Méthodes de discrimination

Partie III - Analyse discriminante

Discriminant analysis

2021

Brigitte Gelein – *bgelein@ensai.fr*

ENSAI

Sommaire

Sommaire

Introduction

L'analyse discriminante :

Discriminer les individus entre deux ou plusieurs groupes, définis par les modalités d'une **variable qualitative** Y **«expliquée»**, à partir de J **variables quantitatives « explicatives »**, ou « prédicteurs ».

Les groupes d'individus sont définis *a priori*, et l'analyse discriminante permet de caractériser ces groupes à l'aide des prédicteurs.

Travaux fondateurs de Fisher R.A. en 1936 :
The Use of Multiple Measurements in Taxonomics Problems,
Annals of Eugenics, 179-188, 1936

Introduction

Approche « descriptive » de l'Analyse Discriminante : Etant donnée une population partagée en plusieurs groupes et disposant par ailleurs de variables quantitatives explicatives relevées sur cette même population, quelle(s) **combinaison(s) linéaire(s) de ces variables quantitatives** permet(tent) de rendre compte de manière optimale de la partition initiale ?

=> représentation graphique sur un (des) axe(s) discriminants

Approche « décisionnelle » de l'Analyse Discriminante : Elaboration d'une **règle de décision** permettant, au vu des valeurs prises par les variables quantitatives, de décider du groupe auquel affecter un nouvel individu pour lequel la réponse n'est pas, cette fois, connue *a priori*.

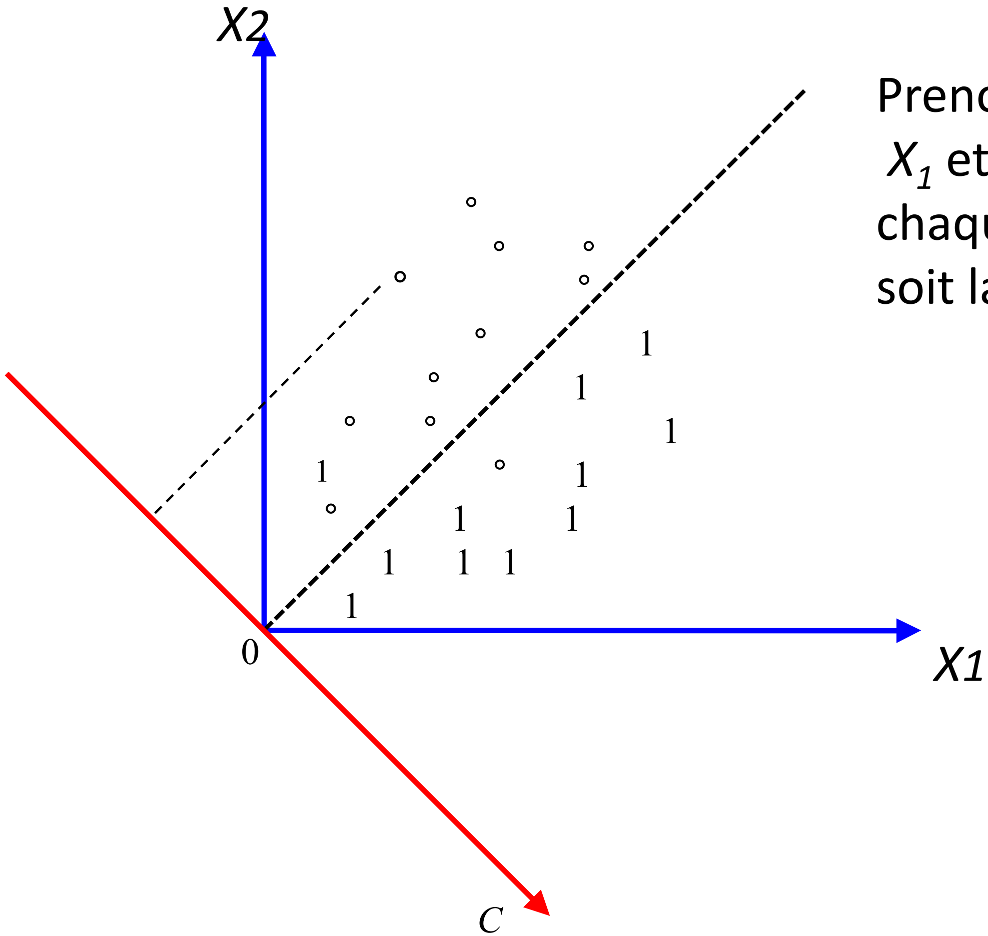
2 - Analyse Factorielle Discriminante linéaire

- **Dans un premier temps :**

  Détermination de la **combinaison linéaire des variables d'origine**, ici quantitatives, qui permette de rendre compte de manière optimale de la partition des individus, induite par les ***R*** modalités d'une variable qualitative *Y*.

- **Dans un second temps :**

  Elaboration d'une **règle de décision** permettant d'affecter un nouvel individu à l'un des ***R*** groupes.



Prenons le cas de deux variables continues X_1 et X_2 et une partition en deux groupes : chaque individu prend soit la valeur Y=« 0 », soit la valeur Y=« 1 ».

Nuage de points des valeurs prises par les individus sur deux variables quantitatives X_1 et X_2.

2 - Analyse Factorielle Discriminante linéaire

2.1 - Notations

Soit un ensemble de n observations indicées par i = 1 à n,
sur lesquelles sont relevées J variables quantitatives X_j, j = 1 à J.

La valeur prise par la variable j sur l'observation i est notée x_{ij}.

Par ailleurs, il existe une partition connue *a priori* des n observations en *R* groupes indexés par r, r=*1 à R*.

Soit p_i le poids associé à l'observation i, avec $\sum_{i=1}^{n} p_i = 1$.

2 - Analyse Factorielle Discriminante linéaire

- g le barycentre du nuage des n individus

  g =[$g_1 \; \dots g_j \dots g_J$] avec $g_j = \sum_{i=1}^{n} p_i x_{ij}$.

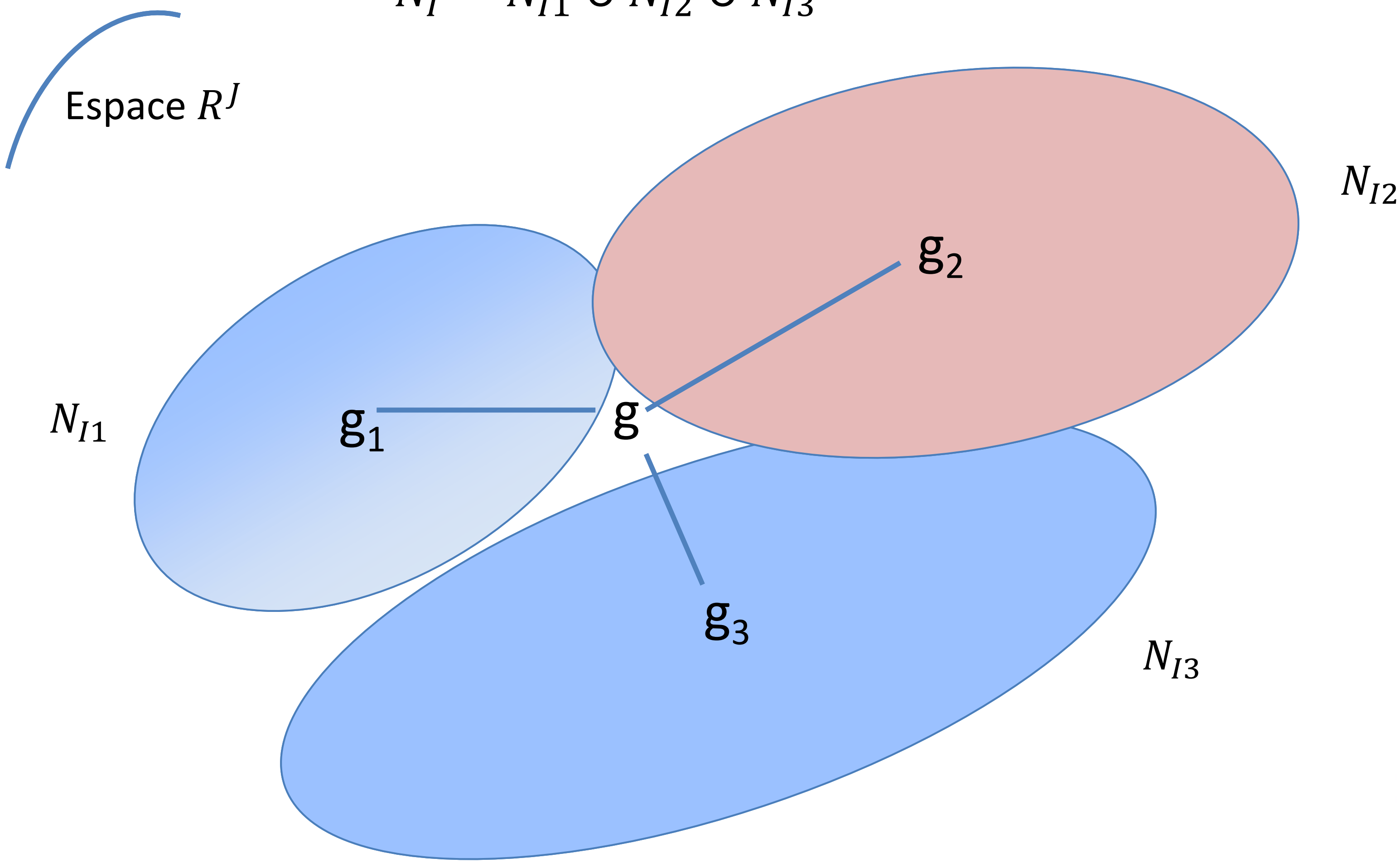
- p_r le poids du groupe r avec $p_r = \sum_{i \in G_r} p_i$

- g_r le barycentre du groupe r , g_r =[$g_{1r} \; \dots g_{jr} \dots g_{Jr}$]

Avec $$g_{jr} = \frac{\sum_{i \in G_r} p_i x_{ij}}{\sum_{i \in G_r} p_i} = \frac{\sum_{i \in G_r} p_i x_{ij}}{p_r} .$$

Exemple d'un nuage de points N_I représentant les n observations dans R^J
N_I partitionné en 3 sous-nuages N_{I1}, N_{I2}, N_{I3}

$$N_I = N_{I1} \cup N_{I2} \cup N_{I3}$$

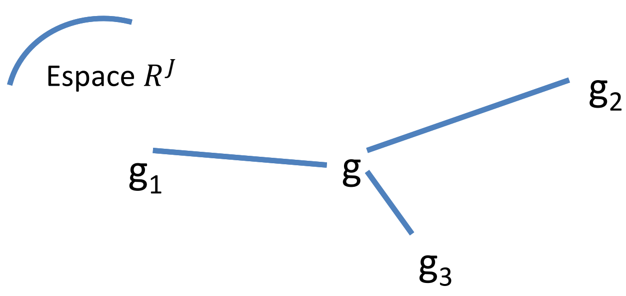


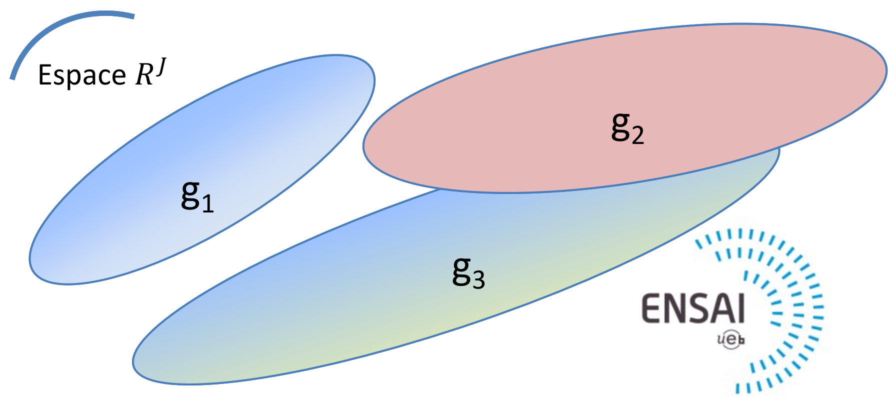
2 - Analyse Factorielle Discriminante linéaire

Décomposition de l'inertie (Huygens) dans R^J :

Inertie totale = Inertie inter-classes+ Inertie intra-classes

$$\sum_{i=1}^{n} p_i \|x_i - g\|_M^2 = \sum_{r=1}^{R} p_r \|g_r - g\|_M^2 + \sum_{r=1}^{R} \sum_{i \in G_r} p_i \|x_i - g_r\|_M^2$$





2 - Analyse Factorielle Discriminante linéaire

Matrice de variance-covariance inter-groupes

Soit B la matrice carrée d'ordre $\boldsymbol{J}$ de terme générique

$$B_{jj'} = \sum_{r=1}^{R} p_r \left(g_{jr} - g_j\right)\left(g_{j'r} - g_{j'}\right)$$

B est la matrice de variance-covariance entre les groupes mesurée sur les variables initiales X_j et $X_{j'}$ car $B_{jj'}$ désigne la covariance empirique entre les variables X_j et $X_{j'}$ mais mesurée sur les barycentres des groupes,

d'où $\boxed{B_{jj'} = COV_{INTER}\left(X_j, X_{j'}\right)}$

2 - Analyse Factorielle Discriminante linéaire

Matrice de variance-covariance intra-groupes

Soit W la matrice carrée symétrique d'ordre $\boldsymbol{J}$ et de terme générique :

$$W_{jj'} = \sum_{r=1}^{R} p_r \frac{1}{p_r} \sum_{i \in G_r} p_i (x_{ij} - g_{jr})(x_{ij'} - g_{j'r})$$

$$W_{jj'} = COV_{INTRA}(X_j, X_{j'})$$

2 - Analyse Factorielle Discriminante linéaire

Matrice de variance-covariance intra-groupes

- La variance intra-groupe totale est la somme des variances intra-groupe de chaque groupe, pondérée par les poids respectifs de chaque groupe.
- Les variances intra-groupes et les poids peuvent être différents d'un groupe à l'autre.

$$W = \sum_{r=1}^{R} p_r W_r$$

d'où

$$W_{r(jj')} = \frac{1}{p_r} \sum_{i \in G_r} p_i \left(x_{ij} - g_{jr}\right)\left(x_{ij'} - g_{j'r}\right)$$

2 - Analyse Factorielle Discriminante linéaire

Matrice de variance-covariance totale

Soit T la matrice carrée d'ordre J, de terme générique

$$T_{jj'} = \sum_{i=1}^{n} p_i \left(x_{ij} - g_j\right)\left(x_{ij'} - g_{j'}\right)$$

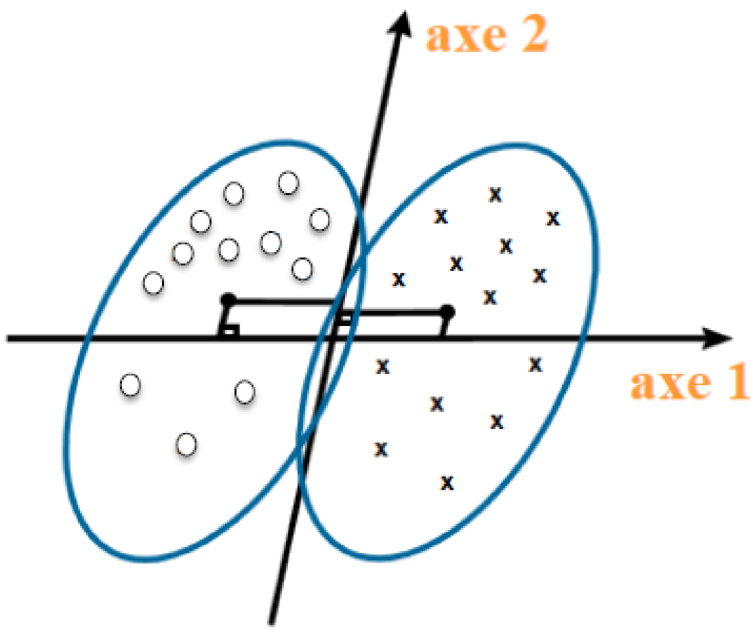
$$T_{jj'} = Cov\left(X_j ; X_{j'}\right)$$

D'après la décomposition de Huygens :

$$\boldsymbol{T = B + W}$$

2 - Analyse Factorielle Discriminante linéaire

Chercher un caractère synthétique C quantitatif (une combinaison linéaire des variables initiales X_j) rendant compte de manière optimale de la partition connue *a priori* en ***R*** groupes. Ici l'axe le plus discriminant est ?



2 - Analyse Factorielle Discriminante linéaire

- Soit u le vecteur d'ordre J de termes u_j .
- Les u_j sont les coefficients de la combinaison linaire des variables centrées donnant la variable synthétique C.

$c_i = \sum_{j=1}^{J} u_j x_{ij}$ est la valeur prise par l'individu i sur la variable *C=Xu*

Soit, sous forme matricielle, $\boldsymbol{C}=\boldsymbol{Xu}$ *avec* $\boldsymbol{X}$ *la matrice [n,J] de termes* x_{ij}

On suppose, sans perte de généralité, les variables X centrées : $\bar{c} = 0$.

$$Var(C) = Var_{INTRA}(C) + Var_{INTER}(C)$$

$$Var(C) = \sum_{r=1}^{R} p_r \frac{1}{p_r} \sum_{i \in G_r} p_i (c_i - \bar{c}_r)^2 + \sum_{r=1}^{R} p_r (\bar{c}_r - \bar{c})^2$$

2 - Analyse Factorielle Discriminante linéaire

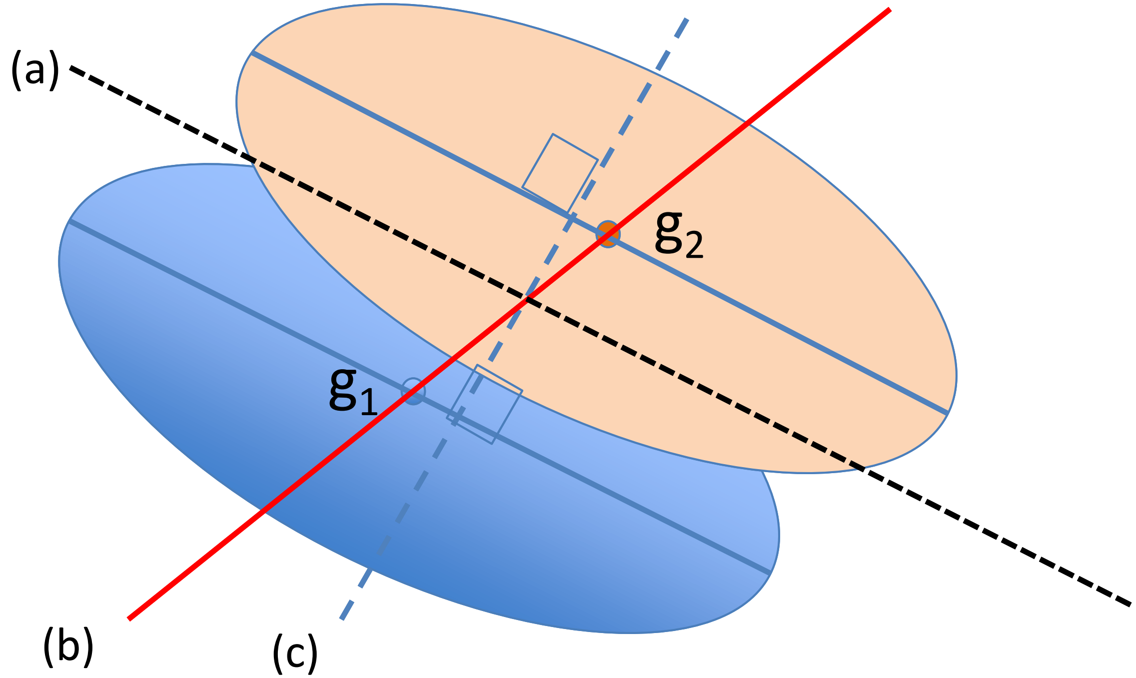
?

Variance totale (C) = variance Intra-groupes (C) + variance Inter-groupes (C)

Le problème revient donc à trouver C qui
(1) ________________________ la variance INTER
(2) ________________________ la variance INTRA.

Quel est l'axe qui vérifie la condition (1) dans le graphique suivant ?
La condition (2) ?

2 - Analyse Factorielle Discriminante linéaire



Nuage des individus dans R^J

Chaque ellipse représente un sous-groupe d'individus prenant une des 2 modalités de la variable qualitative à expliquer

2 - Analyse Factorielle Discriminante linéaire

Variance totale (C) = variance Intra-groupes (C) + variance Inter-groupes (C)

Le problème revient donc à trouver C qui simultanément

________________ la variance INTER

________________ la variance INTRA, soit encore qui

________________ le rapport $\frac{Variance \;\; Inter}{Variance \;\; Intra}$.

Si chaque observation du groupe 1 se **projette** sur C en 1 et en –1 pour le groupe 2 (voir graphique suivant)

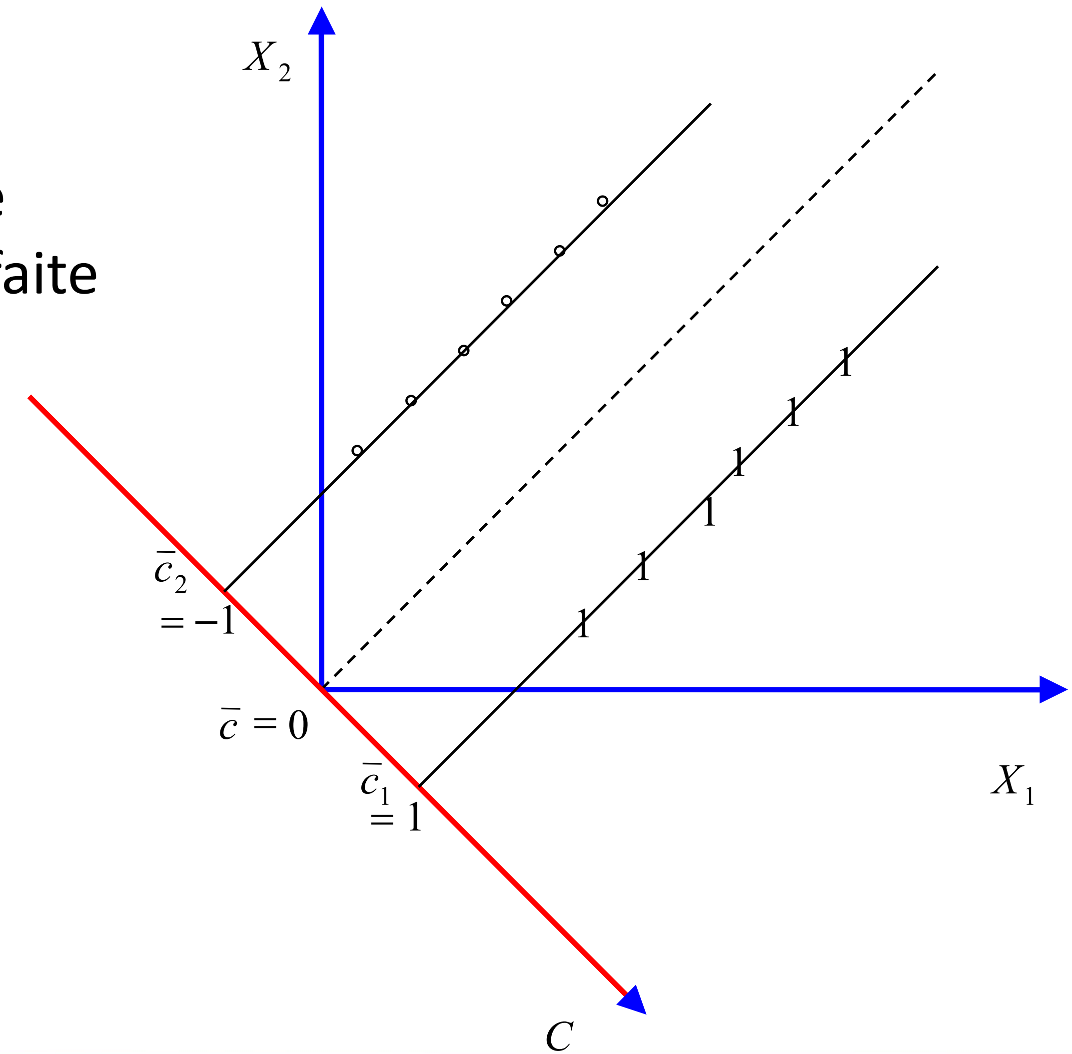
=> Séparation parfaite entre les deux groupes

=> Variance Intra-groupes (C) =

=> Variance Inter-groupes (C) =

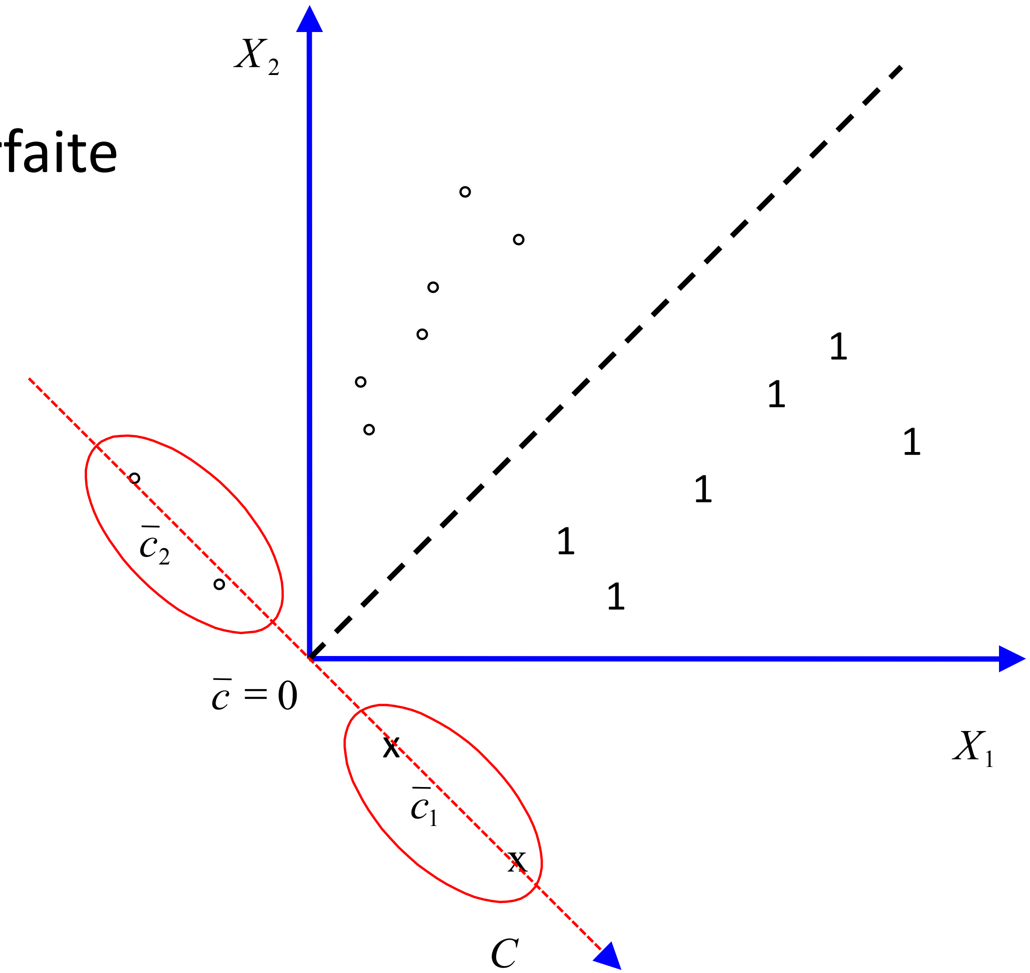
2 - Analyse Factorielle Discriminante linéaire

Cas extrême de séparation parfaite



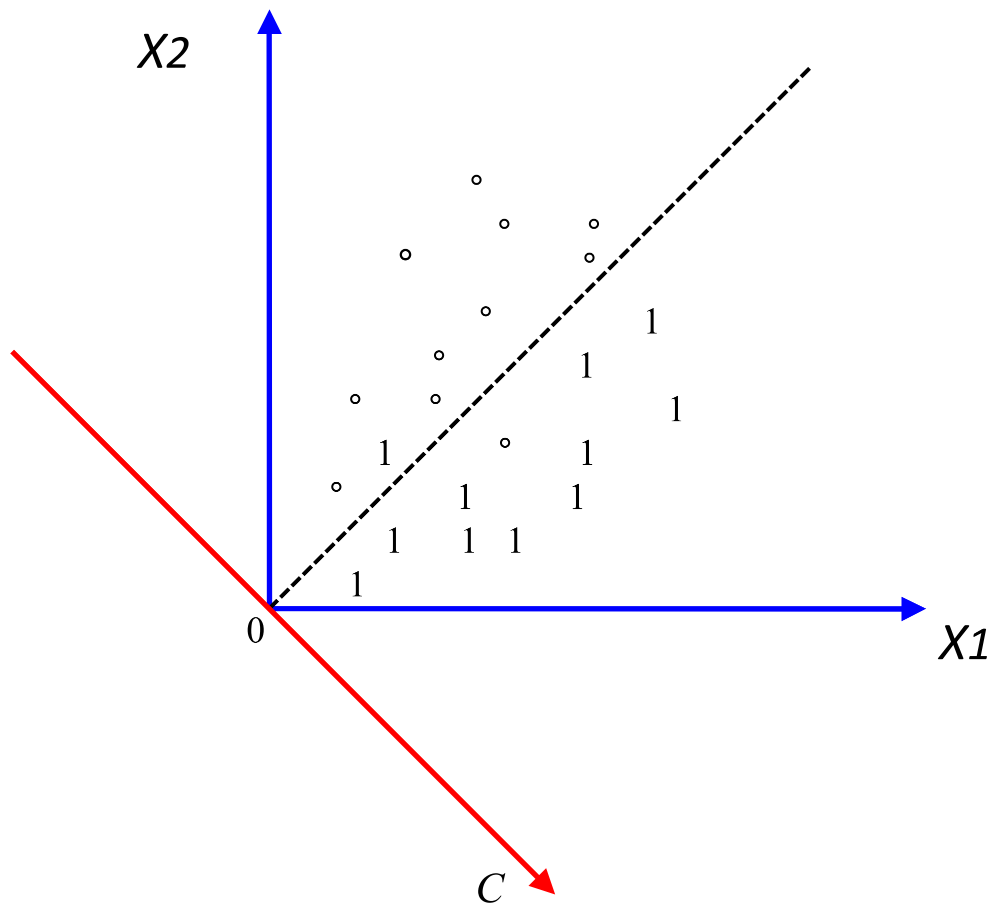
2 - Analyse Factorielle Discriminante linéaire

Séparation parfaite



2 - Analyse Factorielle Discriminante linéaire

Qualité de la séparation ?



2 - Analyse Factorielle Discriminante linéaire

Décomposition de l'inertie projetée sur le facteur U = décomposition de la variance de ***C***

$C = \sum_{j=1}^{J} u_j X_j$ soit sous forme matricielle $C = Xu$ où X est la matrice $[n, J]$ de terme x_{ij}.

Variance de C :

$$Var\,(C) = C^T C = (Xu)^T (Xu) = u^T X^T Xu = u^T Tu$$

Variance inter-classe de C :

$$u^T Bu = \sum_{j=1}^{J} \sum_{j'=1}^{J} u_j u_{j'} \sum_{r=1}^{R} p_r (g_{jr} - g_j)(g_{j'r} - g_{j'}) = Var_{INTER}(C)$$

2 - Analyse Factorielle Discriminante linéaire

Variance intra-classe de C :

$$u^T W u = \sum_{j=1}^{J} \sum_{j'=1}^{J} u_j u_{j'} \sum_{r=1}^{R} \sum_{i \in Gr} p_i (x_{ij} - g_{jr})(x_{ij'} - g_{j'r}) = Var_{INTRA}(C)$$

Variance de C :

$$Var\ (C) = u^T T u = u^T B u + u^T W u$$

Interpréter l'utilisation du rapport $\dfrac{Variance\ \ Inter}{Variance\ \ Intra}$

en termes de liaison entre variables

2 - Analyse Factorielle Discriminante linéaire

2.2 – Résolution du problème

Le programme initial qui consiste à maximiser la quantité $\frac{u^T Bu}{u^T Wu}$ revient à maximiser la seule quantité $u^T Bu$ sous la contrainte $u^T Tu = 1$

Une alternative peut consister à fixer la quantité $u^T Wu$ et à rechercher u rendant $u^T Bu$ maximum.

2 - Analyse Factorielle Discriminante linéaire

Maximiser la quantité $u^T Bu$ sous la contrainte $u^T Tu = 1$.

⇨ la méthode du multiplicateur de Lagrange.

Soit $L(u) = u^T Bu - \lambda(u^T Tu - 1)$

Les conditions du premier ordre s'écrivent :

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{u}} = \mathbf{0} \Rightarrow \boldsymbol{Bu} - \boldsymbol{\lambda Tu} = \mathbf{0}, \text{ soit } \boxed{\boldsymbol{T^{-1}Bu} = \boldsymbol{\lambda u}}$$

u = vecteur propre de $T^{-1}B$ associé à la valeur propre λ.

2 - Analyse Factorielle Discriminante linéaire

Précisons la valeur de λ

L'égalité précédente s'écrit $Bu = \lambda Tu$ soit encore $u^T Bu = \lambda u^T Tu$ en multipliant à gauche par u^T .

Le vecteur u étant normé pour la métrique T , il apparaît que $u^T Bu = \lambda$ et donc que la quantité $u^T Bu$ est maximale pour λ plus grande valeur propre de $T^{-1}B$.

On doit donc **diagonaliser la matrice $T^{-1}B$ et prendre la valeur propre maximale λ .**

Remarque : $\lambda \in [0;1]$

2 - Analyse Factorielle Discriminante linéaire

L'Analyse Discriminante - une Analyse Factorielle particulière.

En effet, elle conduit à chercher le vecteur u rendant maximale la quantité $u^T Bu$ sous la contrainte $u^T Tu = 1$.

Posons $v = Tu$ soit $u = T^{-1}v$.
Le programme précédent s'écrit alors :
$Max_{\{v\}} v^T T^{-1} BT^{-1} v$ sous la contrainte $v^T T^{-1} v = 1$.

On a vu que $B_{jj'} = \sum_{r=1}^{R} p_r (g_{jr} - g_j)(g_{j'r} - g_{j'})$

La matrice B peut donc s'écrire sous la forme $G^T QG$

$G_{[R,J]}$ de terme générique $(g_{jr} - g_j)$ $\qquad Q_{[R,R]} = \text{Diag}\,(p_r)$

2 - Analyse Factorielle Discriminante linéaire

Le programme initial d'analyse discriminante équivaut à la recherche du vecteur v maximisant la quantité $v^T T^{-1} B T^{-1} v$ sous la contrainte $v^T T^{-1} v = 1$

D'où, après simplification, $BT^{-1}v = \lambda v$

soit encore $G^T Q G T^{-1} v = \lambda v$.

Il s'agit donc bien d'une ACP particulière, en l'occurrence celle du triplet :

$$\left(X_{Données} = \qquad\qquad , \; M_{Métrique} = \qquad\qquad , \; P_{Poids} = \qquad\qquad \right)$$

où les unités statistiques sont ______________________________

2 - Analyse Factorielle Discriminante linéaire

Un ou des axes discriminants ?

Répliquer la démarche et rechercher une suite d'axes v_α :

⇨ unitaires pour la norme choisie,

⇨ maximisant l'inertie du nuage des barycentres des groupes, soit $v_\alpha^T T^{-1} B T^{-1} v_\alpha$,

⇨ sous contrainte d'orthogonalité, avec les autres axes ($\langle v_\alpha , v_{\alpha'} \rangle_{M=T^{-1}} = 0$).

Solution obtenue pour v_α vecteur-propre de rang associé à la valeur propre d'ordre α de la matrice BT^{-1} .

Rappel : on peut adopter comme métrique, soit la matrice T^{-1} , soit la matrice W^{-1} (métrique de « Mahalanobis »).

2 - Analyse Factorielle Discriminante linéaire

La valeur propre λ_α est la variance inter-classes du nuage projeté sur l'axe discriminant.

λ_α **mesure le pouvoir discriminant de l'axe** α **.**
Plus λ_α est proche de 1 et plus la séparation (discrimination) sur l'axe α est forte entre les groupes.

Le rang de la matrice B est au plus égal $\min(R,J)-1$.
En général $R<J$ (le nombre de classes est inférieur au nombre de variables)
=> $R-1$ **axes ou fonctions discriminantes.**

2 - Analyse Factorielle Discriminante linéaire

Dans le cas particulier de 2 groupes, il n'y a qu'une seule variable discriminante car R-1=1

L'axe discriminant est alors nécessairement la droite reliant les 2 centres de gravités g_1 et g_2 *de vecteur directeur :*

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

Ou

$$v = W^{-1}(g_1 - g_2)$$

2 - Analyse Factorielle Discriminante linéaire

M=R-1 axes factoriels sont extraits

Choix du nombre d'axes

H0 : les « m» derniers rapports de corrélation sont tous nuls

$$\begin{cases} H0: & \eta^2_{M-m} = \ldots = \eta^2_{M-1} = 0 \\ H1: & non \quad H0 \end{cases}$$

⇨ Statistique de test :

$$\Delta_m = \prod_{k=M-m}^{M-1} (1-\eta_k^2)$$

2 - Analyse Factorielle Discriminante linéaire

?

Plus cette statistique est petite et plus les m derniers axes sont ______________________________
Et plus on __________________________ H0

Si X suit une loi multinormale dans chaque sous-groupe, on peut utiliser les transformations de Bartlett (loi du Khi2) ou de Rao (loi de Fisher).

Transformation de Bartlett :

Sous H0, la statistique $\mathrm{B} = -\left(n - 1 - \frac{J+M}{2}\right)\ln(\Delta_m)$

suit approximativement une loi du Khi2 à *m(J-M+m+1)* degrés de liberté.

2 - Analyse Factorielle Discriminante linéaire

Transformation de Rao :

Sous H0, la statistique $F = \frac{1-\Delta_m^{1/t}}{\Delta_m^{1/t}}$ x $\frac{at-2b}{m(J-M+m+1)}$

où

$a = n - 1 - \frac{J+M}{2}$,

b= $\frac{m(J-M+m+1)-2}{4}$,

$t = \frac{m^2(J-M+m+1)^2-4}{(J-M+m+1)^2+m^2-5}$ si $(J - M + m + 1)^2 + m^2 - 5 > 0$

$t = 1$ sinon ,

suit approximativement une loi de Fisher-Snedecor à *m(J-M+m+1)* et (at-2b) degrés de liberté.

ENSAI

2 - Analyse Factorielle Discriminante linéaire

Pouvoir discriminant global de M-1 régresseurs :

La statistique précédente Δ_m est un cas particulier du Lambda de Wilks :

$$\Lambda_{M-1} = \prod_{k=1}^{M-1} (1 - \eta_k^2)$$

utilisé en MANOVA.

Plus les M-1 régresseurs sont globalement discriminants, plus le Lambda de Wilks est petit.

2 - Analyse Factorielle Discriminante linéaire

2.3 – Analyse factorielle discriminante – optique prédictive

Critère souvent utilisé pour apprécier la qualité de la fonction discriminante obtenue : **le pourcentage de « bien classés »**, i.e. le nombre d'individus que la fonction affecte à leur groupe d'origine.

⇨ calculer la distance entre chaque individu i et les R groupes représentés par leur barycentre g_r , *r= 1 à R*
adopter la règle de décision suivante :
tout individu sera affecté au groupe dont il est le plus proche.

2 - Analyse Factorielle Discriminante linéaire

Calculons la distance, au sens de la métrique W^{-1},
entre l'individu i et le barycentre g_r :

$$\mathrm{d}^2_{\mathrm{W}^{-1}}(x_i, g_r) = (x_i - g_r)^T W^{-1}(x_i - g_r)$$

En développant cette expression nous obtenons :

$$\mathrm{d}^2_{\mathrm{W}^{-1}}(x_i, g_r) = \|x_i\|^2_{W^{-1}} + \|g_r\|^2_{W^{-1}} - 2\ x_i^T W^{-1} g_r$$

Rechercher la distance minimale en fonction de g_r revient
à chercher le groupe pour lequel la quantité

$\boxed{\|g_r\|^2_{W^{-1}} - 2x_i^T W^{-1} g_r \text{ est minimale}}$

puisque $\|x_i\|^2_{W^{-1}}$ est constant quel que soit r.

2 - Analyse Factorielle Discriminante linéaire

Il est équivalent de chercher

$$r = \underset{r \in \{1,\dots,R\}}{\arg\min} \left(\|g_r\|^2_{W^{-1}} - 2 {x_i}^T W^{-1} g_r \right)$$

ou

$$r = \underset{r \in \{1,\dots,R\}}{\arg\max} \left({x_i}^T W^{-1} g_r - \frac{1}{2} \|g_r\|^2_{W^{-1}} \right) = \underset{r \in \{1,\dots,R\}}{\arg\max} \left({x_i}^T W^{-1} g_r - \frac{1}{2} \left({g_r}^T W^{-1} g_r \right) \right)$$

.

La quantité $score(x_i ; r) = {x_i}^T W^{-1} g_r - \frac{1}{2} \left({g_r}^T W^{-1} g_r \right)$ est appelée **score** de l'observation i dans le groupe r .

2 - Analyse Factorielle Discriminante linéaire

Cas particulier de 2 groupes

On compare les scores dans chacun des deux groupes.

La règle de décision :

si $Score_1(x_i) - Score_2(x_i) > 0$ alors l'observation i est affectée au groupe ____

La différence des scores peut encore s'écrire

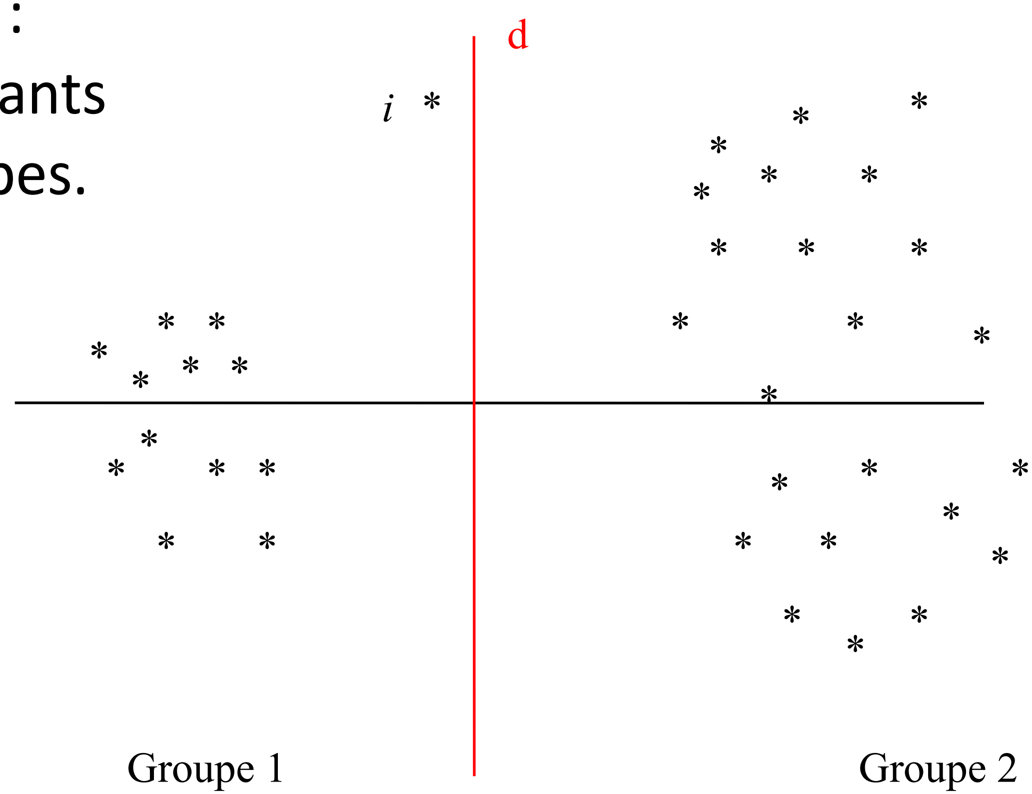
$$x_i^T W^{-1} g_1 - \frac{1}{2}\left(g_1^T W^{-1} g_1\right) - x_i^T W^{-1} g_2 + \frac{1}{2}\left(g_2^T W^{-1} g_2\right)$$

$$= x_i^T W^{-1}(g_1 - g_2) - \frac{1}{2}(g_1 + g_2)^T W^{-1}(g_1 - g_2) = Fish(x)$$

On obtient ainsi, la **fonction linéaire discriminante de Fisher *Fish*(x)**

3 - Analyse Factorielle Discriminante Quadratique

La droite d représente la frontière entre les deux groupes G_1 et G_2 :
le lieu des points équidistants de chacun des deux groupes.



3 - Analyse Factorielle Discriminante Quadratique

L'individu i, situé à gauche de la droite d, sera affecté au groupe 1 par la **règle d'affectation AFD linéaire** or les deux groupes sont très différents.

Le groupe 1 est homogène et l'affectation de i à ce groupe va perturber cette homogénéité. Il serait peut-être préférable d'affecter l'élément i au second groupe, plus hétérogène.

Il est possible d'utiliser une distance spécifique à chaque groupe - **une métrique locale :**
la matrice de variance covariance intra-groupe.

2.2 - Analyse Factorielle Discriminante Quadratique

Le calcul de la distance entre l'élément i et g_r, s'effectue avec la métrique W_r^{-1} au lieu de la métrique indifférenciée W^{-1}.

Cette technique est connue sous le nom d'analyse discriminante quadratique, car le terme en x n'est plus constant sur l'ensemble des groupes.

La règle de décision devient :

$$r = \underset{r \in \{1,\dots,R\}}{\arg\max} \left(x_i^T W_r^{-1} g_r - \frac{1}{2} \|g_r\|^2_{W_r^{-1}} - \frac{1}{2} \|x_i\|^2_{W_r^{-1}} \right).$$

3 - Analyse Factorielle Discriminante Quadratique

Utiliser l'analyse discriminante quadratique si :

- la discrimination linéaire donne des résultats décevants,
- l'hypothèse d'égalité des matrices de variance covariance internes est rejetée.

4 - Analyse Factorielle Discriminante sur régresseurs qualitatifs

Si les variables explicatives sont qualitatives, on ne peut pas appliquer directement l'analyse factorielle discriminante (AFD).

Solution en deux étapes :

1. réaliser au préalable une Analyse des Correspondances Multiples
2. utiliser ensuite l'AFD sur les facteurs issus de l'ACM

Supposons Q variables qualitatives totalisant J modalités.
Une ACM réalisée sur ces modalités permet d'obtenir $J-Q$ axes factoriels non-triviaux.

4 - Analyse Factorielle Discriminante sur régresseurs qualitatifs

Soit F_β la composante factorielle de rang β de l'ACM, c'est-à-dire le vecteur contenant les coordonnées factorielles des observations sur l'axe de rang β :

$$F_{\beta(\beta=1,J-Q)} = \sum_{j=1}^{J} a_{j\beta} Z_j$$

Avec $a_{j\beta}$ la coordonnée de rang j du vecteur propre d'ordre β et Z_j l'indicatrice de la modalité j.

Les F_β étant des variables continues, on peut leur appliquer une AFD.

5 - Analyse Factorielle Discriminante sur régresseurs qualitatifs

Soient $C_{l(l=1,\dots,R-1)}$ les composantes obtenues avec l'AFD alors

$$C_l = \sum_{\beta=1}^{J-Q} u_{\beta\, l}\ F_\beta$$

où $v_{\beta l}$ désigne la composante d'ordre β du vecteur propre de rang l

$$C_l = \sum_{\beta=1}^{J-Q} u_{\beta\, l} \sum_{j=1}^{J} a_{j\beta} Z_j = \sum_{j=1}^{J} \sum_{\beta=1}^{J-Q} u_{\beta\, l} a_{j\beta} Z_j = \sum_{j=1}^{J} w_{jl} Z_j$$

en posant $w_{jl} = \sum_{\beta=1}^{J-Q} u_{\beta\, l} a_{j\beta}$.

5 - Sélection de variables en Analyse Factorielle Discriminante

Critères de sélection de variables les plus utilisés :

- ⇨ La trace de la matrice $T_q^{-1}B_q$ calculée sur les q variables sélectionnées $(q < J)$.
- ⇨ Le « lambda de Wilks » égal à $\dfrac{\det W_q}{\det T_q}$.

Ces critères sont adaptés à l'objectif « descriptif » ou « explicatif » de l'AFD et non à l'objectif prédictif.

Utilisation de procédures de sélection pas à pas

Méthodes contestées pour leur manque de robustesse

6 - Exemple avec R - Spotify

```
#### sans scale
lda(like ~ acousticness+ danceability +duration + energy + instrumentalness+  liveness+ loudness+ speechiness +  tempo+ valence , data = spotify)
```

```
Prior probabilities of groups:
 0          1
0.49      0.51
```

6 - Exemple avec R - Spotify

```
Group means:
   acousticness danceability duration  energy
0   0.22          0.59          234140.5  0.67
1  0.15          0.65          258197.6  0.69
   instrumentalness liveness
0  0.09              0.19
1  0.17              0.19
    loudness  speechiness   tempo   valence
0  -6.81     0.08           120.67  0.47
1  -7.35     0.11           122.52  0.52
```

6 - Exemple avec R - Spotify

```
Coefficients of linear discriminants: LD1
          acousticness      -1.938689e+00
          danceability       2.488930e+00
          duration           3.453592e-06
          energy             5.522830e-01
          instrumentalness   1.600306e+00
          liveness           5.920688e-01
          loudness          -1.379224e-01
          speechiness        4.879219e+00
          tempo              5.003491e-03
          valence            9.812938e-01
```

6 - Exemple avec R - Spotify

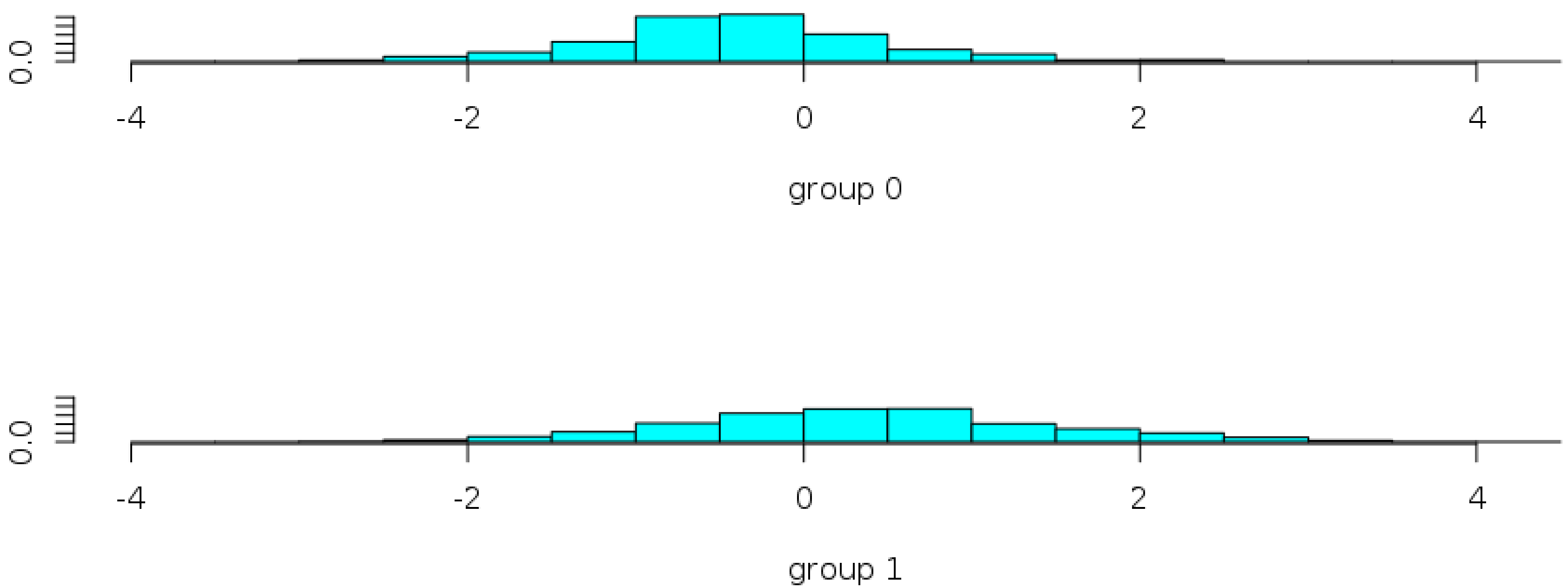
```
prevlda1=lda(like ~ acousticness+ danceability +
duration +energy + instrumentalness +    liveness+
 loudness+ speechiness + tempo+ valence,
data = spotify,CV=TRUE)$class

table(prevlda1,spotify$like )
```

Prevlda1	0	1
0	692	366
1	305	654

6 - Exemple avec R - Spotify

```
#Plot the predictions - first linear discriminant
spot.lda.values <- predict(lda1)l
dahist(spot.lda.values$x[,1], g = spotify$like)
```



6 - Exemple avec R - Spotify

Résultats avec centrage et réduction des régresseurs

```
Coefficients of linear discriminants:LD1
scale(acousticness)      -0.50403834
scale(danceability)       0.40078977
scale(duration)           0.28313176
scale(energy)             0.11613020
scale(instrumentalness)   0.43714300
scale(liveness)           0.09203896
scale(loudness)          -0.51882067
scale(speechiness)        0.43879528
scale(tempo)              0.13352119
scale(valence)            0.24257138
```

6 - Exemple avec R - Spotify

?

Peut-on utiliser les coefficients de la fonction discriminante pour dire qu'un régresseur est plus discriminant qu'un autre ?

6 - Exemple avec R - Spotify

?

Doit-on centrer réduire les régresseurs ?

Quel mode de sélection de variables proposez-vous ?

7 – Analyse discriminante « bayésienne »

7.1 – L'approche probabiliste

On note $P(G_r / X = x_i)$ la probabilité que l'observation i soit issue du groupe r compte tenu des valeurs prises par l'ensemble des variables présentes X.

La démarche = Affecter l'observation i au groupe le plus probable :

⇨ Groupe pour lequel la probabilité *a posteriori* $P(G_r / X = x_i)$ est maximale.

Problème : cette probabilité n'est pas connue.

7 – Analyse discriminante « bayésienne »

D'après le théorème de BAYES :

$$P(G_r / X = x_i) = \frac{P(G_r)P(X = x_i / G_r)}{\sum_{s=1}^{R} P(G_s)P(X = x_i / G_s)}.$$

Estimation :

Les $P(G_r)$, **probabilités *a priori***, seront estimées par Π_r

Les $P(X = x_i / G_r)$ seront estimées par $\hat{f}_r(x_i)$ où $\hat{f}_r$ désigne la densité de X estimée dans le groupe r.

7 – Analyse discriminante « bayésienne »

L'égalité précédente devient avec estimateur plug-in :

$$\hat{P}(G_r / x) = \frac{\Pi_r \hat{f}_r(x)}{\sum_{s=1}^{R} \Pi_s \hat{f}_s(x)}.$$

Comme le dénominateur est identique pour tous les groupes, il suffit de comparer les quantités $\Pi_r \hat{f}_r(x)$.

Le problème est d'**estimer les probabilités *a priori* et les fonctions de densité.**

7 – Analyse discriminante « bayésienne »

7.2 – Résolution du problème

A - <u>Estimation des probabilités *a priori*</u>

Les Π_r doivent refléter la probabilité d'appartenir à chacun des groupes dans la population et non dans l'échantillon :

$\Pi_r = \dfrac{N_r}{N}$ où N_r désigne le nombre d'observations appartenant au groupe dans la population et N l'effectif global.

Prendre les $\Pi_r = \dfrac{n_r}{n}$, proportionnels aux effectifs dans l'échantillon, conduit à favoriser le classement dans le groupe numériquement le plus nombreux.

7 – Analyse discriminante « bayésienne »

Si on ne souhaite pas favoriser le classement dans le groupe ayant l'effectif le plus élevé, on peut retenir des probabilités a priori estimées égales : $\Pi_r = \dfrac{1}{R}$ dans le cas de R groupes

B - <u>Estimation des densités</u>

- ⇨ Approche paramétrique : supposer que la densité suit une loi connue, par exemple une loi normale.
- ⇨ Approche non paramétrique : par exemple utiliser la méthode des noyaux ou celle des plus proches voisins.

7 – Analyse discriminante « bayésienne »

B.1 - Approche paramétrique : fonctions de densité normales

Très souvent utilisées en pratique, les fonctions de densité suivant des lois normales multidimensionnelles :

$$f_r(x) = (2\pi)^{-J/2} \left| \det(\Sigma_r) \right|^{-1/2} \exp\left[-\frac{1}{2}(x-\mu_r)^T \Sigma_r^{-1} (x-\mu_r) \right]$$

où Σ_r désigne la matrice de variance covariance (totale ou intra) théorique du groupe r et μ_r l'espérance dans le groupe r.

Les μ_r seront estimées par g_r, moyennes observées sur chacun des groupes, et les Σ_r par les matrices de variance covariance empiriques totales T ou intra-classe W_r.

7 – Analyse discriminante « bayésienne »

La démarche revient à chercher le groupe r pour lequel $\Pi_r \hat{f}_r(x)$ ou $\log\left(\Pi_r \hat{f}_r(x)\right)$ est maximal, soit dans le cas normal les quantités :

$$\log(\Pi_r) - \frac{1}{2}\log\left(\left|\det\left(\hat{\Sigma}_r\right)\right|\right) - \frac{1}{2}(x-g_r)^T \hat{\Sigma}_r^{-1}(x-g_r)$$

Si on suppose l'identité des matrices de variance-covariance pour chacun des k groupes il suffit de comparer les quantités :

?

____

____

7 – Analyse discriminante « bayésienne »

Cas de deux groupes :

D'après l'expression qui précède, il suffit de considérer la quantité :

$$\log\left(\frac{\Pi_1}{\Pi_2}\right)+x^T\hat{\Sigma}^{-1}(g_1-g_2)-\frac{1}{2}(g_1+g_2)^T\hat{\Sigma}^{-1}(g_1-g_2)$$

Si cette quantité est positive l'individu est affecté au groupe 1 et au groupe 2 si elle est négative.

Si les probabilités *a priori* estimées sont égales, alors l'expression précédente s'écrit $x^T\hat{\Sigma}^{-1}(g_1-g_2)-\frac{1}{2}(g_1+g_2)^T\hat{\Sigma}^{-1}(g_1-g_2)$

et l'on retrouve la fonction discriminante AFD de FISHER.

7 – Analyse discriminante « bayésienne »

Equivalence entre approche géométrique et probabiliste si :

- Deux groupes
- Hypothèse de probabilités *a priori* égales
- Identité des matrices de variance covariance à l'intérieur de chacun des groupes
- Densité normales multidimensionnelles

Remarque sur le rôle des probabilités *a priori* :

Dans le cas de 2 groupes, avec $n_1>n_2$ et des probabilités *a priori* proportionnelles aux effectifs, la règle de classement basée sur $\log\left(\frac{\Pi_1}{\Pi_2}\right)+Fish(x)$ favorise le classement dans le groupe 1 car ajout d'un terme positif à $Fish(x)$.

7 – Analyse discriminante « bayésienne »

Obtention des probabilités *a posteriori*

Pour obtenir les probabilités de classement *a posteriori* il suffit d'utiliser le fait que $\hat{P}(G_1 / X = x) + \hat{P}(G_2 / X = x) = 1$, alors

$$\hat{P}(G_2 / X = x) = \frac{\Pi_2 \hat{f}_2(x)}{\Pi_1 \hat{f}_1(x) + \Pi_2 \hat{f}_2(x)} = \frac{1}{1 + \exp(Fish(x))}$$

Le calcul de $\hat{P}(G_1 / x)$ est immédiat :

$$\hat{P}(G_1 / X = x) = 1 - \hat{P}(G_2 / X = x) = 1 - \frac{1}{1 + \exp(Fish(x))} = \frac{\exp(Fish(x))}{1 + \exp(Fish(x))}$$

Avantage de l'approche probabiliste sur la méthode géométrique : fournir outre le groupe d'affectation, la probabilité avec laquelle cette affectation se réalise.

7 – Analyse discriminante « bayésienne »

B.2 - Méthode des noyaux de PARZEN :

$$\hat{f}_r(x) = \frac{1}{n_r h^J} \sum_{x_i \in G_r} \mathrm{K}\left(\frac{x - x_i}{h}\right)$$

où *h* est appelé largeur de fenêtre ou paramètre de lissage.

La fonction ***K****(z)* appelée noyau doit vérifier les relations :

$\boldsymbol{K}(z) \geq 0$ et $\int \boldsymbol{K}(z)dz = 1$.

On peut prendre pour ***K*** une fonction de densité usuelle, par exemple la loi normale multidimensionnelle, voire la loi continue uniforme.
Le problème fondamental de cette démarche est de déterminer un paramètre de lissage *h* adapté.

7 – Analyse discriminante « bayésienne »

7.3 – Coûts d’erreur

Jusqu’ici les erreurs de classement ont été implicitement considérées comme ayant le même coût.

Or, il est des cas où l’**on peut considérer que certaines erreurs de classement sont plus graves que d’autres.**

<u>Exemple d’un cas avec deux groupes :</u>

Le premier groupe comprend des individus porteurs d’une pathologie grave et le second des individus qui en sont exempts.

Classer des patients du groupe 1 dans le groupe 2 =
ne pas détecter la pathologie est plus grave que l’inverse,
classer des observations du groupe2 dans le groupe 1 =
détecter la pathologie à tort.

7 – Analyse discriminante « bayésienne »

La démarche :

Construire une fonction de coût et affecter chaque élément au groupe pour lequel le coût moyen *a posteriori* est minimum.

Soit c(s|r) le coût de classement incorrect d'une observation appartenant au groupe r dans le groupe s :

- c(s|s)=0
- C(s|r) pour s, $r \in [1, R]$ avec s $\neq r$.

Coût moyen *a posteriori* d'affectation d'une observation au groupe s :

$$CM_s(x) = \sum_{r=1}^{R} c(s/r)\hat{P}(G_r / X = x) \quad = \sum_{r=1}^{R} c(s/r) \frac{\Pi_r \hat{f}_r(x)}{\sum_{l=1}^{R} \Pi_l \hat{f}_l(x)}.$$

7 – Analyse discriminante « bayésienne »

Si nous supposons tous les coûts égaux à une constante c(s/r)=c si $s \neq r$.

Dans ce cas

$$CM_s(x) = \sum_{r=1}^{R} c(s/r)\hat{P}(G_r / X = x) = c \, . \left[\sum_{r=1}^{R} \hat{P}(G_r / X = x) - \hat{P}(G_s / X = x) \right]$$

Retrouvons-nous un résultat vu précédemment ?

?