REVUE DE STATISTIQUE APPLIQUÉE

J. P. NAKACHE
L. DUSSERRE

Étude de problème posés par l'analyse linéaire discriminante en pas à pas

Revue de statistique appliquée, tome 23, nº 3 (1975), p. 59-73
<http://www.numdam.org/item?id=RSA_1975__23_3_59_0>

© Société française de statistique, 1975, tous droits réservés.

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

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

ÉTUDE DE PROBLÈME POSÉS PAR L'ANALYSE LINÉAIRE DISCRIMINANTE EN PAS A PAS [1]

J. P. NAKACHE et L. DUSSERRE (*)
Groupe de recherche U 88 (C.H.U.P.S)

INTRODUCTION

Nous nous proposons dans cet article d'étudier quelques problèmes pratiques posés par l'analyse linéaire discriminante en deux groupes selon la méthode du pas à pas :

– critère du pas à pas
– chute du pourcentage de bien classés au cours du pas à pas
– validité de la fonction linéaire discriminante
– rapport entre le nombre d'individus de l'échantillon total et le nombre de paramètres de l'étude pour obtenir une discrimination significative
– cas des échantillons de taille réduite
– définition d'un intervalle d'indécision
– détermination d'une règle précise permettant d'arrêter la procédure du pas à pas et d'affecter un nouvel individu à l'un des groupes.

Nous donnons en conclusion quelques critères à respecter pour être assuré de l'efficacité d'une analyse linéaire discriminante.

COMPARAISON DE DIFFERENTES PROCEDURES PAS A PAS

Pas à pas ascendant : méthode ADISCRI

C'est la procédure classique utilisée dans la bibliothèque du BMD (BIOMEDICAL COMPUTER PROGRAMS – BMD 6M et 7M).

Le pas à pas dans ADISCRI s'effectue de la façon suivante :

– Au premier pas on sélectionne la variable la plus discriminante parmi toutes les variables de l'étude : c'est la variable pour laquelle les moyennes des deux groupes sont les plus éloignées possibles au sens du D^2 de MAHALANOBIS (Référence 7).

– Au deuxième pas on retient, parmi les couples de variables comprenant la première variable discriminante, celui qui fournit le plus grand D^2 et ainsi

- - - - - - - - - - - - - - -

(1) Article remis en Décembre 1973, révisé en Novembre 1974.

(2) Travail effectué dans le Département de Biomathématiques (F. GREMY et D. SALMON) du C.H. U. Pitié-Salpétrière à PARIS.

de suite. . . De cette façon on ne remet pas en cause à un pas donné ce qui s'est passé aux pas précédents. Si on utilise p variables, cette méthode nécessite le calcul de $p(p + 1)/2$ distances au sens de MAHALANOBIS.

On peut reprocher à cette méthode de ne pas sélectionner à coup sûr le meilleur sous-ensemble de variables à un pas donné. C'est pour cette raison que nous avons programmé les procédures de pas à pas suivantes.

Pas à pas descendant : méthode DDISCRI

A l'inverse de la précédente cette méthode procède par élimination successive des variables les moins discriminantes en opérant ainsi :

– Au premier pas, si p est le nombre de variables à étudier, on calcule les distances au sens de MAHALANOBIS correspondant aux p sous-ensembles de $(p - 1)$ variables et on conserve, pour le pas suivant, les $(p - 1)$ variables qui entrent dans la composition du sous-ensemble fournissant le plus grand D^2 : la variable qui ne figure pas dans cette combinaison est alors éliminée.

– Au deuxième pas, on détermine parmi les $(p - 1)$ sous-ensembles de $(p - 2)$ variables, celui qui fournit le plus grand D^2 en éliminant de la même façon la variable ne faisant pas partie du meilleur sous-ensemble, et ainsi de suite . . .

Comme pour ADISCRI, la méthode DDISCRI nécessite le calcul de $p(p + 1)/2$ distances et présente le même inconvénient, à savoir que la détermination du meilleur sous-ensemble à un pas q dépend de ce qui s'est passé au pas $(q - 1)$.

Pas à pas total : méthode TDISCRI

Cette procédure permet de rechercher à partir de toutes les variables disponibles, et à un pas donné, le meilleur sous-ensemble de variables discriminantes, parmi toutes les combinaisons possibles, en utilisant toujours le critère du D^2 de MAHALANOBIS pour sélectionner les variables.

A l'inverse des précédentes, cette méthode aboutit au meilleur sous-ensemble de variables discriminantes quelque soit le pas. Elle nécessite, par contre, le calcul de $2^p - 1$ distances D^2 au lieu de $p(p + 1)/2$ (le nombre $2^p - 1$ croit très vite quand p dépasse 9).

Remarque

On peut envisager un quatrième procédé de sélection des variables discriminantes en pas à pas basé, non plus sur le critère du D^2 de MAHALANOBIS, mais sur un autre critère : le critère du maximum du pourcentage d'individus bien classés.

Le pourcentage d'individus bien classés est obtenu, à chaque pas et pour chacun des sous-ensembles considérés à ce pas, en calculant la distance des individus de l'échantillon total de l'étude à chacun des deux groupes et en affectant chaque individu au groupe dont il est le plus proche. On obtient ainsi un tableau de classement à deux lignes et deux colonnes contenant, dans la diagonale principale, les nombres d'individus bien classés. Ces nombres,

divisés par l'effectif total de l'échantillon, donnent les pourcentages d'individus bien classés dans les deux groupes.

Nous avons programmé cette procédure dans le seul but de comparer ses résultats à ceux obtenus par les méthodes précédemment définies, mais nous ne la retiendrons pas pour un usage pratique en raison des calculs très longs qu'elle nécessite, calculs qui rendent son coût prohibitif.

Comparaison des résultats des différentes méthodes de pas à pas

Ces méthodes de pas à pas ont été testées sur différents jeux de données médicales. Le tableau I fournit les résultats de l'un d'entre eux, résultats qui semblent particulièrement intéressants pour une comparaison pratique des différentes méthodes.

L'examen du tableau I montre que les résultats de DDISCRI sont très proches de ceux de TDISCRI : ils diffèrent uniquement au premier pas. Par contre, on note une nette différence entre les résultats de TDISCRI et ceux de ADISCRI dans les quatre premiers pas de la procédure. Enfin, à partir du cinquième pas, les résultats des trois méthodes sont identiques.

Le graphique I permet de comparer les différentes méthodes de pas à pas quant à leur temps calcul : il représente les courbes des temps d'exécution des programmes correspondants, en fonction du nombre croissant de variables. L'examen de ce graphique montre que, si les résultats de TDISCRI sont meilleurs que ceux de ADISCRI, du moins pour les premiers pas, ils sont plus coûteux. La différence des temps d'exécution devient considérable au-delà de neuf variables en raison de la disproportion entre $2^p - 1$ et $p(p + 1)/2$.

Dans ADISCRI et DDISCRI on a le même nombre de calculs de distances et pourtant les temps d'exécution sont différents. Cette différence est due au fait que, dans le calcul des distances, on utilise la matrice des variances-covariances commune qui, pour ADISCRI, est déterminée pas à pas en utilisant la méthode de l'escalade (Référence 7) alors que DDISCRI nécessite, au premier pas, le calcul de la matrice totale définie en utilisant toutes les variables de l'étude.

Les différents jeux de données utilisés pour comparer ces procédures ont montré qu'à partir du cinquième ou sixième pas les résultats sont identiques. Par conséquent la méthode ADISCRI, procédure de pas à pas classique, est celle que l'on peut retenir, surtout si le nombre de variables de l'étude est important et que l'on arrête le pas à pas au dixième pas environ.

CHUTE DU POURCENTAGE DE BIEN CLASSES AU COURS DU PAS A PAS

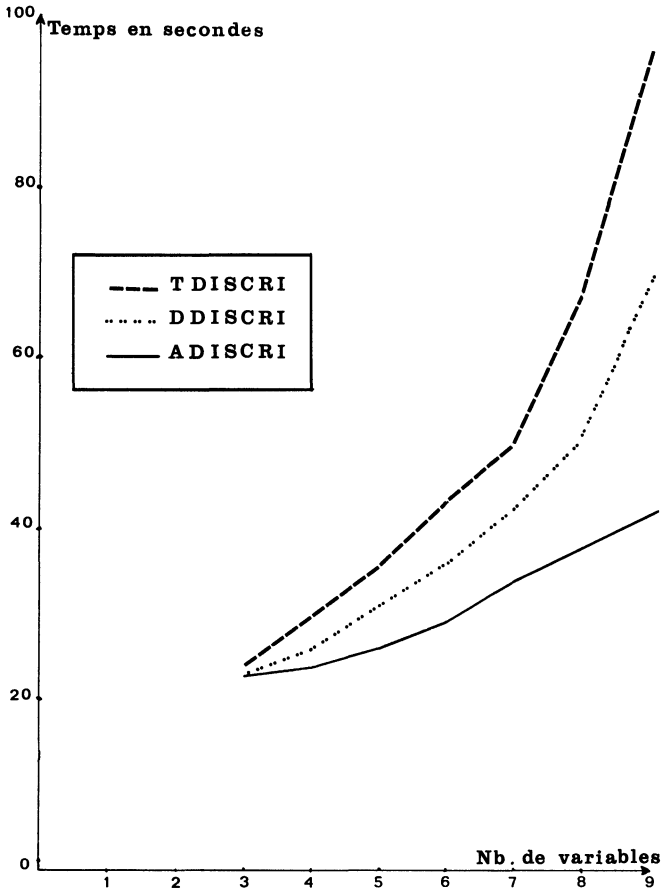
Dans les trois méthodes de pas à pas utilisées on note une chute du pour centage de bien classés au cours du pas à pas. De plus, les pourcentages de bien classés au deuxième pas (85 % dans ADISCRI et 82 % dans TDISCRI) mettent en évidence une discordance entre le D^2 de MAHALANOBIS et le pourcentage de bien classés. Cette discordance apparaît plus nettement à l'examen du tableau II où sont figurés pour l'exemple "Galactose" les résultats complets du pas à pas,

Tableau I

Comparaisons des trois méthodes de discrimination
(pas à pas)

Données Galactose — **48 Normaux** — **56 Pathologiques** — **9 Paramètres**

pas	discr. asc.	% b.c.	discr. tot.	% b.c.	discr. desc.	% b.c.
1	Quick	74	Quick	74	Bsp1	79
2	Quick,Bsp1	85	Bsp1,Accel.	82	Bsp1,Accel.	82
3	Quick,Bsp1,Accel.	81	Age,Bsp1,Accel.	89	Age,Bsp1,Accel.	89
4	Quick,Bsp1,Accel., Age	86	Age,Bsp1,Accel., Ph.Alc.	89	Age,Bsp1,Accel., Ph.Alc.	89
(5)	Quick,Bsp1,Accel., Age,Ph.Alc.	86	Quick,Bsp1,Accel., Age,Ph.Alc.	86	Quick,Bsp1,Accel. Age,Ph.Alc.	86
6 7 8 9	Résultats identiques pour les pas suivants					



Graphique I – Temps d'exécution des programmes en fonctions du nombre de variables.

Tableau II

n° pas	DISCR. ASC.			DISCR. DESC.			DISCR. TOT. (D^2 max.)			DISCR. CLAS. (% BC max.)		
	n° des variables	D^2	% BC	n° des variables	D^2	%BC	n° des variables	D^2	%BC	n° des variables	D^2	%BC
1	4	1.52	74	2	1.50	79	4	1.52	74	2	1.50	79
2	2 4	1.95	85	2 5	2.10	82	2 5	2.10	82	2 4	1.95	85
3	2 4 5	2.11	81	1 2 5	2.20	89	1 2 5	2.20	89	1 2 5	2.20	89
4	1 2 4 5	2.24	86	1 2 5 8	2.24	89	1 2 5 8	2.24	89	1 2 5 6	2.22	90
5	1 2 4 5 8	2.27	86	1 2 4 5 8	2.27	86	1 2 4 5 8	2.27	86	1 2 5 6 7	2.22	90
6	1 2 3 4 5 8	2.28	87	1 2 3 4 5 8	2.28	87	1 2 3 4 5 8	2.28	87	1 2 3 5 6 7	2.24	89
7	1 2 3 4 5 8 9	2.3	87	1 2 3 4 5 8 9	2.30	87	1 2 3 4 5 8 9	2.30	87	1 2 3 5 7 8 9	2.28	89
8	1 2 3 4 5 6 8 9	2.31	88	1 2 3 4 5 6 8 9	2.30	88	1 2 3 4 5 6 8 9	2.31	88	1 2 3 4 5 7 8 9	2.30	88
9	1 2 3 4 5 6 7 8 9	2.32	89	1 2 3 4 5 6 7 8 9	2.32	89	1 2 3 4 5 6 7 8 9	2.32	89	1 2 3 4 5 6 7 8 9	2.32	89

Tableau III

Données simulées ; Echantillon Multinormal

Gr. 1 >> 100 individus $N(m_1, \Sigma)$

Gr. 2 >> 100 individus $N(m_2, \Sigma)$

pas	numéro des variables	correl. (x, f.l.d.)	coeff. de la f.l.d.	% B.C. (ech. tot.)
1	var. 5	.76	.17	85
2	 3	.45	.21	90
3	 2	.41	.18	90
4	 8	-.01	.05	91
5	 6	-.09	-.02	92
6	 4	.57	.03	90
7	 1	-.24	.05	91
8	 9	.01	.01	91
9	 7	.04	.35	91

y compris ceux obtenus en utilisant le critère du maximum du pourcentage de bien classés. Or, théoriquement, si les hypothèses de multi-normalité, d'égalité des matrices des variances covariances et d'égalité des risques de mauvais classement sont vérifiées, on démontre (Référence 1) que le pourcentage de bien classés est une fonction croissante du D^2 de MAHALANOBIS.

Dans le but d'expliquer cette chute du pourcentage de bien classés nous avons simulé, en utilisant l'algorithme ACR 60 (Référence 8), deux échantillons multinormaux à partir de deux vecteurs moyens m_1 et m_2 et d'une même matrice des variances covariances commune Σ. Les résultats

obtenus (tableau III) montrent que cette chute persiste toujours, mais semble beaucoup moins importante que dans les exemples réels utilisés. Elle paraît être en rapport avec des erreurs de précision qui s'accumulent au cours des pas de la procédure.

VALIDITE DE LA FONCTION LINEAIRE DISCRIMINANTE (f.l.d.)

L'intérêt d'une procédure de pas à pas en discrimination est de déterminer, à partir de l'ensemble des variables de l'étude, un sous-ensemble de variables discriminantes de taille réduite qui entrent dans la composition de la f.l.d. utilisée dans la règle de décision.

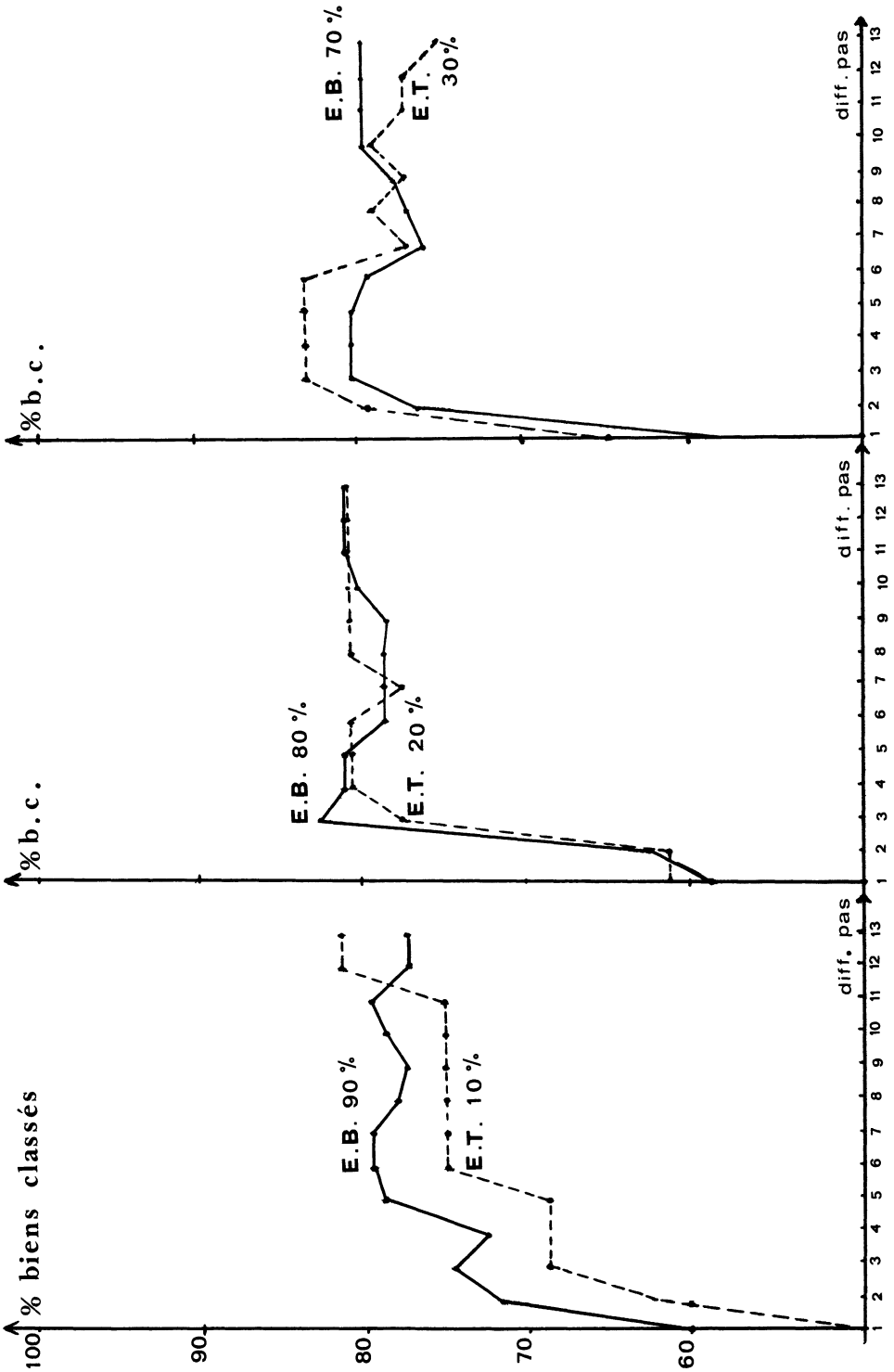
En Médecine par exemple, il est intéressant de pouvoir faire un diagnostic en effectuant le minimum d'examens chez un malade. On diminue ainsi le plus possible le coût de la décision, tant pour le malade, que pour l'organisme qui le prend en charge.

Pour déterminer le sous-ensemble de variables discriminantes il est possible d'arrêter la procédure de pas à pas quand l'introduction de nouvelles variables n'apporte plus d'information supplémentaire pour discriminer entre les deux groupes. Pour cela il faut examiner les pourcentages de bien classés et arrêter le pas à pas quand le pourcentage de bien classés atteint son maximum. Mais ces pourcentages de bien classés déterminés sur l'ensemble des individus de l'échantillon total sont biaisés d'une manière qui n'est pas négligeable : ils viennent du tableau de classement établi par la f.l.d. construite à partir de ces mêmes individus.

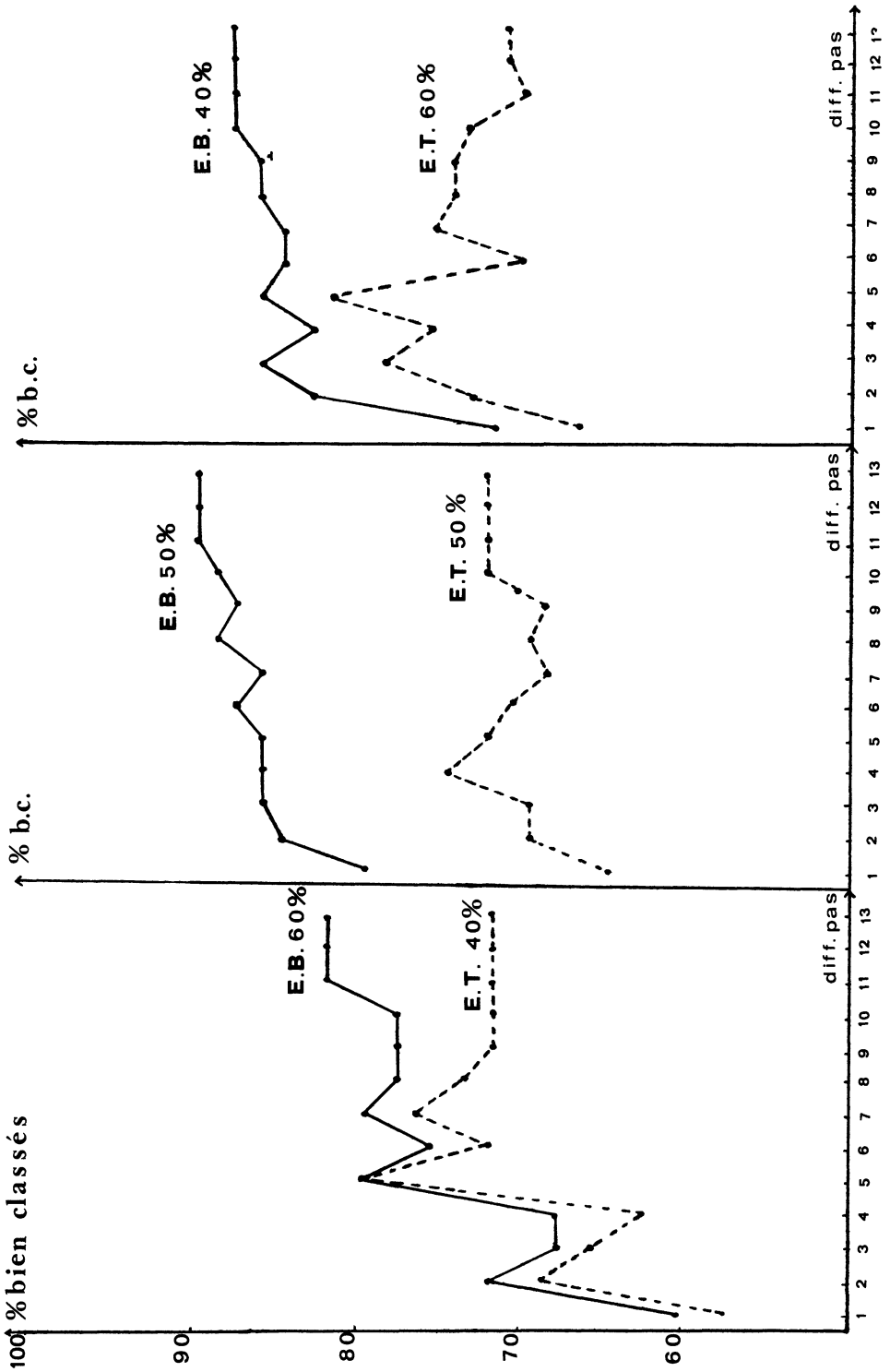
C'est pourquoi on utilise pour valider la f.l.d. la méthode de l'échantillon test (Référence 7) : sur l'ensemble des individus de groupe connu dont on dispose, on prélève au hasard un sous-ensemble d'individus des deux groupes qui constitue l'*échantillon de base,* c'est-à-dire l'échantillon qui sert à calculer la f.l.d., le sous-ensemble restant constitue alors l'*echantillon test,* c'est-à-dire l'échantillon qui sert à tester cette f.l.d. On obtient ainsi à chaque pas un pourcentage de bien classés de l'échantillon test, de valeur plus sûre.

Pour déterminer la taille significative de l'échantillon test, différents essais ont été effectués avec des tirages au hasard de taille croissante, allant de 10 à 90 % de l'échantillon total. Et, pour plusieurs jeux de données, nous avons étudié le retentissement apporté sur le pourcentage de bien classés dans l'échantillon test et dans l'échantillon de base : les courbes représentatives de ces pourcentages en fonction des pas successifs montrent qu'il faut prélever, à partir de l'échantillon de départ, 30 % d'individus pour réaliser un échantillon test de taille valable. Les 70 % qui restent servent d'échantillon de base. Les graphiques II et III montrent en effet que, dans ces conditions, les pourcentages de bien classés deviennent sensiblement les mêmes dans l'échantillon de base et dans l'échantillon test.

Graphique II — Données ictères



Graphique III — Données ictères (suite)



RAPPORT ENTRE LE NOMBRE D'INDIVIDUS DE L'ECHANTILLON TOTAL ET LE NOMBRE DE PARAMETRES DE L'ETUDE POUR OBTENIR UNE DISCRIMINATION SIGNIFICATIVE

L'extraction d'un échantillon test suppose bien entendu un effectif global suffisant. De plus, la taille de l'échantillon de base est, comme le montre T.M. COVER (Référence 3), liée, sous certaines hypothèses, au nombre de paramètres de l'étude, par une relation linéaire. En utilisant les travaux de T.M. COVER repris par J.P. BENZECRI (Référence 2) nous avons dressé le tableau IV qui donne la taille minimale de l'échantillon de base, en fonction du nombre de paramètres, taille minimale nécessaire à l'obtention d'une discrimination significative. Comme les hypothèses qui sont à la base de ces calculs sont très restrictives et rarement vérifiées en pratique, il convient d'utiliser un échantillon de base de taille beaucoup plus importante qu'il n'est indiqué dans ce tableau IV pour s'assurer des résultats valables.

Tableau IV

Relations entre N et P

$N = aP + b$

α (%)	a	b	r
100	1	2	1
90	1,70	0,56	1
80	1,80	1,07	1
70	1,87	1,52	1
60	1,93	2,06	1
50	2	3	1
40	2,06	3,03	0,999
30	2,11	3,79	0,999
20	2,19	4,57	0,999
5	2,39	7,20	0,999
1 %	2,51	10,45	0,999
0,1	2,83	13,38	0,999

Exemple : $\alpha = 1$ % si $P = 15$ --> $N \geq 48$

CAS DES ECHANTILLONS DE TAILLE REDUITE

Dans le cas d'échantillons de taille réduite, de l'ordre de vingt individus dans chacun des groupes, la f.l.d. doit être sévèrement contrôlée en raison de la non représentativité de tels échantillons et du biais entraîné. Or, prélever dans l'échantillon de départ, déjà réduit, un échantillon test suffisant appauvrirait encore l'échantillon de base et il faut renoncer à la méthode classique de l'échantillon test.

P.A. LACHENBRUCH (Référence 6) a préconisé une méthode d'échantillon test utilisable dans le cas d'effectif réduit. Cette méthode conduit à une estimation sans biais du pourcentage de bien classés et fournit de plus un in-

de bon classement). Pour obtenir une telle estimation on opère ainsi : on calcule la f.l.d. pour chacun de tous les sous-ensembles possibles de taille $n_1 + n_2 - 1$ (n_1 et n_2 sont les tailles respectives des deux groupes) obtenus en omettant un seul individu de l'échantillon total. Cette f.l.d. permet alors de classer l'individu mis de côté.

Si m_1 et m_2 sont les nombres respectifs d'individus mal classés dans les deux groupes, m_1/n_1 et m_2/n_2 sont des estimations sans biais des probabilités P_1 et P_3 de mauvais classement dans les deux groupes.

Cette méthode qui a été programmée (Référence 4) et testée sur un exemple médical de 30 individus (15 malades dans chaque groupe) a fourni un pourcentage de bien classés de 70 % contre 80 % obtenus à partir de l'échantillon total. Cet exemple confirme bien qu'il faut être prudent pour apprécier la f.l.d. calculée sur un nombre restreint d'individus. Il montre surtout l'intérêt indiscutable de la méthode de P.A. LACHENBRUCH pour contrôler l'efficacité de la f.l.d. dans le cas de petits échantillons.

DEFINITION D'UN INTERVALLE D'INDECISION

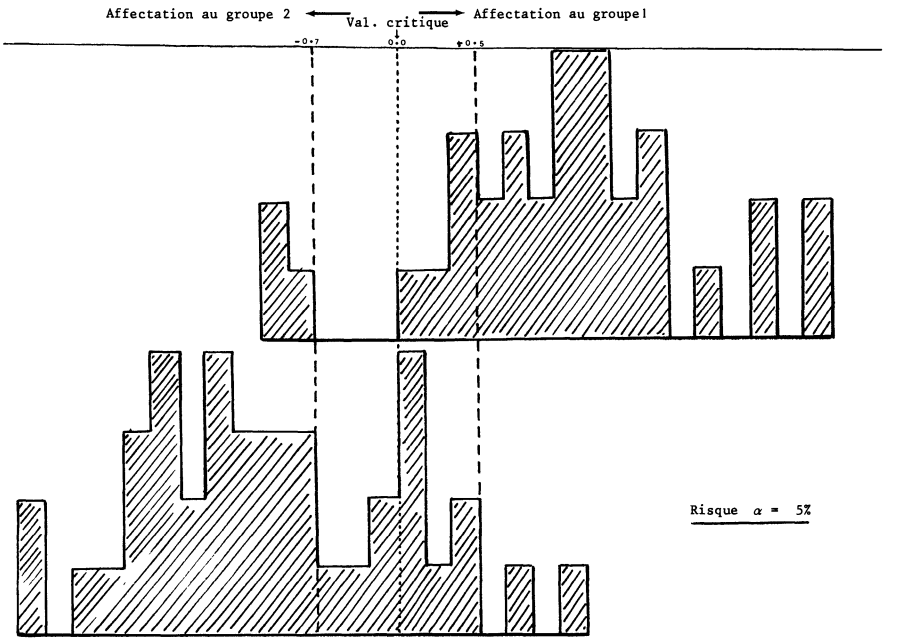
Nous avons appliqué la méthode de P.A. LACHENBRUCH au cas d'échantillons de tailles quelconques pour définir à chaque pas un intervalle de confiance des pourcentages d'individus mal classés dans les deux groupes. En déplaçant la borne critique zéro de la f.l.d. compte tenu des intervalles de confiance des nombres théoriques d'individus mal classés dans les deux groupes, on peut définir un intervalle d'indécision qui permet de modifier et de préciser la règle de décision.

Le graphique IV fournit un exemple de construction d'un intervalle d'indécision : il représente l'histogramme double des valeurs discriminantes individuelles des deux groupes obtenu au neuvième pas en utilisant le jeu de données "Galactose". L'intervalle de confiance au risque 5 % du nombre théorique d'individus mals classés dans le premier groupe étant [1,8] on déplace la borne zéro de la f.l.d. vers les valeurs positives jusqu'à obtenir 8 individus mal classés. On détermine ainsi la borne supérieure 0,5 de l'intervalle d'indécision. De la même manière, en utilisant l'intervalle [2,13] du nombre théorique de mal classés dans le deuxième groupe, on détermine la borne inférieure 0,7 de l'intervalle d'indécision au risque 5 %.

f(a) étant la valeur de la f.l.d. pour l'individu a, la règle d'affectation d'un nouvel individu a devient :

$f(a) > 0{,}5$	affectation de au premier groupe
$f(a) < -0{,}5$	affectation de a au deuxième groupe
$-0{,}7 \leqslant f(a) \leqslant 0{,}5$	on ne prend pas de décision.

On évite ainsi des erreurs de classement d'individus dont les valeurs discriminantes sont très peu positives ou très peu négatives.



Graphique IV – Histogramme double

REGLE D'ARRET DE LA PROCEDURE DU PAS A PAS

L'utilisation de différents jeux de données a montré qu'en pratique les pourcentages d'individus mal classés dans l'échantillon de base et dans l'échantillon test ne s'améliorent pas toujours au fur et à mesure du pas à pas : à partir d'un certain nombre de variables, ces pourcentages restent stables ou subissent même souvent des fluctuations autour d'un maximum local. Les variables introduites par la suite ont un pouvoir de discrimination non significatif et il est alors inutile de les faire entrer dans la f.l.d. qui doit servir de règle d'affectation pour les nouveaux individus : cette règle est, en effet, d'autant moins coûteuse que le nombre de variables de la f.l.d. est plus réduit.

La règle d'arrêt du pas à pas peut être ainsi définie : s'arrêter au pas pour lequel les pourcentages d'individus bien classés dans l'échantillon de base et dans l'échantillon test deviennent sensiblement égaux. Pour préciser cette règle de décision il est intéressant de calculer les coefficients de corrélation entre chacun des variables de l'étude et la f.l.d.

Calcul des coefficients de corrélation (variables – f.l.d.)

$$\text{corr}\,(x_i\,,\,f) = \frac{\text{cov}\,(x_i\,,\,f)}{\sigma_{x_i}\,.\,\sigma_f}$$

où x_i est la i-ième variable et f la f.l.d.

Variance de f :

$$V(f) = E[(f - \bar{f})(f - \bar{f})']$$

Or

$$f = a'x$$

où a est le vecteur des coefficients de la f.l.d. et x le vecteur des variables. Ainsi :

$$V(f) = a' \, E[(x - m_x)(x - m_x)'] \, a$$

$$V(f) = a' \, W \, a$$

W est la matrice des variances covariances commune.

a' s'écrit :

$$a' = d' \, W^{-1}$$

où d est le vecteur des différences des moyennes des variables des deux groupes. D'où :

$$V(f) = d' W^{-1} \, W W^{-1} d$$

$$= d' W^{-1} d$$

$$= D^2 \text{ (MAHALANOBIS)}$$

et

$$\sigma_f = \sqrt{D^2}$$

Covariance de (x_i, f) :

$$\text{cov}(x_i, f) = E[(x_i - m_x)(f - \bar{f})']$$

$$= E[(x_i - m_{x_i})(x - m_x)'] \, a$$

$$= E[(x_i - m_{x_i})(x - m_x)'] W^{-1} d$$

Or

$$\underset{(p,1)}{W W^{-1} d} = \underset{(p,1)}{d}$$

D'où :

$$\text{cov}(x_i, f) = \text{i-ième ligne de } W W^{-1} \, d$$

$$\text{cov}(x_i, f) = \text{i-ième ligne de d, c'est-à-dire } d_i \text{ ou encore}$$

$$\text{cov}(x_i, f) = \bar{x}_{i_1} - \bar{x}_{i_2}$$

D'autre part σ_{x_i} est le terme (i , i) de W, c'est-à-dire s_{ii}, par conséquent :

$$\text{corr}(x_i, f) = \frac{(x_{i_1} - x_{i_2})}{\sqrt{s_{ii} \, D^2}}$$

L'interprétation de ces corrélations jointe à l'examen des courbes des pourcentages de bien classés dans l'échantillon de base et dans l'échantillon test permettent ainsi de définir le choix du sous-ensemble de variables discriminantes. D'ailleurs nous pensons que c'est sur ce sous-ensemble de variables discriminantes

qu'il faut reprendre l'analyse afin de déterminer les paramètres de la f.l.d. ainsi débarassée des parasites dûs aux variables dont le pouvoir discriminant est non significatif.

Le tableau V représente le résumé de l'analyse linéaire discriminante en pas à pas (pas à pas ascendant ADISCRI) fourni par le jeu de données test "Galactose". L'examen de ce tableau montre que seuls les coefficients de corrélation entre les cinq premières variables discriminantes et la f.l.d. sont significativement différentes de zéro et que les pourcentages de bien classés maximaux sont obtenus au quatrième pas. D'autre part la corrélation entre la quatrième variable (Accélérine) et la cinquième variable (Albumine) est forte : on peut donc arrêter la procédure au quatrième pas.

En ordonnant les quatre premières variables discriminantes suivant les corrélations décroissantes on obtient le sous-ensemble hiérarchique de variables discriminantes.

Tableau V
Discrimination ascendante

Données Galactose — **48 Normaux** — **56 Pathologiques** — **Parametres 9**

pas	variables	correl. (x, fld)	coeff. de la fld	% B.C. E.B.	% B.C. E.T.
1	Quick	.69	.008	75	72
2	Bsp1	.66	.154	82	88
3	Age	-.31	-.025	79	94
4	Accelerine	.66	.021	86	94
5	Albumine	.59	.041	85	91
6	Ph.Alcalines	-.03	.012	85	91
7	Transaminases	.02	-.002	85	91
8	Bilirubine	-.14	.001	85	91
9	Cholesterol	.06	-.032	86	91

cste = -.389

Remarque

Le programme fournissant tous les résultats exposés ici a été écrit, en FORTRAN IV, pour le PDP 10 (Digital Equipement) du Centre de Calcul et de Statistique du Centre Hospitalier et Universitaire Pitié-Salpétrière. L'analyse et le dossier de programmation font l'objet d'un fascicule (Référence 5).

CONCLUSION

Nous donnons pour conclure quelques critères que nous pensons utiles et efficaces en analyse linéaire discriminante :

– Elimination dans une première étape des variables redondantes (variables très corrélées) et des variables n'apportant aucune information pour distinguer les groupes

– Respect du rapport entre le nombre d'individus N de l'échantillon total et le nombre de paramètres de l'étude :

$N \geqslant 2p$ (d'après les travaux de T.M. COVER)

– Nécessité d'un échantillon test tiré au hasard et représentant 30 % de l'échantillon total, pour valider la f.l.d. calculée à partir de l'échantillon de base

– Utilisation de la méthode de P.A. LACHENBRUCH dans le cas d'échantillons de taille réduite pour obtenir un pourcentage d'individus bien classés qui soit sans biais

– Utilisation des procédures ADISCRI (pas à pas ascendant) ou DDISCRI (pas à pas descendant) pour l'obtention du sous-ensemble de variables discriminantes de taille réduite dans le cas où le nombre de paramètres de l'étude est grand

– Utilisation de la procédure TDISCRI (pas à pas total) dans le cas d'un petit nombre de paramètres pour obtenir, si nécessaire, la meilleure combinaison de variables dans les trous premiers pas

– Arrêt de la procédure du pas à pas en tenant compte :

– des corrélations entre chaque variable et la f.l.d.

– des courbes de pourcentages de bien classés dans l'échantillon de base et dans l'échantillon test

– Détermination d'un intervalle d'indécision à un risque α fixé pour préciser la règle d'affectation d'un nouvel individu à l'un des groupes.

C'est en tenant compte de ces considérations que l'on peut tirer un maximum de profit de cette méthode de classement et d'aide à la décision qu'est l'analyse linéaire discriminante.

REFERENCES BIBLIOGRAPHIQUES

(Réf. 1) ANDERSON T.W. (WILEY 1958) – Introduction to Multivariate Statistical Analysis.

(Réf. 2) BENZECRI J.P. (1969) – Leçons sur la Reconnaissance des Formes, 3ème édition complètée des notes III' et IV.
Laboratoire du Professeur BENZECRI, Faculté des Sciences, Paris.

(Réf. 3) COVER T.M. (1965) – Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications In Pattern Recognition. IEEE Transactions on Electronic Computers.

(Réf. 4) DUSSERRE L. (1974) – Discrimination Linéaire. Etude de quelques problèmes pratiques. Thèse de Biologie Humaine, Paris. (A paraître).

(Réf. 5) DUSSERRE L., NAKACHE J.P. – Programme DUSDIS. Discrimination linéaire en deux groupes – Méthode du pas à pas – Analyse et programmation.
Fasc. N° Collection MIS – Editions SIMEP

(Réf. 6) LACHENBRUCH P.A. (1967) – Estimation of Probabilities of Misclassification in Discriminant Analysis.
Biometrics, Déc. 1967, vol. 23, N° 4, p. 639.

(Réf. 7) ROMEDER J.M. (1973) – Méthodes et programmes d'analyse discriminante. Ed. Dunod.

(Réf. 8) HURST R.L., KNOPF R.E. (1971) – Algorithm 425 – Collected Algorithms from CACM