Clustering divisif monothétique. Le package divclust

Marie Chavent, Marc Fuentes

University of Bordeaux, France
Inria Bordeaux Sud-Ouest, CQFD Team

Introduction

La fonction principale `divclust` :

- Méthode de classification descendant hiérarchique et monothétique.
- Optimise le même critère que la classification ascendante hiérarchique de WARD.
- Critère différent de la classification descendante hiérarchique de [Kaufman et Rousseeuw, 1990] implémentée dans la fonction `diana` du package **cluster**.
- Implémente la méthode DIVCLUS-T proposée par [Chavent et al., 2007].

  $\hookrightarrow$ Ajout du cas des données mixtes (quantitatives et qualitatives) non traitées dans l'article.

https://github.com/chavent/divclust

Outline

Un petit exemple

```
library(divclust)
data(protein)
treeDIV <- divclust(protein)
sum(treeDIV$height)
```

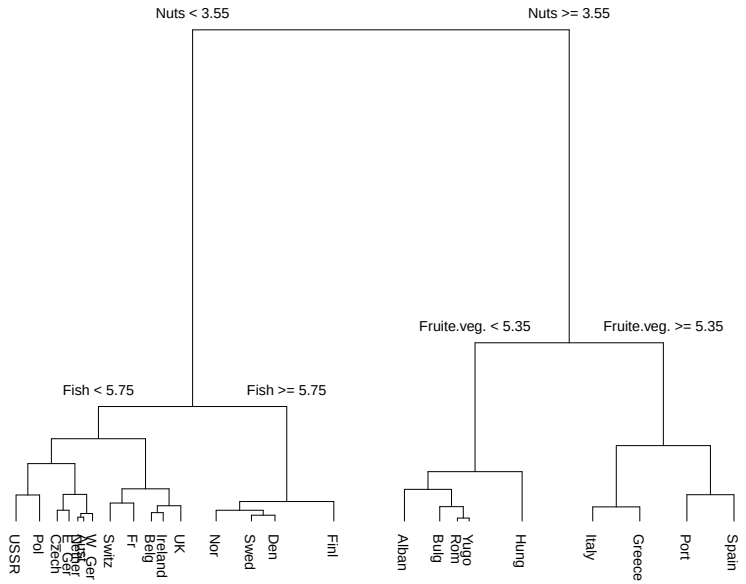
```
## [1] 9
```

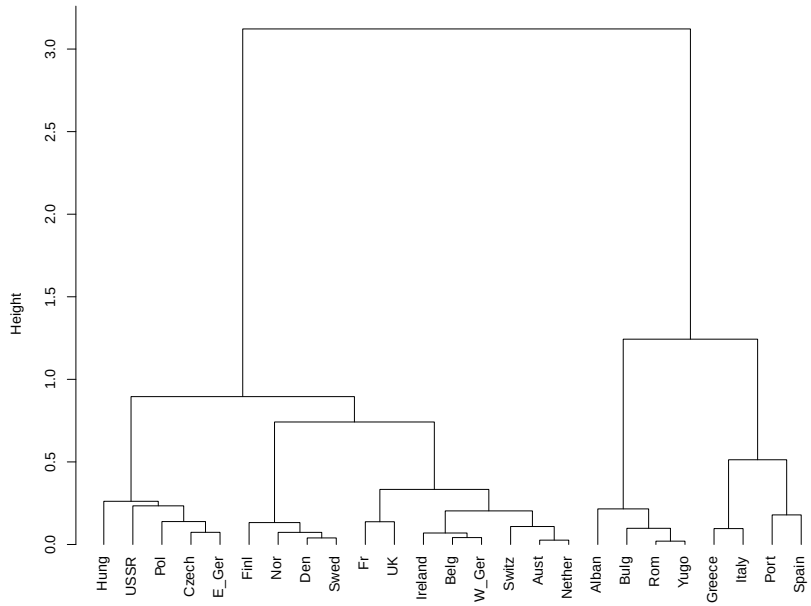
```
plot(treeDIV)
```

```
n <- nrow(protein)
zprotein <- scale(protein)*sqrt(n/(n-1))
treeWard <- hclust(dist(zprotein)^2/(2*n),
                   method="ward.D")
sum(treeWard$height)
```

```
## [1] 9
```

```
plot(treeWard,main="",sub="",xlab ="",hang=-1)
```





Un petit exemple

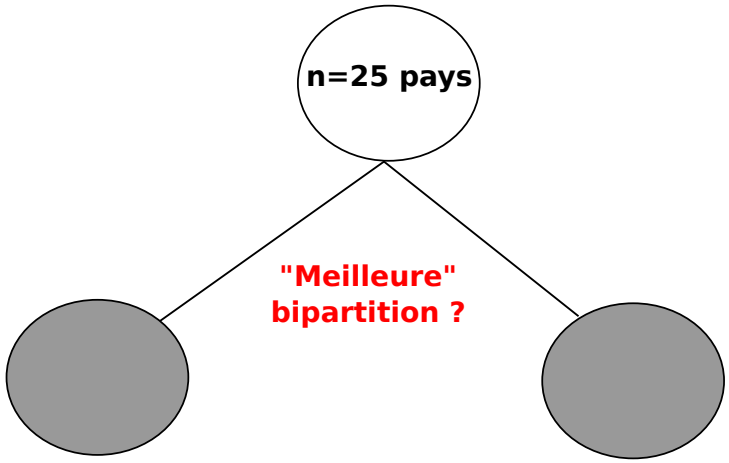
⇒ Comparaison des proportion d'inertie expliquée par les partitions de 2 à 10 classes :

	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
divclust	0.37	0.51	0.59	0.65	0.71	0.76	0.79	0.82	0.84
Ward	0.35	0.49	0.58	0.67	0.72	0.76	0.79	0.82	0.84

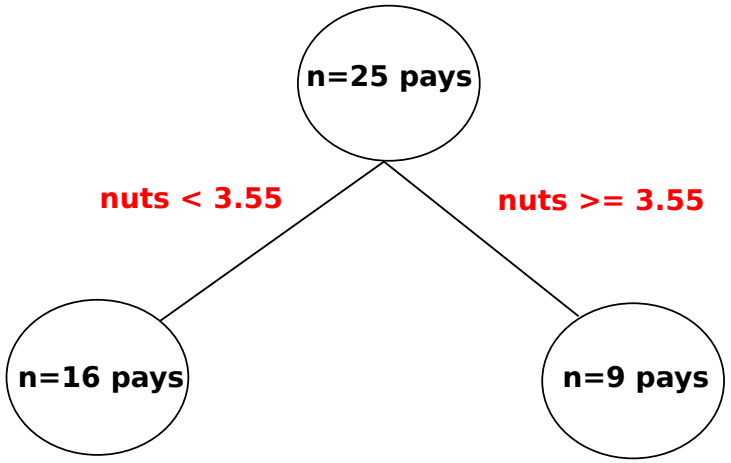
⇒ DIVCLUS-T utilise la distance Euclidienne sur données standardisées.

⇒ Quelles distances pour des données qualitatives ou mixtes ?

Un petit exemple

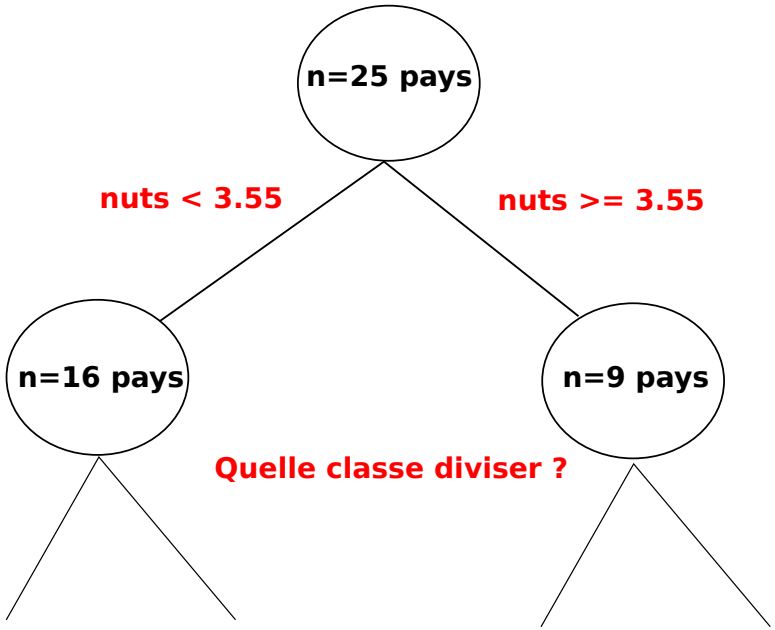


Un petit exemple

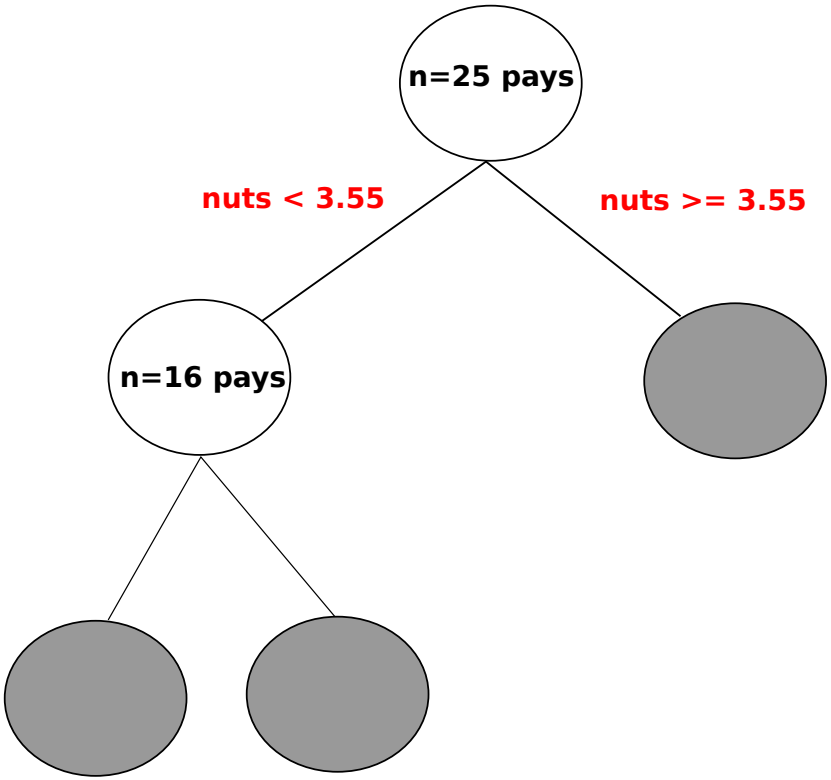


Partition monothétique
de plus petite
inertie intra-classe

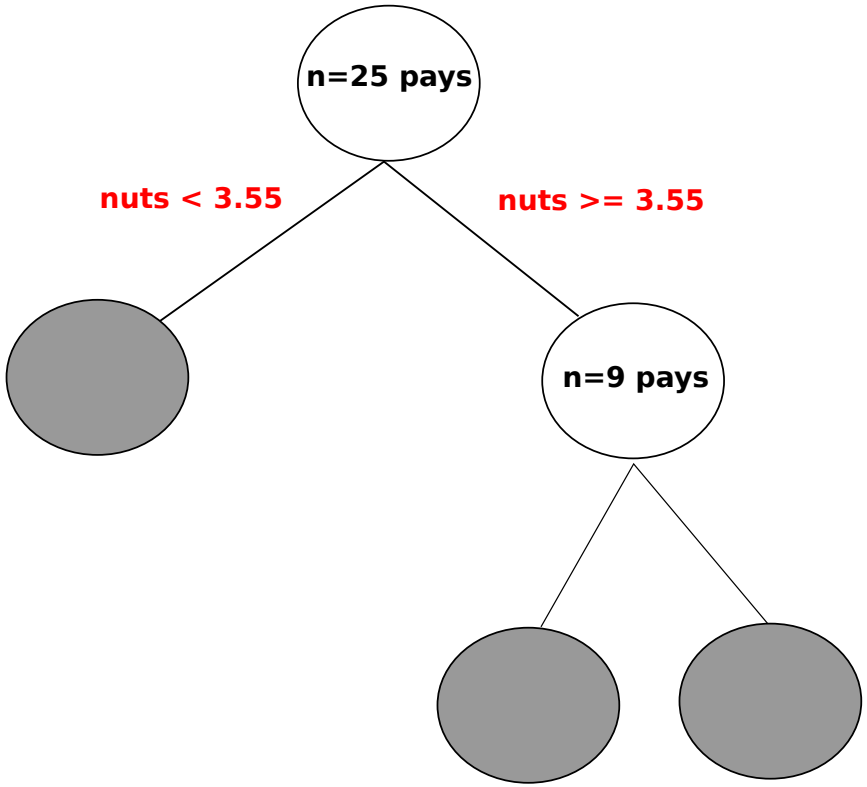
Un petit exemple



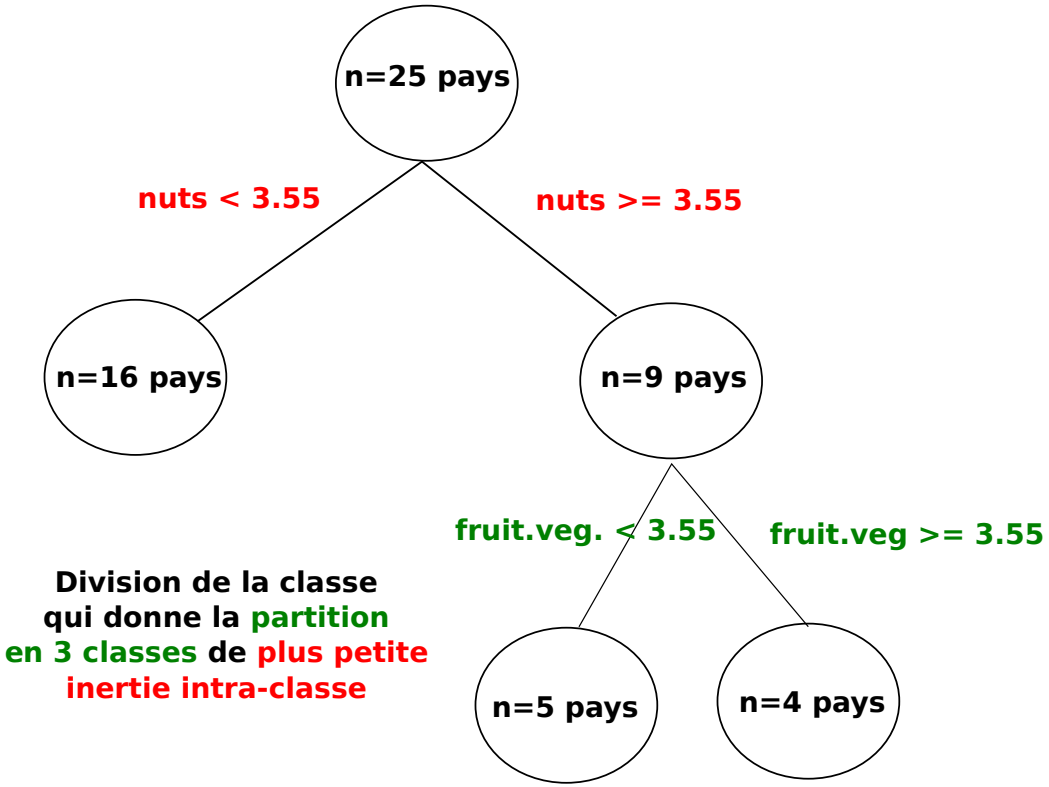
Un petit exemple



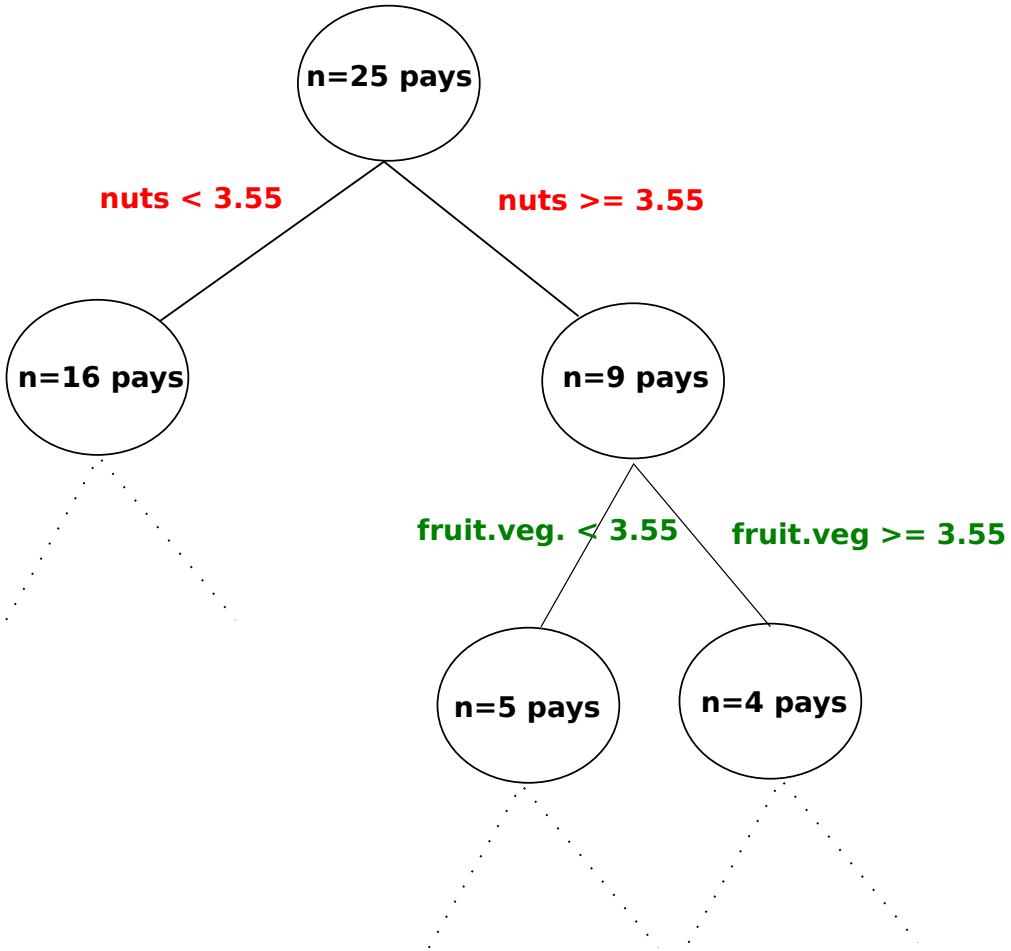
Un petit exemple



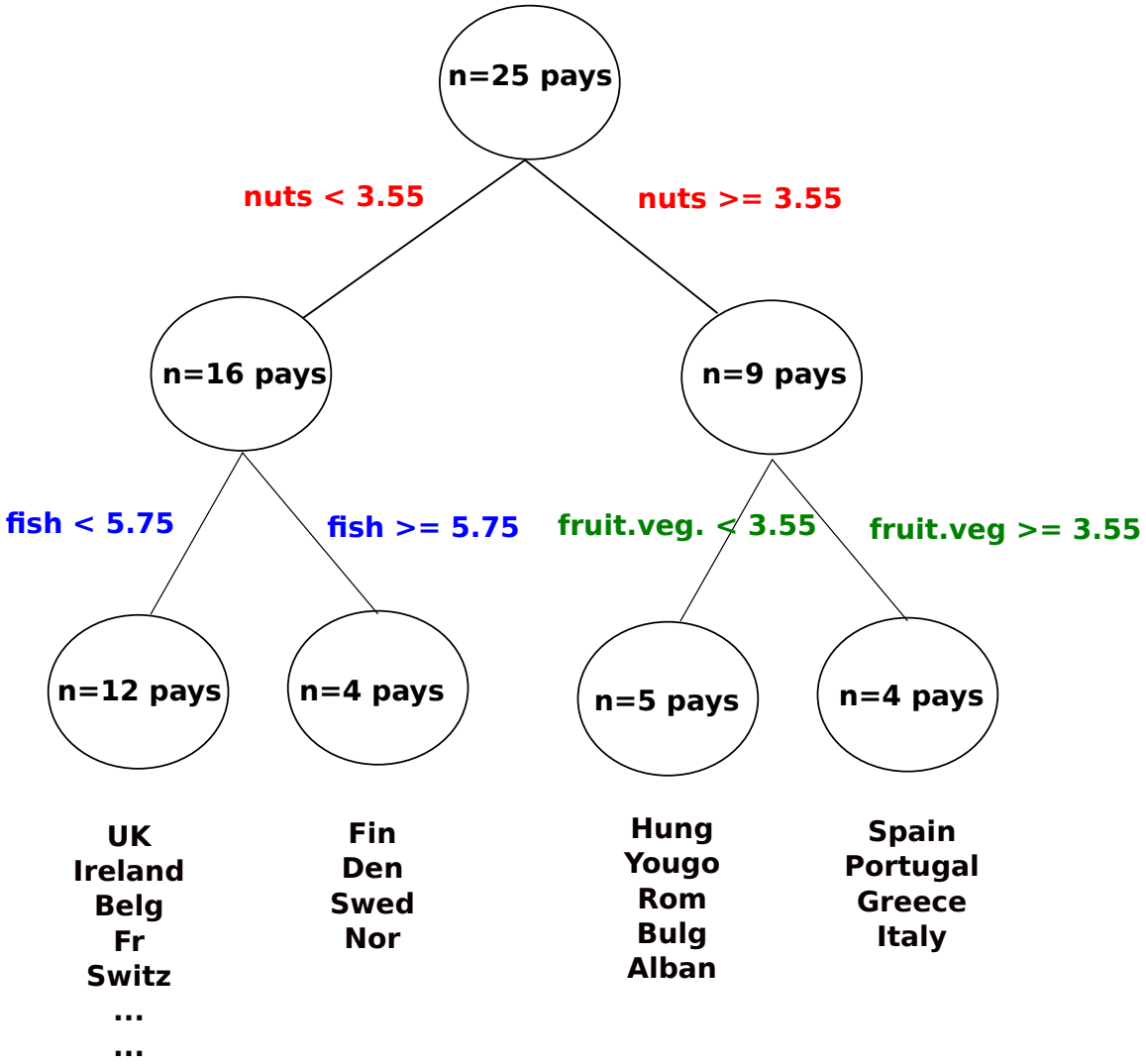
Un petit exemple



Un petit exemple



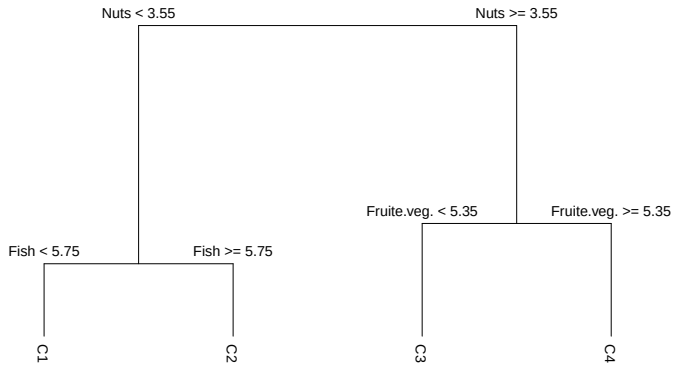
Un petit exemple



Un petit exemple

Haut du dendrogramme (3 divisions).

```
treeDIV_4 <- divclust(protein,K=4)
plot(treeDIV_4)
```

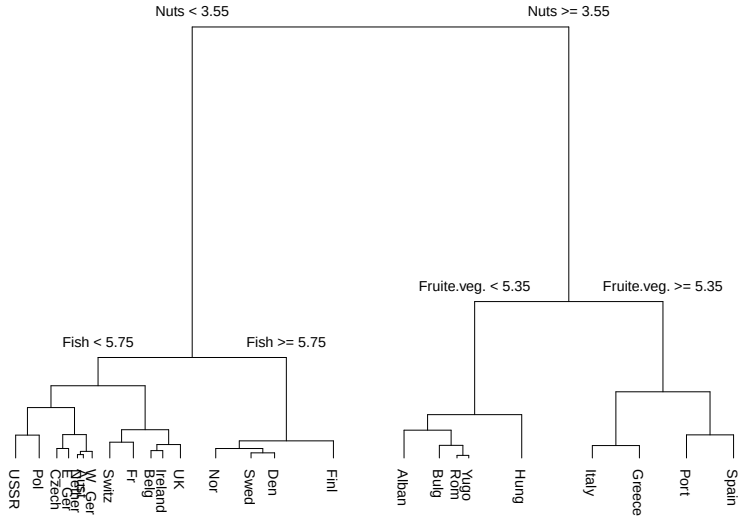



```
treeDIV_4$clusters[[2]]
```

```
## [1] "Den"  "Finl" "Nor"  "Swed"
```

Dendrogramme complet.

```
treeDIV <- divclust(protein)
plot(treeDIV)
```




```
P4 <- cutreediv(treeDIV,K=4)
P4$clusters[[2]]
```

```
## [1] "Den"  "Finl" "Nor"  "Swed"
```

Outline

L'algorithme DIVCLUS-T

En entrée :

- une matrice de donnée $\mathbf{X}$ de dimension $n \times p$ quantitative, qualitative ou mixte .
- un nombre K indiquant le nombre de classes de la dernière partition (par défaut la partition des singletons).

Répétition des deux étapes suivantes

1. division d'une classe en deux : la bipartition doit optimiser un critère W. L'énumération complète est évitée par la contrainte monothétique.
2. choix de la classe à diviser : la nouvelle partition doit optimiser le critère W.

En sortie : une hiérarchie indicée (un dendrogramme) qui se lit comme un arbre de décision.

Division d'une classe

Le principe

Sélectionner parmi toutes les bipartitions induites par toutes les questions binaires (définies sur une unique variable) possibles, celle de plus petite inertie intra-classe W (définie sur toutes les variables).

Pour une variable quantitative X^j, une question binaire est notée $[X^j \leq c]$?

$\Rightarrow$ nombre infini de questions binaires possibles sur X^j mais au plus $n_\ell - 1$ bipartitions d'une classe C_ℓ à n_ℓ observations.

$\Rightarrow$ tri des valeurs des observations de X^j et les valeurs de coupures sont les milieux de deux observations consécutives.

Division d'une classe

Pour une variable qualitative X^j, une question binaire est notée $[X^j \in \mathcal{C}]$?

$\Rightarrow$ le nombre de questions binaires possibles sur X^j est $2^{m_j-1}-1$ où m_j est le nombre de modalités de X^j.

$\Rightarrow$ nombre de bipartitions exponentiel avec le nombre de modalités

Finallement, le nombre maximum de bipartitions est :

$$p_1(n_\ell - 1) + \sum_{j=1}^{p_2} 2^{m_j-1} - 1.$$

Choix de la classe à diviser

Le principe

Choisir la classe $C_\ell = A_\ell \cup \bar{A}_\ell$ de la partition P_k qui permet d'obtenir la partition $P_{k+1} = \{C_1, \ldots, C_{\ell-1}, A_\ell, \bar{A}_\ell, C_{\ell-1}, \ldots, C_k\}$ de plus petite inertie intra-classe $W(P_{k+1})$.

L'inertie intra-classe est un critère additif :

$$W(P_k) = \sum_{\ell=1}^{k} I(C_\ell)$$

$\Rightarrow$ Equivalent de choisir C_ℓ qui maximise la variation de l'inertie :

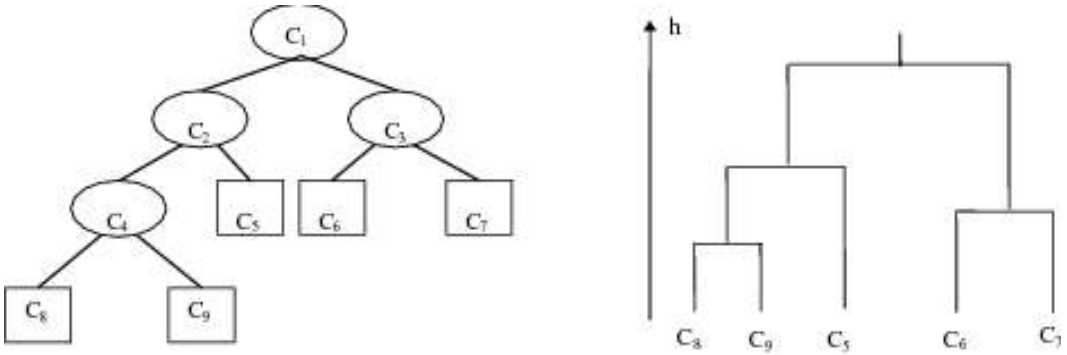
$$h(C_\ell) = I(C_\ell) - I(A_\ell) - I(\bar{A}_\ell).$$

Hauteur des classes dans le dendrogramme

Hiérarchie indicée comme celle de Ward par la variation de l'inertie

$$h(C_\ell) = I(C_\ell) - I(A_\ell) - I(\bar{A}_\ell)$$

$\Rightarrow$ si $K < n$, "haut" du dendrogramme complet.



Hauteur des classes dans le dendrogramme

⇒ La somme des hauteurs de l'arbre complet est l'inertie totale T.

```
#inertie totale
T <- sum(treeDIV$height)
T # nombre de variables quantitatives
```

```
## [1] 9
```

⇒ La somme des $K-1$ hauteurs est l'inertie inter-classe B de la partition en K classes.

```
#inertie inter-classe de la partition en 4 classes
B <- sum(treeDIV$height[1:3])
```

⇒ Pourcentages d'inertie expliquées des partitions du dendrogramme.

```
# pourcentage d'inertie explique de la partition en 4 classes
B/T*100
```

```
## [1] 59.2
```

Inertie pour des données qualitatives ou mixtes

Inertie d'une du matrice des données $\mathbf{Z}$ de dimension $n \times p$:

$$I(\mathbf{Z}) = \frac{1}{n} \sum_{i=1}^{n} d_M(\mathbf{z}_i, g)^2$$

1. Recodage des données qualitatives par les indicatrices des modalités.
2. Métrique M est la matrice diagonale des poids des colonnes :
   - ↪ les colonnes quantitatives sont pondérées par l'inverse de la variance empirique.
   - ↪ les colonnes des indicatrices sont pondérées par l'inverse de la fréquence de la modalités.

Inertie pour des données qualitatives ou mixtes

$\Rightarrow$ Si toutes les variables sont quantitatives, d_M est la distance sur données standardises et $I(\mathbf{Z}) = p$.

$\Rightarrow$ Si toutes les variables sont qualitatives, d_M est une distance de type χ^2 et $I(\mathbf{Z}) = m - p$ où m est le nombre de modalités.

$\Rightarrow$ Si les données sont mixtes, $I(\mathbf{Z}) = p_1 + m - p_2$ où p_1 et p_2 sont les nombres de variables quantitatives et qualitatives.

En pratique, dans la fonction `divclust` du package utilise :

$\hookrightarrow$ la matrice $\mathbf{X}$ des données brutes pour construire les questions binaires.

$\hookrightarrow$ la matrice $\mathbf{Z}$ des de toutes les composantes principales d'une ACP sur données mixtes pour calculer l'inertie avec $\mathbf{M} = I$.

Outline

Un exemple de données mixtes

Les données gironde caractérisent les conditions de vies en Gironde.

↪ 542 communes sont décrites par 27 variables séparées en 4 groupes (Employment, Housing, Services, Environment).

↪ Le groupe housing est mixte avec 3 variables quantitatives et deux variables qualitatives.

```
library(divclust)
data(gironde)
housing <- gironde$housing
```

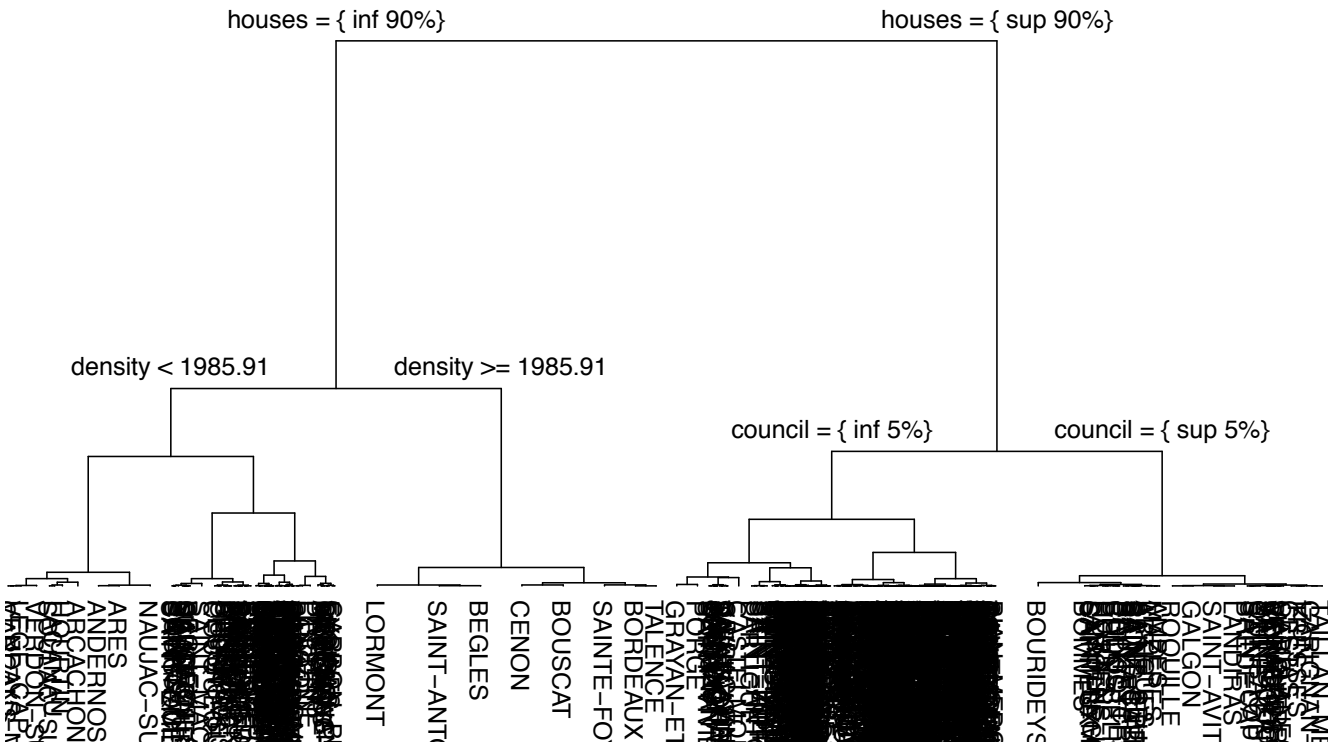
```
head(housing)
```

```
##          density primaryres owners   houses council
## ABZAC      131.7       88.8   64.2  inf 90%  sup 5%
## AILLAS      21.2       87.5   77.1  sup 90%  inf 5%
## AMBARES    532.0       94.9   65.7  inf 90%  sup 5%
## AMBES      101.2       93.8   66.5  sup 90%  sup 5%
## ANDERNOS   551.9       62.1   71.5  inf 90%  inf 5%
## ANGLADE     63.8       81.0   80.5  sup 90%  inf 5%
```

L'arbre complet

```
treeHousing <- divclust(housing)
```

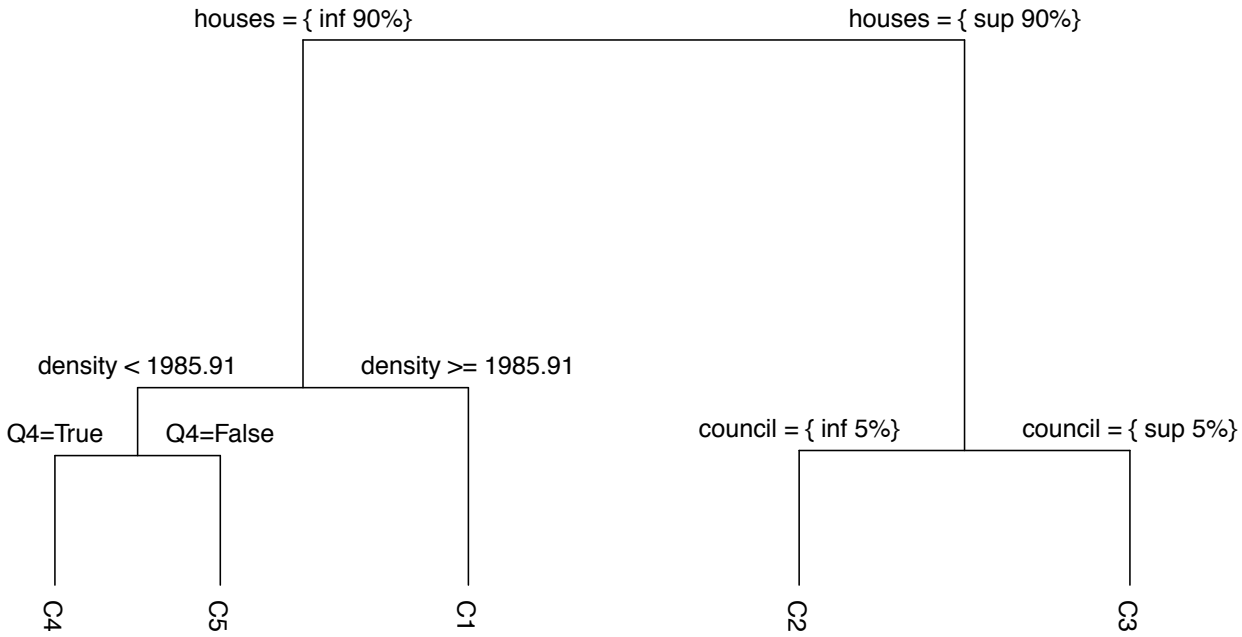
```
plot(treeHousing)
```



La partition en 5 classes

```
treeHousing_5 <- divclust(housing,K=5)
```

```
plot(treeHousing_5,nqbin=4) # Nombre de questions binaires  afficher
```



Ou encore avec la fonction `cutreediv` :

```
P_5_div <- cutreediv(treeHousing,K=5)
```

Quelques sorties

```
##
## Call:
## cutreediv(tree = treeHousing, K = 5)
##
##
##       name              description
## [1,] "$clusters"       "list of observations in each cluster"
## [2,] "$description"    "monothetic description of each cluster"
## [3,] "$which_cluster" "cluster memberships"
## [4,] "$B"              "the proportion of explained inertia"
```

```
  # liste des communes de la classe C4
  P_5_div$clusters$C4
```

```
##  [1] "ANDERNOS"           "ARCACHON"           "ARES"               "CARCANS"
##  [5] "HOURTIN"            "LACANAU"            "LEGE-CAP-FERRET"    "NAUJAC-SUR-MER"
##  [9] "SOULAC-SUR-MER"     "VENDAYS-MONTALIVET" "VERDON-SUR-MER"
```

```
  # Description monothetique de la classe C4
  P_5_div$description$C4
```

```
## [1] " houses = { inf 90%} , density = [-Inf ; 1985.91[ , primaryres = [-Inf ; 68.06["
```

```
  # Proportion d'inertie explique par P_5
  P_5_div$B
```

```
## [1] 0.693
```

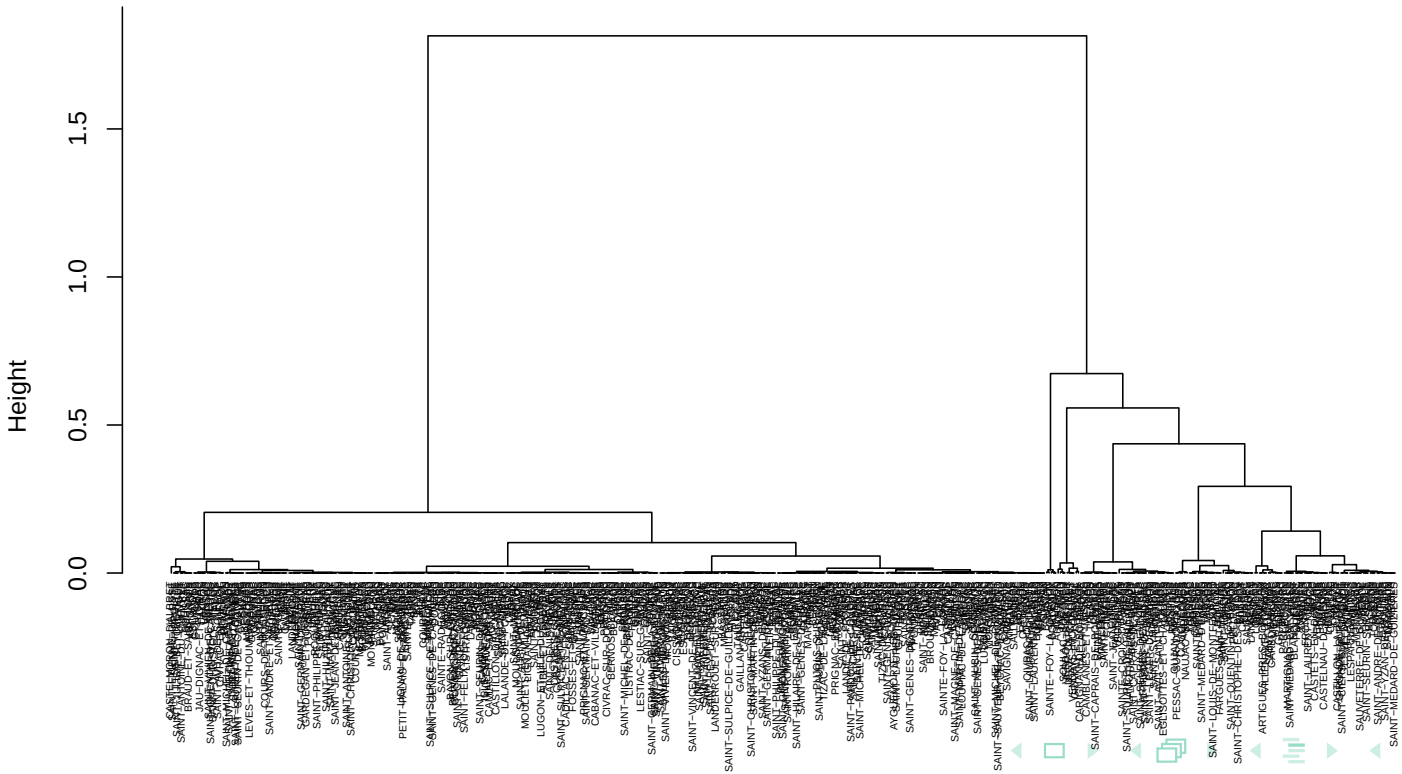
Ward sur données mixtes

⇒ Toutes les coordonnées factorielles de l' ACP mixte :

```
library(PCAmixdata)
X_quanti <- split_mix(housing)$data_quanti
X_quali <- split_mix(housing)$data_quali
zhousing <- PCAmix(X_quanti,X_quali,ndim=5,graph=FALSE)$ind$coord
```

⇒ Hiérarchie de Ward indicé par la variation de l'inertie.

```
treeWard <-  hclust(dist(zhousing)^2/(2*nrow(zhousing)),method="ward.D")
```



Ward sur données mixtes

⇒ Comparaison des partitions en $k = 5$ classes :

```
P_5_ward <- cutree(treeWard,5)
P_5_div<- P_5_div$which_cluster
library(ClustOfVar)
#Indice de Rand entre les deux partitions
rand(P_5_div,P_5_ward)

## [1] 0.981
```

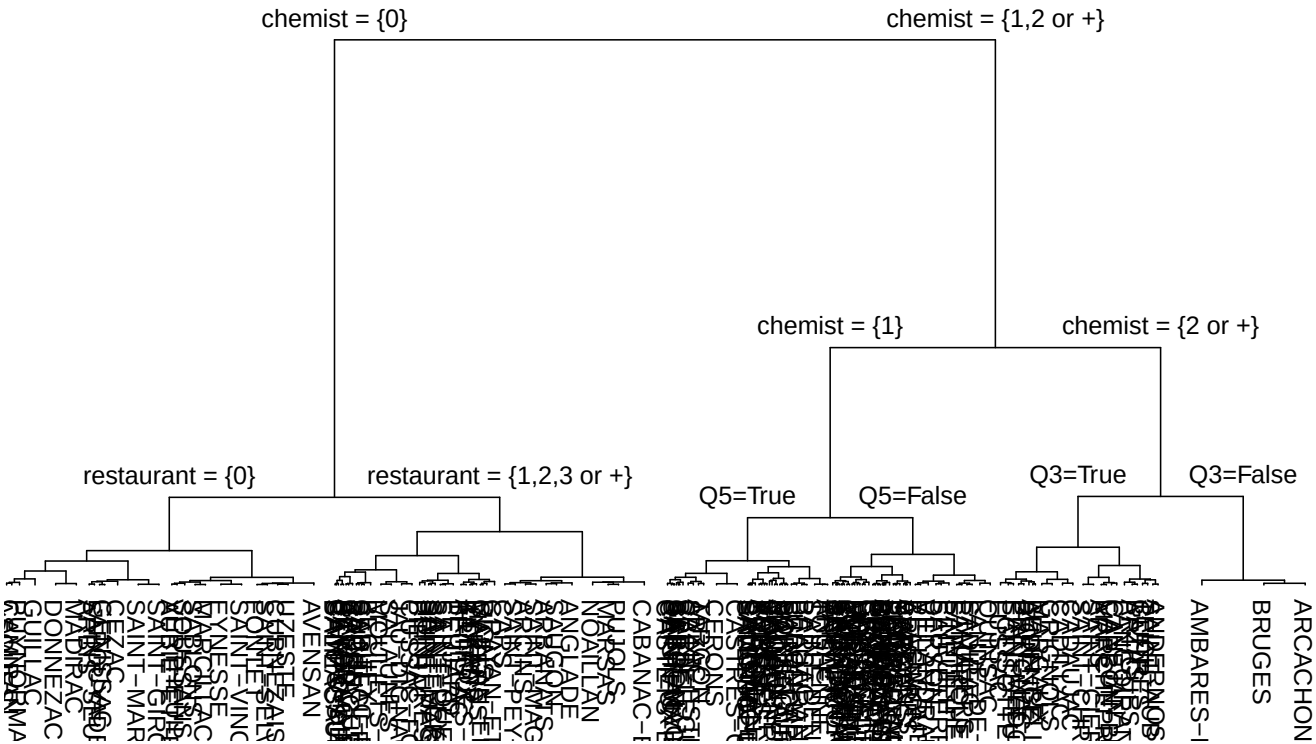
⇒ Comparaison des proportion d'inertie expliquée:

	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
divclust	0.21	0.28	0.34	0.39	0.41	0.44	0.45	0.46	0.47
Ward	0.20	0.28	0.34	0.39	0.42	0.44	0.46	0.47	0.48

Et pour des données qualitatives...

```
services <- gironde$services
treeServices <- divclust(services)
plot(treeServices,nqbin=5)
```

```
## Q5 :  doctor = {3 or +,0} or doctor = {1 to 2}
## Q3 :  nursery = {0} or nursery = {1 or +}
```



Tenir compte de l'ordre des modalités

⇒ Imposer un ordre sur les modalités.

```
#ordonner les modalites
services$doctor <- ordered(services$doctor)
class(services$doctor)

## [1] "ordered" "factor"

treeServices <- divclust(services)
plot(treeServices,nqbin=5)

## Q5 :  doctor = {0,1 to 2} or doctor = {3 or +}
## Q3 :  nursery = {0} or nursery = {1 or +}
```

⇒ Réduction du nombre de questions binaires dans le cas qualitatif.

Références I

M. Chavent, Y. Lechevallier, and 0. Briant

DIVCLUS-T: A monothetic divisive hierarchical clustering method.
Computational Statistics and Data Analysis 52 (2007) 687-701.

L. Kaufman, P. J. Rousseeuw

Divisive Analysis (Program DIANA).
Finding Groups in Data: An Introduction to Cluster Analysis (1990) 253-279.