Le jeu de données MINST ou le *Hello word* de la classification d'images

Mohammed Sedki

Projet à rendre le 28/02/2018

MSP/ES *Apprentissage et agrégation de modèles*

---

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 au regroupement en sous-populations (clustering) du 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 (biallélques). L'accès à ce jeu de données se fait via l'installation du package HGPD.CEPH comme suit

```
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)
```

2. Analyse en composantes principales

L'analyse en composantes principales est communément utilisée pour visualiser des groupes dans un nuage de points. 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 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. Rappeler l'espérance et la variance d'un marqueur X_j observé sous l'équilibre de Hardy-Weinberg. Proposer deux estimateurs intuitifs de ces deux quantités.

b. Le package gaston propose une fonction qui automatise le calcul des estimateurs de ces deux quantités sur la matrice de données. Vérifier que les estimateurs calculés par cette fonction

sur la première colonne de la matrice de données correspondent aux estimateurs proposés en réponse à la question précédente.

On note $\mathbf{X}$ la matrice de données 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 Γ 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[1]. En génétique des populations où $n < p$, [@Cavalli] 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 ξ_i les valeurs propres associées à H et u_i ses vecteurs propres.

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

d. 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.

e. 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.

f. Regrouper le jeu de données en 7 groupes à l'aide d'un kmeans sur les deux premiers vecteurs propres. Comparer la partition obtenue à la partition x@ped@region7 à l'aide d'une matrice de confusion. Évaluer la correspondance entre les deux partitions avec la fonction ARI du package VarSelLCM.

3. Clustering spectral

Pour introduire le sujet nous avons besoin du vocabulaire de la théorie des graphes. Pour n points, on définit un graphe G de sommets $\{1, \ldots, n\}$ On associe à G une matrice W qui décrit la force de lien entre ses sommets : les couples de sommets (i, j) correspondant à une grande valeur $w_{i,j}$ sont fortement connectés. Les entrées nulles $w_{i,j} = 0$ indiquent les sommets non-connectés de G. Ce type de modélisation est d'une grande flexibilité par la variété de matrices de similarité qu'on peut définir sur un jeu de données à condition que celle-ci soit symétrique et à valeurs positives. L'idée

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

du clustering spectral est de détecter les composantes connexes du graphe à l'aide du spectre d'un Laplacien discret. Il existe différentes manières de calculer un Laplacien discret sur un graphe. Nous allons faire appel à la version dite *normalisée* donnée par

$$\mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}},$$

où $L = D - W$ et $D = diag(d_1, \ldots, d_n)$ est la matrice diagonale où $d_i = \sum_{j=1}^{n} w_{ij}$ correspond au degré du sommet i.

L'article de [@lee2010] propose d'adopter l'approche du clustering spectral dans le contexte de génétique des populations. Cette approche permet une détection automatique du nombre de sous-populations ainsi que le clustering des individus.

g. Résumer l'idée à l'origine de l'heuristique du choix du nombre de vecteurs propres du Laplacien ainsi que le nombre de sous-populations.

h. Reprendre le jeu de données restreint aux snp autosomaux avec une fréquence de l'allèle mineur strictement supérieure à 0.05. Implémenter la procédure décrite dans [@lee2010, Algorithm 1 (de 1: à 7:)][2] en appliquant un algorithme des kmeans sur la représentation des données obtenues. Expliciter les étapes ainsi que le résultat obtenu. Comparer la partition obtenue à la partition x@ped@region7 à l'aide d'un matrice de confusion. Évaluer la correspondance entre les deux partitions avec la fonction ARI du package VarSelLCM. Conclure.

[2] Attention: la définition du Laplacien normalisé à l'étape (3:) est erronée . Utiliser la définition donnée dans cet énoncé.