Séance 5: Analyse Factorielle Discriminante

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Cinquième partie V

Analyse Factorielle Discriminante

Données - Objectifs

- Prédire une variable qualitative à k modalités en fonction de p prédicteurs
- n observations réparties en k classes décrites par p variables explicatives
- Objectif descriptif : chercher les c.l. permettant de séparer au mieux les k catégories
- Objectif décisionnel : classer un nouvel individu en fonction des p valeurs des prédicteurs
- Applications : aide à la décision en médecine, prévision en météorologie, finance, ...

Données et Notations

- n individus e_i formant un nuage $E \subset \mathbb{R}^p$, pondérés par p_i
- Partition $E_1, \ldots E_k$ des individus pondérés par q_j

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

- Centre de gravités $g_1, \ldots g_k$
- Matrice de variances $V_1, \ldots V_k$
- Matrice de poids des individus D.

Relations basiques

- $$g_j = \sum_{i \in E_j} \frac{p_i}{q_j} e_i \qquad g = \sum_{j=1}^{k} q_j g_j \qquad V_j = \sum_{i \in E_j} \frac{p_i}{q_j} (e_i - g_j)(e_i - g_j)'$$
- B matrice de variance inter-classe, W matrice de variance intra-classe

$$B = \sum_{j=1}^{k} q_j (g_j - g)(g_j - g)' \quad \text{et} \quad W = \sum_{j=1}^{k} q_j V_j$$

- Formule de reconstitution : $V = B + W$
- Dans la suite de l'étude, on supposera $p_i = 1/n$

Données et Notations

On suppose les données centrées. On écrit les données en :

$$\left(\begin{array}{cccc|c} 1 & 0 & \dots & 0 & \\ 0 & & & & \\ \vdots & & A & & X \\ & & & & \\ 0 & 0 & & 1 & \end{array}\right)$$

X matrice de taille $n \times p$, A tableau logique associé à la variable qualitative.
Les k centres de gravité se lisent en ligne dans la matrice

$$G = (A'DA)^{-1}(A'DX)$$

La matrice des poids des sous-nuages est

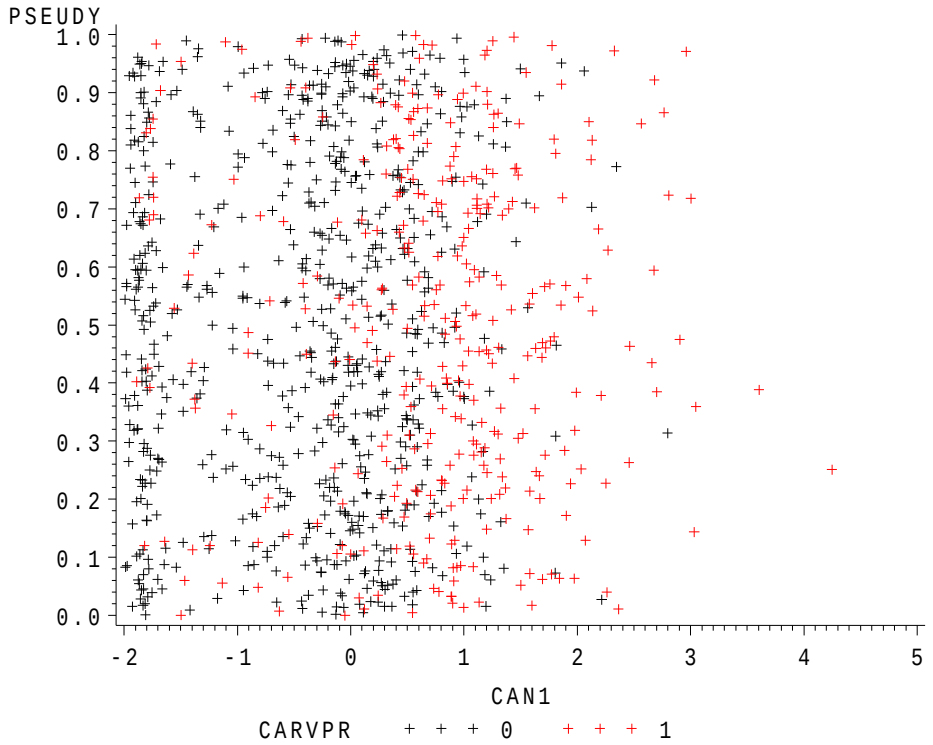
$$D_q = A'DA$$

La matrice de variance interclasse vaut

$$D = (X'DA)D_q^{-1}(A'DX)$$

Analyse Factorielle Discriminante

L'AFD consiste à chercher des nouvelles variables discriminantes séparant au mieux les k groupes.



L'axe 1 ne fournit pas une très bonne séparation. L'axe 2 n'est pas du tout discriminant.

Modèle Analyse Factorielle Discriminante

- Idée : si a désigne un axe de projection, les k centres de gravité doivent être le plus séparés possible alors que l'inertie de chaque nuage projeté issu de E_k doit être la plus faible possible.
- Propriété équivalente : le nuage des projections des g_i doit avoir une inertie maximale et les inerties intra doivent être faible. (Pourquoi ?)
- L'inertie du nuage projeté sur a vaut

$$I_{inter,proj}(a) = a'MBMa$$

- L'inertie intra projetée sur a vaut

$$\forall j \in \{1, \dots k\} \qquad I^j_{intra,proj}(a) = a'MV_jMa$$

- Critère à maximiser :

$$C(a) = \frac{a'MBMa}{a'MVMa}$$

Résolution du modèle

- Maximum réalisé pour a vecteur propre de $(MVM)^{-1}MBM$
- On cherche la plus grande valeur propre λ_1 de :

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

- Facteur discriminant $u = Ma$

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

- $0 \leq \lambda_1 \leq 1$
- $\lambda_1 = 1$ correspond à une dispersion intra-classe nulle pour les projections sur a. Il y a discrimination parfaite si les centres de gravités se projettent en des points différents.
- $\lambda_1 = 0$ correspond au cas où le meilleur axe ne sépare pas les centres de gravités : nuages concentriques et aucune séparation linéaire possible.
- λ est une mesure pessimiste du pouvoir de discrimination d'un axe
- Le nombre de valeurs propres non nulles est égal à $k - 1$ lorsque $n > p > k$

Représentation graphique et interprétations

- Représentation simultanée des individus et barycentres des classes
- Appréciation visuelle de la discrimination
- Un cosinus au carré permet de préciser la qualité de représentation d'un individu
- Projection des variables sur les axes : utilité pour interpréter les axes en fonction des variables initiales
- Si les groupes sont bien discriminés, on peut faire une bonne interprétation des axes *via l'AFD*
- *Sinon, l'AFD est inutile*

Exemple

Les données décrivent trois classes d'insectes sur lesquels ont été réalisées 6 mesures anatomiques. On cherche à savoir si ces mesures permettent de retrouver la typologie de ces insectes. Ce jeu de données "scolaire" conduit à une bien meilleure discrimination que ce que l'on peut obtenir dans une situation concrète.

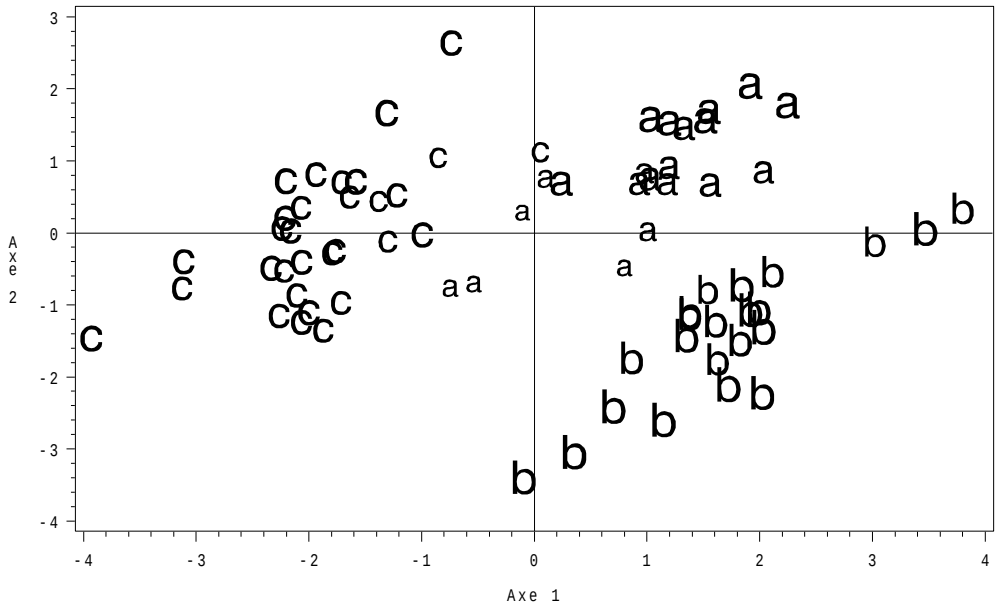


FIG.: *Insectes : premier plan factoriel de l'ACP.*

Exemple

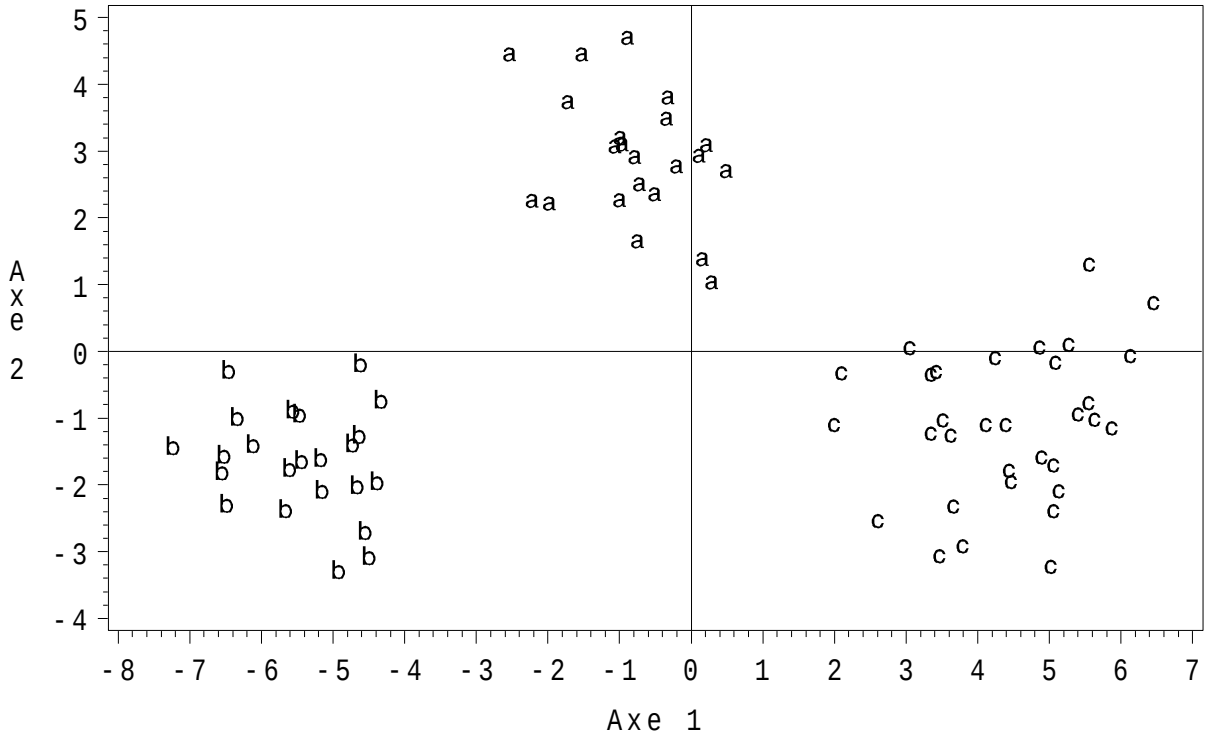


FIG.: *Insectes : premier plan factoriel de l'AFD.*