on a souvent une forme explicite de la loi *a priori* et de la vraisemblance, déterminer la loi *a posteriori* peut s'avérer en pratique compliqué à cause de la constante multiplicative, difficile voir impossible à calculer. Pour contourner ce problème computationel, nous allons utiliser une méthodes de Monte-Carlo par chaînes de Markov (MCMC).

3.2.2 Méthodes MCMC

Les méthodes MCMC sont introduite dans les années 1950 par Metropolis (entre autres) puis dans les années 1990, l'échantillonnage de Gibbs est introduit par les frères Geman pour les appliquer à la restauration bayésienne d'images. Dans cette partie,nous allons montrer comment, à l'aide des méthode MCMC, générer une chaîne de Markov $\left(\theta^{(t)}\right)_{t\geq 0}$ dont la mesure stationnaire est une densité cible f que nous cherchons à approximer.

La structure Markovienne de $\left(\theta^{(t)}\right)_{t\geq 0}$ impose une dépendance entre les variables. Autrement dit, on ne peux pas appliquer la Lois des Grands Nombres pour trouver la loi limite de la chaîne étudiée.
Dans un espace d'état discret, nous savons que si la chaîne $(\theta^{(t)})_{t\geq 0}$ est irréductible, récurrente positive et que si f est l'unique mesure de probabilité invariante associée à la chaîne, alors le Théorème Ergodique nous dit que

$$\forall g \in L^1(f),\ (\text{i.e } \mathbb{E}_f[|g|] < \infty), \qquad \frac{1}{N}\sum_{i=1}^{N} g\left(\theta^{(i)}\right) \xrightarrow[N\to+\infty]{\text{p.s}} \mathbb{E}_f[g] = \int g(\theta)f(\theta)d\theta.$$

Autrement dit, la mesure empirique associée à la chaîne $(\theta^{(t)})_{t\geq 0}$ converge vers la densité cible f presque sûrement, et ce quelque soit sa distribution initiale.

Dans notre cas, la densité cible est la loi *a posteriori* $\pi(\ \cdot \mid o_{1:t})$. Et d'après la structure de notre modèle $\mathcal{M}$, les paramètres que nous allons simuler sont distribués selon des lois continues. Donc il est clair que la chaîne $\left(\theta^{(t)}\right)$ que l'on cherche à construire sera à espace d'états continus. Pour palier ce problème, nous allons nous appuyer sur l'article de Chib and Greenberg [1995]. Nous admettrons certaines propriétés

qui y sont présentées. De plus, on ne rentrera pas dans les détails des conditions à vérifier pour obtenir plusieurs de nos résultats, lesquelles peuvent être trouvés en annexe de Smith and Roberts [1993]. Il utilise en particulier la caractérisation d'une Chaîne de Markov par son noyau de transition que nous allons introduire.

Définition 8 (Noyau de transition d'une Chaîne de Markov) : Soit $\left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^{\mathrm{d}})\right)$ et $\left(A, \mathcal{B}(\mathbb{R}^{\mathrm{d}})\right)$ deux espaces mesurables. La fonction : $P : \mathcal{B}(\mathbb{R}^{\mathrm{d}}) \times \mathcal{B}(\mathbb{R}^{\mathrm{d}}) \longrightarrow [\prime, +\infty]$ est appelée noyau de transition de $\mathbb{R}^d$ à A si :

(i) pour tout A appartenant à $\mathcal{B}(\mathbb{R}^{\mathrm{d}})$, $\quad \theta \longmapsto P(\theta, A)$ est mesurable.

(ii) pour x appartenant à $\mathbb{R}^d$, $\quad A \longmapsto P(\theta, A)$ est une mesure de probabilité.

Le noyau de transition étant une caractérisation de sa chaîne de Markov, nous allons utiliser un théorème qui permet de définir la mesure invariante à partir du noyaux de transition.

Proposition 9 : De plus, sous certaines conditions (Chib and Greenberg [1995]), il existe une densité stationnaire $f^*(.)$ telle que :

$$f^*(d\theta') = \int_{\mathbb{R}^d} P(\theta, d\theta') f(\theta) d\theta'$$

où f la densité selon la mesure de Lebesgue de f^*, et $f^*(d\theta') = f(\theta')d\theta$.

Dans le cas ou f^* est absolument continue par rapport à la mesure de Lebesgue, ce qu'on supposera ici, on a $f^* = f$. Nous allons à présent définir la N ème itération du noyaux de transition qui nous permettra de conclure quand à la convergence du noyau de la Chaîne de Markov.

Définition 9 (N ème itération du noyaux de transition) :

$$P^{(N)}(\theta, A) := \int_{R^d} P^{(N-1)}(\theta, d\theta') P(\theta', A) \quad \text{et} \quad P^{(1)}(\theta, d\theta') := P(\theta, d\theta')$$

La N ième itération du noyaux de transition est la loi de la chaîne de Markov après N unités de temps. Nous pouvons alors énoncer l'analogue du Théorème Ergodique dans le cas d'une chaîne de Markov à espace d'état continus.

Théorème 4 : Sous certaines conditions, $P^{(N)}(\theta, A)$ converge vers la densité stationnaire f^* lorsque N tend vers l'infini.

$$P^{(N)}(\theta, A) \xrightarrow[N \to +\infty]{} f^*(d\theta')$$

Ainsi, pour générer des vecteurs aléatoires de densité f, on utilise le noyau de transition $P(\theta, A)$ dont la N ième itération converge vers f. Cette simulation peut démarrer d'un $\theta^{(0)}$ quelconque. On peut alors affirmer que après un certains nombre d'itérations, les observations auront pour distribution f. De plus, puisque l'opérateur P caractérise la chaîne de Markov, il nous suffit de trouver un opérateur P approprié qui nous permette d'obtenir la mesure invariante souhaitée.

La proposition que nous allons énoncer va nous permettre de poser des conditions sur $P(\theta, d\theta')$ pour que notre densité cible $\pi(\,\cdot \mid o_{1:t})$ soit la mesure de probabilité invariante de $P(\theta, \cdot\,)$.

Proposition 10 : .
(i) On suppose que, pour une fonction $p(\theta, \theta')$, le noyau de transition s'écrit tel que :

$$P(\theta, d\theta') = p(\theta, \theta')d\theta' + r(\theta)\delta_\theta(d\theta')$$

tel que $p(\theta, \theta) = 0$, $\delta_\theta(d\theta')$ soit la mesure de Dirac en θ et $r(\theta)$ représente la probabilité que la chaîne reste dans l'état θ. Alors :

$$\int_{\mathbb{R}^d} p(\theta, \theta^{'})d\theta' = 1$$

(ii) Si de plus, la fonction $p(\theta, \theta')$ satisfait la condition dite d'inversibilité

$$\pi(\theta \mid o_{1:t})p(\theta, \theta') = \pi(\theta' \mid o_{1:t})p(\theta', \theta) \tag{3.2.2}$$

Alors $\pi(\ \cdot \mid o_{1:t})$ est la densité stationnaire de $P(\theta, \cdot\ )$

Preuve : Voir Chib and Greenberg [1995]

Cette propriété nous donne une condition suffisante que doit satisfaire p pour que $\pi(\ \cdot \mid o_{1:t})$ soit la densité stationnaire de $P(\theta, \cdot\ )$. Le but est donc désormais de déterminer une telle fonction p.

A présent, considérons la densité cible propre à $\mathcal{M}$, $\pi(\ \cdot \mid o)$ qui, d'après (3.2.1) s'écrit sous la forme :

$$\pi(\theta \mid o_{1:t}) = \frac{\pi(\theta)\ L(\theta \mid o_{1:t})}{K_\pi}$$

où K_π est la constante de normalisation. On peut donc écrire la densité cible comme le quotient d'une quantité que l'on sait calculer d'une quantité inconnu. On remarque de par cette structure de $\pi(\ \cdot \mid o_{1:t})$, on peut vérifier la condition d'inversibilité sans avoir à calculer K_π. Nous allons à présent montrer comment trouver, à partir de l'algorithme de Métropolis-Hasting, un fonction p satisfaisant la propriété énoncée.

3.2.3 Algorithme de Metropolis-Hasting

Comme dans une méthode d'acceptation-rejet, on suppose que l'on a une densité instrumentale que l'on notera ici $q(\theta, \ \cdot\ )$ qui peut générer des candidats et qui dépend de l'état courant θ de la chaîne de Markov. On suppose que $\int q(\theta, \theta')d\theta' = 1$. Dans notre cas, cela signifie que si le processus est à l'état θ, on génère une valeur θ' à partir de $q(\theta, \ \cdot\ )$.

Si q vérifie la condition d'inversibilité alors choisir p telle que $p(\theta, \theta') = q(\theta, \theta')$ convient.

Sinon, alors cela signifie que l'on est dans un cas où pour certains vecteur (θ, θ'), on a :

$$\pi(\theta \mid o)q(\theta, \theta') > \pi(\theta' \mid o)q(\theta', \theta)$$

On peut écrire cela sans perte de généralité quitte à inverser les rôles de θ et $\theta^{'}$. Cette inégalité s'interprète de la façon suivante : le processus se déplace de θ vers θ' trop souvent, et de θ' vers θ trop rarement. Pour éviter cela, et retrouver la condition d'inversibilité, on introduit une probabilité de déplacement $\alpha(\theta, \theta')$, que l'on définit dans ce cas tel que :

$$\alpha(\theta, \theta') := \begin{cases} \min\left\{\frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}, 1\right\} & \text{si} \quad \pi(\theta)q(\theta, \theta') > 0 \\ 1 \quad \text{sinon} \end{cases}$$

Dans ce cas si l'on définit $p_{MH}(\theta, \theta') := q(\theta, \theta')\alpha(\theta, \theta')$, alors p_{MH} vérifie la condition d'inversibilité. On peut enfin calculer le noyau de transition associé :

$$P_{MH}(\theta, d\theta') = p_{MH}(\theta, \theta')d\theta' + \left(1 - \int_{\mathbb{R}^d} p_{MH}(\theta, \theta')d\theta'\right)\delta_\theta(d\theta')$$

Ainsi, l'algorithme M-H est caractérisé par sa densité instrumentale $q(\theta, \theta')$. A partir de celle-ci, nous pouvons simuler (après beaucoup d'itérations) des paramètres qui sont distribués selon la densité cible. De plus, le calcul de $\alpha(\theta, \theta')$ ne nécessite pas de savoir calculer la constante multiplicative K_π de la densité cible car celle-ci s'annule dans l'expression de $\alpha\left(\theta, \theta'\right)$. Nous avons finalement tous les éléments nécessaires pour simuler selon la loi à posteriori.

Algorithm 3: Algorithme M-H

Input : $\theta^{(0)}$; N
for $j \in [\![1, N]\!]$ *:* **do**
  Generate θ' from $q(\theta^{(j)}, \cdot)$ and u from $\mathcal{U}(0, 1)$
  if $u \leq \alpha(\theta^{(j)}, \theta')$ **then**
    $\theta^{(j+1)} = \theta'$
  else
    $\theta^{(j+1)} = \theta^{(j)}$
  end
end
return $\left\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}\right\}$

.

Un des problème de cet algorithme est qu'il met beaucoup de temps à converger, du fait de la grande dimensions des paramètres. L'algorithme de Gibbs, permet en partie de palier ce problème de grande dimension et ainsi d'améliorer la vitesse de convergence de l'algorithme.

3.2.4 Algorithme de Gibbs

Parmis les méthodes de simulation MCMC, on retrouve l'échantillonnage de Gibbs qui est un cas particulier de Métropolis-Hastings. Dans cette partie nous allons décrire le fonctionnement cet algorithme et pourquoi l'étude des HMM se prête à ce type d'échantillonage.

Dans cette partie, on cherche toujours à estimer la densité cible $\pi(\theta \mid o)$ mais d'une façon plus simple qu'exprimé précédemment. Si l'on reprend modèle bayésien $(L(\theta \mid o), \pi(\theta))$ où $L(\theta \mid o)$ est la vraisemblance des observations et $\pi(\theta)$ la loi *a priori*. On peut réécrire la loi *a priori* de sorte à faire apparaître les lois conditionnelles des paramètres. Si l'on note d comme étant le nombre total de paramètres, on a alors :

$$\pi(\theta) = \pi_1(\theta \mid \theta_1)\, \pi_2(\theta_1 \mid \theta_2) \ldots \pi_d(\theta_{d-1} \mid \theta_d)$$

où $\theta = (\theta_1, \ldots, \theta_d)$. De plus on notera $\theta_{-k} = (\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d)$ qui représente le vecteur θ auquel on a enlever la k $^{\text{ième}}$ coordonée.

On suppose que l'on peut calculer explicitement $\pi(\theta'_k \mid \theta_{-k})$. Dans cette partie, nous noterons $q_G(\theta, \theta')$ la densité instrumentale que l'on va définir à partir de la fonction

$$q_G(\theta, \theta') := \begin{cases} \pi(\theta'_k \mid \theta_{-k}) \quad \text{si} \quad \theta'_{-k} = \theta_{-k} \quad \text{pour } k \in \{1, \ldots, d\} : \\ 0 \quad \text{sinon} \end{cases}$$

Proposition 11 : $q_G(\theta, \theta')$ vérifie la condition d'inversibilité (3.2.2) pour la densité cible $\pi(. \mid o_{1:t})$. Ainsi, le rapport $\alpha(\theta, \theta')$ est toujours égale à 1.

Preuve : Pour simplifier les notations, on notera lors de cette preuve $\hat{\pi}$ la densité cible $\pi(. \mid o_{1:t})$. Pour θ et θ' deux paramètres, on cherche à montrer que

$$\hat{\pi}(\theta) q_G(\theta, \theta') = \hat{\pi}(\theta') q_G(\theta', \theta)$$

Pour $k = 1, \ldots, d$,

Si θ'_{-k} est différent de θ_{-k}, alors par symétrie de l'égalité, on a que

$$\hat{\pi}(\theta) q_G(\theta, \theta') = \hat{\pi}(\theta') q_G(\theta', \theta) = 0$$

Sinon on a

$$\begin{cases} \hat{\pi}(\theta) q_G(\theta, \theta') = \hat{\pi}(\theta) \hat{\pi}(\theta'_k \mid \theta_{-k}) \\ \hat{\pi}(\theta') q_G(\theta', \theta) = \hat{\pi}(\theta') \hat{\pi}(\theta_k \mid \theta'_{-k}) \end{cases}$$

On peut décomposer

$$\hat{\pi}(\theta) = \hat{\pi}(\theta_k \mid \theta_{-k}) \ \hat{\pi}(\theta_{-k})$$

Et :

$$\hat{\pi}(\theta') = \hat{\pi}(\theta'_k \mid \theta'_{-k}) \ \hat{\pi}(\theta'_{-k})$$

Donc :

$$\begin{cases} \hat{\pi}(\theta) q_G(\theta, \theta') = \hat{\pi}(\theta_k \mid \theta_{-k}) \ \hat{\pi}(\theta_{-k}) \ \hat{\pi}(\theta'_k \mid \theta_{-k}) \\ \hat{\pi}(\theta') q_G(\theta', \theta) = \hat{\pi}(\theta'_k \mid \theta'_{-k}) \ \hat{\pi}(\theta'_{-k}) \ \hat{\pi}(\theta_k \mid \theta'_{-k}) \end{cases}$$

Or on sait que $\theta'_{-k} = \theta_{-k}$

Donc finalement

$$\begin{cases} \hat{\pi}(\theta) q_G(\theta, \theta') = \hat{\pi}(\theta_k \mid \theta_{-k}) \ \hat{\pi}(\theta_{-k}) \ \hat{\pi}(\theta'_k \mid \theta_{-k}) \\ \hat{\pi}(\theta') q_G(\theta', \theta) = \hat{\pi}(\theta'_k \mid \theta_{-k}) \ \hat{\pi}(\theta_{-k}) \ \hat{\pi}(\theta_k \mid \theta_{-k}) \end{cases}$$

Ce qui conclut la preuve.

□

Dès lors que l'on a réussit à trouver une densité cible qui vérifie la condition de réversibilité, on peut simuler selon la loi a posteriori. Ainsi on décrit l'algorithme de Gibbs, en reprenant l'algorithme de M-H avec la densité instrumentale q_G.

Algorithm 4: Algorithme de Gibbs

Input : choisir arbitrairement $\theta^{(0)}$ et N la longueur de l'échantillon simulée
for $j \in [\![1, N]\!]$ **do**
  Choisir aléatoirement k parmi $1, ..., d$
  Générer θ' selon la densité $q_G(\theta^{(j)}, .)$
  Poser $\theta^{(j+1)} = \theta'$
end
return $\left\{\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}\right\}$

.

Remarque : Si l'on choisit de poser sur nos paramètres des lois conditionnelles par rapport à d'autre paramètre, on réduit la dimension de Θ. Ce modèle se prête bien à notre étude car dans l'étude de Rydén [2008], le modèle des paramètre proposé est un modèle hiérarchique à 3 niveaux.

3.2.5 Application de l'algorithme de Gibbs

On va maintenant appliquer l'algorithme de Gibbs présenté précédemment dans une perspective d'inférence des paramètres d'un modèle de Markov caché avec une loi d'émission continue. On se place donc dans le cadre d'un HMM à trois états muni d'une loi d'émission gaussienne. Pour rappel, on cherche à estimer les paramètres du modèle introduit dans l'exemple 2 de la partie 3.1.3. On suppose ici que l'on observe une séquence $o_{1:t}$ généré par le modèle considéré.

Dans l'optique d'une approche bayésienne, il convient de choisir des lois a priori pour chacun des paramètres du modèle. Pour ces choix de lois *a priori*, on se basera sur l'article de Rydén [2008]. Ainsi, pour chaque ligne de la matrice de transition A, on suppose que $(a_{d1}, a_{d2}, a_{d3}) \sim \text{Dirichlet}(1,1,1)$ pour $d = 1,2,3$. Pour le vecteur de probabilité initiale μ, on a $(\mu_1, \mu_2, \mu_3) \sim \text{Dirichlet}(1,1,1)$. Pour chaque m_i on choisit la loi *a priori* $\mathcal{N}(u, w^{-1})$ avec $u = (\min o_k + \max o_k)/2$ et $w = 1/(\max o_k - \min o_k)^2$. Enfin, pour σ^{-2} on a $\sigma^{-2} \sim \mathcal{G}a(\alpha, \beta)$ où $\alpha = 2$ et $\beta \sim \mathcal{G}a(f, g)$ avec $f = 0.2$ et $g = 10/(\max o_k - \min o_k)^2$. On fait également l'hypothèse que tous ces paramètres sont *a priori* indépendants. Nous avons également initialisé les paramètres m_i tels que $\forall i \in [\![1,3]\!]$:

$$m_i = \min(o_k) + R/\left[2 \text{ Card}(E)\right] + (i-1)R/ \text{ Card}(E).$$

Avec R définit précédemment. On en déduit une séquence initiale $h_{1:t}$ telle que pour tout $k \in [\![1,t]\!]$, h_k est l'état qui minimise la quantité $(y_k - m_{h_k})^2$ avec m_{h_k} la moyenne de la loi normale associée à l'état courant h_k. Enfin, on estime σ^2 par la quantité $n^{-1} \sum_{k=1}^{n} (o_k - m_{h_k})^2$. Par la suite, on notera "..." pour signifier que l'on conditionne par rapport à tous les paramètres du modèle, la suite cachée $h_{1:t}$ et les observations $o_{1:t}$. Sous ces choix arbitraires de lois *a priori*, on obtient, après calcul (c.f Annexe 4.2), les lois conditionnelles suivantes :

$$(\mu_1, \mu_2, \mu_3) \mid ... \sim \text{ Dirichlet}(\mathbb{I}_{h_1=e_1} + 1, \mathbb{I}_{h_1=e_2} + 1, \mathbb{I}_{h_1=e_3} + 1),$$

$$(a_{d1}, a_{d2}, a_{d3}) \mid ... \sim \text{ Dirichlet}(n_{d1} + 1, n_{d2} + 1, n_{d3} + 1), \quad \text{pour } d = 1,2,3$$

où n_{ij} correspond au nombre de transition entre l'état e_i et l'état e_j dans la séquence $h_{1:t} = (h_1, ..., h_t)$.

$$m_i \mid ... \sim \mathcal{N}\left(\frac{S_i + wu\sigma^2}{n_i + w\sigma^2}, \frac{\sigma^2}{n_i + w\sigma^2}\right),$$

Avec S_i la somme des éléments observés o_k ayant été émis par l'état e_i et n_i le nombre de visite de l'état e_i dans la séquence $h_{1:t} = (h_1, ..., h_t)$.

$$\sigma^{-2} \mid ... \sim \mathcal{G}a\left(\alpha + \frac{t}{2}, \beta + \frac{1}{2}\sum_{k=1}^{t}(o_k - m_{h_k})^2\right)$$

$$\beta \mid ... \sim \mathcal{G}a\left(f + \alpha, g + \sigma^{-2}\right).$$

Concernant la loi conditionnelle de la chaîne de Markov, on obtient une chaîne de Markov non homogène définie par :

$$\mathbb{P}(H_1 = e_j \mid ...) \propto \mu_j \ \phi(o_1, m_j, \sigma^2) \ \beta_j(1)$$
$$\mathbb{P}(H_k = e_j \mid H_{k-1} = e_i) \propto a_{ij} \ \phi(o_k, m_j, \sigma^2) \ \beta_j(k).$$

où $\beta_k(j)$ à la fonction *backward*.

Ainsi, après implémentation de l'algorithme de Gibbs avec les lois définies précédemment, pour une observation de taille $t = 1000$ et 2000 itérations, on obtient pour m_2 le graphique suivant.

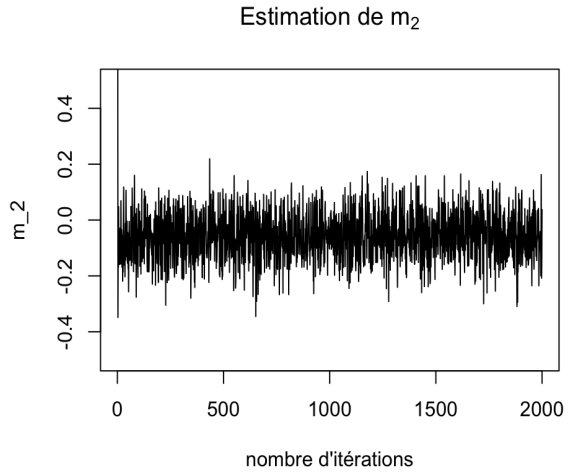

FIGURE 7 – Evolution du paramètre m_2 pour une séquence de longueur $t = 1000$ et 2000 itérations.

On observe une très faible période de chauffe (*burin-in*) et une estimation du paramètre m_2 qui se stabilise autour de la valeur réelle du paramètre, à savoir 0.

Remarque : Lorsque l'on considère une séquence d'observation longue, une des difficultés principales rencontrée est de pouvoir simuler les états d'une chaîne de Markov non-homogène avec des probabilités très petites ($\approx 1.0 \times 10^{-20}$). Pour palier à ce problème, il convient d'utiliser les log-probabilités et d'échantillonner grâce au *Gumbel Max Trick* faisant appel à la loi Gumbel.

4 Algorithme de Viterbi

Dans cette partie on se place dans le cadre d'une loi d'émission discrète et on s'intéresse maintenant à un autre problème lié au modèle de Markov caché qui fait suite à la partie précédente, celui du calcul du chemin optimal. En effet, dans une optique d'inférence complète et après avoir obtenu une estimation des paramètres du HMM dans la partie 3, on pourrait naturellement être tenté de déterminer la séquence sous-jacente $h_{1:t}$ qui maximise la probabilité d'apparition d'une séquence $o_{1:t}$.

On considère un modèle $\mathcal{M}$ de paramètre $\theta = (A, B, \mu)$ et une séquence d'observation $o_{1:t}$.

Définition 10 : On définit le chemin optimal comme la séquence d'états cachés $h^* = (h_1^*, ..., h_t^*)$ la plus probable ayant émis la séquence $o_{1:t}$. Le chemin optimal est donc la meilleure séquence d'états cachés $H_{1:t}$ qui maximise la probabilité d'observation de $o_{1:t}$:

$$h^* = (h_1^*, ..., h_t^*) = \underset{h_{1:t})}{\text{argmax}}\ \mathbb{P}_\theta(O_{1:t} = o_{1:t}, H_{1:t} = h_{1:t})$$

Pour trouver une telle séquence, nous allons utiliser l'algorithme de Viterbi (Forney [1973]) qui repose sur le principe de la programmation dynamique.

4.1 Description de l'algorithme

L'idée de l'algorithme est de maximiser à chaque instant $k \in [\![1, t]\!]$ la probabilité du chemin amenant à l'émission de la séquence $o_{1:k}$. Pour cela, on introduit la variable $v_k(i)$ qui correspond à la probabilité du chemin optimal amenant à l'état e_i à l'instant k, étant donnée la séquence $o_{1:k}$. On a ainsi :

$$v_k(i) = \max_{(h_1,\dots,h_{k-1})} \mathbb{P}_\theta(O_{1:k} = o_{1:k}, H_{1:k} = (h_1, .., h_{k-1}, e_i))$$

Ces quantités se calculent par récurrence de la manière suivante :

$$\begin{cases} v_1(i) = \mu_i b_{h_i}(o_1) & \forall i \in [\![1, n]\!] \\ v_{k+1}(j) = (\max\limits_{1\le i\le n}[v_k(i)a_{ij}])\ b_{h_j}(o_{k+1}) & \forall (k, j) \in [\![1, t-1]\!] \times [\![1, n]\!] \end{cases}$$

Par la suite, on stockera ces valeurs dans une matrice V où $v_k(j)$ correspondera à l'élément de la k-ième colonne et de la j-ième ligne.

Afin de conserver en mémoire l'état caché qui maximise la probabilité d'émission à l'instant $k-1$, on utilise la variable $\phi_k(j)$ définit par :

$$\phi_k(j) = \operatorname*{argmax}_{1\le i\le n} (v_{k-1}(i)a_{ij})$$

Ici, la fonction argmax nous permet de conserver l'indice i, i.e. l'état h_i au temps $t-1$ qui maximise la quantité $(v_{k-1}(i)a_{ij})$. De même, on stockera ces valeurs dans une matrice Φ où $\phi_k(j)$ correspondera à l'élément de la k-ième colonne et de la $j$ième ligne. Enfin, la suite recherchée $h* = (h_1^*, ..., h_t^*)$ est donnée par :

$$\begin{cases} h_t^* = \operatorname*{argmax}\limits_{1\le i\le n} (v_t(i)) \\ h_{k-1}^* = \phi_{k-1}(h_{k-1}^*) \quad \forall k \in [\![2, t]\!] \end{cases}$$

De plus, la probabilité d'émission de $o_{1:t} = (o_1, .., o_t)$ par $h^* = (h_1^*, ..., h_t^*)$ est :

$$p^* = \max_{1\le i\le n} v_t(i)$$

4.2 Exemple d'application

On reprend la structure du modèle $\mathcal{M}$ de l'exemple 1 de la partie 1.3 et on s'intéresse cette fois-ci à la suite d'états cachés optimale qui engendre la séquence "aabb".

Pour $t = 1$:

$$\begin{cases} v_1(1) = \mu_1 b_1(a) = 0.6 \times 1 = 0.6 \\ v_1(2) = \mu_2 b_2(a) = 0.4 \times 0.5 = 0.2 \\ v_1(3) = \mu_3 b_3(a) = 0 \times 0 = 0 \end{cases}$$

$$\begin{cases} \phi_1(1) = 0 \\ \phi_1(2) = 0 \\ \phi_1(3) = 0 \end{cases}$$

Pour $t = 2$:

$$\begin{aligned} v_2(1) &= \max(v_1(1)a_{11},\ v_1(2)a_{21},\ v_1(3)a_{31}) \times b_1(a) \\ &= \max(0.6 \times 0.3,\ 0,\ 0) \times 1 \\ &= 0.18. \end{aligned}$$

On conserve l'indice 1 : $\phi_2(1) = 1$.

$$\begin{aligned} v_2(2) &= \max(v_1(1)a_{12},\ v_1(2)a_{22},\ v_1(3)a_{32}) \times b_2(a) \\ &= \max(0.6 \times 0.5,\ 0.2 \times 0.3,\ 0) \times 0.5 \\ &= 0.15. \end{aligned}$$

On conserve l'indice 1 : $\phi_2(2) = 1$.

$$\begin{aligned} v_2(3) &= \max(v_1(1)a_{13},\ v_1(2)a_{23},\ v_1(3)a_{33}) \times b_3(a) \\ &= \max(0.6 \times 0.2,\ 0.2 \times 0.7,\ 0) \times 0 \\ &= 0. \end{aligned}$$

On conserve l'indice 2 : $\phi_2(3) = 1$.

On obtient finalement les matrices :

$$V = \begin{pmatrix} 0.6 & 0.18 & 0 & 0 \\ 0.2 & 0.15 & 0.045 & 0.0067 \\ 0 & 0 & 0.105 & 0.105 \end{pmatrix} \quad \Phi = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 2 & 1 & 2 \\ 0 & 2 & 2 & 3 \end{pmatrix}$$

Et on en déduit le chemin optimal :

$$\begin{aligned} h_4^* &= \underset{1 \leq i \leq 3}{\text{argmax}}\ (v_4(i)) = \text{argmax}\ (0,\ 0.0067,\ 0.105) = 3 \\ h_3^* &= \phi_4(3) = 3 \\ h_2^* &= \phi_3(3) = 2 \\ h_1^* &= \phi_2(2) = 1 \end{aligned}$$

On en conclut que la suite cachée la plus probable pour la séquence "aabb" est donc $h^* = (1, 2, 3, 3)$ et la probabilité que h^* émette cette séquence est $p^* = 0.105$.

Conclusion

Nous avons à travers ce mémoire abordé plusieurs questions autour du modèle de Markov caché. Lors de l'analyse de ces différentes problématiques nous avons été confronté à des problèmes de complexité combinatoire qui sont directement liées à la structure latente du modèle de Markov caché et qui motivent l'utilisation de méthodes de résolutions particulières.

Ainsi, dans la perspective de déterminer la loi marginale du processus observable, nous avons étudié les algorithmes *Forward-Backward*. De même, pour contourner ce problème computationnel dans le cadre d'une problématique d'inférence des paramètres, nous avons été amené à considérer différentes méthodes d'estimation ; l'algorithme EM dans un contexte fréquentiste et l'algorithme de Gibbs dans un contexte bayésien. La mise en pratique de ces techniques d'estimation a également mis en lumière des difficultés d'implémentation qui ne sont pas anecdotiques lorsque l'on s'intéresse à ce type de modèle statistique. Parmi elles, on retrouve notamment la question du calcul de la probabilité d'une séquence d'observation lorsque celle ci est de grande taille. Enfin, dans une logique d'inférence complète nous avons pu répondre à la problématique de la recherche du chemin optimal dans un cadre de loi d'émission discrète à l'aide de l'application de l'algorithme de Viterbi.

Une suite logique à notre étude serait la généralisation de l'algorithme de Viterbi dans le cadre d'une loi d'émission continue. Dès lors, il serait intéressant d'étudier un éventuel lien entre le chemin optimal donné par Viterbi et la séquence cachée fournie par l'algorithme de Gibbs au cours des itérations. Concernant l'algorithme EM et l'algorithme de Gibbs, il serait pertinent d'approfondir leur analyse en comparant leurs performances selon plusieurs critères tout en étudiant leur sensibilités à la variation des paramètres préalablement choisis.

Annexe

1.Calcul des lois conditionelles :

On a :

$$f_\theta\left(o_{1:t}, h_{1:t}\right) \propto \mu_{h_1} \prod_{i=1}^{t-1} a_{h_i h_{i+1}} \prod_{k=1}^{t} \phi\left(o_k \mid m_{h_k}, \sigma\right).$$

Le calcul de la loi *a posteriori*, étant donné le modèle hiérarchique considéré, donne :

$$\pi(\theta \mid o_{1:t}, h_{1:t}) \propto \mu_{h1} \prod_{k=1}^{t-1} a_{h_k h_{k+1}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(o_k - m_{hk})^2}{2\sigma^2}\right] \left(\prod_{i=1}^{3} \mu_i^{1-1}\right) \left(\prod_{j=1}^{3}\prod_{i=1}^{3} a_{ij}^{1-1}\right)$$
$$\times \frac{1}{\left(\sqrt{2\pi w^{-2}}\right)^3} \exp\left[-\sum_{i=1}^{3} \frac{(m_i - u)^2}{2w^{-2}}\right] \sigma^{-2(\alpha-1)} \beta^\alpha e^{-\beta\sigma^{-2}} \times \beta^{f-1} f^g e^{-g\beta}.$$

On reconnaît ainsi les lois conditionnelles complètes suivantes :

$$\pi(\sigma^2 \mid \dots) \propto \left(\sigma^{-2}\right)^{\alpha+\frac{t}{2}} \exp\left[-\sigma^{-2}\left(\beta + \frac{1}{2}\sum_{k=1}^{t}(o_k - m_{hk})^2\right)\right]$$

Donc

$$\sigma^{-2} \mid \dots \sim \ \mathcal{G}a\left(\alpha + \frac{t}{2}, \beta + \frac{1}{2}\sum_{k=1}^{t}(o_k - m_{h_k})^2\right).$$

$$\pi(\beta \mid \dots) \propto \exp\left[-\beta(g + \sigma^{-2})\right] \beta^{f+\alpha-1}$$

Donc

$$\beta \mid \dots \sim \ \mathcal{G}a\left(f + \alpha, g + \sigma^{-2}\right).$$

$$\pi(\mu \mid \dots) \propto \prod_{i=1}^{n} \mu_i^{\left(\mathbb{I}_{(h_1=e_1)} + \mathbb{I}_{(h_1=e_2)} + \mathbb{I}_{(h_1=e_3)}\right)}$$

Donc

$$(\mu_1, \mu_2, \mu_3) \mid \dots \sim \ \text{Dirichlet}(\mathbb{I}_{h_1=e_1} + 1, \mathbb{I}_{h_1=e_2} + 1, \mathbb{I}_{h_1=e_3} + 1).$$

Pour $d \in [\![1; 3]\!]$,

$$\pi(a_{d1}, a_{d2}, a_{d3} \mid \dots) \propto \prod_{i=1}^{n} a_{d_i}^{n_{di}}$$

Donc

$$(a_{d1}, a_{d2}, a_{d3}) \mid \dots \sim \ \text{Dirichlet}(n_{d1} + 1, n_{d2} + 1, n_{d3} + 1) \ \text{ pour } d = 1, 2, 3$$

$$\mathbb{P}\left(H_1 = e_j \mid \dots\right) \propto \mu_{h1} \phi\left(o_1 \mid m_{h1}, \sigma^2\right) \beta_1(j)$$

Pour $k \in [\![1; t-1]\!]$,

$$\mathbb{P}\left(H_k = e_j \mid H_{k-1} = e_i\right) \propto a_{h_i h_j} \phi\left(o_k \mid m_{hk}, \sigma^2\right) \beta_k(j).$$

.
Pour $i \in [\![1; 3]\!]$,

$$\pi(m_i \mid \dots) \propto \left(\frac{2\pi\sigma^2}{n_i + \kappa\sigma^2}\right)^{-\frac{1}{2}} \exp\left[-\frac{\left(m_i - \frac{S_i + wu\sigma^2}{n_i + w^2}\right)^2}{2\left(\frac{\sigma^2}{n_i + w\sigma^2}\right)}\right]$$

Donc

$$m_i \mid \dots \sim \mathcal{N}\left(\frac{S_i + wu\sigma^2}{n_i + w\sigma^2}, \frac{\sigma^2}{n_i + w\sigma^2}\right).$$

Références

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6) :1554–1563, 1966.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, 1989.

Timo Koski. *Hidden Markov models for bioinformatics*, volume 2. Springer Science & Business Media, 2001.

Ramaprasad Bhar and Shigeyuki Hamori. *Hidden Markov models : applications to financial economics*, volume 40. Springer Science & Business Media, 2004.

Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer Science & Business Media, 2006.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 1977.

Christian P Robert, George Casella, and Joachim Robert. *Méthodes de Monte-Carlo avec R*. Springer, 2011.

G David Forney. The viterbi algorithm. *Proceedings of the IEEE*, 1973.

Tobias Rydén. EM versus Markov chain Monte Carlo for estimation of hidden Markov models : acomputational perspective. *Bayesian Analysis*, 2008.

Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410) :398–409, 1990.

Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 1970.

Ingmar Visser, Maartje EJ Raijmakers, and Peter CM Molenaar. Confidence intervals for hidden markov model parameters. *British journal of mathematical and statistical psychology*, 2000.

Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 1995.

A. F. M. Smith and G. O. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society, Series B (Methodological), vol. 55, no. 1, pp* $3_2 3, 1993$.