REVUE DE STATISTIQUE APPLIQUÉE

J. ULMO

Différents aspects de l'analyse discriminante

Revue de statistique appliquée, tome 21, nº 2 (1973), p. 17-55
<http://www.numdam.org/item?id=RSA_1973__21_2_17_0>

© Société française de statistique, 1973, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

DIFFÉRENTS ASPECTS DE L'ANALYSE DISCRIMINANTE

J. ULMO

Après avoir présenté dans une première partie l'Analyse discriminante du point de vue de l'analyse des données et du point de vue de la décision statistique, sans oublier le problème de la signification de l'influence d'une variable après prise en compte des autres, on traite dans une deuxième partie de l'analyse factorielle discriminante, toujours sous les deux points de vue métrique et statistique.

On montre à la fin de cette deuxième partie que l'analyse factorielle discriminante est un cas particulier de l'analyse canonique.

Une troisième partie est consacrée à l'application de ce qui précède à une extension de la régression linéaire multiple au cas où la variable à expliquer est qualitative.

SOMMAIRE

PREMIERE PARTIE : L'ANALYSE DISCRIMINANTE

1 – INTRODUCTION

1.1 – Généralités

L'analyse discriminante (A.D.) est une technique de classement ou de reconnaissance des formes par opposition à la typologie ou à la classification.

Dans la reconnaissance des formes ou le classement, les formes ou les groupes sont connus ; dans la typologie ou la classification, les types ou groupes ne sont pas connus.

Quand on considère l'analyse discriminante comme une technique d'analyse des données, elle appartient à la catégorie des techniques dans lesquelles l'ensemble des caractères est partitionné en deux groupes : le groupe des caractères à expliquer et le groupe des caractères explicatifs (ce groupe contient également les techniques de régression, de segmentation et d'analyse canonique).

Dans l'A.D., le ou les caractères à expliquer Y (ou Y_i s'il y en a plusieurs) sont qualitatifs ou de type présence-absence, tandis que les caractères explicatifs : x_i, $i = 1,2,\ldots, p$ sont quantitatifs.

Ce sont les modalités de Y (ou de l'ensemble des Y_i) qui définissent les groupes $1,2,\ldots, K$ et il s'agit, étant donné les valeurs prises par les caractères x_i, $i = 1,2,\ldots, p$ pour l'individu j, que l'on peut ainsi représenter après codage par un point $\underline{x}^j$ de R^p, de le classer dans l'un des groupes $1,2,\ldots, K$ définis par Y.

Il s'agit donc de définir une application h de R^p dans la famille des groupes définis par Y telle que le classement défini par cette application pour la population E des individus considérés soit le plus satisfaisant possible.

1.2 – Exemples de problèmes d'analyse discriminante

a) Problèmes de transport

Y est le moyen de transport utilisé et on peut, par exemple, définir 3 groupes :

– utilisation de la voiture particulière

– utilisation des transports en commun

– autres moyens de transport (pieds, bicyclette, etc.).

Il s'agit de discriminer entre ces 3 groupes d'après des critères sociologiques, de logement, etc.

b) Discrimination entre plusieurs types de voitures d'après des critères du type :

– prix de vente

– nombre de places

– vitesse de pointe

– consommation.

– etc.

Remarque : On peut aussi essayer de faire une discrimination entre les acheteurs de ces voitures d'après des critères sociologiques.

c) Discrimination entre plusieurs types de logements urbains d'après des critères relatifs aux ménages tels que catégorie socio-professionnelle du chef, âge du chef, revenu total, nombre d'enfants, etc.

d) Problèmes d'études de marché tel que celui de la télévision. On peut distinguer :

- les possesseurs d'une T.V. couleur
- les possesseurs d'une T.V. ordinaire
- les réfractaires à la T.V.

e) Historiquement, l'analyse discriminante a été, et est encore, très utilisée dans de nombreux problèmes d'identification d'espèces ou de sous-espèces en biologie, ethnologie, anthropologie (problèmes de discrimination entre les races en Inde, d'identification d'ossements d'hommes ou d'animaux fossiles, etc.).

Elle est aussi très utilisée dans de nombreux problèmes de diagnostic médical (par exemple, les problèmes de discrimination entre tumeurs bénignes et malignes et dans les tumeurs malignes entre différents types, d'après des critères cliniques, radiologiques ou histologiques).

1.3 – Formulation plus précise du problème

Le critère de satisfaction doit être défini, et pour ce faire il convient de mieux préciser ce qu'on recherche.

La partition $\mathcal{G}$ de E est définie par l'application canonique $e \rightarrow Y(e)$ de E dans $\mathcal{Y}$, tandis que l'observation de l'ensemble $\underline{x} \in R^p$ des caractères x_i définit l'application

$$e \xrightarrow{X} \underline{x}(e) \xrightarrow{h} y = h[x(e)]$$

et ce que l'on désire, c'est avoir pour tous les individus $e \in E$: $h_0\ x(e) = y(e)$, i.e. $h_0\ X = Y$. h est donc une application qui "factorise" Y à travers X (La représentation des caractères x_i, $i = 1,2,\ldots,p$ et Y comme des fonctions définies sur l'ensemble E des individus et à valeurs dans R et $\mathcal{Y}$ utilisée ici est très précieuse et sera souvent utilisée).

Comme il est pratiquement impossible d'avoir $h_0 X = Y$ si on veut définir h autrement que point par point sur E(1), il convient de définir un critère pour juger de l'écart entre $h_0\ X$ et Y sur E.

Deux types de critères sont généralement envisagés.

Des critères de type métrique, qui font intervenir une notion de distance entre $h_0 . X(e)$ et Y(e), distance qui est généralement définie dans $\mathbf{R}^p = h^{-1}(\mathcal{Y})$ sous forme de distance d'un individu e un groupe G_i (G_i est un groupe d'individus pour lesquels $Y = y_i$).

----

(1) Même dans ce cas ce ne sera pas possible s'il existe plusieurs individus de E ayant même valeur de $\underline{x}$ (i.e. même image par X) et des valeurs différentes de y (i.e. des images différentes par Y). i.e. si pour tout $\underline{x} \in x(E)$ on n'a pas : $\exists\ y_i$ tel que $X^{-1}(\underline{x}) \subset Y^{-1}(y_i)$, i.e. si $Y_0 X^{-1}$ n'est pas une application de X(E) dans $\mathcal{Y}$

Des critères de type probabiliste, qui font intervenir la notion d'erreur de classement d'un individu [il y a erreur de classement pour e si h_0 X(e) $\neq$ Y(e)(1)], et le coût moyen des erreurs de classement pour l'ensemble des individus de E (Ce coût est en relation directe avec le coût total si E est finie, mais il permet aussi de considérer une population infinie, ou non entièrement connue, à condition d'introduire un modèle probabiliste), ou à défaut les risques d'erreurs de classement attachés aux individus de $y_1, y_2, \ldots, y_k$ [plus précisément de $Y^{-1}(y_1)$, $Y^{-1}(y_2), \ldots, Y^{-1}(y_k)$] ou la somme de ces risques.

L'analyse discriminante est en effet un des types les plus simples de problèmes de décision statistique.

Le problème de l'A.D. est la recherche de h à partir d'un échantillon de n-individus e^j, j = 1,2, . . . , n pour lesquels on connait non seulement $\underline{x}^j = X(e^j)$ mais aussi y^j i.e. le groupe G_i auxquels il appartiennent.

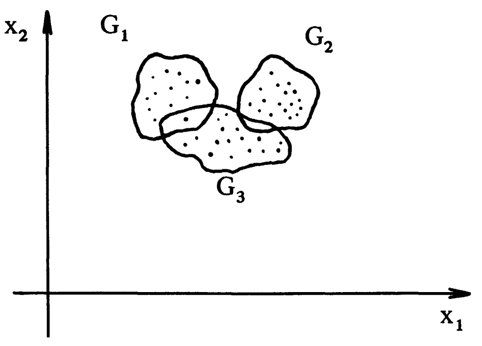
2 – PRINCIPE DES DIFFERENTES TECHNIQUES DE L'ANALYSE DISCRIMINANTE

2.1 – Techniques basées sur l'utilisation d'une distance dans R^p entre un individu e et un groupe G

L'échantillon $\mathcal{E}$ permet de définir dans R^p des nuages de points $\underline{x}^j$ appartenant aux différents groupes G^i. On affecte un individu e à la classe G_i, $1 = 1,2,\ldots, k$ telle que $D[\underline{x}(e), G_i] = \underset{i \in 1,2,\ldots,k}{\mathrm{Min}} [D(\underline{x}(e), G_i)]$

Il reste alors à définir D ($\underline{x}$, G) pour tout groupe $G \in \mathcal{G}$.

On peut définir cette distance de diverses manières (Cf. (3)).



La distance la plus utilisée en raison des simplifications qu'elle apporte et de ses justifications statistiques est celle qui définit D($\underline{x}$, G_i) par la distance euclidienne, pour une certaine métrique M, de $\underline{x}$ au barycentre $\underline{g}_i$ des points de $\mathcal{E}$ appartenant à G_i i.e. au barycentre du "nuage G_i de l'échantillon $\mathcal{E}$" qu'on appellera G_i^* pour le distinguer du nuage G_i de la population totale.

---

(1) Dans les problèmes de typologie, on ne peut plus parler d'erreur de classement puisque les types ne sont pas définis a priori.

Ce barycentre est $\underline{g}_i = \frac{1}{n_i} \sum_{\substack{\{j/Y(e_j) = y_i\} \\ j \in \{1,2,\ldots,n\}}} \underline{x}_j$, n_i = Card. G_i^* = Card ($G_i \cap \mathcal{E}$)

On peut utiliser d'autres définitions de la distance de $\underline{x}$ à G_i, telle que la moyenne des carrés des distances de $\underline{x}$ aux points de G_i (cf. (2)), la plus petite distance de $\underline{x}$ aux points de G_i, mais on n'obtient plus de solutions aussi simples qu'avec la distance barycentrique euclidienne.

La métrique la plus utilisée est celle qui consiste à prendre pour M non pas la métrique euclidienne classique I, mais l'une des deux métriques définies par l'inverse de la matrice de covariance des variables x_1, soit dans les sous-groupes G_i^* de $\mathcal{E}$, soit dans l'échantillon total ,

soit : $M = \Sigma^{-1}$ ou $M = T^{-1}$

avec $\Sigma = \frac{1}{n} \sum_{i=1}^{k} \sum_{x^j \in G_i^*}^{\text{ou } M = T^{-1}} (\underline{x}^j - \underline{g}_i)(\underline{x}^j - \underline{g}_i)'$ matrice de covariance des individus $\underline{x}^j$ dans les groupes

et $T = \frac{1}{n} \sum_{x^j \in \mathcal{E}} (\underline{x}^j - \underline{x})(\underline{x}^j - \underline{x})$, matrice de covariance "totale" des individus.

La distance ou métrique définie par Σ^{-1} dans $\mathbf{R}^p$ est la distance de Mahalanobis relative aux groupes G_i^*. Elle a une justification essentiellement statistique (Cf. 2.2.4-2.). Elle est également susceptible d'une justification géométrique puisqu'on montre (cf. (2)) que la métrique M telle que le moment d'inertie d'un nuage p dimensionnel par rapport à son centre de gravité soit minimum sous la condition $|M| = 1$, est $M^* = |\Sigma|^{1/p} \Sigma^{-1}$, si Σ est la matrice covariance du nuage ($|M|$ = déterminant de M joue le rôle de facteur de normalisation puisque si M est remplacée par λM, le moment d'inertie est multiplié par λ).

Σ^{-1} est donc la métrique M pour laquelle la somme des inerties de chacun des groupes G_i^* par rapport à leurs propres centres de gravité $\underline{g}_i^*$ est minimale, sous la condition $|M| = C^{te}$.

La distance définie par T^{-1} peut être justifiée par des considérations géométriques analogues ou liées à "l'analyse factorielle discriminante". (cf. IIe partie).

C'est donc le signe de la fonction

$$\delta_{i/l(x)} = (\underline{x} - \underline{g}_l)'M(\underline{x} - \underline{g}_l) - (\underline{x} - \underline{g}_i)'M(\underline{x} - \underline{g}_i)$$

soit

$$\delta_{i/l}(\underline{x}) = 2x'M(g_i - g_l) - (g_i + g_l)'M(g_i - g_l)$$

$$= -2\left(\underline{x} - \frac{g_i + g_l}{2}\right)' M(\underline{g}_i - \underline{g}_l)$$

qui est une fonction linéaire de $\underline{x}$ qui permet de décider si $\underline{x}$ est plus proche de G_i^* que de G_l^*

$\delta_{i/l}(\underline{x}) > 0 \Rightarrow \underline{x}$ est plus proche de G_i^* que de G_l^*

$\delta_{i/l}(\underline{x}) < 0 \Rightarrow \underline{x}$ est plus loin de G_i^* que de G_l^*.

Si $\delta_{i/l}(\underline{x}) > 0$, on examine alors le signe de $\delta_{i/m}(\underline{x})$, $m \neq l$, tandis que si $\delta_{i/l}(\underline{x}) < 0$, soit $\delta_{l/i}(\underline{x}) > 0$, puisque $\delta_{l/i}(\underline{x}) = -\delta_{i/l}(\underline{x})$, on examine le signe de $\delta_{l/m}(\underline{x})$, $m \neq i$, etc..

La fonction $\delta_{i/l}(x)$, $i, l \in \{1,2,\ldots, k\}$, $i \neq l$, est appelée fonction discriminante du groupe i par rapport à l et on voit qu'il y en a en principe $\frac{k(k-1)}{2}$ de distinctes.

La règle d'affectation est alors définie par

$$x \rightarrow G_i / \{\delta_{i/l}(x) > 0, \forall\, l \neq i\} \text{ (1)}$$

La manière dont cette règle a été obtenue montre qu'il y a toujours une affectation et une seule, sauf si $\delta_{i/l}(\underline{x}) = 0$, cas auquel on peut choisir indifféremment entre G_i et G_l.

Ainsi dans $\mathbf{R}^p$, les groupes G_i sont limités chacun par $k - 1$ hyperplans passant chacun par les milieux des droites joignant $\underline{g}_i$ aux centres de gravité $\underline{g}_l$ des groupes G_l^* formés par les points de l'échantillon $\mathcal{E}$, et M orthogonaux à ces droites $\underline{g}_i \cdot \underline{g}_l$, c'est-à-dire par les $k - 1$ hyperplans médiateurs de $\underline{g}_i \cdot \underline{g}_l$ pour la métrique M.

Remarques importantes

1/ La forme de la fonction discriminante $\delta_{i/l}$ ne dépend que des indices i et l et par conséquent elle n'est pas modifiée si on décide de réunir plusieurs des autres groupes que i et l en un seul, ou au contraire de diviser certains de ces autres groupes, ou même si on décide de ne considérer que les deux seuls groupes i. et l., dans la mesure où la matrice M n'est pas liée à la répartition dans les groupes de l'ensemble des observations de $\mathcal{E}$

La remarque s'applique ainsi avec $M = T^{-1}$, mais non avec $M = \Sigma^{-1}$, et pourrait justifier le choix de $M = T^{-1}$ plutôt que $M = \Sigma^{-1}$.

Cette remarque peut être très utile dans la pratique. C'est ainsi que dans la discrimination entre les tumeurs on peut dans une 1ère étape réunir un certain nombre de types variés de tumeurs malignes (par exemple, très malignes) sans changer la fonction discriminante entre bénignes et peu malignes.

---

(1) Au point de vue calcul sur ordinateur, il est finalement plus simple de calculer $d^2(\underline{x}, \underline{g}_i)$ pour tous les groupes G_i et de regarder celui qui donne $d^2(\underline{x}, \underline{g}_i)$ minimum que de calculer les $\frac{k(k-1)}{2}$ $\delta_{i/l}(\underline{x})$.

2/ Le pouvoir discriminant de la règle adoptée peut être caractérisé par le pourcentage d'erreurs de classement sur les individus de $\mathcal{E}$ à laquelle elle conduit, ou par le coût total de ces erreurs.

On peut aussi envisager de supprimer des variables explicatives x_1 dont l'influence parait peu sensible pour discriminer entre les groupes. (variables pour lesquelles les centres de gravité $\underline{g}_l$, sont peu distincts les uns des autres en sorte que leur contribution dans la différenciation entre les $d^2(\underline{x}, G_i)$ sera très faible. Il faut cependant faire attention si on n'utilise pas la métrique $M = I$). On pourra juger de l'intérêt de cette suppression par la variation du pourcentage d'erreurs de classement des individus de $\mathcal{E}$, mais si on veut réduire le nombre des variables explicatives il apparait préférable de procéder à une analyse factorielle discriminante qui si $k \leqslant p$, permettra de n'avoir que $k - 1 < p$ variablesdiscriminantes (cf. 2e partie), ou de considérer le problème d'un point de vue probabiliste et de procéder à un test de signification du pouvoir discriminant d'une variable quand les autres sont choisies (Cf. 2.2.5).

2.2 – Techniques probabilistes : l'analyse discriminante, problème de décision statistique [(1) et (4)]

2.2.1 – Les ensembles fondamentaux

Soit $(E, \mathcal{A})$ la population considérée ($\mathcal{A}$ est la tribu d'événements de E) et $X : e \rightarrow X(e)$ une variable aléatoire définie sur E et à valeurs dans $\mathcal{X}$. Dans le cas présent $\mathcal{X} \subset \mathbf{R}^p$ et X est une variable vectorielle de dimension p.

$\mathcal{X}$ constitue l'espace des observations, et sur $\mathcal{X}$ est définie une famille de mesures de probabilité P_θ, $\theta \in \Theta$. Θ est l'espace des états de la nature.

Dans le cas présent, la famille P_θ est finie et de cardinal k et

$$\Theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$$

avec θ_i : l'observation provient d'un élément de E_i, $E_1, E_2, \ldots, E_k$ étant une partition de la population totale E.

Ceci revient à dire que si les observations $\underline{x}^j \in \mathbf{R}^p$ proviennent de la souspopulation E_i, on leur affectera la mesure de probabilité $P_{\theta_i} = P_i$ X induit sur les E_i les applications X_i, ou si on préfère les variables aléatoires X_i dont les lois de probabilité sont P_i, et il s'agit au vu d'une réalisation x de X de savoir de quelle X_i elle provient.

Une règle de décision est une application δ de $\mathcal{X}$ dans l'espace A des décisions (ou des actions). Dans le cas présent A peut être confondu avec Θ. On a en effet $A = \{a_1, a_2, \ldots, a_k\}$ avec a_i : faire comme si l'observation provient de E_i, i.e. décider que l'observation provient d'un élément de E_i.

La règle d'affectation d'un individu à l'un des groupes E_i (ou G_i) à partir d'une observation $x \in \mathcal{X}$ est donc bien une règle de décision statistique δ telle qu'on vient de la définir.

2.2.2 – Critère de choix d'une règle de décision δ et recherche de cette règle : Choix d'une règle de Bayes

La fonction de perte est ici une application de A x Θ dans $\mathbf{R}^+$. $1(a_i, \theta_1)$ représente le coût du classement d'un individu de E_1 dans E_i. On a donc

$$1(a_i, \theta_i) = 0 \quad \forall\, i \in \{1,2, \ldots, k\}$$

et
$$1(a_i, \theta_1) > 0$$

soit
$$1(a_i, \theta_1) = C_{i/1} > 0, \quad i \neq 1.$$

Probabilité a priori sur Θ

Θ étant fini, cette probabilité q est définie par un ensemble de k nombres q_i, $i \in \{1,2, \ldots, k\}$: $q_i = P'(\theta = \theta_i)$

$$0 \leqslant q_i \leqslant 1 \quad \text{et} \quad \sum_i q_i = 1.$$

On sait que la règle de Bayes δ^* associée à la probabilité a priori q et à la fonction de perte 1 est celle qui minimise la perte moyenne :

$$E[1(\delta^*(x), \theta)] = \underset{\delta \in \Delta}{\mathrm{Min}}\ E[1(\delta(x), \theta)]$$

où Δ est l'ensemble des règles de décision considérées (ensemble d'applications de $\mathscr{X}$ dans A), et l'espérance E est prise pour la loi du couple $(\underline{x}, \theta) \in [\mathscr{X} \times \Theta]$. La règle de Bayes est donc celle qui minimise le coût moyen des erreurs de classement pour l'ensemble des valeurs $\underline{x}$ et θ possibles.

Recherche de la règle de Bayes δ^*

A une règle simple δ, application de $\mathscr{X}$ dans A, est associée une partition de $\mathscr{X}$ définie par les ensembles

$$G_\delta(i) = \{x, \delta(x) = a_i\} = \delta^{-1}(a_i) \qquad i \in \{1,2, \ldots, k\}$$

$G_\delta(i)$ est la région de $\mathscr{X}$ formée par les points affectés à E_i par la règle δ.

Remarque :

On pourrait envisager des règles de décision aléatoires, i.e. remplacer l'application δ par une distribution de probabilité sur A ; cette conception peut être utile dans des problèmes voisins de ceux de l'A.D. classique tels que la régression d'une variable qualitative sur des variables quantitatives qui est étudiée en 3ème partie.

Proposition

Si les mesures P_i possèdent des densités p_i par rapport à une mesure μ sur $\mathscr{X}$, les regions $R_i = G_{\delta^*(i)}$, $i \in 1, 2, \ldots, k$ qui minimisent le cout moyen des erreurs de classements sont définies par

$$R_1 = \left\{ x \,/\, \sum_{\substack{i=1 \\ i \neq 1}}^{k} q_i\, p_i(x)\, C_{1/i} < \sum_{\substack{i=1 \\ i \neq m}}^{k} q_i\, p_i(x)\, C_{m/i} \right\}$$

$$1, m \in \{1,2,\ldots,k\} \qquad 1 \neq m$$

Démonstration. – Nous utiliserons la méthode classique de recherche d'une règle de Bayes (4).

On peut écrire $E_{X,\theta}[1(\delta(X), \theta)] = E_X[E_{\theta/x}[1(\delta(x), \theta)]]$ en désignant par $E_{(X,\theta)}$, E_X, $E_{\theta/x}$ les espérances prises respectivement pour la loi du couple (X, θ), la loi marginale de X et la loi de θ conditionnée par $X = x$, c'est-à-dire la loi a posteriori de θ quand on a observé $\underline{x}$, en sorte que s'il existe une règle δ^* minimisant $E_{\theta/x}[1(\delta(x), \theta)]$ pour tout $x \in \mathfrak{X}$, elle minimisera $E_{X,\theta}[1(\delta(X),\theta)]$.

Comme l'espace A des décisions est fini, une telle règle existe. Elle est définie par $\underline{x} \rightarrow a_1$ tel que

$$E_{\theta/x}[1(a_1, \theta)] = \min_{m \in (1,2,\ldots k)} E_{\theta/x}[1(a_m, \theta)]$$

La règle δ^* conduit donc à affecter un $\underline{x}$ donné au groupe 1 minimisant le coût moyen de l'erreur de classement de ce $\underline{x}$.

Calculons $E_{\theta/x}[1(a_m, \theta)]$.

La probabilité, $\underline{x}$ étant donné, que $\theta = \theta_i$ est d'après le théorère de Bayes :

$$p(\theta_i/x) = \frac{p(x/\theta_i)\, p(\theta_i)}{p(x)} = \frac{q_i p_i(x)}{\sum_i q_i p_i(x)}$$

On a donc $E_{\theta/x}[1(a_m, \theta)] = \sum_{i=1}^{k} C_{m/i} \dfrac{q_i p_i(x)}{\sum q_i p_i(x)}$

$$= \frac{\sum_{i \neq m} C_{m/i}\, q_i p_i(x)}{\sum_{i=1}^{k} q_i p_i(x)} \quad \text{puisque } C_{m/m} = 0$$

Minimiser $E_{\theta/x}[1(a_m, \theta)]$ revient donc à minimiser $\sum_{i \neq m} C_{m/i}\, q_i\, p_i(x)$ ce qui démontre la proposition.

2.2.3. – *Etude du cas particulier où on pose $1(a_i, \theta_l) = C > 0$ pour tout $i \neq 1$*

(Toutes les erreurs de classement coûtent le même prix)

Alors le critère de classement est :

$$R_l = \left\{ \underline{x} \Big/ \sum_{\substack{i=1 \\ i \neq l}}^{k} q_i \, p_i(x) < \sum_{\substack{i=1 \\ i \neq m}}^{k} q_i \, p_i(x) \;\; \forall \; l \neq m \right\}$$

En soustrayant $\sum\limits_{\substack{i \neq l \\ i \neq m}} q_i \; p_i(x)$ des 2 membres de l'inégalité on obtient :

$$R_l = \{\underline{x} / q_l \; p_l(x) > q_m \; p_m(x) \;\; \forall \; l \neq m\}$$

On vient de voir que $p(\theta_l / \underline{x}) = \dfrac{q_l \, p_l(\underline{x})}{\sum\limits_i q_i \, p_i(x)}$

Par suite, on affecte x à la population E_l dont la probabilité a posteriori, $\underline{x}$ étant donné, est la plus forte, ce qui paraît très naturel.

Dans la pratique, il est intéressant de calculer les $p(\theta_i / \underline{x})$, $i \in 1, 2, \ldots k$ correspondant à un élément $\underline{x}$ à classer car il se peut que la classe choisie soit très peu préférable à une autre, ou au contraire très nettement préférable à toutes les autres.

Remarques :

1/ Si on fait c = 1, la perte moyenne attachée à une règle de classement δ devient la probabilité d'erreur de classement qui est donc

$$\alpha_\delta = \sum_{l=1}^{k} \int_{G_{\delta_l}} \left[\sum_{\substack{i=1 \\ i \neq l}}^{k} q_i \, p_i(x) \right] d\mu(x) \, ,$$

soit

$$\alpha_\delta = \sum_{l=1}^{k} \left[\sum_{i \neq l} q_i \, P_i(G_{\delta_l}) \right]$$

où $P_i(G)$ désigne la probabilité de l'ensemble G de $\mathfrak{X}$ quand $\theta = \theta_i$, i.e. pour la mesure de probabilité θ_i.

2/ *Si les probabilités a priori ne sont pas connues,* on est souvent conduit à les supposer toutes égales à $\frac{1}{k}$. On est alors conduit à affecter $\underline{x}$ à la population E_l qui lui donne la plus grande probabilité d'être observé. où la plus grande vraisemblance.

La règle d'affectation est alors une règle du Maximum de vraisemblance.

3/ On peut aussi introduire les risques d'erreur de classement attachés à chaque sous-population E_i, $i \in \{1, 2, \ldots, k\}$ soit

$$\alpha_i(\delta) = 1 - P_i [\delta(\underline{x}) = a_i] = 1 - P_i \, [x \in G_\delta \, (i)] \, , \, i = 1, 2, \ldots, k$$

$\alpha_i(\delta)$ est la probabilité des erreurs de classement attachée à la règle δ quand les individus proviennent de E_i (C'est le risque de 1ère espèce de la théorie des tests quand $k = 2$).

On peut vouloir limiter certains α_i ; par exemple dans le diagnostic médical, on peut vouloir limiter le risque de déclarer non-malade ou atteint d'une affectation légère (tumeur bénigne par exemple) un malade grave (tumeur maligne).

On peut aussi chercher une règle $\tilde{\delta}$ qui minimise le plus grand des risques $\alpha_i(\delta)$, $i \in 1, 2, \ldots, k$ i.e. qui minimise $\underset{i \in 1,2,\ldots k}{\text{Max}} \alpha_i(\delta)$.

Une telle règle est appelée règle du Minimax. Elle se place dans l'hypothèse la plus pessimiste (On peut montrer que c'est une règle de Bayes correspondant à une distribution a priori pour θ dite la moins favorable).

On montre que les $\underset{i \in 1,2,\ldots k}{\alpha_i(\tilde{\delta})}$ sont tous égaux, mais la recherche d'une règle du minimax n'est en général pas commode.

2.2.4 – Application pratique de la technique probabiliste

Il faut connaître les probabilités a priori q_i des différentes sous-populations, les valeurs $c_{i/l}$ de la fonction de coût et enfin les lois de probabilité P_i de X pour chacune des sous-populations.

1/ *Cas d'un grand échantillon*

Dans la pratique quand on dispose d'un grand échantillon $\mathcal{E}$, on peut estimer les q_i par les effectifs n_i/n des groupes G_i^*. Il paraît plus difficile d'estimer les distributions P_i par des distributions observées pour $\underline{X}$ sur les groupes G_i^*, si la dimension p de $\underline{X}$ n'est pas très faible ou si $\underline{X}$ est continue, ou a certaines de ses composantes continues et on risque d'obtenir des fréquences nulles pour un bon nombre de classes de valeurs pour $\underline{X} \in R^p$.

Il faut alors faire appel à des méthodes d'estimation des densités de probabilité pour une variable multidimensionnelle, qui peuvent être paramétriques ou non, selon la connaissance a priori que l'on a sur les lois $P_i(\underline{x})$.

On montre, (cf. (5)), que si les estimateurs obtenus sont convergents quand l'ensemble des n_i tend vers l'infini, la règle estimée $\hat{\delta}^*$ converge vers la règle optimale δ^* ; de même les probabilités estimées d'erreurs de classement $\hat{\alpha}$ et $\hat{\alpha}_i$ et la perte moyenne estimée convergent vers les valeurs correspondant à la règle optimale δ^* bien que ces estimateurs soient souvent des estimateurs biaisés par défaut, c'est-à-dire optimistes, des valeurs correspondantes pour δ^*. (ceci a lieu si les estimateurs des densités sont sans biais et plus généralement si

$$E[\hat{q}_i \, \hat{p}_i(\underline{x})] \geqslant q_i \, p_i(\underline{x})$$

pour presque tout $\underline{x} \in \mathcal{X}$).

Le même auteur donne une bibliographie commentée importante concernant les méthodes d'estimation de densités de probabilité, et de probabilités d'erreurs de classement en analyse discriminante.

En ce qui concerne les $c_{i/l}$, valeurs de $l(a_i/\theta_l)$, si on n'a aucune idée de leurs grandeurs respectives, on les supposera toutes égales. Si on a quelqu'idée sur leurs valeurs on peut envisager de faire des essais avec différents systèmes de valeurs et comparer les résultats obtenus, en particulier pour le classement des individus de $\mathcal{E}$ dont on connaît le classement réel.

Il est en effet toujours intéressant de regarder ce que donne la procédure trouvée pour les individus de $\mathcal{E}$, et en particulier de s'intéresser à la proportion $\widetilde{\alpha}(\mathcal{E})$ des erreurs de classement de ces individus et aussi aux proportions $\widetilde{\alpha}_i(\mathcal{E})$ attachées à chaque sous population G_i^*.

On a
$$\widetilde{\alpha} = \frac{\sum_i n_i \widetilde{\alpha}_i}{n} .$$

2/ Cas particulier où on suppose que P_i *est une loi normale* de moyenne μ^i et de matrice *de covariance* Σ *identique pour toutes les valeurs de i, et que le coût des erreurs de classement est constant :* $c_{i/l} = c$, $i \neq 1$.

R_l est définie par $\left\{ x \;/\; \text{Log} \dfrac{p_l(x)}{p_i(x)} > \log \dfrac{q_i}{q_l} \;,\; i \in 1, 2, \ldots, k \;\; {\scriptstyle i \neq l} \right\}$

et
$$\text{Log} \frac{p_l(x)}{p_i(x)} = \left[x - \frac{1}{2} (\mu^l + \mu^i) \right]' \Sigma^{-1} \left[\mu^l - \mu^i \right]$$

On a donc
$$\text{Log} \frac{p_l(x)}{p_i(x)} = 2\, \delta_{l/i}(x), \text{ en prenant } M = \Sigma^{-1}$$

($\delta_{i/}$ a été définie en 2.1.) et $g_i = \mu^i \ \forall$ i.

On retrouve donc les fonctions discriminantes linéaires $\delta_{l/i}(x)$ pour définir les régions R_l, mais le critère d'appartenance à R_l est :

$$\delta_{l/i}(x) > \frac{1}{2} \text{ Log} \frac{q_i}{q_l} \quad \forall\, i \neq 1 ,$$

alors que dans la technique géométrique c'était $\delta_{l/i}(x) > 0$.

On retombe exactement sur le même critère quand $q_i = q_l \ \forall$ i, i.e. quand toutes les populations ont mêmes probabilités a priori.

Dans la pratique les $\underline{\mu}^i$ et Σ sont inconnus et on estime les $\underline{\mu}^i$ par les moyennes observées des groupes soit $\underline{g_i}$ (notées plus souvent $\underline{\bar{x}}^i$) et Σ par la covariance empirique dans les groupes déjà définie en 2.1. sous la dénomination de Σ (On prend aussi parfois
$$\Sigma^* = \frac{n\,\Sigma}{n-k}$$
pour tenir compte du fait qu'on a estimé k moyennes de groupes. Σ^* est un estimateur sans biais de Σ. Cela ne change rien quand les q_i sont tous égaux ou quand n est grand devant k).

Les estimateurs correspondant des densités $p_i(x)$ sont convergents, aussi les résultats énoncés en 1/ s'appliquent-ils, en ce qui concerne la convergence de la solution trouvée, mais non en ce qui concerne le biais des probabilités d'erreurs de classement car les estimateurs des densités sont biaisés. Les études faites dans ce cas particulier (cf. références commentées dans (5) p. 118), semblent indiquer que là encore on a un biais par défaut, c'est-à-dire une estimation optimiste des probabilités d'erreur de classement.

On trouvera dans (6) des estimateurs sans biais et paramétriques des densités $p_i(x)$ dans ce cas.

2.2.5 – *Test de signification de l'influence d'une variable, par exemple* x_p, après *prise en compte des* $p - 1$ *premières variables. Application à l'A.D. stepwise.*

On teste habituellement si la probabilité de x_p conditionnée par les x_i , $i \neq p$ peut être considérée comme étant la même pour les différentes populations θ_i, $i \in 1, 2, \ldots, k$.

Si oui, cela signifie que, après prise en compte de $x_1, x_2, \ldots, x_{p-1}$, x_p n'est d'aucune utilité pour discriminer entre les θ_i.

Sinon, x_p peut apporter une information pour discriminer, mais cette possibilité n'est nullement une certitude.

Le test habituel répond donc à la question : x_p *est-il susceptible* d'apporter une information supplémentaire pour discriminer,et non à la question réelle : x_p améliore-t-il effectivement la discrimination, c'est-à-dire, diminue-t-elle la perte moyenne, ou la probabilité des erreurs de classement dans le cas d'un coût d'erreur constant.

Il est généralement difficile d'estimer la perte moyenne ou la probabilité des erreurs de classement, autrement que directement à partir des résultats obtenus avec la règle estimée $\hat{\delta}$ pour les individus de $\mathcal{E}$, sauf dans le cas de $k = 2$ groupes, sous les hypothèses de normalité, d'égale covariance et d'égale probabilité à priori pour les deux groupes et d'un coût d'erreur constant.

Cas de 2 groupes et d'une distribution normale de même covariance pour les 2 groupes

Il n'y a alors qu'une seule fonction discriminante et la probabilité d'erreur de classement est (1) :

$$\alpha = \Phi\left(-\sqrt{\Delta^2}\right)$$

où Φ est la fonction de répartition de la variable normale centrée réduite et $\Delta^2 = (\underline{\mu}^1 - \underline{\mu}^2)'\Sigma^{-1}(\underline{\mu}^1 - \underline{\mu}^2)$ est le carré de la distance de Mahalanobis entre les deux groupes.

α est donc une fonction décroissante de Δ^2 et on est amené à étudier comment varie Δ^2 quand on passe de p à $p - 1$, et plus généralement à $p - q$, variables X_i (8). On peut calculer explicitement $\omega_p^2 = \Delta_p^2 - \Delta_{p-1}^2$ et montrer que

$\omega_p^2 \geqslant 0$, donc $\alpha_p \leqslant \alpha_{p-1}$. Dans la pratique on ne peut qu'estimer Δ^2 par la distance de Mahalanobis empirique D^2 et en déduire un estimateur $\hat{\alpha}$ pour α qui est, on l'a déjà dit en 2.2.4, 1°, biaisé par défaut, aussi a-t-on remplacé D^2 par d'autres expressions $\hat{D}^2$ tendant à réduire ce biais (8). Ces expressions ne croissent plus nécessairement quand p croît, en sorte qu'on peut prendre comme critère de suppression de x_p, $\hat{D}_p^2 - \hat{D}_{p-1}^2 \geqslant 0$.

Cochran (9) a étudié le pouvoir discriminant de l'ensemble des variables X_i, $i = 1, 2, \ldots, p$, caractérisé par Δ_p^2, en fonction de celui de chacune de ces variables caractérisé par $d_i^2 = \dfrac{(\mu_i^1 - \mu_i^2)^2}{\sigma_i^2}$ en vue d'éliminer, avant de calculer la fonction discriminante, un groupe de variables dont le pouvoir discriminant paraît faible. Un examen de l'effet de corrélations entre variables dans un certain nombre de cas lui suggère les conclusions suivantes :

a) Une variable corrélée négativement aux autres apportera sans doute une contribution complémentaire plus importante que ce qu'on aurait pu attendre de son pouvoir discriminant individuel, tandis qu'une variable corrélée positivement fera l'inverse.

b) Si la contribution d'un groupe de q variables, prédite en supposant l'indépendance des variables, et donc mesurée par $\sum_{i=1}^{q} d_i^2$, est faible, il est peu probable que sa contribution effective soit importante.

C'est la difficulté du calcul de α qui a conduit à essayer de tourner la difficulté avec le test habituel qui est d'une mise en œuvre aisée sous les hypothèses de normalité et d'égale matrice de covariance pour les P_i. En effet on est ramené à tester si les hypersurfaces de régression de X_p sur les X_i, $i < p$, sont ou non confondues (les hypothèses d'égale covariance impliquent leur parallélisme). On est donc ramené à un problème classique de comparaison de k régressions linéaires multiples qu'on résoud par l'analyse de covariance (cf. par exemple (7)).

Si la réponse au test est : acceptation de l'hypothèse nulle : pas d'information supplémentaire, on peut juger inutile de pousser plus loin dans la mesure où on ne le désire pas, car il ne faut pas oublier qu'en théorie classique des tests on est a priori favorable à l'hypothèse nulle.

Si la réponse est : rejet de l'hypothèse nulle, si on souhaite vraiment limiter le nombre des variables explicatives, il faut examiner si x_p améliore effectivement la discrimination.

Proposition de tests pour juger si x_p *améliore effectivement la discrimination, sous l'hypothèse de normalité et d'égale covariance des* P_j, *basés sur le test de Wilks de comparaison sur l'échantillon* $\mathcal{E}_2$ *des moyenne* μ^j, $j = 1, 2, \ldots, k$, *des* k *groupes*

Pour pouvoir trouver une règle de décision δ^* qui permette de discriminer entre les k groupes sans que le risque d'erreur soit trop grand, il faut que les k lois de probabilité P_j soient effectivement différentes.

Avec les hypothèses faîtes, seules leurs moyennes $\underline{\mu}^j$ peuvent diférer, et on dispose entre autres tests, du test de Wilks basé sur le critère $U = \frac{W}{T}$ pour tester l'hypothèse nulle H d'égalité des μ^j, $j = 1, 2, \ldots, k$ [(1) et (10)].

W et T sont respectivement les matrices de covariance empiriques "dans les groupes" ("within") et totale. (cf. 2.1 où W était désignée par Σ).

On a toujours $U \leqslant 1$, et on sait que si H est réalisée, U a une distribution $U_p(n-k, k-1)$ de Wilks, tandis que si H n'est pas réalisée U a tendance à être inférieur à U_p, d'où une région de rejet de seuil α de la forme $U < U_\alpha$ avec

$$P[U_p(n-k, k-1) < U_\alpha] = \alpha.$$

1er test (très statistique)

Pour juger si x_p améliore effectivement la discrimination, nous proposons donc de calculer les statistiques U(p) et U(p−1) correspondant respectivement à l'ensemble des p variables x_i et aux $p - 1$ premières et de déterminer les probabilités

$$P[U_p(n-k, k-1) < U(p)] = \alpha(p) \quad \text{et} \quad P[U_{p-1}(n-k, k-1) < U(p-1)] = \alpha(p-1).$$

Selon que l'on a $\alpha(p) < \alpha(p-1)$ ou non, on conservera la variable x_p ou on l'éliminera.

L'inconvénient de la méthode est d'une part d'exiger le calcul de U(p) et U(p−1), d'autre part le fait que la loi de Wilks n'est généralement pas connue exactement mais seulement approchée.

En fait la plupart des programmes classiques d'analyse discriminante (et en particulier le programme B.M.D. 07. M d'analyse discriminante "stepwise") prévoient le calcul de U et le test de l'hypothèse H [cf. (10) et (11 : 8 c.4) "Test de l'information additionnelle apportée par un groupe de variables"].

Par ailleurs on peut montrer qu'on a toujours $U(p) \leqslant U(p-1)$ (1) et que le rapport $U(p, p-1) = \frac{U(p)}{U(p-1)}$ est tel que $\frac{1 - U(p, p-1)}{U(p, p-1)} \; \frac{n-(k+p-1)}{k-1}$ est égal à la statistique G(p), utilisée en analyse de covariance pour tester l'hypothèse H_1 de même distribution de x_p conditionnée par les x_i, $i = 1, 2, \ldots, p-1$ et qui sous H_1 est distribuée comme $F[k-1, n-(k+p-1)]$.

Ayant calculé cette statistique pour procéder au test habituel de signification de l'influence de x_p, il est facile d'en déduire $U(p, p-1)$ et par suite

$$U(p-1) = \frac{U(p)}{U(p, p-1)}.$$

---

(1) Ce résultat ne permet pas d'admettre qu'en première approximation les lois de $U_p(n-k, k-1)$ et de $U_{p-1}(n-k, k-1)$ diffèrent peu et d'en déduire le critère simple : garder x_p si $\frac{U(p)}{U(p-1)} < 1$.

Remarques

1/ C'est le fait que $\frac{U(p)}{U(p-1)}$ soit au plus égal à 1 et fonction de G(p) qui nous a fait renoncer à définir un test basé sur U(p, p − 1) et qui serait de la forme :
conserver x_p si $U_{p,p-1} < \lambda$ où λ est une constante à déterminer, car ce test serait équivalent au test de H_1, au risque α adopté près.

2ème test (de caractère plus géométrique et empirique)

L'interprétation géométrique de U(p) comme rapport des carrés des volumes de deux parallèlotopes à p dimensions de $\mathbf{R}^n$ [(1) et (10)], qui le rend homogène à la puissance $2p^e$ du rapport de deux longueurs, conduit à comparer $U(p)^{1/p}$ et $U(p-1)^{1/p-1}$ et à adopter une règle de la forme : conserver x_p si $\frac{U_{(p)}^{1/p}}{U_{(p-1)}^{1/p-1}} < 1$, ou plus généralement si $\frac{U_{(p)}^{1/p}}{U_{(p-1)}^{1/p-1}} < \lambda$ où λ pourrait être une constante indépendante de p et au plus égale à 1. (On pourrait prendre λ de telle façon que pour p = 1, $U_{(1)} < \lambda$ corresponde à une différence significative des moyennes de la variables x_p considérée, à un risque α à définir, ce qui conduit à

$$\frac{1-U}{U}\,\frac{n-k}{k-1} < F_\alpha(k-1, n-k)$$

soit

$$\lambda = \frac{1}{F_\alpha(k-1, n-k)\dfrac{k-1}{n-k} + 1}.\Bigg)$$

On peut écrire

$$\left[\frac{U_{(p)}^{1/p}}{U_{(p-1)}^{1/p-1}}\right]^{p-1} = \frac{U(p)}{U(p-1)} \times \frac{1}{U_{(p)}^{1/p}} = \frac{U(p, p-1)}{U_{(p)}^{1/p}}$$

en sorte que la règle serait : conserver x_p si $U_{(p,p-1)} \leqslant \lambda^{p-1} U_{(p)}^{1/p}$ où λ est fixé, $\lambda \leqslant 1$. On peut penser que ce deuxième test est plus "robuste" que le premier, vis-à-vis d'écarts aux hypothèses faites sur les lois de probabilité P_j.

Examen du cas où il y a non corrélation empirique de x_p *avec les* x_i, $i < p$, *à la fois dans les groupes et pour l'ensemble de l'échantillon.*

Ce cas apparemment très particulier, que nous examinons pour mieux comprendre le test proposé, se trouve réalisé pour les "variables canoniques" de l'analyse factorielle discriminante (cf. 2ème partie 2.3)

La statistique G du test de l'hypothèse H_1 de même distribution de x_p conditionnée par les x_i, $i < p$ dans les groupes est

$$G_{(p)} = \frac{b_p}{w_p}\,\frac{n-(k+p-1)}{k-1}$$

où w_p = variance empitique de x_p dans les groupes

b_p = variance empirique de x_p entre groupes (b = between), puisque les surfaces de régression de x_p sur les x_i, $i < p$ dans les groupes sont $x_p^{*j} = \overline{x}_p^j$, et la surface de régression pour l'ensemble est $x_p^* = \overline{x}_p$

et $$\frac{U_p}{U_{p-1}} = \frac{w_p}{t_p} \quad \text{puisque} \quad W(p) = \left\| \begin{matrix} W(p-1) & 0 \\ 0 & w_p \end{matrix} \right\|$$

et $$T(p) = \left\| \begin{matrix} T(p-1) & 0 \\ 0 & t_p \end{matrix} \right\|$$

avec $t_p = w_p + b_p$.

$\frac{b_p}{w_p}$ ou $\frac{w_p}{t_p}$ caractérisent le "pouvoir discriminant" de la variable x_p considérée isolément, et le fait qu'il s'agit du test d'un pouvoir discriminant additionnel intervient dans le 1[er] test par l'intermédiaire de p dans l'expression de G(p) et de sa distribution sous H_1 ; de même dans le deuxième test le seuil imposé à $\frac{w_p}{t_p}$ dépend de p, et même de U(p), et non pas seulement de k et de n comme dans le test d'égalité des moyennes μ_p^j de x_p dans les groupes. Rappelons qu'on procéde à ce test en formant le rapport $G(x_p) = \frac{b_p}{w_p} \times \frac{n-k}{k-1}$ qui sous l'hypothèse nulle a une distribution $F(k-1, n-k)$).

Remarque concernant le 1[er] test.

On a $G(x_p) > G(p)$ tandis que pour un même risque α :

$$F_\alpha(k-1, n-k) < F_\alpha\,[k-1, n-(k+p-1)]$$

Donc, si x_p n'a pas de pouvoir discriminant significatif au risque α [i. e. si

$$G(x_p) < F_\alpha(k-1, n-k)],$$

x_p n'a pas de pouvoir discriminant "additionnel" significatif au même risque

$$[G(p) < G(x_p) \leqslant F_\alpha(k-1, n-k) < F_\alpha\,[k-1, n-(k+p-1)]]),$$

tandis que inversement, x_p peut avoir un pouvoir discriminant significatif sans que son pouvoir discriminant additionnel le soit. On peut avoir

$$G(x_p) > F_\alpha(k-1, n-k), \text{ mais } G(p) < F_\alpha[k-1, n-(k+p-1)].$$

Application à l'A.D. stepwise

On suppose que les distributions dans les groupes sont normales et de même covariance, et que les probabilités a priori et les coûts des erreurs de classement sont tous égaux.

Il existe des méthodes d'introduction successive des variables x_j qui consistent, après avoir introduit la 1ère variable, disons par exemple x_1, à introduire successivement les autres en fonction du niveau de signification de leur apport *conditionnel aux variables déjà introduites* pour discriminer entre les θ_i, i.e. en fonction du niveau de signification de la régression multiple dans les groupes (ou de la corrélation multiple ce qui revient au même) de la variable candidate par rapport aux variables déjà introduites.

Pour la 1ère variable à introduire, l'esprit de la méthode ci-dessus conduit à introduire celle qui donne la plus grande discrimination entre les groupes, i. e. entre leurs moyennes puisque c'est la seule chose en quoi ils diffèrent.

C'est donc la variable qui donne le plus haut niveau de signification au test d'égalité des moyennes de groupes qui sera introduite en premier, i . e. celle pour laquelle le rapport b/w sera le plus grand (1) sous réserve que le pouvoir discriminant de cette variable caractérisé par son rapport $G = \frac{b}{w}\frac{n-k}{k-1}$ soit supérieur au seuil "F pour entrer" fixé.

La deuxième variable introduite sera celle dont le rapport G(p) avec $p = 2$ sera le plus grand, sous réserve que ce rapport G(2) soit supérieur au seuil "F pour entrer". Il est équivalent, on vient de le voir, de dire que c'est celle qui conduit à la plus forte décroissance de la statistique U, c'est-à-dire celle pour laquelle $\frac{U(2)}{U(1)}$ est le plus petit.

La méthode la plus utilisée est la méthode de l'A.D. stepwise analogue à la méthode de régression stepwise (7), dans laquelle une variable après avoir été introduite à une étape j peut être exclue à une étape $j' \geqslant j + 2$ (en fait le principe du test montre que $j' \geqslant j + 3$) et qui fait l'objet du programme BMD 07 M. Elle conduit à introduire un minimum de variables discriminantes.

A chaque étape la statistique U de Wilks est calculée, ainsi que la fonction H(U) qu'on en déduit et qui a sous l'hypothèse nulle d'égalité des moyennes de groupes une distribution sensiblement de Fisher-Snedecor.

A certaines étapes spécifiées et après la dernière étape, on donne les fonctions discriminantes, et pour chaque observation, son classement, sa probabilité a posteriori de provenir de chacun des k groupes et le carré de sa distance de Mahalanobis à chaque groupe (cf. 2.1).

---

(1) Il est équivalent de minimiser le rapport : $u = \frac{w}{t}$ puisque $t = w + b$.

DEUXIEME PARTIE

L'ANALYSE FACTORIELLE DISCRIMINANTE, OU RECHERCHE DES VARIABLES CANONIQUES POUR LA DISCRIMINATION ([2], [13])

1 – INTRODUCTION

Au lieu de procéder directement au classement de $\underline{x}(e)$, il peut être intéressant de chercher les fonctions linéaires de $\underline{x}$ qui sont les plus discriminantes, ce qui permettra en prenant pour variables ces fonctions au lieu des variables initiales x_i d'avoir une meilleure visualisation des groupes G_i dans $\mathbf{R}^p$ ou si possible dans un espace de dimension moindre, par exemple dans $\mathbf{R}^2$ ou dans $\mathbf{R}^3$.

On se propose donc de chercher un axe Δu, ou plus généralement un sous-espace de petite dimension de $\mathbf{R}^p$ engendré par une famille d'axes Δu_i qu'il convient de définir, tel que pour une dimension fixée du sous-espace, la projection sur ce sous-espace du nuage des individus de E réalise la meilleure discrimination entre les k groupes.

La représentation sur ce sous-espace sera plus parlante si les axes Δu_i sont tels que les projections $z_i = u_i'\underline{x}$ sur ces axes soient non corrélées.

Bien que l'analyse factorielle discriminante soit essentiellement une méthode de visualisation donc d'analyse des données, on peut là-encore envisager soit des critères de type métrique, soit des critères de type statistique pour choisir l'axe Δu et plus généralement l'ensemble des axes factoriels discriminants Δu_i sur lesquels on désire projeter le nuage. En fait, comme on va le voir, les différents critères adoptés sont tous équivalents.

2 – DEFINITION PAR DES CRITERES DE TYPE METRIQUE

Dans une telle optique on ne s'intéresse on le sait qu'au seul échantillon $\mathcal{E}$ considéré comme population totale.

2.1 – Choix du premier axe factoriel discriminant Δu

Il est naturel de choisir Δu de telle façon qu'en projetant sur Δu, on écarte au maximum les groupes les uns des autres, en même temps qu'on regroupe au maximum les éléments de chaque groupe.

Avec la distance barycentrique euclidienne entre points et groupes définie en I.1.1.([1]), ceci conduit à choisir Δu tel que les projections des centres de gravité $\underline{g}_i$ soient aussi éloignées que possible du centre de gravité $\underline{g}$ de $\mathcal{E}$, tandis que les projections des éléments de chaque groupe G_i^* sont les plus proches possible de la projection de leur centre de gravité $\underline{g}_i$.

----

(1) I renvoie à la première partie

Si u est un vecteur unitaire de Δu, la projection de $\underline{x}$ sur Δu est la variable $z = u'x$, et il paraît naturel de choisir Δu de façon à maximiser pour l'échantillon $\mathcal{E}$ et la variable $z = u'\underline{x}$, le rapport

$$R_{1(u)} = \frac{\text{Somme des carrés des distances entre groupes}}{\text{Somme des carrés des distances dans les groupes}} = \frac{S_B(u)}{S_W(u)}$$

(B = between, W = within), la distance considérée sur Δu étant la distance barycentrique euclidienne classique entre points et groupes déjà mentionnée en I.2.1.

On rappelle que pour une telle distance, le carré de la distance d'un groupe G_i^* à l'ensemble $\mathcal{E}$ des groupes est $n_i D^2(\underline{g_i}, \underline{g})$ où $\underline{g}$ est le centre de gravité de $\mathcal{E}$ en sorte que

$$S_B = \sum_{i=1}^{k} n_i D^2(\underline{g_i}, \underline{g}).$$

En posant :

S_T = somme des carrés des distances dans l'échantillon total

$= \sum_{\underline{x} \in \mathcal{E}} D^2(\underline{x}, \underline{g})$ = inertie totale du nuage $\mathcal{E}$ par rapport à son barycentre, on a d'après la formule classique de Huyghens pour les moments d'inertie :

$$(1) \qquad S_T = S_W + S_B ,$$

puisque S_W est la somme des inerties de chacun des groupes G_i^* par rapport à son centre de gravité $\underline{g_i}$.

En raison de l'équation (1), il revient au même de maximiser le rapport

$$R_2(u) = \frac{S_B(u)}{S_T(u)} \quad \text{puisque} \quad \frac{1}{R_2} = \frac{1}{R_1} + 1, \quad \text{soit} \quad R_2 = \frac{R_1}{1 + R_1}$$

Calcul de $R_1(u)$ *et* $R_2(u)$

On a :

$$S_T(u) = \sum_{x \in \mathcal{E}} [z(\underline{x}) - \overline{z}]^2 \quad \text{si } \overline{z} = u'\underline{g}$$

soit : $$S_T(u) = \sum_{x \in \mathcal{E}} u'(\underline{x} - \underline{g})(\underline{x} - \underline{g})'u$$

$$= n\, u'Tu$$

$$S_W(u) = \sum_{i=1}^{k} \sum_{x \in G_i^*} (z(x) - \overline{z}_i)^2 = \sum_{i=1}^{k} \sum_{\underline{x} \in G_i^*} u'(\underline{x} - \underline{g_i})(\underline{x} - \underline{g_i})'u$$

avec $$\overline{z}_i = u'g_i$$

soit $$S_W(u) = n\, u'Wu$$

et $$S_B(u) = \sum_{i=1}^{k} n_i(\overline{z}_i - \overline{z})^2 = n\, u'Bu$$

en sorte que :

$$R_1(u) = \frac{u'Bu}{u'Wu} \quad \text{et} \quad R_2(u) = \frac{u'Bu}{u'Tu}$$

où T, W et B désignent respectivement les matrices d'inertie (ou matrices de covariances empiriques) dans l'échantillon total $\mathcal{E}$, dans les groupes et entre groupes.

Le premier axe factoriel discriminant est donc une direction Δu portant les vecteurs u de $\mathbf{R}^p$ tels que $R_1(u) = \frac{u'Bu}{u'Wu}$, ou ce qui est équivalent $R_2(u) = \frac{u'Bu}{u'Tu}$, soit maximum. ($R_1(k\,u) = R_1(u)$ si k est un scalaire non nul en sorte que u est défini à une constante scalaire près).

La recherche d'une direction Δu, rendant maximum le quotient $R_1(\underline{u})$ d'une forme quadratique non négative $u'B\underline{u}$ par une forme quadratique définie positive $u'W\underline{u}$, est un problème classique d'algèbre linéaire qu'on rencontre notamment dans l'analyse en composantes principales par rapport à la métrique définie par W et qu'on peut ramener au cas où W = I, c'est-à-dire à l'analyse en composantes principales classique ([12], [13]).

On sait en effet ([12], [13]) qu'on peut toujours écrire W sous la forme W = C'C où C est régulière puisque W l'est.

On a alors

$$R_1(u) = \frac{u'B\underline{u}}{u'W\underline{u}} = \frac{u'B\underline{u}}{u'C'C\underline{u}}$$

Si on pose $\underline{v} = C\underline{u}$ soit $\underline{u} = C^{-1}\underline{v}$,

il vient

$$R_1(u) = \frac{v'C'^{-1}BC^{-1}\underline{v}}{v'\underline{v}}$$

et on est ramené à la recherche de la direction $\Delta\underline{v}$ rendant maximum le rapport $R'_1(v) = \frac{v'A\underline{v}}{v'\underline{v}}$ avec $A = C'^{-1}BC^{-1}$, symétrique non négative, c'est-à-dire au problème de la recherche du premier axe principal en analyse en composantes principales classique.

On sait que la direction Δv_1 est définie par un vecteur propre de A associé à sa plus grande valeur propre, soit λ_1 et que $R'_1(v_1) = \lambda_1$ – (comme A est symétrique et non négative toutes ses valeurs propres sont positives ou nulles et λ_1 est donc bien définie). Δv_1 est unique si λ_1 est valeur propre simple ce qui est généralement le cas.

A $\underline{v}_1$ correspond $\underline{u}_1 = C^{-1}\underline{v}_1$, et l'axe Δu_1 qui le porte est unique si λ_1 est valeur propre simple, et rend maximum le quotient $R_1(\underline{u})$, qui vaut alors λ_1.

Détermination pratique de Δu_1

Il n'est pas utile de déterminer explicitement C et A pour déterminer $\underline{u}_1$ et λ_1, et plus généralement les autres valeurs propres λ_i et les autres vecteurs $\underline{u}_i = C^{-1}\underline{v}_i$ associés aux vecteurs propres de A.

En effet les $\underline{v}_i$ et λ_i étant définis par $A\underline{v}_i = \lambda_i\underline{v}_i$, soit $C'^{-1}BC^{-1}\underline{v}_i = \lambda_i\underline{v}_i$, on obtient par prémultiplication par C^{-1} puisque $W = C'C$: $W^{-1}B\underline{u}_i = \lambda_i\underline{u}_i$, ce qui montre que les λ_i sont aussi les valeurs propres de $W^{-1}B$ et les $\underline{u}_i$ les vecteurs propres correspondants.

Ce résultat pouvait être obtenu directement en annulant les dérivées de $R_1(\underline{u})$ par rapport aux composantes u^j, $j \in 1, 2, \ldots, p$ de $\underline{u}$, mais $W^{-1}B$ n'étant pas symétrique, on ne pouvait affirmer que ses valeurs propres soient réelles.

On a donc obtenu le résultat suivant :

Théorème 1

Le premier axe factoriel discriminant Δu_1 est porté par un vecteur propre $\underline{u}_1$ de $W^{-1}B$ associé à la plus grande valeur propre λ_1 de cette matrice, dont toutes les valeurs propres sont rélles et positives ou nulles et $R_1(u_1) = \lambda_1$.

Un raisonnement analogue fait à partir du rapport $R_2(u)$, permet d'affirmer que la matrice $T^{-1}B$ a toutes ses valeurs propres réelles et positives ou nulles, sa plus grande valeur propre ν_1 étant telle que $\frac{1}{\nu_1} = \frac{1}{\lambda_1} + 1$, et les vecteurs propres associés à cette valeur propre sont les mêmes que ceux de $W^{-1}B$ associés à λ_1.

2.2. – Définition des axes factoriels discriminants d'ordre supérieur à un

Le mode de raisonnement qui vient d'être fait conduit naturellement à définir les axes factoriels discriminants d'ordre i supérieur à 1, comme ceux qui portent les vecteurs propres u_i de $W^{-1}B$ associés aux valeurs propres successives λ_i, $i \in 1, 2, \ldots, p$, rangées par ordre décroissant, le nombre de vecteurs propres associés à une valeur propre multiple étant égal à son ordre de multiplicité h, ce qui pose le problème du choix de ces vecteurs propres dans le sous-espace propre de dimension h correspondant.

Il est facile de donner une définition géométrique concrète de ces axes en partant de celle des $\underline{v}_i$.

En effet les v_i correspondant à deux valeurs propres distinctes sont deux à deux orthogonaux et par suite deux à deux "A orthogonaux" i.e. tels que

$$v_i' A \underline{v}_j = 0 \qquad \text{si} \qquad \lambda_i \neq \lambda_j.$$

En effet

$$v_i' A v_j = v_i' \lambda_j v_j = \lambda_j v_i' \underline{v}_j = 0.$$

A une valeur propre λ_i multiple d'ordre h, on associera donc encore un ensemble de h vecteurs propres v_{i_h} deux à deux orthogonaux et par suite "A orthogonaux" (cf. démonstration ci-dessus).

Dans ces conditions les valeurs propres successives λ_i, indicées de la même façon que les vecteurs propres auxquels elles correspondent, représentent les maximums successifs de $R_1'(\underline{v})$ en ce sens que $\lambda_i = R_1'(\underline{v}_i) = \underset{v \in E_{i-1}\perp}{\mathrm{Max}} R'(v)$ (1), si E_{i-1} est le sous-espace de $\mathbf{R}^p$ engendré par les $\underline{v}_j$, $j < i$.

Les vecteurs $u_i = C^{-1} v_i$, qui sont les vecteurs propres de $W^{-1}B$ associés aux valeurs propres λ_i, sont alors deux à deux "W orthogonaux" puisque

$$u_j' W u_i = v_j' v_i ,$$

et aussi "Borthogonaux" puisque

$$W^{-1} B \underline{u}_i = \lambda_i \underline{u}_i \Rightarrow B \underline{u}_i = \lambda_i W \underline{u}_i \quad \text{et} \quad \underline{u}_j' B \underline{u}_i = \lambda_i \underline{u}_j' W \underline{u}_i = 0 \quad \text{si} \quad i \neq j,$$

et les valeurs propres λ_i qui leur correspondent sont les maximums successifs de $R_1(\underline{u})$ en ce sens que $\lambda_i = R_1(\underline{u}_i) = \underset{u \in F_{i-1} \overset{W}{\perp}}{\mathrm{Max}} R_1(\underline{u})$, si F_{i-1} est le sous-espace de $\mathbf{R}^p$ engendré par les $\underline{u}_j$, $j < i$. $\overset{W}{\perp}$ signifie "complément orthogonal pour la métrique W". Puisque les u_i sont deux à deux W orthogonaux et B orthogonaux à la fois, il sont aussi deux à deux T orthogonaux puisque $T = W + B$.

Le choix des axes factoriels discriminants ne met donc pas en question le choix de la métrique W ou T, ou encore B, si B est régulière, choisie pour définir leur orthogonalité, puisqu'ils sont 2 à 2 orthogonaux par rapport à ces trois métriques. Il ne met pas davantage en question le choix du rapport $R_1(u)$ ou $R_2(u)$ à maximiser sous la condition d'orthogonalité avec les axes de numéro inférieur.

En effet on a $F_j \overset{W}{\perp} = F_j \overset{T}{\perp}$ (cf. ci-dessous), en sorte que

$$\underset{u \in F_{i-1} \overset{W}{\perp}}{\mathrm{Max}} R_1(\underline{u}) = \underset{u \in F_{i-1} \overset{T}{\perp}}{\mathrm{Max}} R_1(\underline{u}) = \underset{u \in F_{i-1} \overset{T}{\perp}}{\mathrm{Max}} R_2(\underline{u})$$

ce qui montre que $\underline{u}_i$ est également vecteur propre de $T^{-1}B$ associé à la $i^{\text{ème}}$ valeurs propre ν_i, qui est telle que $\frac{1}{\nu_i} = \frac{1}{\lambda_i} + 1$, soit $\nu_i = \frac{\lambda_i}{1 + \lambda_i}$.

Montrons que $F_j \overset{W}{\perp} = F_j \overset{T}{\perp}$. Il suffit de montrer que si $u_l \in F_j$ c'est-à-dire si u_l est vecteur propre de $W^{-1}B$, $u \in \underline{u}_l \overset{W}{\perp} \Leftrightarrow \underline{u} \in \underline{u}_l \overset{T}{\perp}$.

---

(1) $E\perp$ désigne l'orthogonal (ou complément orthogonal) de E dans $\mathbf{R}^p$.

$$u \in u_1 \overset{W}{\perp} \Rightarrow u \in u_1 \overset{B}{\perp}$$

car

$$W^{-1}\, B\underline{u}_1 = \lambda_1\, \underline{u}_1 \Rightarrow B\underline{u}_1 = \lambda_1\, W\underline{u}_1 \Rightarrow \underline{u}'B\underline{u}_1 = \lambda_1 \underline{u}'W\underline{u}_1 = 0$$

On a donc $F_j \overset{B}{\perp} \supset F_j \overset{W}{\perp}$.

Comme $T = W + B \quad \underline{u}'T\underline{u}_1 = \underline{u}'W\underline{u}_1 + \underline{u}'B\underline{u}_1$, donc $F_j \overset{W}{\perp} \subset F_j \overset{T}{\perp}$

Comme T est régulière puisque W l'est, $F_j \overset{W}{\perp}$ et $F_j \overset{T}{\perp}$ ont même dimension (ils sont tous deux supplémentaires de F_j dans $\mathbf{R}^p$) et on a par suite $F_j \overset{W}{\perp} = F_j \overset{T}{\perp}$.

On arrive donc au résultat suivant.

Théorème 2. – Les définitions suivantes sont équivalentes :

1/ Les axes factoriels discriminants d'ordre 1, 2, ..., p sont ceux qui portent des vecteurs propres $\underline{u}_1, \underline{u}_2, \dots, \underline{u}_p$ de $W^{-1}B$ associés aux valeurs propres successives λ_i, $i \in 1, 2, \dots, p$, rangées par ordre décroissant, le nombre de vecteurs propres associés à une valeur propre multiple d'ordre h étant h et ceux-ci étant choisis deux à deux W orthogonaux.

On a $R_1(u_i) = \dfrac{\underline{u}_i'B\underline{u}_i}{\underline{u}_i'W\underline{u}_i} = \lambda_i$, et les $\underline{u}_i$ sont aussi B orthogonaux et T orthogonaux.

2/ Les axes factoriels discriminants d'ordre 1, 2, ..., p sont ceux qui portent des vecteurs $\underline{u}_1, \underline{u}_2, \dots, \underline{u}_p$ de $\mathbf{R}^p$ tels que :

$R_1(\underline{u}_1)$ réalise un maximum de $R_1(\underline{u}) = \dfrac{u'B\underline{u}}{u'W\underline{u}}$

$\underline{u}_2$ soit W orthogonal à $\underline{u}_1$ et réalise sous cette condition un maximum de $R_1(\underline{u})$

$\underline{u}_3$ soit W orthogonal à $\underline{u}_1$ et à $\underline{u}_2$ et réalise sous cette condition un maximum de $R_1(\underline{u})$, etc.

... $\underline{u}_p$ soit W orthogonal à $\underline{u}_1, \underline{u}_2, \dots \underline{u}_{p-1}$ et réalise sous cette condition un maximum de $R_1(\underline{u})$.

Les axes factoriels discriminants sont aussi B orthogonaux et T orthogonaux.

1_T/ et 2_T/ : mêmes définitions en remplaçant dans 1/ et 2/ : W orthogonaux par T orthogonaux.

1'/ Même définition que 1/ en remplaçant $W^{-1}B$ par $T^{-1}B$, λ_i par ν_i, W par T puis : on a $R_2(u_i) = \dfrac{\underline{u}_i'B\underline{u}_i}{\underline{u}_i'T\underline{u}_i} = \nu_i$ et les $\underline{u}_i$ sont aussi B orthogonaux et T orthogonaux.

2') Même définition que 2/ en remplaçant $R_1(\underline{u})$ par $R_2(\underline{u}) = \dfrac{\underline{u}'B\underline{u}}{\underline{u}'T\underline{u}}$ et W par T.

Remarques

1/ *Signification géométrique concrète des axes factoriels discriminants*

La définition 2/ conduit à la recherche de directions conjuguées communes aux familles d'hyperellipsoïdes $\underline{u}'B\underline{u}$ = Cte et $\underline{u}'W\underline{u}$ = Cte c'est-à-dire à la recherche de directions qui sont à la fois B orthogonales et W orthogonales ; ces directions sont alors aussi T orthogonales puisque $T = W + B$, ce qui montre immédiatement l'équivalence entre 2/ et 2').

2/ B est de rang au plus égal à $k - 1$ et généralement strictement égal à $k - 1$, en sorte que $W^{-1}B$ et $T^{-1}B$ sont de rang $r \leqslant \mathrm{Min}\ (p, k - 1)$.

En général on a $k \leqslant p$ et par suite $r \leqslant p$. Les dernières valeurs propres de $W^{-1}B$ et $T^{-1}B$ sont donc nulles et les rapports R_1 et R_2 correspondants à des vecteurs propres associés sont nuls.

En effet les centres de gravité $\underline{g}_i$ sont dans une variété affine de dimension $k - 1$ au plus qui contient $\underline{g}$, et leurs projections sur un sous-espace de $\mathbf{R}^p$ orthogonal à cette variété sont confondues, et confondues avec celle de $\underline{g}$ ce qui donne une distance entre groupes nulle sur ce sous-espace, et par suite sur tout axe de ce sous-espace de dimension $p - (k - 1)$.

Par suite quand $p > k - 1$, il est inutile de rechercher plus de $k - 1$ axes factoriels discriminants et variables "canoniques" associées. Ces axes factoriels discriminants sont W orthogonaux à l'orthogonal de la variété affine engendrée par les centres de gravité $\underline{g}_i$.

3/ Du point de vue pratique il est plus commode de rechercher les maxima successifs de $R_2(\underline{u})$ car on a de par sa définition, $R_2(\underline{u}) \leqslant 1\ \forall\ u$. Les valeurs propres ν_i de $T^{-1}B$ sont donc au plus égales à 1. C'est la raison pour laquelle on *caractérise le pouvoir discriminant d'un axe* $\Delta\underline{u}_i$ *par la valeur propre* $\nu_i = R_2(\underline{u}_i)$ *associée, de préférence à* λ_i. ($\nu_i = 1$ correspond à une discrimination parfaite, c'est-à-dire au cas où, en projection sur Δu_i, tous les points d'un même groupe G_i^* sont confondus avec leur centre de gravité et ceci pour tous les groupes).

On verra plus loin la signification statistique des ν_i.

2.3 – Etude des variables canoniques, associées aux axes factoriels discriminants

2.3.1 – Définition adoptée

Une variable discriminante associée à l'axe Δu_i est de la forme $z_i = u_i'\underline{x}$ où $\underline{u}_i$ est un vecteur porté par $\Delta\underline{u}_i$.

Pour que z_i soit la projection orthogonale de x sur $\Delta\underline{u}_i$, il est nécessaire que $\underline{u}_i$ soit de norme 1, soit $\underline{u}_i'\ \underline{u}_i = 1$, mais toutes les variables proportionnelles à z_i ont le même pouvoir discriminant $R_1(u_i)$ ou $R_2(u_i)$ que z_i.

Il peut en particulier être intéressant de définir des variables discriminantes dont la variance empirique dans les groupes, ou totale soit égale à 1. C'est souvent à de telle variables qu'on réserve la dénomination de variables canoniques.

Dans la suite de cette note on désignera par le terme : i^e variable canonique toute variables $z_i = u_i'\underline{x}$ où u_i est un vecteur porté par le i^e axe factoriel discriminant.

$\underline{u}_i'$ est alors le "facteur" associé à cette variable (comme en analyse en composantes principales, le terme de "facteur" est réservé aux formes linéaires sur R^p c'est-à-dire aux éléments du dual R^{p*} de R^p constitué par les applications linéaires de R^p dans R).

La projection orthogonale de $\underline{x}$ sur Δu_i sera désignée sous le terme de i^e composante canonique de $\underline{x}$, et on utilisera si besoin est les termes de "variable normée dans les groupes, ou dans l'échantillon".

2.3.2 – Propriétés des variables canoniques

Si z_i et z_j sont deux variables canoniques, on a en désignant par cov_W, cov_T, cov_B la covariance empirique: dans les groupes, totale, ou entre groupes :

$$cov_W(z_i , z_j) = u_j'Wu_i \quad , \quad cov_T(z_i , z_j) = u_j'Tu_i \quad , \quad cov_B(z_i , z_j) = u_j'Bu_i$$

Puisque les u_i sont deux à deux W orthogonaux, T orthogonaux et B orthogonaux, on en conclut que *deux variables canoniques d'ordres i et j distincts sont non corrélées à la fois dans les groupes, dans l'échantillon total et entre groupes.* On obtient bien ainsi la non-corrélation souhaitée dans l'introduction pour les variables canoniques.

Remarquant que des conditions telles que $cov_W(z_i , z_j) = 0$ et $u_j'Wu_i = 0$ sont équivalentes, on déduit du théorème 2 de 2.2, les définitions équivalentes suivante des variables canoniques.

Théorème 3. – Les définitions suivantes sont équivalentes et équivalentes à celles du théorème 2.

1_W^G) Une première variable canonique z_1, est, parmi les fonctions linéaires $z = u'\underline{x}$ de $\underline{x}$, une variable qui maximise le rapport

$$R_1(u) = \frac{\text{variance de z entre groupes}}{\text{variance de z dans les groupes}}$$

les variances dont il s'agit étant les variances empiriques.

Une deuxième variable canonique z_2 est, parmi les fonctions linéaires $z = \underline{u}'x$ de $\underline{x}$ non corrélées dans les groupes avec z_1, une variable qui maximise le rapport $R_1(u)$ etc. . .

Une p^e variable canonique z_p est, parmi les fonctions linéaires $z = u'\underline{x}$ de x non corrélées dans les groupes avec $z_1, z_2, \ldots, z_{p-1}$, une variable qui maximise le rapport $R_1(u)$.

Les variables z_i sont également deux à deux non corrélées dans l'échantillon et entre groupes.

1^G_T) Même définition en remplaçant $R_1(u)$ par

$$R_2(u) = \frac{\text{variance de z entre groupes}}{\text{variance de z totale}}$$

et permutant "dans les groupes" et dans "l'échantillon".

Remarques

1/ On peut aussi n'opérer dans 1^G_W qu'une seule des deux modifications ci-dessus.

2/ Ce sont bien entendu les définitions du théorème 3 qui justifient a posteriori le choix des axes factoriels discriminants d'ordre supérieur à 1.

Les variables canoniques ont donc une signification de statistique descriptive très concrète et si l'on n'est pas spécialement attaché aux caractères x_i, le remplacement de x par $z \in \mathbf{R}^p$ de composantes $(z_1, z_2, \ldots, z_p)$ conduit à des matrices W_z, T_z et B_z diagonales, ce qui simplifie notablement l'analyse discriminante elle-même et les critères ou tests qu'on peut vouloir lui associer (cf. 1ère partie). *On n'oubliera pas si* $p > k - 1$, *qu'il suffit de se limiter aux* $k - 1$ *premières variables canoniques puisque les autres ont un pouvoir discriminant nul.*

Par ailleurs si on représente le nuage dans un espace de dimension $s \leqslant p$, par un ensemble de s variables canoniques z_i, on notera qu'on peut par un choix convenable des facteurs u_i', choisir la variance dans les groupes, ou la variance totale des z_i. On a en effet $\mathrm{Var}_W(z_i) = u_i'W\underline{u}_i$ et $\mathrm{Var}_T(z_i) = \underline{u}_i'T\underline{u}_i$ ce qui permet de choisir les z_i de façon à ce qu'ils soient normés dans les groupes, ou dans l'échantillon, en choisissant les u_i de norme 1 soit pour la métrique W, soit pour la métrique T.

A la distance de Mahalanobis de métrique W^{-1} ou T^{-1} dans l'espace des variables initiales x_i, correspondra ainsi la distance euclidienne classique de metrique I dans l'espace des z_i.

On pourra ainsi avoir une représentation des groupes G_i^* dans les premiers plans factoriels discriminants qui sont ceux qui les séparent le mieux, et se faire une idée de la discrimination obtenue.

C'est notamment ce que fait le B M D 07 M qui reporte sur un diagramme plan les valeurs pour tous les individus de $\mathcal{E}$ des deux premières variables canoniques normées dans les groupes, puisque la technique probabiliste classique adoptée revient à adopter la métrique W^{-1} pour x (cf. I. 2.2.4, 2^e). Il determine les λ_i , $i = 1, 2, \ldots, p$, associés qui sont alors égaux aux variances des z_i entre groupes, et les coefficients de corrélation canonique

$$\rho_i = \sqrt{\nu_i} = \sqrt{\frac{\lambda_i}{1 + \lambda_i}}\ , \ i = 1, 2 \text{ (ce terme sera justifié en 4.2).}$$

Pour juger de la qualité de la représentation du nuage des G_i^* par chacune des deux premières variables canoniques et par leur ensemble, il peut être intéressant de calculer les rapports $\dfrac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} = \dfrac{\lambda_i}{\text{trace}\,(W^{-1}B)}$ pour $i = 1$ et 2 et leur somme.

Ces rapports représentent en effet les proportions du carré de la distance entre groupes de l'échantillon $\mathcal{E}$, représenté par l'ensemble $\underline{z} \in R^p$ de ses variables canoniques z_i, expliquées par chacune des deux premières variables et par leur ensemble, puisque les variables canoniques z_i, $i = 1, 2, \ldots, p$ ont été choisies de façon à ce que chacune explique la même fraction du carré de la distance dans les groupes, ou si on préfère, porte la même inertie égale à 1 dans les groupes.

2.4 – Lien entre l'analyse factorielle discriminante et une analyse en composantes principales du nuage des centres de gravité $\underline{g}_i$ des groupes de l'échantillon affectés de poids égaux aux effectifs n_i de ces groupes ([2], [13])

La matrice d'inertie du nuage des $\underline{g}_i$ est B, et si on adopte une métrique euclidienne définie par une matrice M, on sait (cf.[12],[13]) que les axes principaux du nuage sont définis par les vecteurs propres $\underline{v}_i$ de la matrice B M ; les facteurs correspondants $f_i' = v_i' M$, c'est-à-dire les formes linéaires de projection M orthogonale sur ces axes : ($\underline{x}$ a pour projection $\underline{v}_i'$ Mx si $\underline{v}_i$ est de norme 1) sont définis par les vecteurs $\underline{f}_i = M\underline{v}_i$ qui sont les vecteurs propres associés aux mêmes valeurs propres de la matrice M B et de norme 1 pour la métrique duale M^{-1}. En effet $v_i'M\,\underline{v}_i = 1 \Leftrightarrow f_i'M^{-1}\underline{f}_i = 1$, et plus généralement $\underline{v}_i'\,M\underline{v}_j = f_i'M^{-1}\underline{f}_j$ ce qui montre que puisque les $\underline{v}_i$ sont deux à deux M orthogonaux. les $\underline{f}_i$ sont deux à deux M^{-1} orthogonaux.

Si on prend pour M l'une des deux matrices T^{-1} ou W^{-1} on obtient donc pour "facteurs" de l'analyse en composantes principales du nuage des $\underline{g}_i$ avec les métriques T^{-1} ou W^{-1} les facteurs ou formes linéaires discriminantes u_i' associées aux variables canoniques normées dans l'échantillon ou dans les groupes (on a vu que $\text{Var}_T\, z_i = u_i' T\underline{u}_i$ et $\text{Var}_W z_i = u_i' W\underline{u}_i$).

La recherche des facteurs discriminants est donc équivalente à la recherche des facteurs de l'analyse en composantes principales du nuage des $\underline{g}_i$ affectés des poids n_i avec l'une des deux métriques T^{-1} ou W^{-1} (métrique de Mahalanobis pour l'échantillon ou dans les groupes). On obtient ainsi selon la métrique utilisée les facteurs associés aux variables canoniques normées dans l'échantillon ou dans les groupes, et ces variables canoniques sont confondues avec les composantes principales correspondantes.

Les pouvoirs discriminants ν_i ou λ_i des z_i sont égaux aux inerties expliquées par les axes factoriels correspondants.

Remarque – Cette équivalence des deux problèmes rend immédiate la remarque 3) de 2.2 concernant le nombre maximum égal à Min(p , k − 1) de variables discriminantes ayant un pouvoir discriminant non nul.

Application : Recherche de la variable canonique dans le cas où il n'y a que deux groupes

Puisque K = 2, il n'y a que la première variable canonique qui ait un pouvoir discriminant non nul. Le facteur discriminant correspondant est le premier facteur principal dans l'analyse en composantes principales du nuage des deux centres de gravité $\underline{g}_1$ et $\underline{g}_2$ affectés des poids n_1 et n_2 avec l'une des deux métriques T^{-1} ou W^{-1}.

Comme le premier axe principal du nuage est celui qui porte la droite joignant les centres de gravité, on a $\underline{v} = K(\underline{g}_2 - \underline{g}_1)$ où K est un scalaire tel que $v'T^{-1}v = 1$ ou $v'W^{-1}v = 1$ selon la métrique choisie et le premier facteur correspondant est $u' = v'T^{-1}$ ou $u' = v'W^{-1}$ selon le cas.

La variable canonique est donc

$$z = u'\underline{x} = K(g_2 - g_1)'T^{-1}\underline{x} \quad \text{ou} \quad K(g_2 - g_1)'W^{-1}\underline{x}$$

selon qu'on désire qu'elle soit normée dans l'échantillon ou dans les groupes.

C'est la projection T^{-1} orthogonale, ou W^{-1} orthogonale, de $\underline{x}$ sur la droite joignant les centres de gtavité des deux groupes, ce qui correspond bien à l'intuition.

On trouve pour z une fonction affine de la fonction discriminante $\delta_{1/2}(\underline{x})$ définie en I.2.1 ce qui correspond également à l'intuition : on obtient la même fonction discriminante à partir de $z \in \mathbf{R}$ qu'à partir de $x \in \mathbf{R}^p$.

3 - DEFINITION DES VARIABLES CANONIQUES PAR UN CRITERE DE TYPE STATISTIQUE

Si on considère $X \in \mathbf{R}^p$ comme une variable aléatoire dont la distribution dépend du groupe G_i à laquelle appartient l'individu e^j (cf. I.2.2), $Z = u'\underline{X}$ est pour u fixé une variable aléatoire réelle dont la distribution dépend également du groupe G_1 considéré. On prendra comme première variable canonique Z_1, celle qui discrimine le mieux entre les K groupes. Si on fait l'hypothèse d'une distribution de $\underline{X}$ normale et de même covariance pour les K groupes, il en est de même pour Z et la différence entre les K groupes ne peut porter que sur les moyennes μ^i de X et par suite sur les moyennes ζ^i de Z. De même la deuxième variable canonique sera celle qui, parmi les variables Z non corrélées avec Z_1, donc avec les hypothèses faites indépendante en probabilité de Z_1 pour les K groupes, discrimine le mieux entre les moyennes des k groupes. On définira de façon analogue les variables canoniques d'ordre supérieur à 2.

Il paraît alors naturel puisque les moyennes μ^i sont inconnues d'estimer la première variable canonique par une variable Z_1 qui soit la plus favorable au rejet de l'hypothèse nulle d'égalité des moyennes ζ^i quand on procède à ce test sur l'échantillon $\mathcal{E}$.

On sait que sous les hypothèses faites, ce test est basé sur la statistique $R_1(u) = \dfrac{\text{Variance de Z entre groupes } G_i^*}{\text{Variance de Z dans les groupes } G_i^*}$, et plus la valeur observée pour $R_1(u)$ est grande, plus la probabilité d'erreur α attachée au rejet de l'hypothèse nulle au vu de cette valeur sera faible.

La première variable canonique "estimée" $Z_1 = u'\underline{X}$ est donc une des fonctions linéaires réelles de $\underline{X}$ qui sont les plus défavorables à l'hypothèse d'une même moyenne pour les K groupes quand celle-ci est testée sur l'échantillon $\mathcal{E}$, sous l'hypothèse d'une distribution de X normale et de même covariance dans chaque groupe.

La deuxième variable canonique estimée $Z_2 = u_2'\underline{X}$ est, dans la classe des fonctions linéaires réelles de $\underline{X}$, non corrélées avec Z_1 dans les groupes G_i^* de l'échantillon $\mathcal{E}$, définie par la même condition que Z_1.

On définit de façon analogue les variables Z_i , $i > 2$.

On voit immédiatement que cette définition des variables canoniques et des axes factoriels discriminants associés (axes Δu_i portant les vecteurs $u_i \in R^p$ associés aux Z_i) est équivalente à celle précédemment donnée en 2).

On peut, là encore, remplacer la non corrélation dans les groupes G_i^* par la non corrélation dans l'échantillon $\mathcal{E}$, et on trouve comme conséquence de la définition la non corrélation dans l'échantillon $\mathcal{E}$ (ou dans les groupes G_i^* suivant la définition adoptée) et entre les groupes G_i^*.

Définition des variables canoniques de la population totale E

La définition qu'on vient de donner des variables canoniques "estimées" correspond à la définition suivante des variables canoniques de la population totale E :

La première variable réalise parmi les fonctions linéaires de $\underline{X}$, le maximum de puissance pour le test d'égalité des moyennes déjà mentionné quand il est effectué sur des échantillons d'effectifs n_i, pour les groupes G_i, $i = 1, 2, \ldots K$.

En effet on sait que pour un risque α fixé, la puissance de ce test est fonction croissante du paramètre de non centralité

$$\lambda(Z) = \frac{\frac{1}{n} \sum_{i=1}^{k} n_i (\zeta_i - \bar{\zeta})^2}{\sigma_z^2}$$

où

$$\bar{\zeta} = \frac{\sum_{i=1}^{k} n_i \zeta_i}{\sum_{i=1}^{k} n_i}$$

et σ_z^2 désigne la variance de Z dans les groupes G_i, soit ici $\sigma_z^2 = u'\Sigma u$ ([1]).

$\lambda(Z)$ qui ne dépend des n_i que par l'intermédiaire des rapports n_i/n, n'est autre que le rapport $\dfrac{\text{Variance de Z entre groupes } G_i}{\text{Variance de Z dans les groupes } G_i}$, quand les groupes G_i ont des effectifs respectifs proportionnels aux n_i/n, c'est-à-dire des probabilités a priori q_i égales à n_i/n.

$\lambda(Z)$ est inconnu et on l'estime par sa valeur empirique pour l'échantillon $\mathcal{E}$: $R_1(u)$, qu'on cherche à maximiser à sa place. $R_1(u)$ est un estimateur sans doute biaisé de $\lambda(Z)$, mais en tous cas convergent quand la taille $\mathcal{E}$ de l'échantillon augmente indéfiniment, les fréquences respectives n_i/n des groupes G_i^* restant constantes.

Le pouvoir discriminant d'une variable réelle $Z = u'\underline{X}$, dans E formé de groupes G_i en proportions respectives q_i dans lesquels $\underline{X}$ a même matrice de covariance Σ, est ainsi défini par

$$\lambda(Z) = \frac{\text{Variance de Z entre groupes } G_i}{\text{Variance de Z dans les groupes } G_i} = \frac{\sum_{i=1}^{k} q_i(\zeta^i - \bar{\zeta})^2}{\sigma_z^2}$$

avec $\bar{\zeta} = \sum_{i=1}^{n} q_i \zeta^i$ et $\zeta^i = E_{G_i}(Z)$ moyenne de Z dans le groupe G_i, et la première variable canonique estimée correspond à un maximum du pouvoir discriminant, estimé sur $\mathcal{E}$, ou empirique : $R_1(u)$.

Cette définition du pouvoir discriminant, si elle fait intervenir les probabilités a priori des groupes G_i, n'exige plus que $\underline{X}$ ait une distribution normale, et si la matrice de covariance n'est plus la même dans les groupes, $\lambda(Z)$ conserve un sens concret en remplaçant Σ par sa valeur moyenne $\bar{\bar{\Sigma}} = \sum_{i=1}^{k} q_i \Sigma_i$ pour l'ensemble des groupes. Elle permet également de définir les axes factoriels discriminants et les variables canoniques de la population totale E. Il suffit en effet pour obtenir $\lambda(Z)$ de remplacer W et B de $R_1(u)$ respectivement par Σ ou $\bar{\bar{\Sigma}}$ et $\Sigma_B = \sum_{i=1}^{k} q_i (\underline{\mu}^i - \underline{\mu})(\underline{\mu}^i - \underline{\mu})'$.

$R_2(u) = \dfrac{R_1(u)}{1 + R_1(u)}$ est également le pouvoir discriminant empirique correspondant à la deuxième expression du pouvoir discriminant de Z :

$$\nu(Z) = \frac{\text{Variance de Z entre les groupes } G_i}{\text{Variance de Z dans E}}$$

qui est lié à $\lambda(Z)$ par $\nu = \dfrac{\lambda}{1 + \lambda}$, que la covariance de X soit ou non la même dans les groupes (cf. remarque ci-dessous).

Remarque

La loi de probabilité de X donc de Z dans E, et par suite sa variance, est définie à partir du moment où on a introduit sur les G_i [c'est-à-dire en fait sur $\Theta = \theta_1\{\theta_2, \ldots, \theta_k\}$ (cf. I. 2.2.3)] une mesure de probabilité définie par les probabilités a priori q_i. On a alors $\bar{\zeta} = E_E[Z(e)]$ moyenne de Z sur E et

$$V_E(Z) = E_{e\in E}(Z(e) - \bar{\zeta})^2 = \sum_{i=1}^{k} q_i \, E_{e\in G_i}(Z(e) - \bar{\zeta})^2$$

$$= \sum_{i=1}^{k} q_i \, [\underbrace{E_{e\in G_i}(Z(e) - \zeta^i)^2}_{\sigma_z^2(G_i)} + (\zeta^i - \bar{\zeta})^2]$$

soit si

$$\sigma_z^2(G_i) = \sigma_z^2 \ \forall i, \ V_E(Z) = \sigma_z^2 + \sum_i q_i (\zeta^i - \bar{\zeta})^2 .$$

En généralisant à $\underline{X}$, il vient $\Sigma_T = \bar{\Sigma} + \Sigma_B$ si Σ_T est la matrice de covariance de $\underline{X}$ dans E, d'où l'expression $\nu(Z) = \dfrac{u'\Sigma_B u}{u'\Sigma_T U}$.

4 – PRESENTATION DE L'ANALYSE FACTORIELLE DISCRIMINANTE COMME UNE ANALYSE CANONIQUE PARTICULIERE ([13])

On a déjà dit en I.1.1 qu'on pouvait considérer les k groupes comme correspondant aux k modalités d'une variable qualitative Y, ou Θ, qu'on peut représenter elle-même par un ensemble de k variables booléennes t_i mutuellement exclusives définies par

$$t_i = 1 \quad \text{si} \quad Y = Y_i, \qquad t_i = 0 \quad \text{si} \quad Y \neq Y_i \qquad i = 1, 2, \ldots, k$$

Le problème posé se présente alors comme un problème de prévision des valeurs de l'ensemble t_j, $j = 1, 2, \ldots, k$, à l'aide de fonctions linéaires des x_i, $i = 1, 2, \ldots, p$.

On est donc conduit à prendre pour fonctions linéaires des x_i, les variables canoniques qui leur correspondent dans l'analyse des corrélations canoniques entre le groupe des k variables booléennes t_j et celui des p variables x_i ([13], [14]). On peut là encore comme dans 3), raisonner dans la population totale E, ou chaque individu e est caractérisé par un vecteur $\left\{ \dfrac{\underline{t}}{\underline{x}} \right. \in R^{p+k}$, mais comme on ne sait rien de la distribution de ce vecteur dans E, on se trouvera conduit à estimer comme on le fait généralement les variables canoniques de E par celles de $\mathcal{E}$.

4

Nous allons montrer que les facteurs et les variables canoniques définis en 2) sont effectivement ceux que donne l'analyse canonique des deux groupes de variables t_j et x_i.

4.1 – Préliminaire. Etude de la corrélation canonique entre une variable réelle X et un ensemble de k variables booléennes t_i mutuellement exclusives (attachées à une variable qualitative Y à k modalités)

On sait que pour une population E donnée, une première variable canonique attachée aux t_i est une fonction linéaire de ces variables $X^* = \beta_0 + \sum_{i=1}^{k} \beta_i \ t_i$ dont le coefficient de corrélation avec X soit le plus grand possible(1). Une telle variable n'est donc définie qu'à un coefficient de proportionnalité près, mais on sait que la régression linéaire des moindres carrés de X sur les t_i, c'est-à-dire la variable X* telle que $E_{e \in E} [X(e) - X^*(e)]^2$ soit minimum [X* est la plus proche de X "en moyenne quadratique"] satisfait à la condition de corrélation maximale avec X.

Soit E_i la sous-population de E pour laquelle $t_i = 1$ et e_i un individu de E_i.

$$X^*(e_i) = \beta_0 + \beta_i$$

On est donc amené à chercher la fonction constante sur chaque E_i qui soit la plus proche possible de X en moyenne quadratique. Soit x_i^* la valeur de X^* sur E_i.

Comme $$E_{e \in E} [X - X^*]^2 = \sum_{i=1}^{k} q_i \, E_{e \in E_i} [X - x_i^*]^2$$

si $q_i = \text{Prob}[t_i = 1]$, il suffit de rechercher pour chaque i une constante x_i^* telle que $E_{E_i} (X - x_i^*)^2$ soit minimum.

On sait que c'est la moyenne μ^i de X sur E_i, ou moyenne de X liée par $t_i = 1$, qui réalise ce minimum.

On a donc $X^* = \sum_{i=1}^{k} \mu^i t_i$ fonction constante sur chaque groupe E_i et égale à la moyenne μ^i de X pour ce groupe, comme première variable canonique associée aux t_i, et pour coefficient de corrélation canonique correspondant le coefficient de corrélation linéaire multiple de X sur les t_i, R, défini par

$$R = \frac{\text{cov}(X, X^*)}{\sqrt{\text{Var } X, \text{Var } X^*}}$$

Or $$E_E(X^*) = \Sigma_{E_i} q_i \, \mu^i = \mu \qquad \text{si} \qquad E_E(X) = \mu$$

$$\text{Var } X^* = E[X^* - \mu]^2 = \sum_i q_i(\mu^i - \mu)^2$$

----

(1) Les variables canoniques d'ordre supérieur à 1, sont par définition non corrélées avec X, puisque X est rélle et est par suite sa propre première composante canonique.

et $\quad \text{cov}(X, X^*) = \sum_{i=1}^{k} q_i \, E_{E_i}[X - \mu]\,[\mu^i - \mu] = \sum_i q_i[\mu^i - \mu]^2$

puisque

$$X - \mu = X - \mu^i + \mu^i - \mu$$

et

$$E_{E_i}(X - \mu^i) = 0$$

On a donc

$$R^2 = \frac{\sum_i q_i(\mu^i - \mu)^2}{V(X)} = \frac{\text{Variance de X entre groupes } E_i}{\text{Variance totale de X}}$$

Théorème 4. – Le carré du coefficient de corrélation canonique (ou multiple) entre une variable X et un ensemble de k variables booléennes t_i associées à une variable qualitative Y à K modalités est égal au rapport

$$\frac{\text{Variance entre groupes } E_i}{\text{Variance totale}}$$

pour la variable X, les groupes étant définis par les modalités de Y.

4.2 – L'analyse factorielle discriminante est une analyse canonique particulière

Théorème 5. – Les variables canoniques de l'analyse en composantes canoniques sont les variables canoniques définies en 3) et en 2), et les carrés des coefficients de corrélation canonique correspondants sont les valeurs ν_i du rapport $R_2(u_i)$ [ou de $\nu(Z_i)$ défini en 3] si on se place dans une population totale E plus grande que l'échantillon $\mathcal{E}$].

Si $Z = u'\underline{X}$, le carré du coefficient de corrélation canonique de Z avec les t_i dans E est le rapport $R_2(u)$.

La première composante canonique doit donc selon que l'on se place dans E ou dans $\mathcal{E}$ maximiser le rapport $\nu(Z)$ ou $R_2(u)$, tandis que la deuxième doit parmi les variables non corrélées avec la première dans E ou dans $\mathcal{E}$, maximiser ce même rapport, etc... pour les autres variables.

On retrouve donc bien les variables canoniques définies en 2) pour $\mathcal{E}$ seulement et en 3) pour $\mathcal{E}$ et pour E.

[Le fait que Z_1 réalise un maximum de $\nu(Z)$, entraîne l'équivalence entre la non corrélation avec Z_1 dans E, et dans les groupes G_i. La démonstration donnée en 2.3.2 pour $\mathcal{E}$ s'étend de façon évidente à E en remplaçant W, T et B par Σ, Σ_T et $\Sigma_B = \Sigma_T - \Sigma$].

5 – CONCLUSION

La définition de la première variable canonique comme la fonction linéaire de $\underline{X}$ qui présente la plus forte corrélation multiple avec l'ensemble des variables booléennes t_i associées aux groupes G_i, et des variables suivantes comme celles qui parmi les fonctions linéaires de $\underline{X}$ non corrélées avec les variables canoniques d'ordre inférieur, ont la même propriété, nous paraît très intéressante.

Elle se différencie des deux autres définitions par le fait qu'elle considère Y comme un caractère des individus au même titre que X, ce qui conduit à prendre comme donnée initiale l'échantillon global $\mathcal{E}$ ou la population totale E plutôt que les groupes G_i^* ou G_i comme on a tendance à le faire dans les deux premières définitions. L'exemple d'application ci-dessous montre l'intérêt de cette interprétation.

En dehors du lien ainsi établi entre l'analyse factorielle discriminante et l'analyse canonique, le résultat essentiel de cette longue étude nous paraît être la non corrélation des variables canoniques à la fois dans les groupes et dans l'échantillon, l'une impliquant l'autre, ou de manière équivalente pour les deux métriques définies par les matrices de covariance de $\underline{X}$ dans les groupes et dans l'échantillon, l'une impliquant l'autre. (le mot échantillon peut être remplacé par population totale si c'est elle et ses variables canoniques qu'on considère).

C'est ce résultat qui permet d'adopter pour critère de définition d'une variable canonique d'ordre i : $Z_i = u_i' \underline{X}$, indifféremment la maximisation de l'un des deux rapports $R_1(u) = \dfrac{u'Bu}{u'Wu}$ ou $R_2(u) = \dfrac{u'Bu}{u'Tu}$ sous la contrainte de la non corrélation dans les groupes, ou dans l'échantillon $\mathcal{E}$ avec les variables canoniques d'ordre inférieur.

3ème PARTIE
APPLICATION A UNE EXTENSION DE LA REGRESSION LINEAIRE MULTIPLE AU CAS OU LA VARIABLE A EXPLIQUER EST QUALITATIVE

1 – EXPOSE DU PROBLEME ET PREMIERE SOLUTION

Soit y prenant les modalités $y_1, y_2, \ldots, y_k$ la variable à expliquer et $x = (x_1, x_2, \ldots\ x_p)$ l'ensemble des p variables explicatives réelles.

Le problème qui consiste à déduire de l'observation de $\underline{x}$ sur un individu e, la modalité y(e) correspondante pour y, relève de l'analyse discriminante classique et on a vue en I qu'il faisait intervenir $\dfrac{k(k-1)}{2}$ fonctions discriminantes.

Dans le cas présent ce qu'on cherche, c'est à n'utiliser qu'une seule fonction discriminante, et que cette fonction soit une fonction linéaire $z = u'\underline{x}$ de $\underline{x}$.

Il est alors naturel de prendre pour fonction Z la première variable canonique de l'analyse factorielle discriminante définie par $(\underline{x}, y)$ qu'on estimera sur l'échantillon $\mathcal{E}$ dont on dispose.

On peut caractériser la perte d'information due au fait d'avoir remplacé $\underline{x}$ par z par le rapport $\dfrac{\nu_1}{\sum_{i=1}^{p} \nu_i} = \dfrac{\nu_1}{\text{trace}(T^{-1}B)}$ tandis que la "corrélation" entre z et y sera caractérisée par ν_1 carré du coefficient de corrélation canonique entre z et y (cf. II).

La fonction $z = u'x$ étant ainsi définie, il faut définir une règle de décision δ qui à partir d'une observation z(e) permette de prévoir y. C'est cet ensemble (z, δ) qui constitue la fonction de régression linéaire cherchée. La définition de δ est du domaine de l'analyse discriminante classique, si on désire une règle de decision simple, c'est-à-dire une application de $\mathfrak{Z}$, projection de $\mathcal{X}$ sur Δu, dans $\mathcal{Y}$. Une telle règle de décision est, on l'a vu en I, définie par des régions.

$$R_1, R_2, \ldots R_k$$

de $\mathfrak{Z}$ telles que si $z(e) \in R_j$ la prévision $Y^*(e)$ de Y sera Y_j.

La fonction de régression linéaire cherchée pourra alors être définie par l'ensemble de z et des régions R_i.

Si on utilise un critère de type métrique et une métrique euclidienne, il n'y aura plus de problème de choix de cette métrique puisque $p = 1$. Avec la métrique barycentrique, la valeur de $Y^*(e)$ sera l'indice du centre de gravité le plus proche de z(e). Les régions R_i seront donc séparées par les milieux des segments compris entre deux centres de gravité consécutifs. En ces milieux on pourra choisir arbitrairement entre deux valeurs de Y^*. La règle de décision n'est donc pas simple mais mixte en réalité.

On peut aussi utiliser des critères de type probabiliste et puisque $p = 1$, on peut envisager d'estimer sur l'échantillon non seulement les probabilités a priori q_i, mais aussi les lois de probabilité P_i de Z pour chacun des groupes G_i.

La qualité de la régression ou plutôt de la prévision ainsi définie peut être caractérisée par la perte moyenne ou la probabilité des erreurs de classement.

Remarque

Quand les régions R_i sont des intervalles comme c'est le cas pour le critère basé sur la distance barycentrique euclidienne, ce procédé permet d'ordonner les modalités de Y, par l'ordre des intervalles associés aux valeurs de Y^*.

On peut même aller plus loin et affecter à ces modalités des scores qui seraient les valeurs $\overline{z}^i$ de Z pour les centres des groupes correspondants, ou une fonction affine de ces valeurs. Ces procédés n'ont de valeur que dans la mesure où la prévision de y à partir de z est satisfaisante.

2 – CHOIX D'UNE REGLE DE DECISION ALEATOIRE

Dans de nombreux problèmes pratiques le nombre K des modalités de Y est petit (K = 2 ou 3) et l'échantillon relativement grand de sorte qu'à une même valeur de x peuvent correspondre plusieurs valeurs de y. Quand K = 2(1), les données sont du reste souvent rassemblées sous la forme : proportion des valeurs Y_1 pour chaque valeur observée de x. [On a ainsi la proportion des ménages habitant un immeuble par opposition à ceux qui ont un logement individuel en fonction de caractèristiques telles que la catégorie socio-professionnelle du chef, l'âge du chef, son revenu, le nombre d'enfants, etc. . .]

Le remplacement de x par sa projection orthogonale z sur l'axe Δu ne peut qu'amplifier ce phénomène ; aussi paraît-il peu souhaitable de faire correspondre une seule valeur de Y à une valeur donnée de z.

Il est préférable d'associer à z une distribution de probabilité P_y^z sur $\mathcal{Y}$ c'est-à-dire de définir une règle de décision aléatoire.

Si on se place dans un cadre probabiliste, ce qui paraît s'imposer dans une telle optique, il paraît naturel de prendre pour probabilité P_y^r sur Y la probabilité a posteriori de Y quand z est donné, soit P(Y/z) donnée par l'expression

$$P[Y_i/z] = \frac{q_i p_i(z)}{\sum_{i=1}^{k} q_i p_i(z)}$$

(cf. I, 2.2.3 où θ remplace Y et $\underline{x}$ remplace z).

Remerciements

Je tiens à exprimer mes remerciements les plus chaleureux à MM. P. Bertier et J.M. Bouroche de la Direction scientifique de METRA qui ont réveillé mon intérêt pour l'analyse discriminante et m'ont invitée à faire un exposé sur ce sujet, et à M. J.P. Nakache du C.H.U Pitié-Salpêtrière qui a attiré mon attention sur l'équivalence de la définition des axes factoriels discriminants à partir des rapports $R_1(\underline{u})$ et $R_2(\underline{u})$.

REFERENCES BIBLIOGRAPHIQUES

[1] ANDERSON. – An introduction to multivariate statistical Analysis. Wiley 1958, chapitre 6

[2] ROMEDER J.M. – Méthodes de discrimination
Thèse de doctorat de 3ème Cycle, ISUP, Juin 1969

---

(1) Dans ce cas on sait cf. I et II qu'il n'y a qu'une seule fonction discriminante et que seule la première variable canonique a un pouvoir discriminant non nul.

[3] BOUROCHE J.M. – Classification métrique et segmentation.
Note de travail n° 156 de la direction scientifique de la SEMA, Août 1971

[4] ULMO J. – La décision statistique dans le cadre bayesien
Revue de Statistique Appliquée (1971) Vol 19, n° 3 p. 27/66.

[5] GLICK N. – Sample based classification procedures derived from density estimators.
Journal American Statistical Ass. Vol 67 n° 337 (March 1972) p. 116/122

[6] GHURYE S.C. and OLKIN I. – Unbiased estimation of some multivariate probability densities and related functions
Ann. Math. Stat. Vol 40 n° 4 (1969) p. 1261/71

[7] ULMO J. – Problèmes et programmes de régression
Revue de Statistique Appliquée (1971) Vol 19 n° 1 p. 27/39.

[8] URBAKH. – Linear discriminant analysis : loss of discriminating power when a variate is omitted
Biometrics 27, p. 531/4 (Déc. 1971)

[9] COCHRAN. – On the performance of the linear discriminant functions.
Technometrics 6 (1964) p. 179/90

[10] ULMO J. – Le modèle linéaire multidimensionnel et le critère de Wilks pour tester une sous-hypothèse linéaire
Note de travail n° 176 de la direction scientifique de la SEMA, Juillet 1972

[11] RAO R. – Linear statistical inference and its applications. Wiley, 1965

[12] LEBART L. & FENELON J.P. – Statistique et informatique appliquées.
Dunod, 1971

[13] CAILLIEZ F., NAKACHE J.P., MAILLES J.P., PAGES J.P.
Analyse des données multidimensionnelles,
Centre d'Etudes Economiques d'Entreprises
116 Bd Pereire, Paris 17°, 1971

[14] DEMPSTER A.P. – Elements of continuous multivariate analysis (chap. 8, 9, 10)
Addison Wesley Publishing Company, 1969

[15] TATSUOKA M.M. – Discriminant analysis (1970)
(Présentatation élémentaire de l'analyse factorielle discriminante)
Institute for personality and ability testing
1602 Coronado Drive, CHAMPAIGN, Illinois 61820 U.S.A.

[16] TENENHAUS M. – L'analyse discriminante d'un point de vue bayesien.
Application à la segmentation.
Note de travail n° 3 de la direction scientifique de la SEMA, Février 1970.