LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI

Analyse discriminante et analyse factorielle

Les cahiers de l'analyse des données, tome 2, n° 4 (1977), p. 369-406

<http://www.numdam.org/item?id=CAD_1977__2_4_369_0>

Les Cahiers de l'Analyse de Données
Vol. II - 1977 - n° 4 - p. 369-406

ANALYSE DISCRIMINANTE ET ANALYSE FACTORIELLE [SEP. CORR.]

par J.-P. Benzécri (1)
d'après les thèses de M. Danech Pejouh, T. Moussa et J.-M. Romeder

En analyse factorielle, comme en classification automatique, on prétend sans faire usage d'hypothèses a priori, découvrir sur un tableau de données la structure d'un ensemble. Tout autre est le problème de la discrimination : une partition étant reçue à priori, on cherche sur les données une formule - linéaire ou autre - permettant de calculer la classe à laquelle doit être affecté un individu. Ce point de vue ne nous paraît pas être le premier à envisager : une étude affranchie de toute hypothèse est à la fois plus riche et plus sûre. D'une part à aborder les données sans mêler à leur analyse ce qu'on présume de la structure, on peut découvrir un ordre insoupçonné; d'autre part, lors même que les facteurs sont conformes à ce qu'on en attendait, l'hypothèse est à la fois confirmée et précisée. Cependant, après d'autre méthodes, l'analyse discriminante peut être utile : car parfois, la formule même de discrimination a un intérêt pratique; et, généralement, les résultats, comme ceux de tout traitement systématique des données, aident à connaître celles-ci. Chaque regard porté sur un objet complexe n'en saisit-il pas un nouveau profil ?

Il y a plusieurs années, nous avions dans deux notes (cf Benzécri 1967, 1968) critiqué la pratique de l'analyse discriminante, non sans suggérer quelques recherches. Certaines de nos suggestions furent suivies par J.M. Romeder dans sa thèse (1969). Des comparaisons minutieuses avec l'analyse des correspondances se trouvent dans les thèses de M. Danech-Pejouh et T. Moussa (1972). La présente leçon est un sommaire des principes et des résultats de tous ces travaux.

1. *Le problème de la discrimination* :

Ce problème peut en bref s'énoncer ainsi. Soit dans un espace vectoriel E de dimension d (ou espace des descriptions d'objet) un ensemble I de N points répartis en deux classes q et q' : placer dans E une cloison, par exemple une cloison hyperplane, qui sépare les deux classes q et q'. Presque toujours, ces classes ne se limitent pas aux N individus donnés : I n'est au contraire qu'un échantillon actuel fini d'observations, d'après lequel on doit définir une séparation entre les deux classes, potentiellement infinies. Par exemple soient N sujets, tous atteints d'une même maladie. Sur chacun de ces malades on a fait un ensemble J de mesures (dosages sanguins, pression artérielle, etc...), avant de le soumettre à un traitement (e.g. une opération chirurgicale), le même pour tous. De ce traitement, seule une partie des malades, la classe q, ont bénéficié; les autres, la classe q', n'y ont rien gagné. On désire pouvoir à l'avenir, au seul vu de l'ensemble J de mesures, ou mieux encore d'une partie de ces mesures, déterminer si un malade devra ou non être soumis au traitement considéré (e.g. on opérera les malades affectés à la classe q; mais non ceux affectés à la classe q').

(1) *Professeur de statistique. Université P. et M. Curie*

Plus généralement, on considère un ensemble I d'individus dont chacun est d'une part décrit par un vecteur d'un espace E, et d'autre part rapporté à l'une des classes q(i) d'un ensemble Q. On désire affecter tout individu nouveau s à une classe $q \in Q$ (e.g. formuler un diagnostic), au seul vu du vecteur de description. C'est là un problème de régression particulier : il s'agit en effet de trouver une expression approchée de l'application $i \to q(i)$ de I dans Q par une fonction des composantes du vecteur de description de i. Et, comme il est de règle en statistique, on doit non seulement proposer une solution, mais encore en discuter la validité. (Sur les rapports entre régression et discrimination cf *Note , in fine)*

Ainsi dans le cas de la discrimination entre deux classes nous distinguerons trois questions :

1°) Entre les deux classes actuelles q et q', en lesquelles sont répartis les individus de I, peut-on dans E placer une cloison du type fixé (e.g. hyperplane : il est essentiel de fixer a priori le type de cloison; autrement, à moins que deux individus l'un de q, l'autre de q' ne soient représentés par le même point, il y aura toujours une cloison sinueuse qui séparera q de q' !).

2°) Les deux classes potentielles (infinies : e.g. tous les malades) q et q' peuvent-elles être séparées par une cloison du type fixé; et sinon quel sera le taux minimum d'erreur.

3°) A supposer que (cf. 1°) on ait construit une cloison entre les classes actuelles (finies) q et q', que vaut cette cloison pour séparer les classes potentielles (infinies) que représentent les classes finies?

Pour répondre à ces questions, on a longtemps eu recours à l'hypothèse que toutes les classes (potentielles) sont définies par des distributions normales, multidimensionnelles, égales entre elles à une translation près. Cette hypothèse peu réaliste suggère des constructions intéressantes, qu'il est toutefois plus juste de fonder sur la seule considération des formes quadratiques d'inertie (cf. § 2,3); et qu'on peut généraliser au cas où les classes ne sont pas supposées être de même forme (cf. § 4,5). Quant à la validité des résultats, les épreuves de simulation (cf. § 6) apprennent plus que les méthodes paramétriques; mais un modèle probabiliste relevant de la géométrie intégrale (cf. § 7) donne d'utiles ordres de grandeur. Enfin, nous avons rappelé notre préférence pour les méthodes inductives (analyse factorielle, voire classification automatique) qui ordonnent les données sans recourir à des hypothèses structurelles : lors même qu'une structure a priori doit être retrouvée (c'est le cas considéré ici) ces méthodes sont fécondes (cf. §§ 8,9; et 3.2, où l'on suggère d'utiliser la classification automatique pour ramener toute discrimination à une suite hiérarchisée de dichotomies).

2. *La méthode des moindres carrés* :

2.1. *Variance interclasse et variance intraclasse* :

Soit I un ensemble fini dont on note i, i' etc les individus; soit Q une partition de I dont on note q, q'... les classes; on sait que Q est un ensemble de parties de I satisfaisant aux conditions :

$$Q \subset \mathrm{Part}(I);\ I = \cup \{q \mid q \in Q\};$$

$$\forall\ q, q' \in Q : (q \cap q' = \emptyset) \Leftrightarrow (q \neq q').$$

Pour tout élément i de I on note q(i) l'unique classe de Q à laquelle appartient i. Chaque élément i est affecté d'une masse positive μ_i; on note $\mu_q = \Sigma \{\mu_i \mid i \in q\}$, la masse de la classe q; et M la masse totale $\Sigma \{\mu_i \mid i \in I\}$.

On suppose de plus donné pour chaque individu i un vecteur de description dans un espace vectoriel E. Dans la pratique ce vecteur de description est un système de nombres, ou coordonnées de i; en sorte que E est muni d'une base. On notera $E = R_J$ avec J en indice inférieur, non qu'il faille toujours regarder la description $i_J = \{i_j | j \in J\}$ de i comme une mesure sur un ensemble fini J, mais ce choix qui est celui de l'analyse des correspondances et s'impose dans l'étude d'un tableau de contingence, ne nuit pas dans le cas d'un tableau quelconque de nombres; cf [Repr. Eucl.] TII B n° 2, § 5.2. On note q_J le centre de gravité de la classe q :

$$q_J = (1/\mu_q) \Sigma \{\mu_i \, i_J | i \in q\} \in R_J = E$$

et g_J le centre de gravité du nuage total : $g_J = (1/M) \Sigma \{\mu_i \, i_J | i \in I\}$. Dans les calculs, on écrira parfois i, q, g pour i_J, q_J, g_J.

On définit trois tenseurs σ dans $R_J \otimes R_J = E \otimes E \approx L(E^*, E) = L(R^J, R_J)$, (pour ces notations tensorielles, cf [Note Lim.] TII B n° 1) :

$$\sigma(I)_{JJ} = (1/M) \Sigma \{\mu_i (i_J - g_J) \otimes (i_J - g_J) | i \in I\};$$

$$\sigma(Q)_{JJ} = (1/M) \Sigma \{\mu_q (q_J - g_J) \otimes (q_J - g_J) | q \in Q\};$$

$$\sigma(I - Q)_{JJ} = (1/M) \Sigma \{\mu_i (i_J - q(i)_J) \otimes (i_J - q(i)_J) | i \in I\}.$$

à titre de rappel, explicitons la dernière de ces formules en calculant une composante du tenseur σ :

$$\sigma(I - Q)_{jj'} = (1/M) \Sigma \{\mu_i (i_j - q(i)_j)(i_{j'} - q(i)_{j'}) | i \in I\}$$

Les tenseurs $\sigma(I)$ et $\sigma(Q)$ ne sont autres que les formes quadratiques d'inertie des nuages I et Q ramenés à la masse totale 1. Le tenseur $\sigma(I - Q)$ est la moyenne, pondérée par les μ_q, des formes quadratiques d'inertie $\sigma(q)$ des sous nuages q de I. Si on note :

$$\sigma(q)_{JJ} = (1/\mu_Q) \Sigma \{\mu_i (i_J - q_J) \otimes (i_J - q_J) | i \in q\},$$

on a la relation : $\sigma(I - Q)_{JJ} = \Sigma \{(\mu_q/M) \, \sigma(q)_{JJ} | q \in Q\}$.

Soit $u^J \in R^{J} = E^*$ une forme linéaire sur l'espace R_J où sont les i_J; cette forme définit une fonction sur I : $u^J(i) = u^J \, i_J = \Sigma \{u^j \, i_j | j \in J\}$. La variance de u^J sur I n'est autre que :

$$\sigma(I)_{JJ} \, u^J \, u^J = \Sigma \{\sigma(I)_{jj'} \, u^j \, u^{j'} | j \in J, \, j' \in J\}$$

$$= \Sigma \{(\mu_i/M) \, (u^J(i_J - g_J))^2 | i \in I\}$$

De même, $\sigma(q)_{JJ} \, u^J \, u^J$ est la variance de la forme u^J sur le nuage Q des centres des classes affectés des masses μ_q. Enfin $\sigma(I - Q)_{JJ} \, u^J \, u^J$ est la moyenne pondérée par μ_q des variances de u^J à l'intérieur de chacune des classes q. On appellera en bref $\sigma(I)$ variance totale; $\sigma(Q)$ variance interclasse (entre les classes); et $\sigma(I - Q)$ variance intra-classe (intérieure aux classes).

A titre d'exercice, démontrons, par un calcul dont le principe remonte à Huyghens, que la variance totale $\sigma(I)$ est somme de la variance interclasse $\sigma(Q)$ et de la variance intraclasse $\sigma(I - Q)$. Il suffit d'établir que, pour toute classe q de Q on a :

$$\Sigma \{\mu_i \, {}^2\otimes (i - g) \mid i \in q\} = \mu_q \, {}^2\otimes (q - g) + \Sigma \{\mu_i \, {}^2\otimes (i - q) \mid i \in q\},$$

(où, pour abréger les formules, on note ${}^2\otimes x$ pour : $x_j \otimes x_J$). On a en effet :

$$\begin{aligned} {}^2\otimes (i - g) &= {}^2\otimes ((i - q) + (q - g)) \\ &= {}^2\otimes (i - q) + (i - q) \otimes (q - g) + (q - g) \otimes (i - q) + {}^2\otimes (q - g \end{aligned}$$

En sommant cette égalité sur la classe q les termes rectangles disparaissent car $\Sigma \{\mu_i (i_J - q_J) \mid i \in q\}$ est nul, de par la définition même du centre q_J de la classe q; et il reste l'égalité annoncée.

Voici un autre exercice. Nous proposons, sans démonstration la formule :

$$\sigma(Q)_{JJ} = (1/(2M^2)) \Sigma \{\mu_q \mu_{q'} \, {}^2\otimes (q_J - q'_J) \mid q \in Q, q' \in Q\};$$

(on prendra garde que la somme comprend Card Q. (Card Q - 1) termes non nuls, autant qu'il y a de paires ordonnées q, q'; ces termes sont égaux par paires : le terme qq' est égal au terme q'q). Dans le cas de deux classes, où $Q = \{q, q'\}$; $M = \mu_q + \mu_{q'}$, il vient :

$$\sigma(Q)_{JJ} = (\mu_q \mu_{q'}/M^2) \, (q_J - q'_J) \otimes (q_J - q'_J).$$

2.2. *Métrique d'inertie et discrimination* :

Une forme linéaire qui serait constante à l'intérieur de chaque classe, tout en variant d'une classe à une autre, résoudrait le problème de la séparation. Il n'existe pas en général de telle forme linéaire : mais à défaut d'obtenir que u^J soit de variance nulle sur chaque classe, on demandera que la variance intraclasse ($\sigma(I - Q)$ uu) soit minima : c'est le principe général de la méthode des moindres carrés. Plus précisément, car il ne servirait de rien d'avoir ($\sigma(I - Q)$ uu) nul avec une forme u nulle elle-même, on cherche à rendre minimum le rapport ($\sigma(I - Q)$ uu)/($\sigma(Q)$ uu) de la variance intraclasse (mesure de la dispersion interne des classes) à la variance interclasse (mesure de la séparation des classes); ou, ce qui revient au même puisque $\sigma(I) = \sigma(Q) + \sigma(I - Q)$; on cherche à rendre maximum le rapport ($\sigma(Q)$ uu)/($\sigma(I)$ uu) de la variance interclasse à la variance totale. C'est là un problème classique de décomposition simultanée de deux formes quadratiques. Pour en discuter la solution, il convient de préciser si ces formes quadratiques sont de rang maximum (Card J), ou si l'une ou l'autre est dégénérée.

Dans la pratique, Card I (nombre des individus étudiés), doit surpasser Card J (nombre des paramètres de description). S'il n'en est pas ainsi, l'échantillon I ne révèle pas la dispersion potentielle des diverses classes autour de leur centre; de ce qu'on a pu séparer en classes l'ensemble des cas effectivement étudiés, on ne peut conclure qu'on saura reconnaître à quelle classe rattacher un nouvel individu. (Nous y reviendrons au § 6). Au contraire, souvent, Card Q (nombre des classes à séparer) sera inférieur à Card J; ce qui, à la condition près (formulée ci-dessus) que Card I soit assez grand, semble favorable à la discrimination des classes. Donc d'une part le nuage des i_J a pour suppo: R_J tout entier (n'est inclus dans aucun sous-espace strict de R_J); et par conséquent (cf [Repr. Eucl.] TII n° 2 § 2.1) la forme quadratique $\sigma(I)$ est définie positive; et d'autre part il se peut que le nuage des q_J ait pour support un sous-espace strict de R_J; et que la forme $\sigma(Q)$ soit de rang inférieur à Card J.

La forme quadratique $(\sigma(I)^{-1})$ (forme dont la matrice est inverse de celle de $\sigma(I)_{JJ}$) définit sur R_J une métrique euclidienne pour laquelle le nuage des i_J a l'inertie d'une hypersphère (dont tous les moments d'inertie sont égaux). Cette métrique semble propice à l'étude du nuage I et à la discrimination des classes q; peut aussi être étudiée la métrique $(\sigma(I - Q)^{-1})$; là est la base de plusieurs études renommées (citons Pearson, Fisher, Mahalanobis). On peut d'abord à chaque classe q associer le domaine Eq ensemble des points de R_J qui, pour la métrique $\sigma(I)^{-1}$, sont plus proches de q_J que d'aucun autre centre de classe, q'_J. (Cette construction est utilisée dans l'algorithme de classification de E. Diday: cf [Sup. Class.] TI B n° 9 § 2.1). Si, à peu d'exceptions près, on a, pour les individus i étudiés, $i_J \in E_{q(i)}$, on conviendra d'affecter à la classe q tout individu supplémentaire s_J qui tombe dans le domaine E_q.

On peut encore en vue de discriminer les classes, les projeter sur le sous-espace de R_J engendré par les premiers axes principaux d'inertie du nuage $\{q_J | q \in Q\}$ des centres des classes. Soit e_J un vecteur porté par un axe d'inertie; u^J la forme linéaire coordonnée sur cet axe; on sait qu'on a (cf [Repr. Eucl.] § 4.4) :

$$\sigma(Q)_{JJ'} \circ (\sigma(I)^{-1})^{JJ} e_J = \rho\, e_J$$

$$(\sigma(I)^{-1})^{JJ} \circ \sigma(Q)_{JJ} u^J = \rho\, u^J; \text{ ce qui s'écrit encore :}$$

$$\Sigma \{(\sigma(I)^{-1})^{jj'} \sigma(Q)_{j'j''} u^{j''} | j' \in J, j'' \in J\} = \rho\, u^j;$$

formules où on a considéré $\sigma(Q)$ et $\sigma(I)$ comme des applications linéaires de R^J dans R_J ; et où ρ est le moment d'inertie du nuage des centres des classes dans la métrique $\sigma(I)^{-1}$. La forme linéaire u^J relative à la plus forte valeur propre de $\sigma(I)^{-1} \circ \sigma(Q)$ réalise le maximum du rapport de la variance interclasse à la variance totale : nous rencontrons la solution du problème apparu au début de ce paragraphe. Les vecteurs propres e_J de $\sigma(Q) \circ \sigma(I)^{-1}$ relatifs à des valeurs propres non nulles engendrent la direction du support du nuage des centres q_J. Dans l'espace rapporté aux deux ou trois premiers axes ainsi extraits, peuvent apparaître des cloisons, non-nécessairement planes séparant bien les classes données. L'intérêt de ces cloisons est qu'étant construites dans un espace de faible dimension elles sont plus stables : on reviendra plus bas (§§ 6, 7, 8, 9) sur cette condition essentielle de stabilité.

Le moment d'inertie du nuage Q par rapport à son centre, dans R_J muni de la métrique $\sigma(I)^{-1}$ sera appelé, en bref, variance interclasse cumulée : c'est la trace de l'application linéaire $(\sigma(I)^{-1})^{JJ} \circ \sigma(Q)_{JJ}$. Cette variance interclasse cumulée est une fonction croissante de l'ensemble des variables de description employées. De façon précise on a :

$$\forall\ H \subset J : \text{trace}(\sigma(I)_{HH}^{-1})^{HH} \circ \sigma(Q)_{HH} \leqslant \text{trace}(\sigma(I)_{JJ}^{-1})^{JJ} \circ \sigma(Q)_{JJ}$$

Pour démontrer ce résultat, le plus simple est de considérer R_J muni de la métrique euclidienne $(\sigma(I)^{-1})^{JJ}$. Soit $R_{J-H} = N$ le sous-espace ensemble des vecteurs x de R_J dont sont nulles toutes les coordonnées x_j pour $j \in H$. Soit S le supplémentaire orthogonal de N (dans R_J muni de la métrique $(\sigma(I)^{-1})^{JJ}$). Il est équivalent de chercher la variance interclasse cumulée dans S, pour le nuage des points i_S projection orthogonale sur S des points i_J , ou dans R_H pour le nuage des points i_H (obtenus en annulant les coordonnées i_j de i_J pour $j \in J - H$) : en effet les points i_S et i_H se correspondent pas l'isomorphisme de S sur R_H que réalise la projection parallèlement à N. Or dans S la variance interclasse cumulée est moindre que dans R_J; car la métrique d'inertie $\sigma(I)^{-1}$ de S est induite par celle de R_J; donc dans l'inégalité évidente ci-dessous :

$$\Sigma \{(\mu_q/M)\|q_S\|^2 | q \in Q\} \leqslant \Sigma \{(\mu_q/M)\|q_J\|^2 | q \in Q\}$$

(où la norme désigne la métrique $\sigma(I)^{-1}$ de R_J) le premier terme est l'inertie interclasse cumulée dans S et le second est l'inertie interclasse cumulée dans R_J.

2.3. *Discrimination d'après un tableau de correspondance* :

Au paragraphe 2.1 on a défini la variance totale et la variance interclasse, comme associées respectivement à deux nuages I et Q situés dans un même espace R_J. On peut encore comme dans la leçon [Corr. Esp.] (TII B n° 7) construire plusieurs espaces de fonctions munis d'une forme quadratique. Ces constructions sont le plus complètes dans le cas où le tableau de description des éléments de l'ensemble I est un tableau de nombres positifs k_{IJ} ; d'où le titre de ce paragraphe.

Notons L_2^I l'espace R^I des fonctions sur I, muni de la norme :

$$\|g^I\|^2 = (1/M) \Sigma \{\mu_i (g^i)^2 | i \in I\};$$

cette norme, pour une fonction de moyenne nulle, n'est autre que la variance. De même notons L_2^Q l'espace R^Q des fonctions sur Q, muni de la norme :

$$\|g^Q\|^2 = (1/M) \Sigma \{\mu_q (g^q)^2 | q \in Q\}.$$

L'espace L_2^Q , avec sa norme, peut être identifié au sous-espace de L_2^I , ensemble des fonctions qui sont constantes sur chacune des classes q. La projection orthogonale d'une fonction $g^I \in L_2^I$ sur le sous-espace $L_2^Q \subset L_2^I$, n'est autre que la fonction g^Q dont la valeur pour q, est la moyenne de g^I sur cette classe :

$$\forall\, q \in Q : g^q = (1/\mu_q) \Sigma \{\mu_i\, g^i | i \in q\}.$$

On écrira : $g^Q = g^I \circ p_I^Q$ où p_I^Q est la transition qui a q associe le profil de sa classe :

$$\forall\, q \in Q, \forall\, i \in I : p_i^q = \delta^q_{q(i)}\, \mu_i/\mu_q$$

(en particulier : $i \in q \Leftrightarrow p_i^q \neq 0$). Si g^I est de moyenne nulle, sa variance totale est, on l'a dit, $\|g^I\|^2$; sa variance interclasse est $\|g^Q\|^2 = \|g^I p_I^Q\|^2$; sa variance intraclasse est : $\|g^I\|^2 - \|g^Q\|^2 = \|g^I - g^Q\|^2$ (carré de la norme de la différence entre g^I et sa projection orthogonale g^Q). Toutes ces constructions peuvent être regardées comme un cas particulier de celles de l'analyse des correspondances, si on note p_{IQ} la loi suivante :

$$\forall\, i \in I,\ \forall\, q \in Q : p_{iq} = (1/M)\ \delta_{q(i)}{}^q\, \mu_i$$

Considérons maintenant les vecteurs de descriptions des individus i de I. Chaque coordonnée j de E est une fonction sur I qu'on notera p_j^I :

$$\forall\, i \in I,\ \forall\, j \in J : p_j^i = i_j \text{ (coordonnée j de l'individu i).}$$

L'ensemble $\{p_j^I \mid j \in J\}$ de ces fonctions engendre dans L_2^I un sous-espace vectoriel qu'on notera F. Chercher, parmi les combinaisons linéaires des coordonnées p_j^I de la description des éléments de I, celles qui rendent maximum le rapport de la variance interclasse à la variance totale équivaut à chercher dans L_2^I les vecteurs de F faisant avec L_2^Q le plus petit angle. C'est l'étude dans l'espace euclidien L_2^I de la figure formée par les deux sous-espaces L_2^Q et F. Du point de vue géométrique (auquel manquent, toutefois des considérations statistiques dont la fin de ce paragraphe donne un exemple) les meilleures fonctions discriminantes forment dans F un sous-espace vectoriel image de L_2^Q par l'application π_F, projection orthogonale sur F (au sein de l'espace euclidien L_2^I). Nous avons déjà rencontré un autre moyen de définir cet espace vectoriel $\pi_F(L_2^Q)$. En effet, soit G, le support affin dans R_J du nuage des q_J ; tout point i_J (et plus généralement tout point s_J décrivant un élément supplémentaire) admet sur G une projection orthogonale $\pi_G(i_J)$ (au sein de l'espace R_J muni de la métrique d'inertie $\sigma(I)^{-1}$, du nuage I); ainsi toute forme linéaire u sur G définit une forme linéaire $u \circ \pi_h$ sur I (et plus généralement sur tout R_J). Ces fonctions, combinaisons linéaires des coordonnées p_j^I ne sont autres que celles de $\pi_F(L_2^Q) \subset F$.

Supposons de plus que les données consistent en un tableau de contingence k_{IJ} ; notons comme d'usage, k le total général et k(i) le total de la ligne i. Soit p_{IJ} la loi de fréquence associée à ce tableau ($p_{ij} = k(i,j)/k$) : il est naturel de demander que les poids relatifs μ_i des individus s'accordent avec la loi marginale p_I :
$\forall\, i \in I : p_i = k(i)/k = \mu_i/M$. Ceci posé, on a dans L_2^{IxJ} (espace R^{IxJ} des fonctions sur I x J avec comme carré de norme le moment d'ordre 2 pour la loi p_{IJ}) outre les deux sous-espaces L_2^I et L_2^J associés à la correspondance p_{IJ}, un troisième sous-espace L_2^Q défini ci-dessus :

$L_2^Q \subset L_2^I$. On a vu que la recherche du maximum du rapport de la variance interclasse à la variance totale équivaut à l'étude de la figure formée par L_2^Q et $F = \pi_{LI}^{LJ}(L_2^J)$, image de L_2^J dans L_2^I par projection orthogonale. Soit $u^J \in L_2^J$: on a dans L_2^I : $g^I = u^J \circ p_J^I$; et dans L_2^Q : $g^Q = g^I \circ p_I^Q = u^J \circ p_J^Q$ (dans cette formule, on note p_J^Q la transition associée à la correspondance p_{QJ} :

$$\forall\, q \in Q,\ \forall\, j \in J : p_{qj} = \Sigma\, \{p_{ij} | j \in q\}).$$

L'objet de l'analyse discriminante est de trouver u^J rendant minimum l'angle entre g^I et g^Q; tandis que l'analyse factorielle de la correspondance p_{IJ} cherche u^J rendant minimum l'angle entre u^J et g^I.

Sans prétendre étudier en détail la figure $\{L_2^Q \subset L_2^I\ ;\ L_2^J\}$, montrons sur un exemple l'intérêt et la complexité de cette étude.

Soit Q un exemple de poètes, I un ensemble de pièces de théâtres écrites par ceux-ci, J un ensemble de mots : si chaque pièce n'est l'oeuvre que d'un seul auteur, Q définit une partition de I : on pose $i \in q$ si q est l'auteur de i. On considère le tableau de contingence k_{IJ} : k(i,j) est le nombre de fois que le mot j est employé (ou seulement, employé en début de vers...) dans la pièce i. On se propose d'après ce tableau de distinguer des fonctions u^J sur J telles que le calcul des combinaisons linéaires $\Sigma\, \{u^j\ p_j^s | j \in J\}$ des fréquences d'emploi des mots dans une pièce supplémentaire s d'attribution douteuse, aide à en déterminer l'auteur. Ce problème de discrimination ne paraît pas insoluble : par exemple il est facile de s'assurer que Corneille emploie, en début de vers, le Et et le Si bien plus fréquemment que ne le fait Racine; tandis que, dans ses comédies en vers, Molière met encore plus de Et que Corneille, et encore moins de Si que Racine. Mais il existe des fonctions discriminantes parfaites quant aux données du tableau de base k_{IJ} et que les calculs de rapport de variance (cf § 2.2) signaleraient à coup sûr, bien que leur valeur de prédiction soit nulle. Les noms propres fournissent de telles fonctions : on peut définir Racine comme l'auteur chez qui est élevée la somme des fréquences des mots Britannicus, Hermione, Monime etc...; définir Corneille par les mots Chimène, Polyeucte etc.... Tous les faits élémentaires que l'on peut dénombrer (présence de mots, de désinences ou seulement de lettres) ne sont pas également propres à révéler les différences profondes entre auteurs. Jusqu'à présent la variabilité au sein de l'oeuvre d'un même auteur, (ou au sein d'un seul ouvrage), des fréquences des divers traits de langage, n'a fait l'objet que d'hypothèses sommaires, d'après lesquelles on prétend quelquefois autoriser par des calculs statistiques des assertions dont ces calculs n'affermissent pas les bases. Le calcul électronique permettra de découvrir sur des données assez vastes les lois de cette variabilité. Nous nous bornerons à citer en exemple les surprenantes variations des fréquences d'emploi des consonnes que, sous la direction du Pr. G. Weil, A. Salem étudie dans le texte hébraïque de l'Ecriture Sainte.

3. *Le cas d'une dichotomie* :

3.1. Le cas où il n'y a que deux classes : $Q = \{q,q'\}$ est le plus simple; c'est vraisemblablement le plus souvent étudié; et tout autre peut s'y ramener (cf. 3.2). Il n'y a alors qu'un seul axe d'inertie qui est la droite joignant les centres q_J, q'_J. L'aplication linéaire

$\sigma(I)^{-1}\circ\sigma(Q)$, n'a qu'une seule valeur propre non-nulle qui est la variance interclasse dans la direction de l'axe $q_J\, q'_J$ (pour la métrique $\sigma(I)^{-1}$). Cette variance est :

$$\rho = (\mu_q\, \mu_{q'}/M^2)(\sigma(I)^{-1})^{JJ}\ (q'_J - q_J)(q'_J - q_J)$$

$$= \mu_q\, \mu_{q'}\, M^{-2}\|q'_J - q_J\|^2$$

où la norme carrée est comptée dans la métrique $\sigma(I)^{-1}$. La variance totale (variance de I) est 1, dans cette direction comme dans toute autre, du fait de la métrique choisie. La forme linéaire coordonnée sur cet axe (vecteur propre de $\sigma(I)^{-1}\circ\sigma(Q)$) est :

$$u^J = (\sigma(I)^{-1})^{JJ}\ (q'_J - q_J)/\|q'_J - q_J\|;$$

$$u^j = \Sigma\ \{(\sigma(I)^{-1})^{jj'}\ (q'_{j'} - q_{j'})/\|q' - q\|\,|\,j' \in J\}$$

On peut tenter de séparer les classes par l'hyperplan médiateur du segment joignant leur centre : c'est l'hyperplan de Fisher. La place d'un individu supplémentaire s_J est déterminée par le signe du produit scalaire :

$$(\sigma(I)^{-1})^{JJ}(q'_J - q_J)(2s_J - q_J - q'_J) =$$

$$\Sigma\ \{(\sigma(I)^{-1})^{jj'}\ (q'_j - q_j)(2s_{j'} - q_{j'} - q'_{j'})\,|\,j,\ j' \in J\}$$

si ce produit scalaire est positif, s est mis dans la classe q', sinon il est mis dans q. On peut encore choisir un hyperplan séparateur passant par le centre de gravité g_J; en fait, il n'est pas très coûteux (cf infra § 4) de choisir un seuil optimum pour le produit scalaire :

$$(\sigma(I)^{-1})^{JJ}\ (q'_J - q_J)\ s_J\ ,$$

(ou, ce qui revient au même, pour l'abscisse :

$$u^J(s_J - g_J)\ \text{ sur l'axe } q_J q'_J).$$

Une valeur propre ρ de $\sigma(I)^{-1}\circ\sigma(Q)$ est nécessairement inférieure ou égale à 1 (la valeur 1 n'étant atteinte que si sur l'axe principal d'inertie correspondant toute classe q se projette concentrée en un point : condition nécessaire et suffisante pour que s'annule sur cet axe la variance intraclasse). Pour servir de repère dans les applications, donnons quelques valeurs de ρ, pour une distribution continue symétrique de masses sur la droite $p(x)\,dx$ partagée en deux classes par le milieu. Dans les figures 1 on a :

$$\mu_{q'} = \int_0^\infty p(x)\ dx = \int_{-\infty}^0 p(x)\ dx = \mu_q = 1/2\ ;$$

$$\sigma^2 = \int_{-\infty}^{\infty} p(x)\ x^2\ dx = 1\ ;$$

$$|gq| = |gq'| = 2\int_0^\infty x\ p(x)\ dx\ ;\ \rho = |gq|^2/\sigma^2 = |gq|^2.$$

(où gq' est l'abscisse du centre de gravité q' de la partie de la distribution portée par la demi-droite positive).

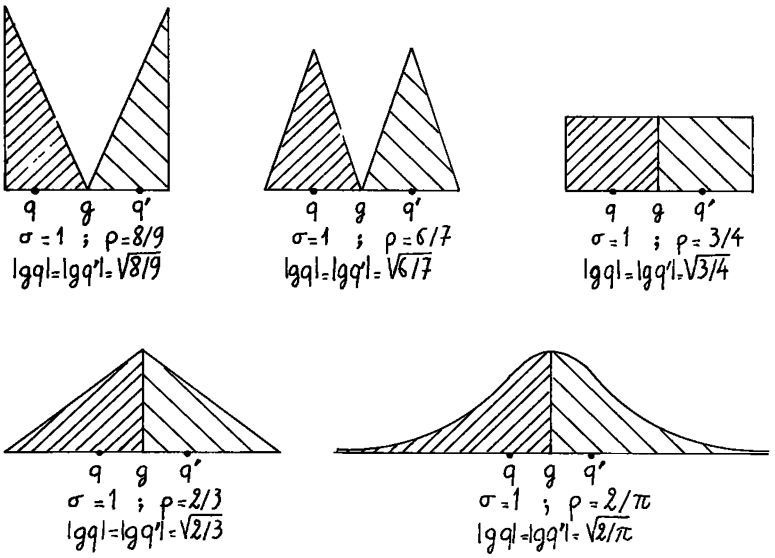


Figure 1 : exemples de distributions continues symétriques de masses divisées en deux classes par le milieu ; on a noté ρ le rapport de la variance interclasse à la variance totale. Pour les lois unimodales, le maximum de ρ est 3/4 ; mais le minimum est 0 (cf fig. 2).

On peut montrer par un calcul de variation que, pour une loi p(x) unimodale, le maximum de ρ est (3/4), valeur donnée par une distribution rectangulaire. Quant au minimum il est nul comme le montre le cas limite esquissé sur la figure 2, que nous commentons brièvement :

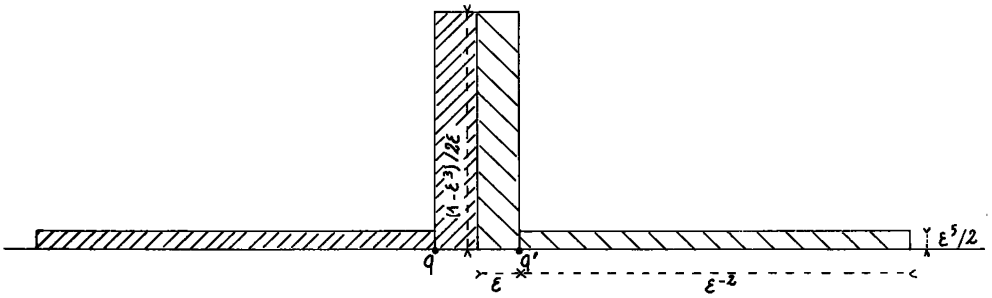


Figure 2 : esquisse d'une distribution unimodale pour laquelle ρ tend vers 0.

Dans cette figure, chaque classe se compose d'un créneau élevé, dont la masse est $(1 - \varepsilon^3)/2$; et d'un palier, dont la masse est $\varepsilon^3/2$. Le centre de gravité de la classe q' tend à avoir pour abscisse ε, parce que à l'intégrale $\int_0^\infty xp(x)\,dx$ le créneau et le palier apportent des contributions équivalentes à $\varepsilon/4$. La variance d'autre part, provient tout entière du palier; et est équivalente à $(1/3)\ \varepsilon^5 \times (\varepsilon^{-2})^3 = (\varepsilon^{-1}/3)$. Variance intraclasse et variance totale diffèrent arbitrairement peu, puisque le centre de gravité général g est à l'origine, arbitrairement proche des centres de classes q et q'. Donc ρ tend vers zéro avec ε.

Revenons au cas général où l'on ne suppose pas que les classes q et q' se partagent symétriquement. On voit sur l'exemple très simple de la figure 3 que le taux d'individus bien classés peut n'être que de 50 % (hasard pur) pour ρ arbitrairement voisin de 0,5. Ici I consiste en quatre points de masse égale;

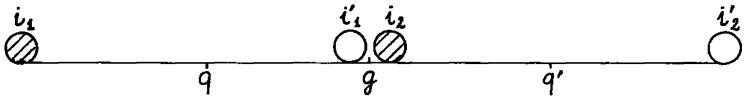


Figure 3: Le rendement de la séparation linéaire est celui du hasard pur ; mais $\rho \neq 0{,}5$.

on a $q = \{i_1 , i_2\}$; $q' = \{i'_1 , i'_2\}$; les points i'_1 et i_2 sont infiniment proches de l'origine; i_1 et i'_2 ont pour abscisse ± 1; variance interclasse et variance intraclasse sont infiniment proches de (1/4).

Cette même figure nous permet de montrer que si les classes q et q' ont même masse le rapport ρ est nécessairement inférieur au taux δ d'individus bien classés et peut prendre toute valeur entre 0 et δ. En effet, cherchons pour δ donné, à rendre ρ maximum. Supposons, ce qui facilitera les calculs, que les centres q' et q ont pour abscisse respective $+\delta$ et $-\delta$: la variance interclasse $\sigma(Q)$ est ainsi fixée à δ^2. Pour minimiser la variance intraclasse, il faut d'abord que les individus mal classés soient aussi proches que possible du centre de leur classe : ce qu'on obtiendra en les mettant tout près de l'origine, mais de l'autre côté que leur centre. Nous aurons donc dans q' un point i'_1 de masse $(1 - \delta)$ et dans q un point i_2 de masse $1 - \delta$ (cf. fig. 3). Le reste, i.e. les individus bien classés des classes q' et q, sont des sous-nuages de masse δ centrés respectivement aux points i'_2 et i_1 d'abscisse ± 1 (car les centres des classes ont les abscisses δ et $-\delta$). On minimisera la variance intraclasse en concentrant en ces points les individus bien classés. D'où :

$$\sigma(I - Q) = \delta^2(1 - \delta) + (1 - \delta)^2\delta = \delta - \delta^2 \ ;$$

$$\sigma(I) = \delta \ ; \quad \sigma(Q)/\sigma(I) = \delta.$$

En déplaçant vers l'origine une partie des masses bien classées, et écartant les autres en sorte que les centres des classes restent en q et q' , on pourra donner à ρ toute valeur entre 0 et δ.

On notera qu'il est équivalent de dire que pour δ donné ρ parcourt $(0, \delta)$ ou de dire que pour ρ donné, δ parcourt $(\rho, 1)$: ainsi ρ fournit une borne inférieure pour δ.

3.2. Affectation par dichotomies à un système quelconque de classes

Supposons l'ensemble Q des classes muni d'une hiérarchie totale binaire de parties AQ (cf [D.M.Cl.] TI B n° 3 § 2) : cette hiérarche a pour éléments terminaux les classes q elles-mêmes, que nous appellerons désormais classes primaires; elle a un ensemble de (CARDQ - 1) noeuds; et son sommet n'est autre que Q tout entier. A chaque noeud, qui est une partie n de Q, est associée une partie I(n) de I, que nous appellerons classe supérieure :

$$I(n) = \cup \{q | q \in n\} \subset I \ ;$$

en particulier, I(Q) = I; et on peut noter $\forall q \in Q : I(q) = q$.

Il est d'usage (cf [C.A.H.] TI B n° 4) de numéroter les classes primaires q de 1 à CARDQ, et les noeuds, ou classes supérieures, de CARDQ + 1 à 2 * CARDQ - 1; le sommet Q (qui en tant que classe n'est autre que I tout entier) recevant le numéro 2 * CARDQ - 1. Chaque noeud a deux successeurs entre lesquels on convient de distinguer un aîné A[n] et un benjamin B[n] (pour un rappel de ces notations, cf. fig. 4).

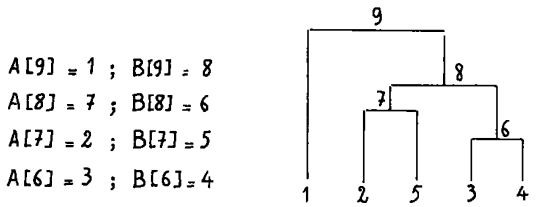


Figure 4 : un exemple d'arborescence pour Card Q = 5 ; Q = {1,2,3,4,5}.

Désormais nous noterons I(N) la classe (primaire ou supérieure) qui a reçu le numéro N : en particulier, I(2 * CARDQ - 1) n'est autre que I tout entier; et l'équation :

$$\forall N \in [\text{CARDI} + 1,\ 2 * \text{CARDI} - 1] \ : \ I(N) := I(A[N]) \cup I(B[N]) ;$$

exprime que la classe supérieure associée au noeud n° N, est réunion des classes associées à l'aîné et au benjamin de ce noeud.

Ceci posé, l'affectation d'un élément (éventuellement, d'un élément supplémentaire s, donné par son vecteur de description) se fera en descendant l'arbre AQ : on décidera d'abord auquel des deux successeurs du sommet, A[2 * CARDQ - 1] et B[2 * CARDQ - 1], affecter cet élément; et, l'affectation faite à un noeud N, on décidera entre A[N] et B[N]; et ainsi de suite jusqu'à parvenir à une classe primaire. (Pour choisir entre deux classes, on utilisera soit l'hyperplan médiateur de Fisher, cf § 3.1, soit une autre technique plus complexe; cf. infra § 4).

Reste à munir l'ensemble Q d'une classification hiérarchique. Pour cela nous recommandons d'utiliser l'algorithme de classification ascendante hiérarchique avec pour critère l'agrégation suivant la variance (cf. [C.A.H.] § 2.5). On prendra donc pour distance DIS[Q,QP] entre les deux classes primaires q = I(Q) et q' = I(QP) (numérotées respectivement Q et QP) la différence :

$$d(q,q') = M^2(q \cup q') - M^2(q) - M^2(q') \ ,$$

où pour toute partie p de I, $M^2(p)$ désigne le moment d'ordre 2 (moment

d'inertie par rapport au centre de gravité) du nuage $\{(i_J , \mu_i) | i \in p\}$ dans R_J muni d'une métrique convenable (qui pour nous sera de préférence $(\sigma(I)^{-1})^{JJ}$); c'est-à-dire l'inertie interclasse de $q \cup q'$:

$$d(q, q') = (\mu_q \mu_{q'}/(\mu_q + \mu_{q'}))\|q_J - q'_J\|^2 ;$$

Le tableau L recevra en $L(Q)$ la masse μ_q de la classe $I(Q) = q$. Et on appliquera le programme usuel.

On remarquera que notre proposition revient à munir l'ensemble I lui-même d'une classification ascendante par agrégation suivant la variance du nuage des $\{(i_J , \mu_i) | i \in I\}$; cette classification se faisant non, comme à l'ordinaire, en partant des éléments $\{i\}$, mais seulement en agrégeant les classes q qui nous sont données a priori.

3.3. *Dichotomie rapide dans le cas d'un tableau de correspondance* :

Considérons le cas suivant : I ensemble d'individus (i. de base) partagé en deux classes disjointes I^+ et I^- ($I = I^+ \cup I^-$; $I^+ \cap I^- \neq \emptyset$); J ensemble de variable; $k_{IJ} = \{k(i,j)\}$ tableau de correspondance décrivant les individus i de I par les variables j de J. On construit un tableau de correspondance à deux lignes décrivant les classes I^+ et I^- :

$$k(I^+ , j) = \Sigma \{k(i, j) | i \in I^+\} ;$$

$$k(I^- , j) = \Sigma \{k(i, j) | i \in I^-\} ;$$

L'analyse de ce tableau k_{QJ} (où $Q = \{I^+ , I^-\}$) ne fournit évidemment qu'un seul axe non trivial (relatif à une valeur propre inconnue λ). Nous allons voir que des calculs linéaires simples permettent de placer sur cet axe les points qui intéressent la discrimination entre les classes I^+ et I^-.

On place d'abord sur une droite I^+ et I^- et entre eux une origine 0 qui est au barycentre de I^+ et I^- affectés des masses $k(I^+)$, $k(I^-)$.

On note $x(I^+)$ et $x(I^-)$ les abscisses de I^+ et I^- calculées avec une échelle arbitraire (qu'il est inutile de normaliser). Pour placer l'ensemble J on pose :

$$x'(j) = (k(I^+ , j) \times (I^+) + k(I^- , j) \times (I^-))/k(j) ;$$

(d'après la formule de transition de l'analyse des correspondances il faudrait poser $x(j) = \lambda^{-1/2} x^1(j)$: mais l'échelle ne nous intéresse pas). Pour placer tout individu i, ou un individu supplémentaire s, on pose :

$$x''(i) = \Sigma \{(k(i, j)/k(i)) x'(j) | j \in J\}$$

(d'après la formule de transition il faudrait poser :

$$x(i) = \lambda^{-1/2} x'(i) = \lambda^{-1} x''(i);$$

mais nous répétons que l'échelle n'importe pas).

Et c'est d'après cette coordonnée x" (dont le calcul est particulièrement facile si $k(i,j)$ est un tableau de description logique ne conte-

nant que des 1 et des 0) que l'on tentera de discriminer entre les deux classes et d'y affecter les individus supplémentaires. Un exemple d'application est donné dans [Aorte] § 4.5 ce cahier pp 415-434.

4. Métrique associée à une classe :

Au paragraphe 2.2 on a proposé d'estimer la proximité d'un individu supplémentaire s à une classe q par la distance de s_J au centre de gravité q_J de cette classe; la distance étant comptée dans la métrique d'inertie $\sigma(I)^{-1}$ du nuage I tout entier. Cette estimation ne tient aucun compte de ce que les classes diffèrent quant au poids à la forme et à la dispersion. Or e.g., un point s_J situé à égale distance du centre q_J d'une classe très concentrée (dont l'écart-type est inférieur à $\|q_J - s_J\|$) et du centre q'_J d'une classe très dispersée devra plutôt être rattaché à cette dernière. Supposons une description probabiliste complète : à chaque classe q, de masse μ_q , est associée une densité continue de probabilité : $p_q(x_J)\, dx_J$, (dont l'intégrale étendue à tout l'espace R_J est 1) : on rattachera l'individu s_J à la classe q pour laquelle est maximum $\mu_q\, p_q(s_J)$. (Eventuellement on modifie cette règle de décision en attribuant aux différentes erreurs possibles des coûts différents). Toutefois, dans la pratique, il n'y a pas de description probabiliste complète connue. Souvent, on associe à une classe q la loi normale spatiale, de centre q_J , ayant même variance que le nuage des descriptions i_J des éléments de q. On a ainsi pour la densité $p_q(x_J)$ une expression analytique simple certes, mais peu sûre, surtout quand on s'éloigne du centre q_J :

$$p_q(x_J) = (2\pi)^{-\mathrm{Card}J/2} (\det \sigma(q)_{JJ})^{-1/2} \exp(-(\sigma(q)^{-1})^{JJ}$$

$$(x_J - q_J)\ (x_J - q_J)/2)$$

(cf [Repr. Eucl.] TII B n° 2 § 3). (Signalons incidemment que si la classe q est de faible effectif, la forme quadratique $\sigma(q)$ peut se trouver inférieure à la variance d'erreur elle-même, σ' résultant de l'imprécision des mesures : en ce cas, il s'impose de substituer à $\sigma(q)$ une forme quadratique $\sigma'(q)$ supérieure à σ' avant d'inverser la matrice d'inertie pour calculer les coefficients de la forme quadratique figurant dans la densité : on posera, par exemple, $\sigma'(q) = \sigma' + \sigma(q)$).

L'hypothèse de normalité de la classe q une fois écartée, l'intérêt de la métrique d'inertie de la classe q subsiste : c'est sans doute de préférence dans cette métrique qu'il convient de compter l'écart d'un point supplémentaire s_J au centre q_J. On posera donc :

$$d^2(s, q) = N_q \|s_J - q_J\|^2_q = N_q((\sigma(q)^{-1})^{JJ}\ (s_J - q_J)(s_J - q_J));$$

dans cette formule on a délibérément omis de préciser un coefficient de normalisation N_q , parce que seuls paraissent sûrs les rapports entre échelles de distance sur deux droites passant par le centre q_J , non les échelles mêmes.:

Pour décider si s doit être rapporté à q ou à q' , on calculera le rapport :

$$(sq/sq') = \|s_J - q_J\|^2_q / \|s_J - q'_J\|^2_{q'},$$

et on comparera (sq/sq') à un seuil S(q/q') choisi de telle sorte que soit minimum le nombre d'erreurs F(S) y correspondant pour les individus i effectivement décrits :

$$F(S) = \text{Card}(\{i \mid i \in q;\ S < (iq/iq')\} \cup \{i' \mid i' \in q';\ (iq/iq') < S\}),$$

(i.e. le nombre des individus i qui sont dans q mais dont le rapport (iq/iq') des distances carrées à q et q' surpasse S; augmenté du nombre des i' qui sont dans q' bien que (iq/iq') < S). Ainsi la surface séparatrice entre q et q' sera non un hyperplan (cf § 3, l'hyperplan de Fisher) mais une quadrique d'équation :

$$\|s_J - q_J\|^2_q - S\|s_J - q'_J\|^2_{q'} = 0.$$

Semblablement on peut prendre pour quantité critère la différence :

$$(sq - sq') = \|s_J - q_J\|^2_q - \|s_J - q'_J\|^2_{q'},$$

qu'on comparera à un seuil S(q - q') choisi pour que soit minimum le nombre d'erreurs.

Si l'on adopte sans réserve le modèle probabiliste, (i.e. qu'on assimile chaque classe à une loi normale dont la densité est estimée d'après l'échantillon des individus de base) on trouve pour logarithme de la densité de la classe q au point s (cf *supra*) :

$$\log p_q(s_J) = -(\text{Card } J/2)\log(2\pi) - (1/2)\log(\det(\sigma(q)_{JJ})) - (1/2)\|s_J - q_J\|^2_q ;$$

le terme en log 2π et le facteur (1/2) étant commun à toutes les classes, on est conduit à mesurer l'écart d'un individu s à une classe q, par la formule :

$$\text{écart}(s,q) = \log(\det(\sigma(q)_{JJ})) + \|s_J - q_J\|^2_q ;$$

d'où pour quantité critère de l'affectation de s à l'une ou l'autre des classes q et q' :

$$\text{diff}(s;\ q,q') = (sq - sq') + \log(\det(\sigma(q)_{JJ})) - \log(\det(\sigma(q')_{JJ}));$$

le terme constant ajouté à (sq - sq') prétend fixer le seuil de séparation entre q et q' (affectation à q si diff < 0; à q' si diff > 0). Cette méthode désignée dans la littérature anglo-saxomme par le sigle NLDA (analyse discriminante non linéaire) peut donner des résultats heureux; nous nous garderons d'y voir une justification du modèle normal, mais soulignerons l'intérêt des cloisons non planes (ici des quadriques) construites d'après les métriques d'inertie propres aux classes (cf. § 4', note à la fin de ce § 4).

On notera que le calcul du seuil S(q/q') (ou S(q - q')) n'est pas d'un coût prohibitif : il suffit d'ordonner la suite des valeurs de (iq/iq') pour $i \in q \cup q'$ (si, c'est rarement le cas, les deux classes q et q' sont trop nombreuses, on peut se borner à un sous-échantillon) et de calculer de proche en proche la fonction F(S) quand S parcourt la suite de ces valeurs. De façon précise notons i^ρ l'élément i de $q \cup q'$ pour lequel le rapport (iq/iq') a rang ρ dans la suite ordonnée en croissant (rang 1 au plus petit)$\{(i'q/i'q') \mid i' \in q \cup q'\}$. Si S est inférieur à (i^1q/i^1q'), tous les éléments de $q \cup q'$, étant au delà du seuil, sont rapportés à la classe q' : il y a donc Card q erreurs. Si $i^1 \in q$, pour S compris entre (i^1q/i^1q') et (i^2q/i^2q'), il y a une erreur

de moins : F(S) = Card q - 1; si au contraire $i^1 \in q'$, on commet une erreur de plus en affectant i^1 à q, et F(S) = Card q + 1. D'où l'algorithme suivant, que nous esquissons en ALGOL.

entier CQ, CAR, RO, FRO, FMIN; réel S;

entier tableau QRO(1,CAR); réel tableau RAP(0,CAR + 1);

Commentaire : CQ := Card q; CAR := Card q + Card q'; en déplaçant le seuil on a un nombre variable d'erreurs, rangé en FRO; le minimum observé pour ce nombre est FMIN, atteint pour la valeur S du seuil; le tableau QRO indique à quelle classe appartient i^ρ : QRO(ρ) := si $i^\rho \in q$ alors 1 sinon - 1; le tableau RAP donne en RAP(ρ) le quotient $(i^\rho_q/i^\rho_{q'})$ (ou équivalemment la différence $(i^\rho_q - i^\rho_{q'})$); on prend, pour la commodité de l'algorithme un RAP(0) inférieur à RAP(1) et un RAP(CAR + 1) supérieur à RAP(CAR).

FRO := FMIN := CQ; S := RAP(0);

Commentaire : pour $S \lesssim_s (i^1_q/i^1_{q'}) = RAP(1)$, le nombre d'erreurs est Q = Card q.

pour RO := 1 pas 1 jusqu'à CAR faire début

FRO := FRO - QRO(RO);

si FRO $\lesssim_s$ FMIN alors début

FMIN := FRO; S := (RAP(RO) + RAP(RO + 1))/2; fin; fin.

Commentaire : si $i^\rho \in q$ le nombre d'erreurs décroît de 1 quand S traverse la valeur $(i^\rho_q/i^\rho_{q'})$; sinon il croît de 1; un nombre d'erreurs FRO, moindre que FMIN est adopté pour nouvelle valeur de FMIN; et le seuil S est pris entre RAP(ρ) et RAP(ρ + 1). Au terme de la boucle on a en S un seuil S(q, q') convenable et un FMIN le nombre de fautes correspondant.

L'algorithme ci-dessous permet de calculer le tableau des seuils S(q/q'); d'après ce tableau l'affectation d'un élément supplémentaire s se peut faire en parcourant la suite des classes, supposée ordonnée $\{q^1, \ldots, q^{Card\ Q}\} = Q$. D'où un algorithme, esquissé ici plus sommairement encore que le précédent :

entier CARQ, Q, CLA; réel tableau S(1: CARQ, 1 : CARQ)

Commentaire : CARQ := Card Q; on essaie successivement les classes q^Q en mettant en CLA le rang de la plus satisfaisante;
S(Q, QP) := $S(q^Q/q^{QP})$: c'est le tableau des seuils;

réel procédure RAP(Q,QP);

Commentaire : sert à calculer RAP(Q, QP) := (sq^Q/sq^{QP});

CLA := 1

pour Q := 2 pas 1 jusqu'à CARQ faire

si RAP(Q, CLA) $\lesssim_s$ S(Q, QP) alors CLA := Q

Commentaire : on adopte CLA := Q si $(sq^{Q}/sq^{CLA}) <_{s} S(q^{Q}/q^{CLA})$, ce qui entraîne à préférer l'affectation $s \in q^{Q}$ à l'affectation prédécente $s \in q^{CLA}$; au terme de la boucle on fait l'affectation : $s \in q^{CLA}$.

L'affectation obtenue par cet algorithme dépend de l'ordre choisi sur l'ensemble Q des classes; il sera intéressant de calculer le nombre d'erreurs faites sur l'ensemble I tout entier.

Une autre voie s'offre pour décider de l'affectation d'un individu à une classe de Q, par une suite ordonnée de comparaisons de cet individu à deux classes : c'est de munir l'ensemble Q d'une classification hiérarchique binaire. De façon précise, reprenons les notations du paragraphe 3.2. Soit S[A[N], B[N]] le seuil relatif à la séparation des deux classes I(A[N]) et I(B[N]); et soit RAP[A[N], B[N]] le rapport des écarts à ces deux classes de l'élément qu'on désire affecter. On a l'algorithme :

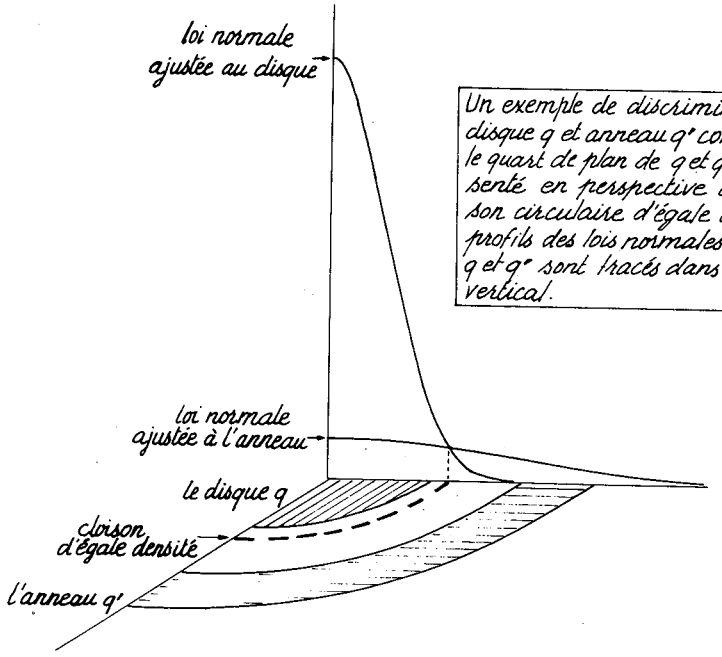
début N := 2*CARDQ - 1 ;

NOEUD;

N := si RAP[A[N], B[N]] < S[A[N], B[N]] alors A[N] sinon B[N] ;

si CARDQ + 1 ⩽ N aller à NOEUD fin

4' Note : un exemple parfois cité en faveur de la méthode NLDA nous fera voir combien le succès de celle-ci est étranger au modèle probabiliste, et dépend seulement comme nous l'affirmons de la richesse de forme offerte par les cloisons non linéaires (cf. *supra*). L'étude détaillée de cet exemple fait l'objet d'un problème publié dans ce même cahier (pp 407-411).



Un exemple de discrimination entre disque q et anneau q' concentriques; le quart de plan de q et q' est représenté en perspective avec la cloison circulaire d'égale densité; les profils des lois normales ajustées à q et q' sont tracés dans un plan vertical.

Soit dans le plan deux classes concentriques : dont l'une q est un disque, l'autre q' un anneau, tous deux de centre O; dans ce cas les distances $\|s - q\|$ et $\|s - q'\|$ sont toutes deux proportionnelles à la distance euclidienne usuelle de s à 0. La méthode NDLA met en place une cloison circulaire qui peut dans les cas favorables séparer l'anneau du disque; ce qu'une cloison linéaire ne fera jamais. Or suivant le modèle normal l'anneau q' est (comme le disque q) assimilé à une distribution unimodale ayant sa densité maxima en 0, ce qui est absurde puisque l'anneau est evidé. La séparation réussit malgré cela parce que la loi normale ajustée à l'anneau q' est plus étalée que celle ajustée au disque q , en sorte que la courbe de séparation (ou cercle d'égalité des densités) se place entre les classes réelles q et q'.

Ce succès n'est toutefois pas général; pour certaines dimensions de l'anneau et du disque, le cercle d'égale densité peut, e.g. être intérieur au disque : au contraire en cherchant un seuil de séparation (cf. *supra* : seuil S(q - q')) on trouve toujours une cloison circulaire convenable (si le disque et le cercle n'empiétent pas).

5. *Ecart d'un individu à une classe, d'après G.S. Sebestyen et J.M. Romeder* :

S'inspirant de G.S. Sebestyen (1962; § 2.5) J.M. Romeder propose de définir l'écart d'un point s_J à une classe q par la formule suivante (où $\|a\|$ est une norme qu'il reste à préciser) :

$$d^2(s,q) = \Sigma \{(\mu_i/\mu_q)\|s_J - i_J\|^2 | i \in q\};$$

il résulte du théorème de Huyghens que $d^2(s,q)$ n'est autre que la somm de $\|s_J - q_J\|^2$ et de la variance totale de la classe q :

$$d^2(s,q) = \|s_J - q_J\|^2 + \Sigma \{(\mu_i/\mu_q)\|i_J - q_J\|^2 | i \in q\}.$$

Quant à la métrique J.M. Romeder demande que soit minima la dispersion de classe qu'il exprime par :

$$D^2(q) = (\text{Card}q\ (\text{Card}q - 1))^{-1} \Sigma \{\|i_J - i'_J\|^2 | i \in q,\ i' \in q\};$$

à supposer que les éléments i aient tous même masse μ_i , $D^2(q)$ n'est que la variance totale de la classe q, (à un coefficient près) :

$$D^2(q) = 2(\text{Card}q/(\text{Card}q - 1))\Sigma \{(\mu_i/\mu_q)\|i_J - q_J\|^2 | i \in q\};$$

Nous dirons donc qu'il s'agit de minimiser la variance totale de la classe q.

Cette condition serait radicalement satisfaite en posant $\forall x: \|x\| =$(

Il faut donc imposer à la forme quadratique de distance une contrai te l'écartant de zéro. J.M. Romeder demande qu'un cube ayant arête 1 pour la métrique choisie (i.e. un parallélépipède dont les arêtes sont les vecteurs d'une base de R_J orthonormée pour cette métrique) ait volume 1 pour l'élément de volume dont est naturellement muni R_J (i.e. l'élément de volume pour lequel a volume 1 le parallélépipède construi sur les vecteurs de la base canonique de R_J ; vecteurs dont chacun a toutes ses composantes nulles sauf une qui vaut 1); ce qui, en formule signifie qu'a valeur 1 le déterminant du tableau carré $\{m(q)^{jj'} | j \in J,\ j' \in J\}$ du tenseur $m(q)^{JJ}$ définissant la métrique associée à la classe q. Comme les métriques $m(q)^{JJ}$ associées aux diverses classes ne servent qu'à effectuer des comparaisons, il reviendrait au

même d'imposer que :

$$\forall q, q' \in Q : \det m(q)^{JJ} = \det m(q')^{JJ} ;$$

en sorte que la condition de normalisation relative des m(q) ne dépend pas d'un changement d'échelle sur les grandeurs mesurées (lequel modifierait l'élément de volume dans l'espace $E = R_J$ des descriptions).

La variance totale du nuage q pour une métrique m s'écrit :

$$\text{trace}(m^{JJ} \circ \sigma(q)_{JJ}) = \Sigma \{m^{jj'} \sigma(q)_{jj'} | j \in J, j' \in J\};$$

c'est la somme des moments principaux d'inertie, ou trace de l'application $m \circ \sigma$; d'autre part le produit de ces moments d'inertie est :

$$\det(m^{JJ} \circ \sigma(q)_{JJ}) = \det(m^{JJ}) \circ \det(\sigma(q)_{JJ}) = \det(\sigma(q)_{JJ})$$

(car det(m) = 1). Les moments d'inertie ayant un produit déterminé, leur somme est minima quand ils sont tous égaux : donc $m \circ \sigma$ est une application diagonale, proportionnelle à l'identité; plus précisément on a :

$$m(q)^{JJ} = \det(\sigma(q)_{JJ})^{1/CardJ} (\sigma(q)^{-1})^{JJ} ;$$

où le scalaire $\det(\sigma)^{1/CardJ}$ a été choisi afin que det(m) = 1. Ainsi dans la métrique m(q) le nuage de la classe q a tous ses moments d'inertie égaux à $\det(\sigma)^{1/CardJ}$ et sa variance totale est $CardJ(\det(\sigma)^{1/CardJ})$ D'où la valeur de l'écart $d^2(s,q)$ adopté par Romeder :

$$d^2(s,q) = \det(\sigma(q)_{JJ})^{1/CardJ} (CardJ + (\sigma(q)^{-1})^{JJ} (s_J - q_J)(s_J - q_J))$$

$$= \det(\sigma(q)_{JJ})^{1/CardJ} (CardJ + \Sigma\{(\sigma(q)^{-1})^{jj'} (s_j - q_j)(s_{j'} - q_{j'}) | j \in J, j' \in J\})$$

Dans la pratique, le calcul répété de matrices inverses σ^{-1} peut être trop coûteux. Romeder propose de choisir alors une métrique md(q) diagonale dans la base canonique de R_J :

$$\forall j, j' \in J : md(q)^{jj'} = \delta^{jj'} md(q)^{jj} ;$$

et il conserve la condition de volume : det(md) = 1. Dans cette classe de métrique, on aura une variance minima en prenant md(q) proportionnelle à l'inverse de la partie diagonale, σd(q), de σ(q). De façon précise, on pose :

$$\sigma d(q)_{JJ} = \{\sigma d(q)_{jj'} | j, j' \in J\}; \quad \sigma d(q)_{jj'} = \delta_{jj'} \sigma(q)_{jj} ;$$

$$md(q)^{JJ} = \det(\sigma d(q)_{JJ})^{1/CardJ} (\sigma d(q)^{-1})^{JJ} ;$$

$$md(q)^{jj'} = \delta^{jj'} \Pi\{\sigma(q)_{ll} | l \in J\}^{1/CardJ} (\sigma(q)^{jj})^{-1}$$

$$d_d^2(s,q) = \Pi\{\sigma(q)_{ll} | l \in J\}^{1/CardJ} (CardJ + \Sigma\{(\sigma(q)^{jj})^{-1} (s_j - q_j)^2 | j \in J\}).$$

Formules qui s'établissent comme les précédentes, et offrent des calculs numériques plus simples.

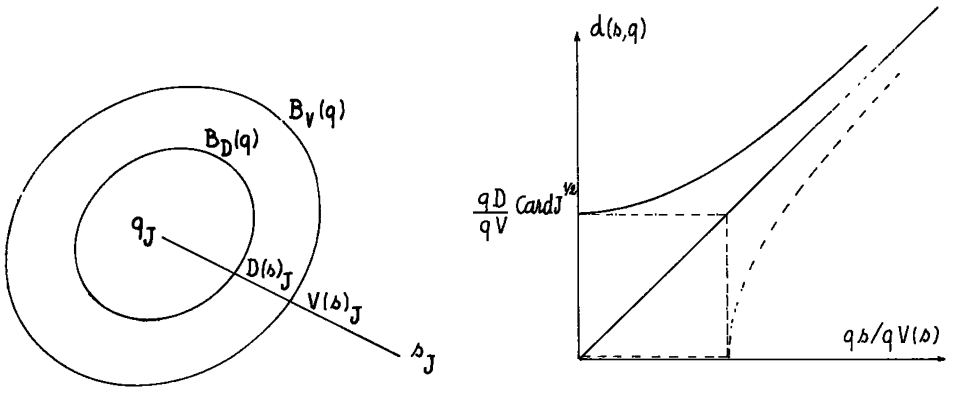


Figure 5. variation de l'indice d'écart d(s,q) calculé par J.M. Romeder en fonction de la position du point s relativement à la boule unité $B_D(q)$ de la métrique d'inertie de la classe q, et à son homothétique $B_V(q)$ qui a même volume que la boule unité de R_J (muni de sa métrique canonique). On a figuré en tireté la variation d'un autre indice d'(s,q) suggéré dans le texte (cf §6 in fine).

Avant de critiquer la fonction d'écart $d^2(s,q)$ utilisée par Romeder, donnons de cet écart une définition géométrique (cf. fig. 1). Soit $B_D(q)$ la boule unité de la métrique d'inertie (ou de dispersion) associée à la classe q :

$$B_D(q) = \{x_J | x_J \in R_J;\ (\sigma(q)^{-1})^{JJ}(x_J - q_J)(x_J - q_J) \leqslant 1\};$$

cette boule découpe sur toute demi-droite issue de q_J un segment égal à la variance du nuage q dans la direction de cette demi-droite. Soit $B_V(q)$ l'homothétique de la boule $B_D(q)$ qui a même volume que la boule unité de R_J (muni de sa métrique canonique) :

$$B_V(q) = \{x_J | x_J \in R_J\ ;\ \det(\sigma(q)_{JJ})^{1/2\ \mathrm{CardJ}}\ x_J \in B_D(q)\}$$

Désignons par $D(s)_J$ et $V(s_J)$ respectivement l'intersection de la demi-droite $q_J s_J$ avec les sphères frontières de $B_D(q)$ et $B_V(q)$. Alors on a :

$$d^2(s,q) = (qD(s)/qV(s))^2\ (\mathrm{CardJ} + (qs/qD(s))^2)\ ;$$

dans cette formule, le rapport qD(s)/qV(s), rapport des rayons des boules $B_D(q)$ et $B_V(q)$ n'est autre que $\det(\sigma(q))^{1/2\ \mathrm{CardJ}}$; et le rapport (qs/qD(s)) est la distance de s à q pour la métrique de dispersion $\sigma(q)^{-1}$. Quand s s'éloigne indéfiniment de q , l'écart d(s,q) tend asymptotiquement vers (qs/qV(s)), i.e. vers la distance de s à q pour la métrique m(q). En fonction de cette dernière distance la courbe représentative de d(s,q) est un arc d'hyperbole ayant pour asymptote la 1° bissectrice.

On peut justifier le choix de la métrique asymptotique m(q) par le désir de suivre la forme du sous-nuage q(en prenant m(q) proportionnel à la métrique d'inertie $\sigma^{-1}(q)$), tout en donnant un même ordre de gran-

deur aux métriques associées aux diverses classes (en demandant que la boule unité $B_V(q)$ ait dans R_J même volume quel que soit q). Mais l'influence de $\sigma(q)$ sur $d^2(s,q)$ pour s_J voisin du centre q_J, ne semble pas heureuse.

Considérons par exemple un système de classes ayant toutes même métrique $m = m(q) = m(q')$, mais différant par la variance. Dans un système de coordonnées orthonormé par la métrique m, $\sigma(q)$ est égal au produit de la matrice unité par une constante que nous noterons $R^2(q)$: R(q) n'est autre que le rayon (qD/qV) de la boule $B_D(q)$ compté dans la métrique m. L'écart $d^2(s,q)$ d'un élément supplémentaire s à la classe q est :

$$d^2(s,q) = |sq|^2 + dR^2(q),$$

(où d = Card J; et $|sq|^2$ est compté dans la métrique m). Entre deux classes q et q' on a un hyperplan de séparation, lieu des points s tels que $d^2(s,q) = d^2(s,q')$. Cet hyperplan se trouve plus proche de celle des deux classes q et q' dont le rayon de giration R(q) est le plus grand. Résultat absurde, qu'on ne peut suffisamment justifier par la nécessité de donner à $d^2(q,q)$ (écart du centre q_J à la classe q) une valeur d'autant plus élevée que la classe est plus dispersée, donc moins dense en son centre.

Il nous paraîtrait plus satisfaisant de poser :

$$d'^2(s,q) = \sup\{|sq|^2 - dR^2(q),\ 0\};$$

cet écart se raccorde asymptotiquement à la métrique m; il est nul à l'intérieur d'une boule homothétique de la boule $B_V(q)$ dans le rapport Card $J^{1/2}$; il est légitime d'associer à q cette boule dont, dans la métrique σ^{-1}, le carré du rayon n'est autre que l'espérance mathématique du carré de la distance d'un point de la classe q à son centre. La formule explicite de d'^2 est :

$$d'^2(s,q)=\sup\{0,\det(\sigma(q)_{JJ})^{1/\mathrm{Card}J}(-\mathrm{Card}J+(\sigma(q)^{-1})^{JJ}(s_J-q_J)(s_J-q_J))\}$$

On traiterait de même de l'écart $d_d(s,q)$, qui ne diffère de d(s,q) qu'en ce qu'on substitue à $\sigma(q)$ sa partie diagonale $\sigma d(q)$.

6. Choix des variables pas à pas et échantillon d'épreuve :

On a dit au paragraphe 1 que le problème de la discrimination est un cas particulier du problème de la régression. Or, en régression, la multiplicité des variables explicatives (ce sont ici les données i_j relatives à l'individu i) prête à erreur (cf [Régr.] à paraître) : une combinaison linéaire de ces variables, affectées de forts coefficients et se soustrayant entre elles, peut n'être qu'un conglomérat d'erreurs dont la coïncidence - sur l'échantillon I des individus étudiés - avec la variable à expliquer, est toute fortuite. Nous avons proposé ailleurs (cf [Régr.] à paraître)diverses méthodes pour se garder des régressions illusoires. Un procédé général très sûr est de soumettre les variables explicatives à l'analyse factorielle : les premiers facteurs obtenus sont de nouvelles variables explicatives qui, gardant l'essentiel de l'information qu'on possède sous une forme peu sensible aux fluctuations

d'erreurs, offrent matière à des formules de régression stables. On verra au paragraphe 8 l'efficacité de l'analyse factorielle dans les problèmes de discrimination. Un autre procédé, qui n'a pas notre faveur, est de choisir parmi l'ensemble (ici J) des variables explicatives, celles qui sont le plus relevantes au problème de régression ou de discrimination que l'on traite. C'est ce que fait J.M. Romeder par la méthode exposée ci-dessous.

Supposons qu'il soit possible de calculer pour toute partie H de J une quantité critère C(H) d'autant plus élevée que la discrimination entre les classes q se fait mieux dans R_H (i.e. en ne considérant de la description i_J de l'individu i que $i_H = \{i_j | j \in H\} \in R_H$). Romeder détermine par récurrence, pas à pas, une suite j(1), j(2),..., j(r), de variables qui sont celles qu'on doit successivement introduire pour obtenir, en un certain sens, la meilleure discrimination possible avec un nombre donné de variables. De façon précise on a :

$$C(\{j(1)\}) = \sup\{C(\{j\}) | j \in J\};$$

$$C(\{j(1), j(2)\}) = \sup\{C(\{j(1), j\}) | j \in J\}; \ldots;$$

$$C(\{j(h) | h = 1, \ldots, r\}) = \sup\{C(\{j(h) | h = 1, \ldots, r-1\} \cup \{j\}) | j \in J\};$$

j(r) est donc la variable dont l'adjonction aux (r-1) déjà choisies est la plus favorable à la discrimination. Ceci n'implique nullement que {j(1), j(2),..., j(r)} soit de tous les sous-ensembles de r variables, celui qui réalise le maximum du critère C. Mais, comme le note Romeder, la recherche du maximum requerrait qu'on calculât $\binom{n}{r}$ le critère C; ce qui, en l'état des moyens de calcul est généralement interdit.

Comme critère C(H), Romeder calcule soit la variance interclasse dans R_H, soit le nombre total des individus bien classés dans R_H. De façon précise (cf § 2), la variance interclasse cumulée dans R_H est la trace de l'endomorphisme $(\sigma(I)^{-1})^{HH} . \sigma(Q)_{HH}$ de R^H; endomorphisme dont les valeurs propres sont les moments principaux d'inertie du nuage des centres q_H, dans R_H muni de la métrique de variance du nuage des i_H (dans le cas particulier le plus souvent traité où Q = {q,q'} la trace se réduit à l'unique moment d'inertie non-nul qui est :

$$(\mu_q \mu_{q'}/M^2)(\sigma(I)^{-1})^{HH} (q_H - q'_H) (q_H - q'_H);$$

c'est au coefficient $(\mu\mu'/M^2)$ près, le carré de la distance d'inertie entre les centres des deux classes q et q', ou distance de Mahalanobis; on rapprochera cette formule de $\sigma(Q)_{HH}$ donné au § 2.2 *in fine*. Quant au nombre total des individus bien classés, il est toujours calculé en rattachant tout individu i à celle des classes q dont il est le plus proche; seul varie cf § 5, le calcul de l'écart de i à q. Un individu $i \in I$ est donc dit bien classé dans R_H si :

$$\forall q \in Q : q \neq q(i) \Rightarrow d^2_H(i,q(i)) < d^2_H(i,q);$$

dans cette formule, l'indice H signifie que le calcul des écarts est fai dans R_H, pour les descriptions limitées au sous-ensemble H des variable explicatives. Romeder indique, de plus, des critères fondés sur des hypc thèses de normalité : malgré cette base peu réaliste, ces critères ont pu dans la pratique aider aux choix des variables.

On notera que le calcul de pas à pas comporte de nombreuses inversions de matrices : calcul coûteux, à moins qu'on ne se restreigne à des matrices diagonales, σd et md, cf § 5. Ce calcul peut être allégé si on remarque que, chaque fois qu'on doit inverser une matrice, r x r, on a déjà inversé au pas précédent une sous-matrice (r - 1) x (r - 1).

Reste à fixer le nombre r de variables explicatives auquel borner les calculs : on sait, en effet qu'une discrimination parfaite réussie sur l'échantillon I dans un espace R_H de dimension élevée est illusoire. On peut tenter ici de recourir à des épreuves de validité fondées sur des hypothèses probabilistes plus ou moins restrictives : et nous donnons au paragraphe 7 un exemple de tel calcul, qui suggère des ordres de grandeur intéressants pour le choix de Card H en fonction de Card I. Toutefois le plus sûr est d'éprouver la stabilité de la règle de discrimination en simulant son application à des éléments supplémentaires (sur le principe général de ces épreuves cf [Epr. Val.] TII B n° 8). Voici comment procède J.M. Romeder. L'ensemble I est situé en deux parties I_1 et I_2 : la règle de discrimination est fondée uniquement sur l'ensemble I_1 ; et on l'éprouve sur l'ensemble I_2 ; c'est pourquoi I_1 est appelé échantillon de base; et I_2 , échantillon d'épreuve. De façon précise, le centre d'une classe q ainsi que la dispersion de cette classe sont étudiés uniquement d'après $q \cap I_1$; c'est aussi d'après I_1 qu'on calcule la quantité critère C(H) pour adjoindre pas à pas des variables explicatives nouvelles. Mais simultanément, on calcule le nombre C_2 des individus de I_2 qui sont bien classés; et on arrête l'adjonction de variables quand C_2 cesse de croître : on verra que dans la pratique (cf. § 8) C_2 croît d'abord puis fluctue : il convient de s'arrêter au seuil de ces fluctuations. Le nombre C_1 des individus de I_1 bien classés fluctue lui-même, quand on a pris pour critère C(H) la variance interclasse; au contraire cette dernière (plus exactement le rapport de la variance interclasse à la variance totale) ne peut que croître quand on adjoint à H une variable (cf. supra § 2.2).

Quant à l'effectif de l'échantillon d'épreuve I_2 , il dépend du nombre CardI des observations disponibles; si Card I semble trop faible pour permettre l'établissement de cloisons sûres, on n'osera pas retrancher beaucoup d'observations pour constituer I_2. Romeder prend, e.g. Card I_2 = 0,25 Card I; et il construit I_2 par tirage au sort. Au fond, se pose ici un problème d'estimation statistique : d'après la fréquence d'erreur, $f(I_2)$, sur I_2 , évaluer la probabilité p d'erreur sur l'ensemble, potentiellement infini, des éléments supplémentaires. Si on admet que le nombre des erreurs est régi par une loi de Poisson, on attribuera à ce nombre un écart-type égal à sa racine carrée. Si l'on n'a pu réserver un échantillon d'épreuve I_2 d'effectif assez élevé, p sera certes mal connu; ce qui incitera à faire plusieurs essais avec des échantillons d'épreuve différents I_2 , I_2' ,... Mais alors on aura

aussi des échantillons-base différents, ($I_1 = I - I_2$, $I_1' = I - I_2'$ ...); et plusieurs systèmes de cloisons; d'où plusieurs probabilités d'erreurs p, p' ... Pour la pratique, le plus utile serait d'avoir une estimation précise de la probabilité d'erreur p associée au système de cloison qu'on a adopté. Cela est impossible. A défaut, en cumulant les erreurs afférentes aux divers essais, on aura une estimation générale E(p). A titre indicatif, on pourra encore, bien que sans fondement théorique précis, attribuer au nombre des erreurs effectuées dans l'ensemble des essais, un écart-type égal à sa racine carrée. Reste un obstacle non négligeable : le coût des calculs limite le nombre des essais.

Malgré ces diverses difficultés, l'échantillon d'épreuve permet un progrès décisif : le taux d'erreur estimé sur I_2 est un repère entre le taux d'erreur sur I (seul calculé auparavant) et la probabilité d'erreur inconnue sur l'ensemble potentiel qu'on vise. (En toute rigueur, on ne peut généralement même pas parler ici de probabilité; car l'infini potentiel des cas est insuffisamment défini; cf [Principes] TII A n° 1 § 1°; ici toutefois le manque de données nous arrête avant de parvenir à cette difficulté là). Quelque méthode de discrimination qu'on utilise on réservera donc un échantillon d'épreuve (cf § 8).

7. Dichotomie aléatoire et séparation linéaire :

Supposons donné dans $E = R_J$ (espace vectoriel de dimension d = CardJ) un nuage I de N points i_J, répartis en deux classes q et q'; plus d sera grand relativement à N, plus il sera facile de séparer q de q' par une cloison hyperplane. Si par exemple l'ensemble I a pour support affin E tout entier (i.e. I n'est inclus dans aucun hyperplan affin de E) et que d = N - 1, une telle cloison h existe sûrement; à titre d'exercice définissons une cloison h par son équation en coordonnées barycentriques relativement aux points de I :

$$h = \{x_J \mid x_J \in R_J;\ \exists\ m_I \in R_I :$$

$$(\Sigma\{m_i \mid i \in q\} = \Sigma\{m_i \mid i \in q'\} = 1/2) \wedge (x_J = \Sigma\{m_i i_J \mid i \in I\})\}$$

(i.e. un point de la cloison est barycentre d'un système m_I de masses, de signe quelconque, réparties moitié sur q, moitié sur q'). Plus généralement, on peut, en fonction de N et de d, calculer, en un certain sens, la probabilité a priori de l'existence d'une cloison. Pour cela, nous utiliserons les résultats de T.M. Cover (1965), exposés par nous en détail ailleurs (cf [Sép. Aléat.]).

Rappelons d'abord ce que Cover entend par dichotomie aléatoire. Soit p_E une loi de probabilité dans E (dim E = d); on fait sur p_E l'hypothèse fort peu restrictive, que la probabilité qu'un point appartienne à un hyperplan h quelconque, donné a priori, est nulle (i.e. $\forall$ h : p(h) = 0) cette hypothèse est en particulier réalisée si la mesure p_E a une densité continue. Tirons au hasard, suivant la loi p_E, indépendamment les uns des autres, un ensemble I de N points de E. Affectons, toujours par une suite de choix indépendants, chacun de ces points équiprobablement, soit à la classe q, soit à la classe q'. On aura ainsi construit une dichotomie aléatoire (q, q') sur un ensemble I d'effectif N. Cover montre que la probabilité p s a (N; d) qu'une telle dichotomie puisse être séparée par un hyperplan affin est (quelle que soit la mesure p_E soumise à la condition précisée; pourvu que l'affectation à q et q' soit

équiprobable) :

$$psa(N;d) = (1/2)^N \mathcal{S}(N,d+1) = (1/2)^{N-1} \Sigma\{\binom{N-1}{d'} | d' = 0, \ldots, d\},$$

formule où $\binom{n}{p}$ est le coefficient binomial $n!/(p!(n-p)!)$ (et où la notation $\mathcal{S}(N,d)$ n'est introduite que pour faciliter la référence à [Sep. Aléa.]).

En particulier, si $N = 2(d+1)$, il y a une chance sur deux pour que la dichotomie soit séparable. (On a déjà vu que si $N = d + 1$ il y a toujours séparabilité!). Supposons que N tende vers l'infini : les premiers points, tirés et affectés aléatoirement à q ou q', forment une dichotomie séparable; mais, en général, à partir d'un certain rang, il y aura trop de points pour qu'existe une cloison hyperplane séparant q de q'. De façon précise, Cover montre que l'espérance mathématique du rang P jusqu'auquel la séparation est possible (la séparation étant impossible pour les P + 1 premiers points) est égale à $2(d+1)$: pour Cover, $2(d+1)$ est la capacité de séparation affine associée à une représentation d-dimensionnelle.

De ces considérations, qui relèvent plutôt de la géométrie intégrale, revenons au problème pratique de la discrimination. Nous croyons pouvoir affirmer ceci : plus, compte-tenu des valeurs de N et d, il sera facile a priori (au sens de la géométrie intégrale) de séparer les classes actuelles q et q' (i.e. plus psa(N;d) sera élevé); moins, de l'existence d'une cloison hyperplane h séparant les classes actuelles on sera autorisé à inférer l'existence d'une cloison entre les classes potentielles infinies, et moins encore à utiliser h pour classer des individus supplémentaires (sur la distinction entre classe actuelle et classe potentielle cf § 1). Si par exemple $N = 32$ (32 individus) et $d = 15$ (15 mesures par individu); il y a, d'après Cover, une chance sur deux pour que les classes actuelles finies q et q' (supposées d'effectifs à peu près égaux, afin que l'affectation à q et q' soit bien, comme le veut Cover, équiprobable) soient séparables par un hyperplan; de l'existente d'un tel hyperplan, on se gardera donc de rien conclure. Dans un cas particulier, qui nous a été soumis par J.M. Romeder, $N = 39$, $d = 16$; un programme d'analyse discriminante fournit une cloison qui sépare q de q' à 6 erreurs près ... : ce résultat doit être regardé avec circonspection.

Dans sa thèse, Romeder propose des tables pour deux fonctions de risque : $N_r(d)$ et $d_r(N)$. La fonction $N_r(d)$ donne le nombre minimum N d'individus à partir duquel la séparabilité totale en dimension d est significative au risque r ($r = 1\%$, $r = 5\%$ etc) :

$$N_r(d) = \inf\{N | N \text{ entier};\ psa(N,d) \leqslant r\} ;$$

(en fait Romeder construit $N_r(d)$ en partant de la définition équivalente:

$$N_r(d) = \sup\{N | N \text{ entier} ;\ r <_s psa(N,d)\},$$

La fonction $d_r(N)$ donne la dimension maxima d jusqu'à laquelle la séparabilité totale d'un ensemble de N individus est significative au risque r :

$$d_r(N) = \sup\{d | d \text{ entier} ;\ psa(N,d) \leqslant r\}.$$

Voici ces deux tables :

d	1	2	3	4	5
$N_{.01}(d)$	12	15	18	20	23
$N_{.05}(d)$	9	12	14	17	19

N	... 8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	2
$d_{.01}(N)$					1	1	1	2	2	2	3	3	4	4	4	
$d_{.05}(N)$		1	1	1	2	2	3	3	3	4	4	5	...	...	...	.

On notera dans la table de d_r que la fonction $d_{.05}$ ne débute qu'à N = 9, parce que, même sur une droite, aucune discrimination significative au risque de 5 % n'est possible pour moins de 9 individus; de même $d_{.01}$ ne débute qu'à N = 12. Le calcul de $d_{.01}$ (fait d'après la table de la fonction $N_{.01}$) n'a pas été poussé au delà de N = 19.

Ces tables fourniront, croyons-nous, d'utiles indications, dont la portée est toutefois limitée pour plusieurs causes que Romeder signale lui-même. Nous citerons :

1) l'épreuve suppose l'équiprobabilité des deux classes.

2) Elle ne considère que le cas de la séparabilité totale.

3) Pratiquement, il se peut qu'il y ait séparabilité totale, bien que la meilleure cloison qu'on ait su construire ne sépare pas parfaitement q de q'.

4) On n'a rien dit du cas de plus de deux classes ($2 <_s$ Card Q) (cf. toutefois § 3.2). Quant à généraliser les travaux de Cover pour répondre à 1, 2 ou 4, cela ne semble pas facile !

8. Etude comparative du potentiel de discrimination :

L'efficacité d'un algorithme de discrimination ne se mesure pas tant par son aptitude à séparer des classes actuelles données que par son rendement potentiel sur l'ensemble infini des individus nouveaux qu'on doit rapporter à ces classes. Faute de pouvoir explorer l'infini, on extrait par tirage au sort, de l'ensemble I des données actuelles, un sous-ensemble I_2 appelé échantillon d'épreuve sur lequel on calcule le taux d'efficacité des cloisons mises en place d'après les données de $I_1 = I - I_2$. (cf. § 6). Ainsi on a pu, sur deux jeux de données, comparer le potentiel de discrimination offert par l'analyse des correspondances à celui des programmes choisissant les variables pas à pas.

8.1. Les professions de foi des députés élus en 1881 :

Ces données sont étudiées ailleurs en détail (cf. [Députés 81] TII C n° 2) : rappelons seulement ici que sur les professions de foi des 550 députés d'un ensemble D, on a compté les occurences des 53 mots d'un ensemble M. L'analyse du tableau de correspondance k_{DM} (k(d,m) = nombre de fois que le mot m apparait dans la profession de foi du député d) donne un premier facteur qui exprime nettement l'opposition droite ≠ gauche. Plus précisément, divisons les députés en trois fractions d'après les étiquettes politiques qur leur attribue le Dictionnaire des parlementaires :

D_G = {radicaux, extrême-gauche} = la gauche ; Card D_G = 87;

D_C={ferristes, gambettistes, républicains modérés}=le centre;Card D_C=382; D_D={monarchistes, bonapartistes, conservateurs}= la droite; Card D_D= 81.

Alors il apparait que, une fois ôté le centre D_C , l'origine sépare presque exactement sur le 1° axe la droite D_D de la gauche D_G : on a, à 6 % d'erreur près environ :

$$D_D \approx \{d | d \in D_D \cup D_G ;\ F_1(d) < 0\};\quad D_G \approx \{d | d \in D_D \cup D_G ;\ F_1(d) > 0\}.$$

Voyant cette belle discrimination, nous avons suggéré à J.M. Romeder de la retrouver par les méthodes dont il a la pratique. Romeder a appliqué deux programmes : l'un, qui court sous le sigle BMD 07 M, utilise l'hyperplan médiateur de Fisher avec, pour choisir les variables pas à pas, un critère suggéré par des hypothèses de normalité; l'autre dû à Romeder lui-même, repose sur les principes de Sebestyen (cf § 5) et a pour critère de choix le taux de succès (cf § 6). Opérant sur l'ensemble $I = D_G \cup D_D$ dont le cardinal est 168, Romeder a réservé des échantillons d'épreuve I_2 comptant de 17 à 50 individus; les résultats des divers essais étant analogues, nous n'en reproduisons qu'une partie. On lit sur les courbes de la figure 6, (relatives à plusieurs cas, différant quant au programme ou à la valeur du rapport β = Card I_2/Card I), la variation des taux d'échecs τ_1 et τ_2 sur I_1 et I_2 respectivement au fur et à mesure que croît le nombre ν des variables explicatives (qui sont ici, les fréquences k(d,m)/k(d) des mots m). Notons que le taux d'échecs, qui sur I_1 devient bientôt inférieur à 8 %, ne descend guère, sur l'échantillon d'épreuve I_2 , au dessous de 15 % (le taux 10 % est atteint deux fois), le minimum étant obtenu avec 10 variables explicatives environ.

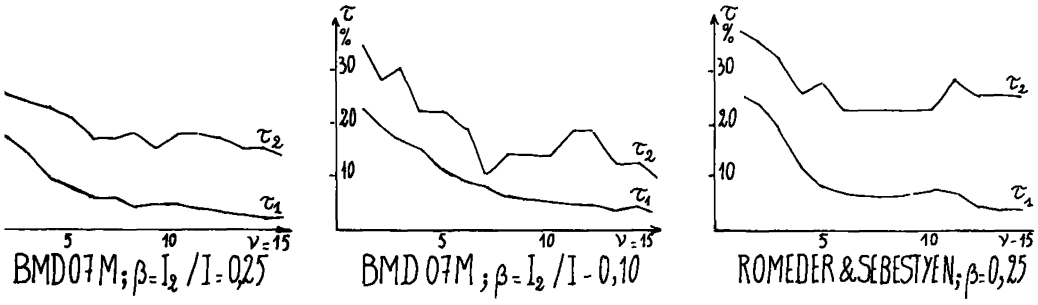


Fig 6 : taux d'échecs τ_1 et τ_2 sur I_1 et I_2 respectivement en fonction du nombre ν des variables explicatives (d'après J.M. Romeder)

Pour achever la comparaison avec l'analyse factorielle, il restait à soumettre cette méthode à des épreuves de stabilité : ce que M. Danech-Pejouh a très bien fait dans sa thèse en analysant des sous-tableaux k_{D1M} , où D_1 est un sous-ensemble de D extrait par tirage au sort. Les facteurs F'_α issus de k_{D1M} sont définis sur D_1 , mais on les étend à D tout entier en traitant comme des éléments supplémentaires les individus de l'échantillon d'épreuve $D_2 = D - D_1$, suivant la formule usuelle :

$$F'_\alpha(d) = \lambda'_\alpha{}^{-1/2} \Sigma\{(k(d,m)/k(d))\ G'_\alpha(m) \mid m \in M\},$$

(où λ'_α est la valeur propre et $G'_\alpha(m)$ le facteur sur M issu de k_{D1M} ; tandis que k(d,m)/k(d) est la fréquence du mot m dans la profession de foi de l'individu supplémentaire d $\in D_2$). Et voici ce qu'on observe : tant que le rapport β = (Card D_2/ Card D) ne dépasse pas 0,5, le facteur F'_1 issu de k_{D1M} demeure corrélé à plus de 0,95 avec le facteur F_1 originel (issu de k_{DM} tout entier); pour β = 0,8 (c'est-à-dire pour un tableau k_{D1M} relatif à 110 députés seulement) la corrélation n'est plus que de 0,85. Un fait curieux est que quand β croît de 0,5 à 0,9 les premières valeurs propres croîssent, ainsi que la part d'inertie qui leur revient. Quant à la discrimination entre D_D et D_G fournie par le 1° facteur, elle reste bonne jusqu'à β = 0,9. La règle :

$$D_D \approx \{d \mid d \in D_D \cup D_G\ ; F'_1(d) < 0\};\ D_G \approx \{d \mid d \in D_D \cup D_G : F'_1(d) > 0\}$$

(appliquée en choisissant convenablement le signe de F'_1) donne sur $(D_D \cup D_G) \cap D_1$ et sur $(D_D \cup D_G) \cap D_2$ des taux d'erreurs τ_1 et τ_2 peu différents entre eux et toujours inférieurs à 15 %. Il est particulièrement frappant qu'un facteur F'_1 calculé par analyse d'un tableau D_1 x M relatif à 55 députés seulement, parmi lesquels il y en a au plus 10 de chacun des deux groupes D_D et D_G de la droite et de la gauche, permette de discriminer entre D_D et D_G sans faire, au total, plus de 24 erreurs.

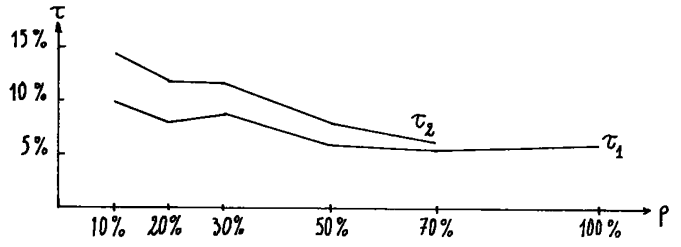


Figure 7 : taux d'échecs τ_1 et τ_2 en fonction du pourcentage p des députés dont les données ont constitué le tableau soumis à l'analyse factorielle (d'après M. Danech Pajouh)

8.2. *Etude comparative du Liban rural en 1960 et 1970* :

Cette étude fait l'objet d'un rapport détaillé (cf. [Liban 60-70] TI: C n° 5); nous nous bornerons ici à des rappels. En 1960 l'association IRFED, qu'animait alors son fondateur, le R.P. L.J. Lebret, fut chargée de dresser un inventaire des possibilités du Liban et de ses besoins. D tout ce travail nous ne considérons ici qu'un tableau de chiffres, 60 x 155, donnant pour 60 villages en réponse à 155 questions, une note comprise entre 0 et 4 : ainsi le village de Qartaba, désigné par le sigle 1QA, a, à la question "Usage des insecticides et des fongicides" (question numérotée 209, parce qu'elle est la 9° du 2° chapitre du questionnaire), la note k(1QA, 209) = 3, etc. L'enquête fut répétée en 1970

sous la direction de R. Delprat, qui était en 1960 l'adjoint du R.P. Lebret : d'où un second tableau de notes, relatif au même ensemble de 60 villages, mais avec un ensemble réduit, R^+, de 59 questions. L'étude comparative repose donc sur un tableau, appelé $k_{VV'R^+}$, qui a 59 colonnes (les 59 questions communes aux deux enquêtes) et 120 lignes dont chacune est une description de village. Chaque village fournit deux lignes : d'une part sa description (vecteurs de 59 notes) pour 1960; et d'autre part sa description pour 1970; on désigne l'état 1960 par un sigle formé de l'un des chiffres 1, 2, 3, 4 suivi de deux lettres et l'état 1970 par ce chiffre augmenté de 4 suivi des deux mêmes lettres : e.g. en 1960 : 1QA; en 1970 : 5QA. On note V l'ensemble des descriptions de villages faites en 1960; et V', l'ensemble des descriptions faites en 1970.

Le tableau $k_{VV'R^+}$ n'est pas soumis tel quel à l'analyse factorielle. A chaque question q^+ de R^+, on adjoint une question complémentaire q^-. Le sigle de la question s'obtient en remplaçant le premier chiffre du sigle de q^+ par la lettre de même rang : ainsi la question $q^+ = 209$, a pour complémentaire $q^- = B09$; quant à la note q^- elle est le complément à 4 de la note q^+ : e.g., puisque k(1QA, 209) = 3 on pose k(1QA, B09) = 4 - 3 = 1. Le tableau des notes de 1960 comporte d'assez nombreuses omissions : quand la note $k(v,q^+)$ manque on pose $k(v,q^+) = k(v,q^-) = 0$; ainsi on a : k(1RE,201) = k(1RE,B01) = 0 parce que, en 1960, le village 1RE ne reçut pas de note à la question 201 (Outillage agricole). Par le dédoublement, les effets des omissions sont atténuées en ce qu'on évite de confondre une omission ($k(v,q^+) = k(v,q^-) = 0$) avec une note nulle ($k(v,q^+) = 0$; $k(v,q^-) = 4$); mais des omissions nombreuses restent très nuisibles : aussi doit-on mettre à part les couples de notes :

102 Non-mortalité infantile ≠ A02 mortalité infantile,

201 Outillage agricole ≠ B01 sans outillage agricole.

ainsi que la description en 1960 du village de Hadeth el Jobbeh (2HE). De plus une question se signale par des omissions systématiques et significatives : c'est 803, "Sans polygamie" : ne sont pas notés à cette question les villages dont la population est quasi-exclusivement chrétienne. Ainsi a-t-on affecté à cette question trois colonnes : 803, H03, CHR : un village, tel que Qartaba, qui n'a pas de note à la question 803 en 1960 est coté en 1960, (1QA), comme en 1970 (5QA) par :

k(1QA, 803) = k(1QA, H03) = 0; k(1QA, CHR) = 4;

k(5QA, 803) = k(5QA, H03) = 0; k(5QA, CHR) = 4.

Un village musulman tel que 1CH (état 1960) reçoit 0 en CHR et des notes complémentaires en 803 et H03 :

k(1CH, 803) = 3; k(1CH, H03) = 1; k(1CH, CHR) = 0

Les trois colonnes {803, H03, CHR} apportent toutes des informations intéressantes; mais elles ont une place à part. Enfin l'analyse factorielle a amené à écarter de certaines analyses les couples :

710 Bibliothèque ≠ G10 sans bibliothèque ;

715 Cinéma ≠ G15 sans cinéma ;

On a donc un tableau de base $k_{VV'R}{}^{+}$: 120 x 59; un tableau complété $k_{VV'R}$: 120 x 119 (119 = 2 x 58 + 3, parce que la question 803 occupe trois colonnes); de quoi il faut écarter la ligne 2HE et les 11 colonnes {102, A02, 201, B01, 710, G10, 715, G15, 803, H03, CHR} : reste donc un tableau $k_{VV'K}$ 119 x 108; (V ∪ V' comprend 119 descriptions : 59 pour 1960 et 60 pour 1970; K comprend 108 questions, i.e. 54 couples).

Après ces rappels sommaires, extrayons de la thèse de T. Moussa ce qui touche à la présente leçon : la discrimination entre les descriptions de village faites en 1960 et celles faites en 1970.

8.2.1. Discrimination par l'analyse factorielle :

Soumettons à l'analyse factorielle le tableau $k_{VV'K}$ 119 x 108 non sans mettre en éléments supplémentaires la ligne 2HE et les 11 colonnes écartées (cf [Liban 60-70] § 2). Sans qu'une discrimination nette semblable à celle notée ci-dessus (§ 8.1) apparaisse sur aucun axe, descriptions de 1960 et descriptions de 1970 se trouvent avec une densité maxima dans des quadrants opposés des plans 1 x 2, 1 x 3, 1 x 4. D'une part sur l'axe 1 les descriptions de 1970 devancent, en moyenne, les descriptions de 1960 dont le niveau général est moindre. D'autre part à niveau égal les descriptions de 1970 n'ont pas même caractère que les descriptions de 1960 : la part relative de l'équipement est plus forte là qu'ici etc; d'où des décalages aussi sur les axes 2, 3, 4. Ce qui laisse espérer qu'on puisse, par une cloison hyperplane, séparer le sous-nuage V des descriptions de 1960, du sous-nuage V' des descriptions de 1970.

Projetons les nuages de descriptions et de questions dans l'espace R_A engendré par un ensemble A d'axes factoriels (e.g. par les quatre premiers axes factoriels : A = {1, 2, 3, 4}). Une description v se projette au point $v_A = \{F_\alpha(v) \mid \alpha \in A\}$; une question q($q^+$ ou q^-) se projette en $q_A = \{G_\alpha(q) \mid \alpha \in A\}$. Le nuage des descriptions $\{v_A \mid v \in V \cup V'\}$ (comme aussi le nuage des questions) a pour axes principaux d'inertie les axes factoriels, et pour moments principaux d'inertie les valeurs propres λ_α : dans R_A , la forme quadratique d'inertie σ_{AA} est diagonale

$$\alpha, \alpha' \in A : \sigma_{\alpha\alpha'} = \delta_{\alpha\alpha'} \lambda_\alpha.$$

De même est diagonale la métrique (σ^{-1}) pour laquelle le nuage V ∪ V' a variance 1 dans toutes les directions.

$$\forall \alpha, \alpha' \in A : (\sigma^{-1})^{\alpha\alpha'} = \delta^{\alpha\alpha'}/\lambda_\alpha$$

Le produit scalaire entre deux vecteurs x_A et y_A est :

$$(\sigma^{-1})^{AA} x_A\, y_A = \langle x_A , y_A \rangle = \Sigma \{x_\alpha\, y_\alpha/\lambda_\alpha \mid \alpha \in A\}.$$

Notons $V_\alpha = F_\alpha(G60)$ la coordonnée sur l'axe factoriel α du centre de gravité des descriptions de villages faites en 1960; on a :

$$V_\alpha = (1/\text{Card } V)\ \Sigma \{F_\alpha(v) \mid v \in V\}; \quad V_A = \{V_\alpha \mid \alpha \in A\}.$$

Notons de même $V'_\alpha = (1/\text{Card } V')\ \Sigma \{F_\alpha(v) \mid v \in V'\} = F_\alpha(G70)$. Dans l'espace R_A , muni de la métrique $(\sigma^{-1})^{AA}$, la variance interclasse est :

$$(\text{Card } V\ \text{Card } V'/(\text{Card } V + \text{Card } V')^2)\ \Sigma \{(V'_\alpha - V_\alpha)^2/\lambda_\alpha \mid \alpha \in A\}$$

Pour A =]5], (on sait qu'on note]n] la suite des entiers 1,...,n), cette variance vaut 0,81; les contributions des facteurs successifs de rang 1 à 5 sont : 0,24; 0,35; 0,13; 0,07; 0,02; l'interprétation s'étant arrêtée au 4° axe, il ne semble pas utile d'aller au delà. Projetons orthogonalement $V \cup V'$ sur l'axe $V_A\ V'_A$ joignant les centres de gravité des deux classes V et V' : l'abscisse d'un point v (comptée dans l'unité de longueur fournie par la métrique $(\sigma^{-1})^{AA}$) est :

$$D(v) = \langle v_A\ ,\ V'_A - V_A \rangle / \|V'_A - V_A\|$$
$$= \Sigma\ \{F_\alpha(v)(V'_\alpha - V_\alpha)/\lambda_\alpha \mid \alpha \in A\} / (\ \Sigma\ \{(V'_\alpha - V_\alpha)^2/\lambda_\alpha \mid \alpha \in A\})^{1/2}.$$

Projeté sur l'axe $V_A\ V'_A$, le nuage $V \cup V'$ a la variance 1 (comme sur tout autre axe de R_A muni de la métrique $(\sigma^{-1})^{AA}$). Le rapport ρ de la variance interclasse à la variance totale est donc 0,81 si A =]5] et 0,79 si A =]4].

Reportons-nous à la figure 1 du paragraphe 3 : le maximum du rapport ρ pour une distribution unimodale est 0,75 : ceci annonce pour D(v), dans les cas A =]4] et A =]5], un histogramme bimodal; et laisse espérer que les deux classes de l'histogramme correspondent à peu près à V et V' :

$$V^- = \{v \mid v \in V \cup V';\ D(v) < 0\} \ \#\ V;\ V^+ = \{v \mid v \in V \cup V';\ 0 < D(v)\} \ \#\ V'$$

Sur la figure 8, on a représenté le cas A =]4] : on voit que : Card $(V^- \cap V') = 5$; Card $(V^+ \cap V) = 2$. Il y a donc, dans $R_{]4]}$, 94 %, (i.e. 112/119), d'individus (descriptions de villages) bien classées si on adopte comme cloison séparatrice l'hyperplan D = 0, perpendiculaire à l'origine à $V_A\ v'_A$. Cette cloison n'est autre que l'hyperplan de Fisher; à ceci près qu'elle passe par l'origine, centre de gravité du nuage, qui ne coïncide pas exactement avec le milieu de $V_A\ V'_A$, parce que, V compte un individu de moins que V' (2HE étant un élément supplémentaire); ce détail minime n'influe pas sur le classement. Avec les trois premiers facteurs, A =]3], on a un classement moins bon : 12 erreurs; à adjoindre le 5° facteur, A =]5], on ne gagne rien : 7 erreurs. Le 5° facteur n'ayant pas été interprété (cf [Liban 60, 70] § 2) on adopte A =]4].

Il est possible de projeter les questions q_A sur l'axe $V_A\ V'_A$, dans R_A muni de la métrique $(\sigma^{-1})^{AA}$; on a comme pour le calcul de D(v) :

$$D(q) = \Sigma\ \{\lambda_\alpha^{-1}(V'_\alpha - V_\alpha)\ G_\alpha(q) \mid \alpha \in A\} / \|V'_A - V_A\|;$$

ainsi doivent apparaître les traits par lesquels 1970 se distingue le plus de 1960. On prendra toutefois garde que pour les projections des nuages sur un axe quelconque tel que $V_A\ V'_A$ qui n'est pas un axe factoriel, il n'y a pas de formule barycentrique rigoureuse. On s'intéressera donc, aussi bien qu'à D(q), au coefficient CD(q) de k(v,q)/k(v) dans l'expression de D(v). On a :

$$D(v) = \Sigma\ \{\lambda_\alpha^{-1}(V'_\alpha - V_\alpha)\ \Sigma\ \{(k(v,q)/k(v))G_\alpha(q)\lambda_\alpha^{-1/2} \mid q \in Q\} \mid \alpha \in A\} / \|V'_A - V_A\|;$$

$$= \Sigma\ \{(k(v,q)/k(v))\ \Sigma\ \{\lambda_\alpha^{-3/2}(V'_\alpha - V_\alpha)G_\alpha(q) \mid \alpha \in A\} \mid q \in Q\} / \|V'_A - V_A\|;$$

$$CD(q) = \Sigma\ \{\lambda_\alpha^{-3/2}(V'_\alpha - V_\alpha)\ G_\alpha(q) \mid \alpha \in A\} / \|V'_A - V_A\|.$$

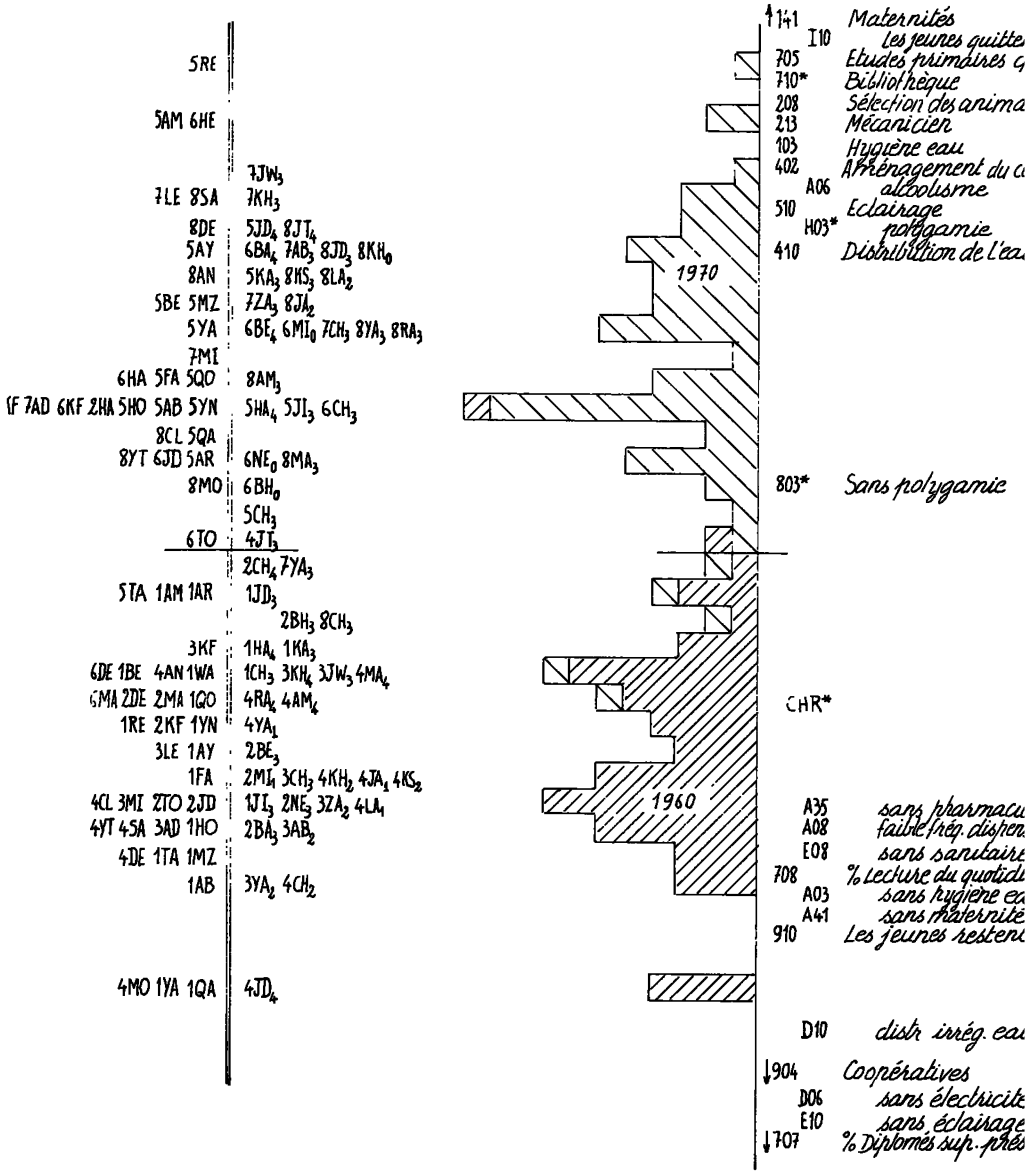


Figure 8 : Projection des deux nuages sur l'axe $V_A V'_A$ (A = 1, 2, 3, 4). On a précisé à gauche le déta des sigles des villages présumés chrétiens et de ceux présumés musulmans : ces derniers ont en indice leur note à la question 803 : polygamie (0 : pol. fréquente ; 4 absence de pol.). A droite on a donné, en face de l'histogramme des villages (diversement hachurés selon l'année) un extrait du nu age Q.

L'une ou l'autre formule (D(q) ou CD(q)) signale ici les mêmes traits. Du côté négatif (1960) on a surtout des déficiences (q^-); du côté positif (1970), on a des avantages (q^+). Mais on voit sur la figure 8 quelques exceptions que nous avions déjà signalées (cf [Liban 60-70]) :

106 (Non-alcoolisme), 707 (Présence de diplômés supérieurs), 708 (Lecture du quotidien), 904 (Coopératives), 910 (Les jeunes restent) vons avec V (1970).

A06, G07, G08, I04, I10; leurs antagonistes, vont avec V' (1970). (Le plus léger des deux points de chaque couple q^+, q^- étant, en vertu du principe du bras de levier, cf e.g. TII C n° 1, le plus écarté).

La place des éléments supplémentaires CHR, 803, H03 mérite un commentaire. On s'étonne que le point CHR s'écarte sensiblement de l'origine du côté de V (1960) : en effet; les villages présumés chrétiens sont les mêmes en 1960 et 1970; il n'y a pas changement en la matière. Mais observons les histogrammes de détail des villages chrétiens et musulmans : au sein de chacune des classes V (1960) et V'(1970), il semble que du centre des villages chrétiens au centre des villages musulmans on se déplace dans la direction V V'. Corrélativement, les points 803 et H03, relatifs aux villages musulmans sont du côté V' (1970), le plus écarté étant H03; ce qui s'explique en partie par un léger accroissement de la polygamie de 1960 à 1970. Cependant sans rejeter les explications que nous proposons, il faut répéter que sur l'axe V_A V'_A le principe barycentrique ne joue pas rigoureusement. Quoiqu'il en soit de l'explication, le lien paradoxal du trait constant CHR avec la discrimination entre 1960 et 1970, réapparaît au paragraphe 8.2.2. où un programme de choix pas à pas propose cette même note (CHR).

8.2.2. Discrimination en fonction de variables choisies pas à pas :

Le programme de discrimination MAHAL 2, dû à Romeder et utilisé par T. Moussa construit un hyperplan de Fisher. En cela, il a même fondement que les calculs de discrimination faits au paragraphe 8.2.1; la différence est dans le choix des variables explicatives : en 8.2.1. ces variables explicatives sont des facteurs; en 8.2.2. ce sont des notes du questionnaire (variables primaires) choisies pas à pas pour que croisse au mieux le rapport ρ de la variance interclasse à la variance totale (cf. § 3 et § 6).

Les résultats fournis par MAHAL 2 sont résumés sur la figure 9. Tareck Moussa a fait plusieurs analyses soit avec un tableau de notes dédoublées soit avec le tableau des 59 notes q^+; il a tantôt gardé, tantôt omis, le village 2HE, ainsi que les 11 notes écartées de l'analyse factorielle.

Dans les analyses fondées sur un tableau dédoublé le taux d'erreur ne descend pas au dessous de 10 % (contre 6 % en analyse factorielle), les premières variables choisies sont généralement parmi celles qui sur l'axe V_A V'_A de l'analyse factorielle (cf. fig. 9) occupent une position écartée; après le 10° choix les notes CHR ou 803 s'introduisent, paradoxe déjà signalé en 8.2.1.

Dans les analyses fondées sur un tableau non-dédoublé, jouent un rôle important, les notes 102 et 201 qui, on l'a dit, présentent en 1960 un grand nombre d'omissions : en effet, l'omission étant codée 0, comme s'il s'agissait d'une déficience extrême, ces notes créent entre 1960 et 1970 une forte dénivellation fictive dont le programme tire parti. Le résultat pour la discrimination est médiocre quand on écarte 2HE (15 % d'erreur); il se trouve excellent (0 % d'erreur au 12° pas) quand on conserve cette description 2HE, qui présente elle-même de nombreuses omissions. Il est clair que ce succès est peu démonstratif.

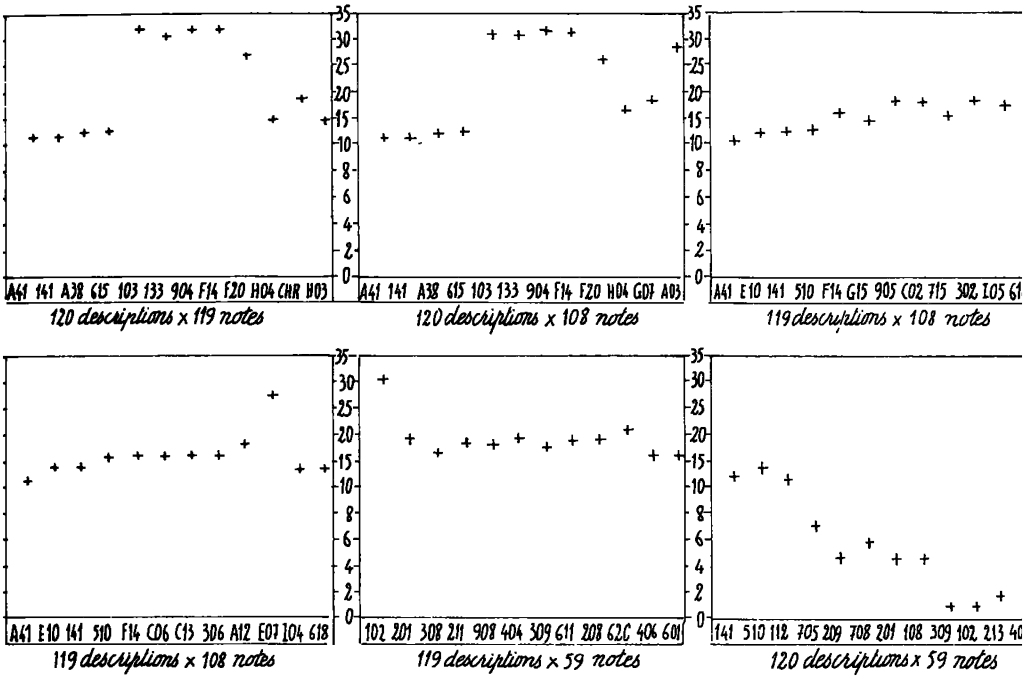


Figure 9 discrimination des villages par le programme MAHAL2 de J.-M. Romeder ; on a noté sur l'axe des abscisses, de gauche à droite, la suite des variables introduites pas à pas; et porté en ordonnée le pourcentage d'erreur après chaque pas. Il importe de prendre garde à la graduation non uniforme des ordonnées.

8.2.3. *Discrimination contrôlée par un échantillon d'épreuves* :

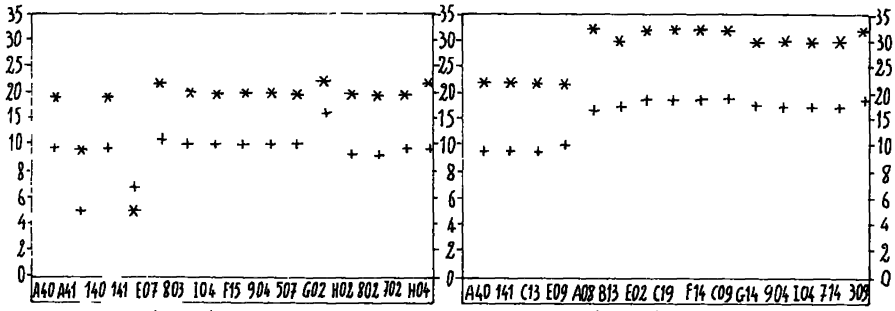
Afin d'estimer la stabilité des taux d'erreurs obtenus par les diverses méthodes, T. Moussa a fait des analyses discriminantes en réservant un échantillon d'épreuves (cf. § 6).

Sur 120 descriptions, on en a réservé 40 dont 20 prises dans V(1960) et 20 prises dans V'(1970). Soit, ce qui laisse un échantillon de base plus équilibré, en tirant au sort 20 villages, et retirant les 2 descriptions de chacun d'eux; soit en tirant au sort successivement 40 villages d'épreuve : 20 pour 1960 et 20 pour 1970.

L'analyse de correspondance a été appliquée soit avec un tableau de 119 notes, soit avec 112 notes seulement (il eût mieux valu n'en garder que 108). D'où 4 essais dont les résultats sont les suivants :

analyse erreurs	(2x(60-20))x112	(2x(60-20))x119	(120-40)x112	(120-40)x119
sur I_1	92,5 %	88,75 %	93,75 %	87,5 %
sur I_2	92,5 %	85 %	85 %	80 %

(ici on a appelé I_1 l'échantillon base; I_2 l'échantillon d'épreuve; on a noté 2 x (60-20) quand les mêmes 20 villages sont retirés de V et de V' et (120-40) quand il y a 40 tirages indépendants).



2(60-20) descriptions x 115 notes (120-40) descriptions x 115 notes

*Figure 10. discrimination par le programme MAHAL 2 avec échantillon d'épreuve; l'ordonnée du point + est le pourcentage d'erreur sur l'échantillon de base; l'ordonnée du point * est le pourcentage d'erreur sur l'échantillon d'épreuve. On a noté 2(60-20) quand les mêmes 20 villages sont retirés de V et de V'; et (120-40) quand il y a 40 tirages indépendants*

Le programme MAHAL 2 n'a été appliqué qu'à des tableaux de notes dédoublés afin d'éliminer l'effet des omissions : les résultats sont dessinés sur la figure 10.

En moyenne, l'analyse factorielle semble fournir à la discrimination la base la plus stable à condition d'écarter les notes comportant des omissions.

8.2.4. Recouvrement entre les deux classes à discriminer :

On a vu en 8.2.1. que dans l'espace engendré par les 4 ou 5 premiers axes factoriels, la discrimination entre descriptions faites en 1960 et descriptions faites en 1970 n'est réalisée qu'à 7 erreurs près :

D'une part les descriptions 5TA, 6MA, 6ME, 7YA, 8CH, faites en 1970 sont affectées à 1960;

d'autre part les descriptions 2HA et 4JT (auxquelles il faut adjoindre la description 2HE qui a de multiples omissions) sont faussement affectées à 1970. La programme de discrimination MAHAL 2 (qui use de variables choisies pas à pas) reproduit la plupart de ces confusions.

Il ne semble pas qu'on doive pour ces erreurs d'affectation répétées blâmer les méthodes d'analyse discriminante : on en conclura plutôt que les classes qu'on a tenté de discriminer se recouvrent quelque peu. Pour confirmer cette interprétation, R. Delprat a demandé à T. Moussa de reprendre l'analyse factorielle en mettant en éléments supplémentaires les 8 descriptions citées ci-dessus (ainsi que les 11 colonnes de notes généralement écartées par nous).

Dans l'espace engendré par les 4 ou 5 premiers axes issus de ce tableau 112 x 108, la discrimination est parfaite pour les 112 descriptions de base; mais les 8 descriptions de villages mises en éléments supplémentaires sont ici encore mal classées (chacune est avec la période qui n'est point la sienne).

9. *Conclusion : discrimination après analyse factorielle* :

Il est pour nous hors de doute que, dans toute étude de discrimination reposant sur des données multidimensionnelles une analyse factorielle doit précéder l'application d'une méthode d'analyse discriminante proprement dite. Après l'analyse factorielle, la discrimination peut être en évidence (cf [Députés 81]) sur un axe ou dans un plan. Sinon on appliquera une méthode de discrimination simple dans l'espace rapporté aux premiers axes factoriels : ainsi on a vu que l'hyperplan médiateur de Fisher sépare de façon satisfaisante (dans R^4 ou R^5) les descriptions de villages faites en 1960 de celles faites en 1970 (cf [Liban 60-70]); de plus, il est très simple de calculer l'équation de cet hyperplan; car si on utilise les facteurs pour coordonnées, la forme d'inertie $\sigma(I)$ est diagonale. Toute méthode de discrimination jouera mieux sur les facteurs, stables et peu nombreux, que sur les multiples mesures primaires. Si on désire choisir parmi les facteurs jugés significatifs, dont le nombre est e.g. 5, ceux en fonction desquels effectuer la discrimination, il ne sera pas nécessaire de recourir au choix pas à pas (cf. § 6) : on pourra généralement rechercher l'optimum absolu d'un critère C calculé sur toutes les parties de l'ensemble des facteurs (qui servent de variables explicatives) car, e.g. $2^5 = 32$.

Parfois, après avoir réduit la dimension (de d = Card J, à moins de 5), on se permettra de l'augmenter, sans pourtant dépasser une borne, (e.g. 10), en prenant comme nouvelles variables explicatives, outre les facteurs, des monômes en ceux-ci (e.g. leurs carrés, leurs produits deux à deux). Une séparation linéaire par rapport à l'ensemble de ces variables équivaut en effet à une séparation polynomiale en les seuls facteurs : ainsi sans cesser d'utiliser un programme de séparation linéaire, on disposera d'une plus grande diversité de cloisons. Nous avons évoqué ailleurs la méthode générale de la régression polynomiale (cf. [Alg. Eucl.] TII B n° 12 § 2.2.2); méthode qu'il faut utiliser avec prudence (cf. [Rég.] § 2) car le nombre des variables explicatives (fussent-elles des monômes les unes des autres) ne peut croître indéfiniment.

En tout cas, on contrôlera la stabilité de la règle de discrimination adoptée en réservant un échantillon d'épreuve.

Mais une fois admis que le calcul des facteurs donne à la discrimination la base la plus sûre, on objectera que, sur un élément supplémentaire s, ce calcul ne peut se faire (selon la formule usuelle d'insertion des éléments supplémentaires cf. e.g. [Prat. Corr.] TII A n° 2 §2.4) qu'en fonction de l'ensemble J de toutes les mesures; tandis qu'en choisissant pour la discrimination un sous-ensemble H de variables explicatives (cf. § 6) on réduit le coût du classement d'un nouvel individu.

Voici comment, dans le cadre de l'analyse factorielle, on réduira l'ensemble J de variables explicatives primaires utilisées pour classer un individu supplémentaire. D'une part (cf. [Prat. Corr.] § 3.5) l'analyse factorielle signale elle même les mesures primaires qui apportent aux facteurs discriminants les plus fortes contributions : on refera donc l'analyse factorielle en écartant les autres mesures, et vérifiera que la discrimination demeure possible (elle sera en fait souvent meilleure après suppression des mesures irrelevantes). D'autre part il est possible de faire, avant toute analyse, un premier tri des variables de l'ensemble J.

Communément, pour une variable primaire continue j, on calcule la variance intraclasse $\sigma_{jj}(Q)$ et la variance interclasse $\sigma_{jj}(I-Q)$ et compare :

$$T_j = \sigma_{jj}(Q)\,(\text{Card } Q - 1)^{-1/2} / (\sigma_{jj}(I - Q)\,(\text{Card } I - \text{Card } Q)^{-1/2}).$$

aux valeurs tabulées d'une loi de Fisher-Snedecor à (Card Q - 1; Card I - Card Q) dimensions; cette première épreuve bien que fondée sur des hypothèses de normalité excessives, garde une valeur indicative. Ou l'on recourt à l'épreuve ordinale de Kruskal-Wallis; comparer :

$$H_j = -3(N+1) + 12/(N(N+1))\ \Sigma\{(\Sigma\{R(i) \mid i \in q\})^2/\text{Card } q \mid q \in Q\}$$

(où N = Card I et R(i) est le rang de la mesure i_j dans l'ensemble $\{i_j \mid i \in I\}$ des valeurs) à un χ^2 à (Card Q - 1) dimensions; épreuve qui ne requiert pas d'hypothèses de normalité; mais si les effectifs Card q de certaines classes sont faibles il faut substituer au χ^2 une loi dont les tables sont peu disponibles. Ici, cependant dans l'une comme dans l'autre épreuve les tables sont à peu près inutiles : en effet nous ne désirons pas choisir les variables explicatives d'après un seuil de signification posé a priori, mais prendre, dans J, des variables en nombre fixé a priori (e.g. 10) qui soient les plus relevantes : ce qui se fera en prenant les variables qui donnent les plus fortes valeurs de T_j ou de H_j.

Pour une variable logique v, dont les modalités sont codées sur un ensemble J_v de colonnes ($J_v \subset J$) on applique l'épreuve du χ^2 au tableau de contingence Q x J_v (où k(q,j) est le nombre des individus de I pour lesquels la variable v prend la valeur j).

Rappelons enfin qu'en analyse discriminante comme en régression, on a pu chercher (principalement dans l'espace rapporté aux premiers axes factoriels : e.g. dans le plan 1 x 2) les individus de l'échantillon de base qui sont les plus proches voisins d'un individu supplémentaire s qu'on cherche à déterminer. Cette méthode souvent utile (dite : discrimination par boule parce que les plus proches voisins sont recherchés dans une boule de centre s) est exposée en ses grandes lignes dans [Histoire V] §3.8.2 (Ca TII n° 1 p. 34) et la notice d'un programme de M.O. Lebeaux pour la recherche des voisins est publiée sous le titre [POUBEL] (cf ce cahier pp. 467-481.

Ainsi sur la base stable de l'analyse factorielle, on pourra, en s'aidant éventuellement d'autres techniques, résoudre le problème de la discrimination non sans satisfaire à des exigences, souvent légitimes, d'économie.

Note : Discrimination et régression :

Quant au but poursuivi le problème de la discrimination est, nous l'avons dit (§ 1 p 370), un cas particulier de celui de la régression : celui où la variable à expliquer prend ses valeurs dans un ensemble fini: l'ensemble Q des classes ; et en effet, lorsqu'on procède par recherche des plus proches voisins, (cf POUBEL , § 1, p 467), l'ensemble Q apparaît comme un cas particulier de l'ensemble des modalités des variables à expliquer. Cependant, comme nous l'a signalé L. Lebart, les techniques utilisées en discrimination dans le cas de plus de deux classes sont au contraire souvent plus générales que celles qui servent à la régression. Par exemple (cf §§ 2.2 & 2.3) quand on cherche une représentation plane du nuage des individus (c'est-à-dire deux coordonnées qui sont des combinaisons linéaires des variables explicatives) rendant max. le rapport de la variance interclasse à la variance totale, on rencontre la décomposition simultanée de deux formes quadratiques (analyse factorielle), question qui, dans une présentation géométrique convenable apparaîtra comme l'étude de la figure formée par deux sous-espaces d'un espace euclidien, avec recherche des couples de vecteurs formant un angle minimal

(ce qui est fait dans l'a. canonique de Hotelling). Toutefois cette même question géométrique se retrouve en régression si l'on pose e.g. le problème statistique suivant : étant donné n variables explicatives x^i et p variables à expliquer y^j ; trouver 1 ou 2 (ou 3) combinaisons linéaires des x^i en fonction desquelles tous les y^j puissent simultanément s'exprimer au mieux. Ainsi le conflit de terminologie se résout si l'on distingue entre le problème statistique qui est le but même de la recherche, et les questions mathématiques qui dans cette recherche ne sont que des outils (pris à la géométrie et à l'algèbre linéaire).

BIBLIOGRAPHIE

J.P. Benzécri : Séparation linéaire et méthode des moindres carrés (1967).

J.P. Benzécri : Sur la stabilité des résultats dans l'analyse discriminante (1968).

T.M. Cover : Geometrical and statistical properties of systems of linear inequalities with applications in Pattern Recognition; *IEEE transactions on electronic computers*, pp. 326-334 Juin (1965). (cf. notre exposé : Dichotomie aléatoire et séparation linéaire; [Sep. Aléat.].

M. Danech Pejouh : Etude de la stabilité des facteurs en analyse des correspondances; Thèse 3° cycle Paris (1972).

R.A. Fisher : The use of multiple measurements in taxonomic problems; *Annals of Eugenics;* Vol. 7; pp. 179-188; (1936).

T. Moussa : Stabilité dans le calcul des facteurs et discrimination en analyse des correspondances. Thèse 3° cycle Paris (1972).

J.M. Romeder : Méthodes de discrimination; Thèse 3° cycle Paris (1969)

J.M. Romeder et Coll. : *Méthodes et programmes de l'analyse discriminante;* Dunod, Paris, Bruxelles, Montréal; (1973); ce livre est une version revue de la thèse de J.M. Romeder

G.S. Sebestyen : *Decision making processer in pattern recognition;* Mac-Millan; N.Y., Londres (1962).

Note : Le texte cité ici [Sép. Aléat.] fut d'abord publié sans titre comme la note III annexée aux "Leçons sur la reconnaissance des formes" (1966-1967).