Fiche TD avec le logiciel : tdr624

Pratique de l'AFC discriminante sur données protéomiques et génomiques

J.R. Lobry

L'analyse factorielle des correspondances discriminante permet d'augmenter le pouvoir de résolution pour discriminer entre des groupes d'individus. On voit ici deux applications, une au niveau protéique et une au niveau nucléique.

Table des matières

1 Données protéomiques

1.1 Lecture des données

Les données sont extraites de [2]. Il s'agit de la composition en acides aminés de 413 protéines de *Escherichia coli* qui sont caractérisées en plus par leur localisation sub-cellulaire (protéines cytoplasmiques, protéines membranaires et protéines périplasmiques). Importer les données sous R avec le script suivant :

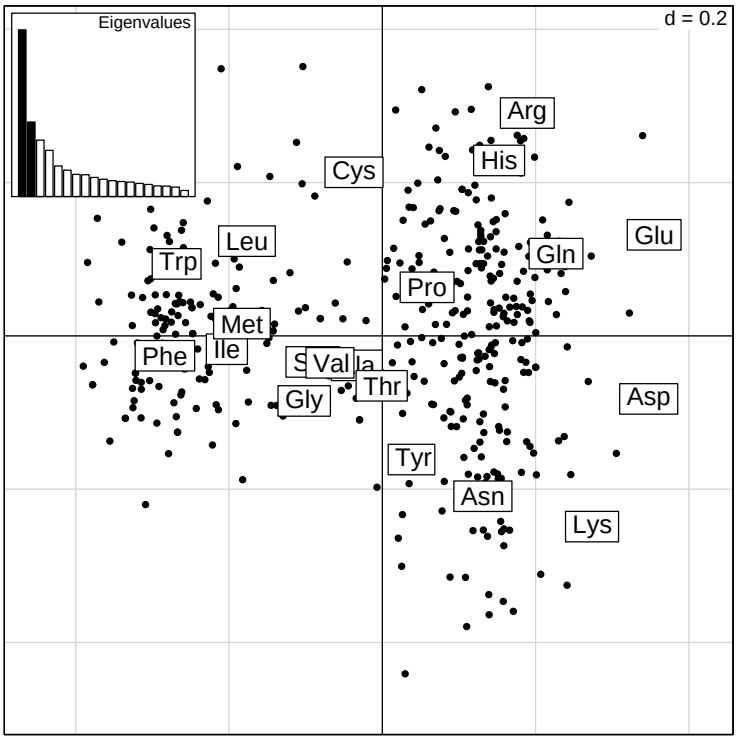
```
path <- "http://pbil.univ-lyon1.fr/members/lobry/repro/cabios96/Prot"
EcMP <- read.table(paste(path, "EcMP.fra", sep = "/"))
EcMPnames <- readLines(paste(path, "EcMP.lst", sep = "/"))
rownames(EcMP) <- paste(EcMPnames, 1:length(EcMPnames), sep = "")
EcCP <- read.table(paste(path, "EcCP.fra", sep = "/"))
EcCPnames <- readLines(paste(path, "EcCP.lst", sep = "/"))
rownames(EcCP) <- EcCPnames
EcPP <- read.table(paste(path, "EcPP.fra", sep = "/"))
EcPPnames <- readLines(paste(path, "EcPP.lst", sep = "/"))
rownames(EcPP) <- EcPPnames
Ec <- rbind(EcMP, EcCP, EcPP)
names(Ec) <- read.table(paste(path, "EcAA.difa", sep = "/"))$V1
locfac <- factor(rep(c("MP", "CP", "PP"), c(nrow(EcMP), nrow(EcCP),
    nrow(EcPP))))
dim(Ec)
```

```
[1] 413  20
```

1.2 AFC simple du tableau

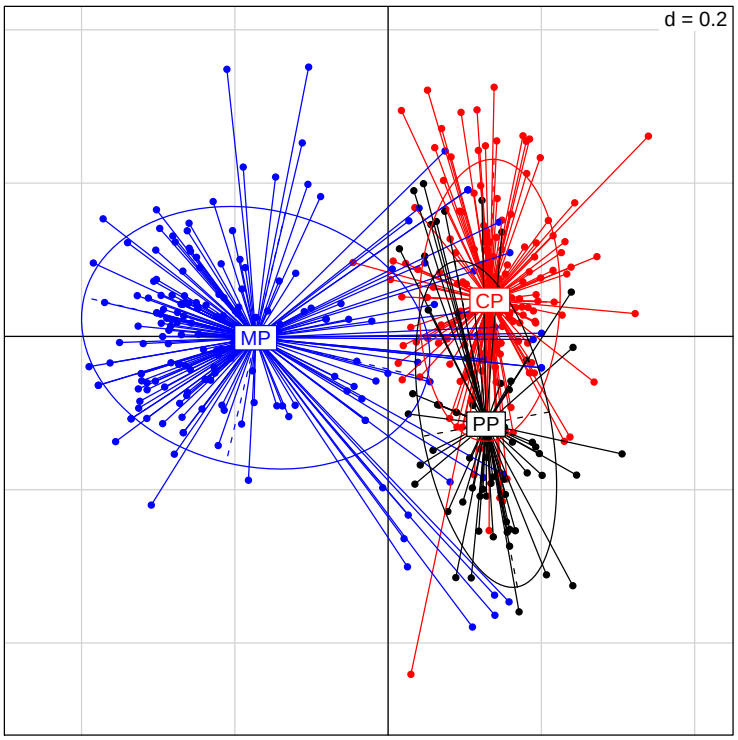
Faire l'AFC du tableau de départ :

```
library(ade4)
afc <- dudi.coa(Ec, scann = FALSE, nf = 2)
scatter(afc, clab.row = 0)
```



Représenter sur le premier plan factoriel l'information supplémentaire sur la localisation subcellulaire des protéines :

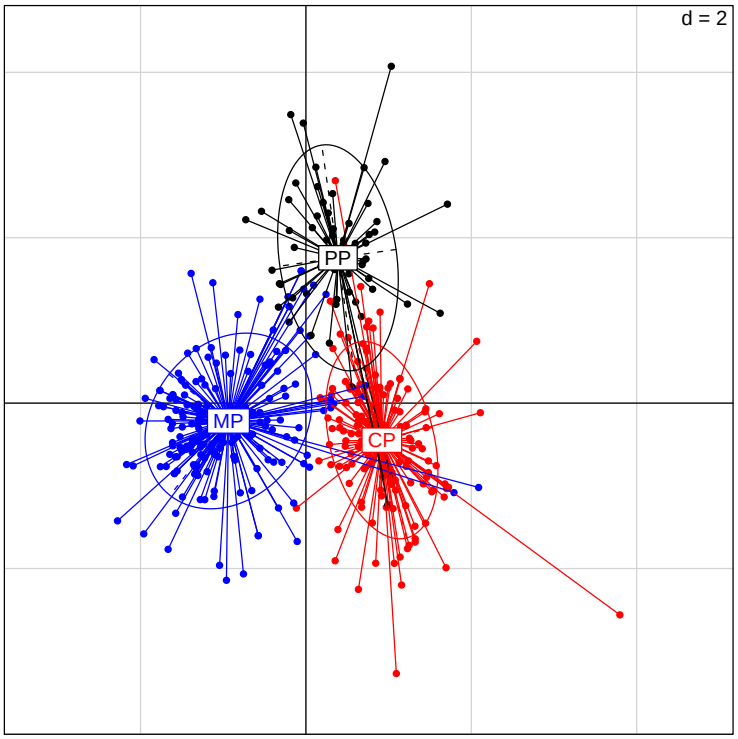
```
s.class(afc$li, locfac, col = c("red", "blue", "black"))
```



Quelle est votre interprétation des résultats ? Qu'est ce qui caractérise chaque groupe du point de vue de la composition en acides aminés ?

1.3 AFC discriminante

```
cda <- discrimin(afc, locfac, scann = FALSE, nf = 2)
s.class(cda$li, fac = locfac, col = c("red", "blue", "black"))
```



Qu'est ce qui a changé par rapport à l'AFC simple ?

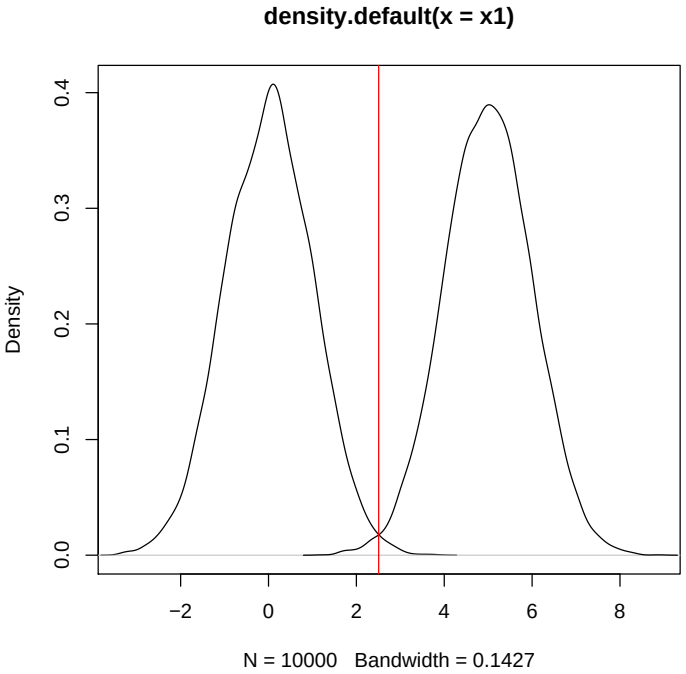
1.4 Choix d'un seuil pour la classification

Pour affecter un individu supplémentaire à un groupe on utilise une valeur seuil, s, qui tient compte de la variabilité intra-groupe :

$$s = \frac{\hat{s}_1 \bar{x}_2 + \hat{s}_2 \bar{x}_1}{\hat{s}_1 + \hat{s}_2}$$

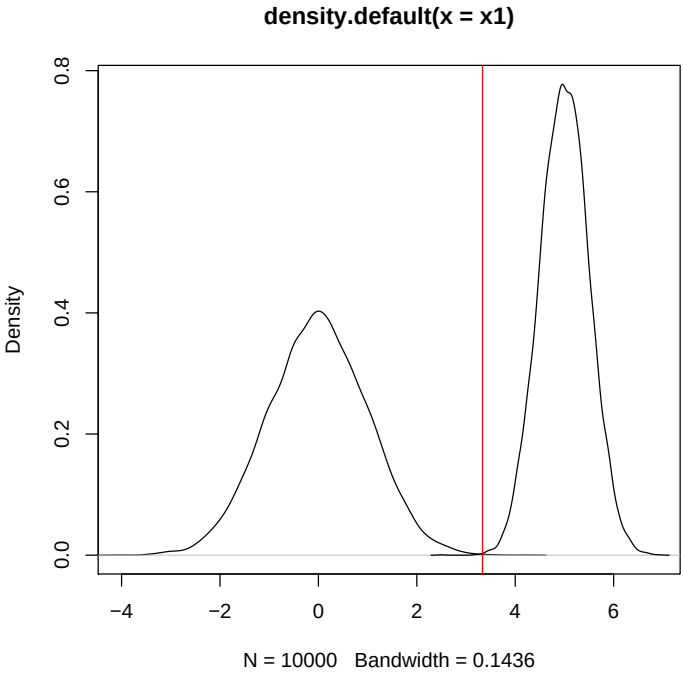
Représenter cette valeur seuil en cas d'homoscédasticité :

```
x1 <- rnorm(10000)
x2 <- rnorm(10000, mean = 5)
dstx1 <- density(x1)
dstx2 <- density(x2)
plot(dstx1, xlim = range(c(x1, x2)), ylim = range(c(dstx1$y, dstx2$y)))
lines(dstx2)
abline(v = (sd(x1) * mean(x2) + sd(x2) * mean(x1))/(sd(x1) + sd(x2)),
    col = "red")
```



Représenter cette valeur seuil en cas d'hétéroscédasticité :

```
x1 <- rnorm(10000)
x2 <- rnorm(10000, mean = 5, sd = 0.5)
dstx1 <- density(x1)
dstx2 <- density(x2)
plot(dstx1, xlim = range(c(x1, x2)), ylim = range(c(dstx1$y, dstx2$y)))
lines(dstx2)
abline(v = (sd(x1) * mean(x2) + sd(x2) * mean(x1))/(sd(x1) + sd(x2)),
    col = "red")
```

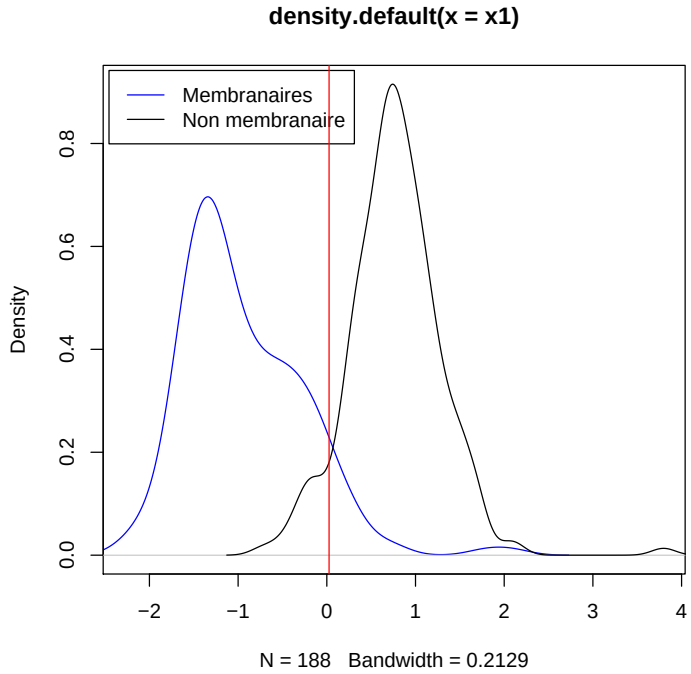


Quel est l'intérêt de calculer ainsi le seuil entre les deux groupes ?

1.5 Application à la prédiction des classes protéiques

On va utiliser le premier facteur de l'analyse factorielle des correspondances discriminante pour prédire si une protéine donnée est à localisation membranaire ou pas. Représenter la distribution des protéines sur le premier facteur et ajouter la valeur du seuil :

```
x1 <- cda$li[locfac == "MP", 1]
x2 <- cda$li[!locfac == "MP", 1]
dstx1 <- density(x1)
dstx2 <- density(x2)
plot(dstx1, xlim = range(c(x1, x2)), ylim = range(c(dstx1$y, dstx2$y)),
    col = "blue")
lines(dstx2)
legend("topleft", inset = 0.01, c("Membranaires", "Non membranaire"),
    col = c("blue", "black"), lty = 1)
s <- (sd(x1) * mean(x2) + sd(x2) * mean(x1))/(sd(x1) + sd(x2))
abline(v = s, col = "red")
```



Quelles sont les protéines membranaires mal classées ?

	DS1	DS2
P31119|AAS_ECOLI1	0.33615693	-0.007219117
P39168|ATMA_ECOLI10	0.21479711	-0.089846087
P08336|CPXA_ECOLI21	0.29620628	0.040299756
P08336|CPXA_ECOLI37	0.29620628	0.040299756
P24184|FDNH_ECOLI62	2.08828621	-1.019742737
P32175|FDOH_ECOLI64	1.79081054	-1.082243427
P06971|FHUA_ECOLI69	0.24319747	1.315028600
P28691|FTSH_ECOLI75	0.70074696	0.083571451
P24205|MSBB_ECOLI112	0.71876848	0.212577623
P31600|NFRA_ECOLI118	0.11746061	0.587463873
P31599|NFRB_ECOLI119	0.04358218	-0.774527800
P21420|NMPC_ECOLI122	0.04415363	0.810868185
P02934|OMPA_ECOLI132	0.09356992	1.425983168
P08400|PHOR_ECOLI142	0.30018018	-0.057151082

Quelles sont les protéines non membranaires mal classées ?

	DS1	DS2
P07365|CHEW_ECOLI	-0.115982100	-1.2697199
P23485|FECR_ECOLI	-0.203617979	0.9793787
P14609|FEPB_ECOLI	-0.160235849	1.2725774
P31697|FIMC_ECOLI	-0.409518476	1.6075088
P09376|HTRA_ECOLI	-0.727430104	2.2179874
Q03961|KSD1_ECOLI	-0.320761109	1.4447210
P42213|KSD5_ECOLI	-0.304716556	1.4349216
P39178|LOLA_ECOLI	-0.032055141	3.3863797
P03841|MALM_ECOLI	-0.183174636	3.4892267
P06875|PAC_ECOLI	-0.140476938	1.8829417
P15319|PAPD_ECOLI	0.009171903	1.3223968
P33364|PBP7_ECOLI	-0.207552294	2.4599200
P06128|PSTS_ECOLI	-0.122346577	2.6622971
P26648|SUFI_ECOLI	-0.332258458	1.1607676
P29679|TESA_ECOLI	-0.200779519	2.0869315
P19935|TOLB_ECOLI	-0.537277659	2.3196654

Utiliser le script suivant pour lire un jeu de données supplémentaire :

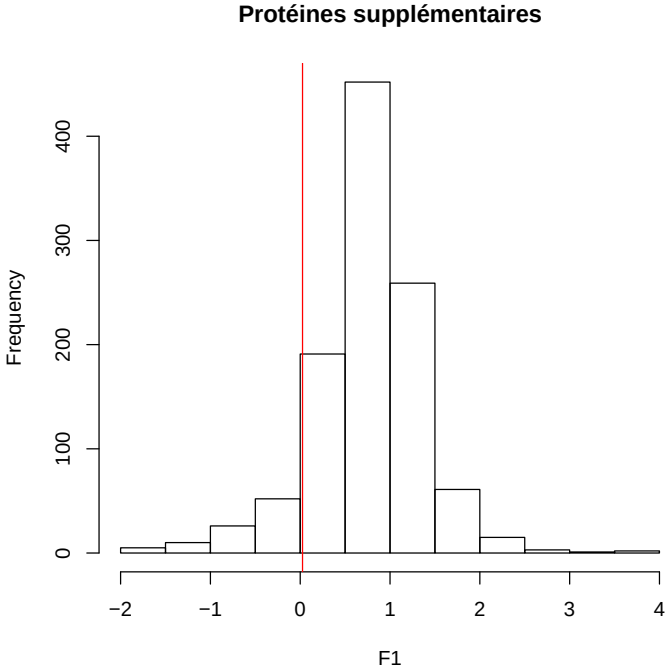
```
Sup <- read.table(paste(path, "Sup.fra", sep = "/"))
Supnames <- readLines(paste(path, "Sup.lst", sep = "/"))
rownames(Sup) <- Supnames
names(Sup) <- read.table(paste(path, "EcAA.difa", sep = "/"))$V1
```

On peut retrouver les coordonnées de la première protéine sur le premier plan factoriel ainsi :

```
all.equal(cda$li[1, 1], sum(Ec) * sum(cda$fa[, 1] * Ec[1, ]/colSums(Ec))/sum(Ec[1,
    ]))
```

```
[1] TRUE
```

Calculer les coordonnées des protéines supplémentaire sur ce premier facteur :



Combien de protéines sont prédites à localisation non membranaires ?

```
[1] 978
```

Combien de protéines sont prédites à localisation membranaires ?

```
[1] 99
```

Cette prédiction est elle compatible avec la proportion théorique de 11.4 % de protéines membranaires chez *Escherichia coli* rapportée [1] par ailleurs ?

```
prop.test(x = 99, n = 1077, p = 0.114)
```

```
        1-sample proportions test with continuity correction
data:  99 out of 1077, null probability 0.114
X-squared = 4.9812, df = 1, p-value = 0.02562
alternative hypothesis: true p is not equal to 0.114
95 percent confidence interval:
 0.07566352 0.11115885
sample estimates:
      p
0.091922
```

2 Données génomiques

2.1 Lecture des données

Les données sont extraites de [2]. Il s'agit de la composition en codons des séquences codantes dans quatre génomes bactériens. Ces séquences peuvent appartenir au groupe "brin précoce" (leading) ou bien au groupe "brin retardé" (lagging) en fonction de leur orientation vis à vis de la réplication du chromosome. Importez les données dans R avec le script suivant :

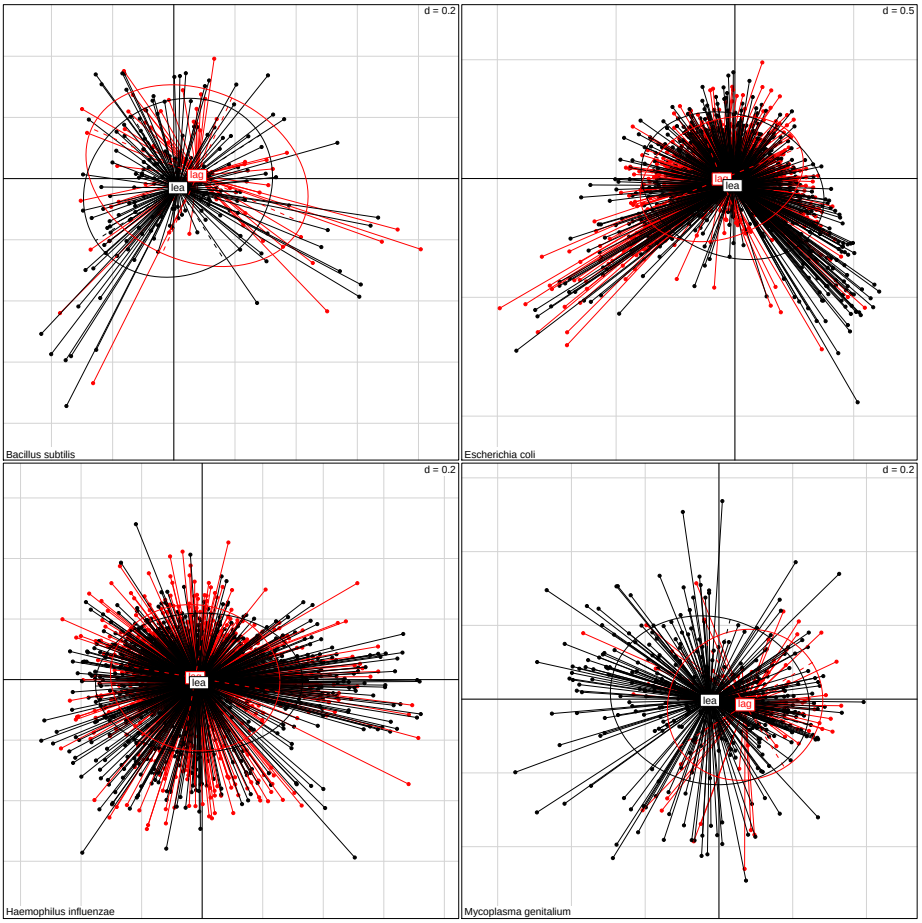
```
cat(readLines("http://pbil.univ-lyon1.fr/members/lobry/repro/cabios96/readNucl.r"),
    sep = "\n")
```

```
#
# Import under R data from the Nucl folder at
# http://pbil.univ-lyon1.fr/members/lobry/repro/cabios96/Nucl/
#
path <- "http://pbil.univ-lyon1.fr/members/lobry/repro/cabios96/Nucl"
#
# Bacillus subtilis data:
#
Bslea <- read.table(paste(path, "Bslea.frc", sep = "/"))
rownames(Bslea) <- readLines(paste(path, "Bslea.lst", sep = "/"))
Bslag <- read.table(paste(path, "Bslag.frc", sep = "/"))
rownames(Bslag) <- readLines(paste(path, "Bslag.lst", sep = "/"))
Bs <- rbind(Bslea, Bslag)
Bsfac <- factor(rep(c("lea", "lag"), c(nrow(Bslea), nrow(Bslag))))
names(Bs) <- read.table(paste(path, "codons", sep = "/"))$V1
#
# Escherichia coli data:
#
Eclea <- read.table(paste(path, "Eclea.frc", sep = "/"))
rownames(Eclea) <- readLines(paste(path, "Eclea.lst", sep = "/"))
Eclag <- read.table(paste(path, "Eclag.frc", sep = "/"))
rownames(Eclag) <- readLines(paste(path, "Eclag.lst", sep = "/"))
Ec <- rbind(Eclea, Eclag)
Ecfac <- factor(rep(c("lea", "lag"), c(nrow(Eclea), nrow(Eclag))))
names(Ec) <- read.table(paste(path, "codons", sep = "/"))$V1
#
# Haemophilus influenzae data:
#
Hilea <- read.table(paste(path, "Hilea.frc", sep = "/"))
rownames(Hilea) <- readLines(paste(path, "Hilea.lst", sep = "/"))
Hilag <- read.table(paste(path, "Hilag.frc", sep = "/"))
rownames(Hilag) <- readLines(paste(path, "Hilag.lst", sep = "/"))
Hi <- rbind(Hilea, Hilag)
Hifac <- factor(rep(c("lea", "lag"), c(nrow(Hilea), nrow(Hilag))))
names(Hi) <- read.table(paste(path, "codons", sep = "/"))$V1
#
# Mycoplasma genitalium data (variant genetic code):
#
Mglea <- read.table(paste(path, "Mglea.frc", sep = "/"))
rownames(Mglea) <- readLines(paste(path, "Mglea.lst", sep = "/"))
Mglag <- read.table(paste(path, "Mglag.frc", sep = "/"))
rownames(Mglag) <- readLines(paste(path, "Mglag.lst", sep = "/"))
Mg <- rbind(Mglea, Mglag)
Mgfac <- factor(rep(c("lea", "lag"), c(nrow(Mglea), nrow(Mglag))))
names(Mg) <- read.table(paste(path, "codonsmg", sep = "/"))$V1
```

```
source("http://pbil.univ-lyon1.fr/members/lobry/repro/cabios96/readNucl.r")
```

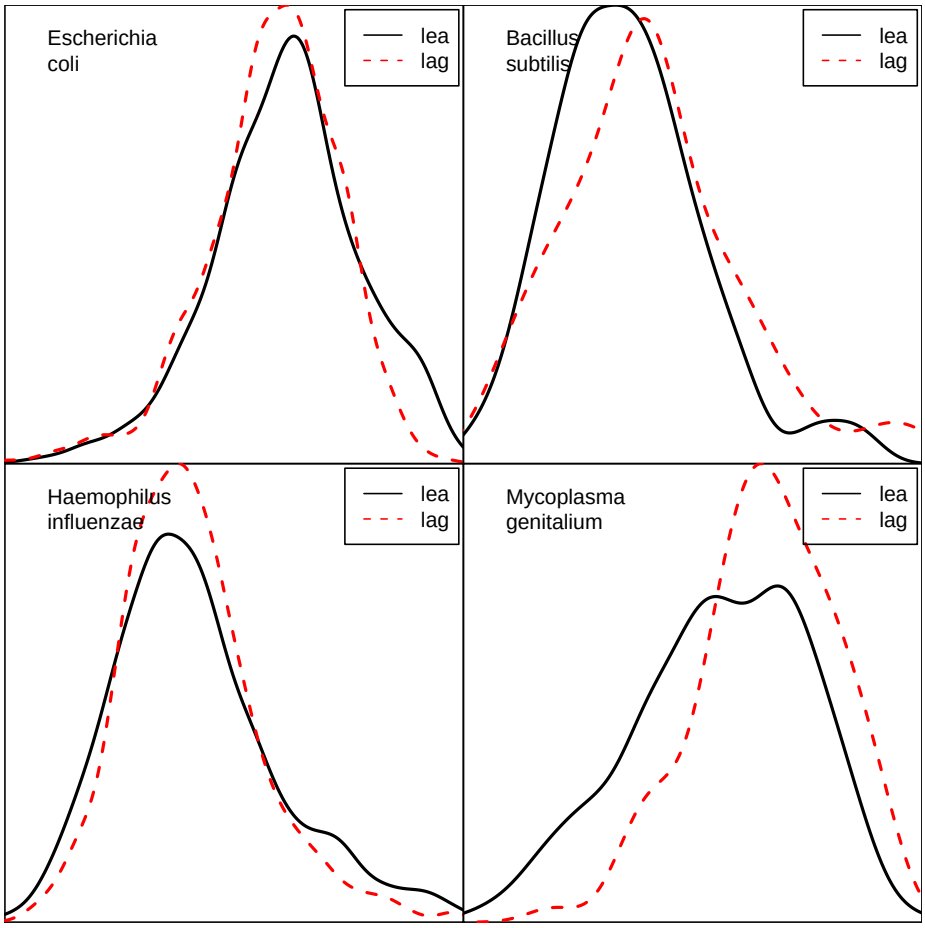
2.2 AFC simples

Faire les AFC simples des 4 tableaux et représenter les groupes en information complémentaire. Notez la différence par rapport au cas précédent : ici la structuration en groupe ne saute pas aux yeux. C'est dans ce type de situation que l'analyse discriminante, en tant que technique descriptive, prend tout son intérêt. Comment discriminer au mieux entre les deux groupes ?



Représenter la distribution des individus sur le premier facteur de l'AFC en fonction des deux groupes :

```
opar <- par(no.readonly = T)
par(mfrow = c(2, 2), mar = rep(0, 4))
f <- function(x, fac, title = "titre") {
    dstlea <- density(x$li[fac == "lea", 1])
    dstlag <- density(x$li[fac == "lag", 1])
    plot(dstlea, xlim = range(x$li[, 1]), ylim = range(c(dstlea$y,
        dstlag$y)), xaxs = "i", xaxt = "n", yaxt = "n", yaxs = "i",
        lwd = 2, main = "")
    lines(dstlag, col = "red", lwd = 2, lty = 2)
    legend("topleft", inset = 0.01, title, bty = "n")
    legend("topright", inset = 0.01, c("lea", "lag"), lty = 1:2,
        col = c("black", "red"))
}
f(Ecafc, Ecfac, "Escherichia\ncoli")
f(Bsafc, Bsfac, "Bacillus\nsubtilis")
f(Hiafc, Hifac, "Haemophilus\ninfluenzae")
f(Mgafc, Mgfac, "Mycoplasma\ngenitalium")
par(opar)
```



Le premier facteur de l'AFC permet-il de bien discriminer entre les deux groupes ? Les facteurs suivants le permettent ils ? Est-ce l'objectif de l'AFC ?

2.3 AFC discriminantes

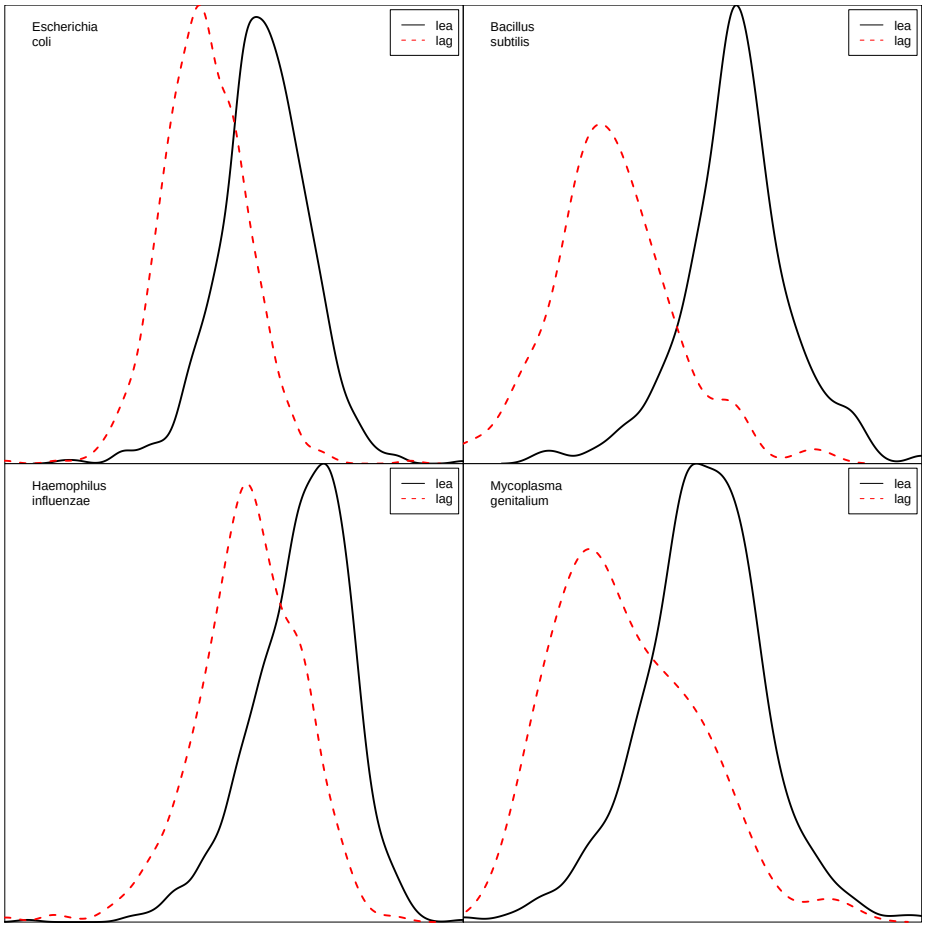
Lancer les calculs avec ce script :

```
cat(readLines("http://pbil.univ-lyon1.fr/members/lobry/repro/cabios96/runNucl.r"),
    sep = "\n")
```

```
library(ade4)
Bscda <- discrimin(dudi.coa(Bs, scann = FALSE, nf = 1), Bsfac, scann = FALSE, nf = 1)
Eccda <- discrimin(dudi.coa(Ec, scann = FALSE, nf = 1), Ecfac, scann = FALSE, nf = 1)
Hicda <- discrimin(dudi.coa(Hi, scann = FALSE, nf = 1), Hifac, scann = FALSE, nf = 1)
Mgcda <- discrimin(dudi.coa(Mg, scann = FALSE, nf = 1), Mgfac, scann = FALSE, nf = 1)
```

```
source("http://pbil.univ-lyon1.fr/members/lobry/repro/cabios96/runNucl.r")
```

Représenter la distribution des coordonnées des individus sur l'axe discriminant en fonction des deux groupes.



2.4 Interprétation des résultats

Qu'est ce qui caractérise les séquences codantes des deux groupes ? Indice : examiner par acide aminé les scores des codons en fonction de la nature de la base en troisième position.

Références

[1] J.R. Lobry and C. Gautier. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Research*, 22 :3174–3180, 1994.

[2] G. Perrière, J.R. Lobry, and J. Thioulouse. Correspondence discriminant analysis : a multivariate method for comparing classes of protein and nucleic acid sequences. *Computer Applications in the Biosciences*, 12 :519–524, 1996.