Projet ADD

2024-07-19

ACP sur des données de vin

On dispose d'un jeu de données (vin.csv). Nous avons 1143 vins décrits par 11 variables (les 11 premières colonnes). La dernière colonne est une note indiquant la qualité du vin ; nous l'utiliserons comme variable supplémentaire dans notre analyse.

Réalisez une ACP sur ces données, choisissez le nombre de composantes, analysez les composantes principales, et représentez les individus. Trouvez-vous visuellement des groupes de vins ? Que pouvez-vous en dire ?

La qualité du vin vous semble-t-elle explicable par les autres variables ?

On aimerait faire une heatmap avec les différents vin mais il sont trop nombreux, faites d'abord 20 groupes dans les données puis faites une heatmap avec le centre des classes. Analysez là.

Étude des données sur le nombre de copies d'ADN

Les données sont disponnibles dans "data_cnv.Rdata".

Introduction

Dans une cellule saine, nous avons deux copies de chaque segment de chromosome (une venant du père et une venant de la mère). Cependant, certaines maladies, comme le cancer, peuvent attaquer des segments de chromosomes, les supprimant ou, au contraire, les dupliquant. Ainsi, nous pouvons avoir 0, 1, 2, 3, 4, etc., copies d'un même segment de chromosome.

Ce jeu de données fournit pour 50 patientes atteintes du cancer de l'ovaire le nombre de copies d'ADN sur 1000 positions différentes du chromosome 7. Ici, un individu est une patiente, et la variable est le nombre de copies à une position donnée (la variable 1 correspond à la première position,2 à la seconde, etc). Nous avons donc 50 individus et 1000 variables correspondant aux positions sur le chromosome 7.

Nous allons examiner si différents groupes de patientes atteintes du cancer de l'ovaire se distinguent en fonction du nombre de copies d'ADN aux différentes positions du chromosome 7.

Identifier ces groupes peut avoir un intérêt thérapeutique. En effet, une thérapie peut fonctionner pour des patientes d'un groupe mais pas pour celles d'un autre groupe.

Analyse

Réalisez une ACP sur ce jeu de données. Nous n'analyserons pas ici les composantes principales car il y a beaucoup trop de variables. (Si on veut ne mettre que quelques variables juste pour voir on peut faire fviz_pca_var(pc1, select.var = list(contrib =20)), qui ne représentera que les 20 variables qui contribuent le plus aux compolsantes.)

Vous pouvez également représenter la contribution des variables à la première composante en faisant plot(pc1varcontrib[,1]). En abcisse vous aurez les numéro de la variable (donc la position le long du génome) et en ordonnées sa contribution. Si vous le faites qu'observez vous ?

Représentez la projection des individus sur les deux premières composantes principales. Concluez : observe-t-on plusieurs groupes distincts ?

Nous conserverons les 5 premières composantes principales. Effectuez un clustering sur les individus en utilisant ces 5 premières composantes principales.

Réalisez une heatmap sur les données et analysez les groupes obtenus. (On attend ici un commentaire tel que : “On observe trois groupes distincts. Le premier est élevé pour la première composante et faible pour les autres. Le second est élevé pour les deux premières composantes et faible pour les autres. Le dernier est, à l’inverse, faible pour les deux premières composantes et élevé pour les autres.”)

(Le problème est réel mais pour des problèmes d’accessibilité aux données les données sont ici simulées à l’aide du package jointseg qui simule des données en se basant sur des données réelles de nombre de copies d’ADN)