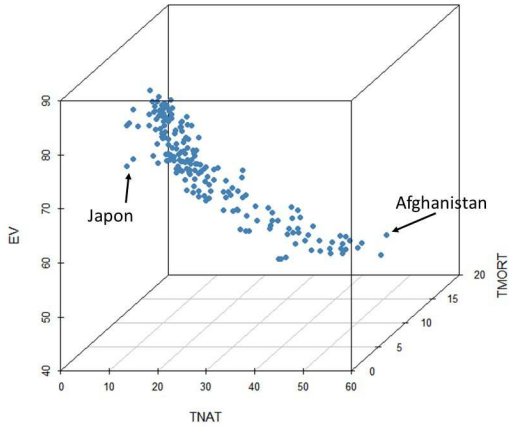
TP3 : Réduction de la dimension

Durée : 3h

L'objectif de ce TP est d'utiliser l'analyse en composantes principales pour réduire la dimension pour la recherche de clusters dans un jeu de données.

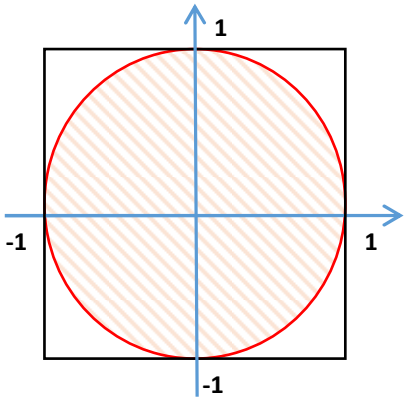
Les méthodes de datamining (supervisées ou non supervisées) souffrent du fléau de la dimension. La dimension est donnée par le nombre de variables dans le jeu de données. Supposons que toutes les variables soient quantitatives alors une ligne dans le jeu de données correspond à un point dans un espace dont les axes sont les variables. Ses coordonnées sont les valeurs que prend l'individu observé pour ces variables. L'espace est assimilé à un (hyper)cube car chaque variable varie entre une valeur minimale et une valeur maximale.

	TNAT	TMORT	EV
Afghanistan	45,49	18,6	44,97
Japon	7,7	9,66	83,37
…			



La plupart des algorithmes considèrent qu'un point donne de l'information « autour » de lui (voisinage). Considérons un point dans le milieu de l'espace ramené au cube unité, et le plus grand voisinage possible selon un distance euclidienne (boule). La surface couverte par ce voisinage correspond à 78,5% de la surface de l'espace en dimension 2. Si on augmente la dimension à 10 (c'est-à-dire 10 variables dans le jeu de données, ce qui n'est pas beaucoup), cette surface tombe à 0,2%. Cela signifie que plus la dimension augmente et plus les points sont éloignés dans l'espace, et ceci même s'il y a beaucoup de points. Est-ce que cela signifie encore quelque chose d'attribuer un point au cluster le plus proche sachant que tous les points du cluster sont très éloignés ? Il est donc nécessaire de réduire la dimension. Une des méthodes les plus connues est l'analyse en composantes principales.

	Recouvrement		
d	Vol. hypercube	Vol. sphère	%
2	4	3,1	78,5%
4	16	4,9	30,8%
6	64	5,2	8,1%
8	256	4,1	1,6%
10	1024	2,6	0,2%



Exercice 1 : Pluviométrie des villes françaises

Le jeu de données FrenchCities.csv contient des 34 caractéristiques météorologiques de 32 villes françaises.

```
CLIMAT : Kind of climate (Continental=1/Mediterranean=2/Oceanic=3/semi-oceanic=4)
NO2 : Nitrogen dioxide
DENSITY : People per km2
RAINFALL : Average annual rainfall (mm)
"MONTH"+r : Average monthly rainfall (12 variables)
DAY_RAINFALL : Average annual number of rainy days
"MONTH"=dr : Average monthly number of rainy days (12 variables)
TEMP : Average annual temperature (d° Celcius)
TEMP_RANGE : Temperature variation
SUNSHINE : Average of sunny days (hours per day)
LATITUDE : Latitude
LONGITUDE : Longitude
```

1) Centrer et réduire les variables

```
Dataset=read.table("FrenchCities.csv",
header=T,sep=';',row.names=1)
Scaled.Data=Dataset
Scaled.Data[,-1]=scale(Dataset[,-1])
Scaled.Data=as.data.frame(Scaled.Data)
```

2) Effectuer un clustering avec la méthode CAH. Combien choisissez-vous de clusters ?

```
d=dist(Scaled.Data,method="euclidian")
res=hclust(d,method="ward.D2")
plot(res)
plot(res$height)

groupes1=cutree(res,k=???)
```

3) Effectuer une ACP. Combien d'axes retenez-vous ? Récupérer les coordonnées des villes sur ces nouveaux axes.

```
library("FactoMineR")
library("explor")
res.PCA=PCA(Scaled.Data[,-1])
explor(res.PCA)

Scaled.Data2=scale(res.PCA$ind$coord[,1:2]) # si on ne retient que 2
axes
```

4) Effectuer un clustering avec les nouvelles coordonnées des villes sur les axes retenus. Est-ce que cela a changé le résultat par rapport à la question 2 ?

```
d=dist(Scaled.Data2,method="euclidian")
res=hclust(d,method="ward.D2")
plot(res)
plot(res$height)
```

```
groupes2=cutree(res,k=???)

# Graphique des clusters avant réduction
X11()
plot(res.PCA$ind$coord[,1],res.PCA$ind$coord[,2],
col=groupes1,main="Avant Reduction")
text(res.PCA$ind$coord[,1],res.PCA$ind$coord[,2],
row.names(Scaled.Data), col=groupes1)

# Graphique des clusters après réduction
X11()
plot(res.PCA$ind$coord[,1],res.PCA$ind$coord[,2], col=groupes2,
main="Après Reduction")
text(res.PCA$ind$coord[,1],res.PCA$ind$coord[,2],
row.names(Scaled.Data), col=groupes2)
```

5) Caractériser chaque cluster (pluvieux, ensoleillé, …) ?.

```
# on ajoute une colonne au jeu de données avec le numéro du cluster
Scaled.Data=cbind(Scaled.Data,groupes2)
Scaled.Data=as.data.frame(Scaled.Data)
Scaled.Data$groupes2=as.factor(Scaled.Data$groupes2)
# on effectue une ACP avec le nouveau jeu de données en précisant que
la dernière colonne est une variable qualitative supplémentaire
res.PCA=PCA(Scaled.Data[,-1],quali.sup=34)
explor(res.PCA)

# On peut aussi faire un croisement entre le Climat et les clusters
table(Scaled.Data$CLIMAT,groupes2)
```

Exercice 2 : Fromages français

Faites la même analyse que dans l'exercice 1 avec le jeu de données sur les fromages français, fromages.txt