Expectation Maximization pour les chaînes de Markov cachées discrètes

Soit $(X_k)_{0\le k\le n}$ une chaîne de Markov discrète à valeurs dans $\{1,\ldots,r\}$, de matrice de transition Q et de loi initiale ν. On considère que cette chaîne est uniquement observée au travers des variables $(Y_k)_{0\le k\le n}$, indépendantes conditionnellement à $(X_k)_{0\le k\le n}$ et telles que pour tout $0\le \ell\le n$, la loi de Y_ℓ sachant $(X_k)_{0\le k\le n}$ est une gaussienne de moyenne μ_{X_ℓ} et de variance v_{X_ℓ}. Le paramètre inconnu est ici $\theta=\{\mu_1,\ldots,\mu_r,v_1,\ldots,v_r\}$.

1. Écrire la logvraisemblance jointe de $(X_{0:n},Y_{0:n})$: $\theta\mapsto \log p_\theta(X_{0:n},Y_{0:n})$.

   Notons φ_{μ,σ^2} la densité de la loi gaussienne de moyenne μ et de variance σ^2. Nous avons, pour tout 1

$$
\begin{aligned}
\log p_\theta(X_{0:n},Y_{0:n}) &= \log p_\theta(X_{0:n}) + \log p_\theta(Y_{0:n}|X_{0:n})\,,\\
&= \log p_\theta(X_0) + \sum_{k=1}^n \log p_\theta(X_k|X_{k-1}) + \sum_{k=0}^n \log p_\theta(Y_k|X_k)\,,\\
&= \sum_{i=1}^r \mathbb{1}_{X_0=i}\log \nu_i + \sum_{k=1}^n\sum_{i,j=1}^r \mathbb{1}_{X_{k-1}=i,X_k=j}\log Q_{i,j} + \sum_{k=0}^n\sum_{i=1}^r \mathbb{1}_{X_k=i}\log\varphi_{\mu_i,v_i}(Y_k)\,.
\end{aligned}
$$

2. Écrire la quantité intermédiaire de l'EM $Q(\theta,\theta')$ pour tout θ, θ' :

$$
Q(\theta,\theta') = \mathbb{E}_{\theta'}\left[\log p_\theta(X_{0:n},Y_{0:n})|Y_{0:n}\right]\,.
$$

   Par la question précédente,

$$
\begin{aligned}
Q(\theta,\theta') &= \mathbb{E}_{\theta'}\left[\log p_\theta(X_{0:n},Y_{0:n})|Y_{0:n}\right]\,,\\
&= \mathbb{E}_{\theta'}\left[\sum_{i=1}^r \mathbb{1}_{X_0=i}\log\nu_i + \sum_{k=1}^n\sum_{i,j=1}^r \mathbb{1}_{X_{k-1}=i,X_k=j}\log Q_{i,j} + \sum_{k=0}^n\sum_{i=1}^r \mathbb{1}_{X_k=i}\log\varphi_{\mu_i,v_i}(Y_k)\middle| Y_{0:n}\right]\,,\\
&= \sum_{i=1}^r \mathbb{E}_{\theta'}\left[\mathbb{1}_{X_0=i}|Y_{0:n}\right]\log\nu_i + \sum_{k=1}^n\sum_{i,j=1}^r \mathbb{E}_{\theta'}\left[\mathbb{1}_{X_{k-1}=i,X_k=j}\middle|Y_{0:n}\right]\log Q_{i,j}\\
&\qquad + \sum_{k=0}^n\sum_{i=1}^r \mathbb{E}_{\theta'}\left[\mathbb{1}_{X_k=i}|Y_{0:n}\right]\log\varphi_{\mu_i,v_i}(Y_k)\,.
\end{aligned}
$$

3. Écrire cette quantité en faisant apparaître les probabilités

$$
\omega^\theta_{k-1,k}(i,j) = \mathbb{P}_\theta\left(X_{k-1}=i,X_k=j|Y_{0:n}\right)\,,
$$

   pour $1\le k\le n$ et

$$
\tilde\omega^\theta_k(i) = \mathbb{P}_\theta\left(X_k=i|Y_{0:n}\right)\,,
$$

   pour $0\le k\le n$.

   Il suffit d'appliquer la question précédente :

$$
Q(\theta,\theta') = \sum_{i=1}^r \tilde\omega_0^{\theta'}(i)\log\nu_i + \sum_{k=1}^n\sum_{i,j=1}^r \omega^{\theta'}_{k-1,k}(i,j)\log Q_{i,j} + \sum_{k=0}^n\sum_{i=1}^r \tilde\omega_k^{\theta'}(i)\log\varphi_{\mu_i,v_i}(Y_k)\,.
$$

4. À l'itération $p \geq 0$, on dispose de l'estimation $\hat{\theta}^{(p)}$. Écrire l'estimateur $\hat{\theta}^{(p+1)}$ en maximisant $\theta \mapsto Q(\theta, \hat{\theta}^p)$, donner la valeur de $\theta^{(p+1)}$ en fonction des $\omega_{k-1,k}^{\theta^{(p)}}(i,j)$, $1 \leq k \leq n$, $1 \leq i,j \leq r$.

   On peut montrer que la fonction $\theta \mapsto Q(\theta, \theta^{(p)})$ admet un maximum unique obtenu en résolvant l'équation $\nabla_\theta Q(\theta, \theta^{(p)}) = 0$.

   - *Pour tout $1 \leq i \leq r$ et tout $1 \leq j \leq r-1$, en remarquant que $Q_{i,r} = 1 - \sum_{\ell=1}^{r-1} Q_{i,\ell}$,*

     $$\partial_{Q_{i,j}} Q(\theta, \theta^{(p)}) = \sum_{k=1}^{n} \frac{\omega_{k-1,k}^{\theta^{(p)}}(i,j)}{Q_{i,j}} - \sum_{k=1}^{n} \frac{\omega_{k-1,k}^{\theta^{(p)}}(i,r)}{Q_{i,r}} .$$

     On obtient donc que pour tout $1 \leq i \leq r$ et tout $1 \leq j \leq r-1$,

     $$\sum_{k=1}^{n} \frac{\omega_{k-1,k}^{\theta^{(p)}}(i,j)}{Q_{i,j}} = \sum_{k=1}^{n} \frac{\omega_{k-1,k}^{\theta^{(p)}}(i,r)}{Q_{i,r}} ,$$

     puis que

     $$Q_{i,j}^{(p+1)} = \frac{\sum_{k=1}^{n} \omega_{k-1,k}^{\theta^{(p)}}(i,j)}{\sum_{k=1}^{n} \tilde{\omega}_{k-1}^{\theta^{(p)}}(i,j)}$$

   - *Pour tout $1 \leq i \leq r$,*

     $$\partial_{\mu_i} Q(\theta, \theta^{(p)}) = -\frac{1}{v_i} \sum_{k=0}^{n} \tilde{\omega}_k^{\theta^{(p)}}(i) \left(\mu_i + Y_k\right) .$$

     Ainsi,

     $$\mu_i^{(p+1)} = \frac{\sum_{k=1}^{n} \tilde{\omega}_k^{\theta^{(p)}}(i) Y_k}{\sum_{k=0}^{n} \tilde{\omega}_k^{\theta^{(p)}}(i)} .$$

   - *Pour tout $1 \leq i \leq r$,*

     $$\partial_{v_i} Q(\theta, \theta^{(p)}) = \sum_{k=0}^{n} \tilde{\omega}_k^{\theta^{(p)}}(i) \left(-\frac{1}{2v_i} + \frac{1}{2v_i^2}(Y_k - \mu_i)^2\right) .$$

     Ainsi,

     $$v_i^{(p+1)} = \frac{\sum_{k=0}^{n} \tilde{\omega}_k^{\theta^{(p)}}(i)(Y_k - \mu_i^{(p+1)})^2}{\sum_{k=1}^{n} \tilde{\omega}_k^{\theta^{(p)}}(i)} .$$

5. Dans le cas où on souhaite également apprendre la loi initiale de la chaîne de Markov, i.e. si l'on note $\nu_i = \mathbb{P}(X_0 = i)$, $1 \leq i \leq r$ et que $\theta = \{\mu_1, \ldots, \mu_r, v_1, \ldots, v_r, \nu_1, \ldots, \nu_r\}$, donner les équations de mise à jour de ν.

   En utilisant la question 4, on obtient, pour tout $1 \leq i \leq r-1$,

   $$\partial_{\nu_i} Q(\theta, \theta^{(p)}) = \frac{\tilde{\omega}_0^{\theta^{(p)}}(i)}{\nu_i} - \frac{\tilde{\omega}_0^{\theta^{(p)}}(r)}{\nu_r} ,$$

   et

   $$\nu_i^{(p+1)} = \tilde{\omega}_0^{\theta^{(p)}}(i) .$$

6. Calculer le gradient de la logvraisemblance des observations : $\theta \mapsto \nabla_\theta \log p_\theta(Y_{0:n})$.

On remarque que pour tout θ, θ'

$$
\begin{aligned}
\log p_\theta(Y_{0:n}) &= \sum_{i=1}^{n} \log p_\theta(Y_{0:n}) p_{\theta'}(X_{0:n} = x_{0:n}|Y_{0:n}) \\
&= \sum_{i=1}^{n} \log \frac{p_\theta(X_{0:n} = x_{0:n}, Y_{0:n})}{p_\theta(X_{0:n} = x_{0:n}|Y_{0:n})} p_{\theta'}(X_{0:n} = x_{0:n}|Y_{0:n})\,, \\
&= \sum_{i=1}^{n} \log p_\theta(X_{0:n} = x_{0:n}, Y_{0:n}) p_{\theta'}(X_{0:n} = x_{0:n}|Y_{0:n}) \\
&\qquad - \sum_{i=1}^{n} \log p_\theta(X_{0:n} = x_{0:n}|Y_{0:n}) p_{\theta'}(X_{0:n} = x_{0:n}|Y_{0:n})\,,
\end{aligned}
$$

Ainsi,

$$
\nabla_{\theta=\theta'} \log p_\theta(Y_{0:n}) = \mathbb{E}_{\theta'}\left[\nabla_{\theta=\theta'} \log p_\theta(X_{0:n}, Y_{0:n}) | Y_{0:n}\right]\,,
$$

car

$$
\begin{aligned}
&\nabla_{\theta=\theta'} \sum_{i=1}^{n} \log p_\theta(X_{0:n} = x_{0:n}|Y_{0:n}) p_{\theta'}(X_{0:n} = x_{0:n}|Y_{0:n}) \\
&\qquad = \nabla_{\theta=\theta'} \sum_{i=1}^{n} p_\theta(X_{0:n} = x_{0:n}|Y_{0:n}) = 0\,.
\end{aligned}
$$

Le score se calcule à l'aide d'une espérance conditionnelle du même type que la quantité intermédiaire de l'EM.

7. En déduire un algorithme de mise à jour des paramètres de type "descente de gradient".

   Si l'on souhaite utiliser une méthode du premier ordre on peut écrire, pour $p \geq 0$,

   $$
   \tilde{\theta}^{(p+1)} = \tilde{\theta}^{(p)} + \gamma_p \mathbb{E}_{\theta^{(p)}}\left[\nabla_{\theta=\theta^{(p)}} \log p_\theta(X_{0:n}, Y_{0:n}) | Y_{0:n}\right]\,,
   $$

   où les $\{\gamma_p\}_{p\geq 0}$ sont des pas positifs.

8. **Bonus:** Calcul des $\omega^\theta_{k-1,k}(i,j)$, $1 \leq k \leq n$, $1 \leq i,j \leq r$.

   (a) Montrer que l'on peut calculer récursivement $\mathbb{P}_\theta\left(X_k = i|Y_{0:k}\right)$, $0 \leq k \leq n$, $1 \leq i,j \leq r$.

   (b) Montrer que l'on peut calculer récursivement, de $k = n$ à $k = 0$, $\mathbb{P}_\theta\left(X_k = i|Y_{0:n}\right)$, $0 \leq k \leq n$, $1 \leq i \leq r$.

   (c) Conclure.