Analyse Factorielle Discriminante

Ghazi Bel Mufti

belmufti@yahoo.com

ESSAI-3 / DATA MINING

Plan

Introduction

Soit y une variable qualitative à q modalités, que l'on souhaite expliquer par $x^1, \ldots, x^p$: p variables quantitatives ; $x^1, \ldots, x^p$ sont les variables **explicatives** et y la variable **à expliquer**. Ces variables sont observées sur un ensemble d'individus. L'ensemble des individus possédant une même modalités de y est appelé ”classe”.

L'A.D. comporte deux étapes :

- on cherche d'abord à séparer de façon optimale les classes dans l'espace explicatif $\mathbb{R}^p$ (on parle alors d'**Analyse Factorielle Discriminante**).
- Puis, on prévoit y connaissant les x^j, en d'autres termes on affecte tout individu, pour lequel on connaît les x^j, à l'une des classes (**Analyse Discriminante Décisionnelle**)

1. Notations

- Soit I un échantillon de n individus pour lesquels les $p+1$ variables $x^1, \ldots, x^p$ et y ont été observées.
- On note
$$X = (x^1, \ldots, x^p) = (x_1, \ldots, x_n)'$$
où x_i $(1 \leq i \leq n)$ est le vecteur des valeurs prises par l'individu i pour les p variables explicatives, et où x^j $(1 \leq j \leq p)$ est le vecteur des valeurs prises par la variable x^j pour les n individus.

Nuage des individus

- p_i : masse de l'individu i avec $\sum_i p_i = 1$
- D_p : Diag(p_i)
- $\mathcal{M}_X$ = nuage des x_i
- g = centre de gravité de $\mathcal{M}_X = \sum_{i \in I} p_i x_i$
- On suppose $g = 0$, i.e. $\overline{x}^j = \sum_{i \in I} p_i x_i^j = 0$, on a alors :

$$V = X' D_p X$$

.

Nuages associés aux q classes

- On note $y_1, \dots y_k, \dots y_q$ les modalités de y.
- Pour tout k ($1 \leq k \leq q$), l'ensemble I_k des individus ayant adopté la modalité y_k est appelé "classe k".
- De plus on note :

$$n_k = |I_k| \quad ; \quad m_k = \sum_{i \in I_k} p_i \quad ; \quad g_k = \frac{1}{m_k} \sum_{i \in I_k} p_i x_i$$

  i.e. g_k est le c.d.g. du nuage des individus de I_k.
- On note V_k la matrice variance associée au nuage des individus à I_k.

Nuages des c.d.g. des classes

- On note $\mathcal{M}_G = \{(g_k, m_k) \mid 1 \leq k \leq q\}$ où G désigne le tableau des coordonnées des g_k :

$$G = (g_1, \dots, g_k, \dots, g_q)'$$

Le centre de gravité de $\mathcal{M}_G$ est égal à :

$$\sum_{k=1}^{q} m_k g_k = \sum_{k=1}^{q} m_k \Big(\frac{1}{m_k} \sum_{i \in I_k} p_i x_i\Big) = \sum_{i \in I} p_i x_i = g = 0$$

Donc G est un tableau centré.

- La matrice variance B de G s'écrit

$$B = G' D_m G$$

où $D_m = \text{Diag}(m_k)$. On dit que B est la matrice variance interclasses.

- La matrice variance intraclasses, notée W, est définie par :

$$W = \sum_{k=1}^{q} m_k V_k$$

Relation fondamentale

$$V = B + W$$

On dit que V est la matrice variance totale.

Preuve :

$$V = (cov(x^j, x^{j'}))_{j,j'}$$

$$\begin{aligned} cov(x^j, x^{j'}) &= \sum_i p_i(x_i^j - \overline{x^j})(x_i^{j'} - \overline{x^{j'}}) \\ &= \sum_{k=1}^{q} \sum_{i \in I_k} p_i(x_i^j - \overline{x^j})(x_i^{j'} - \overline{x^{j'}}) \end{aligned}$$

On a : $x_i^j - \overline{x^j} = (x_i^j - g_k^j) + (g_k^j - \overline{x^j})$ pout $i \in I_k$
Or :

$$\sum_{i \in I_k} p_i(x_i^j - g_k^j)(g_k^{j'} - \overline{x^{j'}}) = (g_k^{j'} - \overline{x^{j'}}) \sum_{i \in I_k} p_i(x_i^j - g_k^j) = 0$$

$$\sum_{i \in I_k} p_i(g_k^j - \overline{x^j})(x_i^{j'} - g_k^{j'}) = (g_k^j - \overline{x^j}) \sum_{i \in I_k} p_i(x_i^{j'} - g_k^{j'}) = 0$$

Donc

$$cov(x^j, x^{j'}) = \sum_{k=1}^{q} \sum_{i \in I_k} p_i (x_i^j - g_k^j)(x_i^{j'} - g_k^{j'}) + \sum_{k=1}^{q} \sum_{i \in I_k} p_i (g_k^j - \overline{x^j})(g_k^{j'} - \overline{x^{j'}})$$

$$cov(x^j, x^{j'}) = E_1 + E_2$$

avec

$$E_1 = \sum_{k=1}^{q} m_k \left(\frac{1}{m_k} \sum_{i \in I_k} p_i (x_i^j - g_k^j)(x_i^{j'} - g_k^{j'})\right) = \sum_{k=1}^{q} m_k V_k(j, j') = W_{jj'}$$

$$E_2 = \sum_{k=1}^{q} m_k (g_k^j - \overline{x^j})(g_k^{j'} - \overline{x^{j'}}) = \sum_{k=1}^{q} m_k (g_k^j - g^j)(g_k^{j'} - g^{j'}) = B_{jj'}$$

D'où le résultat.

Description du problème

On cherche $z = b_1 x^1 + b_2 x^2 + \ldots + b_p x^p$ telle que :

c1) Les classes sont séparées de façon optimale ; i.e. : la variance de z entre les classes (interclasses) est maximale.

c2) La variance de z est minimale à l'intérieure des classes (variance intraclasses).

La variable z ainsi définie est appelée **score**. On a :

$$\begin{aligned} var(z) &= cov(z,z) \\ &= cov(\sum_j b_j x^j, \sum_{j'} b_{j'} x^{j'}) \\ &= \sum_j \sum_{j'} b_j b_{j'} cov(x^j, x^{j'}) \end{aligned}$$

$cov(.,.)$ étant une forme bilinéaire.

Donc $var(z) = b' Vb$ où $b' = (b_1, \ldots, b_p)$. Or $V = B + W$ donc

$$var(z) = b' Bb + b' Wb$$

$$
\begin{aligned}
b'Bb &= \sum_{j,j'} b_j b_{j'} B_{jj'} \\
&= \sum_{j,j'} b_j b_{j'} \sum_k m_k (g_k^j - g^j)(g_k^{j'} - g^{j'}) \\
&= \sum_k m_k \sum_j b_j (g_k^j - g^j) \sum_{j'} b_{j'} (g_k^{j'} - g^{j'}) \\
&= \sum_k m_k (\sum_j b_j g_k^j - \sum_j b_j g^j)(\sum_{j'} b_{j'} g_k^{j'} - \sum_{j'} b_{j'} g^{j'}) \\
&= \sum_k m_k (\sum_j b_j g_k^j - \sum_j b_j g^j)^2
\end{aligned}
$$

où $\sum_j b_j g_k^j$ est la moyenne de z dans la classe k et $\sum_j b_j g^j$ n'est autre que $\bar{z}$. Donc $b'Bb$ est égale à la variance interclasses de z.

$$
\begin{aligned}
b'Wb &= b'(\sum_k m_k V_k)b \\
&= \sum_k m_k (b' V_k b) \\
&= \sum_k m_k \sum_{j,j'} b_j b_{j'} V_k(j,j') \\
&= \sum_k m_k \sum_{j,j'} b_j b_{j'} (\frac{1}{m_k} \sum_{i \in I_k} p_i (x_i^j - g_k^j)(x_i^{j'} - g_k^{j'})) \\
&= \sum_k m_k (\frac{1}{m_k} \sum_{i \in I_k} p_i (\sum_j b_j x_i^j - \sum_j b_j g_k^j)^2)) \\
&= \sum_k m_k (\frac{1}{m_k} \sum_{i \in I_k} p_i (z_i - \overline{z^k})^2)
\end{aligned}
$$

où $\overline{z^k}$ désigne la moyenne de z dans la calsse k. Donc $b'Wb$ est la variance intraclasses de z.

Pour satisfaire c1) et c2), on cherche à maximiser

$$\eta^2 = \frac{b'Bb}{b'Vb}$$

Solution

Comme η^2 est inchangé si b est remplacé par γb, γ désignant un scalaire quelconque, on supposera que $var(z) = b'Vb = 1$, et par conséquent

$$var(z) = 1 = b'Bb + b'Wb$$

Donc maximiser η^2 revient à maximiser $b'Bb$ sous la contrainte $b'Vb = 1$.

Posons $\mathcal{L} = b'Bb - \lambda(b'Vb - 1)$. En annulant les dérivées de $\mathcal{L}$ par rapport à b et λ, on obtient

$$\left\{ \begin{array}{lll} 2Bb - 2\lambda Vb & = 0 & (1) \\ b'Vb & = 1 & (2) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{ll} Bb & = \lambda Vb \\ b'Vb & = 1 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{ll} V^{-1}Bb & = \lambda b \\ b'Vb & = 1 \end{array} \right.$$

- D'après (1), on $b'Bb = \lambda b'Vb$, donc $\lambda = \dfrac{b'Bb}{b'Vb}$, ce qui prouve que $\lambda \in [0,1]$ et que la solution du problème est obtenue en prenant la plus grande valeur propre de $V^{-1}B$, notée λ_1, par la suite. On notera $z_{(1)}$ le score solution, $b_{(1)}$ le vecteur propre (normé pour V) associé à λ_1.
- On cherchera ensuite un second score $z_{(2)} = \sum_j b_{(2)j} x^j$ centré, réduit, non corrélé à $z_{(1)}$, et tel que $\dfrac{b'_{(2)} B b_{(2)}}{b'_{(2)} V b_{(2)}}$ soit maximum. On montre de la même manière que $b_{(2)}$ est vecteur propre de $V^{-1}B$, normé pour V, associé à la deuxième plus grande valeur propre λ_2 de $V^{-1}B$.

Le $\alpha^{\text{ème}}$ score centré réduit, et non corrélé aux précédents, noté z_α se déduit de façon analogue, à partir de la $\alpha^{\text{ème}}$ plus grande valeur propre de $V^{-1}B$, par la formule :

$$z_{(\alpha)} = \sum_j b_{(\alpha)j} x^j$$

Notons r le rang de $V^{-1}B$. On a $rg(V^{-1}B) = rg(B) = rg(G'D_mG) = rg(G) = rg(g^1, \ldots, g^q)$. Or $\sum_{k=1}^{q} m_k g^k = g = 0$, donc $r \leq min(p, q-1)$.

Les vecteurs $b_{(1)}, \ldots, b_{(r)}$ sont appelés **facteurs discriminants**, ou **formes linéaires discriminantes**. Les valeurs propres $\lambda_1, \ldots, \lambda_r$ sont appelées **pouvoirs dscriminants**. On a :

$$V^{-1}Bb_{(\alpha)} = \lambda_\alpha b_{(\alpha)}$$

avec $b'_{(\alpha)}Bb_{(\beta)} = \lambda_\alpha \delta_\alpha^\beta$ et $b'_{(\alpha)}Vb_{(\beta)} = \delta_\alpha^\beta$.

Axes factoriels discriminants I

- Posons :

$$u^\alpha = Vb_{(\alpha)}$$

Donc

$$BV^{-1}u^\alpha = Bb_{(\alpha)} = VV^{-1}Bb_{(\alpha)} = \lambda_\alpha Vb_{(\alpha)} = \lambda_\alpha u^\alpha \quad (1)$$

De plus

$$(u^\alpha)'V^{-1}u^\beta = (u^\alpha)'V^{-1}VV^{-1}u^\beta = b'_{(\alpha)}Vb_{(\beta)} = \delta_\alpha^\beta \quad (2)$$

- Les relations (1) et (2) montrent que les u^α sont les vecteurs axiaux factoriels de l'analyse factorielle du nuage $\mathcal{M}_G$ (associé à G), muni des masses m_k, et de la métrique $M = V^{-1}$. Cette analyse factorielle est dite "discriminante" et les axes Δu^α sont appelés **axes factoriels discriminants**.

Axes factoriels discriminants II

- Considérons la projection de l'individu x_i sur l'axe Δu^α :

$$(x_i)' V^{-1} u^\alpha = (x_i)' b_{(\alpha)} = \sum_j x_i^j b_{(\alpha)j} \quad (3)$$

qui s'interprète comme le $\alpha^{\text{ème}}$ score de l'individu i. Donc le $\alpha^{\text{ème}}$ axe factoriel discriminant est associé au $\alpha^{\text{ème}}$ score. On déduit de (3) que :

$$z_\alpha = XV^{-1} u^\alpha = Xb_{(\alpha)} \quad (4)$$

Autrement dit, le vecteur des cores z_α s'obtient en plaçant les lignes du tableau X en éléments supplémentaires dans l'A.F. du tableau G avec la métrique V^{-1}.

Axes factoriels discriminants III

- Les composantes principales ψ_α sont définies par :

$$\psi_\alpha = GV^{-1}u^\alpha = Gb_{(\alpha)} \quad (5)$$

On vérifie que l'on a :

$$\begin{cases} GV^{-1}G'D_m\psi_\alpha = & \lambda_\alpha\psi_\alpha \\ \psi'_\alpha D_m\psi_\beta = & b'_{(\alpha)}Bb_{(\beta)} = \lambda_\alpha\delta_\alpha^\beta \end{cases}$$

Remarque : On a

$$z'_\alpha D_p z_\beta = (Xb_\alpha)'D_pXb_\beta = b'_\alpha X'D_pXb_\beta = b'_\alpha Vb_\beta = \delta_\alpha^\beta.$$

On retrouve ainsi le fait que les scores sont réduits et non corrélés.

Effet du choix de la métrique W^{-1} I

Nous examinons comment sont changés les résultats de l'analyse factorielle discriminante si on utilise la métrique W^{-1} au lieu de V^{-1}.

Désignons respectivement par w^{α}, c_{α}, γ_{α} et μ_{α} les vecteurs axiaux factoriels, les facteurs, les composantes principales et les valeurs propres lorsqu'on utilise la métrique W^{-1}.

Remarquons tout d'abord que l'on a :

$$\begin{aligned} ||b_{(\alpha)}||^2_W &= b'_{(\alpha)} W b_{(\alpha)} = b'_{(\alpha)}(V-B)b_{(\alpha)} = 1-\lambda_{\alpha} \\ Bb_{(\alpha)} &= \lambda_{\alpha} V b_{(\alpha)} = \lambda_{\alpha}(B+W)b_{(\alpha)} \\ W^{-1}Bb_{(\alpha)} &= \lambda_{\alpha} W^{-1} B b_{(\alpha)} + \lambda_{\alpha} b_{(\alpha)} \end{aligned}$$

Effet du choix de la métrique W^{-1} II

d'où

$$W^{-1}Bb_{(\alpha)} = \frac{\lambda_\alpha}{1-\lambda_\alpha} b_{(\alpha)}$$

Par conséquent

$$\begin{aligned} c_{(\alpha)} &= \frac{b_{(\alpha)}}{||b_{(\alpha)}||_W} = \frac{b_{(\alpha)}}{\sqrt{1-\lambda_\alpha}} \\ \mu_\alpha &= \frac{\lambda_\alpha}{1-\lambda_\alpha} \end{aligned}$$

On en déduit la composante principale γ_α :

$$\gamma_\alpha = Gc_{(\alpha)} = \frac{Gb_{(\alpha)}}{\sqrt{1-\lambda_\alpha}} = \frac{\psi_\alpha}{\sqrt{1-\lambda_\alpha}}$$

Effet du choix de la métrique W^{-1} III

puis le score :

$$t_\alpha = Xc_{(\alpha)} = \frac{Xb_{(\alpha)}}{\sqrt{1-\lambda_\alpha}} = \frac{z_\alpha}{\sqrt{1-\lambda_\alpha}}.$$

Enfin, le vecteur axial factoriel w^α s'exprime par :

$$w^\alpha = \frac{GD_m\gamma_\alpha}{\mu_\alpha} = \frac{GD_m\gamma_\alpha(1-\lambda_\alpha)}{\lambda_\alpha} = \frac{GD_m\psi_\alpha\sqrt{1-\lambda_\alpha}}{\lambda_\alpha} = \sqrt{1-\lambda_\alpha}\,u^\alpha$$

Analyse discriminante décisionnelle

- Il s'agit ici d'affecter un individu à une classe k connaissant les valeurs prises par les variables explicatives $x^1, \ldots, x^p$.
- Il existe plusieurs critères d'affectation :
  1. le critère géométrique linéaire qui est directement dérivé de l'AFD,
  2. le critère quadratique qui généralise le critère géométrique linéaire en associant la métrque V_k^{-1} à chaque groupe,
  3. le critère fondé sur des hypothèses gaussiennes,
  4. le critère fondé sur des hypothèses bayésiennes qui est le plus général.

Critère géométrique linéaire I

- On se place dans $\mathbb{R}^p$ qui contient $\mathcal{M}_X$ et $\mathcal{M}_G$. On suppose que $\mathbb{R}^p$ est muni de la métrique $M = V^{-1}$ ou $M = W^{-1}$ qui est équivalente à V^{-1}.
- **Critère :** un individu $x = \begin{pmatrix} x^1 \\ \vdots \\ x^p \end{pmatrix}$ est affecté au groupe k_0 s'il est plus proche de g_{k_0} que du c.d.g. de tout autre classe :

$$||x - g_{k_0}||^2_M = \min_{1 \leq k \leq q} ||x - g_k||^2_M$$

Comme $||x - g_k||^2_M = ||x||^2_M - (g_k)'M(2x - g_k)$, le critère s'écrit aussi :

$$(g_{k_0})'M(g_{k_0} - 2x) = \min_{1 \leq k \leq q} (g_k)'M(g_k - 2x)$$

Critère géométrique linéaire II

ou

$$(g_{k_0})'M(x - \tfrac{1}{2}g_{k_0}) = \max_{1 \leq k \leq q} (g_k)'M(x - \frac{1}{2}g_k)$$

- On voit bien qu'il s'agit d'un critère linéaire.
- Désignons par R_k la région dont tous les points sont plus proches de $g_k \quad (1 \leq k \leq q)$ que de tout autre c.d.g. de classe ; un individu x est donc affecté à la classe k_0 ssi $x \in R_{k_0}$.
- Le critère revient alors à découper $\mathbb{R}^p$ selon les q régions $R_1, \ldots, R_q$ dont les frontières sont les $q(q-1)/2$ hyperplans médiatiques des segments $g_k g_l$, avec $(1 \leq k < l \leq q)$ (le nombre de frontières est souvent très inférieur à $q(q-1)/2$).

Critère géométrique quadratique

- On associe à chaque classe k une métrique M_k pour tenir compte de la forme de cette classe : en général $M_k = V_k^{-1}$.
- Un individu x est affecté à la classe k_0 si :

$$||x - g^{k_0}||^2_{M_{k_0}} = \min_{1 \leq k \leq q} ||x - g^k||^2_{M_k}$$

- Ce critère est quadratique car, ici, le terme $||x||^2_{M_k}$ ne peut être éliminé du critère comme dans le cas précédent où $M_k = M = V^{-1}$.

a) Hypothèse de lois, cas de la loi normale I

- L'hypothèse de lois consiste à supposer que x appartient à la classe k ssi x suit une loi de probabilité, dont la densité sera noté f_k.
- Le critère consiste à affecter x à la classe k_0 pour laquelle la densité f_{k_0} est maximale :

$$f_{k_0}(x) = \max_{1 \leq k \leq q} f_k(x)$$

Cas particulier : si les lois des classes sont des lois gaussiennes de densité

$$f_k(x) = ((2\pi)^p \det \Sigma_k)^{-1/2} \exp\{-\frac{1}{2} ||x - \mu_k||^2_{\Sigma_k^{-1}}\}$$

a) Hypothèse de lois, cas de la loi normale II

alors on raisonnera sur les quantités $-2\ln f_k(x)$, et on affectera x à la classe k qui minimise

$$(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)+\ln(\det\Sigma_k)$$

En estimant μ_k par g^k et Σ_k par V_k, ce critère s'écrit :

$$(x-g^k)'V_k^{-1}(x-g^k)+\ln(\det V_k)$$

Remarques :

1) Ce critère est quadratique mais diffère du précédent par l'ajout de $\ln(\det V_k)$
2) Dans le cas où $\Sigma_k=\Sigma$ (on dit qu'il y'a homoscédasticité), on retrouve le critère linéaire.

b) Hypothèses bayésiennes I

- L'hypothèse de lois étant supposée vraie, on suppose de plus que l'on affecte à chaque classe k une probabilité a priori, notée P_k avec $\sum P_k = 1$.
- Comme la probabilité a posteriori d'être dans la classe k sachant x s'écrit
$$\frac{P_k f_k(x)}{f(x)}$$
où $f(x) = \sum_{k=1}^{q} P_k f_k(x)$, x sera affecté à la classe qui maximise $P_k f_k(x)$.

b) Hypothèses bayésiennes II

- Dans le cas où la loi de la classe k est la loi $\mathcal{N}(\mu_k, \Sigma_k)$, x est affecté à I_{k_0} ssi :

$$(x - g^k)' V_k^{-1} (x - g^k) + \ln(\det V_k) - 2 \ln P_k$$

  est minimum pour $k = k_0$, en estimant μ_k par g^k et Σ_k par V_k.
- On remarque que l'on retrouve les critères précédents si la loi sur les classes est uniforme ($P_k = 1/q$).

Qualité de l'affectation I

Pour mesurer la qualité de la règle d'affectation retenue, on considère le tableau de contingence (ou de classement) croisant la classe d'appartenance avec la classe d'affectation.

- On appelle tableau de classement le tableau T de contingence croisant la classe d'appartenance avec la classe d'affectation. Le terme général $(n_{kk'})_{1 \leq k,k' \leq q}$ de ce tableau est le nombre d'individu de I_k affectés à $I_{k'}$.
- Les éléments bien classés se trouvent dans la diagonale et on peut donc calculer le pourcentage t de bien classés :

$$t = \frac{1}{n} \sum_k n_{kk}$$

qui est un premier indice de qualité.

Qualité de l'affectation II

- On peut aussi examiner le pourcentage de bien classés dans la classe k qui est égal à

$$t_k = \frac{n_{kk}}{n_k}.$$

- Cette mesure de qualité est biaisée (trop optimiste) car elle est fondée sur les individus qui ont servi à calculer la règle d'affectation.
- Une mesure plus réaliste s'obtient de la façon suivante :
  - On divise l'échantillon initial I en deux partie I_A et I_T : l'échantillon I_A (80 à 90% de I) est appelé échantillon d'apprentissage, et I_T (10 à 20% de I) est appelé échantillon test.
  - La procédure consiste alors à construire la règle d'affectation à partir de I_A et à calculer le pourcentage de bien classé sur I_T.

Qualité de l'affectation III

- **Validation croisée :** dans le cas d'un petit échantillon, on calcule le pourcentage de bien classés t_{vc} par validation croisée :
  - on enlève un individu i de léchantillon I,
  - on construit la règle d'affectation sur $I \setminus \{i\}$,
  - on affecte i selon cette règle.

  Après avoir procédé ainsi pour chaque individu de I, on calcule le pourcentage de bien classés sur les n individus.

4. Cas de deux groupes

Il s'agit d'un cas fréquent en pratique, et pour lequel les résultats sont simplifiés puisque les trois points $g = (0)$, g_1 et g_2 sont alignés, et que donc il n'y a qu'un seul axe factoriel discriminant.

4.1 Matrice variance interclasses I

$$
\begin{aligned}
B &= G' D_m G \\
&= (g_1 g_2) \begin{pmatrix} m_1 & 0 \\ 0 & m_2 \end{pmatrix} \begin{pmatrix} (g_1)' \\ (g_2)' \end{pmatrix} \\
&= (g_1 g_2) \begin{pmatrix} m_1 (g_1)' \\ m_2 (g_2)' \end{pmatrix} \\
&= (m_1 g_1^j g_1^{j'} + m_2 g_2^j g_2^{j'})_{j,j'} \\
&= (m_1 g_1 (g_1)' + m_2 g_2 (g_2)')
\end{aligned}
$$

Or $m_1 g_1 + m_2 g_2 = 0$ et $m_1 + m_2 = 1$, d'où :

$$
\begin{aligned}
g_1 &= g_1 - g = g_1 - (m_1 g_1 + m_2 g_2) = m_2 (g_1 - g_2) \\
g_2 &= g_2 - g = g_2 - (m_1 g_1 + m_2 g_2) = -m_1 (g_1 - g_2)
\end{aligned}
$$

4.1 Matrice variance interclasses II

Donc

$$\begin{aligned} B &= m_1 m_2^2 (g_1 - g_2)(g_1 - g_2)' + m_2 m_1^2 (g_1 - g_2)(g_1 - g_2)' \\ &= m_1 m_2 ((g_1 - g_2)(g_1 - g_2)' \end{aligned}$$

4.2 Axe factoriel discriminant I

- Un seul axe factoriel discriminant puisque le nuage $\mathcal{M}_G$ est réduit aux deux points g_1 et g_2. Donc le seul vecteur axial discriminant s'écrit :

$$u^1 = \frac{g_1 - g_2}{||g_1 - g_2||_{V^{-1}}}$$

- Le facteur discriminant $b_{(1)}$ et la valeur propre associé λ_1 sécrivent

$$b_{(1)} = \frac{V^{-1}(g_1 - g_2)}{||g_1 - g_2||_{V^{-1}}}$$

et $\lambda_1 = b'_{(1)} B b_{(1)} = (g_1 - g_2) V^{-1} B V^{-1} (g_1 - g_2) / ||g_1 - g_2||^2_{V^{-1}}$.

- En utilisant $B = m_1 m_2((g_1 - g_2)(g_1 - g_2)'$, on trouve

$$\lambda_1 = m_1 m_2 ||g_1 - g_2||^2_{V^{-1}}$$

.

4.3 Affectation I

- Notons $S_k(x)$ le critère d'affectation qui affecte x à la classe I_{k_0} si $S_k(x)$ est minimum pour $k = k_0$.
- Notons $S(x) = S_2(x) - S_1(x)$: on affecte donc x à la classe I_1 si $S(x)$ est positif, et à I_2 sinon.
- Dans le cas du critère linéaire, les deux régions R_1 et R_2 sont délimitées par l'hyperplan médiateur du segment $g_1 g_2$, appelé hyperplan séparateur de Fisher lorsque $M = V^{-1}$.
- On a vu que l'on peut choisir S_k sous la forme $S_k(x) = (g_k)'M(g_k - 2x)$. Donc

$$S(x) = S_2(x) - S_1(x) = (g_2)'M(g_2 - 2x) - (g_1)'M(g_1 - 2x)$$

Comme $(g_2)'M(g_2) - (g_1)'M(g_1) = -(g_1 - g_2)'M(g_1 + g_2)$, on a :

$$S(x) = 2(g_1 - g_2)'M(x - \frac{g_1 + g_2}{2})$$

4.3 Affectation II

- Posons $f(x) = (g_1 - g_2)'Mx$. On affecte x à la classe I_1 si $f(x) > f(\frac{g_1 + g_2}{2})$ et à I_2 sinon.
- Dans le cas où $M = V^{-1}$, alors $b = \frac{V^{-1}(g_1 - g_2)}{||g_1 - g_2||_{V^{-1}}}$, et par conséquent x est affecté à la classe I_1 si le score $z = b'x$ est supérieur à celui du milieu de $g_1 g_2$, c'est à dire à $\frac{1}{2}b'(g_1 + g_2)$; et x sera affecté à I_2 sinon.
- Dans le cas où $M = W^{-1}$, cela ne change pas l'hyperplan médiateur de $g_1 g_2$ et la règle d'affectation.
- La fonction $x \mapsto (g_1 - g_2)'W^{-1}x$ est appelé **fonction discriminante de Fisher**.

5. Multicolinéarité I

- En analyse discriminante comme en régression, on doit inverser la matrice variance V des variables explicatives, ce qui peut poser problème en cas de multicolinéarité des variables explicatives.
- Si le déterminant de V est nul, il existe une ou plusieurs relations entre les variables explicatives. Il suffit alors de garder r' variables explicatives linéairement indépendantes puis de faire l'analyse discriminante sur ces variables, r' $(r' < p)$ étant le rang de V.
- S'il existe une ou plusieurs relations approchées entre les variables explicatives, V est en théorie inversible mais cette matrice qui va comporter des termes de valeur très élevée (puisque dans le calcul de V^{-1}, le déterminant de V, qui intervient au dénominateur, est petit) risque d'être très sensible aux fluctuations d'échantillonnage et de conduire à des résultats peu stables.

5. Multicolinéarité II

Pour pallier ce problème de multicolinéarité, on se sert des mêmes techniques que celles utilisées en régression, à savoir :

a) Analyse discriminante pas à pas ascendante ou stepwise ; critère de sélection : trace de $V^{-1}B$, pourcentage de bien classés. On arrêtera la procédure de sélection sur la base d'un graphique indiquant le pourcentage de bien camssés en fonction du pas (nombre de sélectionnées).

b) A.F. du tableau X des variables explicatives (A.C.P., A.C.) et réalisations de l'A.F.D. sur les premiers facteurs donnant un pourcentage d'inertie suffisant. La matrice variance associée est diagonale, donc facilement inversible.

c) Effectuer dans l'espace explicatif (espace initial $\mathbb{R}^p$ ou dans l'espace des premiers axes factoriels de l'A.F. de X), une discrimination par boule : on affecte x à la classe majoritaire dans le voisinage de x, ce voisinage étant défini comme les u points x_i les plus proches de x (avec $u = 5$ ou 10).

6. Analyse Discriminante sur variables qualitatives

Méthode DISQUAL. Si toutes les variables explicatives sont qualitatives, on effectue l'A.F.C. du tableau disjonctif complet T associé, puis on fait l'A.D. sur les facteurs non triviaux (i.e. associés à une valeur propre non nulle) ou sur les facteurs non triviaux supérieurs à un pourcentage d'inertie donné, issus de cette A.F.C.

7. Cas d'un échantillon de taille n

Les paramètres μ_k, Σ_k, Σ et P_k sont estimés par leurs correspondants paramétriques g^k, V_k, V et $\frac{n_k}{n}$ (avec $n_k = |I_k|$).
Dans le cas $\Sigma_k = \Sigma$, Σ peut être estimée par la matrice variance intraclasse :

$$W = \sum_{k=1}^{q} \frac{n_k}{n} V_k$$

ou mieux :

$$W^* = \frac{n}{n-q} W$$

qui est un estimateur sans biais de Σ.