Analyse factorielle discriminante (AFD)

Résumé

Méthode factorielle de réduction de dimension pour l'exploration statistique de variables quantitatives et d'une variable qualitative. Construction du modèle statistique associé, estimation. Représentation graphique optimale des classes des individus, liens avec d'autres définitions de l'AFD.

Travaux pratiques *de complexité croissante par l'études de données socio-économiques.*

Retour au plan du cours.

1 Introduction

1.1 Données

Les données sont constituées de

- p variables *quantitatives* $X^1, \ldots, X^p$ jouant le rôle de variables explicatives comme dans le modèle linéaire,
- une variable *qualitative* T, à m modalités $\{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$, jouant le rôle de variable à expliquer.

La situation est analogue à celle de la régression linéaire multiple mais, comme la variable à expliquer est qualitative, on aboutit à une méthode très différente. Les variables sont observées sur l'ensemble Ω des n individus affectés des poids $w_i > 0$, $(\sum_{i=1}^n w_i = 1)$, et l'on pose

$$\mathbf{D} = \text{ diag}(w_i \,; i = 1, \ldots, n).$$

La variable T engendre une partition $\{\Omega_\ell \,; \ell = 1, \ldots, m\}$ de l'ensemble Ω des individus dont chaque élément est d'effectif n_ℓ.

On note $\mathbf{T}$ $(n \times m)$ la matrice des indicatrices des modalités de la variable T ; son terme général est

$$t_i^\ell = t^\ell(\omega_i) = \begin{cases} 1 & \text{si } T(\omega_i) = \mathcal{T}_\ell \\ 0 & \text{sinon} \end{cases}.$$

En posant

$$\overline{w_\ell} = \sum_{i \in \Omega_\ell} w_i,$$

il vient

$$\overline{\mathbf{D}} = \mathbf{T}'\mathbf{D}\mathbf{T} = \text{ diag}(\overline{w_1}, \ldots, \overline{w_m}).$$

1.2 Objectifs

Deux techniques cohabitent sous la même appellation d'analyse discriminante :

descriptive : cette méthode recherche, parmi toutes les ACP possibles sur les variables X^j, celle dont les représentations graphiques des individus *discriminent* "au mieux" les m classes engendrées par la variable T (e.g. recherche de facteurs de risque en statistique médicale) ;

décisionnelle : connaissant, pour un individu donné, les valeurs des Y^j mais pas la modalité de T, cette méthode consiste à affecter cet individu à une modalité (e.g. reconnaissance de formes). Cette méthode est décrite dans la partie *modélisation* de ce cours.

Remarque. — Lorsque le nombre et les caractéristiques des classes sont connues, il s'agit d'une *discrimination* ; sinon, on parle de *classification* ou encore, avec des hypothèses sur les distributions, de *reconnaissance de mélanges*.

1.3 Notations

On note $\mathbf{X}$ la matrice $(n \times p)$ des données quantitatives, $\mathbf{G}$ la matrice $(m \times p)$ des barycentres des classes :

$$\mathbf{G} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{X} = \begin{bmatrix} \mathbf{g}_1{}' \\ \vdots \\ \mathbf{g}_m{}' \end{bmatrix} \text{ où } \mathbf{g}_\ell = \frac{1}{\overline{w_\ell}} \sum_{i \in \Omega_\ell} w_i \mathbf{x}_i,$$

et $\mathbf{X}_e$ la matrice $(n \times p)$ dont la ligne i est le barycentre $\mathbf{g}_\ell$ de la classe Ω_ℓ à laquelle appartient l'individu i :

$$\mathbf{X}_e = \mathbf{T}\mathbf{G} = \mathbf{P}\mathbf{G} \,;$$

$\mathbf{P} = \mathbf{T}\overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}$ est la matrice de projection $\mathbf{D}$-orthogonale sur le sous-espace engendré par les indicatrices de T ; c'est encore l'espérance conditionnelle sachant T.

Deux matrices "centrées" sont définies de sorte que $\overline{\mathbf{X}}$ se décompose en

$$\overline{\mathbf{X}} = \overline{\mathbf{X}}_r + \overline{\mathbf{X}}_e$$

avec

$$\overline{\mathbf{X}}_r = \mathbf{X} - \mathbf{X}_e \text{ et } \overline{\mathbf{X}}_e = \mathbf{X}_e - \mathbf{1}_n\overline{\mathbf{x}}'.$$

On note également $\overline{\mathbf{G}}$ la matrice centrée des barycentres :

$$\overline{\mathbf{G}} = \mathbf{G} - \mathbf{1}_m\overline{\mathbf{x}}'.$$

On appelle alors variance intraclasse (within) ou résiduelle :

$$\mathbf{S}_r = \overline{\mathbf{X}}_r{}'\mathbf{D}\overline{\mathbf{X}}_r = \sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell} w_i(\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)',$$

et variance interclasse (between) ou expliquée :

$$\mathbf{S}_e = \overline{\mathbf{G}}'\overline{\mathbf{D}\mathbf{G}} = \overline{X}_e'\mathbf{D}\overline{X}_e = \sum_{\ell=1}^{m}\overline{w_\ell}(\mathbf{g}_\ell - \overline{\mathbf{x}})(\mathbf{g}_\ell - \overline{\mathbf{x}})'.$$

PROPOSITION 1. — *La matrice des covariances se décompose en*

$$\mathbf{S} = \mathbf{S}_e + \mathbf{S}_r.$$

2 Définition

2.1 Modèle

Dans l'espace des individus, le principe consiste à projeter les individus dans une direction permettant de mettre en évidence les groupes. À cette fin, Il faut privilégier la variance interclasse au détriment de la variance intraclasse considérée comme due au bruit.

En ACP, pour chaque effet $\mathbf{z}_i$ à estimer, on ne dispose que d'une observation $\mathbf{x}_i$; dans le cas de l'AFD on considère que les éléments d'une même classe Ω_ℓ sont les observations répétées n_ℓ fois du même effet $\mathbf{z}_\ell$ pondéré par $\overline{w_\ell} = \sum_{i\in\Omega_\ell} w_i$. Le modèle devient donc :

$$\begin{array}{l} \{\mathbf{x}_i\,;i=1,\ldots,n\},\ n \text{ vecteurs indépendants de } E,\\ \forall\ell,\forall i\in\Omega_\ell, \mathbf{x}_i = \mathbf{z}_\ell + \boldsymbol{\varepsilon}_i \text{ avec } \left\{\begin{array}{l} E(\boldsymbol{\varepsilon}_i)=0,\ \text{var}(\boldsymbol{\varepsilon}_i)=\boldsymbol{\Gamma},\\ \boldsymbol{\Gamma} \text{ régulière et inconnue,}\end{array}\right.\\ \exists A_q,\ \text{sous-espace affine de de dimension } q \text{ de } E \text{ tel que}\\ \forall\ell, \mathbf{z}_\ell \in A_q,\ (q<\min(p,m-1)).\end{array} \qquad (1)$$

Remarque. — Soit $\overline{\mathbf{z}} = \sum_{\ell=1}^{m}\overline{w_\ell}\mathbf{z}_\ell$. Le modèle entraîne que $\overline{\mathbf{z}} \in A_q$. Soit E_q le sous-espace de dimension q de E tel que $A_q = \overline{\mathbf{z}} + E_q$. Les paramètres à estimer sont E_q et $\{\mathbf{z}_\ell\,;\ell=1,\ldots,m\}$; $\overline{w_\ell}$ est un paramètre de nuisance qui ne sera pas considéré.

2.2 Estimation

L'estimation par les moindres carrés s'écrit ainsi :

$$\min_{E_q,\mathbf{z}_\ell}\left\{\sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell} w_i\left\|\mathbf{x}_i - \mathbf{z}_\ell\right\|_{\mathbf{M}}^2\ ;\ \dim(E_q)=q, \mathbf{z}_\ell - \overline{\mathbf{z}} \in E_q\right\}.$$

Comme on a

$$\sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell} w_i\left\|\mathbf{x}_i - \mathbf{z}_\ell\right\|_{\mathbf{M}}^2 = \sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell} w_i\left\|\mathbf{x}_i - \mathbf{g}_\ell\right\|_{\mathbf{M}}^2 + \sum_{\ell=1}^{m}\overline{w_\ell}\left\|\mathbf{g}_\ell - \mathbf{z}_\ell\right\|_{\mathbf{M}}^2,$$

on est conduit à résoudre :

$$\min_{E_q,z_\ell}\left\{\sum_{\ell=1}^{m}\overline{w_\ell}\left\|\mathbf{g}_\ell - \mathbf{z}_\ell\right\|_{\mathbf{M}}^2\ ;\ \dim(E_q)=q, \mathbf{z}_\ell - \overline{\mathbf{z}} \in E_q\right\}.$$

La covariance $\sigma^2\boldsymbol{\Gamma}$ du modèle (1) étant inconnue, il faut l'estimée. Ce modèle stipule que l'ensemble des observations d'une même classe Ω_l suit une loi (inconnue) de moyenne $\mathbf{z}_e ll$ et de variance $\boldsymbol{\Gamma}$. Dans ce cas particulier, la matrice de covariances intraclasse ou matrice des covariances résiduelles empiriques $\mathbf{S}_r$ fournit donc une estimation "optimale" de la métrique de référence :

$$\mathbf{M} = \widehat{\boldsymbol{\Gamma}}^{-1} = \mathbf{S}_r^{-1}$$

.

PROPOSITION 2. — *L'estimation des paramètres E_q et $\mathbf{z}_\ell$ du modèle 1 est obtenue par l'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$. C'est l'Analyse Factorielle Discriminante (AFD) de $(\mathbf{X}|\mathbf{T}, \mathbf{D})$.*

3 Réalisation de l'AFD

Les expressions matricielles définissant les représentations graphiques et les aides à l'interprétation découlent de celles de l'ACP.

3.1 Matrice à diagonaliser

L'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ conduit à l'analyse spectrale de la matrice positive $\mathbf{S}_r^{-1}$-symétrique :

$$\overline{\mathbf{G}}'\overline{\mathbf{D}}\,\overline{\mathbf{G}}\mathbf{S}_r^{-1} = \mathbf{S}_e\mathbf{S}_r^{-1}.$$

Comme $\mathbf{S}_r^{-1}$ est régulière, cette matrice est de même rang que $\mathbf{S}_e$ et donc de même rang que $\mathbf{G}$ qui est de dimension $(m \times p)$. Les données étant centrées lors de l'analyse, le rang de la matrice à diagonaliser est

$$h = \text{rang}(\mathbf{S}_e\mathbf{S}_r^{-1}) \leq \inf(m-1, p),$$

qui vaut en général $m-1$ c'est-à-dire le nombre de classes moins un.

On note $\lambda_1 \geq \cdots \geq \lambda_h > 0$ les valeurs propres de $\mathbf{S}_e\mathbf{S}_r^{-1}$ et $\mathbf{v}^1, \ldots, \mathbf{v}^h$ les vecteurs propres$\mathbf{S}_r^{-1}$-orthonormés associés. On pose

$$\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_h) \text{ et } \mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^h].$$

Les vecteurs $\mathbf{v}^k$ sont appelés *vecteurs discriminants* et les sous-espaces vectoriels de dimension 1 qu'ils engendrent dans $\mathbb{R}^p$ les *axes discriminants*.

3.2 Représentation des individus

L'espace des individus est ($\mathbb{R}^p$, b. c., $\mathbf{S}_r^{-1}$). Une représentation simultanée des individus $\mathbf{x}_i$ et des barycentres $\mathbf{g}_\ell$ des classes par rapport aux mêmes axes discriminants est obtenue dans cet espace au moyen des coordonnées :

$$\begin{aligned} \mathbf{C} &= \overline{\mathbf{X}}\mathbf{S}_r^{-1}\mathbf{V} \text{ pour les individus et} \\ \overline{\mathbf{C}} &= \overline{\mathbf{G}}\mathbf{S}_r^{-1}\mathbf{V} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{C} \text{ pour les barycentres.} \end{aligned}$$

Les individus initiaux sont projetés comme des individus supplémentaires dans le système des axes discriminants. Comme en ACP, on peut calculer des cosinus carrés pour préciser la qualité de représentation de chaque individu.

Il est utile de différencier graphiquement la classe de chaque individu afin de pouvoir apprécier visuellement la qualité de la discrimination.

3.3 Représentation des variables

L'espace des variables est ($\mathbb{R}^m$, b. c., $\overline{\mathbf{D}}$). Chaque variable X^j est représenté par un vecteur dont les coordonnées dans le système des axes factoriels est une ligne de la matrice $\mathbf{V}\mathbf{\Lambda}^{1/2}$.

3.4 Interprétations

Les interprétations usuelles : la norme est un écart-type, un cosinus d'angle est un coefficient de corrélation, doivent être faites en termes d'écarts-types et de corrélations *expliquées* par la partition.

La représentation des variables est utilisée pour interprétée les axes en fonction des variables initiales conjointement avec la matrice des corrélations expliquées variables×facteurs : $\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$. La matrice $\mathbf{\Sigma}_e^{-1}$ étant la matrice diagonale des écarts-types expliqués σ_e^j c'est-à-dire des racines carrées des éléments diagonaux de la matrice $\mathbf{S}_e$.

Le point pratique essentiel est de savoir si la représentation des individus-barycentres et des individus initiaux permet de faire une bonne discrimination entre les classes définies par la variable T. Si ce n'est pas le cas, l'AFD ne sert à rien, les X^j n'expliquent pas T. Dans le cas favorable, le graphique des individus permet d'interpréter la discrimination en fonction des axes et, celui des variables, les axes en fonction des variables initiales. La synthèse des deux permet l'interprétation de T selon les X^j.

4 Variantes de l'AFD

4.1 Individus de mêmes poids

L'AFD peut être définie de différentes façon. Dans la littérature anglo-saxonne, et donc dans la version standard d'AFD du logiciel SAS (procédure `candisc`), ce sont les estimations sans biais des matrices de variances "intra"

(within) et “inter” (between) qui sont considérées dans le cas d’individus de mêmes poids $1/n$.

Dans ce cas particulier,

$$\mathbf{D} = \frac{1}{n}\mathbf{I}_n \text{ et } \overline{\mathbf{D}} = \frac{1}{n}\text{diag}(n_1, \ldots, n_m) \text{ où } n_\ell = \text{card}(\Omega_\ell)$$

et les matrices de covariances empiriques ont alors pour termes généraux :

$$\begin{aligned}
(\mathbf{S})_j^k &= \frac{1}{n}\sum_{i=1}^{n}(x_i^j - \overline{x}^j)(x_i^k - \overline{x}^k),\\
(\mathbf{S}_e)_j^k &= \frac{1}{n}\sum_{\ell=1}^{m} n_\ell (g_\ell^j - \overline{x}^j)(g_\ell^k - \overline{x}^k),\\
(\mathbf{S}_r)_j^k &= \frac{1}{n}\sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell}(x_i^j - g_\ell^j)(x_i^k - g_\ell^k).
\end{aligned}$$

Du point de vue de le Statistique inférentielle, on sait que les quantités calculées ci-dessus ont respectivement $(n-1)$, $(m-1)$ et $(n-m)$ degrés de liberté. En conséquence, ce point de vue est obtenu en remplaçant dans les calculs

$$\begin{aligned}
\mathbf{S} \quad &\text{par} \quad \mathbf{S}^* = \frac{n}{n-1}\mathbf{S},\\
\mathbf{S}_e \quad &\text{par} \quad \mathbf{S}_e^* = \mathbf{B} = \frac{n}{m-1}\mathbf{S}_e,\\
\mathbf{S}_r \quad &\text{par} \quad \mathbf{S}_r^* = \mathbf{W} = \frac{n}{n-m}\mathbf{S}_r.
\end{aligned}$$

Les résultats numériques de l’AFD se trouvent alors modifiés de la façon suivante :

– matrice à diagonaliser :	$\mathbf{S}_e^*\mathbf{S}_r^{*-1}$	$=$	$\frac{n-m}{m-1}\mathbf{S}_e\mathbf{S}_r^{-1}$,
– valeurs propres :	$\mathbf{\Lambda}^*$	$=$	$\frac{n-m}{m-1}\mathbf{\Lambda}$,
– vecteurs propres :	$\mathbf{V}^*$	$=$	$\sqrt{\frac{n}{n-m}}\mathbf{V}$,
– représentation des barycentres :	$\overline{\mathbf{C}}^*$	$=$	$\sqrt{\frac{n-m}{n}}\overline{\mathbf{C}}$,
– représentation des variables :	$\mathbf{V}^*\mathbf{\Lambda}^{*1/2}$	$=$	$\sqrt{\frac{n}{m-1}}\mathbf{V}\mathbf{\Lambda}^{1/2}$,
– corrélations variables-facteurs :	$\mathbf{\Sigma}_e^{*-1}\mathbf{V}^*\mathbf{\Lambda}^{*1/2}$	$=$	$\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

Ainsi, les représentations graphiques sont identiques à un facteur d’échelle près tandis que les parts de variance expliquée et les corrélations variables-facteurs sont inchangées.

4.2 Métrique de Mahalanobis

L’AFD est souvent introduite dans la littérature francophone comme un cas particulier d’Analyse Canonique entre un ensemble de p variables quantitatives et un ensemble de m variables indicatrices des modalités de T. La proposition suivante établit les relations entre les deux approches :

PROPOSITION 3. — *l’ACP de* $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ *conduit aux mêmes vecteurs principaux que l’ACP de* $(\mathbf{G}, \mathbf{S}^{-1}, \overline{\mathbf{D}})$. *Cette dernière est l’ACP des barycentres des classes lorsque l’espace des individus est muni de la métrique dite de Mahalanobis* $\mathbf{M} = \mathbf{S}^{-1}$ *et l’espace des variables de la métrique des poids des classes* $\overline{\mathbf{D}}$.

Les résultats numériques de l’AFD se trouvent alors modifiés de la façon suivante :

– matrice à diagonaliser :	$\mathbf{S}_e\mathbf{S}^{-1}$,
– valeurs propres :	$\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}$,
– vecteurs propres :	$\mathbf{V}(\mathbf{I}+\mathbf{\Lambda})^{1/2}$,
– représentation des barycentres :	$\overline{\mathbf{C}}(\mathbf{I}+\mathbf{\Lambda})^{-1/2}$,
– représentation des variables :	$\mathbf{V}\mathbf{\Lambda}^{1/2}$,
– corrélations variables-facteurs :	$\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

Les représentations graphiques des individus (voir ci-dessus) ne diffèrent alors que d’une homothétie et conduisent à des interprétations identiques, les corrélations variables-facteurs ainsi que les représentations des variables sont inchangées.

5 Exemples

5.1 Les insectes de Lubitsch

Cette méthode est illustrée par une comparaison des sorties graphiques issues d’une ACP et d’une AFD. Les données décrivent trois classes d’insectes

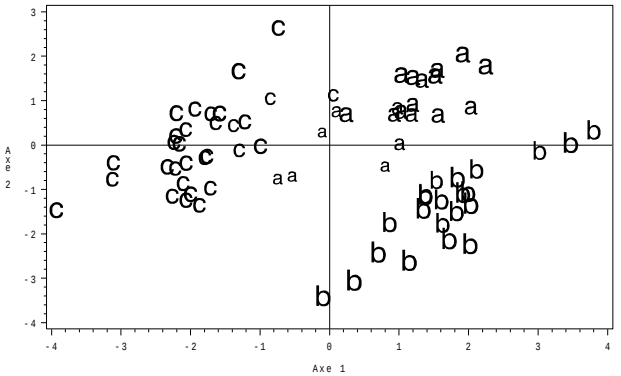

FIGURE 1 – *Insectes : premier plan factoriel de l'ACP.*

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", comme les fameux iris de Fisher conduit à une discrimination assez évidente. La comparaison entre l'ACP et l'AFD met clairement en évidence le rôle de la distance $\mathbf{S}_R^{-1}$ que la forme des nuages de chaque classe en analyse discriminante.

5.2 Données génomiques

Les données génomiques pose évidemment des problèmes à l'analyse discriminante ; le grand nombre de gènes/variables par rapport au nombre de souris/individus rend impossible l'inversion de la matrice des covariances intra-classes. Aussi, en s'aidant de la sélection de variables suggérée par l'analyse en composantes principales, une analyse factorielle discriminante a été calculée sur les seules souris sauvages (`WR`) pour qui les régimes apparaissaient déjà bien différenciés sur l'ACP. Les variables ne sont pas représentées mais les rapprochements déjà évoqués pour l'ACP sont confirmés et précisés.

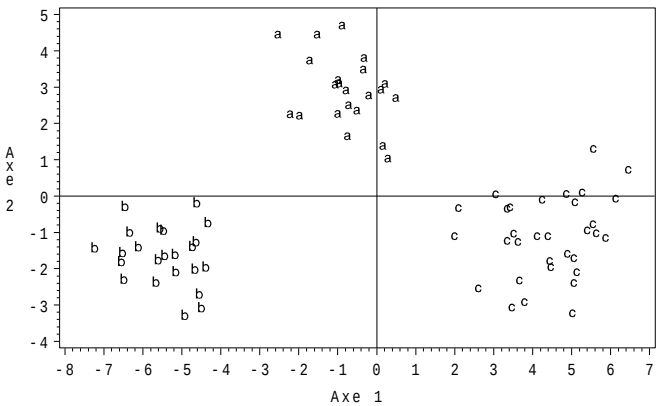

FIGURE 2 – *Insectes : premier plan factoriel de l'AFD.*

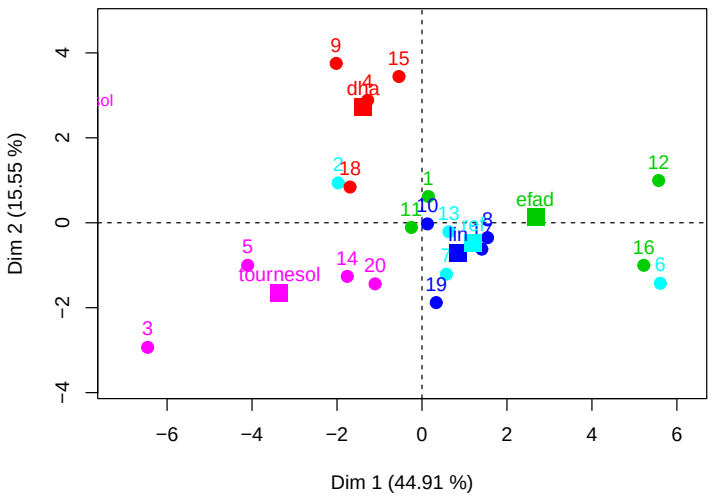

FIGURE 3 – *Souris : Les souris de génotype* `WT` *dans le premier plan factoriel de l'AFD calculée avec une sélection de variables d'expression de gènes conditionnellement au régime.*