

I. Analyse factorielle discriminante

a) Objectifs

La question est de savoir si plusieurs groupes d'individus pour lesquels on a observé p caractéristiques (variables) sont bien discriminés par ces variables. Si c'est bien le cas, on peut effectivement espérer classer correctement les individus en observant ces p caractéristiques.

Nous chercherons à ce niveau à comprendre le rôle que joue les variables dans la discrimination des individus. La méthode proposée permet également de classer des individus (de déterminer de quel groupe est le plus proche un nouvel individu) mais cela ne sera pas abordé à ce stade.

Les notations utilisées sont les suivantes :

g : groupe.

X_g : matrice des données du groupe g (n_g individus, p variables).

\bar{X}_k : vecteur des moyennes ($p \times 1$).

S_k : matrice variance/covariance d'un groupe ($p \times p$).

b) Analyse factorielle

Pour comprendre comment les p variables observées contribuent à discriminer les g groupes, on commence par décomposer la variance :

- variance totale : $T = \frac{1}{n} X^t M X = \frac{1}{n} \{X^t X - n \bar{X} \bar{X}^t\}$.
- variance within : $W = \frac{1}{n} \sum_{k=1}^g X_k^t M_k X_k = \frac{1}{n} \left\{ \sum_{k=1}^g (X_k^t X - n_k \bar{X}_k \bar{X}_k^t) \right\}$.
- variance between : $B = \frac{1}{n} \sum_{k=1}^g n_k (\bar{X}_k - \bar{X})(\bar{X}_k - \bar{X})^t$.

$$\rightarrow T = W + B$$

On cherche ensuite les axes factoriels discriminants : il s'agit des axes qui discriminent le mieux les groupes et non les individus. Si v est le premier axe factoriel, la projection des individus est $z = X v \in \mathbb{R}^n$. La dispersion des individus est donc :

$$v^t T v = v^t W v + v^t B v$$

L'objectif poursuivi est :

$$v = \operatorname{argmax}_v \frac{v^t B v}{v^t W v}$$
$$\text{sc } v^t T v = 1$$

Comme pour l'ACP, pour les axes suivants, on rajoute la contrainte d'orthogonalité. Le vecteur v est le vecteur propre de $T^{-1}B$ associé à la plus grande valeur propre de $T^{-1}B$ et :

$z^1 = X v$: première variable discriminante.

$$I = \frac{v^t B v}{v^t T v} \text{ est la part de la dispersion de } X \text{ qui est due à la différence entre les groupes.}$$

Pour interpréter les variables factorielles (qui représentent à présent les individus), on peut à présent étudier les corrélations entre les variables initiales X^i et ces nouvelles variables en projetant ces variables dans l'espace factoriel :

