2 Analyse d'un déplacement animal par chaîne de Markov cachée

Problème

On cherche à analyser les déplacements d'un zèbre au cours d'une journée. Pour cela, pendant 24 heures, toutes les 8 minutes environ, on a relevé sa position géographique. De ces mesures on a déduit, pour chaque temps t, sa vitesse et l'angle relatif de sa direction par rapport à la direction précédente. Dans la suite, on notera Y_t le déplacement au temps t avec

$$Y_t = \left[\begin{array}{ll} Y_{1t} = & \text{vitesse au temps } t \\ Y_{2t} = & \text{angle relatif au temps } t \end{array} \right]$$

Les données, issues de Patin et al. [2019], sont disponibles dans le fichier `dryad_zebra.csv` (accessible sur le moodle du cours).

Objectif. On cherche à distinguer quelques comportements typiques dans les déplacements de l'animal au cours de la journée.

2.1 Modèle de Markov caché

On se propose d'utiliser le modèle de Markov caché à K états suivant

$$\begin{aligned} Z = \{Z_t\}_{1 \leq t \leq n} &\sim CM_K(\nu, \pi), \\ \{Y_t\}_{1 \leq t \leq n} \text{ indépendants } \mid Z : \quad Y_t \mid Z_t = k &\sim \mathcal{N}(\mu_k, \Sigma_k) \end{aligned} \tag{5}$$

où ν désigne la distribution initiale de la chaîne cachée Z, π sa matrice de transition, $\mu_k \in \mathbb{R}^2$ l'espérance du déplacement dans l'état k et $\Sigma_k \in \mathcal{M}_2$ sa variance dans l'état k. Les paramètres du modèle à K états sont réunis dans

$$\theta = (\nu, \pi, (\mu_k)_{1 \leq k \leq K}, (\Sigma_k)_{1 \leq k \leq K}).$$

Estimation des paramètres.

1. Écrire la vraisemblance complète de ce modèle.
2. En déduire son espérance conditionnelle aux données observées pour une valeur courante du paramètre notée $\theta^{(h)}$. On notera $\tau_{tk}^{(h)} = \mathbb{P}_{\theta^{(h)}}\{Z_t = k \mid Y\}$ et $\eta_{tk\ell}^{(h)} = \mathbb{P}_{\theta^{(h)}}\{Z_{t-1,k} Z_{t,\ell} \mid Y\}$.
3. En supposant les quantités $\tau_{tk}^{(h)}$ et $\eta_{tk\ell}^{(h)}$ connues, en déduire la valeur $\theta^{(h+1)}$ qui maximise $\mathbb{E}_{\theta^{(h)}}(\log p_\theta(Z, Y) \mid Y)$ en θ.
4. Déterminer le critère BIC permettant de choisir le nombre d'état K.

2.2 Implémentation de l'algorithme EM

1. Écrire une fonction `Mstep` prenant en arguments les données Y et la valeur courante des probabilités conditionnelles $\tau_{ik}^{(h)}$ et $\eta_{tk\ell}^{(h)}$ et qui retourne les estimations obtenues à la question 3.
2. Écrire une fonction `Forward` prenant en arguments les données Y et la valeur courante du paramètre $\theta^{(h)}$ et qui retourne
   — les probabilités conditionnelles $F_{tk} = \mathbb{P}_{\theta^{(h)}}\{Z_t = k \mid Y_1^t\}$,
   — les densités estimées $\phi_{tk}^{(h)} = \mathcal{N}(Y_t, \mu_k^{(h)}, \Sigma_k^{(h)})$ et
   — la vraisemblance $\log p_{\theta^{(h)}}(Y)$.
3. Écrire une fonction `Backward` prenant en arguments la valeur courante du paramètre $\theta^{(h)}$ et le résultat de la fonction `Forward` et qui retourne
   — les probabilités conditionnelles $\tau_{tk}^{(h)}$ et
   — les probabilités conditionnelles $\eta_{tk\ell}^{(h)}$.
4. A partir de méthodes que vous connaissez, proposer une initialisation des espérances conditionnelles τ_{tk}^0 et $\eta_{tk\ell}^0$. Écrire une fonction `InitHMM` prenant en arguments les données Y et le nombre d'états K et qui retourne ces valeurs.
5. Écrire une fonction `HMM` prenant en arguments les données Y et le nombre d'états K et utilisant l'algorithme EM et qui retourne
   — l'estimation par maximum de vraisemblance $\widehat{\theta}$ de θ,
   — les espérances conditionnelles $\widehat{\tau}_{tk}$ et $\widehat{\eta}_{tk\ell}$ correspondantes et
   — la log-vraisemblance $\log p_{\widehat{\theta}}(Y)$.

2.3 Application

1. Appliquer la fonction `HMM` aux données de Patin et al. [2019] pour $K = 2$ et interpréter les paramètres.
2. Utiliser le critère BIC pour choisir le nombre d'états K et interpréter les résultats. Quels grands types de comportements pouvez-vous distinguer chez l'animal ?

Question subsidiaire.

3. Écrire une fonction `Viterbi` prenant en arguments l'estimation finale du paramètre $\widehat{\theta}$ et les espérances conditionnelles $\widehat{\tau}_{tk}$ et $\widehat{\eta}_{tk\ell}$ et qui retourne le chemin caché le plus probable

$$\widehat{Z} = \underset{z \in \{1, \dots K\}^n}{\arg\max} \; \mathbb{P}_{\widehat{\theta}}\{Z = z \mid Y\}.$$

   On pourra s'aider des notes de cours disponibles sur le moodle du cours.

Rendu attendu

Vous enverrez à l'adresse `stephane.robin@sorbonne-universite.fr` un fichier '.R' contenant l'implémentation en R de l'algorithme EM. Ce programme devra :

— prendre en entrée un fichier de la même forme que `dryad_zebra.csv`,
— tracer l'évolution de la log-vraisemblance au cours des itérations de l'algorithme EM,
— afficher la valeur du critère BIC pour $K = 1 \dots 5$,
— afficher les estimations des paramètres du modèles pour le K optimal,
— représenter la trajectoire de l'animal coloriée en fonction de l'état caché.