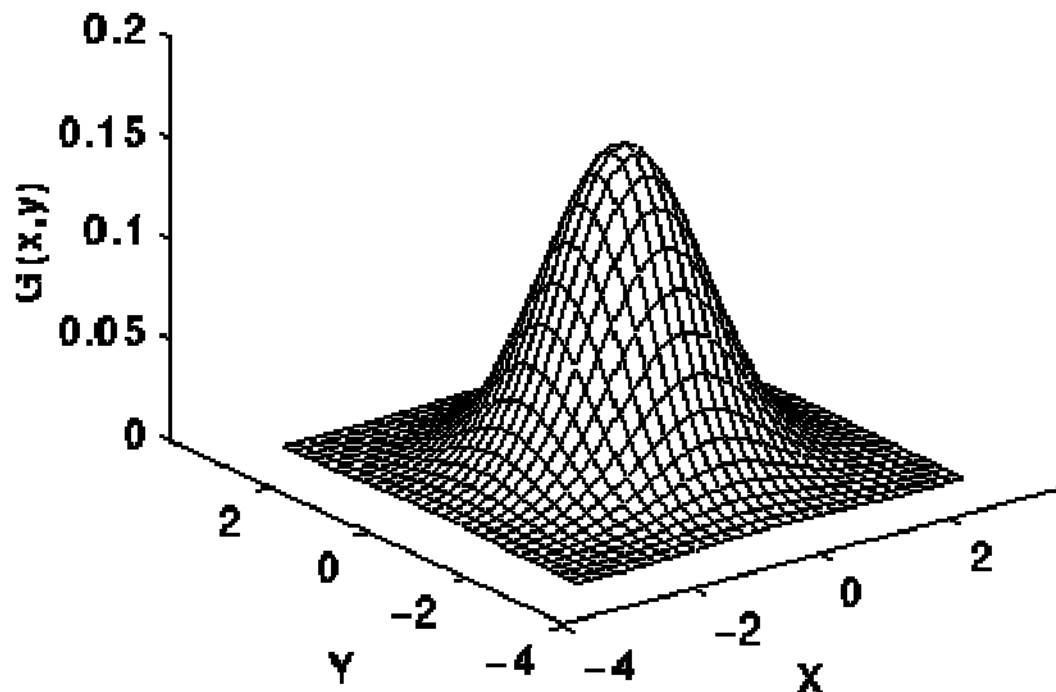


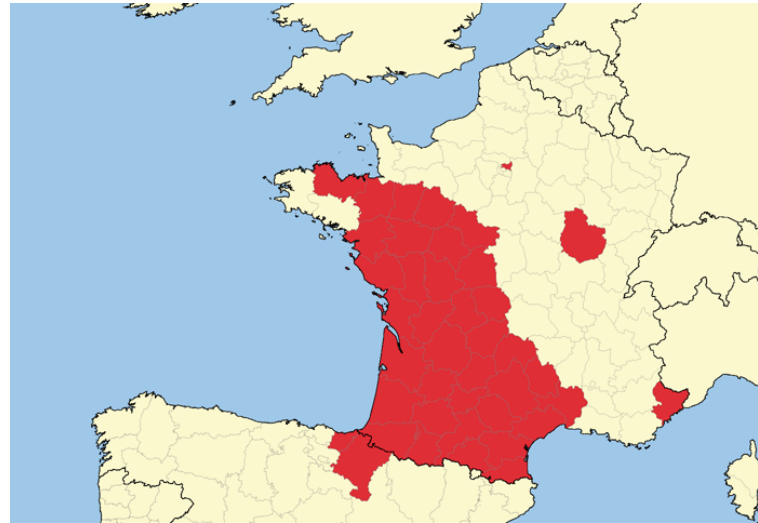


Mclust :

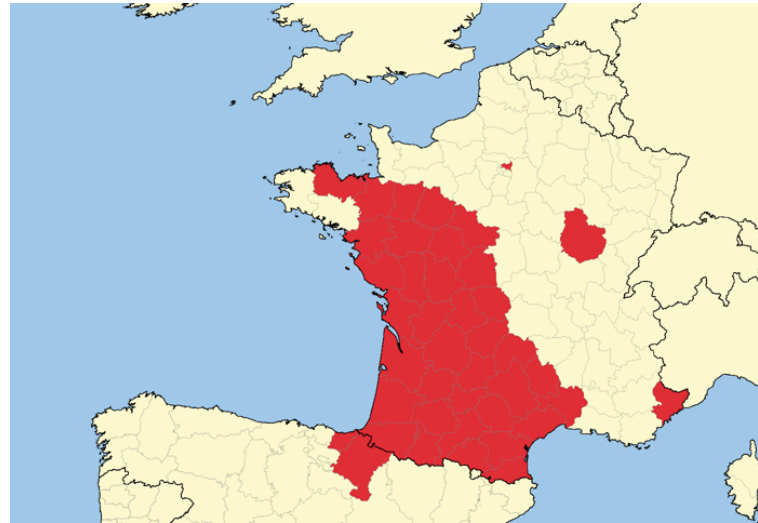
Déceler des groupes dans un jeu de données grâce aux mélanges gaussiens.



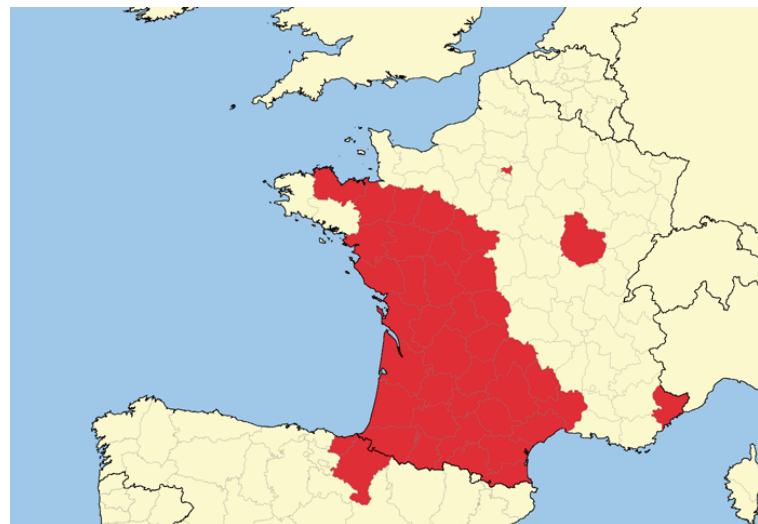
Partition et mélanges gaussiens



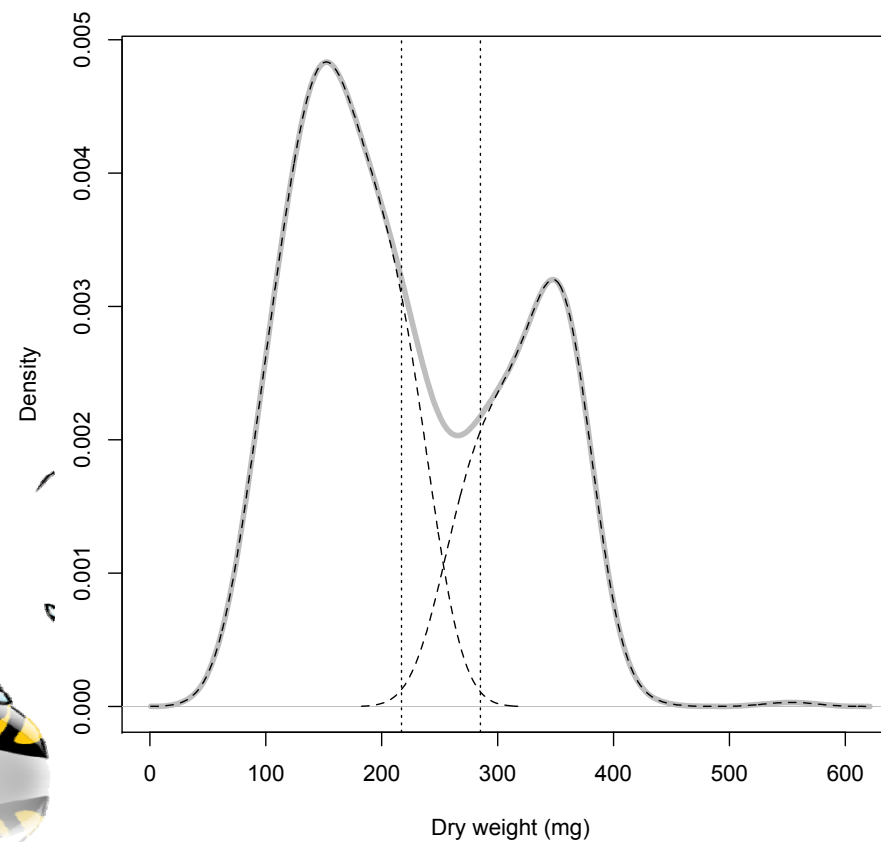
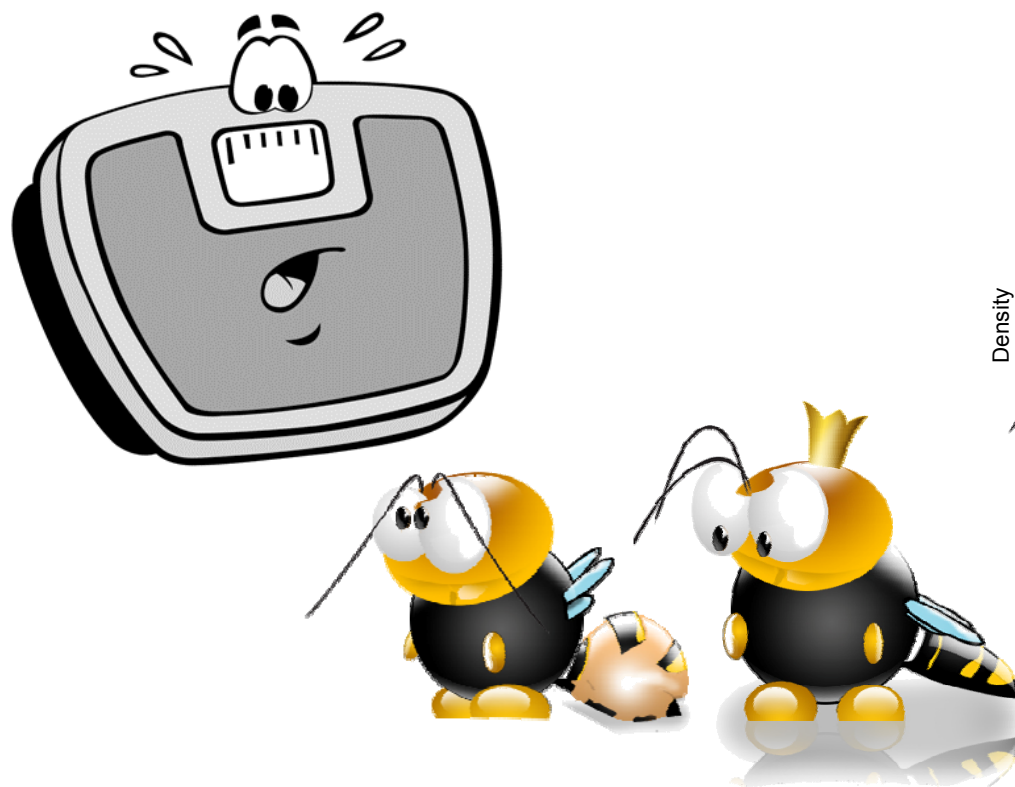
Partition et mélanges gaussiens



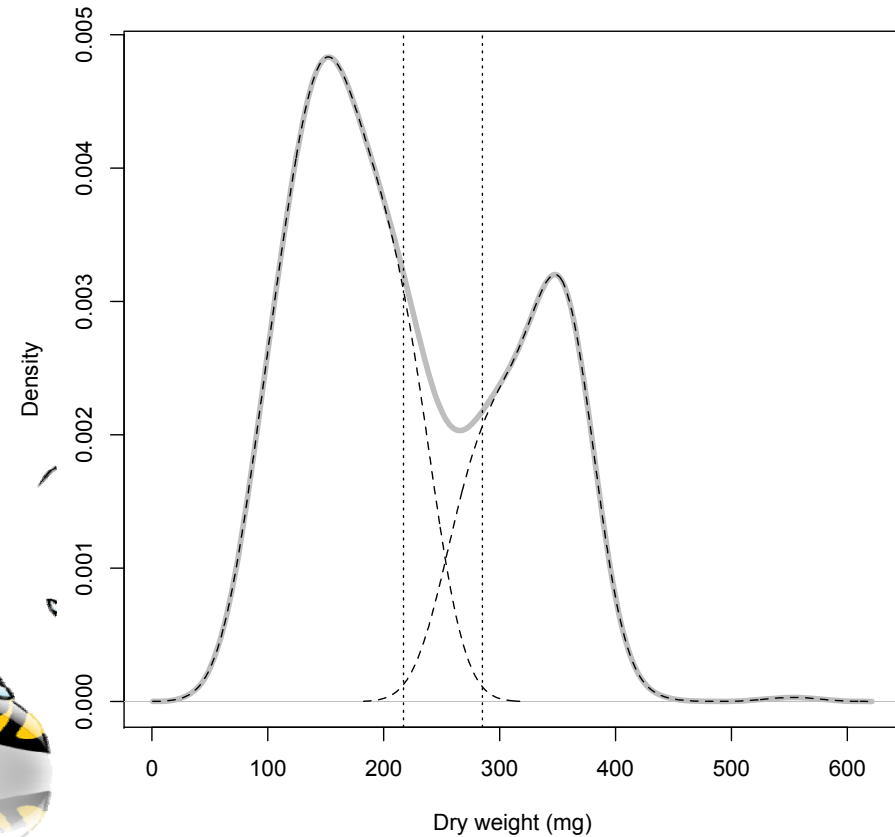
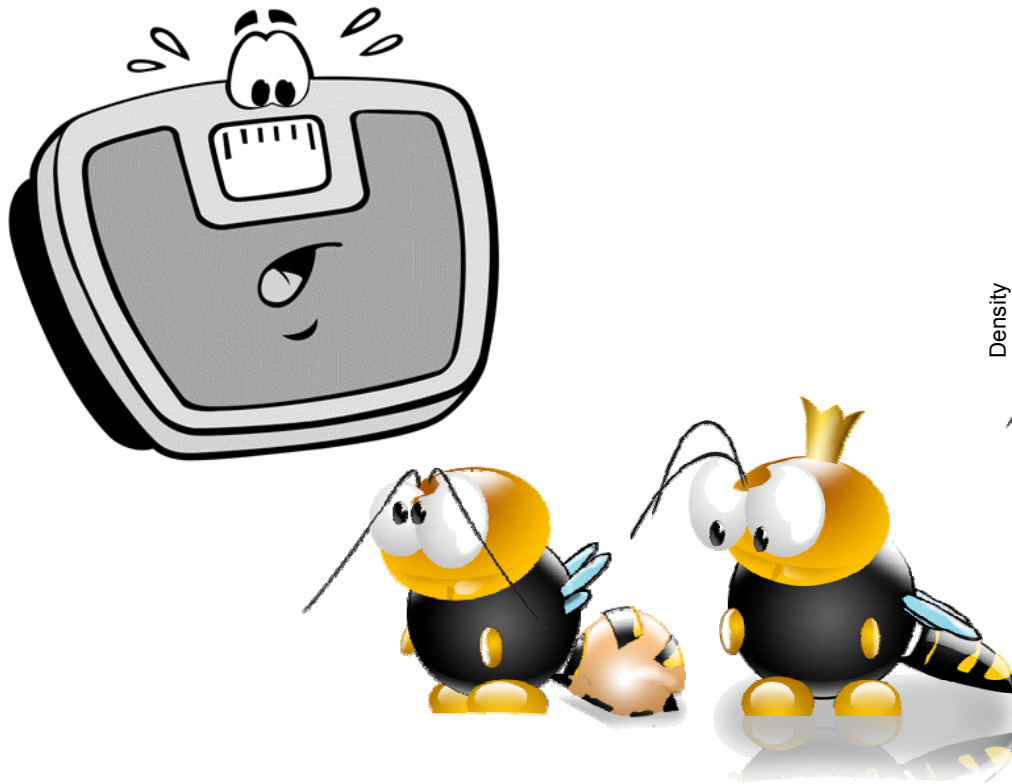
Partition et mélanges gaussiens



Partition et mélanges gaussiens



Partition et mélanges gaussiens

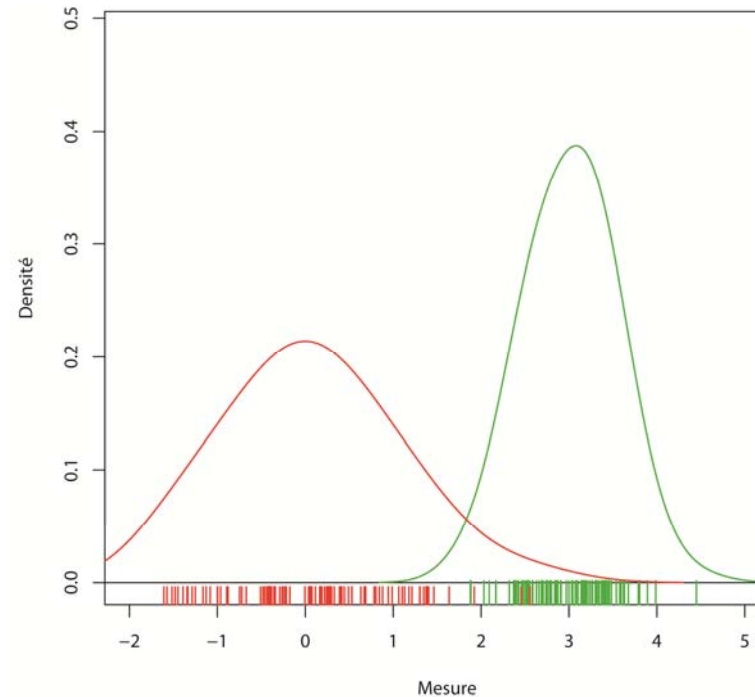


Méthodes de partition (« clustering »):

- classification hiérarchique (hclust)
- moyennes mobiles (kmeans)
- mélanges gaussiens (**Mclust**)

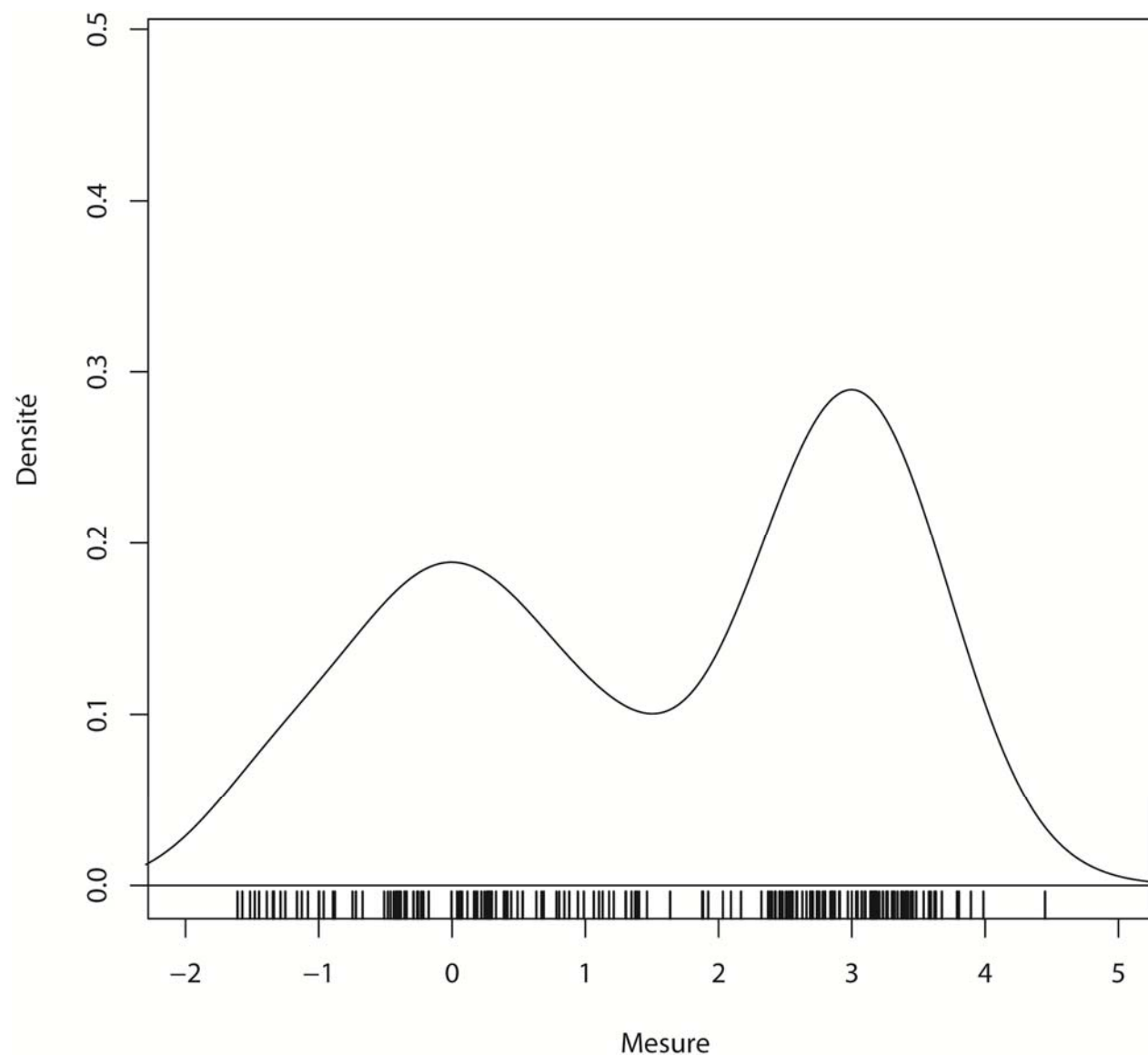
Partition et mélanges gaussiens

Ici, on a 2 groupes bien définis dans nos « mesures » ...



```
A<-rnorm(100,0,2)
B<-rnorm(100,3,0.5)
N<-c(A,B)
plot(N,xlim=c(-2,5),ylim=c(0,0.5),type="n")
abline(h=0);
rug(A,col=2);rug(B,col=3)
lines(density(A),col=2);
lines(density(B),col=3);
```

Partition et mélanges gaussiens



Partition et mélanges gaussiens

Fonction « Mclust »

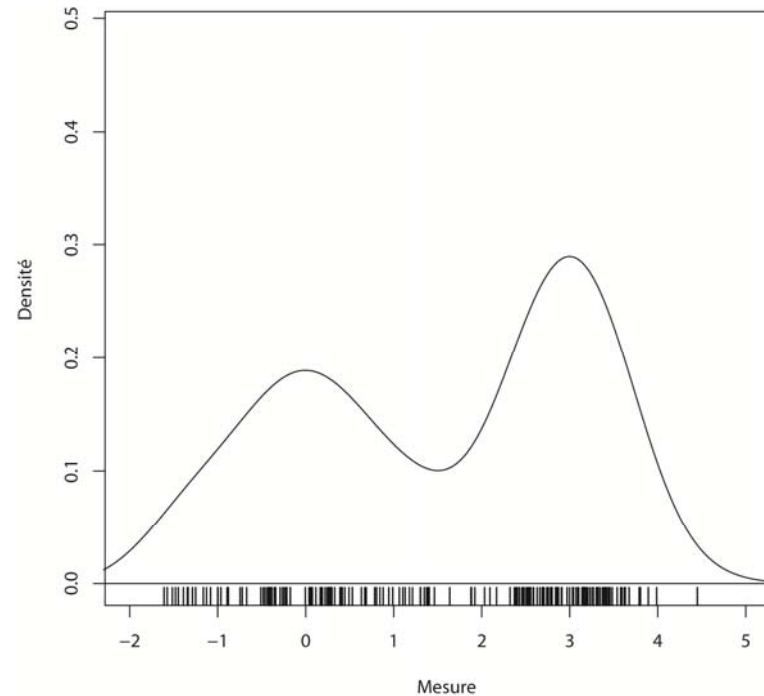
En supposant que les groupes aient une variation suivant une distribution Normale :

```
library(mclust)
```

```
Mclust(N)->McN
```

```
> str(McN)
```

```
List of 11
 $ modelName      : chr "V"
 $ n              : int 200
 $ d              : num 1
 $ G              : int 2
 $ BIC            : num [1:9, 1:2]
 $ bic            : num -715
 $ loglik         : num -344
 $ parameters     :List of 4
 $ z              : num [1:200, 1:2]
 $ classification: num [1:200]
 $ uncertainty    : num [1:200]
```



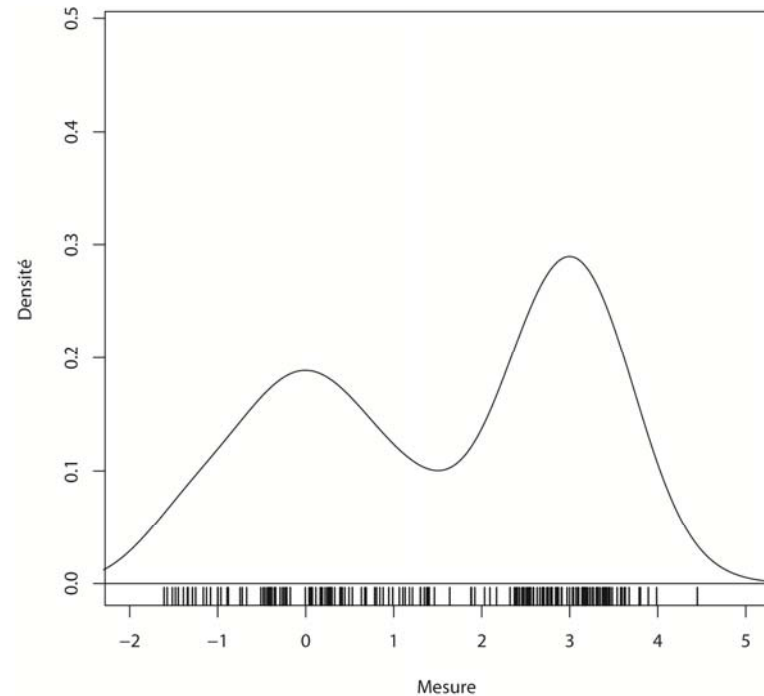
```
plot(N,xlim=c(-2,5),ylim=c(0,0.5),type="n")
abline(h=0)
rug(N)
lines(density(N))
```

Partition et mélanges gaussiens

Fonction « Mclust »

En supposant que les groupes aient une variation suivant une distribution Normale :

- Combien de groupes ?



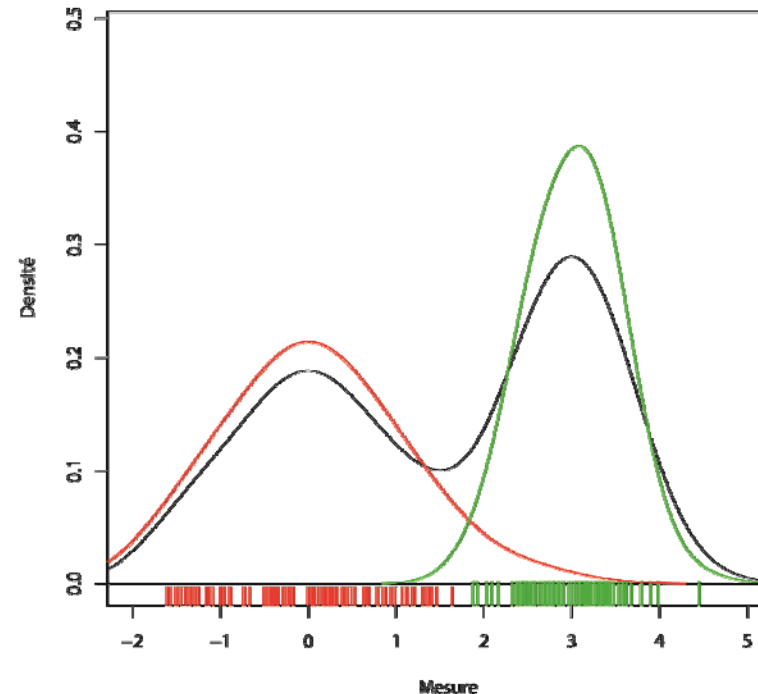
```
> McN$G  
[1] 2
```

Partition et mélanges gaussiens

Fonction « Mclust »

En supposant que les groupes aient une variation suivant une distribution Normale :

- Combien de groupes ?
- Classification des individus ?



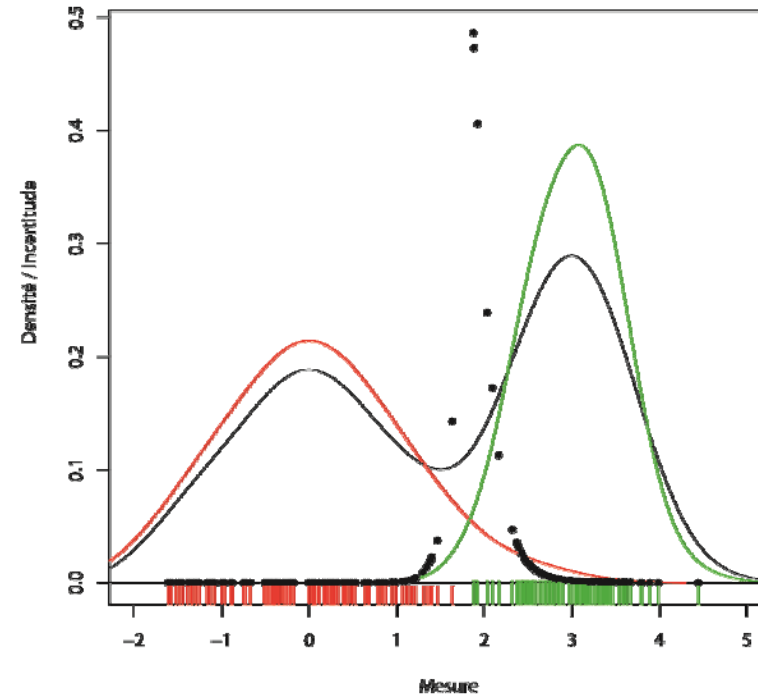
```
rug(N[which(McN$classification==1)],col=2)  
rug(N[which(McN$classification==2)],col=3)  
  
> classError(McN$class,gl(2,300))$errorRate  
[1] 0.025
```

Partition et mélanges gaussiens

Fonction « Mclust »

En supposant que les groupes aient une variation suivant une distribution Normale :

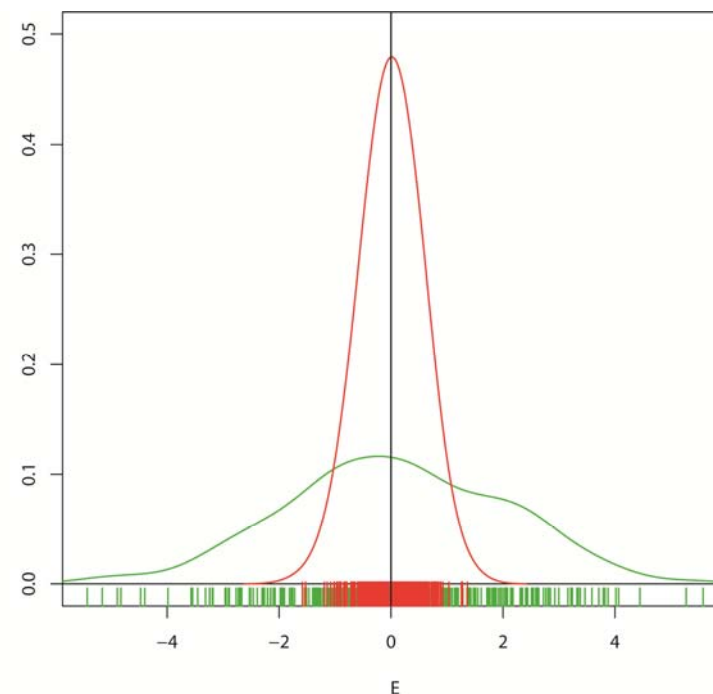
- Combien de groupes ?
- Classification des individus ?
- Degré d'incertitude d'attribution des individus aux groupes ?



```
points(N,McN$uncertainty,pch=20)
```

Partition et mélanges gaussiens

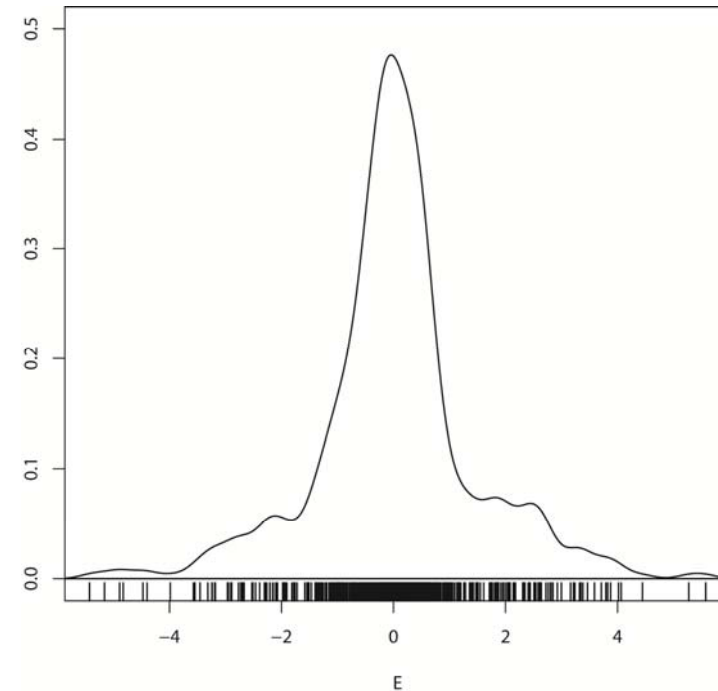
Différents groupes dont la moyenne est équivalente?



```
C<-rnorm(300,0,2)
D<-rnorm(300,0,0.5)
E<-c(C,D)
plot(E,ylim=c(0,0.5),type="n")
abline(h=0,v=0)
rug(C,col=3);rug(D,col=2)
lines(density(A),col=2);
lines(density(B),col=3);
```

Partition et mélanges gaussiens

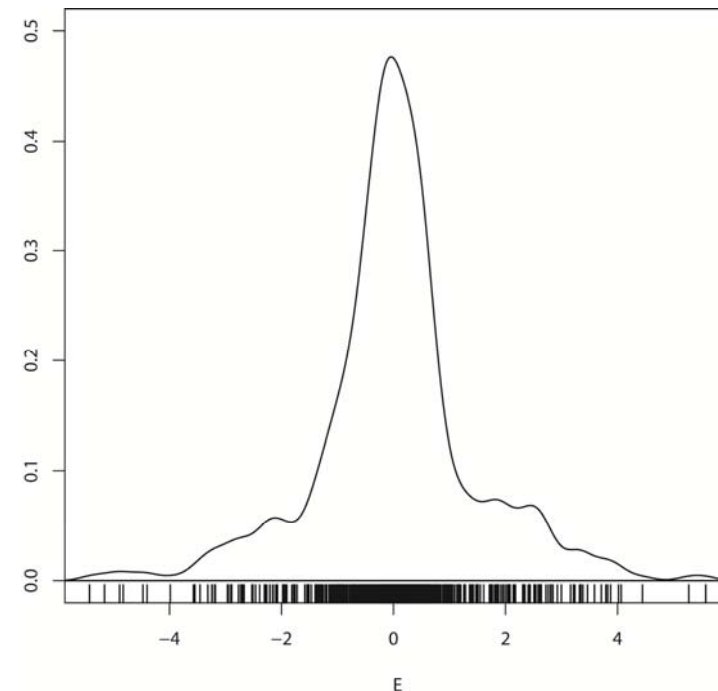
Différents groupes dont la moyenne est équivalente?



Partition et mélanges gaussiens

Différents groupes dont la moyenne est équivalente?

- Les groupes sont détectés



```
Mclust(E) -> McE
```

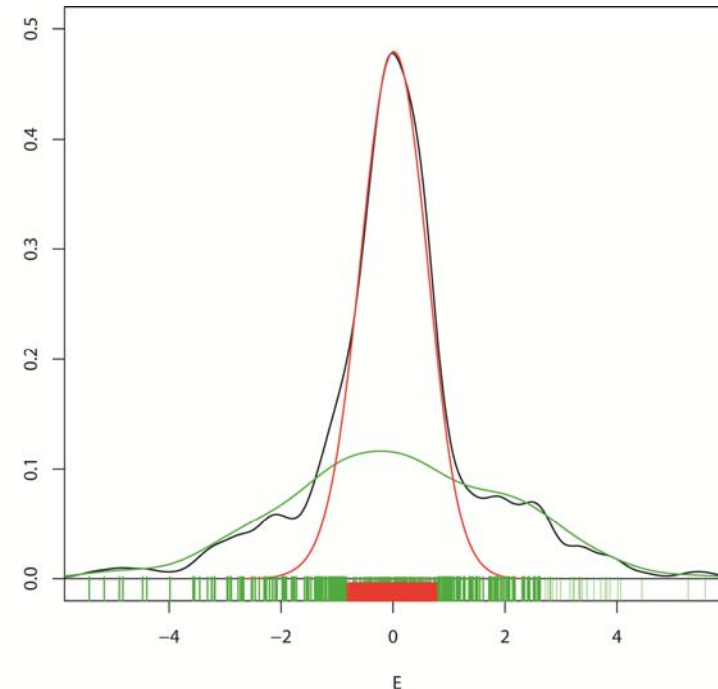
```
> McE$G
```

```
[1] 2
```

Partition et mélanges gaussiens

Différents groupes dont la moyenne est équivalente?

- Les groupes sont détectés
- Mais le taux de classification correcte est plus faible



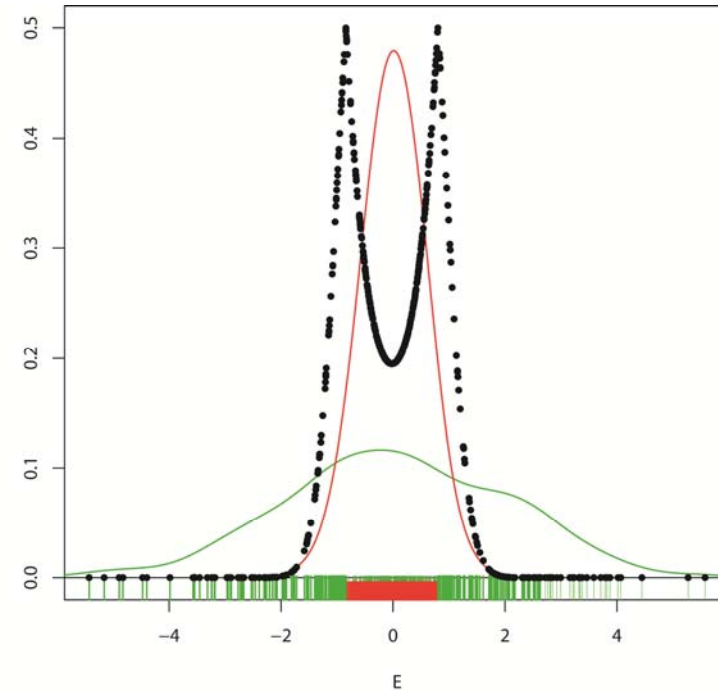
```
rug(E[which(McE$classification==1)],col=2)
rug(E[which(McE$classification==2)],col=3)

> classError(McE$class,gl(2,300))$errorRate
[1] 0.21
```


Partition et mélanges gaussiens

Différents groupes dont la moyenne est équivalente?

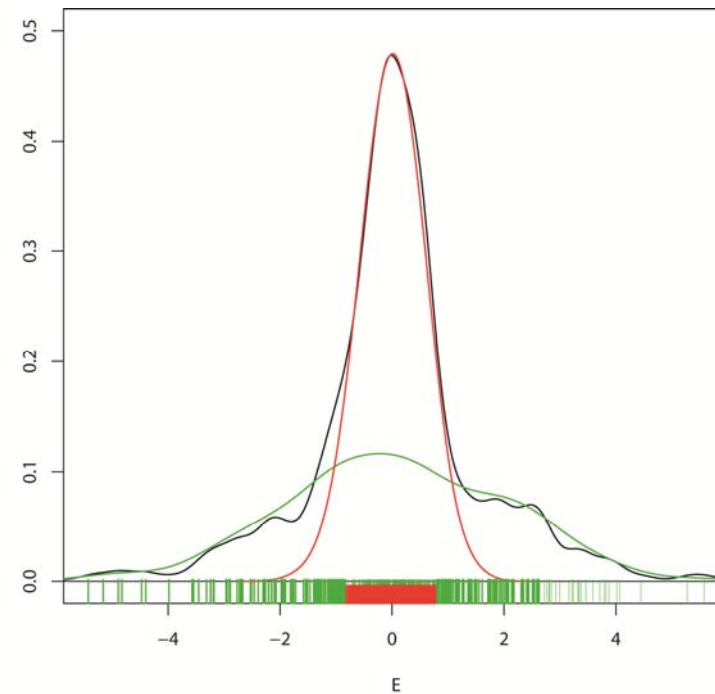
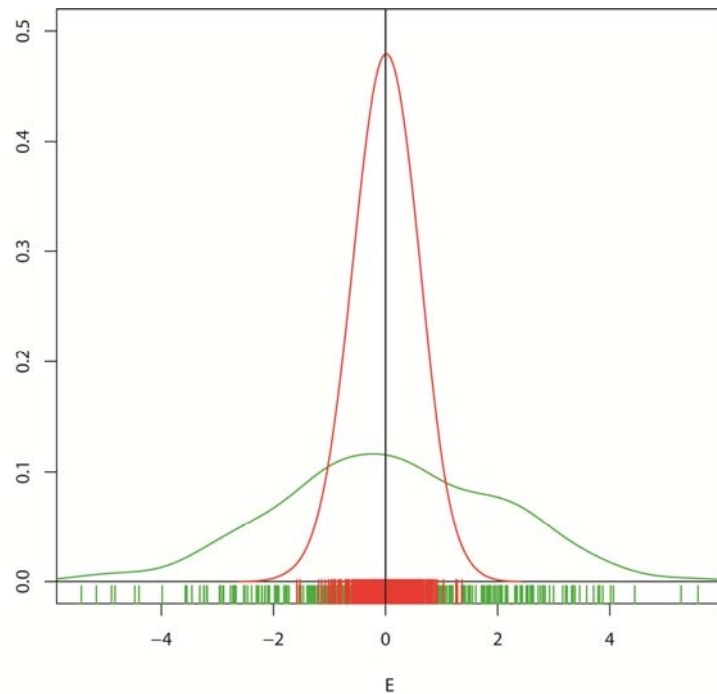
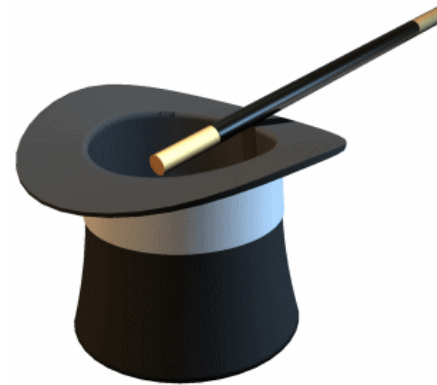
- Les groupes sont détectés
- Mais le taux de classification correcte est plus faible
- Et l'incertitude d'attribution plus forte



```
points(E, MCE$uncertainty, pch=20)
```

Partition et mélanges gaussiens

Mais quand même !



Dans les coulisses...

Mélanges gaussiens (Gaussian mixture models) :

Maximiser la vraisemblance d'une partition par une méthode de mélange (= «mixture») : attribution à un groupe $K \rightarrow$ une probabilité (τ_k)
(méthode de « classification » \rightarrow valeurs discrètes)

Dans les coulisses...

Mélanges gaussiens (Gaussian mixture models) :

Maximiser la vraisemblance d'une partition par une méthode de mélange (= «mixture») : attribution à un groupe K \rightarrow une probabilité (τ_k)
(méthode de « classification » \rightarrow valeurs discrètes)

$$l(\theta_k, \tau_k, z_{ik} | \mathbf{x}) = \sum_{i=1}^n \sum_{k=1}^G z_{ik} [\log \tau_k f_k(x_i | \theta_k)].$$

Modèle du groupe K \rightarrow θ_k

Probabilité qu'une observation appartienne au groupe K \rightarrow τ_k

Probabilité que l'observation i appartienne au groupe K \rightarrow z_{ik}

Fonction de densité du groupe K \rightarrow $f_k(x_i | \theta_k)$

Dans les coulisses...

Mélanges gaussiens (Gaussian mixture models) :

Maximiser la vraisemblance d'une partition par une méthode de mélange (= «mixture») : attribution à un groupe K \rightarrow une probabilité (τ_k)
(méthode de « classification » \rightarrow valeurs discrètes)

$$l(\theta_k, \tau_k, z_{ik} | x) = \sum_{i=1}^n \sum_{k=1}^G z_{ik} [\log \tau_k f_k(x_i | \theta_k)].$$

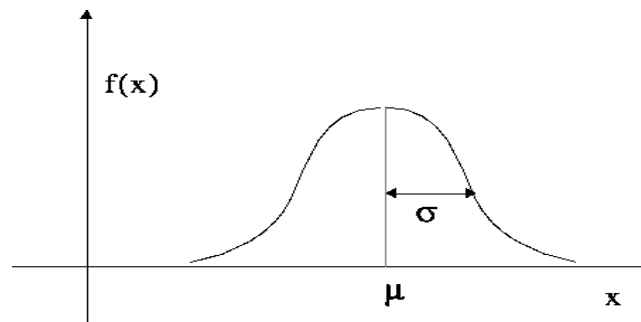
Modèle du
groupe K

Probabilité qu'une observation
appartienne au groupe K

Probabilité que l'observation i
appartienne au groupe K

Fonction de densité
du groupe K

Le modèle θ est défini
par la moyenne (μ) et
la variance (σ)



Dans les coulisses...

Algorithme EM « Expectation-Maximization »

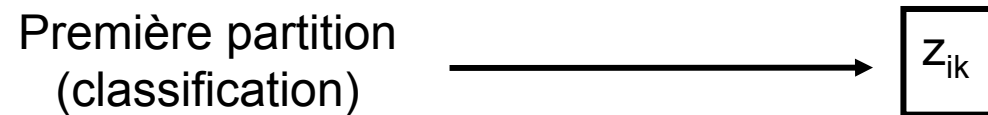
Optimisation du maximum de vraisemblance de la partition des données

$$\sum_{i=1}^n \sum_{k=1}^G z_{ik} [\log \tau_k f_k(x_i | \theta_k)].$$

Dans les coulisses...

Algorithme EM « Expectation-Maximization »

Optimisation du maximum de vraisemblance de la partition des données



$$\sum_{i=1}^n \sum_{k=1}^G z_{ik} [\log \tau_k f_k(x_i | \theta_k)].$$

Dans les coulisses...

Algorithme EM « Expectation-Maximization »

Optimisation du maximum de vraisemblance de la partition des données

Première partition
(classification)



z_{ik}

τ_k, μ_k, σ_k

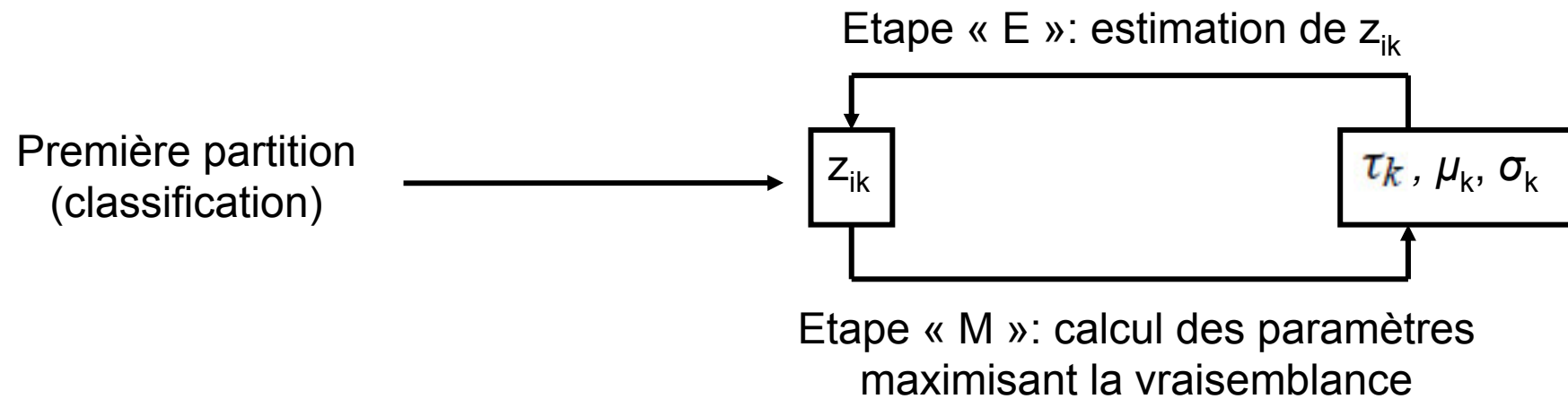
Etape « M »: calcul des paramètres
maximisant la vraisemblance

$$\sum_{i=1}^n \sum_{k=1}^G z_{ik} [\log \tau_k f_k(x_i | \theta_k)].$$

Dans les coulisses...

Algorithme EM « Expectation-Maximization »

Optimisation du maximum de vraisemblance de la partition des données



$$\sum_{i=1}^n \sum_{k=1}^G z_{ik} [\log \tau_k f_k(x_i | \theta_k)].$$

Répété jusqu'à la convergence

Dans les coulisses...


Sélection du modèle par critère Bayésien (BIC $\sim 2 \log(\text{Bayes factor})$)

Dans les coulisses...

Sélection du modèle par critère Bayésien (BIC $\sim 2 \log(\text{Bayes factor})$)

$$\text{BIC} \equiv 2 \log \text{lik}_M(x, \theta_k^*) - (\#_{\text{params}})M \log(n)$$

Vraisemblance maximisée
du mélange gaussien du modèle M

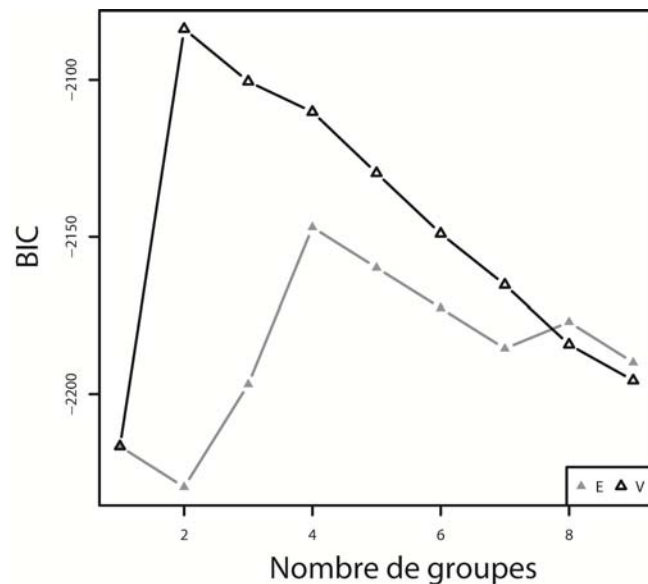


Dans les coulisses...

Sélection du modèle par critère Bayésien (BIC $\sim 2 \log(\text{Bayes factor})$)

$$\text{BIC} \equiv 2 \log \text{lik}_M(\mathbf{x}, \theta_k^*) - (\#_{\text{params}})M \log(n)$$

Vraisemblance maximisée
du mélange gaussien du modèle M



Sélection du modèle M

Sélection du nombre de groupes

```
plot(McE) #Apparait après validation dans  
la barre des tâches du terminal.
```

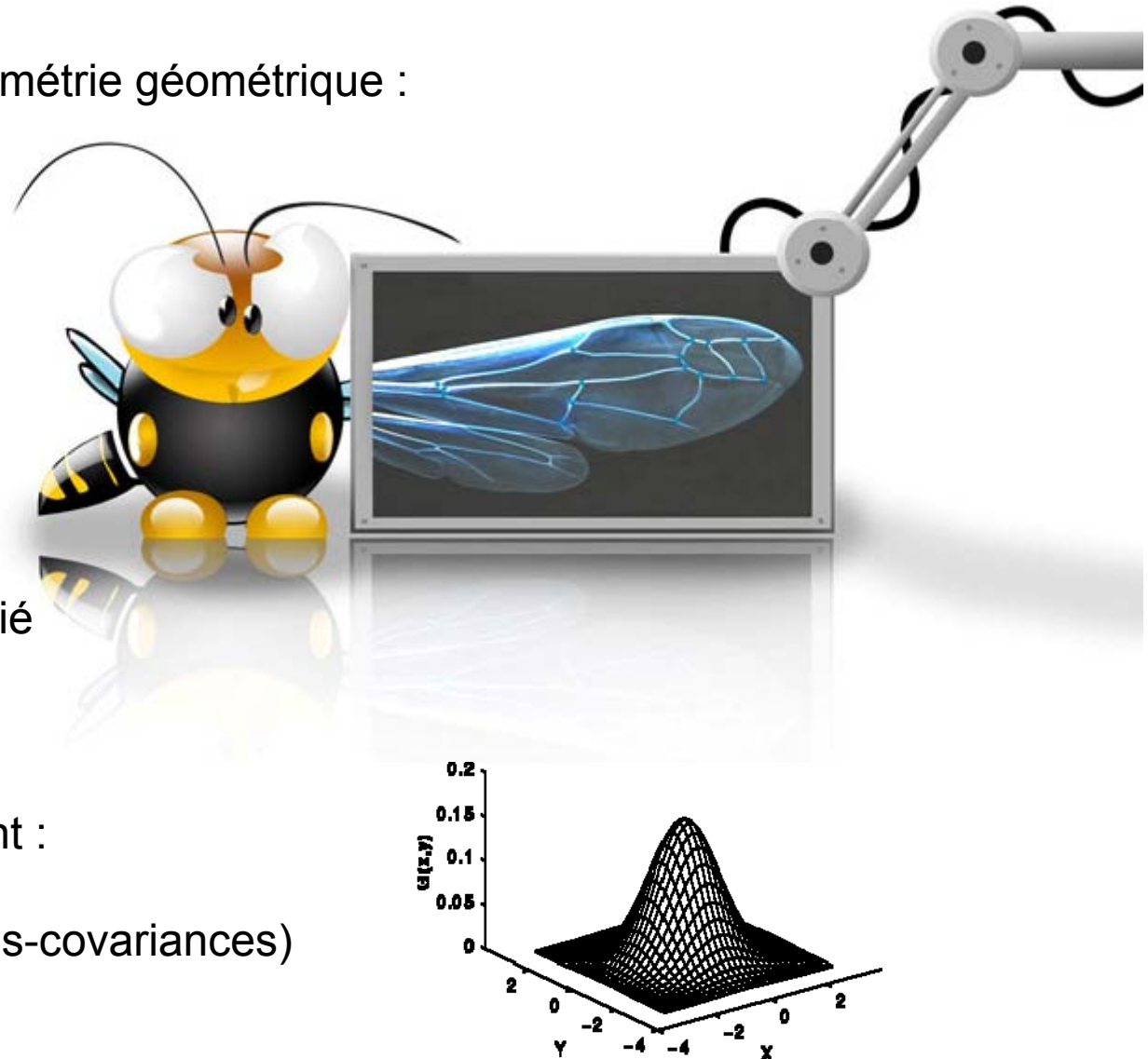
Et avec plus de dimensions...

Mesure des ailes en morphométrie géométrique :

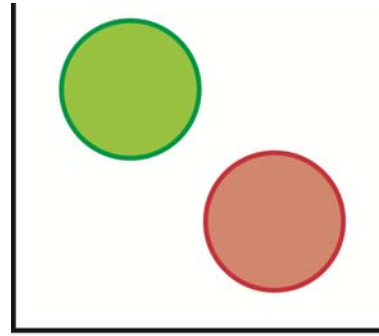
38 variables

Mélanges gaussiens :
Même principe qu'en uni-varié

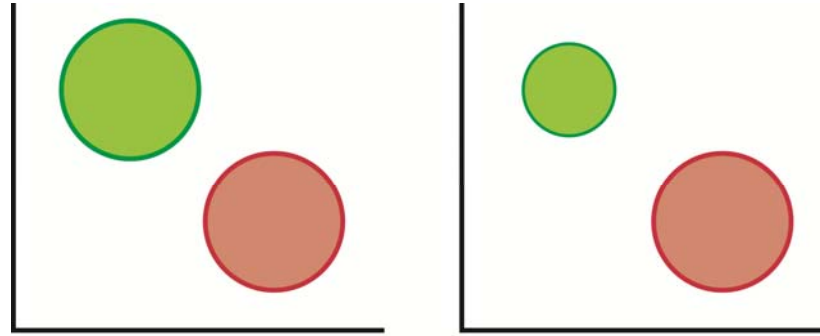
Seul les modèles θ_k changent :
 μ_k (scalaire) $\rightarrow \mu_k$ (vecteur)
 $\sigma_k \rightarrow \Sigma_k$ (Matrice de variances-covariances)



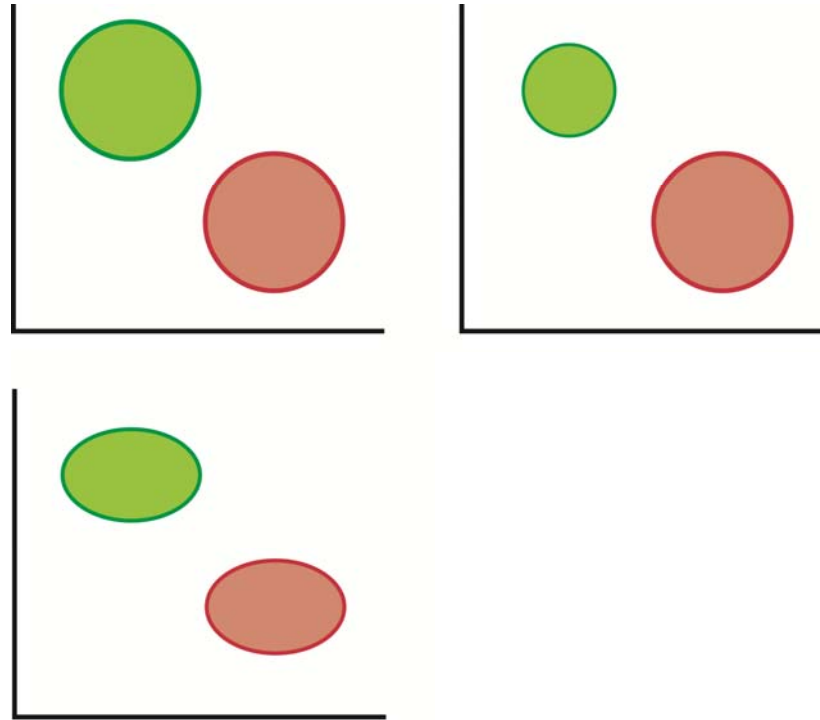
Et avec plus de dimensions...



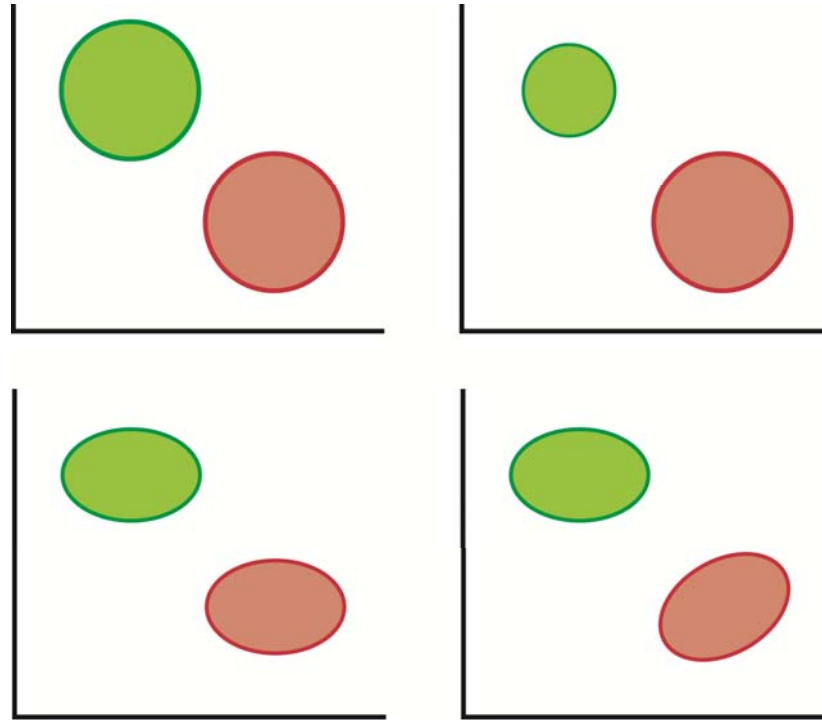
Et avec plus de dimensions...



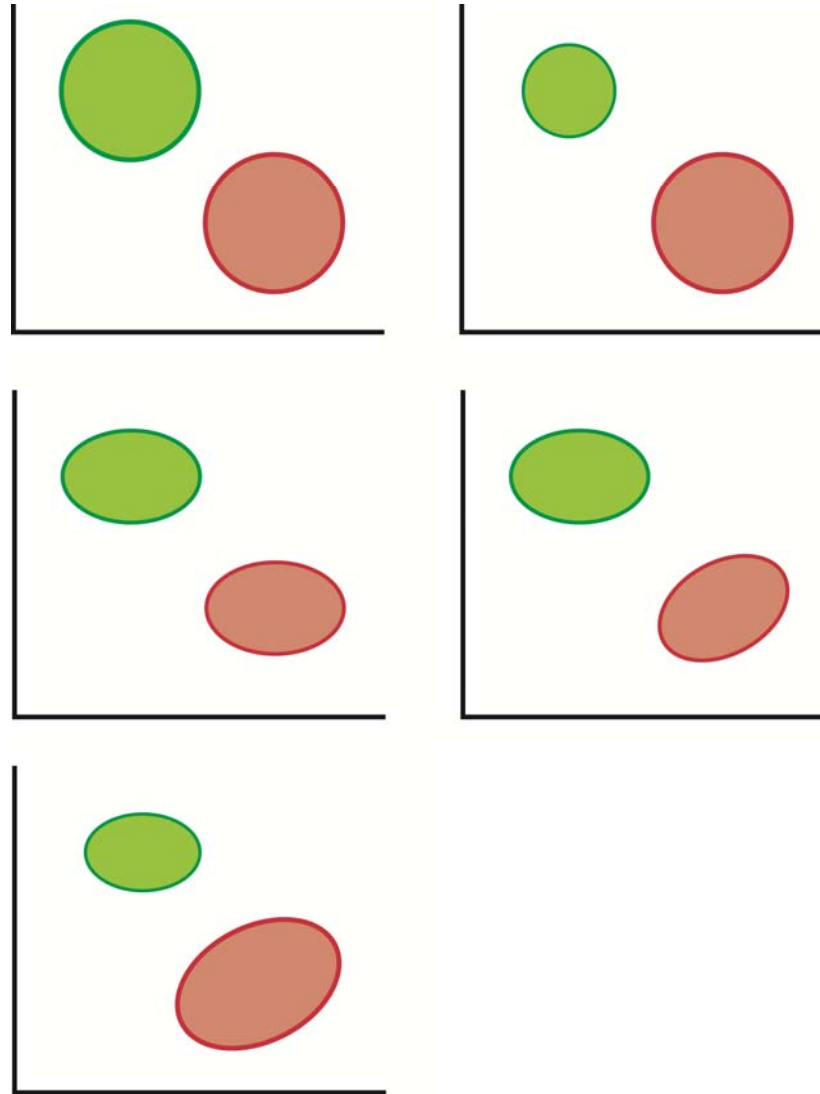
Et avec plus de dimensions...



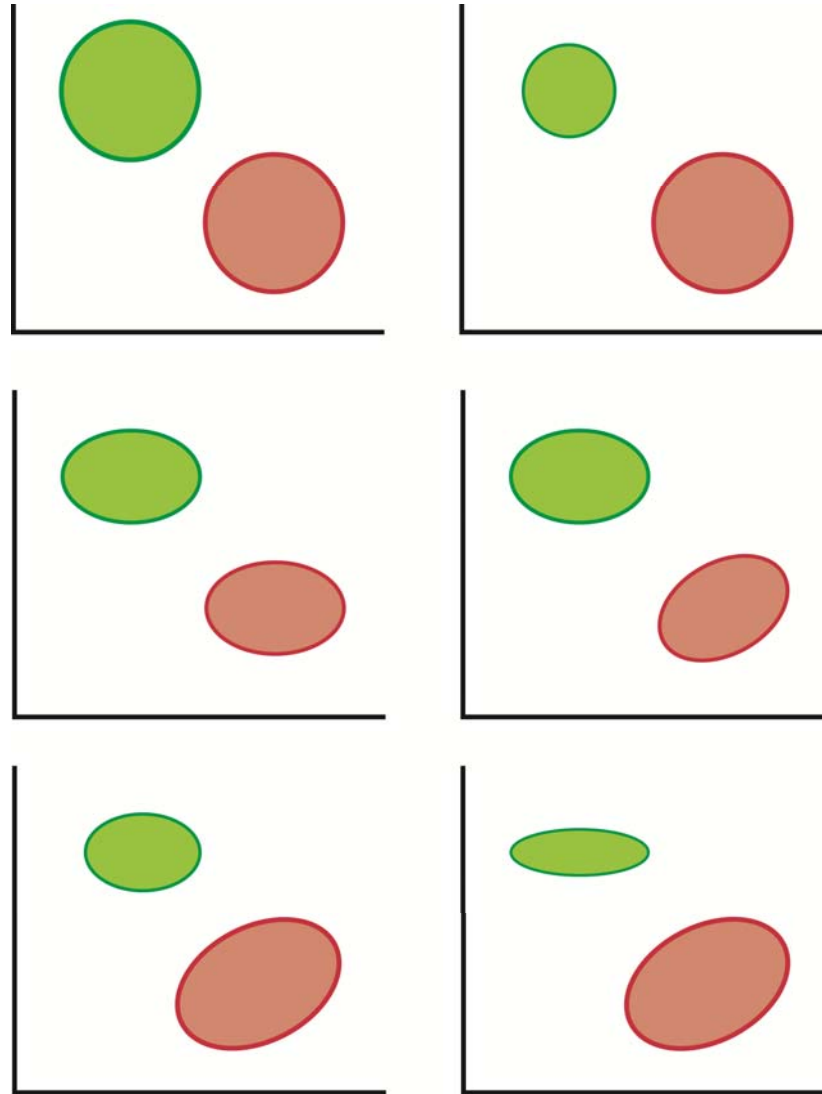
Et avec plus de dimensions...



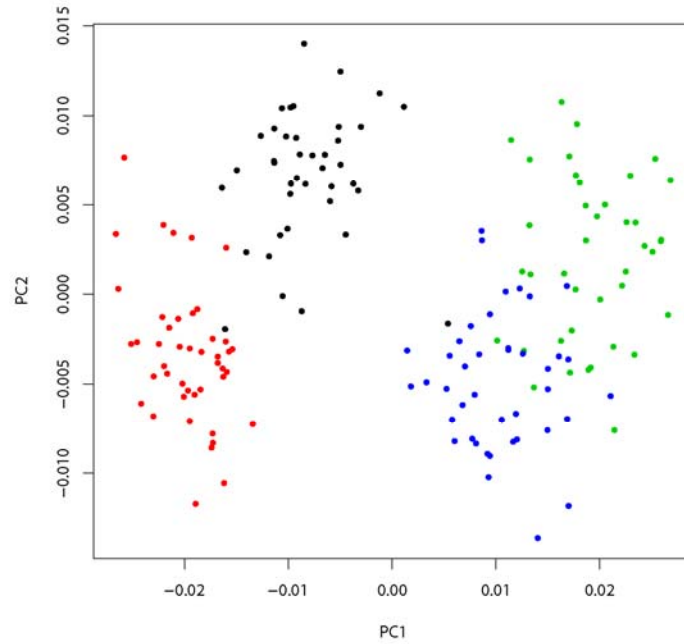
Et avec plus de dimensions...



Et avec plus de dimensions...

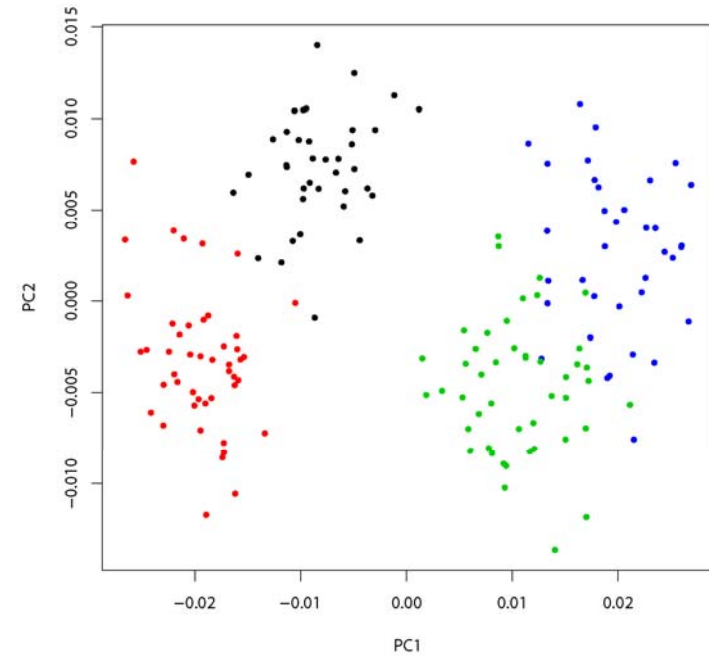
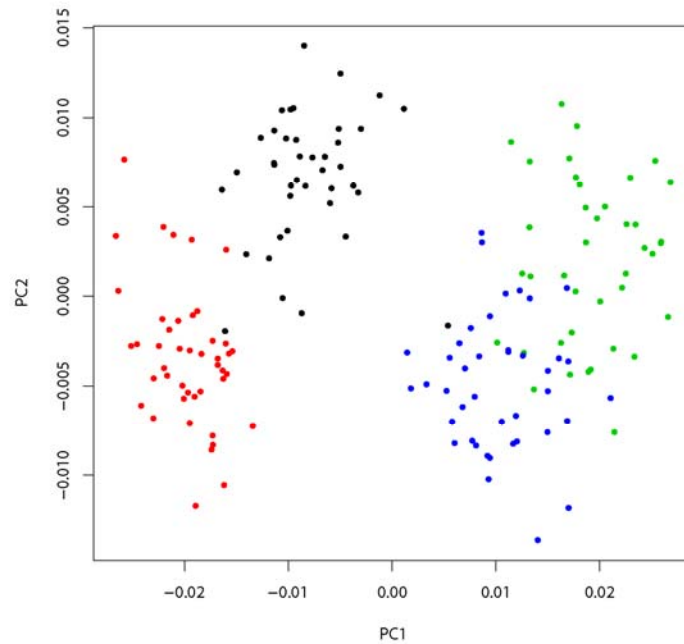


Et avec plus de dimensions...



Et avec plus de dimensions...

Correspond aux groupes biologiques (erreur de classification ~ 5%)



Groupes	1	2	3	4
Sp1 ouv.	37	2	1	
Sp1 reine		44		
Sp2 ouv.			5	35
Sp2 reine			40	1

Warnings et limites de la méthode

Nécessité de données de plein rang

Messages d'avis :

```
1: In summary.mclustBIC(Bic, data, G = G, modelNames = modelNames) :  
  best model occurs at the min or max # of components considered  
2: In Mclust(Vcaste2$scores) :  
  optimal number of clusters occurs at min choice
```

Tailles des échantillons

Nombre de groupe maximal

Temps de calcul...

Autres fonctions de mclust

Partition hiérarchique:

- hc

Estimation de densité:

- densityMclust

Analyses discriminantes

- MclustDA(trainData, trainClass)

- cv.MclustDA(ObjMclustDA)

- predict.MclustDA(ObjMclustDA,otherdata)

Plot

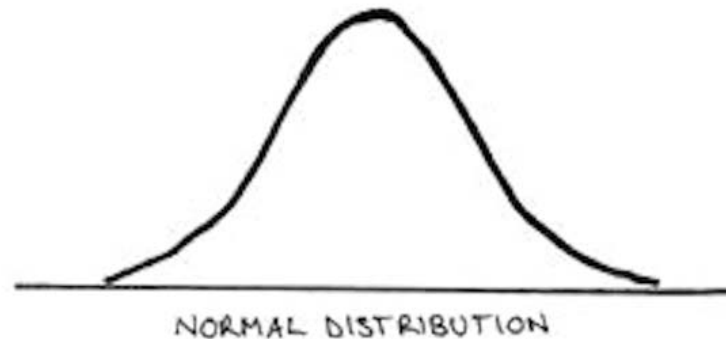
- mclust2Dplot / surfacePlot # plot pour données bivariées

- coordProj / MclustDR # plot pour données multivariées

Autres:

- classError (McObj\$classification,groups)

Merci de votre attention !



Banfield, J. D., & Raftery, A. E. 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**: 803–821.

Fraley, C., & Raftery, A. E. 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal* **41**: 578–588.

Fraley, C., & Raftery, A. 2009. mclust: Model-based clustering/normal mixture modeling. *R package version 3*.