Reduction dimensionnelle à l’aide de l’analyse par composante principale.

L’ACP va essayer de trouver la relation dans notre donnée qui explique le plus la variance entre les données, et crée de nouvelles variables en les reliant. Qui devient une première composante, puis une autre variable qui sera liée et cette composante de variables, et ainsi de suite.

Objectif :

1. Etude des individus (*i.e.* classification dans un cluster) : deux lignes (deux individus) sont proches s'ils ont des résultats similaires ? pour y répondre, on s'intéresse à la variabilité entre individus.
   * Y a-t-il des similarités entre les individus pour toutes les variables ?
   * Peut-on établir des profils d’individus ?
   * Peut-on opposer un groupe d'individus à un autre ?
2. Etude des variables : on étudie les liaisons linéaires entre les variables. Les objectifs sont de résumer la matrice des corrélations et de chercher des variables synthétiques
   * Peut-on résumer les 5000 variables par un petit nombre de variables ?
3. Lien entre les deux études : peut-on caractériser des groupes d'individus par des variables ?

J’ai choisi pour montrer la puissance de l’ACP, une table de données qui contient 6000 lignes, et 5000 colonnes.

Toutes les colonnes sont quantitatives et n’ont pas de signification exacte.

On a un vecteur réponse qui détermine l’appartenance à 2 clusters distincts.

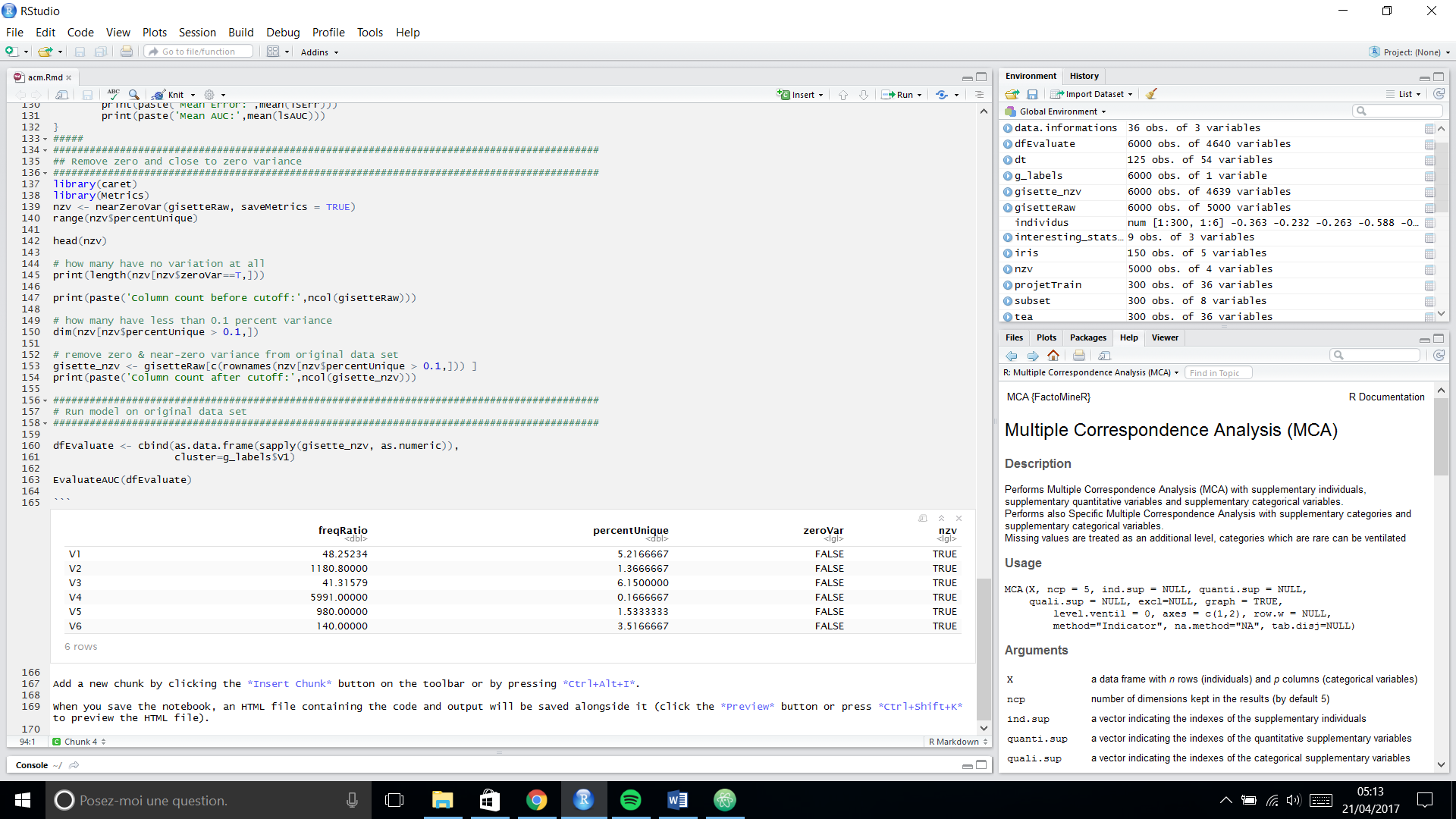
Etapes à suivre pour aboutir à la réduction de dimension :

1. On supprime les zéro et les variables qui présentent une variance faible car elle ne nous apportera pas beaucoup d’informations sur les composantes principales qui vont avoir une importance significative sur la réponse attendue.

On utilise la fonction « nearZeroVar du package Caret, qui donne ceci :

On a pu vérifier que le « percentUnique » qui est le paramètre d’unicité d’une variable, se trouve dans un intervalle de **[0.17 ; 8.6].**

On observe ci-dessous que l’algorithme indique si c’est une dimension a variance nulle, ou alors proche de zero (nzv), mais nous allons nous contenter de se débarrasser des dimensions au pourcentage unique < 0.1. On passe ainsi de **5000 colonnes à 4639**



1. On construit un modèle rapide de référence, xgboost sur notre dataset.

* AUC : 0.970094573718909"

1. Analyse par composante principale

On utilise la fonction Prcomp qui va nous permettre d’obtenir toutes les variables associées qui représentent la majeure partie de la variation de nos données.

Après l’obtention de ces variables composées, nous allons les utiliser pour essayer de construire un modèle sur un nombre de variable inferieur à celle du départ, mais tout en étant aussi efficace sur la prédiction.

4 – Prédiction en utilisant la **première** composante, la plus dominante :

On construit un modèle à partir du vecteur « composante principale 1 », et en testant on obtient les résultats suivants :

* AUC : 0.718743151498416"

5 – Prédiction en utilisant les **cinq** **premières** composantes les plus dominante :

* AUC : 0.927671419929747

**Conclusion :**

Ce résultat est remarquable, car nous sommes passés de 5000 variables, à 5, qui nous permet d’obtenir quasiment le même résultat.

L’analyse par composante principale est très intéressante dans ce cas de figure de réduction de dimension et de généralisation de modèle.

D’autres pistes d’exploitation sur la sélection des composantes principales sont d’analyser graphiquement la séparation de ces composantes,

Mais pour cela la recherche doit être ciblée, ou alors beaucoup plus affinée pour comprendre de quoi est ce que sont composée les composantes principales qui sont une association de variables qui maximisent la variance entre les composantes.

Annexes :

PC1 PC2 PC3 PC4 PC5

[1,] 0.09733800 -0.01733936 -0.0601171454 0.06236975 0.0929287784

[2,] 0.06140186 0.03708471 0.0347182565 -0.02810667 -0.0081968581

[3,] 0.03489463 0.04911855 0.0331244368 0.01281559 -0.0010768176

[4,] 0.03928385 0.01858043 -0.0124523262 0.02657117 -0.0554238798

[5,] -0.04302395 0.02043370 0.0002652201 -0.01785960 -0.0006091771

[6,] -0.03183122 0.02253993 -0.0059015150 -0.03051497 0.0037355813