Compte-rendu du devoir 3

Vincent Matthys

1 HMM implementation

1.0 Notations

On désignera par M le nombre d'états cachés (4 dans notre cas). On notera $\boldsymbol{u} = \{u_t\}_{t\in[\![1;T]\!]}$, $\boldsymbol{q} = \{q_t\}_{t\in[\![1;T]\!]}$, $\boldsymbol{\theta} = (\pi, A, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ où $\boldsymbol{\mu} = \{\mu_k\}_{k\in[\![1;M]\!]}$ et $\boldsymbol{\Sigma} = \{\Sigma_k\}_{k\in[\![1;M]\!]}$. On utilisera également le *one-hot encoding* sur $\boldsymbol{q}$, de sorte que q_t^i, variable binaire, représente la i^{e} composante de q_t.

1.1

Le problème de l'inférence dans la chaine de Markov modélisée, à savoir le calcul de $p(\boldsymbol{q} \mid \boldsymbol{u})$ est appréhendable par le calcul de la marginale $p(q_t \mid \boldsymbol{u})$ suivant :

$$\begin{aligned}
p(q_t \mid \boldsymbol{u}) &= \frac{p(\boldsymbol{u} \mid q_t)p(q_t)}{p(\boldsymbol{u})} \\
&= \frac{p(u_1,\dots,u_t \mid q_t)p(u_{t+1},\dots,u_T \mid q_t)p(q_t)}{p(\boldsymbol{u})} \qquad u_1,\dots,u_t \perp\!\!\!\perp u_{t+1},\dots,u_T \mid q_t \\
&= \frac{p(u_1,\dots,u_t, q_t)p(u_{t+1},\dots,u_T \mid q_t)}{p(\boldsymbol{u})} \\
&= \frac{\alpha_t(q_t)\beta_t(q_t)}{p(\boldsymbol{u})} \\
&= \frac{\alpha_t(q_t)\beta_t(q_t)}{\sum_{q_t}\alpha_t(q_t)\beta_t(q_t)} \\
&= \frac{\alpha_t(q_t)\beta_t(q_t)}{\sum_{q_T}\alpha_T(q_T)} \quad \text{puisque} \quad p(\boldsymbol{u}) = \sum_{q_T} p(q_T, \boldsymbol{u}) = \sum_{q_T}\alpha_T(q_T)
\end{aligned} \tag{1}$$

avec $\alpha_t(q_t) = p(u_1,\dots,u_t,q_t)$ et $\beta_t(q_t) = p(u_{t+1},\dots,u_T \mid q_t)$
On peut alors aussi exprimer $p(q_t, q_{t+1} \mid \boldsymbol{u})$ en fonction de α_t et β_t :

$$
\begin{aligned}
p(q_t, q_{t+1} \mid \boldsymbol{u}) &= \frac{p(\boldsymbol{u} \mid q_t, q_{t+1}) p(q_t, q_{t+1})}{p(\boldsymbol{u})} \\
&= \frac{p(\boldsymbol{u} \mid q_t, q_{t+1}) p(q_{t+1} \mid q_t) p(q_t)}{p(\boldsymbol{u})} \\
&= \frac{p(u_1, \ldots, u_t \mid q_t, q_{t+1}) p(u_{t+1}, \ldots, u_T \mid q_t, q_{t+1}) p(q_{t+1} \mid q_t) p(q_t)}{p(\boldsymbol{u})} \\
&= \frac{p(u_1, \ldots, u_t \mid q_t) p(u_{t+1}, \ldots, u_T \mid q_{t+1}) p(q_{t+1} \mid q_t) p(q_t)}{p(\boldsymbol{u})} \\
&= \frac{p(u_1, \ldots, u_t \mid q_t) p(u_{t+1} \mid q_{t+1}) p(u_{t+2}, \ldots, u_T \mid q_{t+1}) p(q_{t+1} \mid q_t) p(q_t)}{p(\boldsymbol{u})} \\
&= \frac{p(u_1, \ldots, u_t, q_t) p(u_{t+1} \mid q_{t+1}) p(u_{t+2}, \ldots, u_T \mid q_{t+1}) p(q_{t+1} \mid q_t)}{p(\boldsymbol{u})} \\
&= \frac{\alpha_t(q_t) p(u_{t+1} \mid q_{t+1}) \beta_t(q_{t+1}) A_{q_t, q_{t+1}}}{p(\boldsymbol{u})}
\end{aligned} \tag{2}
$$

où $A_{q_t,q_{t+1}} = p(q_{t+1} \mid q_t)$ et $p(u_{t+1} \mid q_{t+1})$ suit une loi gaussienne.

À des fins d'implémentation, on considèrera les logarithmes des α et β, $\tilde{\alpha}$ et $\tilde{\beta}$ de sorte que les formules de récursion peuvent se réecrire :

$$
\begin{cases}
\ln \alpha_t(q_t) = \ln p(u_t \mid q_t) + \ln \sum_{q_{t-1}} \left(e^{\ln \alpha_{t-1}(q_{t-1})} A_{q_{t-1}, q_t} \right) \\
\ln \beta_t(q_t) = \ln \sum_{q_{t+1}} \left(e^{\ln \beta_{t+1}(q_{t+1})} A_{q_t, q_{t+1}} p(y_{t+1} \mid q_{t+1}) \right)
\end{cases} \tag{3}
$$

$$
\begin{cases}
\tilde{\alpha}_t(q_t) = \ln p(u_t \mid q_t) + \ln \sum_{q_{t-1}} \left(e^{\tilde{\alpha}_{t-1}(q_{t-1})} A_{q_{t-1}, q_t} \right) \\
\tilde{\beta}_t(q_t) = \ln \sum_{q_{t+1}} \left(e^{\tilde{\beta}_{t+1}(q_{t+1})} A_{q_t, q_{t+1}} p(y_{t+1} \mid q_{t+1}) \right)
\end{cases} \tag{4}
$$

et que la tâche d'inférence dite de *smoothing* en équation (1) s'écrit :

$$
\begin{aligned}
\ln p(q_t \mid \boldsymbol{u}) &= \ln \alpha_t(q_t) + \ln \beta_t(q_t) - \ln \sum_{q_T} \alpha_T(q_T) \\
\ln p(q_t \mid \boldsymbol{u}) &= \tilde{\alpha}_t(q_t) + \tilde{\beta}_t(q_t) - \ln \sum_{q_T} e^{\tilde{\alpha}_T(q_T)}
\end{aligned} \tag{5}
$$

La tâche d'inférence dite de *filtering* est également ajoutée en normalisant α_t :

$$
\alpha_t^{\text{norm}}(q_t) = p(q_t \mid u_1, \ldots, u_t) = \frac{\alpha_t(q_t)}{\sum_{q_t} \alpha_t(q_t)}
$$

et est disponible par le calcul des α_{norm} dans le code fourni.

1.2

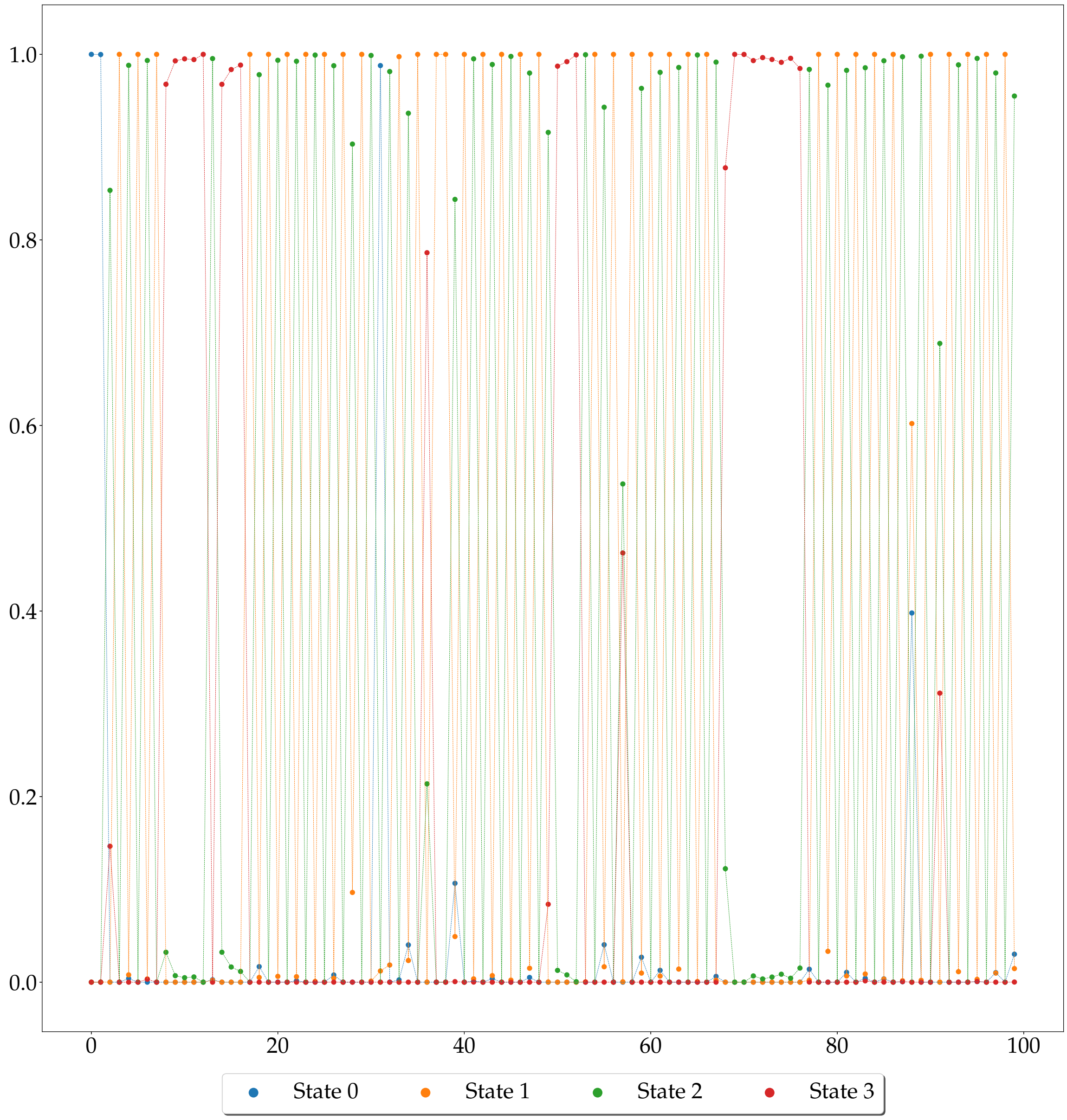

FIGURE 1 – Smoothing result for the first 100 datapoints

1.3

On exprime la log-vraisemblance incomplète de la manière suivante :

$$\begin{aligned}
\ell(\boldsymbol{\theta};\boldsymbol{u}) &= \ln \sum_{\boldsymbol{q}} p\left(\boldsymbol{q},\boldsymbol{u} \mid \boldsymbol{\theta}\right) \\
&= \ln \sum_{\boldsymbol{q}} r(\boldsymbol{q} \mid \boldsymbol{u}) \frac{p\left(\boldsymbol{q},\boldsymbol{u} \mid \boldsymbol{\theta}\right)}{r(\boldsymbol{q} \mid \boldsymbol{u})} \quad \text{pour r quelconque} \\
&\geq \sum_{\boldsymbol{q}} r(\boldsymbol{q} \mid \boldsymbol{u}) \ln \frac{p\left(\boldsymbol{q},\boldsymbol{u} \mid \boldsymbol{\theta}\right)}{r(\boldsymbol{q} \mid \boldsymbol{u})} \quad \text{par l'inégalité de Jensen} \\
&\geq \mathbb{E}_r[\ell_c(\boldsymbol{\theta};\boldsymbol{q},\boldsymbol{u})] - H(r(\boldsymbol{q} \mid \boldsymbol{u})) = \mathcal{L}(r,\boldsymbol{\theta},\boldsymbol{u})
\end{aligned} \tag{6}$$

avec $\ell_c(\boldsymbol{\theta}; \boldsymbol{q}, \boldsymbol{u})$, log-vraisemblance complète, dont l'expression est la suivante :

$$
\begin{aligned}
\ell_c(\boldsymbol{\theta}; \boldsymbol{q}, \boldsymbol{u}) &= \ln p\left(\boldsymbol{q}, \boldsymbol{u} \mid \boldsymbol{\theta}\right) \\
&= \sum_{t=1}^{T} \ln p(u_t \mid q_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \sum_{t=2}^{T} \sum_{i=1}^{M} \sum_{j=1}^{M} q_{t-1}^i q_t^j \ln A_{ij} + \sum_{i=1}^{M} q_1^i \ln \pi_i \\
&= \sum_{t=1}^{T} \sum_{k=1}^{M} q_t^k \left(- \ln 2\pi - \frac{1}{2} \ln |\Sigma_k| - \frac{1}{2} (u_t - \mu_k)^\intercal \Sigma_k^{-1} (u_t - \mu_k) \right) \\
&+ \sum_{t=2}^{T} \sum_{i=1}^{M} \sum_{j=1}^{M} q_{t-1}^i q_t^j \ln A_{ij} + \sum_{i=1}^{M} q_1^i \ln \pi_i
\end{aligned} \tag{7}
$$

On a donc une borne inférieure de la log-vraisemblance incomplète, $\mathcal{L}(r, \boldsymbol{\theta}, \boldsymbol{u})$. L'algorithme EM consiste en une montée alternée de cette borne inférieure par rapport à r et $\boldsymbol{\theta}$. En choisissant $r(\boldsymbol{q} \mid \boldsymbol{u}) = p(\boldsymbol{q} \mid \boldsymbol{u}, \boldsymbol{\theta})$, on maximise celle-ci, de sorte que les deux étapes de l'algorithme se résument de la manière suivante :

$$
\begin{aligned}
E - step \quad r^{(s+1)} &= \arg\max_{r} \mathcal{L}(r, \boldsymbol{\theta}^{(s)}, \boldsymbol{u}) = p(\boldsymbol{q} \mid \boldsymbol{u}, \boldsymbol{\theta}^{(s)}) \\
M - step \quad \boldsymbol{\theta}^{(s+1)} &= \arg\max_{\boldsymbol{\theta}} \mathcal{L}(r^{(s+1)}, \boldsymbol{\theta}, \boldsymbol{u}) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{r^{(s+1)}}(\ell_c(\boldsymbol{\theta}; \boldsymbol{q}, \boldsymbol{u}))
\end{aligned} \tag{8}
$$

où s dénote le nombre d'itérations de l'algorithme.

D'après l'équation (7), on observe que $m_{ij} := \sum_{t=2}^{T} q_{t-1}^i q_t^j$ est la statistique exhaustive pour A_{ij}, que q_1^i est la statistique exhaustive pour π_i. Les estimateurs de maximum de log-vraisemblance complète pour A et π sont donc :

$$
\widehat{\pi}_i = q_1^i \tag{9}
$$

$$
\widehat{A}_{ij} = \frac{m_{ij}}{\sum_{k=1}^{M} m_{ik}} \tag{10}
$$

D'autre part, grâce aux calculs déjà effectués pour les estimateurs de vraisemblance de la mixture de gaussiennes, on obtient les estimateurs de log-vraisemblance complète des paramètres d'émission :

$$
\widehat{\mu_k} = \frac{\sum_{t=1}^{T} q_t^k u_t}{\sum_{t=1}^{T} q_t^k} \tag{11}
$$

$$
\widehat{\Sigma_k} = \frac{\sum_{t=1}^{T} q_t^k \left(u_t - \mu_k\right) \left(u_t - \mu_k\right)^\intercal}{\sum_{t=1}^{T} q_t^k} \tag{12}
$$

Ces estimateurs doivent être pris sous l'espérance précise obtenue lors de l'étape E de l'algorithme EM, donnée par l'équation (8), afin que la maximisation de la log-vraisemblance complète induise une augmentation de la log-vraisemblance incomplète lors de l'itération $s + 1$.

En reprenant l'estimateur de π obtenu en équation (9), on obtient l'estimateur à l'itération $(s + 1)$:

$$
\begin{aligned}
\widehat{\pi}_i^{(s+1)} &= \mathbb{E}_{\boldsymbol{q} \mid \boldsymbol{u}, \boldsymbol{\theta}^{(s)}}[q_1^i] \\
&= \mathbb{E}_{q_1 \mid \boldsymbol{u}, \boldsymbol{\theta}^{(s)}}[q_1^i] \\
&= p(q_1^i = 1 \mid \boldsymbol{u}, \boldsymbol{\theta}^{(s)}) := \gamma_1^{i,(s+1)}
\end{aligned} \tag{13}
$$

En reprenant l'estimateur de A_{ij} obtenue en équation (10), il faut déterminer $\mathbb{E}_{\boldsymbol{q}|\boldsymbol{u},\boldsymbol{\theta}^{(s)}}[m_{ij}]$. On a :

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{q}|\boldsymbol{u},\boldsymbol{\theta}^{(s)}}[m_{ij}] &= \sum_{t=2}^{T} \mathbb{E}_{\boldsymbol{q}|\boldsymbol{u},\boldsymbol{\theta}^{(s)}}\left[q_{t-1}^{i} q_{t}^{j}\right] \\
&= \sum_{t=2}^{T} p\left(q_{t-1}^{i} = 1, q_{t}^{j} = 1 \mid \boldsymbol{u}, \boldsymbol{\theta}^{(s)}\right) \\
&:= \sum_{t=2}^{T} \xi_{t-1,t}^{ij,(s+1)}
\end{aligned}$$

D'où l'estimateur à l'itération $(s+1)$ pour A :

$$\widehat{A}_{ij}^{(s+1)} = \frac{\sum_{t=2}^{T} \xi_{t-1,t}^{ij,(s+1)}}{\sum_{k=1}^{M} \sum_{t=2}^{T} \xi_{t-1,t}^{ik,(s+1)}} = \frac{\sum_{t=2}^{T} \xi_{t-1,t}^{ij,(s+1)}}{\sum_{t=2}^{T} \gamma_{t-1}^{i,(s+1)}} \tag{14}$$

De la même façon, on obtient les estimateurs à l'itération $(s+1)$ pour les μ_k et Σ_k :

$$\begin{aligned}
\widehat{\mu_k}^{(s+1)} &= \frac{\sum_{t=1}^{T} \gamma_t^{k,(s+1)} u_t}{\sum_{t=1}^{T} \gamma_t^{k,(s+1)}} \quad \forall k \in [\![1; M]\!] \\
\widehat{\Sigma_k}^{(s+1)} &= \frac{\sum_{t=1}^{T} \gamma_t^{k,(s+1)} \left(u_t - \mu_k^{(s)}\right)\left(u_t - \mu_k^{(s)}\right)^{\intercal}}{\sum_{t=1}^{T} \gamma_t^{k,(s+1)}} \quad \forall k \in [\![1; M]\!]
\end{aligned} \tag{15}$$

L'implémentation du calcul des γ et ξ est succintement détaillée et fait intervenir le calcul des α_{norm} du problème de *filtering*.

$$\begin{aligned}
\gamma_t(q_t) &= p(q_t \mid u_1, \ldots, u_T) \\
&= \sum_{q_{t+1}} p(q_t, q_{t+1} \mid u_1, \ldots, u_T) \\
&= \sum_{q_{t+1}} \frac{\alpha_t(q_t) A_{q_t,q_{t+1}}}{\sum_{q_t} \alpha_t(q_t) A_{q_t,q_{t+1}}} \gamma_{t+1}(q_{t+1})
\end{aligned} \tag{16}$$

$$\begin{aligned}
\xi_{t,t+1}(q_t, q_{t+1}) &= p(q_t, q_{t+1} \mid \boldsymbol{u}) \\
&= \frac{\alpha_t(q_t) p(u_{t+1} \mid q_{t-1}) \gamma_{t+1}(q_{t+1}) A_{q_t,q_{t+1}}}{\alpha_{t+1}(q_{t+1})}
\end{aligned} \tag{17}$$