Expectation Maximization pour les chaînes de Markov cachées discrètes

Soit $r$ un entier strictement positif. Notons $\mathbb{S}_r$ l'ensemble des matrices de transition de taille $r\times r$, i.e. l'ensemble des matrices de taille $r\times r$ de réels positifs telles que la somme de chaque ligne soit égale à 1. Considérons un paramètre $\theta = \{P,\mu_1,\dots,\mu_r,v_1,\dots,v_r\}\in\Theta$, où $\Theta = \mathbb{S}_r\times\mathbb{R}^r\times(\mathbb{R}_+^*)^r$ et le modèle suivant paramétré par $\theta$.

Soit $(X_k)_{0\le k\le n}$ une chaîne de Markov discrète à valeurs dans $\{1,\dots,r\}$, de matrice de transition $P$ et de loi initiale $\nu$. Cela signifie que pour tout $1\le j\le r$, $\mathbb{P}_\theta(X_0=j)=\nu_j$ et pour tout $0\le k\le n-1$ et tout $1\le i,j\le r$, $\mathbb{P}_\theta(X_{k+1}=j|X_k=i)=P_{i,j}$. 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 $Y_\ell$ sachant $(X_k)_{0\le k\le n}$ est une gaussienne de moyenne $\mu_{X_\ell}$ et de variance $v_{X_\ell}$.

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

   *Notons $\varphi_{\mu,\sigma^2}$ la densité de la loi gaussienne de moyenne $\mu$ et de variance $\sigma^2$. Nous avons, pour tout $\theta$,*

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

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

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

   *Par la question précédente,*

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

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 :*

$$Q(\theta,\theta') = \sum_{i=1}^r \tilde\omega^{\theta'}_0(i)\log\nu_i + \sum_{k=1}^n\sum_{i,j=1}^r \omega^{\theta'}_{k-1,k}(i,j)\log P_{i,j} + \sum_{k=0}^n\sum_{i=1}^r \tilde\omega^{\theta'}_k(i)\log\varphi_{\mu_i,v_i}(Y_k)\,.$$

4. À l'itération $p \geq 0$, on dispose de l'estimation $\hat{\theta}^{(p)}$. Écrire l'estimateur $\hat{\theta}^{(p+1)}$ en maximisant $\theta \mapsto Q(\theta, \hat{\theta}^{p})$.

   *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 $P_{i,r} = 1 - \sum_{\ell=1}^{r-1} P_{i,\ell}$,*

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

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

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

   *puis que*

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

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

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

   *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 et que $\theta = \{P, \mu_1, \ldots, \mu_r, v_1, \ldots, v_r, \nu_1, \ldots, \nu_r\}$, donner les équations de mise à jour de $\nu$.

   *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 $\theta, \theta'$ et en notant $\mathsf{X} = \{1, \ldots, r\}$,*

$$\begin{aligned}
\log p_\theta(Y_{0:n}) &= \sum_{x_{0:n} \in \mathsf{X}^{n+1}} \log p_\theta(Y_{0:n}) \mathbb{P}_{\theta'}(X_{0:n} = x_{0:n} | Y_{0:n}) \\
&= \sum_{x_{0:n} \in \mathsf{X}^{n+1}} \log \frac{p_\theta(x_{0:n}, Y_{0:n})}{\mathbb{P}_\theta(X_{0:n} = x_{0:n} | Y_{0:n})} \mathbb{P}_{\theta'}(X_{0:n} = x_{0:n} | Y_{0:n}) \,, \\
&= \sum_{x_{0:n} \in \mathsf{X}^{n+1}} \log p_\theta(x_{0:n}, Y_{0:n}) \mathbb{P}_{\theta'}(X_{0:n} = x_{0:n} | Y_{0:n}) \\
&\qquad - \sum_{x_{0:n} \in \mathsf{X}^{n+1}} \log \mathbb{P}_\theta(x_{0:n} | Y_{0:n}) \mathbb{P}_{\theta'}(X_{0:n} = x_{0:n} | Y_{0:n}) \,,
\end{aligned}$$

*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*

$$\begin{aligned}
&\nabla_{\theta=\theta'} \sum_{i=1}^{n} \log \mathbb{P}_\theta(X_{0:n} = x_{0:n} | Y_{0:n}) \mathbb{P}_{\theta'}(X_{0:n} = x_{0:n} | Y_{0:n}) \\
&\qquad = \nabla_{\theta=\theta'} \left( \sum_{x_{0:n} \in \mathsf{X}^{n+1}} \mathbb{P}_\theta(X_{0:n} = x_{0:n} | Y_{0:n}) \right) = 0 \,.
\end{aligned}$$

*Le score (gradient de la logvraisemblance) 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.