5 Exercice 5

On s'intéresse à la modélisation d'une population composée de K groupes et l'on suppose que celle-ci peut être modélisée par un mélange de lois normales :

$$f(x) = \sum_{k=1}^{K} \pi_k \phi(x; \mu_k, \sigma_k^2)$$

où ϕ est la densité de probabilité de la loi normale :

$$\phi(x; \mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

1. Pour $\theta = \left((\mu_1, \sigma_1^2), \cdots, (\mu_K, \sigma_K^2), \pi_1, \cdots, \pi_{K-1}\right)$, donnez l'expression de la log-vraisemblance $\log(L_\theta(\mathcal{D}))$ pour des données complètes $\mathcal{D} = \{(x_1, z_1), \cdots, (x_n, z_n)\}$ où $z_i = (z_{i1}, \cdots, z_{iK})$ avec $z_{ik} = 1$ si x_i appartient au groupe k et 0 sinon.
2. Montrez que l'estimateur du maximum de vraisemblance de μ_k est $\hat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{n} z_{ik} x_i$ où $n_k = \sum_{i=1}^{n} z_{ik}$.
3. Calculez l'estimateur du maximum de vraisemblance pour σ_k^2.
4. Calculez l'estimateur du maximum de vraisemblance pour π_k.
5. Refaites ces calculs dans le cadre de la loi normale multi-dimensionnelle.

6 Exercice 6

On considère les données suivantes :

```
> donnees<-matrix(c(-3.3,-4.4,-1.9,3.3,2.5,3.2,0.3,0.1,-0.1,-0.5,
+                   1,1,1,2,2,2,2,2,1,1,
+                   1,3,2,1,3,2,1,3,2,1),nrow=10,ncol=3)
> donnees<-as.data.frame(donnees)
> names(donnees)<-c("Var","partition1","partition2")
> t(donnees)
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
Var	-3.3	-4.4	-1.9	3.3	2.5	3.2	0.3	0.1	-0.1	-0.5
partition1	1.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0	1.0	1.0
partition2	1.0	3.0	2.0	1.0	3.0	2.0	1.0	3.0	2.0	1.0

1. Faites 3 itérations de l'algorithme EM pour proposer une segmentation en $k = 2$ groupes des données ci-dessus. Vous utiliserez la partition1 ci-dessus comme initialisation et calculerez la vraisemblance à chaque itération.
   (a) Etape E : calculez les probabilités conditionnelles $t_{ik} = P(Z = k | X = x_i)$ pour $i = 1, \cdots, 10$ et $k = 1, 2$.
   (b) Etape M : calculez les estimateurs du maximum de pseudo-vraisemblance de μ_k et σ_k^2.
2. Calculez le critère BIC pour $k = 2$ et 3. Quel nombre de groupes est-il le plus vraisemblable ?
3. Comparez sur cet échantillon le comportement des algorithmes k-means et EM.