Analyse discriminante

I Introduction

L'analyse discriminante est une ~~branche~~ méthode statistique qui consiste à construire une variable qualitative à k modalités (ou catégories) en utilisant p variables explicatives (quantitatives).

On considère un ensemble Ω (l'ensemble des individus) muni d'une partition a priori $P(P_1, \dots, P_k)$ en k classes. Chaque individu $\omega \in \Omega$ est caractérisé par un ensemble de p variable descriptives (ou prédicteurs). (var quanti). Le but général, de l'analyse discriminante est de séparer au mieux les k classes a priori sur Ω à partir des p prédicteurs.

Exp 1 : Analyse linguistique.

Ω = ensemble des députés élus en 1981. La partition P sur Ω comprend deux classes : les députés de droite et les députés de gauche. Les descripteurs sont les fréquences d'utilisation de 53 mots figurant dans les discours des députés. Le but de l'étude est de voir si les mots utilisés permettent de retrouver la séparation droite-gauche.

Il s'agit ici d'une étude descriptive. Le but n'est pas de découvrir la tendance de députés mais de faire une étude politico linguistique.

Exp 2 : Prévision de risque d'avalanche (Der Megredichian 1982)

Ω est l'ensemble des stations de ski des Alpes. La partition a priori comprend 3 classes : Avalanche à la station, couloir d'avalanche, et pas d'avalanche. Les descripteurs sont au nombre 274 et concernent des mes météorologiques.

Il s'agit d'anal discri à but prévisionnel : à partir de l'étude du passé le but est de déterminer chaque jour la station à haut risque d'avalanche. Le but de l'analyse prévisionnelle ou décisionnelle est de construire à partir d'un échantillon d'inds connus des règles de décisions stat. Ces règles serviront à affecter de le futur, les inds à une classe a priori et devront minimiser

les erreurs de prevision ou de diagnostique.
En resume : affecter un ind à une classe a priori au moyen d'une fonctions des vars explicatives avec une proba d'erreur minimale.
Les fonctions peuvent être explicite (lineaire, quadratique ou logique) ou implicité (estimation de densité).
Ds ce cours nous nous limitons à analyse factorielle discriminante, qui consiste à rechercher un ensemble de s ($s < p$) variables de classement (axes discriminants) qui soient des combinaisons linéaires des p vars initiales les plus discriminants possibles au sens d'un certains criteres à fixer.
Ce qui signifie que les valeurs des inds d'une même classe par ces vars discriminantes soient le plus concentré possibles et celles prises par les inds de classes $\neq$ sont les plus eloignée possibles.
Un critere naturel utilisant les données est d'examiner les variances intraclasse (dans la classe) et variances interclasses (entre classes)

Notations et definitions.
Soit Ω l'ensemble des individus et $P(P_1, \dots, P_k)$ la partition definie a priori sur Ω. On dispose d'un échantillon de taille n (échantillon d'apprentissage) d'elts de Ω. On designe par X la matrice (n, p) representant le tableau de données pour les p variables descriptibles.
$X = (x_{ij}, i=1,\dots,n ; j=1,\dots,p)$ où x_{ij} est la valeur prise par le $i^{ème}$ ind pour la $j^{ème}$ variable.
A chaque ind ω_i on associe un vecteur $\underline{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^p$ qui correspond à la $i^{ème}$ ligne de X. $\mathbb{R}^p$ est l'espace des inds.
Et à chaque var on associe un vecteur $\underline{x}^j = \begin{pmatrix} x^j_1 \\ \vdots \\ x^j_n \end{pmatrix} \in \mathbb{R}^n$ qui correspond à la $j^{ème}$ colonne du tableau. $\mathbb{R}^n$ est l'espace des vars.
$\underline{x}_i$ le $i^{ème}$ ind et $\underline{x}^j$ la $j^{ème}$ var.
On suppose que chaque ind est muni d'un poids p_i ($\forall p_i \leq 1$, $\sum_{i=1}^{n} p_i = 1$)
$D_p = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix}$ metrique de poids

Elements descriptifs de l'espace des variables.

a/ Moyenne $\bar{x}^j$ $\quad \bar{x}^j = \langle \underline{x}^j, \mathbb{1} \rangle_D$.

b) Variance $Var(\underline{x}^j) = \sigma_{x^j}^2 = \sum p_i (x_i^j - \bar{x}^j)^2 = \langle \underline{x}^j - \mathbb{1}\bar{x}^j \,;\, \underline{x}^j - \mathbb{1}\bar{x}^j \rangle_D$

c) ecart type $\sigma_j = \sqrt{Var(x^j)}$.

Var mesure la dispersion des autour de la moyenne.

d) covariance ; $cov(\underline{x}^j, \underline{x}^{j'}) = \sum_{i=1}^{n} p_i (x_i^j - \bar{x}^j)(x_i^{j'} - \bar{x}^{j'}) = \langle \underline{x}^j - \bar{x}^j \mathbb{1} \,,\, \underline{x}^{j'} - \bar{x}^{j'} \mathbb{1} \rangle_D$

Rq le tableau centré $\tilde{X} = (x_i^j - \bar{x}^j \,;\, i = 1, \dots, n \,,\, j = 1, \dots, p)$

Elements descriptifs de l'espace des inds.

a/ Le nuage des p^{ts} $\quad N = \{ (\underline{x}_i, p_i) \,/\, i = 1, \dots, n \;\; \underline{x}_i \in \mathbb{R}^p \}$

b/ Le centre de gravité. ou barycentre des w_i muni de poids p_i.

$$g = \sum_{i=1}^{n} p_i \underline{x}_i = \begin{pmatrix} \bar{x}^1 \\ \vdots \\ \bar{x}^p \end{pmatrix} \in \mathbb{R}^p.$$

Utilité de la variable qualitative.

La var quali correspond à la partition apriori $\underline{P} = (P_1, \dots, P_k)$ de centres de gravité resp $G = (g_1, \dots, g_k)$. et de matrice de var-covar $(W_1, \dots, W_k)$ W_ℓ designe la dispersion à l'intérieur de la classe P_ℓ

l'ensemble des realisations de ℓ est donné par

$$\Phi(\Omega) = \{ y_i^\ell \,,\, i = 1, \dots, n \,,\, \ell = 1, \dots, k \}$$

avec $y_i^\ell = \begin{cases} 1 & \text{si } w_i \in P_\ell \\ 0 & \text{sinon.} \end{cases}$

Poids des classes

Comme chaque ind est muni d'un poids p_i ; alors on attribue à chaque classe P_ℓ un poids $P_\ell = \sum_{w_i \in P_\ell} p_i$

Relation entre g et g_ℓ.

On a $\quad g = \sum_{\ell=1}^{k} P_\ell g_\ell \quad$ où $\quad g_k = \frac{1}{P_\ell} \sum_{w_i \in P_\ell} p_i \underline{x}_i$.

En effet :

$$\sum_{\ell=1}^{k} P_\ell g_k = \sum_{\ell=1}^{k} P_\ell \cdot \frac{1}{P_\ell} \sum_{w_i \in P_\ell} p_i \underline{x}_i = \sum_{\ell=1}^{k} \sum_{w_i \in P_\ell} p_i \underline{x}_i = \sum_{i=1}^{n} p_i \underline{x}_i = g$$

Matrice de variance covariance.

a/ Interieur des classes : Matrice var-covar intra-classe.

On note par $\underset{(p,p)}{W}$ (du mot anglais "within") la matrice var-covar intra-classe

$$W = \sum_{\ell=1}^{k} P_\ell W_\ell \quad \text{ou} \quad W_\ell = \frac{1}{P_\ell} \sum_{w_i \in P_\ell} p_i (x_i - g_\ell)^t (x_i - g_\ell)$$

W est la moyenne des matrice W_ℓ $\ell = 1, \dots, k$.

$$= \frac{1}{P_\ell} \sum_{\ell=1}^{k} \sum_{w_i \in P_\ell} p_i (x_i - g_\ell)^t (x_i - g_\ell)$$

b/ Entre les classes : Matrice de var-cov inter-classes.

On note par $\underset{(p,p)}{B}$ (du mot anglais "between") la matrice var-covar inter-classe

$$B = \sum_{\ell=1}^{k} P_\ell (g_\ell - g)^t (g_\ell - g)$$

Traduit l'eloignement entre les classes : B mesure la separation de classes.

Relation entre B et G.

$$G = \left(g_1^t, \dots, g_k^t \right)$$

On a $B = \underset{(p,k)}{G} \underset{(k,k)}{D_P} \underset{(k,p)}{G^t}$ $\qquad D_P = \begin{pmatrix} P_1 & & 0 \\ & \ddots & \\ 0 & & P_k \end{pmatrix}$

Matrice de variance-covariance totale

Soit $V = \left(\text{cov}(X^j, X^{j'}) \right)_{\substack{1 \le j \le p \\ 1 \le j' \le p}}$ la matrice de var covar.

$$= \sum_{i=1}^{n} p_i (x_i - g)^t (x_i - g)$$

La matrice de var-covar totale V mesure la dispersion totale de tous les inds

Relation entre V et X

Si les vars sont centrées $V = \sum_{i=1}^{n} p_i x_i^t x_i$ $\qquad g = 0_{1p}$.

$V = \underset{(p,n)}{X^t} \underset{(n,n)}{D_p} \underset{(n,p)}{X}$ (sinon on pose $\tilde{X}$ la matrice centrée) $\qquad D_p = \begin{pmatrix} p_1 & & \\ & \ddots & \\ & & p_n \end{pmatrix}$

Théorème de Huygens.

on a $V = W + B$.

Preuve. ~~On suppose $g = 0_{1p}$~~

$$V = \sum_{i=1}^{n} p_i (x_i - g)^t (x_i - g) = \sum_{\ell=1}^{k} \sum_{w_i \in P_\ell} p_i (x_i - g_k - g + g_k)^t (x_i - g_k - g + g_k)$$

$$= \sum_{\ell=1}^{k} \sum_{w_i \in P_\ell} p_i \left[(x_i - g_k) + (g_k - g) \right]^t \left[(x_i - g_k) + (g_k - g) \right]$$

$$V = \sum_{l=1}^{k} \sum_{w_i \in P_l} p_i (x_i - g_l)(x_i - g_l)' + \sum_{l=1}^{k} \sum_{w_i \in P_l} p_i (g_l - g)(x_i - g_l)' +$$

$$\sum_{l=1}^{k} \sum_{w_i \in P_l} p_i (g_l - g)(x_i - g_l)' + \sum_{l=1}^{k} \sum_{w_i \in P_l} p_i (g_l - g)(g_l - g)'$$

$$= W + \sum_{l=1}^{k} (g_l - g) \sum_{w_i \in P_l} p_i (x_i - g_l)' + \sum_{l=1}^{k} \left(\sum_{w_i \in P_l} p_i (x_i - g_l) \right)' (g_l - g) + B$$

$$= W + B + \sum_{l=1}^{k} (g_l - g) \left[\sum_{w_i \in P_l} p_i x_i' - P_l g_l' \right] + \sum_{l=1}^{k} \left[\sum_{w_i \in P_l} p_i x_i - P_l g_l \right]$$

avec $g_l = \frac{1}{P_l} \sum_{w_i \in P_l} p_i x_i$

D'où $V = W + B$

Commentaire : La décomposition d'Huygens produit un nouvel espace de représentation qui permet de distinguer au mieux les k classes en construisant une suite de vars discriminantes z^h (donc de nouveaux axes discriminants) tell que les vars z^h soient non corrélées 2 à 2 et la projection des inds d'une m^{e} classe sur les axes discriminant soient le plus proches possible. et ceux de classe $\neq$ soient les plus éloignés possible.

Les nouvelles vars z^h sont des combinaisons linéaire de vars initiales.

Pour des raisons de simplicité de notations nous supposons dans ce qui suit que les vars sont centrées $g = 0_{\mathbb{R}^p}$.

Variances et covariances des vars discriminantes.

Si les vars discriminantes z^h sont données par $z^h = \sum_{j=1}^{p} u_h^j x^j \in \mathbb{R}^n$ (combinaisons linéaire des vars initiales), alors

$$\text{Var}(z^h) = {}^t u_h V u_h \quad \text{avec} \quad u_h = \begin{pmatrix} u_h^1 \\ \vdots \\ u_h^p \end{pmatrix}$$

et $\text{cov}(z^h, z^l) = {}^t u_h V u_l = {}^t u_l V u_h$

Démonstration

$${}^t u_h V u_h = (u_h^1, \dots, u_h^p) \begin{pmatrix} \sum_{i=1}^{n} p_i (x_i^1 - 0) x_i^1 & \cdots & \sum_{i=1}^{n} p_i (x_i^1 - 0) x_i^p \\ \vdots & & \vdots \\ \sum_{i=1}^{n} p_i x_i^p x_i^1 & \cdots & \sum_{i=1}^{n} p_i x_i^p x_i^p \end{pmatrix} \begin{pmatrix} u_h^1 \\ \vdots \\ u_h^p \end{pmatrix}$$

$$= \left(\sum_{j=1}^{p} \left(\sum_{i=1}^{n} p_i x_i^1 x_i^j \right) u_h^j, \dots, \sum_{j=1}^{p} \left(\sum_{i=1}^{n} p_i x_i^p x_i^j \right) u_h^j \right) \begin{pmatrix} u_h^1 \\ \vdots \\ u_h^p \end{pmatrix}$$

$$= \sum_{j=1}^{p} u_h^j \left(\sum_{i=1}^{n} p_i x_i^1 x_i^j \right) u_h^1 + \cdots + \sum_{j=1}^{p} u_h^j \left(\sum_{i=1}^{n} p_i x_i^p x_i^j \right) u_h^p$$

$$= \sum_{i=1}^{n} p_i \left(\sum_{j=1}^{p} u_h^j x_i^j \right) x_i^1 u_h^1 + \cdots + \sum_{i=1}^{n} p_i \left(\sum_{j=1}^{p} u_h^j x_i^j \right) x_i^p u_h^p$$

$$= \sum_{i=1}^{n} p_i \left(\sum_{j=1}^{p} u_h^j x_i^j \right) \left(x_i^1 u_h^1 + \cdots + x_i^p u_h^p \right)$$

$$= \sum_{i=1}^{n} p_i \left(\sum_{j=1}^{p} u_h^j x_i^j \right)^2 = \sum_{i=1}^{n} p_i z_i^2 = \text{Var}(z^h)$$

$$\mathrm{Var}(z^h) = \sum_{i=1}^{n} p_i (z_i^h)^2 = \sum_{i=1}^{n} p_i \left(\sum_{j=1}^{p} u_h^j x_i^j \right)^2$$

$$= \sum_{i=1}^{n} p_i \, (u_h^1, \dots, u_h^p) \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^q \end{pmatrix} (x_i^1, \dots, x_i^p) \begin{pmatrix} u_h^1 \\ \vdots \\ u_h^p \end{pmatrix}$$

$$= (u_h^1, \dots, u_h^p) \sum_{i=1}^{n} p_i \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^q \end{pmatrix} (x_i^1, \dots, x_i^p) \begin{pmatrix} u_h^1 \\ \vdots \\ u_h^p \end{pmatrix}$$

$${}^t u_h V u_h. \qquad \underline{\text{Cqfd}}$$

Pour la cov

$$\mathrm{cov}(z^h, z^\ell) = \sum_{i=1}^{n} p_i z_i^h z_i^\ell = \sum_{i=1}^{n} p_i \left(\sum_{j=1}^{p} u_h^j x_i^j \right) \left(\sum_{j=1}^{p} u_\ell^j x_i^j \right)$$

$$= \sum_{i=1}^{n} p_i \left[(u_h^1, \dots, u_h^p) \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^q \end{pmatrix} \right] \left[(x_i^1, \dots, x_i^p) \begin{pmatrix} u_\ell^1 \\ \vdots \\ u_\ell^q \end{pmatrix} \right] \qquad (1,q)(q,q)(q,1)$$

$$= {}^t u_h \sum_{i=1}^{n} p_i \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^q \end{pmatrix} (x_i^1, \dots, x_i^p) u_\ell = {}^t u_h V u_\ell$$

Rq : comme on a $V = W + B$ (var totale = variance intra classe + var inter class)

alors ${}^t u_h V u_h = {}^t u_h W u_h + {}^t u_h B u_h$.

ce qui signifie que la variance totale de z^h est égale à la variance de z^h intra classe + variance de z^h inter classe.

Choix des variables discriminantes

c'est la détermination des axes discriminants.

On cherche les variables discriminantes z^h de variance totale $V(z^h)$ donnée dont la variance inter class B est maximale. (de manière à séparer au mieux les classes). Par la relation ${}^t u_h V u_h = {}^t u_h W u_h + u_h B u_h$.

$$\max_{u_h} {}^t u_h B u_h \iff \min_{u_h} {}^t u_h W u_h \qquad \text{(classes les + homogène minimum)}$$

On peut formuler le pb d'optimisation de plusieurs façons

Pb 1 : Trouver $u_h \in \mathbb{R}^p$ tq $\begin{cases} \dfrac{{}^t u_h B u_h}{{}^t u_h V u_h} \max \text{ sous la contrainte} \\ {}^t u_h V u_h = C^{te} \end{cases}$

La discrimination est alors impossible.

La statistique de decision est Λ de wilks

$$\Lambda = \frac{|W|}{|V|} \qquad |.| \text{ determinant.}$$

Sous H_0 Λ suit la loi de wilks de parametres $(n-1)$, p, $k-1$ notée $\Lambda(n-1, p, k-1)$

La region critique du test $RC = [\Lambda < \Lambda^{\alpha}_{tab}]$ Λ est tabulée.

au niveau α fixé.

Si H_0 est rejetter, il est possible de projeter les inds ds un nouvel espace factoriel permettant de distinguer au mieux les classes.

Variance expliquée

Le pouvoir discriminant d'un axe factoriel est evaluée par le rapport $\frac{\lambda}{\sum_{j=1}^{k-1} \lambda_j}$ inertie ou proportion de variance expliquée par l'axe.

L'inertie ou proportion de variance expliquée par le plan est obtenue par les deux premiers axes factoriels est : $\frac{\lambda_1 + \lambda_2}{\sum_{j=1}^{k-1} \lambda_j}$.

de maniere generale l'inertie explique par le s.epace factoriel de dim s

est $$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_s}{\sum_{j=1}^{k-1} \lambda_j} \qquad s \leq p-1$$

Rq!

La somme est jusqu'à $k-1$ car il y a au plus $(k-1)$ vps non nulles si $k \leq p$

En effet, la matrice B de variance_covariance intra_classe est la matrice de associé aux centres de gravité $g_1, \ldots, g_k$

Or on a $\sum_{k=1}^{k} p_k g_k = g$. d'où B est de rang $\leq k-1$

$\Rightarrow rg(V^{-1}B) \leq \min(p, k-1)$

si $k < p \Rightarrow k-1 < p \Rightarrow \min(p, k-1) = k-1$

$\Rightarrow rg(V^{-1}B) \leq (k-1)$