

# STATISTIQUE ET ANALYSE DES DONNÉES

C. HEUCHENNE

## **La procédure linéaire de classement la plus sûre sans hypothèse distributionnelle ?**

*Statistique et analyse des données*, tome 6, n° 2 (1981), p. 22-48

[http://www.numdam.org/item?id=SAD\\_1981\\_\\_6\\_2\\_22\\_0](http://www.numdam.org/item?id=SAD_1981__6_2_22_0)

© Association pour la statistique et ses utilisations, 1981, tous droits réservés.

L'accès aux archives de la revue « Statistique et analyse des données » 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/>

LA PROCEDURE LINEAIRE DE CLASSEMENT LA  
PLUS SURE SANS HYPOTHESE DISTRIBUTIONNELLE ?

C. HEUCHENNE

Institut de Psychologie de l'Université de Liège  
au Sart-Tilman - B 4000

Résumé : Quand  $\vec{g}$  est un vecteur de prédicteurs, pour n'importe quelles fonctions  $f_j$ , soit la règle : attribuer  $\vec{g}$  à la classe  $i$  qui maximise  $f_i(\vec{g})$  parmi  $f_1(\vec{g}), \dots, f_c(\vec{g})$ . Des bornes inférieure et supérieure de mauvais classement sont établies sous différentes conditions. Ces résultats conduisent à la règle de Bayes quand les distributions sont connues et, sinon, suggèrent que  $f_j$  soit la régression linéaire de l'indicatrice 0-1 de la classe  $j$  sur  $\vec{g}$ . Cette simple procédure sans hypothèse distributionnelle minimise la borne supérieure de la probabilité de mauvais classement parmi les règles linéaires et est la meilleure approximation linéaire de la règle de Bayes. Elle se réduit à un algorithme de régression et permet d'agréger rationnellement des classes en cours de route.

Summary : When  $\vec{g}$  is a vector of predictors, for any functions  $f_j$ , let be the rule : attribute  $\vec{g}$  to that class  $i$  which maximizes  $f_i(\vec{g})$  among  $f_1(\vec{g}), \dots, f_c(\vec{g})$ . G.l.b. and l.u.b. of the probability of misclassification are stated under miscellaneous conditions. These results lead to Bayes rule when distributions are known and, if not, suggest that  $f_j$  would be the linear regression of the 0-1 indicator of class  $j$  onto  $\vec{g}$ . This simple distribution-free procedure minimizes the l.u.b. of the probability of misclassification among linear rules and is the best linear approximation of Bayes rule. It reduces to a regression algorithm and allows rationally aggregating some classes on the way.

Mots clés : *Analyse discriminante, classement, regression linéaire.*

Toute procédure de classement en  $c$  catégories exclusives sur base de  $w$  variables prédictives se ramène au schéma suivant :

- à chaque classe  $E_j$  est attachée une fonction  $f_j$  de  $R^w$  dans  $R$ ;
- à un individu portant  $\vec{g} = (g_1, g_2, \dots, g_w)$ , tiré au hasard de  $E = \bigcup_{j=1}^c E_j$ , est pronostiquée la catégorie  $E_i$  si  $f_i(\vec{g})$  est la plus grande des  $c$  valeurs  $f_j(\vec{g})$ .

S'il y a plusieurs  $f_j(\vec{g})$  maxima, un tirage au sort ou une convention a priori doit décider de l'attribution à un  $E_i$ .

Considérons les indicatrices des catégories :

$h_j = 1$  si l'individu appartient à  $E_j$ ,  
 $= 0$  sinon,

et leur moyenne conditionnée par  $\vec{g}$

$$\hat{h}_j(\vec{g}) = \mathbb{E}(h_j/\vec{g}) = \text{pr}(E_j/\vec{g}).$$

Les  $\hat{h}_j$  sont des fonctions de  $R^w$  dans  $[0, 1]$ . Pour simplifier l'écriture, on notera aussi  $f_j$  la variable statistique qui associe la valeur  $f_j(\vec{g})$  à un individu qui porte  $\vec{g}$ .

Relativement à un ensemble précisé, le terme fourchette désignera l'intervalle de variation d'une probabilité entre sa borne inférieure et sa borne supérieure.

Dans l'ensemble des fonctions  $f_j$  appliquées à une distribution multivariée dans  $E$ , spécifiée par ses  $\hat{h}_j$ , la fourchette de la probabilité de classement correct est

$$[\mathbb{E}(\inf_j \hat{h}_j), \mathbb{E}(\sup_j \hat{h}_j)].$$

Conditionnellement au vecteur  $\vec{g}$ , la probabilité de classement correct est  $\text{pr}(E_i/\vec{g})$  si  $f_i(\vec{g})$  est le maximum des  $c$  valeurs  $f_j(\vec{g})$ .

De

$$\inf_j \hat{h}_j(\vec{g}) \leq \hat{h}_i(\vec{g}) \leq \sup_j \hat{h}_j(\vec{g}),$$

on tire

$$\mathbb{E}(\inf_j \hat{h}_j) \leq \mathbb{E}(\hat{h}_i) = \text{pr}(c.c) \leq \mathbb{E}(\sup_j \hat{h}_j).$$

La borne inférieure est atteinte quand les fonctions classificatrices  $f_j$  sont les  $(-\hat{h}_j)$ ; dans ce cas, en effet, l'indice  $i$  qui maximise  $-\hat{h}_j(\vec{g})$  est l'indice  $i$  qui minimise  $\hat{h}_j(\vec{g})$ . C'est la pire des procédures quand les  $\hat{h}_j$  sont données.

La borne supérieure est atteinte quand les  $f_j$  sont les  $\hat{h}_j$  elles-mêmes puisqu'alors l'indice  $i$  est celui qui maximise  $\hat{h}_j(\vec{g})$ . C'est la meilleure règle quand les  $\hat{h}_j$  sont données. Elle n'est autre que la règle de Bayes puisque

$$\hat{h}_j(\vec{g}) = \frac{\text{pr}(E_j)\text{pr}(\vec{g}/E_j)}{\text{pr}(\vec{g})} ;$$

quand  $\vec{g}$  est connu, prédire  $E_i$  qui donne le maximum des  $\text{pr}(E_j)\text{pr}(\vec{g}/E_j)$  [Rao, 7d.3].

Dans l'ensemble des distributions telles que  $\sum_j \hat{h}_j^2$  est un nombre fixé  $r$ , la fourchette de la probabilité de classement correct est

$$[\sup\{0, c^{-1}(1-\sqrt{(c-1)(cr-1)})\}, \frac{1-r}{c-1}]$$

par la procédure du  $\hat{h}_j$  minimum, et celle pour la procédure du  $\hat{h}_j$  maximum est

$$[r, c^{-1}(1+\sqrt{(c-1)(cr-1)})].$$

Notons immédiatement que  $r \in [c^{-1}, 1]$ , puisque  $\sum_{j=1}^c \hat{h}_j = 1$  et  $\hat{h}_j \in [0, 1]$  entraînent

$$c^{-1} \leq \sum_{j=1}^c \hat{h}_j^2 \leq 1.$$

On a

$$\sum_j \hat{h}_j^2 \leq (\sup_j \hat{h}_j) \sum_j \hat{h}_j = \sup_j \hat{h}_j$$

$$(c-1)(\inf_j \hat{h}_j) + \sum_j \hat{h}_j^2 \leq \sum_{j \neq i} \hat{h}_j + \hat{h}_i = \sum_j \hat{h}_j = 1$$

si  $\hat{h}_i = \sup_j \hat{h}_j$ , donc

$$\inf_j \hat{h}_j \leq \frac{1 - \sum_j \hat{h}_j^2}{c-1}$$

Par [Rao, 1f.1.1], on a pour un indice  $k$  quelconque

$$1 + \hat{h}_k^2 - 2\hat{h}_k = (1-\hat{h}_k)^2 = (\sum_{j \neq k} \hat{h}_j)^2 \leq (c-1) \sum_{j \neq k} \hat{h}_j^2$$

puis successivement

$$1 + c\hat{h}_k^2 - 2\hat{h}_k \leq (c-1) \sum_j \hat{h}_j^2$$

$$1 + c + c^2\hat{h}_k^2 - 2c\hat{h}_k \leq 1 + c(c-1) \sum_j \hat{h}_j^2$$

$$(c\hat{h}_k - 1)^2 \leq (c-1)(c \sum_j \hat{h}_j^2 - 1)$$

$$-\sqrt{(c-1)(c\sum_j \hat{h}_j^2 - 1)} \leq c\hat{h}_k - 1 \leq \sqrt{(c-1)(c\sum_j \hat{h}_j^2 - 1)}$$

$$c^{-1}(1 - \sqrt{(c-1)(c\sum_j \hat{h}_j^2 - 1)}) \leq \hat{h}_k \leq c^{-1}(1 + \sqrt{(c-1)(c\sum_j \hat{h}_j^2 - 1)})$$

Comme  $\sqrt{(c-1)(cx-1)}$  est une fonction concave de  $x$ , à l'aide de l'inégalité de Jensen [Rao, le.5.6], on obtient

$$\frac{1 - \sqrt{(c-1)(c\sum_j \hat{h}_j^2 - 1)}}{c} \leq \&(\inf_j \hat{h}_j) \leq \frac{1 - \&(\sum_j \hat{h}_j^2)}{c-1}$$

$$\&(\sum_j \hat{h}_j^2) \leq \&(\sup_j \hat{h}_j) \leq \frac{1 + \sqrt{(c-1)(c\sum_j \hat{h}_j^2 - 1)}}{c}$$

Si  $r = c^{-1}$ , presque sûrement  $\hat{h}_j = c^{-1}$  pour tout  $j$ , donc  $\text{pr}(c.c)$  n'est pas définie. Les deux fourchettes se réduisent à  $c^{-1}$  qui est la probabilité de classement correct par tirage au sort.

Si  $r = 1$ , à un ensemble de probabilité nulle près, pour tout  $\vec{g}$ , il existe  $i$  tel que  $\hat{h}_i(\vec{g}) = 1$  et  $\hat{h}_j(\vec{g}) = 0$  si  $j \neq i$ . Dans cet autre cas extrême,  $\text{pr}(c.c) = 0$  avec la règle du  $\hat{h}_j$  minimum et  $\text{pr}(c.c) = 1$  avec la règle du  $\hat{h}_j$  maximum; les deux fourchettes se réduisent précisément à 0 et 1.

Il reste à montrer que si  $r \in ]c^{-1}, 1[$ , les minorants et majorants sont atteints ou arbitrairement approchés par certaines distributions.

Pour  $r \in ]\frac{1}{c}, \frac{1}{c-1}[$  (resp.  $\in ]\frac{1}{c}, 1[$ ), soient

$$p = c^{-1}(1 - \sqrt{(c-1)(cr-1)}) \quad (\text{resp. } c^{-1}(1 + \sqrt{(c-1)(cr-1)}))$$

qui est dans  $]0, \frac{1}{c}[$  (resp.  $] \frac{1}{c}, 1[$ )

et un seul vecteur  $\vec{g}$  en proportion  $p$  dans  $E_1$ , en proportion  $\frac{1-p}{c-1}$  dans  $E_2, \dots, E_c$ .

$$\sum_j \&(\hat{h}_j^2) = p^2 + (c-1)\left(\frac{1-p}{c-1}\right)^2 = r$$

Avec la procédure du  $\hat{h}_j$  minimum (resp. maximum),  $E_1$  est toujours pronostiqué puisque

$\hat{h}_1(\vec{g}) = p < (\text{resp. } >) \frac{1}{c} < (\text{resp. } >) \frac{1-p}{c-1} = \hat{h}_j(\vec{g})$  pour  $j \neq 1$ , donc  $\text{pr}(c.c) = \text{pr}(E_1) = p$ .

Pour  $r \in ]\frac{1}{c-1}, 1[$ , donc  $c > 2$ , on prend

$$p = \frac{1 + \sqrt{(c-2)(cr-1)}}{c-1}$$

qui est dans  $[\frac{1}{c-1}, 1[$ ,

un vecteur  $\vec{g}_1$  en proportion  $\frac{p}{2}$  dans  $E_1$ , en proportion  $\frac{1-p}{2(c-2)}$  dans

$E_3, \dots, E_c$  et un autre vecteur  $\vec{g}_2$  en proportion  $\frac{p}{2}$  dans  $E_2$ , en propor-

tion  $\frac{1-p}{2(c-2)}$  dans  $E_3, \dots, E_c$ .

$\hat{h}_2(\vec{g}_1) = \hat{h}_1(\vec{g}_2) = 0 < \frac{1-p}{c-2} = \hat{h}_j(\vec{g}_1) = \hat{h}_j(\vec{g}_2) \leq p = \hat{h}_1(\vec{g}_1) = \hat{h}_2(\vec{g}_2)$   
pour  $j \neq 1, 2$ .

$$\sum_j \mathcal{E}(\hat{h}_j^2) = \frac{p^2+0^2}{2} + \frac{0^2+p^2}{2} + \frac{(c-2)}{2} \left( \left(\frac{1-p}{c-2}\right)^2 + \left(\frac{1-p}{c-2}\right)^2 \right) = r$$

La règle du  $\hat{h}_j$  minimum prédit  $E_2$  à  $\vec{g}_1$  issu de  $E-E_2$  et  $E_1$  à  $\vec{g}_2$  issu de  $E-E_1$ , donc  $pr(c.c)$  nulle est atteinte quand  $r \geq \frac{1}{c-1}$ , c'est-à-dire  $c^{-1}(1-\sqrt{(c-1)(cr-1)}) \leq 0$ .

Pour  $r \in ]\frac{1}{c}, 1[$ , on prend  $\eta$  positif et arbitrairement petit de telle sorte que

$$p = \frac{r-c^{-1}-2\eta^2}{1-c^{-1}+(c^2-3c)\eta^2-2(c-2)\eta}$$

soit dans  $]0, 1[$ , un vecteur  $\vec{g}_1$  en proportion  $p(1-(c-2)\eta)$  dans  $E_1$ , en proportion  $p\eta$  dans  $E_3, \dots, E_c$ , et un autre vecteur  $\vec{g}_2$  en proportion  $(1-p)(\frac{1}{c} + \eta)$  dans  $E_1$ , en proportion  $(1-p)(\frac{1}{c} - \eta)$  dans  $E_2$  et en proportion  $\frac{1-p}{c}$  dans  $E_3, \dots, E_c$ . Pour  $j \neq 1, 2$ ,

$$\hat{h}_2(\vec{g}_1) = 0 < \hat{h}_j(\vec{g}_1) = \eta < \hat{h}_1(\vec{g}_1) = 1 - (c-2)\eta$$

$$\hat{h}_2(\vec{g}_2) = \frac{1}{c} - \eta < \hat{h}_j(\vec{g}_2) = \frac{1}{c} < \hat{h}_1(\vec{g}_2) = \frac{1}{c} + \eta$$

$$\sum_j \mathcal{E}(\hat{h}_j^2) = p((1-(c-2)\eta)^2 + (c-2)\eta^2) + (1-p)\left(\left(\frac{1}{c} + \eta\right)^2 + \left(\frac{1}{c} - \eta\right)^2 + \frac{c-2}{c^2}\right) = r$$

Avec la décision du  $\hat{h}_j$  minimum,  
 $pr(c.c) = pr(E_2) = (1-p)(\frac{1}{c} - \eta)$

qui est arbitrairement proche de

$$\left(1 - \frac{r-c^{-1}}{1-c^{-1}}\right)c^{-1} = \frac{1-r}{c-1}$$

Avec la décision du  $\hat{h}_j$  maximum,  
 $pr(c.c) = pr(E_1) = p(1-(c-2)\eta) + (1-p)(\frac{1}{c} + \eta)$

qui est arbitrairement voisin de

$$\frac{r-c^{-1}}{1-c^{-1}} + (1 - \frac{r-c^{-1}}{1-c^{-1}})c^{-1} = r.$$

\* \* \*

Si on ne connaît pas les fonctions  $\hat{h}_j$  (ou n'en veut rien connaître par souci de simplicité), on peut cependant, avec un peu d'information, en trouver de bonnes approximations.

Soit  $\hat{h}_j$  la régression linéaire de l'indicatrice  $h_j$  de la classe  $E_j$  sur le vecteur prédictif  $\vec{g}$ .

Au sens des moindres carrés,  $\hat{h}_j$  est la meilleure approximation linéaire de  $\hat{h}_j$ .

Les propriétés caractéristiques des modèles de régression sont [Rao, 4g.1] :

$$\mathcal{E}(\hat{h}_j) = \mathcal{E}(\hat{h}_j) = \mathcal{E}(h_j),$$

$$\mathcal{E}(\hat{h}_j, f) = \mathcal{E}(h_j, f) \text{ pour toute fonction } f \text{ de } \vec{g},$$

$$\mathcal{E}(\hat{h}_j, f) = \mathcal{E}(h_j, f) \text{ pour toute fonction linéaire } f \text{ de } \vec{g}.$$

Il en résulte

$$\mathcal{E}(\hat{h}_j, \hat{h}_j) = \mathcal{E}(h_j, \hat{h}_j) = \mathcal{E}(\hat{h}_j^2)$$

puis, pour une fonction linéaire  $f$  quelconque de  $\vec{g}$ ,

$$\mathcal{E}((f - \hat{h}_j)(\hat{h}_j - \hat{h}_j)) = \mathcal{E}(f\hat{h}_j) - \mathcal{E}(f\hat{h}_j) - \mathcal{E}(\hat{h}_j^2) + \mathcal{E}(\hat{h}_j\hat{h}_j) = \mathcal{E}(f\hat{h}_j) - \mathcal{E}(f\hat{h}_j) = 0$$

enfin

$$\mathcal{E}(f - \hat{h}_j)^2 = \mathcal{E}(f - \hat{h}_j + \hat{h}_j - \hat{h}_j)^2 = \mathcal{E}(f - \hat{h}_j)^2 + \mathcal{E}(\hat{h}_j - \hat{h}_j)^2 \geq \mathcal{E}(\hat{h}_j - \hat{h}_j)^2$$

$\hat{h}_j$  est ainsi la fonction linéaire  $f$  de  $\vec{g}$  qui minimise  $\mathcal{E}(f - \hat{h}_j)^2$ .

Une bonne procédure linéaire de classement est donc : pronostiquer  $E_j$  à un individu qui porte  $\vec{g}$  si  $\hat{h}_j(\vec{g})$  est le maximum des  $\hat{h}_j(\vec{g})$  pour  $j=1, 2, \dots, c$ .

Soient  $\pi_j$  la proportion de  $E_j$ ,  $\vec{\mu}_j$  le vecteur moyen de  $\vec{g}$  dans  $E_j$ ,  $\vec{\mu} = \sum_{j=1}^c \pi_j \vec{\mu}_j$  le vecteur moyen global,  $V$  la matrice (totale) des covariances de  $\vec{g}$  dans  $E$ . Puisque

$$\mathbb{E}(h_j) = \pi_j, \text{ cov}(h_j, \vec{g}) = \mathbb{E}(h_j \vec{g}) - \mathbb{E}(h_j) \mathbb{E}(\vec{g}) = \pi_j (\vec{\mu}_j - \vec{\mu})$$

on a

$$\hat{h}_j(\vec{g}) = \pi_j + \pi_j (\vec{\mu}_j - \vec{\mu})' V^{-1} (\vec{g} - \vec{\mu})$$

d'où

$$\sum_j \hat{h}_j = 1.$$

Puisque

$$\mathbb{E}((h_j - \hat{h}_j) \hat{h}_j) = 0, \text{ il vient encore}$$

$$\mathbb{E}(\hat{h}_j^2) + \mathbb{E}(h_j - \hat{h}_j)^2 = \mathbb{E}(h_j^2)$$

puis l'analyse de variance

$$\text{var}(\hat{h}_j) + \text{var}(h_j - \hat{h}_j) = \text{var}(h_j) = \pi_j(1 - \pi_j)$$

La procédure est liée à l'analyse factorielle discriminante (vecteurs propres de  $V^{-1}B$ ) et à l'analyse de variance multivariée par

$$\text{tr}(V^{-1}B) = \sum_{j=1}^c \pi_j^{-1} \text{var}(\hat{h}_j)$$

où

$$B = \sum_{j=1}^c \pi_j (\vec{\mu}_j - \vec{\mu})(\vec{\mu}_j - \vec{\mu})'$$

est la matrice des covariances interclasse. En effet,

$$\begin{aligned} \sum_{i=1}^w (V^{-1}B)_{ii} &= \sum_{i,k} (V^{-1})_{ik} B_{ki} = \sum_{i,k} (V^{-1})_{ik} \sum_{j=1}^c \pi_j (\mu_{kj} - \mu_k)(\mu_{ij} - \mu_i) \\ &= \sum_j \pi_j \sum_{i,k} (\mu_{ij} - \mu_i)(V^{-1})_{ik} (\mu_{kj} - \mu_k) \\ &= \sum_j \pi_j (\vec{\mu}_j - \vec{\mu})' V^{-1} (\vec{\mu}_j - \vec{\mu}) \end{aligned}$$

et

$$\begin{aligned} \text{var}(\hat{h}_j) &= \mathbb{E}(\hat{h}_j - \pi_j)^2 \\ &= \pi_j^2 (\vec{\mu}_j - \vec{\mu})' V^{-1} \mathbb{E}((\vec{g} - \vec{\mu})(\vec{g} - \vec{\mu})') V^{-1} (\vec{\mu}_j - \vec{\mu}) \\ &= \pi_j^2 (\vec{\mu}_j - \vec{\mu})' V^{-1} (\vec{\mu}_j - \vec{\mu}) \end{aligned}$$

On arrive à un coefficient de détermination de la classification par les variables prédictives, variant dans  $[0,1]$ , en divisant la quantité ci-dessus par  $(c-1)$  :

$$\rho^2 = (c-1)^{-1} \sum_{j=1}^c \pi_j^{-1} \text{var}(\hat{h}_j)$$

En effet,  $\rho^2 = 0$  si et seulement si  $\text{var}(\hat{h}_j) = 0$  pour tout  $j$ , c'est-à-dire  $\hat{h}_j(\vec{g}) = \pi_j$  pour tout  $j$  et presque tout  $\vec{g}$ .



C'est principalement le cas lorsque  $\vec{\mu}_j = \vec{\mu}$  pour tout  $j$ .

En outre,  $\rho^2 \leq 1$  car

$$\text{var}(\hat{h}_j) \leq \pi_j(1-\pi_j)$$

d'où

$$\sum_{j=1}^c \pi_j^{-1} \text{var}(\hat{h}_j) \leq \sum_{j=1}^c (1-\pi_j) = c-1$$

Il s'ensuit que la condition nécessaire et suffisante pour que  $\rho^2 = 1$  est  $\text{var}(h_j - \hat{h}_j) = 0$  pour tout  $j$ , soit  $\hat{h}_j = h_j$  pour tout  $j$  et presque tout individu. C'est le cas notamment de la détermination linéaire complète des indicatrices  $h_j$  par  $\vec{g}$ . Des considérations algébri-co-géométriques montrent que ceci arrive quand, au sein de l'espace  $R^w$  des  $\vec{g}$ , tout individu de  $E_j$  se situe dans un hyperplan de dimension  $(w-c+1)$ , les  $c$  hyperplans ainsi définis étant parallèles et distincts.

Si  $c = 2$ ,  $\rho^2$  se réduit à

$$\frac{\text{var}(\hat{h}_1)}{\text{var}(h_1)} = \pi_1 \pi_2 (\vec{\mu}_1 - \vec{\mu}_2)' V^{-1} (\vec{\mu}_1 - \vec{\mu}_2)$$

Si, en outre,  $w = 1$ ,  $\rho$  est la corrélation bisériale de point.

Il est encore intéressant de noter que, si les variables prédictives  $g_i$  sont les  $w$  premières indicatrices d'une classification à  $(w+1)$  catégories  $F_i$ , croisée avec celle des  $c$  classes  $E_j$ ,

$$\begin{aligned} \hat{h}_j &= \text{pr}(E_j/F_{w+1}) + \sum_{i=1}^w (\text{pr}(E_j/F_i) - \text{pr}(E_j/F_{w+1})) g_i \\ &= \sum_{i=1}^{w+1} \text{pr}(E_j/F_i) g_i \end{aligned}$$

et que, par conséquent, l'appartenance à  $F_i$  (soit  $g_i=1$ , les autres nuls), prédit, comme il se doit, la catégorie  $E_j$  qui maximise  $\text{pr}(E_j/F_i)$ . Dans ce cas,

$$\rho^2 = \frac{1}{c-1} \left( \sum_{i=1}^{w+1} \sum_{j=1}^c \frac{\text{pr}^2(F_i \cap E_j)}{\text{pr}(F_i) \text{pr}(E_j)} - 1 \right)$$

et  $\rho$  est l'indice de contingence de Cramer quand  $c-1 \leq w$ .

\* \* \*

On peut examiner le problème de classement d'un autre point de vue en faisant appel aux résidus  $\epsilon_j = h_j - f_j$ .

Le maximum  $f_i$  des fonctions classificatrices correspond au minimum de  $\sum_{j=1}^c \epsilon_j^2$  conditionnel à  $E_i$  puisque, dans  $E_i$ ,

$$\sum_j \epsilon_j^2 = (1-f_i)^2 + \sum_{j \neq i} (0-f_j)^2 = 1-2f_i + \sum_j f_j^2$$

Dans l'ensemble des distributions et des fonctions  $f_j$  telles que  $\sum_j \epsilon_j^2 = v$ , nombre fixé, la fourchette de la probabilité de mauvais classement est  $[0, \inf\{1, 2v\}]$ .

Il y a erreur de classement quand un individu de  $E_j$  obtient, via son  $\vec{g}$ , un  $f_i$  maximum avec  $i \neq j$ . Dans ce cas,  $2(\epsilon_j^2 + \epsilon_i^2) \geq \epsilon_j^2 + \epsilon_i^2 - 2\epsilon_j \epsilon_i = (\epsilon_j - \epsilon_i)^2 = (1-f_j + f_i)^2 \geq 1$  puisque  $h_j = 1$ ,  $h_i = 0$  et  $f_i \geq f_j$ .

Soit  $l$  l'indicatrice d'erreurs :  $l = 1$  s'il y a erreur,  $l = 0$  sinon. De

$$l \leq 2 \sum_j \epsilon_j^2$$

on tire

$$\text{pr}(m.c) = \mathbb{E}(l) \leq \mathbb{E}(2 \sum_j \epsilon_j^2) = 2v.$$

Il reste à montrer que, si  $v \in [0, 0.5]$  (resp.  $> 0.5$ , resp.  $\geq 0$ ), il existe des fonctions  $f_j$  et des distributions de  $\vec{g}$  dans des  $E_j$  telles que  $\sum_j \epsilon_j^2 = v$  et telles que la probabilité de mauvais classement soit arbitrairement voisine de  $2v$  (resp. égale à 1, resp. égale à 0).

Soient  $f_j$  la projection de  $R^c$  sur son  $j^e$  axe, c'est-à-dire  $f_j(\vec{g}) = g_j$ , et  $i = j + 1$  (modulo  $c$ ).

Si  $v \in [0, 0.5]$ , pour  $\eta$  arbitrairement petit et positif, on prend

$$p = \frac{2v}{(1+2\eta)^2}$$

qui est donc  $\in [0, 1]$  et arbitrairement proche de  $2v$ .

Pour  $j = 1, 2, \dots, c$ , dans  $E_j$ , une proportion  $\frac{1-p}{c}$  d'individus portent  $g_j = 1$ , les autres  $g_k = 0$ , et une proportion  $\frac{p}{c}$  d'individus portent  $g_j = 0.5 - \eta$ ,  $g_i = 0.5 + \eta$ , les autres  $g_k = 0$ .

La probabilité d'erreurs est donc  $p$ .

$$\mathbb{E}(\epsilon_j^2) = c^{-1}((1-p)(1-1)^2 + p(1-0.5+\eta)^2 + (1-p)(0-0)^2 + p(0-0.5-\eta)^2 + (c-2)((1-p)(0-0)^2 + p(0-0)^2)) = 2 c^{-1} p (0.5+\eta)^2.$$

Ainsi

$$\sum_j \mathcal{E}(\epsilon_j^2) = 2 p(0.5+\eta)^2 = v.$$

Dans les deux autres cas de  $v$ , pour  $j = 1, 2, \dots, c$ , dans  $E_j$  de proportion  $c^{-1}$ , tous les individus portent  $g_j = 1 \mp \sqrt{0.5v}$ ,  $g_i = \pm \sqrt{0.5v}$ , les autres  $g_k = 0$ .

$$\mathcal{E}(\epsilon_j^2) = c^{-1}((1 - 1 \pm \sqrt{0.5v})^2 + (0 \mp \sqrt{0.5v})^2 + (c-2)(0-0)^2) = vc^{-1}, \text{ donc}$$

$$\sum_j \mathcal{E}(\epsilon_j^2) = v.$$

Si  $v > 0.5$ , avec  $g_j = 1 - \sqrt{0.5v}$  et  $g_i = \sqrt{0.5v}$ , la probabilité d'erreurs est 1 puisque  $g_j < 0.5 < g_i$ .

Si  $v \geq 0$ , avec  $g_j = 1 + \sqrt{0.5v}$  et  $g_i = -\sqrt{0.5v}$ , la probabilité d'erreurs est 0 puisque  $g_i \leq 0 < 1 \leq g_j$ .

Avec un tel résultat, il est naturel de prendre comme critère de choix d'une procédure la minimisation de  $\sum_j \mathcal{E}(\epsilon_j^2)$  avec l'information dont on dispose.

Si l'on connaît la distribution de  $\vec{g}$  liée à chaque  $E_j$  et la proportion  $\pi_j$  de (dans un choix au hasard d'un individu, la probabilité a priori d'appartenance à)  $E_j$ , le principe des moindres carrés procure encore  $\hat{h}_j$  comme fonction classificatrice optimum.

$$\mathcal{E}(\epsilon_j^2) = \mathcal{E}(h_j^2) - 2\mathcal{E}(h_j \hat{h}_j) + \mathcal{E}(\hat{h}_j^2) = \mathcal{E}(h_j) - \mathcal{E}(\hat{h}_j^2) = \pi_j - \mathcal{E}(\hat{h}_j^2)$$

donne

$$\sum_j \mathcal{E}(\epsilon_j^2) = 1 - \sum_j \mathcal{E}(\hat{h}_j^2)$$

ce qui permet de transposer la fourchette de pr(c.c) donnée ci-avant.

Avec la règle du  $\hat{h}_j$  maximum, la fourchette de pr(m.c) est

$$\left[ \frac{c-1}{c} \left( 1 - \sqrt{1 - \frac{cv}{c-1}} \right), v \right]$$

dans l'ensemble des distributions telles que  $\sum_j \mathcal{E}(\epsilon_j^2) = v$ .

Si on recherche le minimum de  $\sum_j \mathcal{E}(h_j - f_j)^2$  au sein des fonctions linéaires  $f_j$ , le même principe des moindres carrés fournit encore  $\hat{h}_j$  comme meilleure fonction classificatrice.

Avec une information plus étendue que celle des deux premiers moments, rien n'empêcherait d'ailleurs de considérer d'autres types

de fonctions classificatrices  $f_j$  - par exemple polynomiales - en résolvant les équations normales qui donnent le minimum de  $\mathbb{E}(\sum_j (h_j - f_j)^2)$ .

Notons que pour les fonctions  $\hat{h}_j$  ou  $\hat{h}_j$ , on a  $\sum_j \varepsilon_j = 0$  car  $\sum_j \hat{h}_j = \sum_j \hat{h}_j = \sum_j h_j = 1$  et  $\mathbb{E}(\varepsilon_j) = 0$  car  $\mathbb{E}(\hat{h}_j) = \mathbb{E}(\hat{h}_j) = \mathbb{E}(h_j) = \pi_j$ .

Le théorème général donnant la fourchette de pr(m.c) en termes de  $v$  pourrait inclure ces conditions d'échelle à titre d'hypothèse sur les  $f_j$  puisque les exemples qu'il contient satisfont à ces conditions, comme on le vérifie aisément.

\* \* \*

Pour construire des exemples de régressions linéaires, on utilisera le lemme suivant.

Si la distribution de  $\vec{g} = (g_1, g_2, \dots, g_{c-1})$  est telle que  $\mathbb{E}(g_j) = \pi_j$ ,  $\mathbb{E}(g_j \vec{g}) = \pi_j \vec{\mu}_j$  pour  $j=1, 2, \dots, (c-1)$  et que les  $(c-1)$  vecteurs  $\pi_j (\vec{\mu}_j - \vec{\mu})$  sont linéairement indépendants, en posant  $g_c = 1 - \sum_{j=1}^{c-1} g_j$ , la régression linéaire de  $h_j$  sur  $\vec{g}$  est  $g_j$ , quel que soit  $j=1, 2, \dots, c$  et  $\sum_{j=1}^c \mathbb{E}(\varepsilon_j^2) = 1 - \sum_{j=1}^c \mathbb{E}(g_j^2)$

$\text{cov}(g_j, \vec{g}) = \mathbb{E}(g_j \vec{g}) - \mathbb{E}(g_j) \mathbb{E}(\vec{g}) = \pi_j (\vec{\mu}_j - \vec{\mu})$   
est la  $j^e$  colonne de la matrice inversible  $V$ , donc  $[0 \dots 0 \ \pi_j^{-1} \ 0 \dots 0] V = (\vec{\mu}_j - \vec{\mu})'$

Pour  $j=1, 2, \dots, (c-1)$  :

$$\begin{aligned} \hat{h}_j(\vec{g}) &= \pi_j + \pi_j (\vec{\mu}_j - \vec{\mu})' V^{-1} (\vec{g} - \vec{\mu}) \\ &= \pi_j + \pi_j [0 \dots 0 \ \pi_j^{-1} \ 0 \dots 0] (\vec{g} - \vec{\mu}) \\ &= \pi_j + g_j - \pi_j = g_j \end{aligned}$$

$$\text{et } \hat{h}_c = 1 - \sum_{j=1}^{c-1} \hat{h}_j = 1 - \sum_{j=1}^{c-1} g_j = g_c.$$

Enfin

$$\mathbb{E}(\varepsilon_j^2) = \mathbb{E}(h_j^2) - 2\mathbb{E}(h_j \hat{h}_j) + \mathbb{E}(\hat{h}_j^2) = \mathbb{E}(h_j) - \mathbb{E}(\hat{h}_j^2) = \pi_j - \mathbb{E}(g_j^2)$$

Avec la règle du  $\hat{h}_j$  maximum, la fourchette de  $pr(m.c)$  est  
 $[0, \inf\{1, 2v\}]$

dans l'ensemble des distributions telles que  $\sum_j \mathcal{E}(\epsilon_j^2) = v$ .

On a encore

$$\sum_j \mathcal{E}(\epsilon_j^2) \leq \sum_j \pi_j (1 - \pi_j) \leq 1 - c^{-1}$$

car

$$\mathcal{E}(\epsilon_j^2) = \text{var}(\epsilon_j) = \pi_j (1 - \pi_j) - \text{var}(\hat{h}_j)$$

Evidemment les bornes générales 0 et  $\inf\{1, 2v\}$  restent mino-  
 rant et majorant. Si  $v = 0$ , les deux bornes sont nulles car pour  
 tout  $j$ ,  $\mathcal{E}(\epsilon_j^2) = 0$  exige presque sûrement  $h_j = \hat{h}_j$ , donc une probabi-  
 lité d'erreurs nulle. Si  $v = 1 - c^{-1}$ ,  $\hat{h}_j(\vec{g}) = c^{-1} j$  pour tout  $j$  et pres-  
 que tout  $\vec{g}$  : la probabilité de mauvais classement n'est pas définie.

Il reste à montrer que si  $v \in ]0, 0.5]$  (resp.  $\in ]0.5, 1 - c^{-1}[$ ,  
 resp.  $\in ]0, 1 - c^{-1}[$ ), il existe une distribution de  $\vec{g}$  vérifiant le  
 lemme, telle que  $\sum_j \mathcal{E}(g_j^2) = 1 - v$  et que la probabilité d'erreurs soit  
 arbitrairement proche de  $2v$  (resp. vaille 1, resp. vaille 0).

En ce qui concerne la borne supérieure, on considère

$$y = 0.25 + 0.5c^{-1} \pm 0.25c^{-1} \sqrt{(c-2)^2 - 8c(c-2)\eta(1+2\eta)}$$

On a

$$-2cy^2 - 2(c-2)\eta^2 + 4y + c - 3 = (c-2)(1 + \eta - y)$$

Quand  $\eta$  parcourt  $]0, 0.25(\sqrt{2(1-c^{-1})} - 1)]$ ,  $y$  est réel et la fonction  
 continue  $1 + \eta - y$  parcourt

$]0.5, 0.5 - 0.5c^{-1} + 0.25\sqrt{2(1-c^{-1})}]$  si  $y$  est pris avec le signe + et

$[0.5 - 0.5c^{-1} + 0.25\sqrt{2(1-c^{-1})}, 1 - c^{-1}[$  si  $y$  est pris avec le signe -.

Avec  $p \in ]0, 1]$ , on répartit les  $c^2$  vecteurs de dimension  $c$   
 qui suivent.

Dans  $E_j$  de proportion  $c^{-1}$ , une proportion  $\frac{1-p}{c}$  d'individus  
 portent  $g_j = 1$ , les autres  $g_k = 0$ , et pour chacun des  $i \neq j$ , une pro-  
 portion  $\frac{p}{c(c-1)}$  d'individus portent  $g_j = y - \eta$ ,  $g_i = y + \eta$ , les autres  
 $g_k = \frac{1-2y}{c-2}$  (si  $c > 2$ ).

Des considérations algébriques rudimentaires et fastidieuses mon-  
 trent que les  $(c-1)$  premières variables ainsi créées satisfont aux  
 conditions du lemme et que

$$\sum_j \mathcal{E}(g_j^2) = 1 - p(1 + \eta - y)$$

Il y a mauvais classement quand, dans  $E_j$ ,  $g_j = y-\eta$ ,  $g_i = y+\eta$ , les autres  $g_k = \frac{1-2y}{c-2}$ . La probabilité d'erreur est donc  $p$ .

On procède alors au choix des paramètres  $\eta$  et  $p$ .

a) Si  $v \in ]0,0.5]$ , on prend  $\eta$  positif arbitrairement petit,  $y$  avec le signe + et  $p = v(1+\eta-y)^{-1}$ .

$1+\eta-y$  est arbitrairement proche de 0.5,  $p$  est arbitrairement voisin de  $2v$  et  $\sum_j g_j^2 = 1-v$ .

b) Si  $v \in ]0.5, 0.5-0.5c^{-1}+0.25\sqrt{2(1-c^{-1})}]$ , on prend  $y$  avec le signe +  $\eta$  tel que  $1+\eta-y = v$  et  $p = 1$ .

c) Si  $v \in ]0.5-0.5c^{-1}+0.25\sqrt{2(1-c^{-1})}, 1-c^{-1}[$ , on prend  $y$  avec le signe -  $\eta$  tel que  $1+\eta-y = v$  et  $p = 1$ .

En ce qui concerne la borne inférieure, on prend  $p = \frac{c-1+cv}{2(c-1)}$

qui est donc  $\in ]0.5, 1 [$

$$a = v+(c-1-cv)\sqrt{vc^{-1}(c-1+cv)^{-1}}$$

$$b = v-\sqrt{vc^{-1}(c-1+cv)}$$

Pour  $j=1,2,\dots,c$ , dans  $E_j$ , le vecteur dont la  $j^e$  composante est  $1-a$ , les  $(c-1)$  autres  $a(c-1)^{-1}$  apparaît en proportion  $pc^{-1}$  et le vecteur dont la  $j^e$  composante est  $1-b$ , les autres  $b(c-1)^{-1}$  apparaît en proportion  $(1-p)c^{-1}$ .

En notant que

$$pa + (1-p)b = v$$

$$pa^2 + (1-p)b^2 = v(1-c^{-1}),$$

l'algèbre élémentaire indique que les  $(c-1)$  premières variables ainsi créées respectent les conditions du lemme et que

$$\sum_j g_j^2 = 1-v$$

Il n'y a aucune erreur de classement car

$$a(c-1)^{-1} < 1-a$$

$$b(c-1)^{-1} < 1-b$$

donc  $pr(m.c)$  peut atteindre 0.

\*

\*

\*

Le coût de l'erreur qui confond deux catégories peut varier suivant les classes considérées. Il est donc intéressant d'avoir aussi un renseignement sur la probabilité de ce type de mauvais classement.

Si on utilise des  $\hat{h}_j$  telles que  $\text{var}(\varepsilon_k - \varepsilon_i) = v$  pour deux catégories spécifiées  $E_k$  et  $E_i$ , la fourchette de la probabilité d'attribuer un individu de  $E_k$  à  $E_i$ , ou vice-versa est  $[0, v]$ .

Notons que  $v \in [0, 1]$  car

$$\begin{aligned} \text{var}(\varepsilon_k - \varepsilon_i) &\leq (\sqrt{\text{var}(\varepsilon_k)} + \sqrt{\text{var}(\varepsilon_i)})^2 \leq (\sqrt{\pi_k(1-\pi_k)} + \sqrt{\pi_i(1-\pi_i)})^2 \\ &\leq (\sqrt{0.25} + \sqrt{0.25})^2 = 1 \end{aligned}$$

Par conséquent, si  $v = 1$ , la probabilité ci-dessus n'est pas définie puisqu'alors  $\pi_k = \pi_i = 0.5$  et  $\hat{h}_k(\vec{g}) = \hat{h}_i(\vec{g}) = 0.5$  pour presque tout  $\vec{g}$ .

Pour un individu de  $E_k$  assigné à  $E_i$ ,

$$\varepsilon_k - \varepsilon_i = 1 - \hat{h}_k + \hat{h}_i \geq 1$$

et pour un individu de  $E_i$  assigné à  $E_k$ ,

$$\varepsilon_k - \varepsilon_i = -\hat{h}_k - 1 + \hat{h}_i \leq -1$$

L'indicatrice de ce double type d'erreur est inférieure ou égale à  $(\varepsilon_k - \varepsilon_i)^2$ . Le passage par la moyenne fournit un minorant 0 et un majorant  $v$  qui sont des bornes comme le montre l'exemple suivant pour  $v \in [0, 1[$ .

Avec  $\eta$  arbitrairement petit en valeur absolue tel que

$$1 - v > |\eta|(1 + v)$$

soient

$$a = 1 - v - \eta^2(1 + v)$$

$$b = (1 + 2\eta)(1 - v) + \eta^2(1 + v)$$

$$d = 1 - v + \eta(1 + v)$$

Pour  $j=1, 2, \dots, (c-1)$ , dans  $E_j$ , le vecteur  $\vec{g}$  de dimension  $c$  tel que  $g_j = 1$ , les autres  $g_k$  nuls, apparaît en proportion  $0.5(1-v)(c-1)^{-1}$ ; dans  $E_{c-1}$ , le vecteur tel que  $g_{c-1} = 0.5(1-\eta)$ ,  $g_c = 0.5(1+\eta)$ , les autres nuls, apparaît en proportion  $0.5v(1+\eta)^{-1}$ ; dans  $E_c$  enfin, le vecteur tel que  $g_{c-1} = 0.5(1+\eta)$ ,  $g_c = 0.5(1-\eta)$ , les autres nuls, apparaît en proportion  $0.5vab^{-1}(1+\eta)^{-1}$  et le vecteur tel que  $g_{c-1} = v\eta d^{-1}$ ,  $g_c = 1 - v\eta d^{-1}$ , les autres nuls, apparaît en proportion  $0.5d^2b^{-1}$ .

On montre que les variables respectent les conditions du lemme et que  $\text{var}(\varepsilon_c - \varepsilon_{c-1}) = v$ .

Si  $\eta > 0$ ,  $E_c$  et  $E_{c-1}$  sont confondus en proportion  $0.5vab^{-1}(1+\eta)^{-1} + 0.5v(1+\eta)^{-1} = v(1-v)b^{-1}$  quantité arbitrairement proche de  $v$ .

Si  $\eta < 0$ , il n'y a d'erreur d'aucune sorte.

Si on emploie des  $\hat{h}_j$  telles que  $\text{var}(\varepsilon_k - \varepsilon_i) = v$  pour deux classes spécifiées  $E_k$  et  $E_i$ , la fourchette de la probabilité d'attribuer un individu de  $E_k$  à  $E_i$  est  $[0, (1+v^{-1})^{-1}]$

Quels que soient la variable  $g$  et l'ensemble  $A$ , la variance de  $g$  est supérieure ou égale à sa variance entre  $A$  et  $E-A$ , qui est  $\text{pr}(A)(\&(g/A) - \&(g))^2 + (1 - \text{pr}(A))(\&(g/E-A) - \&(g))^2$

On en déduit

$$\frac{(\&(g/E-A) - \&(g))^2}{\text{var}(g) + (\&(g/E-A) - \&(g))^2} \leq \text{pr}(A) \leq \frac{\text{var}(g)}{\text{var}(g) + (\&(g/A) - \&(g))^2}$$

Appliquons la dernière inégalité à la variable  $\varepsilon_k - \varepsilon_i$  et à l'ensemble  $A$  des individus de  $E_k$  qui portent  $\hat{h}_i$  maximum.

Dans  $A$ ,

$$\varepsilon_k - \varepsilon_i = 1 - \hat{h}_k + \hat{h}_i \geq 1$$

donc

$$\&(\varepsilon_k - \varepsilon_i / A) \geq 1$$

alors que

$$\&(\varepsilon_k - \varepsilon_i) = 0$$

On obtient

$$\text{pr}(A) \leq \frac{\text{var}(\varepsilon_k - \varepsilon_i)}{\text{var}(\varepsilon_k - \varepsilon_i) + (\&(\varepsilon_k - \varepsilon_i / A) - \&(\varepsilon_k - \varepsilon_i))^2} \leq \frac{v}{1+v}$$

L'exemple précédent montre que la probabilité peut atteindre 0. Pour la borne supérieure concernant  $v \in [0, 1[$ , soient  $\eta$  positif arbitrairement petit tel que

$$1 - v > \eta(c-2)(1-v) + (c-1)\eta^2$$

$$a = 1 + \eta$$



$$b = c + (c-2)v$$

$$d = 1 + (c-2)v.$$

Pour  $j=1,2,\dots,(c-2)$  (si  $c > 2$ ), dans  $E_j$  de proportion  $a(1-v)b^{-1}$ , il y a le seul vecteur tel que  $g_j = 1-v(1-v)a^{-1}b^{-1}$ ,  $g_{c-1} = -vda^{-1}b^{-1}$ ,  $g_c = v(c-1)a^{-1}b^{-1}$ , les autres  $g_k = -v(1-v)a^{-1}b^{-1}$ ; dans  $E_{c-1}$  de proportion  $adb^{-1}$ , il y a le seul vecteur tel que  $g_k = -v(1-v)a^{-1}b^{-1}$  pour  $k=1,2,\dots,(c-2)$ ,  $g_{c-1} = 1-vda^{-1}b^{-1}$  et  $g_c = v(c-1)a^{-1}b^{-1}$ ; dans  $E_c$ , le vecteur tel que  $g_k = -v(1-v)a^{-1}b^{-1}$  pour  $k=1,2,\dots,(c-2)$ ,  $g_{c-1} = -vda^{-1}b^{-1}$  et  $g_c = 1+v(c-1)a^{-1}b^{-1}$  apparaît en proportion  $(1+va^{-2})^{-1}-a(c-1)b^{-1}$ , et le vecteur tel que  $g_k = a(1-v)b^{-1}$  pour  $k=1,2,\dots,(c-2)$ ,  $g_{c-1} = adb^{-1}$ ,  $g_c = 1-a(c-1)b^{-1}$  apparaît en proportion  $(1+a^2v^{-1})^{-1}$ .

Les  $(c-1)$  premières variables ainsi créées satisfont aux conditions du lemme et on a  $\text{var}(\varepsilon_{c-1} - \varepsilon_c) = v$ . Comme  $adb^{-1}$  est supérieur à  $1-a(c-1)b^{-1}$  et  $a(1-v)b^{-1}$ , la probabilité d'attribuer un individu de  $E_c$  à  $E_{c-1}$  est celle du dernier vecteur, soit  $(1+a^2v^{-1})^{-1}$  arbitrairement proche de  $(1+v^{-1})^{-1}$ .

Le fait que des probabilités n'atteignent pas leur borne provient visiblement de l'ambiguïté de l'attribution quand  $f_i$  n'est pas seul maximum. Ainsi, rien n'empêche de prendre  $\eta = 0$  dans tous les exemples donnés jusqu'ici;  $\sum_j \mathbb{E}(\hat{h}_j^2) = r$ ,  $\sum_j \mathbb{E}(\varepsilon_j^2) = v$  ou  $\text{var}(\varepsilon_k - \varepsilon_i) = v$  sont conservés; mais la probabilité d'erreurs ne serait plus définie car il y aurait plusieurs  $f_j$  maxima.

\* \* \*

Outre sa justification théorique comme meilleure approximation linéaire de la règle de Bayes et comme règle minimax au sein des prédictions linéaires, la procédure de  $\hat{h}_j$  maximum a un aspect pratique extrêmement attractif. Elle rentre en effet dans le schéma général des régressions linéaires et, par conséquent, ne pose aucun problème de calcul.

Supposons plus précisément qu'on utilise un algorithme qui, en modifiant la matrice des covariances résiduelles, fait passer une variable de l'état de critère à celui de prédicteur. A chaque stade de la progression, on dispose de  $\text{cov}(\varepsilon_i, \varepsilon_j)$ ,  $\varepsilon_i$  étant le

résidu du  $i^e$  critère sur l'ensemble des explicateurs du moment.

L'objectif peut être alors d'obtenir

$$v = \sum_j \mathbb{E}(h_j - \hat{h}_j)^2 = \sum_j \text{var}(\varepsilon_j)$$

aussi petit que possible.

Au départ, en l'absence de prédicteurs,  $\hat{h}_j$  est  $\pi_j$ , donc

$$v = \sum_j \text{var}(h_j) = \sum_j \pi_j(1-\pi_j).$$

A ce stade initial, il y a déjà une décision possible, moins absurde qu'il n'y paraît : elle consiste à attribuer n'importe quel individu à la classe la plus importante, en ignorant son  $\vec{g}$ .

A chaque stade, la méthode permet de diminuer  $v$  au mieux par

1) Adjonction de nouvelles variables prédictives.

Si on adjoint les variables  $g_l$  pour  $l \in L$ ,  $v$  diminue de

$$\sum_{j=1}^c \vec{\sigma}_j' S^{-1} \vec{\sigma}_j \quad \text{où } S = [\text{cov}(\varepsilon_i, \varepsilon_l)]_{i \in L, l \in L} \text{ et } \vec{\sigma}_j = [\text{cov}(\varepsilon_1, \varepsilon_j)]_{l \in L}$$

En effet,  $\text{var}(\varepsilon_j)$  devient  $\text{var}(\varepsilon_j) - \vec{\sigma}_j' S^{-1} \vec{\sigma}_j$ .

2) Fusion de classes entre elles, si cette option est admise.

Si on agrège les catégories  $E_j$  pour  $j \in J$ ,  $v$  devient

$$v + \sum_{\substack{j, k \in J \\ j \neq k}} \text{cov}(\varepsilon_j, \varepsilon_k)$$

En effet, la régression linéaire de l'indicatrice  $\sum_{j \in J} h_j$  de  $\cup_{j \in J} E_j$  sur  $\vec{g}$  est

$$(\sum_{j \in J} \pi_j) + (\sum_{j \in J} \pi_j) ((\sum_{j \in J} \pi_j)^{-1} \sum_{j \in J} \pi_j \vec{\mu}_j - \vec{\mu})' v^{-1} (\vec{g} - \vec{\mu})$$

$$= \sum_{j \in J} (\pi_j + \pi_j (\vec{\mu}_j - \vec{\mu})' v^{-1} (\vec{g} - \vec{\mu})) = \sum_{j \in J} \hat{h}_j$$

donc  $\sum_{j \in J} \varepsilon_j$  est le résidu correspondant à  $\cup_{j \in J} E_j$ .

Par le regroupement des  $E_j$  pour  $j \in J$ ,  $\sum_j \text{var}(\varepsilon_j)$  devient donc

$$\sum_{j \in J} \text{var}(\varepsilon_j) + \text{var}(\sum_{j \in J} \varepsilon_j) = \sum_j \text{var}(\varepsilon_j) + \sum_{\substack{j, k \in J \\ j \neq k}} \text{cov}(\varepsilon_j, \varepsilon_k)$$

On dispose ainsi d'un algorithme très souple qui permet de contrôler la borne supérieure de la probabilité totale d'erreurs. Dans les domaines médicaux et psychologiques en particulier, on a parfois affaire à des classifications assez contestables.

Une bonne stratégie serait de partir d'un classement fin, quitte à abandonner en route une part des différenciations primitives, au vu des maigres progrès obtenus par seule considération d'explicateurs quantitatifs.

Certes l'étude théorique a été faite au niveau strictement descriptif et les commentaires pratiques présents ne tiennent pas compte des aléas de l'échantillonnage. Mais on peut escompter qu'un échantillon de taille raisonnable fournirait de bonnes estimations des paramètres de la population.

Sans prétendre arriver aux meilleures solutions, on se contentera dans l'exemple numérique suivant de n'examiner chaque fois que

1) l'adjonction d'un seul prédicteur; on choisira parmi le jeu encore disponible la variable  $g_1$  qui maximise

$$\text{var}^{-1}(\varepsilon_1) \sum_{j=1}^c \text{cov}^2(\varepsilon_1, \varepsilon_j);$$

2) la réunion d'un couple de classes; on confondra  $E_k$  et  $E_i$  qui maximisent  $-2 \text{cov}(\varepsilon_k, \varepsilon_i)$ .

Il est intéressant de noter que, ce faisant, on neutralise la distinction de deux catégories qui, par les derniers théorèmes, se discriminaient mal puisque

$$\text{var}(\varepsilon_k - \varepsilon_i) = \text{var}(\varepsilon_k) + \text{var}(\varepsilon_i) - 2 \text{cov}(\varepsilon_k, \varepsilon_i).$$

Malgré l'insuffisance notoire de l'effectif, l'exemple qui vient, réalisé sur une calculatrice de poche TI 59 avec le programme TISOFT 20591003, révèle les possibilités de la procédure. Il concerne un jeu de 5 prédicteurs possibles, 4 catégories primitives et 20 individus. Le critère d'arrêt est que la borne supérieure de la probabilité d'erreurs soit inférieure à 0.5, soit que  $v < 0.25$ . Cet objectif peut sembler bien lâche; mais les exemples théoriques précédents montrent assez que les extrémités des fourchettes ne sont approchées que dans les cas limites de distributions pathologiques qui ne sont pas rencontrées dans la réalité.

Pour passer des régressions de  $g_k, \dots, g_w, h_1 = g_{w+1}, \dots, h_c = g_{w+c}$  sur  $g_1, \dots, g_{k-1}$  aux régressions de  $g_{k+1}, \dots, g_w, g_{w+1}, \dots, g_{w+c}$  sur

$g_1, \dots, g_{k-1}, g_k$ , les formules bien connues sont, en termes des sommes de produits d'écartes  $SP_{ij}$  (covariances résiduelles multipliées par l'effectif) et pour  $i=k+1, \dots, w+c$  :

$$SP_{ki} \text{ devient } \beta_{ik} = SP_{ki} SP_{kk}^{-1},$$

$$\alpha_i \text{ devient } \alpha_i - \beta_{ik} \alpha_k,$$

$$\beta_{ij} \text{ devient } \beta_{ij} - \beta_{ik} \beta_{kj} \text{ pour } j=1, 2, \dots, k-1,$$

$$SP_{ij} \text{ devient } SP_{ij} - \beta_{ik} SP_{kj} \text{ pour } j=i, i+1, \dots, w+c.$$

Les résultats sont présentés avec 3 décimales, bien que les calculs aient été exécutés avec 13 chiffres significatifs. L'indicatrice de la 4<sup>e</sup> classe, redondante, n'a pas été introduite. Le tableau A des données individuelles est :

Individu	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$	$g_6=h_1$	$g_7=h_2$	$g_8=h_3$
1	3	6	4	9	7	0	0	1
2	8	2	9	7	0	0	0	0
3	5	8	5	4	5	0	1	0
4	6	5	6	2	8	0	0	0
5	4	2	6	4	4	1	0	0
6	1	9	5	6	8	0	0	1
7	8	1	5	3	6	0	0	0
8	8	3	7	5	9	0	0	0
9	0	6	2	8	5	0	0	1
10	5	9	3	0	2	0	1	0
11	9	0	8	4	5	0	0	0
12	4	5	9	8	5	0	0	1
13	2	9	4	8	9	0	0	1
14	9	0	6	5	4	0	0	0
15	4	6	8	3	7	0	1	0
16	2	7	6	9	8	0	0	1
17	8	4	2	6	0	1	0	0
18	7	4	9	0	5	0	0	0
19	2	9	8	1	6	0	1	0
20	9	1	6	6	3	0	0	0

Le vecteur des  $\mu_j$  et la matrice des  $SP_{ij}$  constituent le tableau B :

j	1	2	3	4	5	6	7	8	9
$\mu_j$	5.2	4.8	5.9	4.9	5.3	0.1	0.2	0.3	0.4
SP <sub>1j</sub>	163.2	-138.2	34.4	-45.6	-66.2	1.6	-4.8	-19.2	22.4
SP <sub>2j</sub>		185.2	-39.4	-0.4	55.2	-3.6	12.8	13.2	-22.4
SP <sub>3j</sub>			91.8	-27.2	8.6	-3.8	0.4	-5.4	8.8
SP <sub>4j</sub>				151.8	15.6	0.2	-11.6	18.6	-7.2
SP <sub>5j</sub>					132.2	-6.6	-1.2	10.2	-2.4
SP <sub>6j</sub>						1.8	-0.4	-0.6	-0.8
SP <sub>7j</sub>							3.2	-1.2	-1.6
SP <sub>8j</sub>								4.2	-2.4
SP <sub>9j</sub>									4.8

La dernière colonne, relative à  $h_4 = g_9$ , se trouve aisément à partir des trois précédentes :

$$\mu_9 = \&(h_4) = 1 - \&(h_1) - \&(h_2) - \&(h_3) = 1 - \mu_6 - \mu_7 - \mu_8$$

$$SP_{i9} = -SP_{i6} - SP_{i7} - SP_{i8}$$

A ce stade initial,  $\sum_{j=6}^9 SP_{jj}$  est 14, très supérieur au seuil fixé  $0.25 \times 20 = 5$ . En prenant  $g_1$  comme premier prédicteur,

$\sum_{j=6}^9 SP_{jj}$  diminuerait de

$$SP_{11}^{-1} \sum_{j=6}^9 SP_{1j}^2 = 163.2^{-1} (1.6^2 + 4.8^2 + 19.2^2 + 22.4^2) = 5.490$$

Pour  $g_2, g_3, g_4, g_5$ , il s'agirait respectivement de 4.605, 1.320, 3.507, 1.171.

En fusionnant  $E_1$  et  $E_2$ ,  $\sum_{j=6}^9 SP_{jj}$  diminuerait de  $-2 SP_{67} = 0.8$ .

Pour  $E_1 \cup E_3, E_1 \cup E_4, E_2 \cup E_3, E_2 \cup E_4, E_3 \cup E_4$ , il s'agirait respectivement de 1.2, 1.6, 2.4, 3.2 et 4.8.

On est amené à faire entrer  $g_1$  au titre de premier prédicteur, ce qui fait descendre  $\sum_{j=6}^9 SP_{jj}$  à 8.510. Par les formules ci-dessus, le tableau B devient le tableau C :

j	2	3	4	5	6	7	8	9
$\alpha_j$	9.203	4.804	6.353	7.409	0.049	0.353	0.912	-0.314
$\beta_{j1}$	-0.847	0.211	-0.279	-0.406	0.010	-0.029	-0.118	0.137
$SP_{2j}$	68.170	-10.270	-39.015	-0.859	-2.245	8.735	-3.059	-3.431
$SP_{3j}$		84.549	-17.588	22.554	-4.137	1.412	-1.353	4.078
$SP_{4j}$			139.059	-2.897	0.647	-12.941	13.235	-0.941
$SP_{5j}$				105.347	-5.951	-3.147	2.412	6.686
$SP_{6j}$					1.784	-0.353	-0.412	-1.020
$SP_{7j}$						3.059	-1.765	-0.941
$SP_{8j}$							1.941	0.235
$SP_{9j}$								1.725

Maintenant, en prenant  $g_2, g_3, g_4, g_5$  comme prédicteur supplémentaire ou en faisant  $E_1 \cup E_2, E_1 \cup E_3, E_1 \cup E_4, E_2 \cup E_3, E_2 \cup E_4, E_3 \cup E_4,$   
 $\sum_{j=6}^9 SP_{jj}$  diminuerait respectivement de 1.503, 0.444, 2.473, 0.910  
 ou 0.706, 0.824, 2.039, 3.529, 1.882, -0.471. Le meilleur choix est de réunir  $E_2$  et  $E_3$ , ce qui atteint l'objectif :

$$\sum_j SP_{jj} \text{ devient } 4.980.$$

Les paramètres de

$$g_{(78)} = g_7 + g_8 = h_2 + h_3$$

sont obtenus par les formules générales de fusion de  $E_k$  et  $E_i$  :

$$\alpha_{(ki)} = \alpha_k + \alpha_i \quad \beta_{(ki)1} = \beta_{k1} + \beta_{i1}$$

$$SP_{(ki)1} = SP_{k1} + SP_{i1} \text{ si } 1 \neq k, i$$

$$SP_{(ki)(ki)} = SP_{kk} + SP_{ii} + 2SP_{ki}$$

La partie droite du tableau C devient :

j	6	78	9
$\alpha_j$	0.049	1.265	-0.314
$\beta_{j1}$	0.010	-0.147	0.137
$SP_{2j}$	-2.245	5.676	-3.431
$SP_{3j}$	-4.137	0.059	4.078
$SP_{4j}$	0.647	0.294	-0.941
$SP_{5j}$	-5.951	-0.735	6.686
$SP_{6j}$	1.784	-0.765	-1.020
$SP_{(78)j}$		1.470	-0.706
$SP_{9j}$			1.725

Avec les régressions

$$\hat{h}_1 = 0.049 + 0.010 g_1$$

$$\hat{h}_{(23)} = 1.265 - 0.147 g_1$$

$$\hat{h}_4 = -0.314 + 0.137 g_1,$$

les vingt individus sont bien reclassés sauf les deux de  $E_1$  qui

obtiennent :	$\hat{h}_1$	$\hat{h}_{(23)}$	$\hat{h}_4$
n°5	0.088	0.676	0.235
n°17	0.127	0.088	0.784

Essayons d'autres solutions en introduisant d'abord la seconde meilleure variable prédictive  $g_2$ . Le tableau B devient le tableau D :

j	1	3	4	5	6	7	8	9
$\alpha_j$	8.782	6.921	4.910	3.869	0.193	-0.132	-0.042	0.981
$SP_{1j}$	60.072	4.999	-45.898	-25.009	-1.086	4.752	-9.350	5.685
$\beta_{j2}$	-0.746	-0.213	-0.002	0.298	-0.019	0.069	0.071	-0.121
$SP_{3j}$		83.418	-27.285	20.343	-4.566	3.123	-2.592	4.035
$SP_{4j}$			151.799	15.719	0.192	-11.572	18.629	-7.248
$SP_{5j}$				115.747	-5.527	-5.015	6.266	4.276
$SP_{6j}$					1.730	-0.151	-0.343	-1.235
$SP_{7j}$						2.315	-2.112	-0.052
$SP_{8j}$							3.259	-0.803
$SP_{9j}$								2.091

On vérifie que  $\sum_{j=6}^9 SP_{jj} = 14 - 4.605 = 9.395$ . La meilleure décision ultérieure est de fondre  $E_2$  et  $E_3$ , ce qui transforme la partie droite du tableau D en :

j	6	78	9
$\alpha_j$	0.193	-0.174	0.981
$SP_{1j}$	-1.086	-4.598	5.685
$\beta_{j2}$	-0.019	0.140	-0.121
$SP_{3j}$	-4.566	0.531	4.035
$SP_{4j}$	0.192	7.056	-7.248
$SP_{5j}$	-5.527	1.251	4.276
$SP_{6j}$	1.730	-0.495	-1.235
$SP_{(78)j}$		1.350	-0.855
$SP_{9j}$			2.091

Les pertes apportées à  $\sum_j SP_{jj}$  respectivement par l'adjonction de  $g_1, g_3, g_4, g_5$  et les réunions  $E_1 \cup (E_2 \cup E_3), E_1 \cup E_4, (E_2 \cup E_3) \cup E_4$  sont 0.910, 0.449, 0.674, 0.435 et 0.989, 2.471, 1.711. Après l'agrégation de  $E_1$  et  $E_4$ , la partie droite du tableau D est

j	69	78
$\alpha_j$	1.174	-0.174
$SP_{1j}$	4.598	-4.598
$\beta_{j2}$	-0.140	0.140
$SP_{3j}$	-0.531	0.531
$SP_{4j}$	-7.056	7.056
$SP_{5j}$	-1.251	1.251
$SP_{(69)j}$	1.350	-1.350
$SP_{(78)j}$		1.350

Puisque  $\sum_j SP_{jj} = 2.700$ , le seuil est largement dépassé. Les régressions appliquées aux données donnent une seule erreur pour l'individu n° 4 de  $E_1 \cup E_4$  :

$$\hat{h}_{(14)} = 0.472 \quad \hat{h}_{(23)} = 0.528$$

Si, après la considération du prédicteur  $g_2$ , on refuse de confondre  $E_2$  et  $E_3$ , on sera amené à adjoindre la variable  $g_4$ . Le tableau D devient le tableau E :

j	1	3	5	6	7	8	9
$\alpha_j$	10.267	7.804	3.361	0.187	0.243	-0.645	1.215
$SP_{1j}$	46.194	-3.251	-20.256	-1.028	1.253	-3.717	3.493
$\beta_{j2}$	-0.747	-0.213	0.298	-0.019	0.069	0.072	-0.121
$SP_{3j}$		78.514	23.169	-4.531	1.043	0.757	2.732
$\beta_{j4}$	-0.302	-0.180	0.104	0.001	-0.076	0.123	-0.048
$SP_{5j}$			114.120	-5.547	-3.817	4.337	5.027
$SP_{6j}$				1.730	-0.137	-0.367	-1.226
$SP_{7j}$					1.433	-0.692	-0.604
$SP_{8j}$						0.973	0.086
$SP_{9j}$							1.745

L'objectif n'est pas encore atteint car  $\sum_{j=6}^9 SP_{jj} = 5.881$ . Il s'impose alors de réunir  $E_1$  et  $E_4$ , ce qui change la partie de droite du tableau E en :



j	69	7	8
$\alpha_j$	1.402	0.243	-0.645
$SP_{1j}$	2.465	1.253	-3.717
$\beta_{j2}$	-0.140	0.069	0.072
$SP_{3j}$	-1.800	1.043	0.757
$\beta_{j4}$	-0.046	-0.076	0.123
$SP_{5j}$	-0.520	-3.817	4.337
$SP_{(69)j}$	1.022	-0.741	-0.281
$SP_{7j}$		1.433	-0.692
$SP_{8j}$			0.973

$\sum_j SP_{jj} = 3.428$  tombe en-dessous de 5. Les régressions reclas-  
sent bien tous les individus dans  $E_1 \cup E_4$ ,  $E_2$  et  $E_3$ .

On peut enfin désirer garder distinctes les quatre classes.  
Après  $g_2$  et  $g_4$ , on adjoindra successivement  $g_5$  et  $g_1$ , ce qui trans-  
forme le tableau E en :

j	3	6	7	8	9
$\alpha_j$	6.902	0.864	0.208	-0.021	-0.051
$\beta_{j1}$	0.020	-0.047	0.014	-0.069	0.103
$\beta_{j2}$	-0.260	-0.038	0.088	0.012	-0.063
$SP_{3j}$	73.792	-3.364	1.806	-0.064	1.622
$\beta_{j4}$	-0.195	-0.007	-0.069	0.099	-0.023
$\beta_{j5}$	0.207	-0.057	-0.031	0.026	0.062
$SP_{6j}$		1.365	-0.295	-0.295	-0.775
$SP_{7j}$			1.298	-0.507	-0.495
$SP_{8j}$				0.604	0.198
$SP_{9j}$					1.072

Puisque  $\sum_j SP_{jj} = 4.339$ , le seuil est dépassé. Les régressions

$$\hat{h}_1 = 0.864 - 0.047 g_1 - 0.038 g_2 - 0.007 g_4 - 0.057 g_5$$

$$\hat{h}_2 = 0.208 + 0.014 g_1 + 0.088 g_2 - 0.069 g_4 - 0.031 g_5$$

$$\hat{h}_3 = -0.021 - 0.069 g_1 + 0.012 g_2 + 0.099 g_4 + 0.026 g_5$$

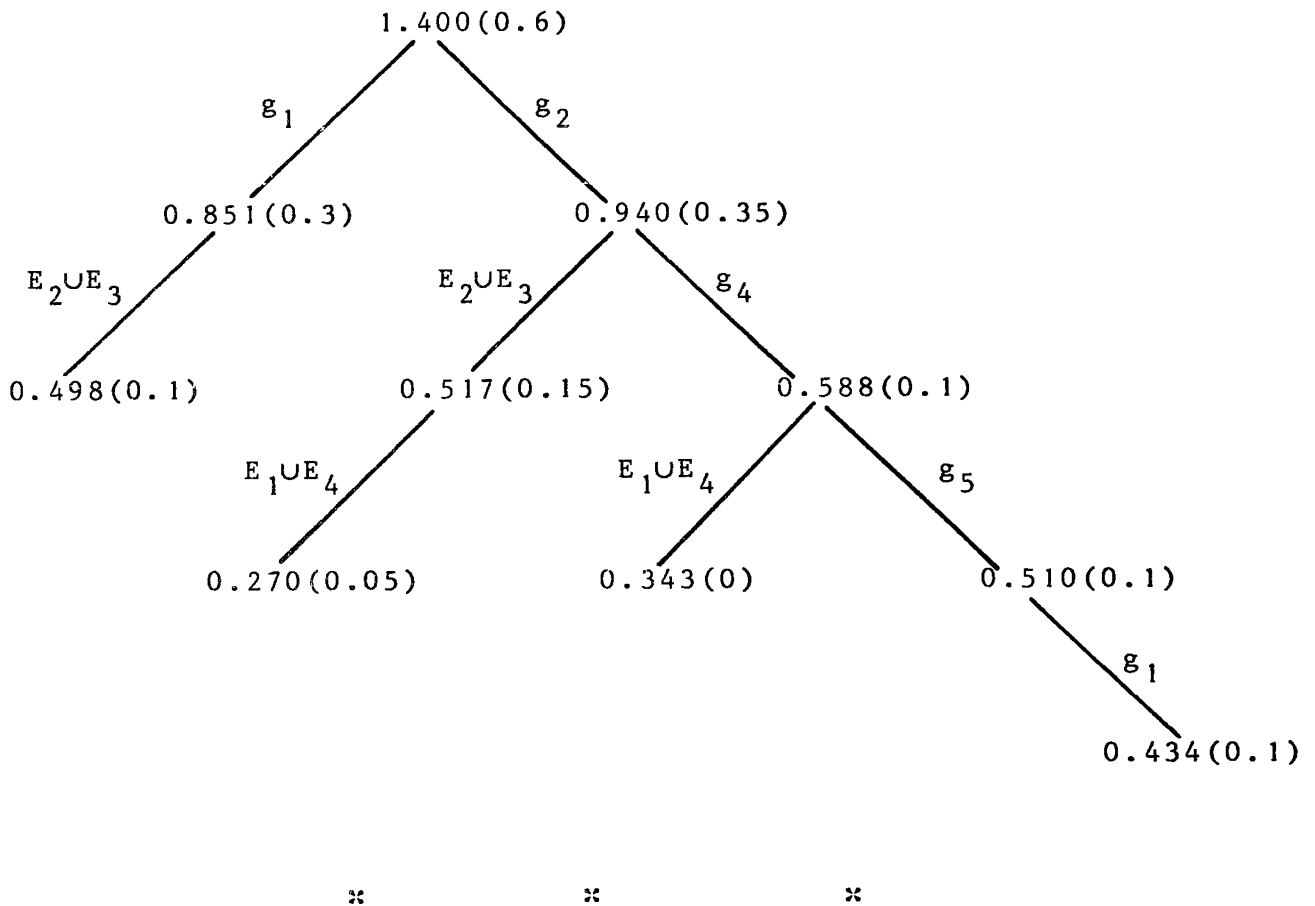
$$\hat{h}_4 = -0.051 + 0.103 g_1 - 0.063 g_2 - 0.023 g_4 + 0.062 g_5$$

font commettre deux erreurs sur les vingt données. Elles attribuent  
les deux individus de  $E_1$  à  $E_4$  :

	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_3$	$\hat{h}_4$
n°5	0.343	0.039	0.226	0.392
n°17	0.292	0.256	0.069	0.383

Cette mauvaise discrimination entre  $E_1$  et  $E_4$  est manifestée par la forte valeur persistante de  $-SP_{69}$ .

Les différentes solutions envisagées sont présentées dans l'arbre suivant où la cote d'un sommet est  $2 \sum_j \text{var}(\epsilon_j) = 0.1 \sum_j SP_{jj}$ . et le nombre entre parenthèses le taux d'erreur obtenu en reclassant les vingt individus selon les régressions correspondantes.



Les résultats de [Heuchenne] concernant une dichotomie et sont résumés comme suit.

La probabilité d'attribuer un individu de  $E_1$  à  $E_2$  (resp. de  $E_2$  à  $E_1$ ) est notée  $p_1$  (resp.  $p_2$ ); la probabilité totale d'erreurs est  $p = p_1 + p_2$ .

Dans l'ensemble des distributions pour lesquelles  $\pi_1$  et  $\rho^2$  (ou  $\text{var}(\varepsilon_1) = \text{var}(\varepsilon_2) = \pi_1\pi_2(1-\rho^2)$ ) sont donnés,  $q$  désigne la valeur attendue de  $p$ . Les fourchettes de  $p_1$ ,  $p_2$  et  $p$  concernent ce même ensemble.

Soit

$$a = \frac{\pi_1\pi_2 \rho^2(1-\rho^2)}{0.25 - \pi_1\pi_2(1-\rho^2)^2}$$

1) si  $\pi_1^{-1} \leq 2(1-\rho^2) \leq \pi_2^{-1}$ ,  $q = \pi_2$  et les fourchettes de  $p_1$ ,  $p_2$ ,  $p$  sont respectivement  $[0, a]$ ,  $[\pi_2 - a, \pi_2]$ ,  $[\pi_2 - a, \pi_2 + a]$ .

2) Si  $2(1-\rho^2) \leq \inf\{\pi_1^{-1}, \pi_2^{-1}\}$ ,  $q = 2\pi_1\pi_2(1-\rho^2)$ , la fourchette de  $p_1$  et  $p_2$  est  $[0, a]$ , la fourchette de  $p$  est  $[0, 4\pi_1\pi_2(1-\rho^2)]$ .

3) Si  $\pi_2^{-1} \leq 2(1-\rho^2) \leq \pi_1^{-1}$ ,  $q = \pi_1$  et les fourchettes de  $p_1$ ,  $p_2$ ,  $p$  sont respectivement  $[\pi_1 - a, \pi_1]$ ,  $[0, a]$ ,  $[\pi_1 - a, \pi_1 + a]$ .

En tenant compte de ce que  $a$ , toujours inférieur à  $\pi_1$  et  $\pi_2$ , atteint un maximum

$$\frac{1-\rho^2}{2-\rho^2} = \frac{4 \text{ var}(\varepsilon_1)}{1+4 \text{ var}(\varepsilon_1)}$$

pour  $\pi_1 = \pi_2 = 0.5$ , on peut voir que ces énoncés sont (évidemment) compatibles avec les précédents. Ils sont cependant plus précis parce qu'ils font intervenir explicitement les probabilités a priori  $\pi_1$  et  $\pi_2$ . Pour  $c > 2$ , cela ouvre les questions :

1) Comment améliorer les fourchettes si l'on tient compte des  $\pi_j$  ?

2) Avec ou sans cette information, pourrait-on définir une probabilité attendue d'erreurs ? S'il se révélait que cette dernière est fonction croissante de  $\sum_j \varepsilon_j^2$ , on pourrait parler de la meilleure procédure linéaire  $j$  en moyenne sans hypothèse distributionnelle.

3) Quelle est la fourchette de la probabilité d'attribuer un individu de  $\bigcup_{k \in K} E_k \geq \bigcup_{j \in J} E_j$  si  $K \cap J = \emptyset$  ? En ce sens, nous

n'avons pu obtenir que des résultats très partiels et peu satisfaisants.

L'auteur se propose, dans un prochain travail, de comparer numériquement sur des données réelles la méthode proposée à des concurrentes (analyse factorielle discriminante, règle de Bayes en multinormalité,...).

\*

\*

\*

C.Heuchenne. *La meilleure discrimination linéaire sans hypothèse distributionnelle ?* Bull.Soc.Sc.Liège, 49, 3-4, 1980, pp.128-143.

C.R.Rao. *Linear statistical inference and its applications.* Wiley, 1973.