Université de Lille 1 – U.F.R. de Mathématiques
M1 MAS – Ingénierie Mathématique – Semestre 1, 2017

TISD [1] – DM 2

Travail à faire en binôme. Un rapport et un script R par binôme doivent être déposés sur Moodle avant le vendredi 22 décembre 23h55. Le nom des étudiants doit apparaître en commentaire au début du script. Le rapport doit être au format PDF (travaillez avec R Markdown, ou Latex, ou OpenOffice, ou Word, etc, puis exportez le fichier en format PDF non éditable). Vous y détaillerez les réponses aux questions, les résultats graphiques, et y ferez part de vos commentaires. Il n'est pas nécessaire d'y inclure les programmes. La clarté et la présentation des rapports & scripts seront appréciés dans la note.

Partie A : Théorique

Dans la suite, $\mathcal{S}^+$ est l'espace des matrices réelles 2×2 symétriques définies positives. On note $\gamma_{\mu,\Sigma}(x)$ la densité d'un vecteur gaussien $\mathcal{N}(\mu, \Sigma)$ sur $\mathbb{R}^2$ de moyenne $\mu \in \mathbb{R}^2$ et de matrice de covariance $\Sigma \in \mathcal{S}^+$.

A.1. Expliquer comment obtenir un vecteur gaussien $\mathcal{N}(\mu, \Sigma)$ par une transformation affine d'un vecteur gaussien $\mathcal{N}(0, I_2)$.

A.2. On veut calculer les estimateurs du maximum de vraissemblance pour μ et Σ. Soit un échantillon $X_1, \dots, X_n$ i.i.d de loi $\mathcal{N}(\mu, \Sigma)$ avec $(\mu, \Sigma) \in \mathbb{R}^2 \times \mathcal{S}^+$ inconnus, et on note $\mathcal{L}(\mu, \Sigma | X_1, \dots, X_n) := \log \prod_{i=1}^n \gamma_{\mu,\Sigma}(X_i)$ la log-vraissemblance du modèle. On introduit également

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^n X_i, \qquad \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})\, {}^t(X_i - \hat{\mu}).$$

1. On pose $\phi(\mu) := \mathcal{L}(\mu, \Sigma | X_1, \dots, X_n)$. Montrer que [2] $\mathrm{d}\phi_\mu = 0$ si et seulement si $\mu = \hat{\mu}$.
2. On pose $\psi(\Sigma) := \mathcal{L}(\mu, \Sigma | X_1, \dots, X_n)$. Justifier que

$$\sum_{i=1}^n {}^t(X_i - \hat{\mu}) \Sigma^{-1} (X_i - \hat{\mu}) = n \mathrm{Tr}\big(\Sigma^{-1} \hat{\Sigma}\big)$$

et donc que

$$\psi(\Sigma) = -\frac{n}{2} \left(\log\det(\Sigma) + \mathrm{Tr}\big(\Sigma^{-1}\hat{\Sigma}\big) + \log(2\pi) \right).$$

Montrer que [3] $\mathrm{d}\psi_\Sigma = 0$ si et seulement si $\Sigma = \hat{\Sigma}$.

---

1. Responsable : Adrien Hardy. Laboratoire Paul Painlevé, Université des Sciences et Technologies de Lille, Bâtiment M3, Bureau 306. Email : adrien.hardy@math.univ-lille1.fr
2. Bien sûr, $\mathrm{d}\phi_\mu$ est la différentielle de l'application ϕ en $\mu \in \mathbb{R}^2$.
3. Aide : On pourra utiliser sans preuve que la différentielle du déterminant en l'identité donnée par la formule $\det(I + \varepsilon A) = 1 + \varepsilon \mathrm{Tr}(A) + O(\varepsilon^2)$ quand $\varepsilon \to 0$.

On fixe maintenant $k \geq 2$, on considère l'espace des paramètres

$$\Theta := \Big\{\mu_1, \ldots, \mu_k \in \mathbb{R}^2,\ \Sigma_1, \ldots, \Sigma_k \in \mathcal{S}^+,\ \pi_1, \ldots, \pi_k > 0 \text{ tels que } \sum_{j=1}^{k} \pi_j = 1\Big\},$$

ainsi que les densités de mélanges de vecteurs gaussiens

$$f_\theta(x) := \sum_{j=1}^{k} \pi_j\, \gamma_{\mu_j,\Sigma_j}(x). \tag{1}$$

Ayant à disposition un échantillon i.i.d. $X_1, \ldots, X_n$ de densité f_{θ^*} avec $\theta^* \in \Theta$ inconnu, on veut appliquer l'algorithme EM pour obtenir une suite d'estimateurs (θ_m) pour θ^*.

A.4. Après avoir introduit une variable Z appropriée, en utilisant les mêmes notations que dans le cours, donner les formules des densités $h_\theta(x, z)$, $g_\theta(z)$, et $g_\theta(z|x)$, et préciser quelles sont les mesures de référence pour chaque variable. Ensuite, donner une formule pour $Q(\theta, \theta_m)$; on notera $H_{iz}(\theta) := g_\theta(z|X_i)$ par commodité.

A.5. Montrer que les estimateurs (θ_m) de l'algorithme EM sont donnés par les formules de récurrence :

$$\begin{aligned}
\pi_j^{(m+1)} &= \frac{1}{n}\sum_{i=1}^{n} H_{ij}(\theta_m),\\
\mu_j^{(m+1)} &= \frac{\sum_{i=1}^{n} X_i H_{ij}(\theta_m)}{\sum_{i=1}^{n} H_{ij}(\theta_m)},\\
\Sigma_j^{(m+1)} &= \frac{\sum_{i=1}^{n} (X_i - \mu_j^{(m+1)})\, {}^t(X_i - \mu_j^{(m+1)}) H_{ij}(\theta_m)}{\sum_{i=1}^{n} H_{ij}(\theta_m)}.
\end{aligned}$$

Partie B : Implémentation sous R

B.1. A l'aide de la réponse à la question **A.1.**, créer une fonction `Vgauss` qui renvoie n réalisations indépendantes d'un vecteur gaussien $\mathcal{N}(\mu, \sigma)$ quand on lui donne n, μ, Σ.

B.2. Créer une fonction qui renvoie n réalisations indépendantes du mélange gaussien (1) étant donné n et un paramètre $\theta \in \Theta$.

B.3. Créer une fonction qui, d'un jeu de données $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^2$, un paramètre initial $\theta_0 \in \Theta$ et $M \geq 1$, renvoie les estimateurs $(\theta_m)_{m=1\ldots M}$ de l'algorithme EM jusqu'à l'ordre M.

B.4. On considère le vecteur $\mathbf{x}_1 = (x_1, y_1), \ldots, \mathbf{x}_n = (x_n, y_n)$ où $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$ correspondent respectivement aux variables `Petal.Length` et `Petal.Width` du jeu de données `Iris` disponible sur R.

1. Appliquer l'algorithme EM à ces données pour un mélange de $k = 3$ vecteurs gaussiens.

2. On attribue à chaque observation $\mathbf{x}_i$ le label $\hat{z}_i \in \{1, 2, 3\}$ défini par

$$\hat{z}_i := \underset{z \in \{1,2,3\}}{\operatorname{argmax}}\, g_{\theta_M}(z|\mathbf{x}_i).$$

   Interprétez cette formule, puis donner une représentation graphique des observations $\mathbf{x}_i$ où l'on représentera les trois groupes $\hat{z}_i = 1, 2, 3$ d'une couleur différente.

3. Commenter vos résultats.