Exercices avec le logiciel R

Épreuve MADG

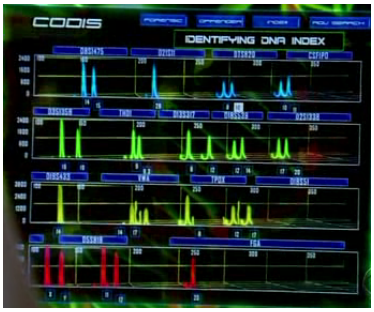
J.R. Lobry

Seconde session - Juillet 2021

Tous documents autorisés - échanges strictement interdits

Répondre directement sur la copie

Numéro d'intercalaire :



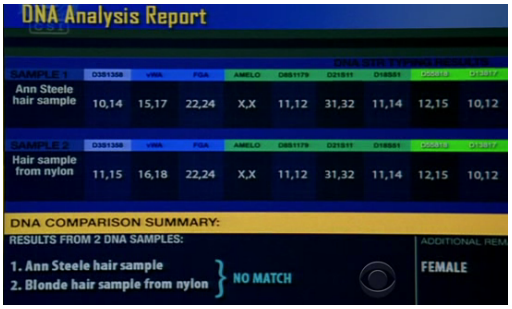


Figure 1: Copies d'écran de séries de télévision présentant des profils génétiques utilisés par la police scientifique. À gauche, la version variable quantitative (extrait de la série CSI Miami). Les allèles sont caractérisés par leur taille déterminée par leur temps de migration dans une électrophorère capillaire. Les quatre couleurs correspondent aux quatre fluorochromes utilisés. Les 16 rectangles bleus donnent les amplitudes de taille attendues pour les 16 loci utilisés (`D8S1179`, `D21S11`, `D7S820`, `CSF1PO`, `D3S1358`, `TH01`, `D13S317`, `D16S539`, `D2S1338`, `D19S433`, `vWA`, `TPOX`, `D18S51`, `AMEL`, `D5S818`, `FGA` du kit AmpFℓSTR© Identifiler© [1]). À droite la version variable qualitative (extrait de la série CSI New-York). Les allèles sont codés par une variable qualitative. Il s'agit ici d'échantillon de cellules diploïdes, il y a donc deux allèles pour chaque locus. Les conventions de nomemclature pour les modalités de ces variables seront données dans le texte.

1 Introduction

L'utilisation de profils génétiques par la police scientifique a été popularisée par de nombreuses séries de télévision parfois remarquablement exactes dans les détails (*cf* figure 1). Les données utilisées ici [2] sont les profils génétiques

de 699 individus (uniquement des Etats-Unis d'Amérique) ayant auto-déclaré leur *ethnicity* ou *race* (*cf* figure 2). La variable `ethnicity` est une variable qualitative non ordonnée dont la signification [2] est la suivante :

`Afric`	*African American*
`Cauc`	*U.S. Caucasian*
`Hisp`	*Hispanic*

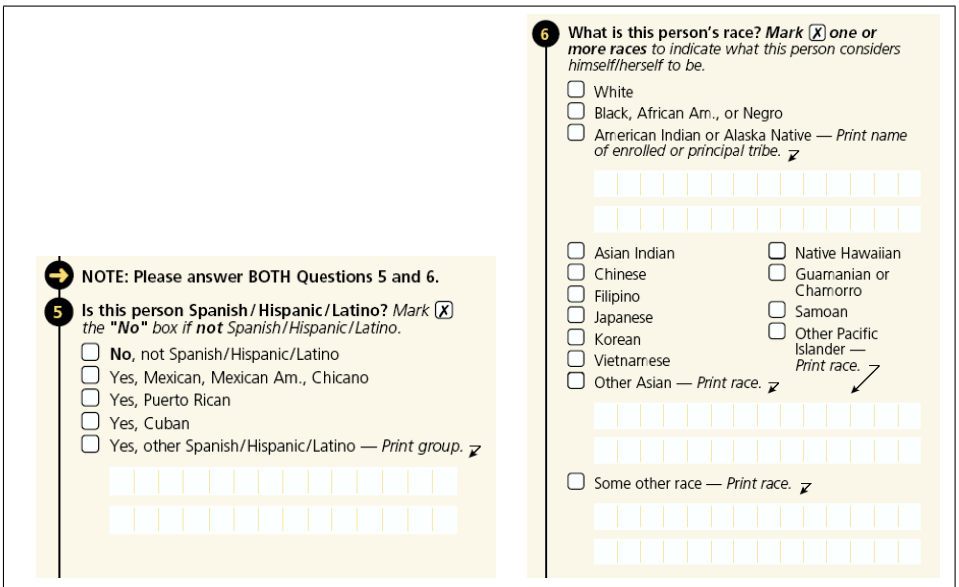


Figure 2: Les Etats-Unis d'Amérique utilisent un concept assez flou d'*ethnicity* (question 5) et de *race* (question 6). Cette figure est une copie du formulaire utilisé pour le recencemment de l'année 2000. Notez la possibilité pour un individu de cocher plusieurs cases pour *race*. La collecte de ce type d'informations est interdite en France par la loi informatique et libertés du 6 janvier 1978 qui dispose en son article 8 qu'*[i]l est interdit de collecter ou de traiter des données à caractère personnel qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou qui sont relatives à la santé ou à la vie sexuelle de celles-ci.*

2 Étude d'un artefact

On décide d'analyser les données du point de vue de la longueur des allèles, l'objet `codis.q` contient la longueur des allèles :

```
codis.q[1:5,1:5]
        D8S1179.1 D8S1179.2 D21S11.1 D21S11.2 D7S820.1
GA05071        52        56       122       130       28
GT37306        52        56       124       126       40
GT37312        52        56       116       120       40
GT37349        52        56       120       124       40
GT37351        40        48       112       116       40
```

Les individus étudiés sont diploïdes, ils ont donc tous deux allèles (éventuellement le même en cas d'homozygotie) pour chaque locus. Par exemple le premier individu a un allèle de taille 52 et un allèle de taille 56 au locus `D8S1179`.

```
codis.q[1, 1:2]
```
```
D8S1179.1 D8S1179.2
       52        56
```

On s'intéresse à la relation entre la taille des deux allèles. Au vu de la figure 3 page 3, que peut-on dire sur la relation entre la longueur des deux allèles pour chaque locus ?

Réponse:

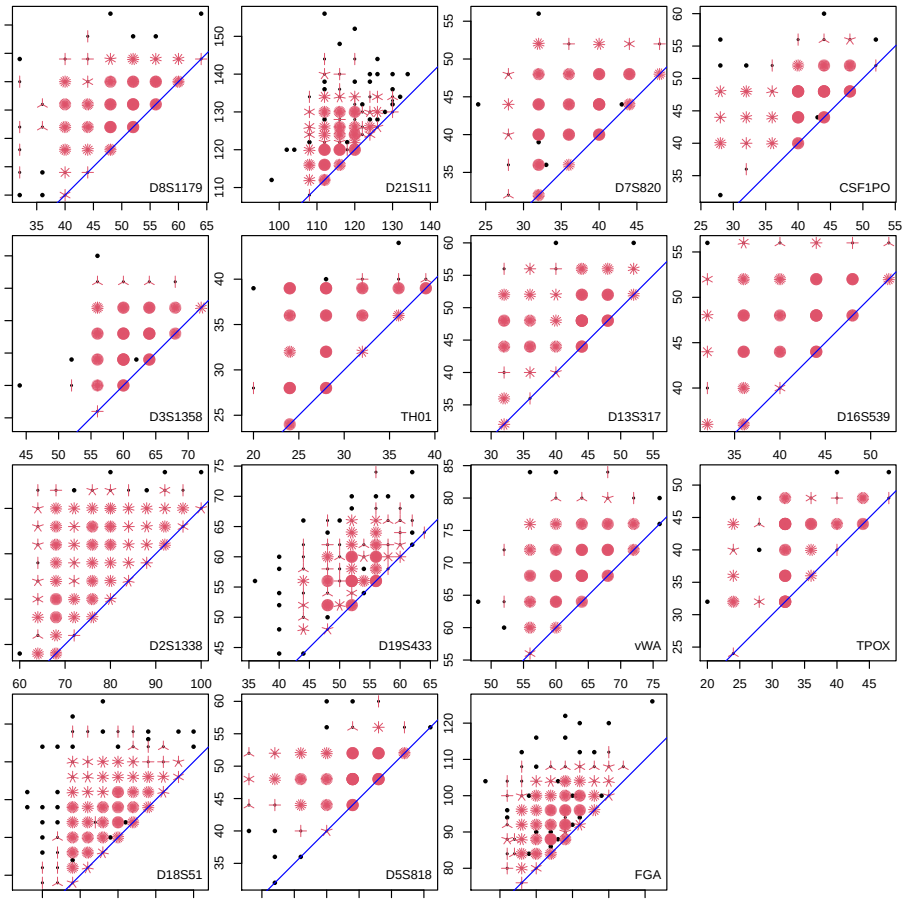


Figure 3: Il y a 15 graphiques pour les 15 loci étudiés (leur nom figure en bas à droite). Les 699 individus sont positionnés à la taille de leur premier allèle en abscisse et à celle de leur deuxième en ordonnée. La superposition des points est gérée ici par une représentation graphique de type tournesol. La ligne en bleu correspond à la première bissectrice ($y = x$).

Cette structure dans les données est complètement artificielle : chaque indi-

vidu possède un allèle issu de son géniteur et un allèle issu de sa génitrice, mais on ne peut pas dire lequel est lequel. L'ordre utilisé ici est purement conventionnel. On se demande si une telle structure dans les données est susceptible de conduire à des artefacts. Pour ce faire, on fait l'expérience suivante :

```
x <- rnorm(699)
head(x)
```

```
[1] -0.10283889 -0.96090053 -1.08888748 -0.08211635 -0.34245417  0.70478064
```

```
y <- rnorm(699)
head(y)
```

```
[1] -0.36238243 -0.09583669 -0.44373304  0.06972701 -2.27604194 -1.51435608
```

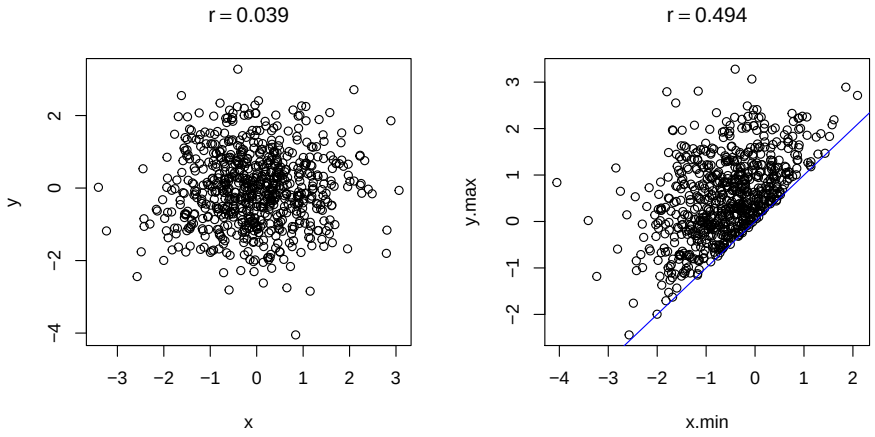
```
x.min <- pmin(x,y)
head(x.min)
```

```
[1] -0.36238243 -0.96090053 -1.08888748 -0.08211635 -2.27604194 -1.51435608
```

```
y.max <- pmax(x,y)
head(y.max)
```

```
[1] -0.10283889 -0.09583669 -0.44373304  0.06972701 -0.34245417  0.70478064
```

```
par(mfrow = c(1,2))
plot(x,y, main = bquote(r == .(round(cor(x,y),3))))
plot(x.min, y.max, main = bquote(r == .(round(cor(x.min,y.max),3))))
abline(c(0,1), col = "blue")
```



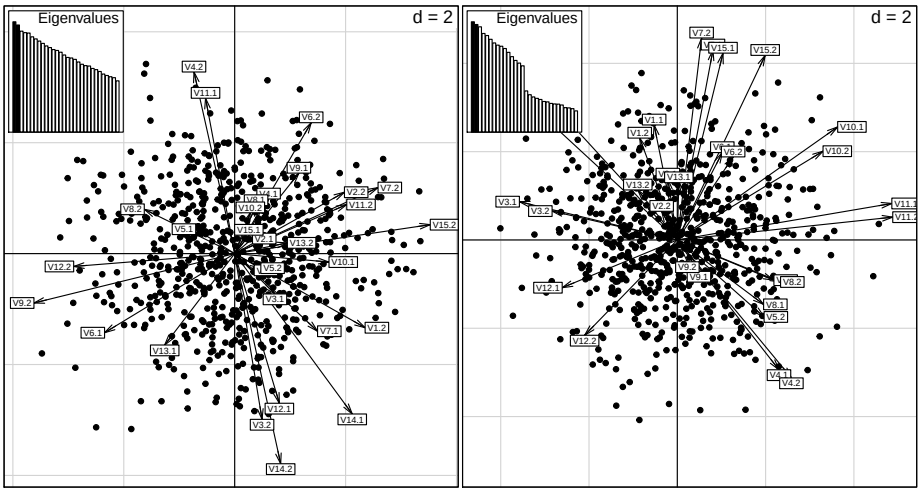
Au vu de ces résultats, pensez-vous que la structure artificielle dans les données soit susceptible de conduire à des artefacts ?

Réponse:

Toujours pour étudier l'effet de la structure artificielle des données, on simule un jeu de données de même dimension que le jeu étudié que l'on résume par une ACP :

```
set.seed(1)
rndtab <- as.data.frame(matrix(rnorm(30*699), ncol = 30))
colnames(rndtab) <- paste(rep(paste("V",1:15, sep = ""),each=2),1:2, sep = ".")
ordtab <- rndtab
for(i in seq(1,30,by=2)){
  ordtab[,i] <- pmin(rndtab[,i], rndtab[,i+1])
    ordtab[,i+1] <- pmax(rndtab[,i], rndtab[,i+1])
```

```
}
library(ade4)
rndtab.acp <- dudi.pca(rndtab, scann=FALSE)
ordtab.acp <- dudi.pca(ordtab, scann=FALSE)
par(mfrow = c(1,2))
scatter(rndtab.acp, clab.row = 0, clab.col = 0.5)
scatter(ordtab.acp, clab.row = 0, clab.col = 0.5)
```

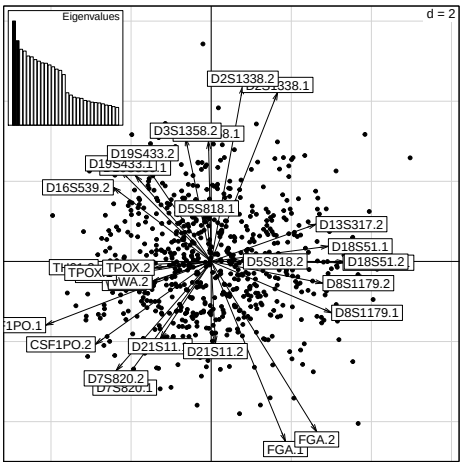


Au vu de ces résultats, quel est l'effet attendu de la structure artificielle des données sur le résultat de l'ACP ?

Réponse:

On effectue maintenant l'ACP du jeu de données pour voir si l'artefact est présent.

```
acp <- dudi.pca(codis.q, scann = FALSE)
scatter(acp, clab.row = 0)
```

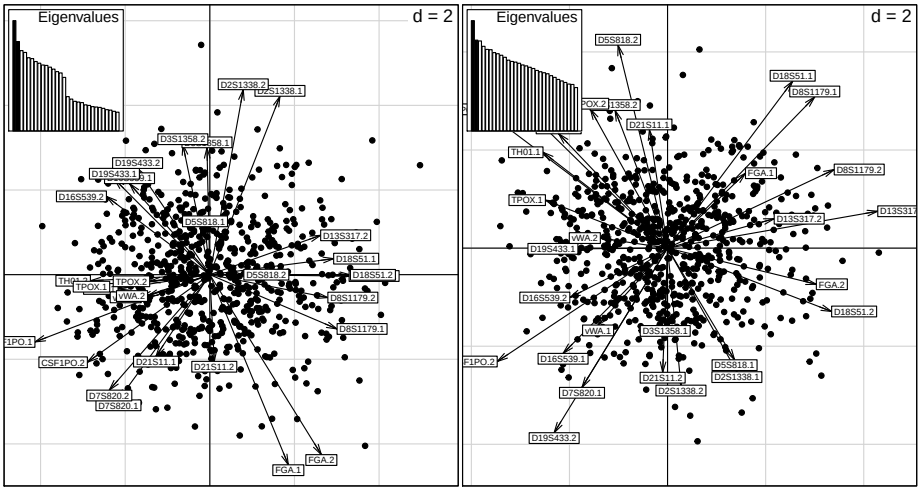


Au vu de ces résultats, l'artefact attendu est-il présent ?

Réponse:

Pour pallier cet inconvénient, on décide alors de faire la chose suivante :

```
codis.q2 <- codis.q
for(j in seq(1,30,by=2)){
  for(i in 1:699)
    codis.q2[i, c(j,j+1)] <- sample(codis.q[i, c(j,j+1)])
}
acpq2 <- dudi.pca(codis.q2, scann=FALSE)
par(mfrow = c(1,2))
scatter(acp, clab.row = 0, clab.col = 0.5)
scatter(acpq2, clab.row = 0, clab.col = 0.5)
```



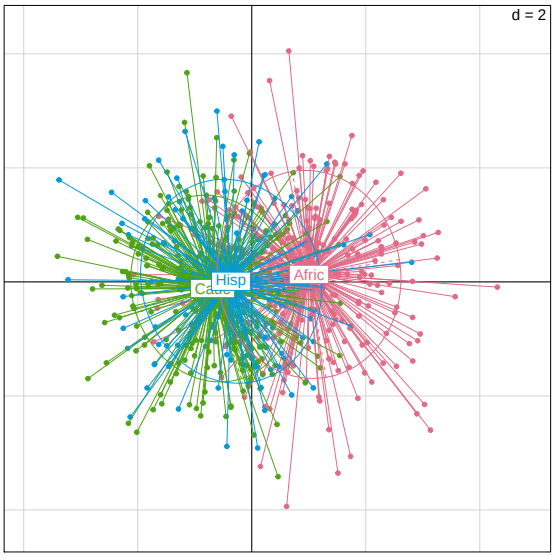
Expliquer ce que l'on a fait ici et pourquoi.

Réponse:

3 Interprétation de l'ACP

Pour aider l'interprétation du premier plan factoriel, on utilise les groupes comme variables illustratives :

```
s.class(acpq2$li, codis$ethnicity, col = hcl.colors(3, "Dark 3"))
```



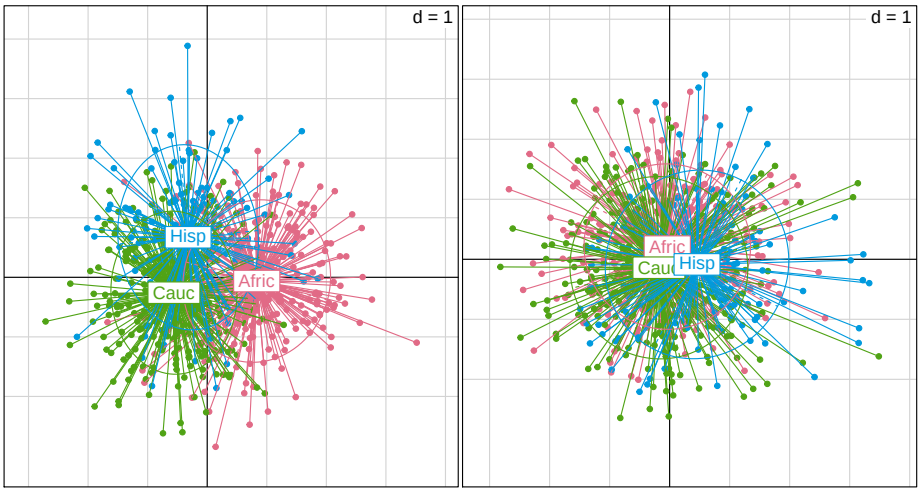
Au vu de ce résultat, quelle est votre interprétation du premier facteur de l'ACP ?

Réponse:

4 Pouvoir prédictif

Les loci étudiés ici sont ceux qui sont utilisés dans les bases de données ADN telles que le CODIS (*Combined DNA Index System*) géré par le FBI (*Federal Bureau of Investigation*) aux Etats-Unis d'Amérique ou le FNAEG (Fichier National Automatisé des Empreintes Génétiques) géré par la police nationale et la gendarmerie nationale en France. La question de la prédictibilité de l'appartenance d'un individu à un groupe ethnique à partir de la connaissance de ses allèles est une question sociétale sensible. On fait l'expérience suivante pour étudier le pouvoir prédictif de ce type de données :

```
ad2 <- discrimin(acpq2, codis$ethnicity, scann = FALSE)
codis.q3 <- codis.q2
for(j in 1:ncol(codis.q2))
  codis.q3[,j] <- sample(codis.q2[,j])
acpq3 <- dudi.pca(codis.q3, scann = FALSE)
ad3 <- discrimin(acpq3, codis$ethnicity, scann = FALSE)
par(mfrow = c(1,2))
s.class(ad2$li, codis$ethnicity, col = hcl.colors(3, "Dark 3"))
s.class(ad3$li, codis$ethnicity, col = hcl.colors(3, "Dark 3"))
```



Au vu de ces résultats, que pouvez-vous dire du pouvoir prédictif des loci utilisés en sciences forensiques ? Est-il beaucoup plus important que celui obtenu avec des données aléatoires ?

Réponse:

References

[1] Anonymous. *AmpFℓSTR© Identifiler© PCR Amplification Kit. User's Manual.* Applied Biosystems, Foster City, CA, USA, 2006. PN 4323291D.

[2] J.M. Butler, R. Schoske, P.M. Vallone, J.W. Redman, and M.C. Kline. Allele frequencies for 15 autosomal STR loci on U.S. caucasian, african american, and hispanic populations. *Journal of Forensic Sciences*, 48:908–911, 2003.