

Sujet n°4 : Mise en place d'une méthode de prédiction du volume apparent de distribution

Jérôme LAURENT, Julie SAMYDE

Saturday, November 08, 2014

Sélection des variables

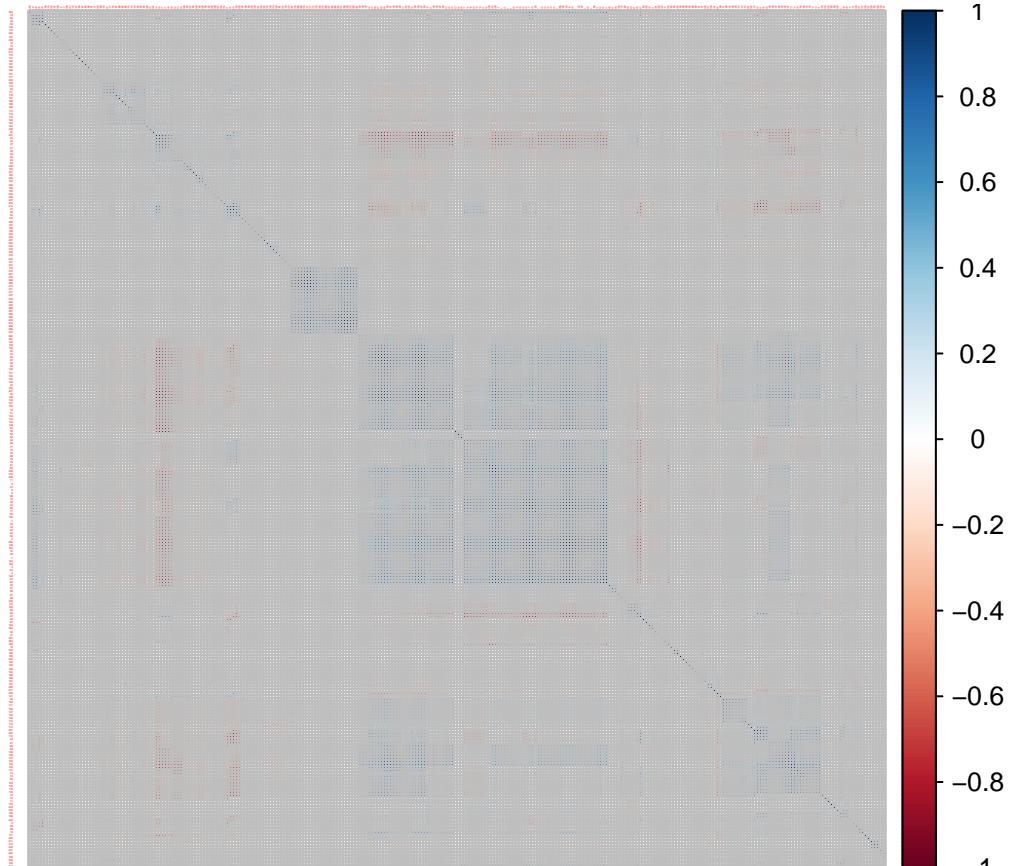
Une première étape simple consiste en l'application d'un filtre afin de repérer et d'enlever les variables fortement corrélées. En chémoinformatique, les descripteurs moléculaires présentent fréquemment de fortes corrélations entre eux.

```
# Centrage-réduction de toutes les variables explicatives
dat.scale<- scale(dataset[1:(ncol(dataset)-1)],center=TRUE,scale=TRUE)

# Calcul des corrélations
corMat <- cor(dat.scale)

# Représentation graphique
corrplot(corMat, order = "hclust", tl.cex = 0.1)

## Loading required package: corrplot
```



```

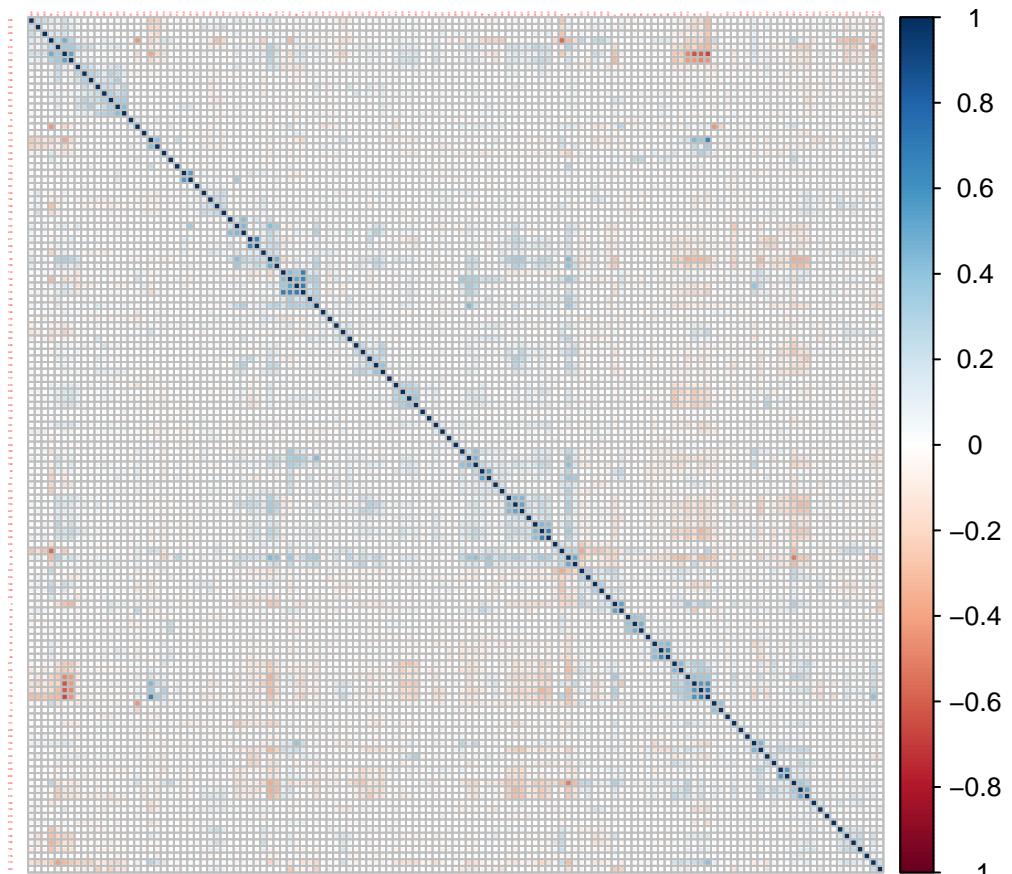
## Loading required package: caret
## Loading required package: lattice
## Loading required package: ggplot2

# Sélection des variables dont la corrélation est supérieure à 0,70
highCor <- findCorrelation(corMat, 0.70)

# Suppression de ces variables du jeu de données
datFiltered.scale <- dat.scale[,-highCor]

# Calcul et représentation de la matrice de corrélation
corMat.Filt <- cor(datFiltered.scale)
corrplot(corMat.Filt, order = "hclust", tl.cex = 0.1, bg = "white")

```



Contenant initialement 278 variables, le jeu de données n'en contient plus que 129 après application du filtre.