ANALYSE FACTORIELLE DISCRIMINANTE

CHERKAOUI Manal
ROUSSAFI Mariam
BELLOUALI Anas

Université Hassan 2 – Faculté des sciences Ain Chock

Juin 2022

PLAN :

Introduction

Le résumé de l'information a toujours été une chose d'intérêt dans plusieurs domaines tels que la biologie, la physique, l'informatique, l'économie ou la gestion. Ainsi, plusieurs méthodes d'analyse des données ont été développées à l'effet de réduire les nombres des variables, voire même le nombre d'individus sans perte d'information et afin de rendre l'interprétation plus facile et la représentation graphique plus aisée.

Définition :

L'analyse discriminante est une technique d'analyse des données qui vise à décrire, expliquer et prédire l'appartenance d'un individu à des groupes prédéfinis. À l'origine, cette méthode a été étudiée par Ronald Fisher dès 1936, dans le but de reconnaitre le type d'iris (setosa, virginica, et versicolor) à l'aide de la longueur et la largeur de ses pétales et sépales.

Aspects :

Précisons aussi que, la technique d'analyse discriminante donne lieu à deux principales approches.

(i) D'une part, l'analyse factorielle discriminante (ou analyse discriminante descriptive), qui est une méthode factorielle ou descriptive (comme l'ACP et l'AFC), qui a pour but de proposer un nouveau système de représentation, des variables latentes formées à partir de combinaisons linéaires des variables prédictives, qui permettent de discerner le plus possible les groupes d'individus.

(ii) D'autre part, l'analyse discriminante linéaire, est une méthode prédictive consistant à construire une fonction de classement (règle d'affectation, . . .) permettant de prédire la classe dans lequel appartient un individu à partir des valeurs prises par les variables prédictives.
Ces deux approches correspondent grosso-modo à la distinction entre méthodes géométriques et méthodes probabilistes.

Méthodes Géométriques

Ces méthodes, essentiellement descriptives, ne reposent que sur des notions de distance et ne font pas intervenir d'hypothèses probabilistes.
L'idée s'agit de calculer la distance entre la nouvelle observation et le centre de chacun des groupes. On classera la nouvelle observation dans le groupe pour lequel cette distance est minimale.

Variances Interclasse et Intraclasse

Données-Notations :

Les n individus e_i de l'échantillon constituent un nuage E de R^p, partagé en k sous-nuages : $E_1, E_2, ..., E_k$, de centres de gravité $g_1, g_2, ..., g_k$ et de matrices de de variances $V_1, V_2, ..., V_k$.

Avec :

g : centre de gravité de E (centre global)

V : matrice de variance de E (variance totale)

D : matrice diagonale des poids $p_1, p_2, ..., p_n$ de n individus e_i.

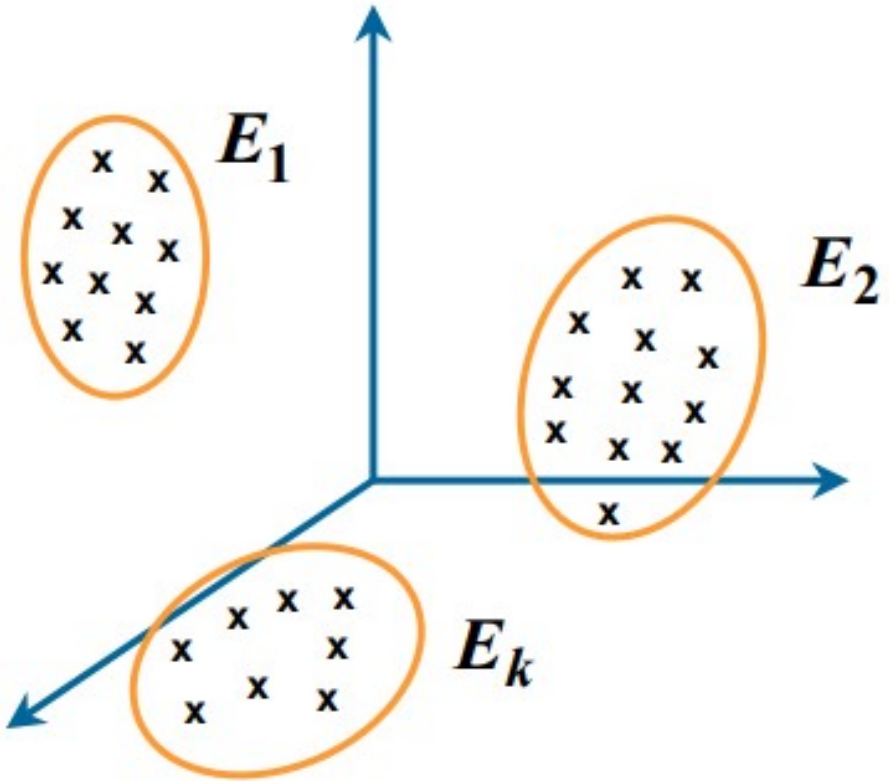
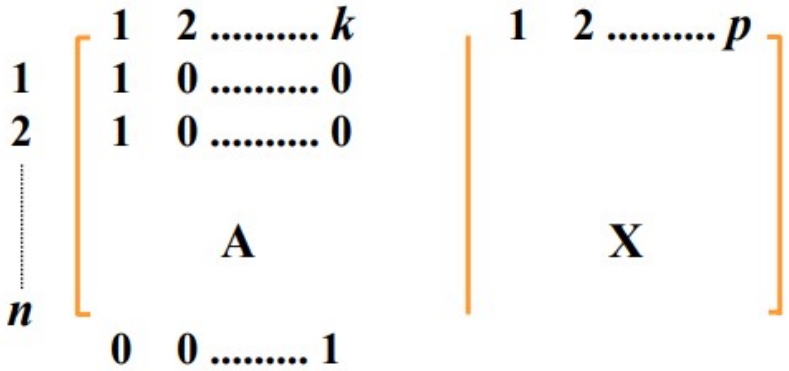


Figure – Représentation des nuages des points

Notations Matricielles

Tableau de données :



A : matrice des indicatrices de la variable qualitative à prédire
X : matrice des p variables explicatives
$D_q = A'DA$: matrice diagonale des poids q_j des sous nuages
$(A'DA)^{-1}(A'DX)$: ses lignes sont les coordonnées des k centres de gravité $g_1, g_2, ..., g_k$

- Poids de la j ème classe (sous-nuage) :

$$q_j = \sum_{e_i \in E_j} p_i$$

- Centres de gravité :

$$g_j = \frac{1}{q_j} \sum_i p_i e_i, \quad pour \quad e_i \in E_j$$

$$g = \sum_{j=1}^{k} q_j g_j$$

- Matrice de variance-covariance de la classe E_j :

$$V_j = \frac{1}{q_j} \sum_{e_i \in E_j} p_i (e_i - g_j)(e_i - g_j)'$$

Matrice de variance interclasse :

La matrice de variance B des k centres de gravité affectés des poids q_j :

$$B = \sum_{j=1}^{k} q_j (g_j - g)(g_j - g)'$$

Cette matrice rend compte de la dispersion des centres de gravité des sous-nuages autour du centre global g.

Matrice de variance intraclasse W :

La moyenne des matrices V_j :

$$W = \sum_{j=1}^{k} q_j V_j$$

Cette matrice rend compte de la dispersion à l'intérieur des sous-nuages.

On a alors la relation suivante :

$$V = W + B$$

Variance totale = Moyenne des variances + Variance des moyennes

Dans le cas où $p_i = 1/n$ et $g = 0$ les expressions précédentes se simplifient et en introduisant les effectifs $n_1, n_2, ..., n_k$ des k sous-nuages, on a :

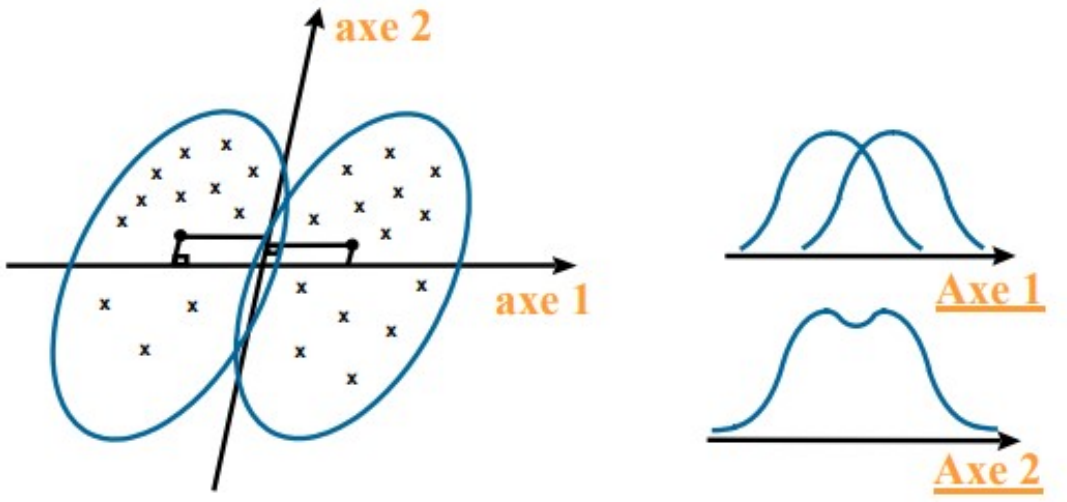
$$B = \frac{1}{n} \sum_j n_j g_j g_j' \ ; \ g_j = \frac{1}{n_j} \sum_{E_j} e_i$$

$$W = \frac{1}{n} \sum_j n_j V_j$$

Nous supposerons désormais être dans ce cas.

Les axes et variables discriminantes

L'AFD consiste à rechercher de nouvelles variables (les **variables discriminantes**) correspondant à des directions de $\mathbb{R}^p$ qui séparent le mieux possible en projection les k groupes d'observations.



L'axe 1 de la figure possède un bon pouvoir discriminant tandis que l'axe 2 (qui est l'axe principal usuel) ne permet pas de séparer en projection les deux groupes.

Supposons $\mathbb{R}^p$ muni d'une métrique M. On notera a l'axe discriminant, u le facteur associé tel que u = Ma et la variable discriminante sera Xu.
En projection sur l'axe a, les k centres de gravité doivent être aussi séparés que possible, tandis que chaque sous-nuage doit se projeter de manière groupée autour de la projection de son centre de gravité.
En d'autres termes, l'inertie du nuage des g_j projetés sur a doit être maximale. La matrice d'inertie du nuage des g est MBM, l'inertie du nuage projeté sur a est $a'MBMa$ si a est M-normé à 1 ($a^{'}Ma = 1$).

Il faut aussi qu'en projection sur a, chaque sous-nuage reste bien groupé, donc que $a'MV_jMa$ soit faible pour $j = 1, 2, ..., k$.
On cherchera donc à minimiser la moyenne $\sum_{j=1}^{k} q_j a'MV_jMa$ soit $a'MWMa$.
Or la relation $V = B + W$ entraîne que $MVM = MBM + MWM$, donc que $a'MVMa = a'MBMa + a'MWMa$

Critère

On prendra alors comme critère, la maximisation du rapport de l'inertie interclasse à l'inertie totale :

$$\max_a \frac{a'MBMa}{a'MVMa}$$

Ce maximum est atteint si **a** est vecteur propre de $(MVM)^{-1}(MBM)$ associé à sa plus grande valeur propre λ_1 :

$$M^{-1}V^{-1}BMa = \lambda_1 a$$

A l'axe discriminant **a** on associe le facteur discriminant **u** tel que $u = Ma$
On a alors

$$V^{-1}Bu = \lambda_1 u$$

Donc les variables discriminantes **Xu** sont indépendantes de la métrique **M**.

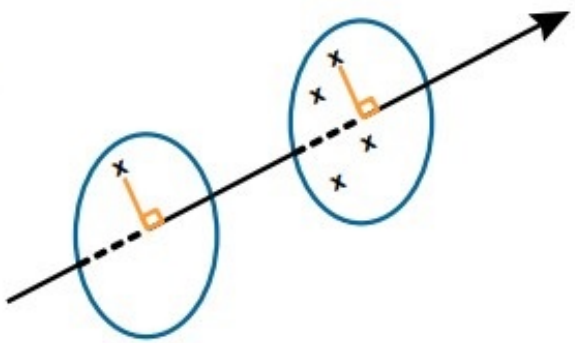
On choisira par commodité $M = V^{-1}$

$$\begin{cases} BV^{-1}a = \lambda_1 a. \\ V^{-1}Bu = \lambda_1 u \end{cases}$$

On a toujours $0 \leq \lambda_1 \leq 1$ car λ_1 est la quantité à maximiser.

Cas Particuliers

Cas $\lambda_1 = 1$:

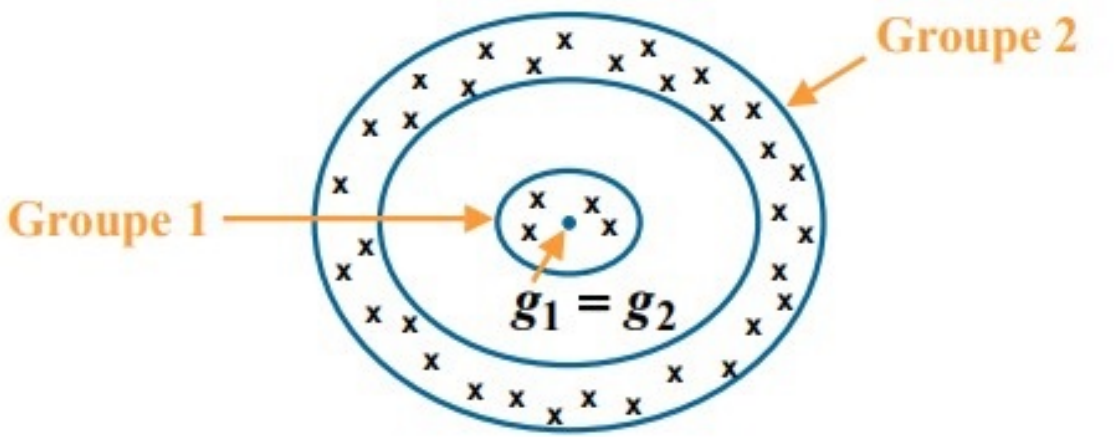


En projection sur **a** les dispersions intra-classes sont nulles. Les **k** nuages sont donc chacun dans un hyperplan orthogonal à **a**.
Il y a discrimination parfaite si les centres de gravité se projettent en des points différents.

Cas $\lambda_1 = 0$:

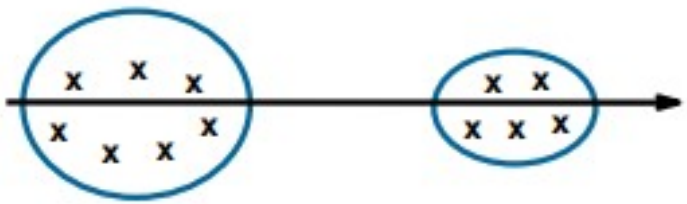
Le meilleur axe ne permet pas de séparer les centres de gravité g_i, c'est le cas où ils sont confondus.

Les nuages sont donc concentriques et aucune séparation linéaire n'est possible.



Autres Propriétés

La valeur propre λ mesure le pouvoir discriminant d'un axe.



$\lambda < 1$ mais les groupes sont bien séparés

Le nombre des valeurs propres non nulles, donc d'axes discriminants est égal à $k-1$ dans le cas habituel où $n > p > k$ et où les variables ne sont pas liées par des relations linéaires.

Remarque : Le cas de deux groupes

Il n'y a qu'une seule variable discriminante puisque $2 - 1 = 1$.
L'axe discriminant est alors nécessairement la droite reliant les deux centres de gravité g_1 et g_2 :

$$a = g_1 - g_2$$

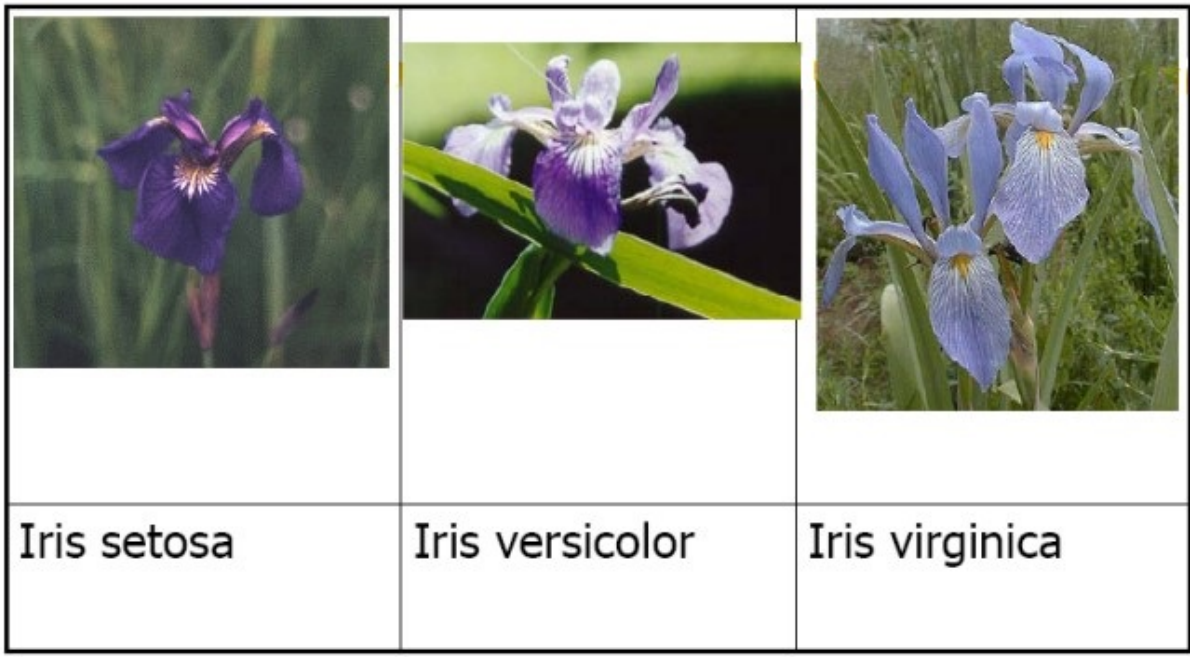
Le facteur discriminant **u** vaut donc :

$$u = V^{-1}(g_1 - g_2)$$

ou $u = W^{-1}(g_1 - g_2)$ qui lui est proportionnel.
$W^{-1}(g_1 - g_2)$ est la fonction de Fisher

Exemple : Les iris de Ficher

Ce fameux exemple sert de jeu d'essai. Les données concernent trois espèces d'iris (setosa, versicolor, virginica) représentées chacune par 50 individus décrits par 4 variables (longueur et largeur des pétales et sépales).



Il y a donc uniquement deux axes discriminants ce qui permet une représentation plane.

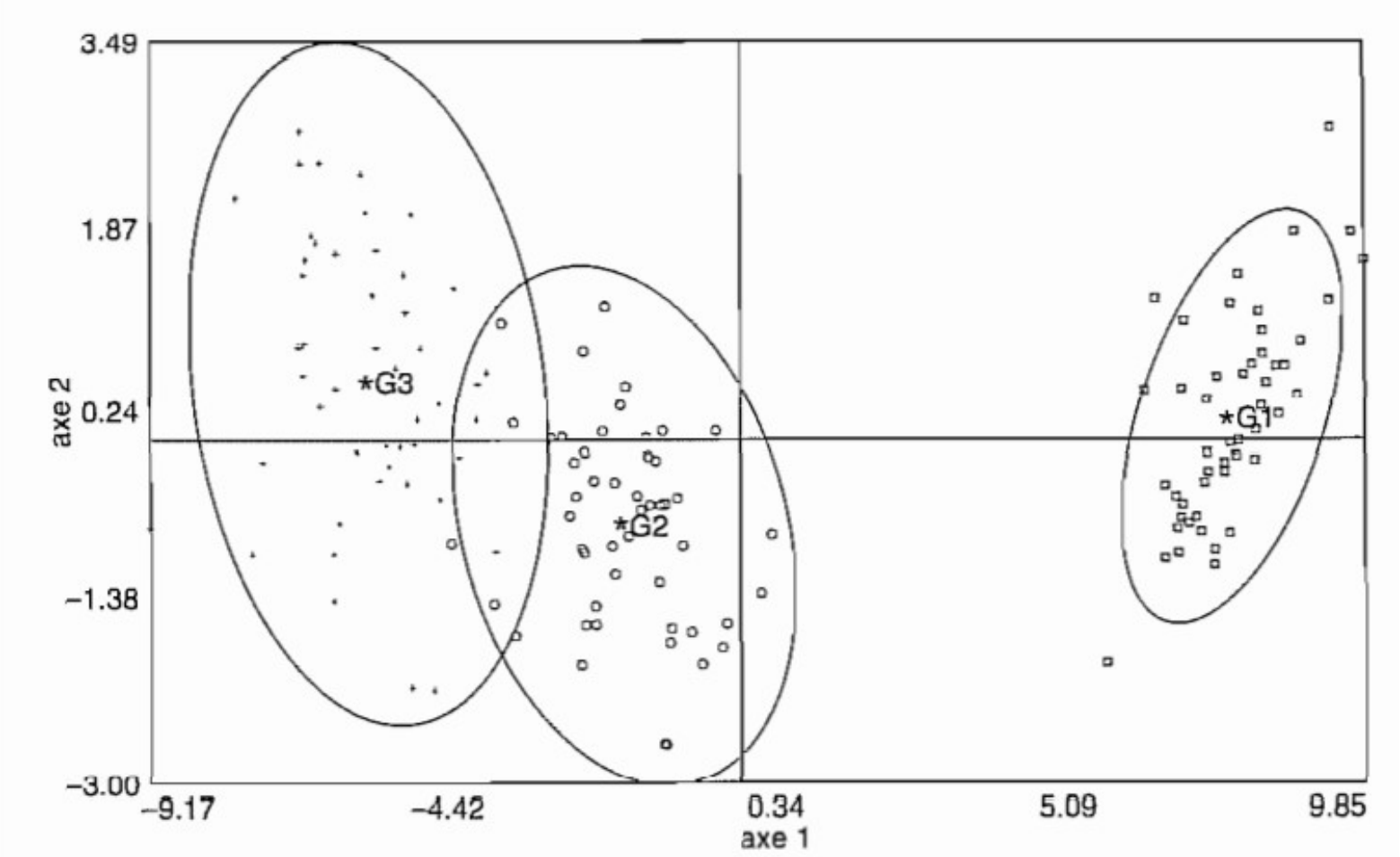


Figure – Plan discriminant des iris de Fisher

Règle Géométrique d'affectation

Ayant trouvé la meilleure représentation de la séparation en k groupes des n individus, on peut alors chercher à affecter une observation e à l'un des groupes.
La règle naturelle consiste à calculer les distances de l'observation à classer à chacun des k centres de gravité et à affecter selon la distance la plus faible.

Question posée : Quelle métrique à utiliser ?

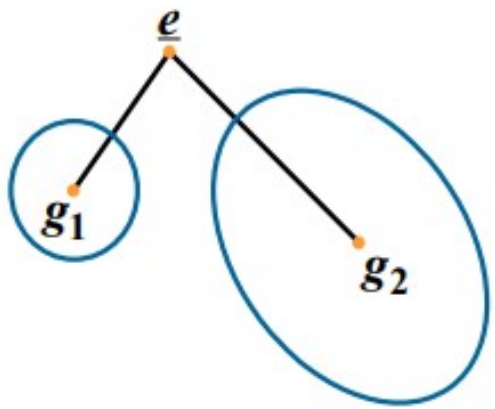
Règle de Mahalanobis Fisher :
On utilise W^{-1}

$$d^2(e, g_i) = (e - g_i)' W^{-1} (e - g_i)$$

Remarque

L'utilisation de la règle précédente conduit à des affectations incorrectes lorsque les dispersions des groupes sont très différentes entre elles : rien ne justifie alors l'usage de la même métrique pour les différents groupes.

Exemple :



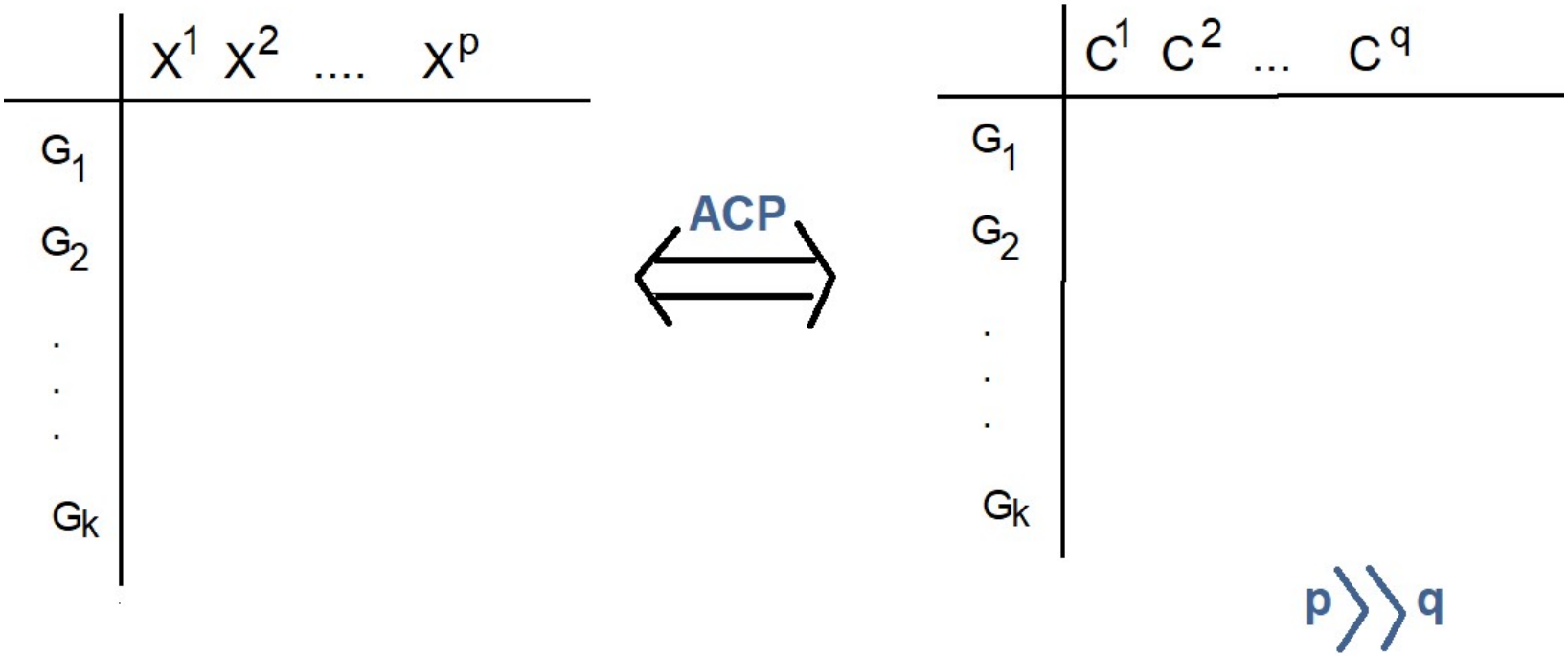
e plus proche de g_1 que de g_2 au sens habituel. Pourtant, il est plus naturel d'affecter e à la deuxième classe qu'à la première dont le pouvoir d'attraction est moindre.

Solution du Problème : Chercher une métrique locale M_i.

Dans la plupart des cas, on choisit M_i proportionnel à V_i^{-1}

Remarque

L'AFD peut être vue comme une ACP (Analyse en Composantes Principales) des centres de gravités g_k avc la métrique V^{-1}



Introduction

L'ensemble de données Breast Cancer Wisconsin du [* UCI Machine learning repo *] est un ensemble de données de 32 variables mesurant la taille et la forme des noyaux cellulaires, l'objectif est de créer un modèle qui nous permettra de prédire si une cellule cancéreuse du sein est bénigne ou malin. nous commencerons notre analyse par une technique de réduction de dimensionnelle L'Analyse en Composantes Principales (ACP) . puis l'Analyse Factorielle Discriminante (AFD) .

Données

Les caractéristiques sont calculées à partir d'une image numérisée d'une aspiration à l'aiguille fine (FNA) d'une masse mammaire. Ils décrivent les caractéristiques des noyaux cellulaires présents dans l'image. Notre ensemble de données comprend 569 observations et 32 variables. Il y a une variable d'identification, une variable de diagnostic révélant s'ils étaient bénins ou malins, et 30 variables de mesure détaillant la taille et la forme des noyaux cellulaires. Le diagnostic, variable catégorielle, est notre variable réponse et les 30 variables de mesure, toutes continues, sont nos variables explicatives potentielles de notre modèle.

Données

Jetons un coup d'oeil :

data.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id	diagnosis	radius_me	texture_m	perimeter_	area_mea	smoothne	compactn	concavity_	concave p	symmetry_	fractal_dir	radius_se	textu
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7
4	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7
5	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1
6	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7
7	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8
8	844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7
9	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1
10	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1
11	84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1
12	845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1
13	84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9
14	846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3
15	846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1

Données

importation de données

Notre variable de réponse est le diagnostic : bénin (B) ou malin (M). Nous avons 3 0 variables numériques
Collectons d'abord toutes les 30 variables numériques dans une matrice et créons un vecteur de diagnostic où M = 1 et B = 0

```
5
6  wdbc <- read.csv("data.csv")
7  # Convertir les caractéristiques des données dans la matrice wdbc.data
8  wdbc.data <- as.matrix(wdbc[,c(3:32)])
9  # Définir les noms de ligne de wdbc.data
10 row.names(wdbc.data) <- wdbc$id
11 # Créer un vecteur de diagnostic
12 diagnosis <- as.numeric(wdbc$diagnosis == "M")
13
```

L'analyse exploratoire

Moyennes et écarts-types

Répondons à quelques questions de base :

```
15 #Combien d'observations ont un diagnostic bénin ou malin ?
16 table(wdbc$diagnosis)
17 #Quelle est la moyenne de chacune des colonnes numériques ?
18 round(colMeans(wdbc.data),2)
19 #Quelle est l'écart type de chacune des colonnes numériques ?
20 roundSD <- function(x){
21   round(sd(x), 2)
22 }
23 apply(wdbc.data, 2, roundSD)
24
```

L'analyse exploratoire

Moyennes et écarts-types

les moyennes :

```
  B   M
357 212
> round(colMeans(wdbc.data),2)
            radius_mean            texture_mean          perimeter_mean               area_mean
                  14.13                   19.29                   91.97                  654.89
         smoothness_mean         compactness_mean          concavity_mean     concave.points_mean
                   0.10                    0.10                    0.09                    0.05
          symmetry_mean  fractal_dimension_mean               radius_se              texture_se
                   0.18                    0.06                    0.41                    1.22
           perimeter_se                 area_se            smoothness_se           compactness_se
                   2.87                   40.34                    0.01                    0.03
           concavity_se      concave.points_se               symmetry_se    fractal_dimension_se
                   0.03                    0.01                    0.02                    0.00
          radius_worst           texture_worst         perimeter_worst              area_worst
                  16.27                   25.68                  107.26                  880.58
        smoothness_worst        compactness_worst         concavity_worst    concave.points_worst
                   0.13                    0.25                    0.27                    0.11
         symmetry_worst fractal_dimension_worst
                   0.29                    0.08
> roundSD <- function(x){
```

L'analyse exploratoire

Moyennes et écarts-types

les écartes type :

```
> apply(wdbc.data, 2, roundSD)
          radius_mean           texture_mean         perimeter_mean              area_mean
                 3.52                   4.30                  24.30                 351.91
      smoothness_mean        compactness_mean        concavity_mean   concave.points_mean
                 0.01                   0.05                   0.08                   0.04
       symmetry_mean  fractal_dimension_mean           radius_se            texture_se
                 0.03                   0.01                   0.28                   0.55
        perimeter_se                area_se         smoothness_se        compactness_se
                 2.02                  45.49                   0.00                   0.02
        concavity_se       concave.points_se          symmetry_se  fractal_dimension_se
                 0.03                   0.01                   0.01                   0.00
        radius_worst          texture_worst        perimeter_worst             area_worst
                 4.83                   6.15                  33.60                 569.36
     smoothness_worst       compactness_worst       concavity_worst  concave.points_worst
                 0.02                   0.16                   0.21                   0.07
       symmetry_worst fractal_dimension_worst
                 0.06                   0.02
```

L'analyse exploratoire

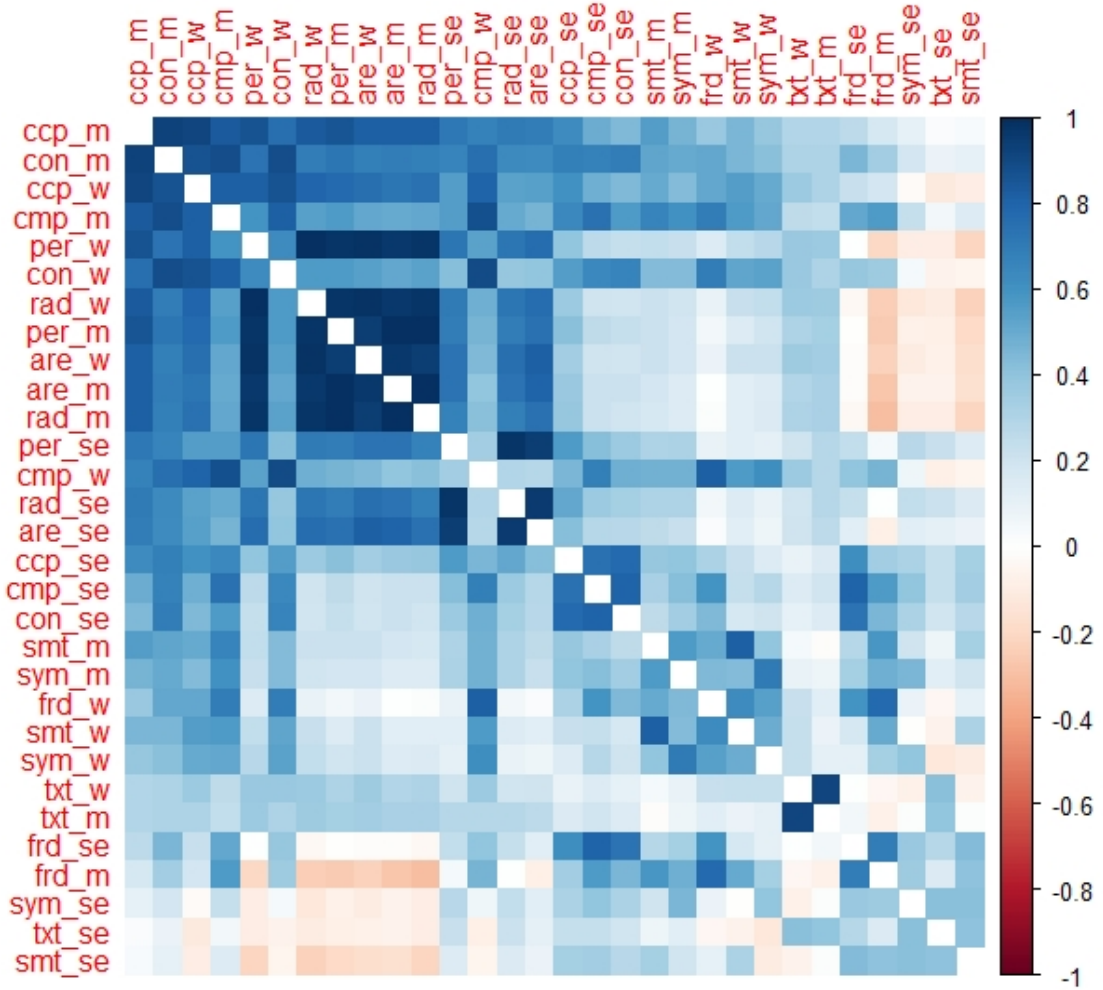
corrélation

Comment les variables sont-elles liées les unes aux autres ?

```
29
30 library(corrplot)
31
32 corMatrix <- wdbc[,c(3:32)]
33 cNames <- c("rad_m","txt_m","per_m",
34              "are_m","smt_m","cmp_m","con_m",
35              "ccp_m","sym_m","frd_m",
36              "rad_se","txt_se","per_se","are_se","smt_se",
37              "cmp_se","con_se","ccp_se","sym_se",
38              "frd_se","rad_w","txt_w","per_w",
39              "are_w","smt_w","cmp_w","con_w",
40              "ccp_w","sym_w","frd_w")
41 colnames(corMatrix) <- cNames
42
43 # Créer la matrice de corrélation
44 M <- round(cor(corMatrix), 2)
45
46 #Créer corrplot
47 corrplot(M, diag = FALSE, method="color", order="FPC", tl.srt = 90)
48
```

L'analyse exploratoire

corrélation



ACP

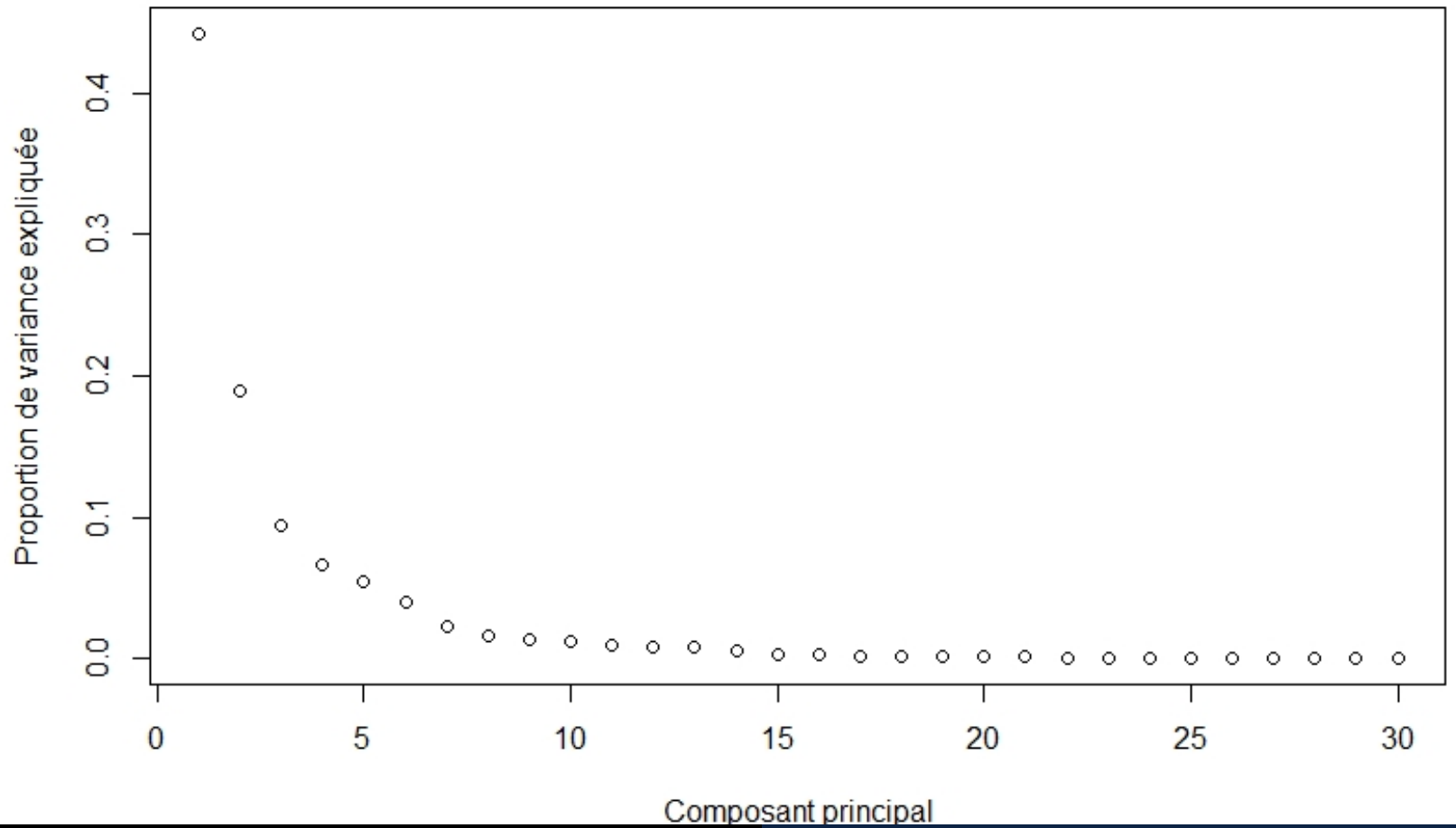
Script

En utilisant l'ACP l'aide de prcomp() , nous pouvons combiner nos nombreuses variables en différentes combinaisons linéaires qui expliquent chacune une partie de la variance du modèle.

```
51
52 # ACP à l'aide de prcomp() fonction
53 wdbc.pr <- prcomp(wdbc.data, scale = TRUE, center = TRUE)
54 summary(wdbc.pr)
55 par(mfrow = c(1, 2))
56 # Calculer la variabilité de chaque composant
57 pr.var <- wdbc.pr$sdev ^ 2
58 # Variance expliquée par chaque composante principale : pve
59 pve <- pr.var/sum(pr.var)
60 # Valeurs propres
61 round(pr.var, 2)
62 # Pourcentage d'écart expliqué
63 round(pve, 2)
64 # Pourcentage cumulé expliqué
65 round(cumsum(pve), 2)
66
67 plot(pve, xlab = "Composant principal",
68      ylab = "Proportion de variance expliquée",
69      ylim = c(0, 1), type = "b")
70
```

ACP

Résultats



ACP

Résultats

le pourcentage d'information explique par chaque composantes est :

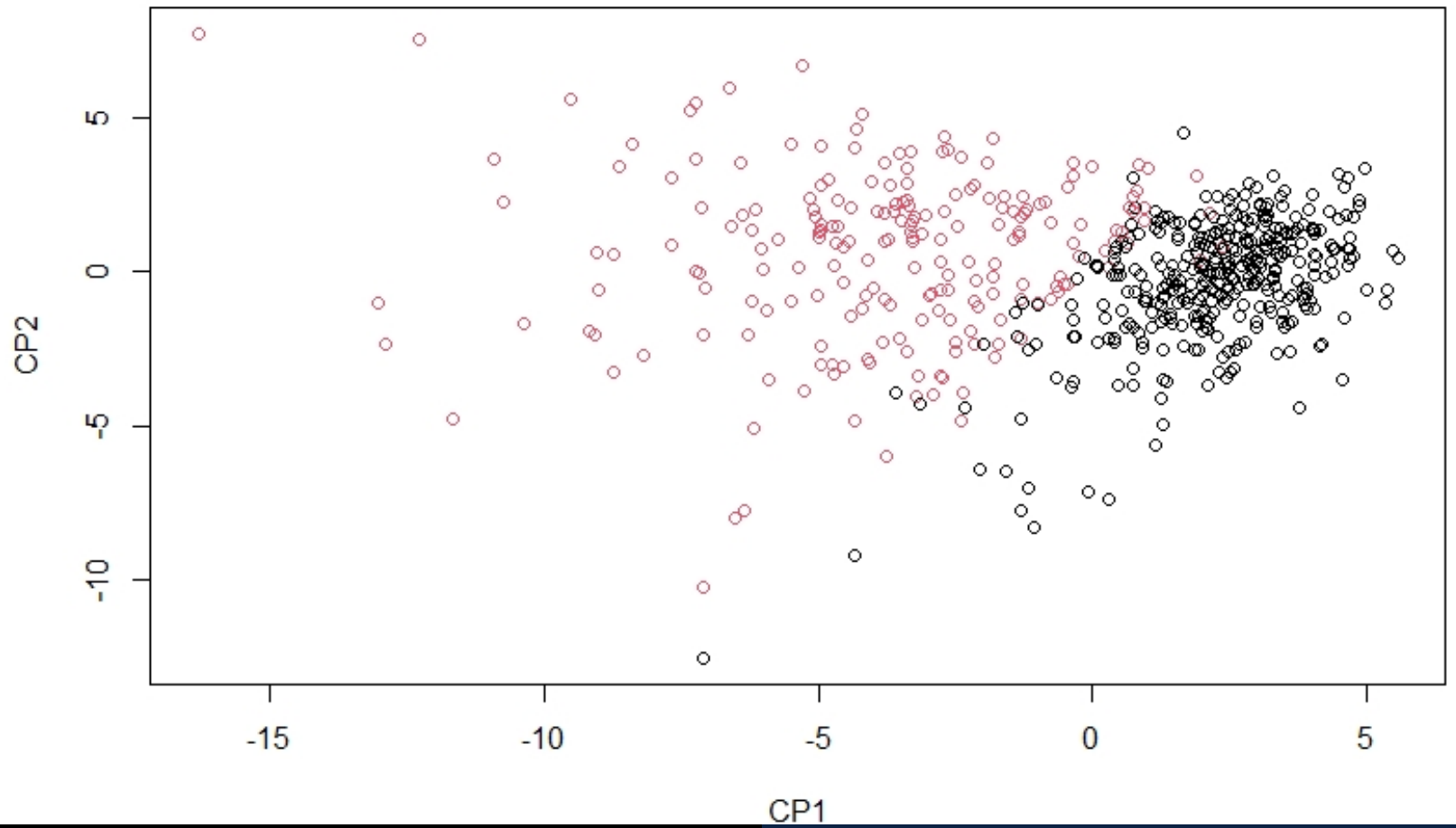
```
> round(pve, 2)
 [1] 0.44 0.19 0.09 0.07 0.05 0.04 0.02 0.02 0.01 0.01 0.01 0.01 0.01 0.01 0.00 0.00 0.00 0.00
[19] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
```

Remarque

89 pour-cent de l'information est expliquée par les six premiers PC. De plus, les valeurs propres associées aux 6 premiers PC sont supérieures à 1. Nous utiliserons ce critère pour décider du nombre de PC à inclure dans la phase de construction du modèle linéaire.

ACP

Résultats



AFD

D'après les diagrammes de dispersion de la composante principale, il est évident qu'il existe un certain regroupement de points bénins et malins. Ceci suggère que nous pourrions construire une fonction discriminante linéaire en utilisant ces composantes principales. Maintenant que nous avons nos composantes principales choisies, nous pouvons effectuer l'analyse discriminante linéaire.

AFD

Script

la première étape consiste à projeter vos données sur les composants principaux, puis nous effectuons LDA sur le diagnostics

```
ls(wdbc.pr)
#projeter les données sur les composantes principale
wdbc.pcs <- wdbc.pr$x[,1:6]
head(wdbc.pcs, 20)
wdbc.pcst <- wdbc.pcs
#ajouter les diagnostics
wdbc.pcst <- cbind(wdbc.pcs, diagnosis)
head(wdbc.pcst)

#Effectuer LDA sur les diagnostics
library(MASS)
wdbc.pcst.df <- as.data.frame(wdbc.pcst)
wdbc.lda <- lda(diagnosis ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6, data = wdbc.pcst.df)
wdbc.lda
```

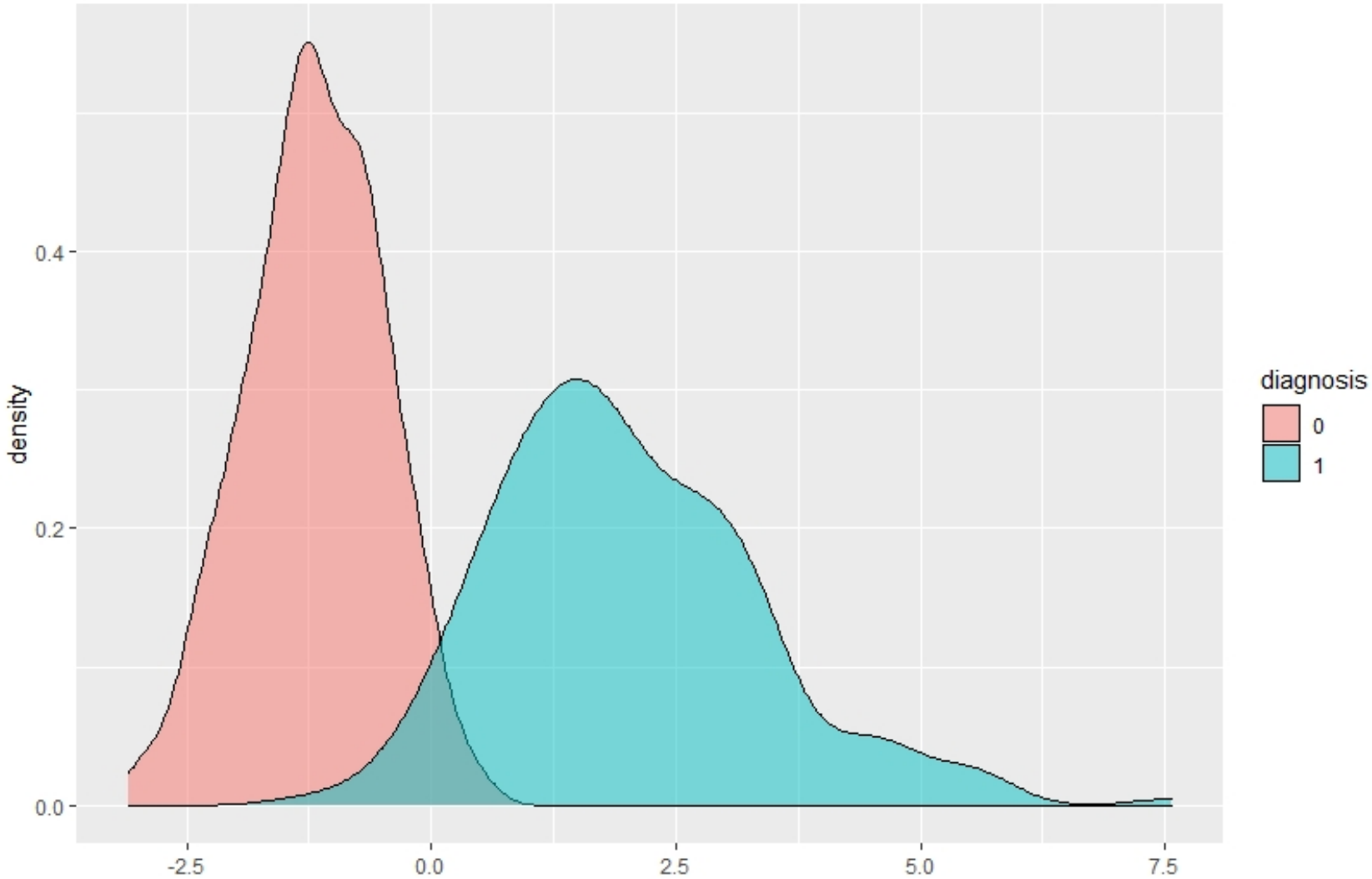
AFD

Résultats

```
Call:
lda(diagnosis ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6, data = wdbc.pcst.df)

Prior probabilities of groups:
        0         1
0.6274165 0.3725835

Group means:
        PC1        PC2        PC3        PC4         PC5          PC6
0  2.204035 -0.3459829  0.2129742  0.1382257 -0.09815733  0.009665194
1 -3.711511  0.5826221 -0.3586405 -0.2327669  0.16529323 -0.016275822

Coefficients of linear discriminants:
           LD1
PC1 -0.4726182
PC2  0.1731337
PC3 -0.2152467
PC4 -0.1987586
PC5  0.1695570
PC6 -0.0227991
```

AFD

Résultats



Conclusion

L'AFD est une technique statistique qui vise à décrire, expliquer et prédire l'appartenance d'un ensemble d'observations à des groupes prédéfinis et est utilisée dans de nombreux domaines tel que :
- La médecine, par exemple pour détecter les groupes à hauts risques cardiaques à partir de caractéristiques telles que l'alimentation, le fait de fumer ou non, les antécédents familiaux, etc.
- Le domaine bancaire, lorsque l'on veut évaluer la fiabilité d'un demandeur de crédit à partir de ses revenus, du nombre de personnes à charge, des encours de crédits qu'il détient, etc.

Références bibliographiques

[1] G.SAPORTA , Probabilités, analyse de données et statistique, TECHNIP, 27 rue Ginoux, 75737 PARIS Cedex 15, France, 2006.
[2] G.KAMINGU , Analyse factorielle discriminante , Université de Kinshasa , Mai 2016 (Article).
[3] https ://www.techno-science.net/glossaire-definition/Analyse-discriminante.html.
[4] http ://www.jybaudot.fr/Analdonnees/afdlin.html