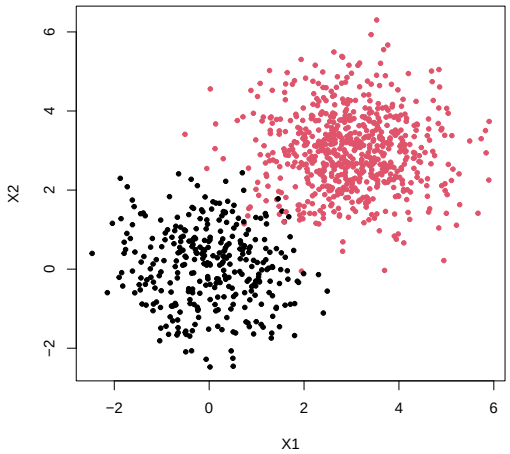
TP3 : partitionnement avec les GMM

Reprendre le cours sur le clustering via les modèles de mélanges Gaussiens (Gaussian Mixture Models-GMM).

1. Simuler une matrice de données X à partir du code ci-dessous :

```
library(mvtnorm)
n <- 1000
set.seed(50)
U =runif(n)
X <-  matrix(NA,n,2)
Z <- rep(NA,n)
for(i in 1:n) {
  if(U[i] < .3) {
    X[i,] <- rmvnorm(1,c(0,0),diag(c(1,1)))
    Z[i] <- 1
  } else {
    X[i,] <- rmvnorm(1,c(3,3),diag(c(1,1)))
    Z[i] <- 2
  }
}
Z <- as.factor(Z)
levels(Z) <- c("classe1", "classe2")
colnames(X)=c("X1","X2")
plot(X, pch=20, col=Z)
```



2. Ces données ont été simulées à partir d'un mélange de mélange Gaussien à deux classes :

$$\pi_1 \times \mathbb{N}(\mu_1, \Sigma) + \pi_2 \times \mathbb{N}(\mu_2, \Sigma).$$

   Quels sont d'après le code ci-dessus les paramètres θ de ce modèle ?

3. On veut maintenant appliquer l'algorithme EM pour estimer ces paramètres à partir de la matrice X.
   a. Quel échantillon est utilisé pour estimer θ ?
   b. Initialiser l'algorithme EM à partir de la partition en 2 classes de la CAH de Ward.

c. Appliquer l'étape M pour estimer θ. Calculer avec cette valeur de θ la vraisemblance de l'échantillon. Vous pouvez utiliser la fonction `dmvnorm` du package **mvtnorm**.
d. Appliquer l'étape E pour estimer les probabilités à posteriori.
e. Recommencer c. et d.
f. A votre avis, l'algorithme a-t-il convergé ? Quel critère de convergence proposez-vous ?

4. En déduire une partition en deux classes et la comparer à la "vraie" partition.

5. En pratique, le nombre de classe est choisi en utilisant un critère de vraisemblance pénalisé :

$$BIC(\mathcal{M}) = -2 \ln f_{\mathcal{M}}(\mathbf{x}_1, ..., \mathbf{x}_n; \theta) + |\mathcal{M}| \ln n$$

où $\mathcal{M}$ est un modèle de mélange à K classes et $|\mathcal{M}|$ le nombre de paramètres de ce modèle.

a. Quel est le nombre de paramètres pour un mélange de K gaussiennes de dimension 2.
b. En déduire la valeur du BIC de notre modèle.
c. Comment pourrait-on alors procéder pour choisir le nombre de classes ? Tester à l'aide du package **mclust** que cette procédure recommande le choix de $K = 2$ classes.