Compte rendu 3 : Analyse en composantes principales

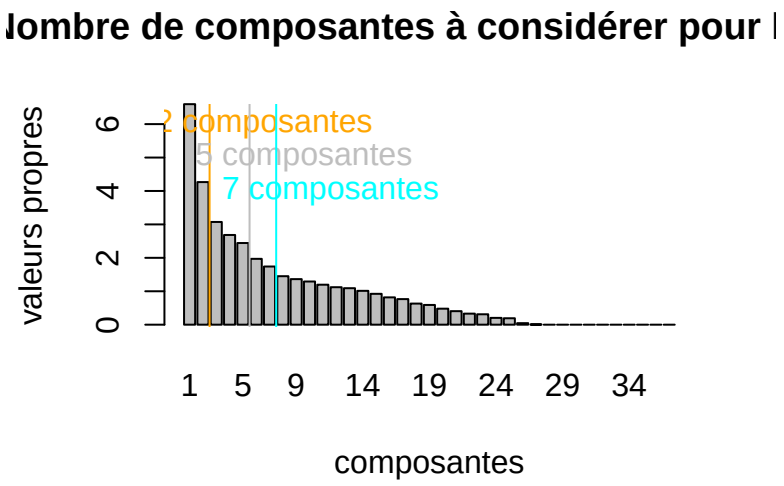
Etienne JEAN

Introduction / Contexte

Analyse en composantes principales du jeu de données

Anayse préliminaire

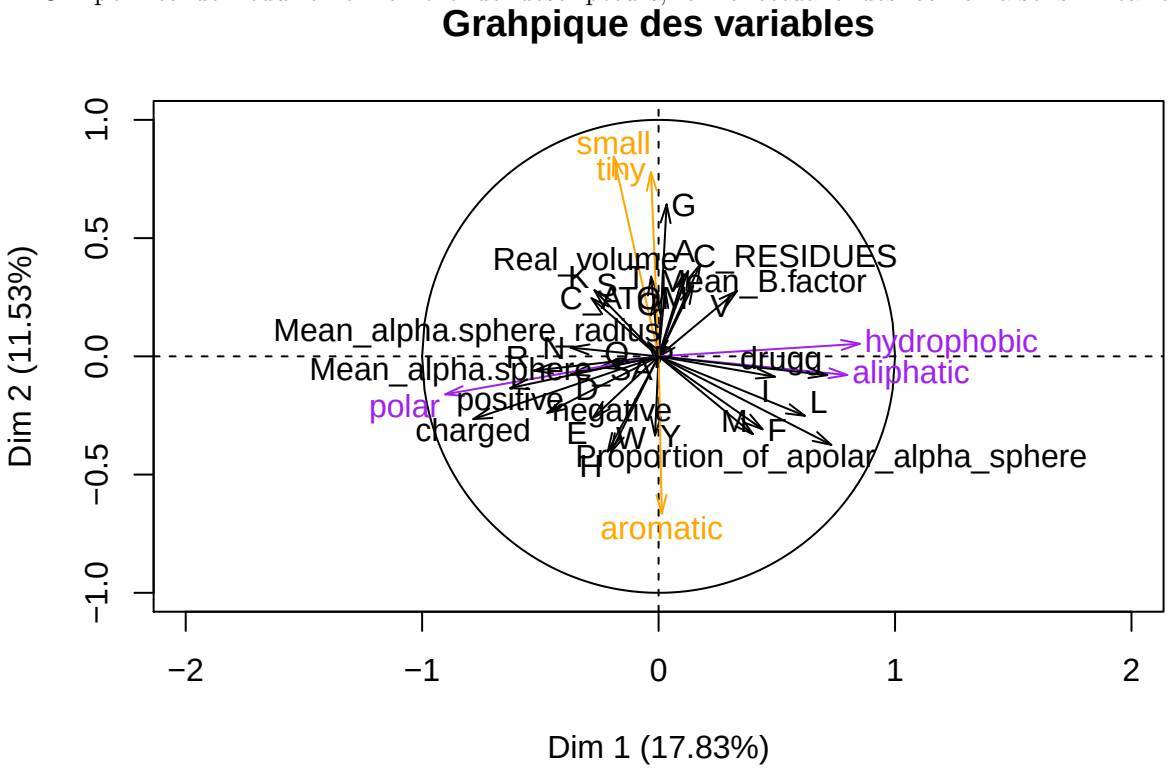
Afin de déterminer le nombre de composantes qu'il faudra considérer lors de l'ACP, il est judicieux de regarder les valeurs propres du jeu de données. Plus une valeur propre est grande, plus la composante qui lui est associée représente une grande dispersion du jeu de données, et donc plus elle permet de discriminer les différentes poches. Ainsi, dans l'optique de limiter le nombre de composantes à analyser, l'histogramme des valeurs propres met en évidence les "chutes" de valeurs propres. Il est donc intéressant de ne considérer que les composantes qui correspondent aux valeur propres qui précèdent ces chutes (fig. 1).



On peut au mieux distinguer trois sauts ici, après la 2ème composante, après la 5ème, et après la 7ème. Le saut le plus significatif se trouve après la 2ème composante, c'est pourquoi nous ne considèreront dans notre ACP que les composantes 1 et 2, qui à elles deux représentent près de 30% de la variabilité du jeu de données.

Graphique des variables

L'ACP permet de réduire le nombre de descripteurs, en effectuant des combinaisons linéaires de ceux-ci.



La première composante représente plus de 17% de la variabilité du jeu de données. Elle correspond aux fortes proportions dans le poches d'acides aminés polaires, à gauche du graphe, ou hydrophobes et aliphatiques, à droite du graphe. Le caractère polaire d'un acide aminé étant opposé aux caractères hydrophobe et aliphatique, il est normal d'observer que ce descripteur est anti-corrélé avec les deux autres. La deuxième composante représente plus de 11% de la variabilité. On retrouve une forte proportion d'acides aminés petits (*small* et *tiny*) dans les poches en haut du graphe, et les acides aminés aromatiques en bas du graphe.

Graphique des individus

Le grpahie des individus nous permet de d'observer la répartition des poches suivant les deux composantes. On remarque que les poches druggables sont plus hydrophobes et aliphatiques, alors que les poches non druggables sont plutot polaires (fig. 3). La deuxième dimension cependant ne permet pas de discriminer les poches druggables des poches non-druggables.

Grahpique des individus