Discriminant

La méthode a été introduite en 1936 par R. A. Fisher. Il s'intéressait à la taxonimie végétale, c'est-à-dire déterminer l'espèce de fleurs à partir de diverses mesures.

Notation :

Soit $X = (X_{ij})$, qui est une matrice de dimension $n \times p$, où n est le nombre d'individus dans l'échantillon, p est le nombre de variables et X_{ij} est la valeur de la je variable pour le ie individus.

Identification des groupes :

— $I_k = $ ensemble des individus du groupe k
— $n_k = |I_k| = $ cardinalité de I_k.
— $n_1 + \cdots + n_q = n$, où q est le nombre de groupes.

Score de l'analyse discriminante : on a des observations dans R^p. Pour faire de la classification à partir de $X_1, \dots, X_p$, on doit partionner R^p en q sous-ensembles de sorte que chacun des q sous-ensembles est associé à un des q groupes.

On va chercher à passer de la dimension p à la dimension 1 en calculant un score $f(x_1, \dots, x_p) \in \mathbb{R}$ pour chaque observation et ensuite utiliser ce score pour déterminer le groupe d'appartenance (et donc partionner R). Le score proposé par Fisher est une combinaison linéaire des variables, c'est-à-dire

$$f(X_1, \dots, X_p) = a^\top X + b = a_1 X_1 + \cdots + a_p X_p + b.$$

On en déduira q intervalles de décision $S_1, \dots, S_q$ associés aux groupes.

Remarque

Sans perte de généralité, on peut choisir

$$-b = a_1 \overline{X}_1 + \cdots + a_p \overline{X}_p = a^\top \overline{X}$$

ce qui permet de centrer les variables en enlevant le vecteur de moyenne

$$\overline{X} = \left(\overline{X}_1, \dots, \overline{X}_p\right).$$

Il ne reste plus qu'à choisir le vecteur $a = (a_1, \dots, a_p)$.

On voudrait choisir le vecteur a de sorte que les scores soient, à la fois, très différents entre les groupes et très similaires à l'intérieur d'un groupe. On s'intéresse donc à la variabilité des scores à l'intérieur des groupes et entre les groupes.

Étant donné $a \in R^p$, on a :

$$\mathrm{Var}(f(X_1, \dots, X_p)) = \mathrm{Var}(a^\top X) = a^\top \mathrm{Var}(X) a,$$

que nous estimons à partir des n observations par

$$\widehat{\mathrm{Var}}(f(X_1, \dots, X_p)) = \frac{1}{n} a^\top S a.$$

La base de l'analyse discriminante repose sur le fait que

$$S = W + B,$$

où W est la matrice de variance intragroupe et B est la matrice de variance intergroupe.

On peut prouver ce résultat en considérant la définition des matrices S, W et B. La moyenne de la variable j pour tous les individus de l'échantillon est

$$\overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}.$$

La moyenne de la variable j pour les individus du groupe k est

$$\overline{X}_{kj} = \frac{1}{n_k} \sum_{i \in I_k} X_{ij}.$$

La somme des carrés totale est

$$s_{jj'} \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)(X_{ij'} - \overline{X}_{j'}).$$

On tirerait de la matrice S une estimation de $\mathrm{Cov}(X_j, X_{j'})$ si toutes les observations provenaient d'un même groupe. On définit $s_{jj'}$ comme étant

$$s_{jj'} = w_{jj'} + b_{jj'},$$

où

$$w_{jj'} = \sum_{k=1}^{q} \sum_{i \in I_k} (X_{ij} - \overline{X}_{kj})(X_{ij'} - \overline{X}_{kj'}),$$

$$b_{jj'} = \sum_{k=1}^{q} n_k(\overline{X}_{kj} - \overline{X}_j)(\overline{X}_{kj'} - \overline{X}_{j'}).$$

Preuve :

1. Poser $X_{ij} - \overline{X}_j = X_{ij} - \overline{X}_{kj} + \overline{X}_{kj} - \overline{X}_j$, dans la définition de $s_{jj'}$, iden pour $X_{ij'}$.
2. Développer les produits.
3. Remplacer $\sum_{i=1}^{n}$ par $\sum_{k=1}^{q} \sum_{i\in I_k}$.
4. Faire les simplications appropriées.

On obtient

$$\widehat{\mathrm{Var}}(a^\top X) = \frac{1}{n} a^\top S a = \frac{1}{n}\left(a^\top W a + a^\top B a\right).$$

On se rappelle que l'on veut choisir le vecteur a pour que les scores puissent facilement séparer les groupes. En d'autres mots, on veut des scores les plus similaires possible à l'intérieur d'un groupe et des scores les plus différents possible entre les groupes.

On propose de choisir le vecteur $ain\mathbb{R}^p$ pour maximiser

$$\frac{a^\top B a}{a^\top W a} \quad \text{où} \quad \frac{a^\top B a}{a^\top S a}.$$

ce vecteur est unique à une constante près.

Ce problème peut être reformuler des façons suivantes :

— Maximiser $a^\top Ba/a^\top Sa$ sous la contrainte que $a^\top a = 1$.
— Maximiser $a^\top Ba$ sous la contrainte que $a^\top Sa = 1$.
— Maximiser $c^\top S^{-1/2} B S^{-1/2} c$ sous la contrainte que $c^\top c = 1$, où $c = S^{1/2} a$.

En écrivant la troisième formulation

$$c^\top \left(S^{-1/2} B S^{-1/2}\right) c \quad \text{s.c.} c^\top c = 1,$$

on peut prendre $a = S^{-1/2} c$, où c est un vecteur propre normé associé à λ_1 la première valeur propre de $S^{-1/2}BS^{-1/2}$. De façon équivalente, de la deuxième formulation, on peut prendre a, un vecteur propre normé associé à λ_1 la première valeur propre de $S^{-1}B$. Notons que comme

$$S^{-1/2} B S^{-1/2} c = \lambda c \quad \text{et} \quad a = S^{-1/2} c,$$

alors

$$S^{-1/2} B a = \lambda S^{1/2} a \Rightarrow S^{-1} B a = \lambda a.$$

Les valeurs propres de $S^{-1}B$ et de $S^{-1/2}BS^{-1/2}$ sont donc les mêmes.

La fonction discriminante de Fisher est donc

$$f(x) = a^\top (x - \overline{X}),$$

où a est le vecteur propre normé associé à la plus grande valeur propre de $S^{-1}B$. Les scores $U_i = a^\top (X_i - \overline{X})$ sont les scores linéaires en X_i qui ont le rapport (variance inter) / (variance intra) le plus élevé. On peut aussi prendre $U_i = a^\top X_i$, car ajouter la même constante à toutes les observations $i = 1, \dots, n$ ne change rien.

— Pouvoir discriminant

Puisque la matrice $S^{-1/2}BS^{-1/2}$ est symétrique et définie positive, ses valeurs propres sont toutes réelles et positives. De plus, on a que $S^{-1}Ba - \lambda_1 a$. Ainsi,

$$Ba = \lambda_1 Sa \Rightarrow a^\top Ba = \lambda_1 a^\top Sa \Rightarrow \lambda_1 = \frac{a^\top Ba}{a^\top Sa}.$$

On a donc $0 \leq \lambda_1 \leq 1$. La valeur propre λ_1 peut donc être vue comme le pouvoir discriminant de f :

- $\lambda_1 = 1 \Rightarrow a^\top Ba = a^\top Sa$, donc 100% de la variabilité entre les groupes et 0 variabilité à l'intérieur des groupes.
- $\lambda_1 = 0 \Rightarrow a^\top Ba = 0$, donc 0 variabilité entre les gorupes et 100% de la variabilité à l'intérieur des groupes.

Règle de classification

- Score moyen des groupes : Après avoir défini la fonction discriminante $f(x)$, on peut calculer le score moyen de chaque groupe défini comme étant

$$m_k = a^\top \left(\overline{X}_{k1}, \dots, \overline{X}_{kp}\right)^\top,$$

  où $\overline{X}_{kj}$ est la moyenne de la je variable pour les individus appartenant au ke groupe.
- Stratégie de classement des individus. Considérons une nouvelle observations $X_0 \in \mathbb{R}^p$. Pour classer ce nouvel individu dans un groupe de la population, on calcule son score $f(X_0) = a^\top X_0$. Ensuite, on l'assigne au groupe k_0 qui lui ressemble le plus, c'est-à-dire le groupe tel que

$$\left|a^\top X_0 - m_{k_0}\right| = \min_{1<=k<=q} \left|a^\top X_0 - m_k\right|.$$

En applicant cette règle à l'échantillon $X_1, \dots, X_n$, on peut estimer les risques de mauvaise classification avec la matrice de confusion.

Cas particulier de la classification binaire

On peut montrer que le vectuer propre de l'analyse discriminante dans la cas où il n'y que deux populations peut être défini ainsi :

$$a = S^{-1}C = \sqrt{\frac{n_1 n_2}{n}} S^{-1}(\widetilde{X}_1 - \widetilde{X}_2),$$

où

$$C = \sqrt{\frac{n_1 n_2}{n}} (\tilde{x}_1 - \tilde{x}_2) \quad \text{et} \quad B = CC^\top.$$

et $\tilde{x}_i, i = 1, 2$ sont les moyennes des caractériques x dans chaque groupe.

Supposons que

$$m_1 = a^\top \tilde{x}_1 > a^\top \tilde{x}_2 = m_2.$$

Alors, on classe un individu dans le premier groupe si

$$a^\top x > \overline{m} = \frac{m_1 + m_2}{2} = a^\top \left(\frac{\tilde{x}_1 + \tilde{x}_2}{2} \right).$$

Ceci est équivalent à

$$(\tilde{x}_1 - \tilde{x}_2)^\top S^{-1} x > (\tilde{x}_1 - \tilde{x}_2)^\top S^{-1} \left(\frac{\tilde{x}_1 + \tilde{x}_2}{2} \right).$$

Example