# Apprentissage en génétique des populations

Mohammed Sedki

Projet à rendre le 31/03/2018

*SID, Com et Mark* *ENSAI 2019*

---

## 1. Jeu de données

L'étude de la diversité génétique humaine présente un intérêt pour divers domaines allant de la compréhension génétique des maladies aux applications en criminologie. La détection de sous-populations ou clusters est la clé pour la reconstruction de l'histoire démographique d'une population. On s'intéresse ici à la prédiction de la sous-population d'appartenance d'un individu à partir de l'observation de bio-marqueurs génétiques de cet individus. Pour mettre en place un modèle d'apprentissage de la sous-population d'un individu à partir de l'observation des marqueurs dits SNP, nous allons utiliser le jeu de données *Human Genome Diversity Panel* disponible sur le site http://www.cephb.fr/hgdp/. Ce jeu de données est composé de 1043 individus et 660918 marqueurs SNP (bialléliques). L'accès à ce jeu de données se fait via l'installation du package HGPD.CEPH comme suit[1]

```
install.packages("HGDP.CEPH", repos="https://genostats.github.io/R/")
require(HGDP.CEPH)
# lire données
filepath <-system.file("extdata", "hgdp_ceph.bed", package="HGDP.CEPH")
x <- read.bed.matrix(filepath)
# données SNP et individus
head(x@snps)
head(x@ped)
head(x@ped$region)
head(x@ped$region7)
```

## 2. Réduction de dimension

L'analyse en composantes principales (ACP) est communément utilisée pour visualiser un nuage de points. Ici nous allons faire appel à une ACP pour réduire la dimension des covariables en retenant la totalité des axes en guise de nouvelle représentation des individus. Avant d'appliquer une telle procédure, nous avons besoin de normaliser les données. Rappelons que les données sont sous forme d'une matrice de tailles $(n, p)$ où $n$ est le nombre d'individus observés et $p$ le nombre de marqueurs observés pour chaque individu. Chaque marqueur est un SNP, qui possède

[1] Attention, l'installation du package HGDP.CEPH nécessite l'installation préalable du package gaston.

deux allèles possibles. Ainsi, le génotype d'un marqueur particulier peut être codé sur la base du nombre d'allèles (0, 1 ou 2).

a. Le package gaston automatise le calcul des estimateurs de l'espérance et de la variance de chaque variable $X_j$ sur la matrice de données. Plus précisément set.stats permet de centrer et réduire la matrices des covariables x après avoir calculé les moyennes et variances des colonnes de la matrice.

On note $\mathbf{X}$ la matrice obtenue après standardisation des colonnes du jeu de données. La décomposition en valeurs singulière de $\mathbf{X}$ donnée par

$$\mathbf{X} = U\Gamma V^t$$

où $\Gamma$ est une matrice diagonale formée par les valeurs dites *sigulières* $\gamma_1, \gamma_2, \ldots$.

On note $S$, la matrice de covariance empirique des marqueurs de taille $p \times p$ définie par

$$S = \frac{1}{n-1}\mathbf{X}^t\mathbf{X}$$

Les vecteurs propres $\mathbf{v}_1, \mathbf{v}_2, \ldots$ associés aux valeurs propres $\lambda_1 \geq \lambda_2 \geq \ldots$ où $\lambda_i = \gamma_i^2$ sont les composantes principales[2]. En génétique des populations où $n < p$, (Cavalli-Sforza, Menozzi, and Piazza 1994) se sont intéressés à la matrice dite *duale* de taille $n \times n$ donnée par

$$H = \frac{1}{p}\mathbf{X}\mathbf{X}^t,$$

pour mettre en place une analyse en composantes principales. On notera $\xi_i$ les valeurs propres associées à $H$ et $u_i$ ses vecteurs propres.

b. Rappeler le lien entre les éléments propres des matrices $S$ et $H$.

c. La fonction GRM du package gaston permet le calcul de la matrice $H$. À l'aide de la fonction select.snps, restreindre l'étude aux snp autosomaux avec une fréquence de l'allèle mineur strictement supérieure à 0.05.

d. Calculer les vecteurs propres associés à la matrice $H$ à l'aide de la fonction eigen. Représenter les individus du jeu de données dans le premier plan factoriel. Colorier chaque point en fonction de sa région indiqué dans x@ped@region7.

## 3. Apprentissage

Nous souhaitons mettre en place un modèle d'apprentissage pour prédire $y$ correspondant à la région de l'individu disponible dans le champ region7 de l'objet x@ped. Les variables explicatives du modèle d'apprentissage sont l'ensemble des axes de l'ACP retenus précédemment[3].

[2] La matrice $S$ correspond à la matrice de corrélation lorsque les colonnes de $\mathbf{X}$ sont centrées et réduites (standardisées).

[3] À la fin de la partie précédente je vous conseille de stocker $y$ et la nouvelle matrice $X$ dans un fichier .rda pour éviter de refaire les calculs précédents.

e. Quelle est la nouvelle taille du jeu de données[4] ?

f. À l'aide d'une d'une partition en jeux de données apprentissage-test à 75% et 25%, à l'aide de l'erreur de prédiction du jeu de données test, comparer l'ensemble des modèles suivants[5].

| Modèle | Package | Conditions d'apprentissage |
|---|---|---|
| Gradient Boosting | caret et gbm | CV à 5 folds répétée 1 fois |
| Random Forest | caret et randomForest | CV à 5 folds répétée 1 fois |
| SVM radial | caret et e1071 | CV à 5 folds répétée 1 fois |
| SVM linéaire | caret et e1071 | CV à 5 folds répétée 1 fois |
| Réseau de neurones à une couche cachée | keras | mini-lots de taille 100 et 20 cycles d'apprentissage |

Cavalli-Sforza, Luigi Luca, Paolo Menozzi, and Alberto Piazza. 1994. *The History and Geography of Human Genes*. Book. Princeton, N.J. : Princeton University Press.

[4] Plus précisément, le nombre de variables après la réduction de dimension

[5] Attention, les calculs peuvent être très lents, il est fortement conseillé de faire des tests avec des petites grilles avant de lancer le calcul final. Il est aussi conseillé de travailler sur vos machines personnelles qui sont plus performantes que les VM. Si la capacité de votre machine le permet, penser à paralléliser le calcul et sauvegarder les calculs au fur et à mesure si le script nécessite des heures entières de calcul.