Classification automatique

TP3

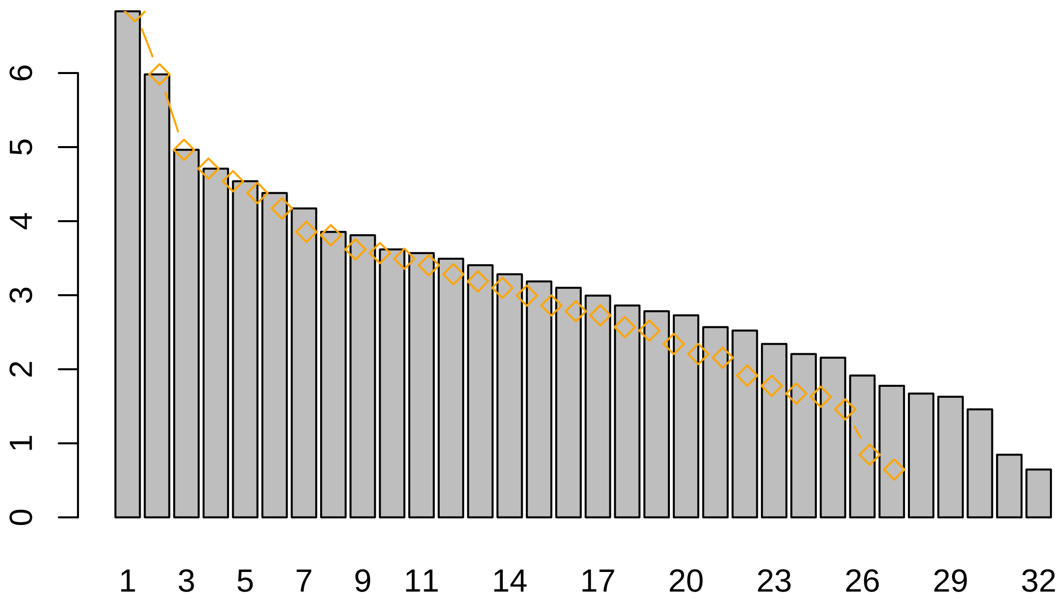
Clustering sur données qualitatives

```r
bc_data<-read.csv('breast_cancer.csv')
colnames(bc_data) <- c("class","age","menopause","tumor_size","inv_node",
                       "node_capes","deg_malig","breast","breast_quad","irradiat")


z <- data.frame(variable = names(bc_data),
               classe = sapply(bc_data, class),
                number_distinct_value = sapply (bc_data,
                                                function(x) paste0(length(unique(x)))),
                distinct_values = sapply(bc_data,
                                         function(x) paste0(unique(x)[1:4], collapse = ', ')),
               row.names=NULL)
xtable(z)
```

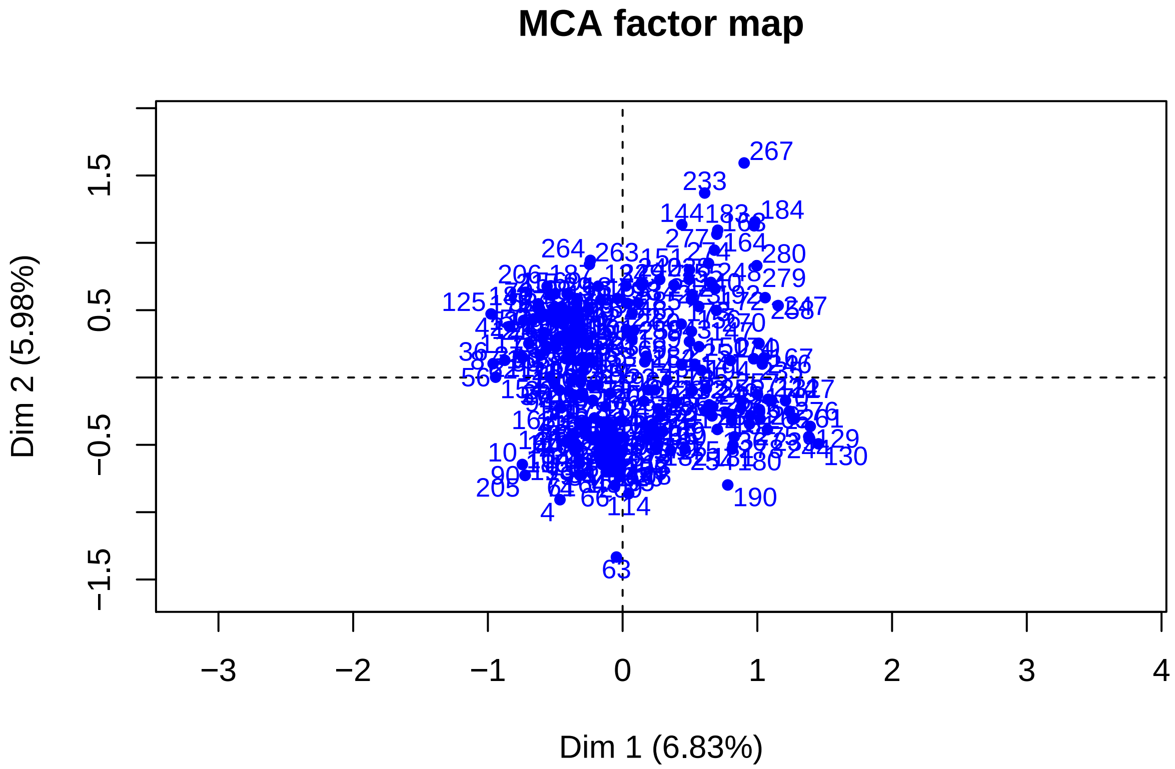
variable	classe	number_distinct_value	distinct_values
class	factor	2	no-recurrence-events, recurrence-events, NA, NA
age	factor	6	40-49, 60-69, 50-59, 30-39
menopause	factor	3	premeno, ge40, lt40, NA
tumor_size	factor	11	20-24, 15-19, 0-4, 25-29
inv_node	factor	7	0-2, 6-8, 9-11, 3-5
node_capes	factor	3	no, yes, ?, NA
deg_malig	integer	3	2, 1, 3, NA
breast	factor	2	right, left, NA, NA
breast_quad	factor	6	right_up, left_low, left_up, right_low
irradiat	factor	2	no, yes, NA, NA

2. ACM du jeu de données

```r
k_bc_data<-bc_data[,2:10]

res.mca <- MCA(k_bc_data, quali.sup =6,  graph=FALSE)

barplot(res.mca$eig[, 2],
        names.arg = 1:nrow(res.mca$eig))

lines(x = 1:nrow(res.mca$eig), res.mca$eig[, 2],
      type = "b", pch = 5, col = "orange")
```



```
plot(res.mca, invisible = c("var", "quali.sup", "quanti.sup"),
     cex = 0.8,  autoLab = "yes")
```



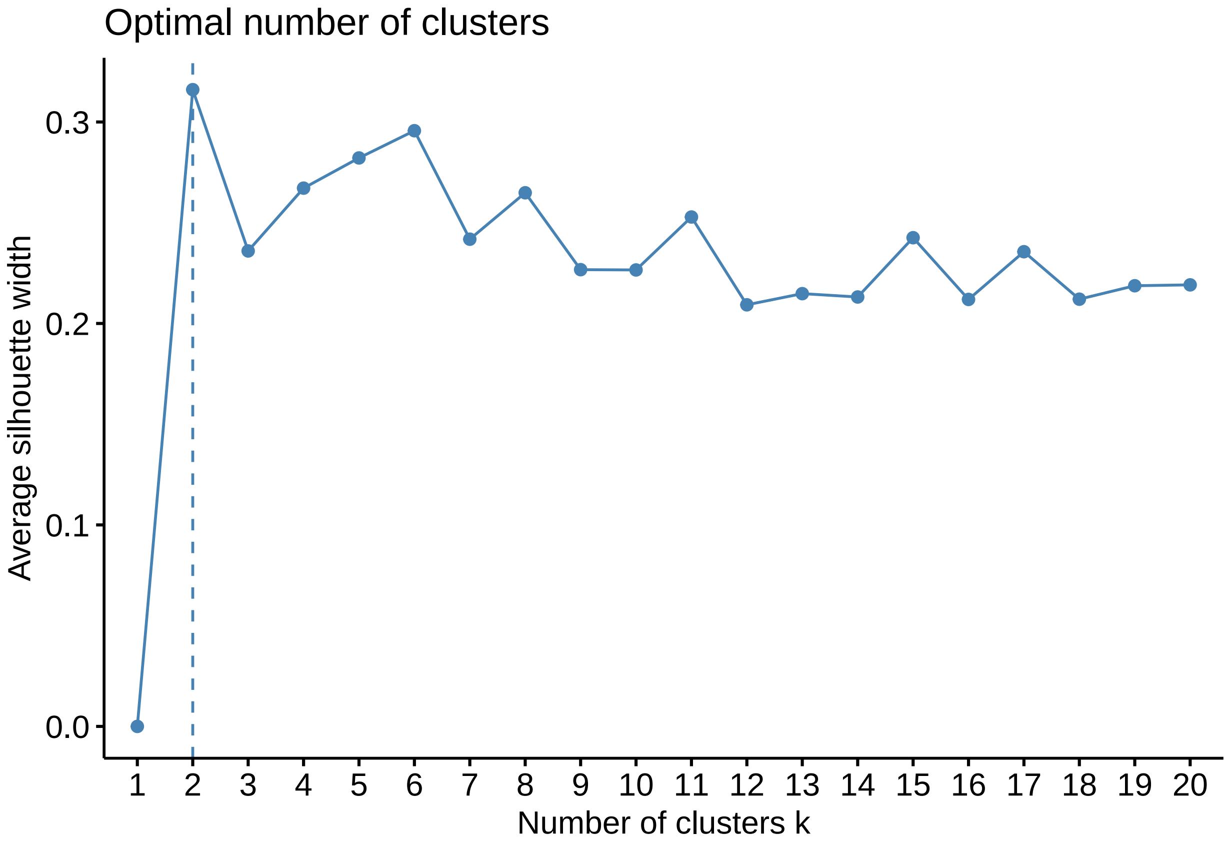
3. Kmeans à partir des coordonnées des axes factoriels

```
xtable(head(res.mca$ind$coord))
```

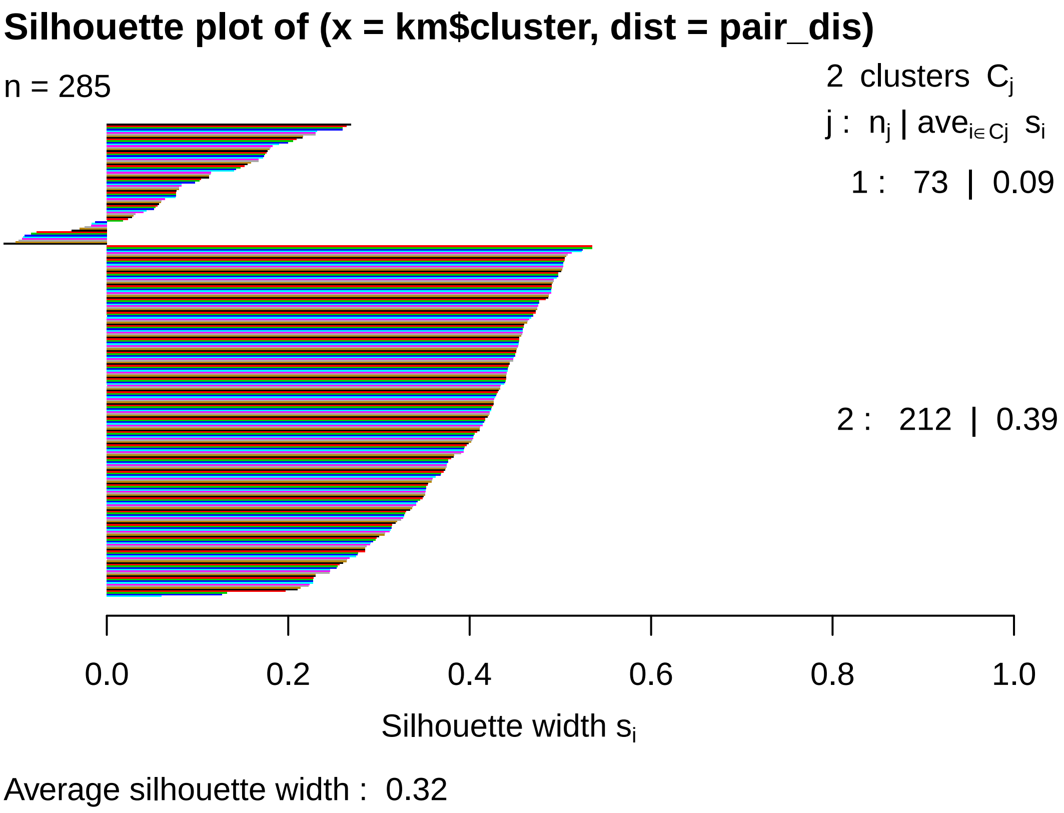
Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
-0.1201535	-0.6966364	-0.5267443	0.0174677	-0.1496198

Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
-0.1931814	-0.4367858	0.1178870	-0.1832874	-0.3366270
-0.5239355	0.4823525	-0.3006123	-0.0621047	-0.2159676
-0.4649799	-0.9074301	0.8822874	0.3725640	0.3053015
-0.5359814	0.6127724	0.1913333	-0.0171689	-0.1469078
-0.2213696	-0.1702961	0.2516085	-0.1779277	-0.0154044

```
fviz_nbclust(res.mca$ind$coord, kmeans, method = "silhouette",k.max=20)
```



```
km <- kmeans(res.mca$ind$coord,2,nstart = 10)

pair_dis<-daisy(res.mca$ind$coord)
sc<-silhouette(km$cluster, pair_dis)
plot(sc,col=1:8,border=NA)
```



```
res.bc_data <- cbind(bc_data,km$cluster)
xtable(table(res.bc_data[,1],res.bc_data[,11]))
```

	1	2
no-recurrence-events	36	164
recurrence-events	37	48

4. Kmode

```
k_bc_data[,"deg_malig"] <- as.factor(k_bc_data[,"deg_malig"])
k.mode<-kmodes(k_bc_data,2,iter.max = 100)
res.bc_data <- cbind(res.bc_data,k.mode$cluster)
xtable(k.mode$modes)
```

age	menopause	tumor_size	inv_node	node_capes	deg_malig	breast	breast_quad	irradiat
50-59	ge40	30-34	0-2	no	2	left	left_low	no
40-49	premeno	25-29	0-2	no	2	right	left_up	no

```
xtable(table(res.bc_data[,1],res.bc_data[,12]))
```

	1	2
no-recurrence-events	117	83
recurrence-events	52	33

5. Indices de rand

```
ir.1 <- cluster_similarity(km$cluster, k.mode$cluster)
ir.2 <- cluster_similarity(km$cluster, res.bc_data$class)
ir.3  <- cluster_similarity(k.mode$cluster, res.bc_data$class)
ir <- cbind(c("kmean vs. kmode","kmean vs. class"," kmode vs. class"),round(c(ir.1,ir.2,ir.3),3))
xtable(ir,digits = 2)
```

1	2
kmean vs. kmode	0.391
kmean vs. class	0.483
kmode vs. class	0.373