LES DIFFÉRENTES MÉTHODES DE DISCRIMINATION

Olivier SAUTORY

Les pages qui suivent constituent une version provisoire d'un manuel d'utilisation de l'analyse discriminante avec SAS. Certains de ses paragraphes sont rédigés de façon très concise, et de nombreuses démonstrations sont omises. Le lecteur désireux d'en savoir plus pourra se reporter aux divers ouvrages présentés dans la bibliographie, et en particulier au premier d'entre eux.

Présentation de l'analyse discriminante

1. Objectifs

Expliquer ou prédire une variable qualitative à q modalités (variable "expliquée") par un ensemble de p variables quantitatives ("explicatives"), appelées aussi prédicteurs. Chaque modalité de la variable qualitative définit une classe dans la population, constituée des individus prenant cette modalité : ces classes sont définies *a priori* (contrairement aux méthodes de classification), et l'objectif de l'analyse discriminante est de caractériser ces classes avec les prédicteurs.

Remarques : L'analyse discriminante peut donc être considérée comme une extension du problème de la régression, dans le cas où la variable expliquée est qualitative. On verra que l'on peut également effectuer une analyse discriminante lorsque les variables explicatives sont qualitatives.

2. Deux catégories de problèmes

L'analyse discriminante est un ensemble de méthodes que l'on classe généralement en deux catégories :

a) Discrimination à but descriptif

On cherche parmi les variables quantitatives, ou parmi les combinaisons linéaires de ces variables, celles qui permettent de séparer le mieux possible (i.e. de discriminer au mieux) les différentes classes.

b) Discrimination à but décisionnel

On dispose d'un nouvel individu, pour lequel on connaît les valeurs des variables quantitatives, mais pas la classe à laquelle il appartient : il s'agit d'affecter l'individu à l'une des classes, de le "classer".

La plupart des problèmes traités par l'analyse discriminante rentrent dans cette catégorie.

3. Exemples

a) Discrimination à but descriptif

Analyse historique (exemple traité par Benzecri)

Les données concernent les députés de la III^{e} République en 1881 : il s'agit de différencier les députés de droite de ceux de gauche par les fréquences d'utilisation dans leurs discours de 53 mots, tels que public, indépendance, autorité, menace, famille, Dieu, république, paix, programme ...

Anthropométrie (exemple traité par C.R. Rao)

Discrimination entre différentes castes ou sous-castes de l'Inde, fondée sur des données anthropométriques : taille, profondeur du nez, hauteur du nez ...

b) Discrimination à but décisionnel

Anthropologie (information perdue)

On dispose de données anthropométriques relatives à un échantillon de crânes d'hommes et de crânes de femmes, et on cherche à déterminer le sexe (inconnu) d'un individu dont on a retrouvé le crâne lors de fouilles.

Reconnaissance des formes

La lecture automatique, qui nécessite la reconnaissance des lettres de l'alphabet ou des dix chiffres, peut s'opérer de la façon suivante :

Chaque chiffre est écrit dans un carré que l'on quadrille, en 36 petits carrés par exemple : à chacun d'eux on attribue une valeur numérique, comme l'intensité du noir ; si l'on dispose d'un échantillon de chiffres écrits par un grand nombre d'individus, on pourra discriminer entre les 10 chiffres, d'après les 36 intensités de noir, et ainsi définir une règle d'affectation pour la lecture automatique d'un chiffre.

Prédiction

En médecine :

- on dispose de mesures cliniques, biologiques ... caractérisant des malades, atteints de la même maladie, avant une opération chirurgicale, et on discrimine, ensuite, entre guéris et non-guéris : on peut ainsi "prévoir" le résultat d'une opération pour un nouveau malade ;

- l'analyse discriminante permet aussi, en utilisant des mesures de laboratoire, d'aider à diagnostiquer une maladie, à condition de disposer d'un échantillon de patients dont on a observé l'évolution.

En finance :

- les banques sont intéressées à prévoir le comportement de demandeurs de crédits, en fonction de caractéristiques qui doivent discriminer entre les "bons" et les "mauvais" clients.

En météorologie :

- prévision des avalanches (par exemple), à partir de mesures liées à l'atmosphère et à la neige.

I. Généralités

I.1. Données

On dispose d'une population $\mathcal{P}$ de n individus ($1 \ldots i \ldots n$) munis de poids p_i positifs (égaux à $1/n$ en général), avec $\sum_i p_i = 1$. Pour chacun d'eux, on connaît :

- la valeur d'une variable qualitative γ de modalités $Y_1 \ldots Y_k \ldots Y_q$. γ définit une partition de $\mathcal{P}$ en q classes $\mathcal{P}_1 \ldots \mathcal{P}_k \ldots \mathcal{P}_q$, avec :

$$\mathcal{P}_k = \{ i \in \mathcal{P},\ \gamma(i) = Y_k \}$$

- les valeurs de p variables numériques $X^1 \ldots X^j \ldots X^p$: $X^j(i) = x_i^j \in \mathrm{R}$.

On note :

$$\underset{(n,p)}{X} = \begin{pmatrix} x_1^1 & \ldots & x_1^p \\ & x_i^j & \\ x_n^1 & \ldots & x_n^p \end{pmatrix}$$

$$x_i = \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^j \\ \vdots \\ x_i^p \end{pmatrix} \in \mathrm{R}^p \qquad x^j = \begin{pmatrix} x_1^j \\ \vdots \\ x_i^j \\ \vdots \\ x_n^j \end{pmatrix} = \in \mathrm{R}^n$$

individu i variable X^j

I.2. Nuages de points

I.2.1. Nuage associé à $\mathcal{P}$

$N = \{ (x_i , p_i) , i \in \mathcal{P} \}$

Centre de gravité : $g = \sum_i p_i \ x_i = \begin{pmatrix} \bar{x}^1 \\ \cdot \\ \cdot \\ \bar{x}^j \\ \cdot \\ \cdot \\ \bar{x}^p \end{pmatrix}$

où $\bar{x}^j = \sum_i p_i \ x_i^j$ = moyenne de X^j.

On supposera le nuage centré : $g = 0$

I.2.2. Nuages associés aux $\mathcal{P}_k$

$N_k = \left\{ (x_i , \frac{p_i}{P_k}) , i \in \mathcal{P}_k \right\}$ avec $P_k = \sum_{i \in \mathcal{P}_k} p_i$ = poids de la classe $\mathcal{P}_k$

$$= \frac{n_k}{n} \ si \ p_i = \frac{1}{n} \ (n_k = \text{card } \mathcal{P}_k)$$

Centre de gravité : $g_k = \sum_{i \in \mathcal{P}_k} \frac{p_i}{P_k} x_i$

I.2.3. Nuage des centres de gravité

$N^{(g)} = \{ (g_k, P_k), k = 1 \ \ldots \ q \}$, de centre de gravité : $\sum_k P_k \ g_k = g = 0$

I.3. Matrices de variance

I.3.1. Matrice de variance totale

$$\underset{(p,p)}{V} = \begin{pmatrix} \cdot \\ \cdot \\ \cdot \\ \dots v_{jj'} \dots \\ \cdot \\ \cdot \\ \cdot \end{pmatrix} \quad \text{où } v_{jj'} = \text{Cov}\,(X^j, X^{j'}) = \sum_{i=1}^{n} p_i \,(x_i^j - \bar{x}^j)\,(x_i^{j'} - \bar{x}^{j'})$$

$$V = \sum_{i=1}^{n} p_i \; x_i \; {}^t x_i = {}^t X \, D \, X$$

$$\text{où } \underset{(n,n)}{D} = \begin{pmatrix} p_1 & & & \\ & \cdot & & \\ & & p_i & \\ & & & \cdot \\ & & & & p_n \end{pmatrix}$$

I.3.2. Matrice de variance intraclasse

Matrice de variance dans $\mathcal{P}_k$: $W_k = \sum_{i \in \mathcal{P}_k} \frac{p_i}{P_k} \,(x_i - g_k)\,{}^t(x_i - g_k)$ de terme général :

$$\sum_{i \in \mathcal{P}_k} \frac{p_i}{P_k} \,(x_i^j - \bar{x}_k^j)\,(x_i^{j'} - \bar{x}_k^{j'}) \quad \text{où } \bar{x}_k^j = \text{moyenne de } X^j \text{ dans } \mathcal{P}_k$$

Matrice de variance intraclasse :

$$W = \sum_{k=1}^{q} P_k \; W_k = \sum_{k=1}^{q} \sum_{i \in \mathcal{P}_k} p_i \,(x_i - g_k)\,{}^t(x_i - g_k)$$

I.3.3. Matrice de variance interclasse

C'est la matrice de variance de $X^1 \ldots X^p$ dans $N^{(g)}$:

$$B = \sum_{k=1}^{q} P_k \; g_k \; {}^t g_k$$

de terme général : $\sum_{k=1}^{q} P_k \quad \bar{x}_k^j \quad \bar{x}_k^{j'}$

I.3.4. Relation "de Huyghens"

$$V = W + B$$

II. Analyse factorielle discriminante

II.1. Variables discriminantes

II.1.1. Première variable discriminante

Problème : déterminer une variable $C = \sum_{j=1}^{p} a_j \; X^j$ qui discrimine au mieux les q classes $\mathcal{P}_1 \ldots \mathcal{P}_q$

On a la relation :

Variance totale = variance intraclasse + variance interclasse

→ Choisir C qui $\begin{cases}\text{maximise la variance interclasse}\\ \text{minimise la variance intraclasse}\end{cases}$

sous une contrainte de normalisation, par exemple : variance totale = 1.

Solution : C est représentée dans R^n par le vecteur

$$c = \sum_{j=1}^{p} a_j \quad x^j = X\ a = \begin{pmatrix} c_1 \\ \cdot \\ \cdot \\ \cdot \\ \\ \cdot \\ \cdot \\ \cdot \\ c_n \end{pmatrix}, \text{ avec } a = \begin{pmatrix} a_1 \\ \cdot \\ \cdot \\ \cdot \\ a_j \\ \cdot \\ \cdot \\ \cdot \\ a_p \end{pmatrix}$$

. $V_{tot}(C) = \sum_{i=1}^{n} p_i (c_i)^2 = {}^t c\, D\, c = {}^t(X\,a)\ D\ (X\,a) = {}^t a\, V\, a$

. $V_{intra}(C) = \sum_{k=1}^{q} P_k \left(\sum_{i \in \mathcal{P}_k} \frac{p_i}{P_k} \left(c_i - \bar{c}^k \right)^2 \right) = {}^t\, a\ W\ a$

($\bar{c}^k$ = moyenne de C dans $\mathcal{P}_k$)

. $V_{inter}(C) = \sum_{k=1}^{q} P_k\ (\bar{c}^k)^2 = {}^t\, a\ B\ a$

On maximise ${}^t A\ B\ A$ sous la contrainte ${}^t\, a\ V\ A = 1$

→ $a = a_1$: vecteur propre de $V^{-1}\ B$ associé à la plus grande valeur propre λ_1.

$c^1 = X\, a_1$ = première variable discriminante

a_1 = premier facteur discriminant

$\lambda_1 = {}^t\, a_1\ B\ a_1 = V_{inter}(C^1)$ = pouvoir discriminant de C^1.

* $\lambda_1 = 1 \Leftrightarrow V_{intra}(C^1) = 0 \Leftrightarrow C^1$ constante dans chaque classe
* $\lambda_1 = 0 \Leftrightarrow V_{inter}(C^1) = 0 \Leftrightarrow$ les moyennes $\bar{c}^{1k}$ sont toutes égales.

II.1.2. Autres variables discriminantes

- On cherche $c = X\,a$, non corrélée avec C^1, de variance égale à 1, qui maximise ${}^t a\,B\,a$ $\rightarrow C^2 = X\,a_2$, $a_2 =$ vecteur propre de V^{-1} B associé à la 2^e plus grande valeur propre λ_2, et on a ${}^t a_2\,B\,a_2 = \lambda_2$, etc.

- On détermine au total r variables discriminantes, où $r = rg\,(V^{-1}\,B) = rg\,B = q - 1$ en général (si $q < p < n$).

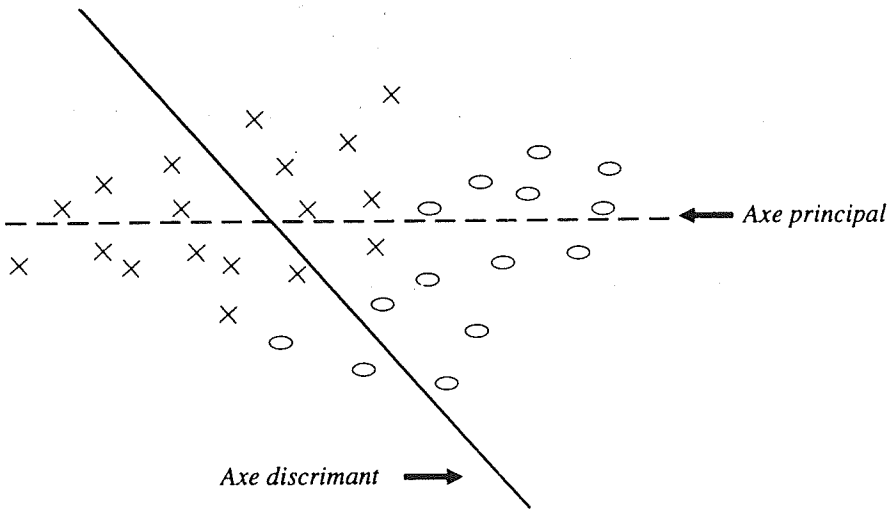
II.2. Axes discriminants

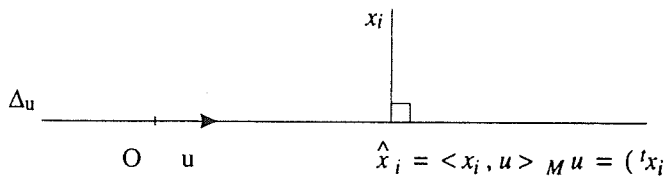
On suppose R^p muni d'une métrique M

II.2.1. Premier axe discriminant

Problème : déterminer un axe tel que lorsque l'on projette sur cet axe le nuage N des individus, les nuages N_k soient aussi séparés que possible, chacun des N_k étant le plus groupé possible.

Ce problème est différent de celui de l'analyse en composantes principales, où l'on cherche l'axe tel que les n individus de N soient en projection sur cet axe le plus dispersés possible.





u M-normé

$({}^t u M u = 1)$

Inertie totale du nuage sur Δ_u :

$$I_{tot} = \sum_i p_i \quad ({}^t x_i \; M \; u)^2 = {}^t u M \; V \; M \; u$$

Inertie interclasse sur Δ_u :

$$I_{inter} = \sum_k \frac{p_i}{P_k} \quad ({}^t g_k \; M \; u)^2 = {}^t u M \; B \; M \; u$$

Inertie intraclasse sur Δ_u :

$$I_{intra} = \sum_k P_k \sum_{i \in P_k} \frac{p_i}{P_k} \left[{}^t (x_i - g_k) \; M u \right]^2 = {}^t u M \; W \; M \; u$$

On cherche à maximiser :

$$\frac{I_{inter}}{I_{tot}} = \frac{{}^t u M B M u}{{}^t u M V M u}$$

$\rightarrow$ $u = u_1$ vecteur propre de $(M V M)^{-1} \; M B M = M^{-1} \; V^{-1} B M$ associé à la plus grande valeur propre λ_1.

$\Leftrightarrow$ $M u_1$ vecteur propre de $V^{-1} B$ associé à la plus grande valeur propre λ_1.

$\rightarrow$ $M u_1 = a_1$: on retrouve le premier facteur discriminant.

Le vecteur des coordonnées des x_i sur le premier axe discriminant Δ_u est :

$$\begin{pmatrix} {}^t x_1 \; M \; u_1 \\ \vdots \\ {}^t x_n \; M \; u_1 \end{pmatrix} = \begin{pmatrix} {}^t x_1 \; a_1 \\ \vdots \\ {}^t x_n \; a_1 \end{pmatrix} = X \; a_1 = c^1$$

→ on retrouve les valeurs de la première variable discriminante.

Ces coordonnées ne dépendent pas de M. On prendra par exemple $M = V^{-1}$, ou $M = W^{-1}$ (métrique de Mahalanobis).

Avec $M = V^{-1}$: $\begin{cases} B \; V^{-1} \; u_1 = \lambda_1 \; u_1 \\ V^{-1} \; B \; a_1 = \lambda_1 \; a_1 \end{cases}$

Remarque : $V^{-1} \; B$ et $W^{-1} \; B$ ont les mêmes vecteurs propres

$$W^{-1} \; B \; a = \mu \; a \Leftrightarrow B \; a = \mu \; W \; a = \mu \; (V - B) \; a$$

$$\Leftrightarrow V^{-1} \; B \; a = \frac{\mu}{\mu + 1} \; a = \lambda a$$

$$\lambda = \frac{\mu}{\mu + 1} \qquad \mu = \frac{\lambda}{\lambda - 1}$$

II.2.2. Autres axes discriminants

- On cherche u M-normé, M-orthogonal à u_1, tel qu'en projection sur Δ_u les nuages N_k soient aussi séparés que possible.

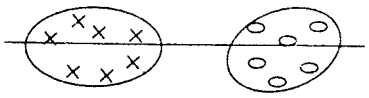
→ $u = u_2$ vecteur propre de $M^{-1} \; V^{-1} \; B \; M$ associé à la deuxième plus grande valeur propre λ_2

→ $M \; u_2 = a_2$, $c^2 = X \; a_2 = X \; M \; u_2 =$ deuxième variable discriminante

- On trouve q-1 axes discriminants associés à des valeurs propres non nulles.

Remarque : $\lambda_1 = \dfrac{I_{inter}}{I_{tot}}$ (sur Δ_{u_1})

On peut avoir discrimination parfaite avec $\lambda < 1$ (graphique)



II.2.3. ACP du nuage des centres de gravité

Les axes discriminants sont les axes principaux issus de l'analyse en composantes principales du nuage $N^{(g)}$ des centres de gravité, avec la métrique V^{-1}.

II.2.4. Cas de 2 groupes

Il n'y a qu'une seule variable discriminante, un seul axe discriminant = droite reliant les deux centres de gravité g_1 et g_2.

$u = g_1 - g_2$ (à une constante multiplicative près)

$a = V^{-1} (g_1 - g_2)$

ou $W^{-1} (g_1 - g_2)$ = fonction de Fisher

$\lambda = P_1 P_2 {}^t(g_1 - g_2) V^{-1} (g_1 - g_2)$

$\mu = P_1 P_2 {}^t(g_1 - g_2) W^{-1} (g_1 - g_2)$

II.3. Règles d'affectation

Dans une optique "prédictive", on peut définir des règles géométriques d'affectation d'un individu représenté par $x \in \mathrm{R}^p$, à l'une des q classes.

II.3.1. Règle de Mahalanobis-Fisher

On utilise la métrique W^{-1} (ou V^{-1}) :

$$d^2_{W^{-1}}(x, g_k) = {}^t(x-g_k) W^{-1} (x-g_k) = {}^t x W^{-1} x + {}^t g_k W^{-1} g_k - 2\, {}^t x W^{-1} g_k$$

→ on cherche le maximum de :

$$ {}^t x\, W^{-1}\, g_k - \frac{1}{2}\, {}^t g_k\, W^{-1}\, g_k $$

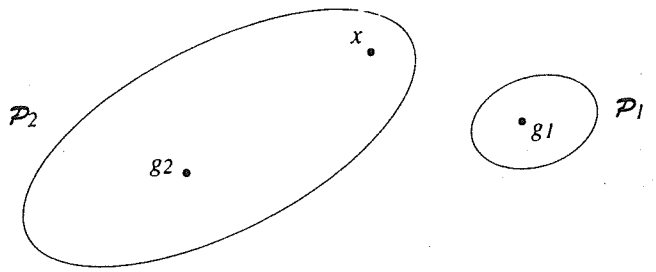
Cas de 2 classes :

On affecte x à la classe 1 si :

$$ {}^t x\, W^{-1}\, (g_1 - g_2) > \frac{1}{2}\, {}^t (g_1 + g_2)\, W^{-1} (g_1 - g_2) $$

${}^t x\, W^{-1}\, (g_1 - g_2)$ est la valeur de la fonction de Fisher au point x.

II.3.2. Métriques locales



Il semble plus naturel d'affecter x à la classe $\mathcal{P}_2$ qu'à la classe $\mathcal{P}_1$, bien que x soit plus proche de g_1 que de g_2

→ utiliser des métriques locales tenant compte de la dispersion des classes.

Par exemple : M_k proportionnelle à W_k^{-1}

→ on minimise ${}^t(x - g_k)\, M_k\, (x - g_k)$

III. Analyse canonique discriminante

III.1. Analyse canonique

2 paquets de variables numériques mesurées sur n individus

$$\underset{n}{\overset{1}{}} [\underset{p}{X_1} \mid \underset{q}{X_2}]$$

$Z_1 = \{\xi \in R^n , \xi = X_1 a, a \in R^p\}$

(combinaisons linéaires des colonnes de X_1)

$Z_2 = \{\eta \in R^n , \eta = X_2 b, b \in R^q\}$

Problème :

- déterminer $\xi_1 \in Z_1$ et $\eta_1 \in Z_2$ telles que $Cor^2 (\xi_1, \eta_1)$ maximal

- déterminer $\xi_2 \in Z_1$ non corrélé avec ξ_1, et $\eta_2 \in Z_2$, non corrélé avec η_1,telles que $Cor^2 (\xi_2, \eta_2)$ max. etc .

(ξ_1, η_1) , (ξ_2, η_2) . . . = variables canoniques.

Solution (dans R^p et R^q)

$V_{11} = {}^t X_1 D X_1$ $\qquad$ $V_{12} = {}^t X_1 D X_2$

$V_{21} = {}^t X_2 D X_1$ $\qquad$ $V_{22} = {}^t X_2 D X_2$

où $D = \begin{pmatrix} p_1 & & \\ & p_i & \\ & & p_n \end{pmatrix}$

$$\begin{cases} V_{11}^{-1} \; V_{12} \; V_{22}^{-1} \; V_{21} \; a_h = \lambda_h \; a_h \\ V_{22}^{-1} \; V_{21} \; V_{11}^{-1} \; V_{12} \; b_h = \lambda_h \; b_h \end{cases}$$

et $\lambda_h = \text{Cor}^2 \; (\xi_h, \; \eta_h)$

où $\begin{cases} \xi_h = X_1 \; a_h \\ \eta_h = X_2 \; b_h \end{cases}$ (a_h, b_h) = facteurs canoniques.

Normalisation : ${}^t a_h \, V_{11} \, a_h = 1$
${}^t b_h \, V_{22} \, b_h = 1$

→ variables canoniques de variance 1.

III.2. L'analyse discriminante est une analyse canonique particulière

$$\left[\begin{array}{c|c} \begin{matrix} 1 \ldots k \ldots q \\ 0 \ldots 1 \ldots 0 \\ 1 \ldots 0 \ldots 0 \end{matrix} & \\ A & X \end{array}\right]$$

indicatrices associées à y — p variables numériques

L'équation déterminant les facteurs canoniques s'écrit (pour $a \in R^p$) :

$$\underset{V^{-1}}{({}^tXDX)^{-1}} \; ({}^tXDA) \; ({}^tADA)^{-1} \; ({}^tADX) \, a = \lambda a$$

Or $B = {}^tG \; D_p \, G$

avec $\begin{cases} G = \begin{bmatrix} \ldots \\ t_{g_k} \\ \ldots \end{bmatrix} = ({}^tADA)^{-1} \; {}^tADX \\ D_p = \begin{bmatrix} P_1 & & \\ & \ddots & \\ & & P_q \end{bmatrix} = {}^tADA \end{cases}$

$\rightarrow B = ({}^t XDA)\,({}^t ADA)^{-1}\,({}^t ADX)$

$\rightarrow V^{-1} B_a = \lambda_a$

facteurs canoniques = facteurs discriminants

IV. Méthodes probabilistes

On considère que les individus sont issus d'une population Ω ; la variable qualitative $\mathcal{Y}$ définit q classes $\Omega_1 \ldots \Omega_k \ldots \Omega_q$ dans Ω.

Règle de décision, ou règle d'affectation

Connaissant le vecteur $x = {}^t(x^1 \ldots x^p)$ des valeurs de $X^1 \ldots X^p$ prises par un individu ω de Ω, on cherche à affecter cet individu à l'une des classes Ω_k. Ce vecteur x sera parfois noté $x(\omega)$.

Une **règle de décision** est une application δ de R^p dans $\{1 \ldots k \ldots q\}$ telle que :

$$\delta\ (x) = k \Leftrightarrow \text{on affecte } \omega \text{ à la classe } \Omega_k .$$

La donnée de δ est équivalente à celle d'une partition de R^p en q classes $A_1 \ldots A_q$ telles que :

$$\delta(x) = k \Leftrightarrow x \in A_k$$

Si la connaissance de x (ω) contient de l'information quant à l'appartenance de ω à l'une des classes Ω_k (ce qui est le cas si les variables $X^1 \ldots X^p$ sont bien choisies), la règle d'affectation doit l'utiliser au mieux, i.e. de telle façon que l'affirmation "$\omega \in \Omega_k$" soit "le plus souvent" exacte.

IV.1. Règle bayésienne

On pose : $p_k = P(\omega \in \Omega_k)$ $\qquad (\sum_{k=1}^{q} p_k = 1)$.

"P" signifie ici "probabilité", ce qui suppose que Ω est probabilisé, et que les Ω_k appartiennent à la tribu associée à Ω . En général, Ω est fini, la tribu associée est l'ensemble des parties de Ω et P est une proportion :

$$p_k = \frac{card\,\Omega_k}{card\,\Omega}$$

On suppose que les variables quantitatives $X^1 \ldots X^p$ sont des variables aléatoires réelles définies sur Ω ; $X = (X^1 \ldots X^p)$ est un vecteur aléatoire à valeurs dans R^p .

On note $f_k\ (x)\ \ (1 \leq k \leq q)$ la densité de probabilité du vecteur X défini sur la sous-population Ω_k , et $f\,(x)$ la densité de probabilité de X sur Ω :

$$f(x) = \sum_{k=1}^{q} p_k\, f_k\ (x)$$

($f_k(x)$ est la densité conditionnelle de $x(\omega)$ sachant $\omega \in \Omega_k$).

On choisit des coûts d'erreur de décision (a priori) :

C_{lk} = coût d'affectation de ω à Ω_l alors qu'en réalité $\omega \in \Omega_k$

($C_{lk} \geq 0$, avec $C_{kk} = 0,\ \forall\ k = 1 \ldots q$)

On pose :

C_k = coût moyen d'affectation lorsque l'individu $\omega \in \Omega_k$

$$= \sum_{l=1}^{q} C_{lk}\ P(\ x \in A_l \mid \omega \in \Omega_k\ ) = \sum_{l=1}^{q} C_{lk} \int_{R^p} 1_{A_l}(x) f_k\ (x)\ dx$$

$$C = \text{coût moyen} = \sum_{k=1}^{q} C_k\ P(\ \omega \in \Omega_k\ ) = \sum_{k=1}^{q} p_k\ C_k$$

La règle de décision bayésienne consiste à déterminer les q régions $A_1 \ldots A_k \ldots A_q$ de R^p telles que le coût moyen C soit minimum.

$$C = \sum_{k=1}^{q} p_k \left(\sum_{l=1}^{q} C_{lk} \int_{R^p} 1_{A_l}(x) f_k(x)\, dx \right)$$

$$= \int_{R^p} \sum_{l=1}^{q} \left(\sum_{k=1}^{q} C_{lk}\, p_k\, f_k(x) \right) 1_{A_l}(x)\, dx$$

En tout point x de R^p, une seule des valeurs $1_{A_l}(x)$ $(1 \le l \le q)$ n'est pas nulle ; minimiser C revient donc à choisir pour valeur non nulle la valeur $1_{A_j}(x)$ telle que $\sum_{k=1}^{q} C_{jk}\, p_k\, f_k(x)$ soit le minimum des $\sum_{k=1}^{q} C_{lk}\, p_k\, f_k(x)$ $(1 \le l \le q)$

La quantité :

$$\frac{\sum_{k=1}^{q} C_{lk}\, p_k\, f_k(x)}{\sum_{k=1}^{q} p_k\, f_k(x)} = C_m(l)$$

est **le coût moyen *a posteriori* (i.e. sachant x (ω)) d'affectation** de ω à Ω_l.

La règle bayésienne consiste donc à affecter ω à la classe Ω_j qui minimise le coût moyen *a posteriori* d'affectation.

Cas où tous les coûts *a priori* sont égaux

Si $C_{kl} = 1$ si $k \neq l$, et $C_{kk} = 0$, alors :

$$C_m(l) = \frac{\sum_{k \neq l} p_k\, f_k(x)}{\sum_{k=1}^{q} p_k\, f_k(x)} = 1 - \frac{p_l\, f_l(x)}{\sum_{k=1}^{q} p_k\, f_k(x)}$$

On affecte ω à la classe Ωj qui maximise la quantité

$$\frac{p_l f_l(x)}{\sum_{k=1}^{q} p_k f_k(x)}$$

appelée **probabilité *a posteriori*** (i.e. sachant $x(\omega)$) d'appartenance de ω à Ω_l. Il est bien sûr équivalent de maximiser $p_l\ f_l\ (x)$.

IV.2. Le modèle normal multidimensionnel

On suppose que chaque loi de probabilité f_k est une loi normale de dimension p, $N(\mu_k, \Sigma_k)$, où μ_k est le vecteur-moyenne, et Σ_k la matrice des variances-covariances :

$$f_k(x) = \frac{1}{(2\Pi)^{p/2}\ (\det \sum_k)^{1/2}} \exp\left(-\frac{1}{2}\ {}^t(x-\mu_k) \sum_k^{-1} (x-\mu_k)\right)$$

La règle bayésienne (dans le cas où les coûts a priori sont égaux), revient à maximiser, en k, la quantité :

$$p_k\ f_k(x) \quad \text{ou} \quad \log(p_k\ f_k(x))$$

Or :

$$\log(p_k\ f_k\ (x)) = \log p_k - \log\left((2\Pi)^{p/2}\ (\det \sum_k)^{1/2}\right) - \frac{1}{2}\ {}^t(x-\mu_k) \sum_k^{-1} (x-\mu_k)$$

Il faut donc minimiser :

$${}^t(x-\mu_k) \sum_k^{-1} (x-\mu_k) - 2 \log p_k + \log(\det \Sigma_k)$$

La règle d'affectation est donc une règle **quadratique** : il faut comparer q fonctions quadratiques du vecteur x.

En général, on estime :

- μ_k par le vecteur des moyennes empiriques, noté g_k.
- Σ_k par la matrice des variances-covariances empiriques, notée V_k, (ou par $\frac{n}{n-1} V_k$)
- P_k par la fréquence empirique $\frac{n_k}{n}$

Cas d'égalité des matrices de variances-covariances

On suppose $\Sigma_1 = \ldots = \Sigma_k = \ldots = \Sigma_q = \Sigma$.

Il faut minimiser :

$$^t x \ \Sigma^{-1} x - 2 \ ^t \mu_k \ \Sigma^{-1} x + \mu_k \ \Sigma^{-1} \mu_k - 2 \log p_k + \log (\det \Sigma)$$

i.e. maximiser :

$$^t \mu_k \ \Sigma^{-1} x - \frac{1}{2} \mu_k \ \Sigma^{-1} \mu_k + \log p_k$$

La règle d'affectation devient donc **linéaire**.

En particulier, si on estime Σ par $\frac{n}{n-k} W$, et si les probalités *a priori* p_k sont égales, on retrouve la règle géométrique vue précédemment.

Cas où q = 2, et $\Sigma_1 = \Sigma_2$

On affecte x au groupe 1 si :

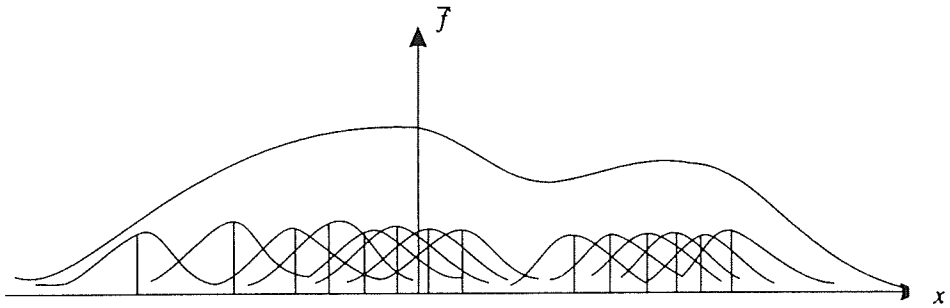
$$^t x \ \Sigma^{-1} (\mu_1 - \mu_2) > \frac{1}{2} \ ^t (\mu_1 + \mu_2) \ \Sigma^{-1} (\mu_1 - \mu_2) + \log \frac{p_2}{p_1}$$

Si $p_1 = p_2$, on retrouve la règle de Fisher.

IV.3. Méthodes non-paramètriques

IV.3.1. Méthode des noyaux (Parzen)

Estimation non paramètrique des densités $f_k(x)$, fondée sur des variantes multidimensionnelles de la méthode du noyau.



Approximation d'une densité à l'aide de noyau de Gauss

IV.3.2. Méthode des K plus proches voisins

On cherche les K points de l'échantillon $\mathcal{P}$ les plus proches du point x à affecter, au sens d'une certaine métrique, et on affecte x à la classe majoritaire parmi ces points.

V. Estimation des taux d'erreur de classement

Si la règle de décision est issue d'un modèle paramétrique (ex : discrimination linéaire), on peut faire une estimation paramétrique des probabilités d'erreur de classement ... mais peu robuste.

→ méthodes d'estimation non paramétriques.

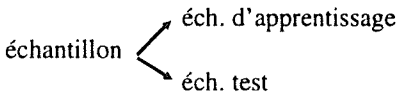
V.1. Resubstitution

On applique la règle de décision à l'échantillon "d'apprentissage" (données ayant servi à l'élaboration de la règle) : on observe le % de "biens classés", de "mal-classés" correspondant aux différentes erreurs possibles

→ estimateurs trop optimistes, sous évaluation des taux d'erreur (surtout si l'échantillon est petit).

V.2. Échantillon-test

Par tirage aléatoire :



Les taux d'erreur sont estimés sur l'échantillon-test (mais la règle de décision est estimée de manière moins précise).

V.3. Validation croisée

On retire successivement chacune des observations de l'échantillon, on estime la règle de décision sur le nouvel échantillon de taille n-1, et on affecte l'observation retirée selon cette règle.

Remarque : la technique du bootstrap peut être utilisée pour l'estimation du biais des taux d'erreur "apparents" (resubstitution).

VI. Sélection de variables

VI.1. Principe

Lorsque les variables descriptives X^1 ... X^p sont trop nombreuses :

- par rapport à la taille de l'échantillon
- pour qu'il n'y ait pas de risque que certaines soient fortement corrélées entre elles
- pour être toutes "discriminantes"
- pour la puissance de calcul disponible

il est nécessaire de sélectionner au préalable les "meilleures" variables pour obtenir une discrimination fiable, à l'aide d'un critère C mesurant la qualité de la discrimination.

L'examen de tous les m-uples parmi p variables est impossible dès que p est grand

→ procédure de sélection

PAS À PAS

ascendantes

. on sélectionne la meilleure variable $X^{(1)}$

. on sélectionne la variable $X^{(2)}$ qui, couplée avec $X^{(1)}$, fournit la meilleure valeur de C.

. on sélectionne la variable $X^{(3)}$ qui, jointe à $X^{(1)}$ et $X^{(2)}$, forme le meilleur triplet ...

On s'arrête lorsque l'adjonction d'une nouvelle variable n'améliore plus significativement C.

ascendantes avec remise en cause

. possibilité à chaque étape d'éliminer des variables déjà introduites si elles perdent leur "pouvoir discriminant" à cause de la sélection de la nouvelle variable.

descendantes

On part de toutes les variables, et à chaque étape on élimine la variable qui produit la plus faible détérioration "non significative" du critère.

VI.2. Critère du Λ de Wilks

VI.2.1. Lambda de Wilks

$$\Lambda = \frac{|W|}{|V|} = \frac{|W|}{|W+B|} = \frac{1}{|I+W^{-1}B|} = \prod_{j=1}^{q-1} \frac{1}{1+\theta_j} \quad \in [0,1]$$

où θ_j = valeurs propres non nulles de $W^{-1}B$.

Si $q = 2$

$$\Lambda = \frac{1}{1+\theta} = \frac{1}{1+\dfrac{n_1 n_2}{n\,(n_1+n_2-2)}\,D^2}$$

où $D^2 = \dfrac{n_1+n_2-2}{n_1+n_2}\,{}^t(g_1 - g_2)\,W^{-1}\,(g_1 - g_2)$

= distance de Mahalanobis entre les 2 classes.

VI.2.2. Critère (C1)

A l'étape $r + 1$:

- variables sélectionnées $X^1 \ldots X^r$

- variable candidate $Y \in \{ X^{r+1} \ldots X^p \}$

Matrice de variance totale de $X^1 \ldots X^r, Y$:

$$\underset{(r+1,\,r+1)}{V} = \begin{pmatrix} V_{11} & V_{12} \\ (r,r) & (r,1) \\ V_{21} & V_{22} \\ (1,r) & (1,1) \end{pmatrix}$$ où V_{11} = matrice de variance de $X^1 \ldots X^r$

Matrice de variance intra de $X^1 \ldots X^r, Y$:

$$\begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}$$

$$\Lambda\ (X^1 \ldots X^r, Y) = \frac{|W|}{|V|} = \frac{|W_{11}|\,(W_{22} - W_{21}\,W_{11}^{-1}\,W_{12})}{|V_{11}|\,(V_{22} - V_{21}\,V_{11}^{-1}\,V_{12})}$$

$$= \Lambda\ (X^1 \ldots X^r)\ \frac{W_{22} - W_{21}\,W_{11}^{-1}\,W_{12}}{V_{22} - V_{21}\,V_{11}^{-1}\,V_{12}} \leq \Lambda\ (X^1 \ldots X^r)$$

. Critère de sélection :

introduire la variable qui minimise $\Lambda\ (X^1 \ldots X^r, Y)$

. Critère d'arrêt :

$\frac{\Lambda_{r+1}}{\Lambda_r}$ pas significativement inférieur à 1.

$$F = \frac{n - q - r}{q - 1} \left(\frac{\Lambda_r}{\Lambda_{r+1}} - 1 \right) \rightarrow F(q\text{-}1,\ n\text{-}q\text{-}r)$$

. Critère d'élimination : idem.

Dans le cas de 2 classes :

minimiser Ω <===> maximiser D^2

VI.3. Critère de SAS

VI.3.1. Modèle d'analyse de la covariance

$$Y = \sum_{k=1}^{q} \alpha_k \ 1_k \qquad + \qquad \sum_{j=1}^{r} \beta_j \ X^j + \varepsilon$$

effet discriminant propre liaison avec les variables déjà sélectionnées

Test H_0 : $(\alpha_1 = \ldots = \alpha_k = \ldots \alpha_q)$

$$F = \frac{(SCR_0 - SCR_1) / (q-1)}{SCR_1 / (n-q-r)}$$

$\rightarrow F(q\text{-}1,\ n\text{-}q\text{-}r)$

. Critère de sélection (C_2) :

introduire la variable qui maximise F.

. Critère d'arrêt : F maximal non significatif.

. Critère d'élimination : idem.

VI.3.2. Équivalence avec (C_1)

$SCR_0 = T_{22} - T_{21}\ T_{11}^{-1}\ T_{12}$

$SCR_1 = W_{22} - W_{21}\ W_{11}^{-1}\ W_{12}$

(cf RAO : "Linear statistical inference and its applications" 1973).

$\rightarrow F(C2) = F(C1)$

VI.4. Autre critère

Trace de Pillaï $= Tr\ (V^{-1}\ B)$

= inertie du nuage des centres de gravité avec la métrique V^{-1}

= somme des "pouvoirs discriminants" des variables discriminantes.

= somme des carrés des coefficients de corrélation canonique.

que l'on cherche à maximiser (équivaut à $(C1)$ si $q = 2$).

BIBLIOGRAPHIE

G. SAPORTA : Probabilités, analyse des données et statistique, Technip (1990).

E. DIDAY, J. LEMAIRE, J. POUGET, F. TESTU : Éléments d'analyse de données, Dunod (1983).

R. TOMASSONE, M. DANZART, J.J. DAUDIN, J.P. MASSON : Discrimination et classement, Masson (1988).

G. CELEUX : (éditeur scientifique) Analyse discriminante sur variables continues, Inria (1990).

J.M. ROMEDER : Méthodes et programmes d'analyse discriminante, Dunod (1973).

P. DAGNELLE : Analyse statistique à plusieurs variables, Presses agronomiques de Gembloux (1977).

T.W. ANDERSON : An introduction to multivariate statistical analysis, Wiley (1984).

A.M. KSHIRSAGAR : Multivariate analysis, Marcel Dekker (1972).

P.A. LACHENBRUCH : Discriminant analysis, Hafner Press (1975).

A N N E X E

Présentation succincte des procédures d'analyse discriminante dans SAS 6.06

La procédure DISCRIM

La procédure DISCRIM réalise une analyse discriminante sur variables numériques, avec nombre de classes quelconques.

Elle permet la mise en œuvre de méthodes :

paramétriques (hypothèse de normalité des distributions des variables) :

- méthode linéaire ;
- méthode quadratique.

non-paramétriques

- méthode du noyau (estimation locale de densité)
- méthode des K plus proches voisins.

Elle évalue la méthode utilisée par des estimations de taux d'erreur de classement :

- par comptage des erreurs de classement (sur échantillon-test, ou par validation croisée) ;
- ou à partir des probabilités *a posteriori*.

Elle peut également effectuer une analyse canonique discriminante (comme CANDISC).

Syntaxe

PROC DISCRIM options ;

CLASS variable définissant les classes ;

PRIORS probabilités *a priori* ;

(EQUAL, ou PROP, ou liste de probabilités)

VAR variables numériques ;

ID variable ;

FREQ variable ;

WEIGHT variable ;

BY variables ;

TESCLASS variable ;
TESTFREQ variables ;
TESTID variable ;
} avec l'option TESTDATA

Quelques options de PROC DISCRIM

1. METHOD = { NORMAL (paramétrique) / NPAR (non-paramétrique) }

METRIC = { DIAGONAL (diag. variances intra) / FULL (variances intra) / IDENTITY (identité) }

POOL = { YES (W^{-1}) / NO (W_k^{-1}) / TEST (test de BOX) }

K = k (k plus proches voisins)

R = r (rayon de la méthode du noyau)

KERNEL = { BIWEIGHT / EPANECHNIKOV / NORMAL / TRIWEIGHT / UNIFORM } (méthode du noyau)

2. { TESTDATA = table SAS / TESTLIST liste des résultats / TESTLISTERR liste des erreurs } (échantillon test)

3. { LIST résultats de la ré-affectation pour toutes les observations / LISTERR seulement pour les erreurs / NOCLASIFY pas de ré-affectation pour l'échantillon de base }

4. { CROSSVALIDATE / CROSSLIST / CROSSLISTERR } Validation croisée

5. POSTERR estimation des taux d'erreurs (probabilités a posteriori)

6. CANONICAL analyse canonique

7. Impressions diverses et variées (matrices de corrélations, covariances, SSCP, totale, intra ..., statistiques univariées, multivariées ...)

8. Tables en sortie :

OUT = : table en entrée + probas a posteriori + classe d'affectation par la méthode de réaffectation (+ variables canoniques avec CANONICAL)

OUTSTAT = : table contenant différentes statistiques + résultats de la discrimination (fonctions discriminantes, corrélations canoniques ...)

OUTD = : table en entrée + estimations des densités de chaque classe en chaque point

OUTCROSS = : table en entrée + probas a posteriori + classes d'affectation par la méthode de validation croisée

TESTOUT = : comme OUT, pour la table TESTDATA

TESTOUTD = : comme OUTD, pour la table TESTDATA.

La procédure STEPDISC

La procédure STEPDISC réalise une sélection de variables numériques en vue d'une analyse discriminante, par une méthode pas à pas :

- ascendante ;
- ascendante avec remise en cause ;
- descendante.

Le critère utilisé pour entrer, ou sortir, une variable, est l'un des suivants :

1. le niveau de signification d'un test F d'une analyse de covariance ;

2. le carré du coefficient de corrélation partielle entre la variable et les variables de classe, les autres variables du modèle étant fixées .

Ces 2 critères ne diffèrent que par le nombre de variables introduites.

Syntaxe :

- PROC STEPDISC options ;
- CLASS variable définissant les classes ;
- VAR variables numériques ;

- FREQ variable ;
- WEIGHT variables ;
- BY variables ;

Quelques options de PROC STEPDISC

METHOD = { FORWARD | STEPWISE | BACKWARD }

SLENTRY = p : niveau du test F pour entrer (0,15 par défaut)

SLSTAY = p : niveau du test F pour rester (0,15 par défaut)

PR2ENTRY = p : R^2 partiel pour entrer

PR2STAY = p : R^2 partiel pour rester

INCLUDE = n : inclut les n premières variables de VAR dans chaque modèle

MAXSTEP = n : nombre maximum d'étapes

START = n : les n premières variables de VAR constituent le modèle de départ

STOP = n : nombre de variables du modèle final

La procédure CANDISC

La PROC CANDISC réalise une analyse canonique discriminante. Elle permet de créer une table SAS contenant les valeurs des variables (canoniques) discriminantes.

Syntaxe :

- PROC CANDISC options ;
- CLASS variable définissant les classes ;
- VAR variables numériques ;
- FREQ variables ;
- WEIGHT variables ;
- BY variables ;

Quelques options de PROC CANDISC

NCAN = n : nombre de variables canoniques calculées

OUT = : nom de la table SAS contenant les données en entrée + les variables canoniques

OUTSTAT = : nom de la table SAS contenant les résultats de l'analyse

DISTANCE imprime les distances de Mahalanobis entre les classes.