CAHIER DU LAMSADE

Laboratoire d'Analyse et Modélisation de Systèmes pour l'Aide à la Décision
(Université de Paris-Dauphine)
Equipe de Recherche Associée au C.N.R.S. N° 656

DISCRIMINATION PAR L'AJUSTEMENT DE VARIABLES INDICATRICES

CAHIER N° 48
septembre 1983

J.M. DEVAUD

SOMMAIRE

DISCRIMINATION BY FITTING DUMMY VARIABLES

ABSTRACT

Among multivariate statistical methods, discrimination is still a wide opened field. A possible way to deal with that problem consists in estimating the probability to belong to a definite population for an individual. Our purpose is to describe a method giving linear estimates of these probabilities with linear regression. This method has already been proposed and called "Multivariate Nominal scale Analysis (M.N.A.) but the approach presented will allow to consider each subject as a point of a $(k - 1)$-dimensional space (for a discrimination over k groups). Furthermore this approach gives convenient rules to allocate a new individual to one of the k populations. The categories of the qualitative variables can also be graphically displayed as points in the $(k - 1)$-dimensional space and quantitative variables can be introduced among the explanatory variables. We also study how unbalanced groups can be taken into account and, finally, an example is discussed.

DISCRIMINATION PAR L'AJUSTEMENT DE VARIABLES INDICATRICES

RESUME

Parmi les méthodes statistiques multidimensionnelles, la discrimination est un domaine encore largement ouvert. Une façon possible de traiter cette question consiste à estimer la probabilité d'appartenance à une classe pour un individu donné. L'objet de ce cahier est de décrire une méthode donnant des estimations linéaires de ces probabilités à l'aide de régressions multiples. Cette méthode a déjà été proposée sous le nom de "Multivariate Nominal scale Analysis" (M.N.A.) mais l'approche présentée ici permettra de considérer chaque individu comme un point d'un espace de dimension $k - 1$ (pour une discrimination sur k classes). De plus, cette approche fournit des règles commodes d'affectation d'un nouvel individu à l'une des k sous-populations. Les modalités des variables qualitatives peuvent aussi être représentées graphiquement dans l'espace de dimension $k - 1$ et il est possible d'introduire des variables quantitatives parmi les variables explicatives. On montre aussi comment traiter le cas où les classes ont des effectifs déséquilibrés et, enfin, un exemple d'application est présenté.

INTRODUCTION

La discrimination est une préoccupation déjà ancienne des théoriciens et des praticiens de la statistique et de l'analyse des données. La vogue qu'elle a pu connaître tient sans doute d'une part à la richesse des problèmes méthodologiques qu'elle conduit à résoudre et, d'autre part, au nombre et à la variété des applications concrètes qui entrent dans son cadre. Plusieurs travaux de synthèse ont déjà vu le jour sur ce sujet. On citera en particulier ROMEDER (1973), ULMO (1973), GOLDSTEIN et DILLON (1978) et NAKACHE (1980).

Les méthodes de discrimination où des prédicteurs sont qualitatifs constituent un cas particulièrement intéressant, là aussi tant sur le plan méthodologique que sur celui des applications. Ce problème a également fait l'objet de beaucoup de recherches ; on évoquera à ce propos des travaux de TENENHAUS et BOUROCHE (1970) et MORGAN et SONQUIST (1963) concernant la segmentation, ceux présentés par LEBART, MORINEAU et TABARD (1977) visant à adapter l'analyse des correspondances multiples à la discrimination ainsi que ceux de SAPORTA (1977) et de DAUDIN (1978). Malgré toutes les approches qui ont été proposées, il semble que le problème de la discrimination sur prédicteurs qualitatifs mixtes reste encore assez largement ouvert.

Les propos qui suivent ont pour but de présenter une approche dont le principe consiste à "expliquer" les indicatrices associées à la variable définissant les classes à discriminer. Après avoir exposé le cadre dans lequel s'insère cette méthode, on en donnera une interprétation géométrique permettant la représentation simultanée, dans un espace de faible dimension, des individus étudiés et des modalités des variables qualitatives. Enfin, on évoquera les problèmes posés par le cas où les classes à discriminer ont des effectifs trop déséquilibrés et on présentera un exemple d'application.

I - MODELISATION DU PROBLEME DE DISCRIMINATION

Soit a_j une variable qualitative prenant k modalités. a_j définit une partition de la population étudiée E de n individus ($|E| = n$) en k sous-populations notées $E^1, \ldots, E^k$. On cherche à "expliquer" a_j par p variables explicatives $X_1, \ldots, X_p$. $X_1, \ldots, X_p$ pourront être quantitatives ou des indicatrices associées à des variables qualitatives. Dans le cas où des indicatrices figurent parmi les variables X_i, on supposera avoir pris les "précautions d'usage" pour que la matrice X, $X = [1\!\!1_n, X_1, \ldots, X_p]$, soit de plein rang (*).

On notera Y^h l'indicatrice associée à la h^e modalité de a_j (donc attachée à sous-population E^h).

Plaçons-nous dans le cas simple où $k = 2$. On peut alors, pour le i^e individu de E, considérer la variable de Bernoulli Y_i^1 définie comme suit :

$Y_i^1 = 1$ si le i^e individu de E appartient à E^1,

$Y_i^1 = 0$ sinon

avec $\text{Prob}[Y_i^1 = 1] = \pi_i$.

On constate immédiatement que les réalisations des Y_i^1 ($i = 1, \ldots, n$) sont les valeurs prises par l'indicatrice attachée à la population E^1, π_i étant, pour sa part, la probabilité d'appartenance à E^1 du i^e individu.

Notre problème d'analyse discriminante va donc consister à estimer π_i en présupposant que π_i dépend des valeurs prises par les variables explicatives sur le i^e individu de E. On notera x_{ji} la valeur prise par X_j sur le i^e individu de E ($j = 1, \ldots, p$).

(*) Dans la suite, on supposera que chaque colonne correspondant à une modalité explicative est égale à la différence entre la colonne de l'indicatrice associée et celle de l'indicatrice associée à la dernière modalité de la variable qualitative initiale.

Or, on sait bien que $E(Y_i^1) = \pi_i$. Par analogie avec la régression linéaire où le problème est précisément d'estimer $E[Y_i^1 \mid X_1 = x_{1i}, \ldots, X_p = x_{pi}]$, nous pouvons donc envisager d'estimer la probabilité d'appartenance à E^1 en régressant Y^1 sur $X_1, \ldots, X_p$. L'inconvénient de cette approche tient à ce que, la variable à expliquer étant une indicatrice, d'une part il n'y a plus homoscédasticité, ce qui fait perdre à l'estimateur des moindres carrés ordinaires son caractère optimal au sens du théorème de Gauss-Markov, d'autre part les hypothèses de normalité n'étant plus vérifiées, les statistiques utilisées habituellement dans les tests perdent beaucoup de leur intérêt.

Enfin, les probabilités estimées à l'aide de la régression linéaire n'appartiendront pas forcément à l'intervalle [0, 1]. Pour remédier à ce dernier inconvénient, on peut proposer de faire une régression logistique au lieu d'une régression linéaire. Cette approche revient à estimer π_i non plus à l'aide d'une fonction linéaire des x_{ji} mais avec une fonction logistique :

$$\pi_i : \frac{1}{1 + e^{-(\beta_o+\beta_1 x_{1i}+\ldots+\beta_p x_{pi})}}$$

L'estimation de π_i sera alors par construction comprise entre 0 et 1. On estime alors les paramètres $\beta_o, \beta_1, \ldots, \beta_p$ par le maximum de vraisemblance et on teste la validité du modèle par le logarithme du rapport de vraisemblance. Tous ces aspects sont développés par COX (1970). On peut aussi trouver des justifications plus satisfaisantes à l'utilisation d'une fonction logistique dans DAY et KERRIDGE (1967).

Malgré les inconvénients évoqués plus haut, l'approche consistant à régresser linéairement Y^1 sur $X_1, \ldots, X_p$ nous a paru digne d'intérêt. On peut en effet interpréter les probabilités estimées comme des "indicateurs d'appartenance aux classes" et en déduire ainsi des règles d'affectation d'un nouvel individu. Nous verrons aussi qu'on peut trouver une interprétation géométrique fournissant une représentation simultanée des individus et des modalités des variables qualitatives. Enfin, et sur le plan pratique, ce point n'est pas négligeable, un simple programme informatique de régression suffit pour pouvoir utiliser cette méthode.

II - ECRITURE DU MODELE DE BASE

II.1 Différentes expressions du problème

Revenons au cas général où l'on doit discriminer sur k classes. L'approche proposée consiste donc à opérer k régressions multiples. Chacune d'elles consistera à régresser une indicatrice de $\mathcal{Y}$ sur $X_1, \ldots, X_p$. On cherchera donc k jeux de paramètres $\hat{\beta}^1, \ldots, \hat{\beta}^h, \ldots, \hat{\beta}^k$ -

$\hat{\beta}^h = \begin{pmatrix} \hat{\beta}^h_0 \\ \hat{\beta}^h_1 \\ \vdots \\ \hat{\beta}^h_p \end{pmatrix}$ - répondant aux problèmes :

$$(1) \begin{cases} \underset{\beta^h}{\text{Min}} \, \| Y^h - X \beta^h \|^2_{I_n} \\ h = 1, \ldots, k \end{cases}$$

où $\| \ \|_{I_n}$ est la norme euclidienne associée à la matrice identité d'ordre n, notée I_n.

Cette méthode a été proposée par ANDREWS et MESSENGER (1973) sous le nom de "Multivariate Nominal scale Analysis" (MNA). Nous proposons pour notre part quelques compléments qui, nous semble-t-il, devraient permettre de la rendre plus exploitable dans la pratique.

Il est possible de donner à ce problème deux interprétations équivalentes.

D'une part, en se plaçant sur $\mathbb{R}^{nk}$, on peut constater que la recherche de $\hat{\beta}^1, \ldots, \hat{\beta}^k$ se ramène à rechercher la matrice $\hat{\beta}$, $\hat{\beta} = [\hat{\beta}^1 \mid \ldots \mid \hat{\beta}^k]$ permettant d'ajuster au mieux le tableau disjonctif complet Y associé à $\mathcal{Y}$, $Y = [Y^1 \mid \ldots \mid Y^k]$.

En effet, on doit alors résoudre

$$\underset{\beta}{\text{Min}} \, \| Y - X\beta \|^2_{I_{nk}} . \tag{2}$$

$$\text{Or, } \| Y - X\beta \|^2_{I_{nk}} = \sum_{h=1}^{k} \sum_{e \in E} [Y^h(e) - X(e)\beta^h]^2 \tag{3}$$

en posant : $Y^h(e)$ = valeur prise par Y^h sur l'individu e de E

$X(e) = (1, X_1(e), \ldots, X_p(e))$ = ligne associée à l'individu e dans X

et $(3) = \sum_{h=1}^{k} \| Y^h - X\beta^h \|^2_{I_n}$.

Le minimum de (2) sera donc atteint pour la matrice $\hat{\beta}$ dont les colonnes $\hat{\beta}^h$ $(h = 1, \ldots, k)$ sont les vecteurs de paramètres estimés dans la régression de Y^h sur les colonnes de X.

D'autre part, (3) peut encore s'écrire, en inversant les signes sommes :

$$(3) = \sum_{e \in E} \sum_{h=1}^{k} [Y^h(e) - X(e)\, \beta^h]^2$$

$$= \sum_{e \in E} \| Y(e) - X(e)\beta \|^2_{I_k} \tag{4}$$

en posant $Y(e) = (Y^1(e), \ldots, Y^k(e))$.

$$\text{Soit encore } (4) = \sum_{h=1}^{k} \sum_{e \in E^h} \| Y(e) - X(e)\beta \|^2_{I_k} \tag{5}$$

Or, pour tout individu e de E^h, Y(e) a toutes ses composantes nulles à l'exception de la h^e qui vaut 1. Nous noterons C^h ce k-uplet caractérisant l'appartenance à la classe E^h. Donc :

$$e \in E^h \iff Y(e) = C^h .$$

Minimiser (5) s'écrit alors :

$$\text{(5 bis)} \quad \min_{\beta} \sum_{h=1}^{k} \sum_{e \in E^h} \| C^h - X(e)\beta \|^2_{I_k} .$$

On constate donc que MNA a une interprétation immédiate dans $\mathbb{R}^k$.

II.2 Une propriété du modèle

Soit W le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de X rapporté à la base correspondante. On sait que $\hat{Y}^h$, $\hat{Y}^h = X\,\hat{\beta}^h$ est la projection orthogonale de Y^h sur W au sens du produit scalaire usuel $\hat{Y}^h = \text{Proj}_W^{\perp} Y^h$.

Donc $\sum_{h=1}^{k} \hat{Y}^h = \text{Proj}_W^{\perp}(\Sigma\, Y^h) = \text{Proj}_W^{\perp}(\mathbb{1}_n) = \mathbb{1}_n$, ce qui montre que $\sum_{h=1}^{k} \hat{Y}^h(e) = 1$. D'autre part, $\hat{Y}^h = \hat{\beta}_o^h\, \mathbb{1}_n + \hat{\beta}_1^h\, X_1 + \ldots + \hat{\beta}_p^h\, X_p$. Donc, comme $\sum_{h=1}^{k} \hat{Y}^h = \mathbb{1}_n$, on peut conclure que $\sum_{h=1}^{k} \hat{\beta}_o^h = 1$ et $\sum_{h=1}^{k} \hat{\beta}_j^h = 0$ $(j = 1, \ldots, p)$.

II.3 Remarque

Si, en théorie, la présence de variables X_j quantitatives ne pose pas de problèmes, en pratique, si les valeurs prises par ces variables sont grandes devant 1, le fait d'ajuster les valeurs en 0-1 des indicatrices conduira à trouver des β_j^h quasiment nuls, voire nuls à l'arrondi de l'ordinateur près.

Pour contourner cette difficulté, on peut bien sûr travailler sur des X_j centrées-réduites mais on peut aussi leur faire subir la transformation suivante :

$$X_j \to X_j^* = 2\,\frac{X_j - X_{j\min}}{X_{j\max} - X_{j\min}} - 1$$

où X_{jmin} (resp. X_{jmax}) représente la plus petite (resp. grande) valeur observée de X_j. On a alors :

$$-1 \leq X_j^* \leq 1.$$

Les X_j^* sont alors "comparables" aux colonnes correspondant aux indicatrices dans X. Il est alors possible de constater que :

$$X_j^* = 2 \frac{X_j - \left(\frac{X_{jmax} + X_{jmin}}{2}\right)}{X_{jmax} - X_{jmin}}.$$

X_j^* est donc une variable "centrée-réduite" en prenant la demi-étendue comme indicateur de tendance centrale et l'étendue comme indicateur de dispersion.

III - ASPECTS GEOMETRIQUES ET VISUALISATIONS

On a vu en (5) que MNA pouvait trouver une interprétation dans $\mathbb{R}^k$; nous allons développer cette interprétation.

III.1 Principe de l'interprétation dans $\mathbb{R}^k$

A chaque individu e de E, on peut donc associer un k-uplet $(\hat{Y}^1(e), \ldots, \hat{Y}^k(e))$ que nous noterons $\hat{Y}(e)$. $\hat{Y}(e)$ peut être interprété comme un point d'un espace affine de dimension k. Nous pourrons donc représenter chaque individu e par le point $\hat{Y}(e)$. De plus, on a vu que, pour tout e de E, $\sum_{h=1}^{k} \hat{Y}^h(e) = 1$; tous les $\hat{Y}(e)$ sont donc dans un espace de dimension $k - 1$ que nous noterons $\mathcal{L}_k$ et que nous appellerons "espace des classes".

De plus, les k-uplets $C^1, \ldots, C^h, \ldots, C^k$ qui caractérisent l'appartenance respective des individus aux différentes classes $E^1, \ldots, E^h, \ldots, E^k$ peuvent, eux aussi, être considérés comme des points de $\mathcal{L}_k$. Nous pourrons donc visualiser, dans l'"espace des classes", tous les individus de E et ils seront positionnés par rapport aux points représentatifs de leur classe d'appartenance. D'autre part, on voit maintenant apparaître que le critère (5 bis) traduit que les $\hat{Y}(e)$ seront construits de telle façon qu'ils soient globalement aussi proches que possible des points représentatifs des classes d'appartenance.

Lorsque $k = 2$, $\mathcal{L}_k$ sera bien sûr une droite (cf. figure 1) et on pourra illustrer les résultats de l'analyse en utilisant un histogramme représentant l'effectif des individus de E^1 et E^2 dont le codage se situe sur un segment donné de la droite $\mathcal{L}_2$ (figure 2).

Lorsque $k = 3$, $\mathcal{L}_k$ sera alors un plan (cf. figure 3) et on pourra visualiser différemment les individus suivant leur classe d'origine (figure 4).

Figure 1

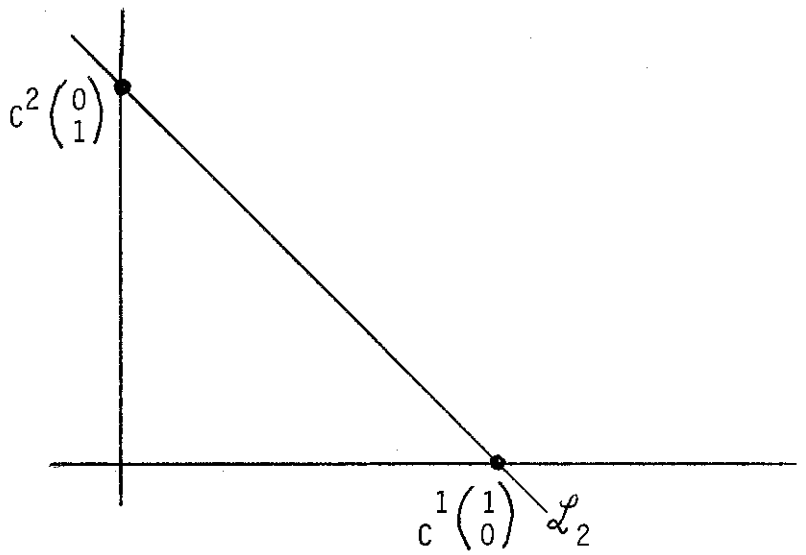


Figure 2

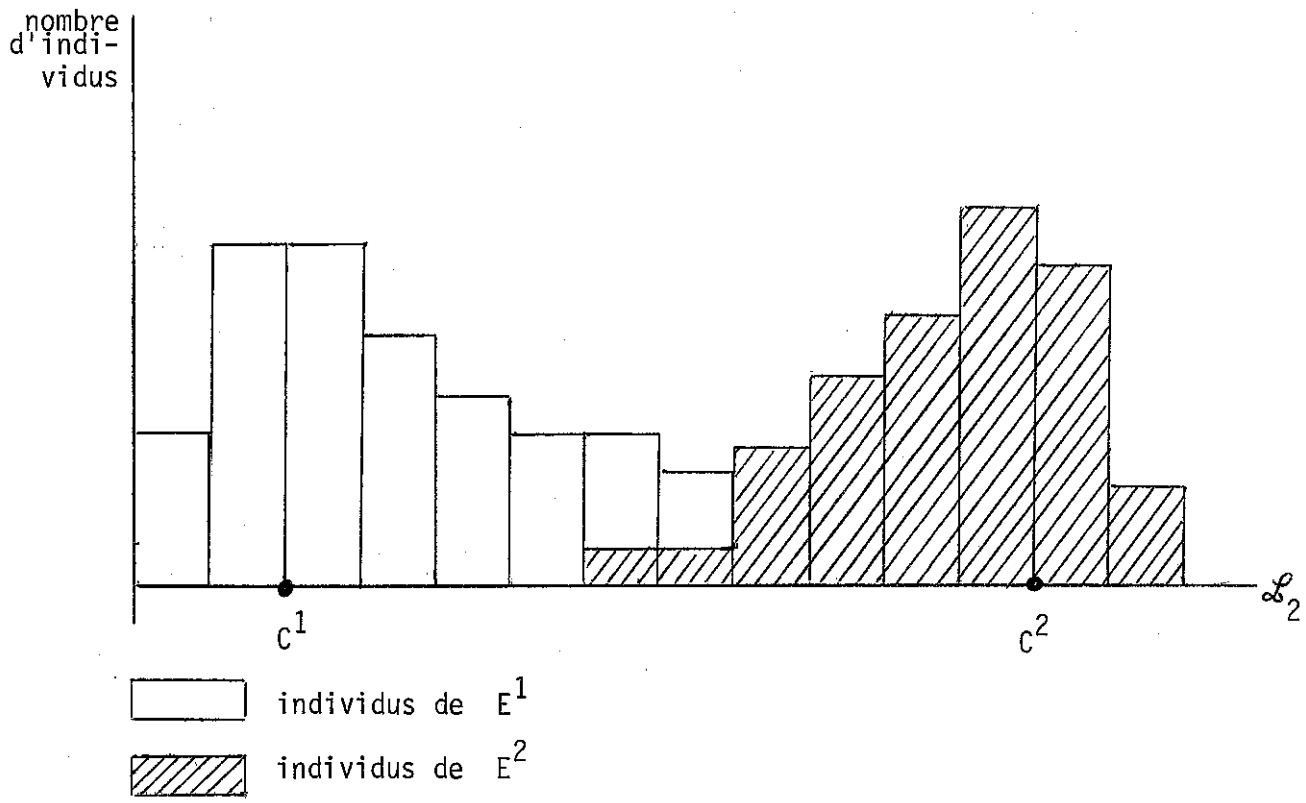


Figure 3

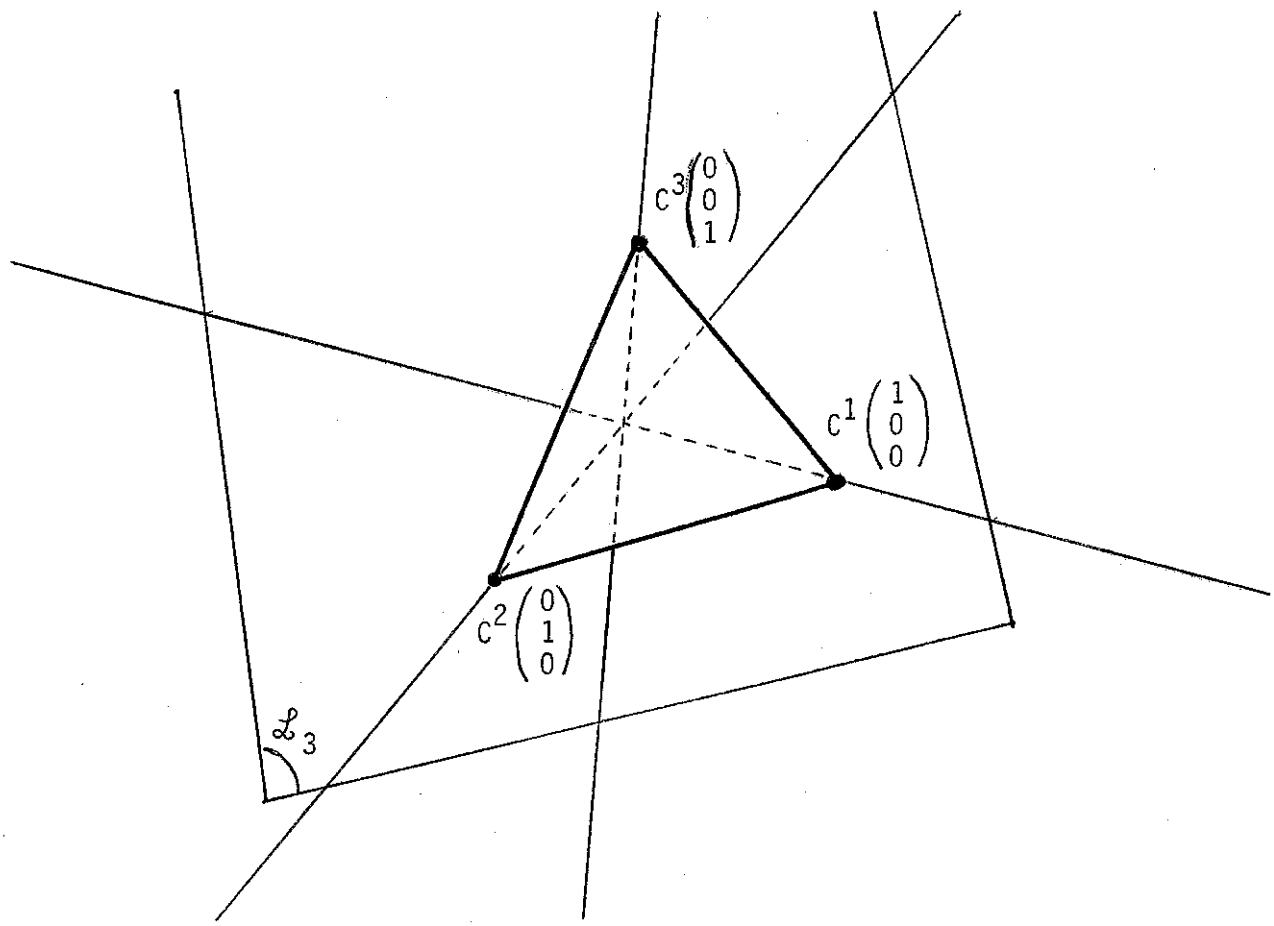
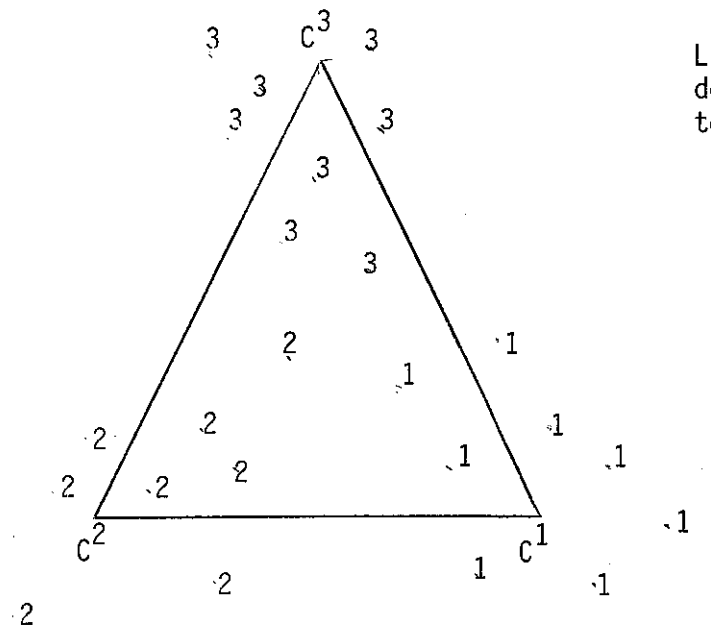


Figure 4



Les numéros sont ceux de la classe d'appartenance

Si $k > 3$, $\mathcal{L}_k$ sera plus difficile à représenter. Il est cependant possible d'utiliser une analyse en composantes principales pour obtenir une représentation la moins déformée possible.

III.2 Réaffectation des individus

L'interprétation géométrique permet d'induire une règle de réaffectation des individus. Dans le cas où le modèle est bon, lorsque $e \in E^h$, $\hat{Y}(e)$ devrait être proche de C^h. Donc, il semble naturel d'affecter un nouvel individu e à la classe E^h si C^h est le point de $\{C^1, \ldots, C^k\}$ le plus proche de $\hat{Y}(e)$. Autrement dit :

$$(6) \quad e \text{ affecté à } E^h \iff d(\hat{Y}(e), C^h) = \operatorname{Min}_l \{d(\hat{Y}(e), C^l)\}.$$

Il se pose alors le choix de la distance d. Il paraît assez cohérent avec le reste de la méthode de choisir la distance définie par la norme utilisée dans (5 bis).

A cet égard, il nous semble utile de signaler deux résultats ; ceux-ci sont justifiés de façon plus détaillée en annexe.

Tout d'abord, on peut montrer que, lorsque d est la distance associée à la norme euclidienne canonique, la règle de réaffectation devient :

$$(7) \quad e \text{ affecté à } E^h \iff \hat{Y}^h(e) = \operatorname{Max} \{\hat{Y}^1(e), \ldots, \hat{Y}^k(e)\}.$$

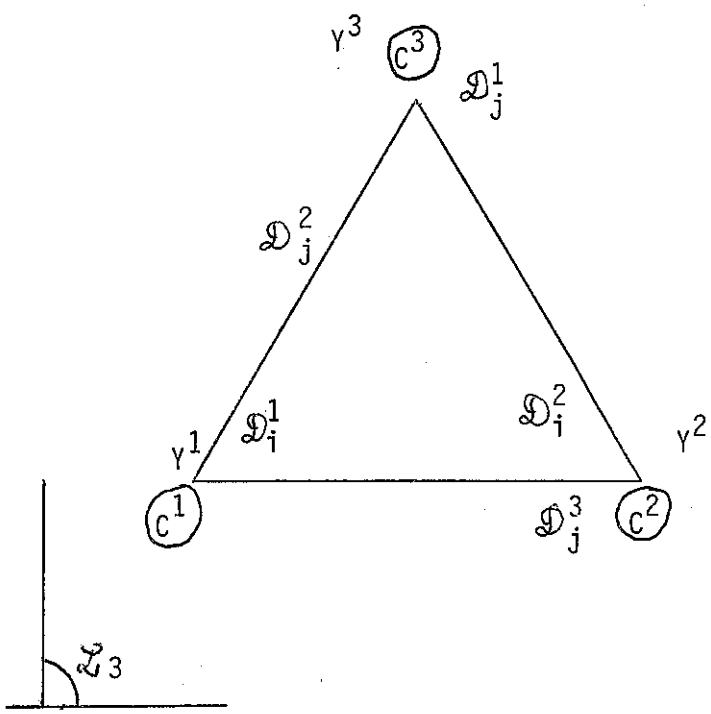
Par ailleurs, lorsque $k = 2$, on peut montrer que la règle de réaffectation est invariante quelle que soit la norme euclidienne utilisée et que la règle de réaffectation est la suivante :

$$(8) \quad \begin{cases} e \text{ affecté à } E^1 \iff \hat{Y}^1(e) > \frac{1}{2} \\ e \text{ affecté à } E^2 \iff \hat{Y}^1(e) < \frac{1}{2} \end{cases}$$

III.3 Visualisation des modalités des variables qualitatives dans $\mathcal{L}_k$

Puisque chaque individu e de E peut être visualisé par un point $\hat{Y}(e)$ de $\mathcal{L}_k$, on peut envisager de représenter chaque modalité des variables qualitatives par le point moyen des $\hat{Y}(e)$ des individus e prenant cette modalité. Cette idée, fréquemment utilisée en analyse factorielle, permettra de "juger visuellement" du pouvoir discriminant des différentes variables qualitatives selon la proximité des points représentatifs des modalités des points $C^1, \ldots, C^k$. La figure 5 illustre ces propos dans le cas où $k = 3$.

Figure 5



$\mathcal{D}_i$ et $\mathcal{D}_j$ sont deux variables explicatives qualitatives.

$\mathcal{D}_i^1, \mathcal{D}_i^2$ sont des modalités de $\mathcal{D}_i$.

$\mathcal{D}_j^1, \mathcal{D}_j^2, \mathcal{D}_j^3$ sont des modalités de $\mathcal{D}_j$.

Y^1 et Y^2 sont des modalités de $\mathcal{Y}$.

$\mathcal{D}_i^1, \mathcal{D}_i^2, \mathcal{D}_j^1, \mathcal{D}_j^2, \mathcal{D}_j^3, Y^1, Y^2$ sont positionnés aux centres de gratité respectifs des individus prenant ces modalités.

C^1, C^2, C^3 sont définis comme précédemment.

IV - PROBLEMES POSES PAR DES DONNEES DESEQUILIBREES

Nous allons voir que de trop grandes disparités entre les effectifs $|E^1|, \ldots, |E^k|$ des classes $E^1, \ldots, E^k$ entraînent quelques inconvénients puis nous étudierons comment tenter de les éviter.

IV.1 Position du problème

Plaçons-nous dans le cas où $k = 2$. Supposons que $|E^1| = 80$ et $|E^2| = 20$. On sait qu'en régression, la moyenne des observations est égale à la moyenne des estimations, donc :

$$\frac{1}{n} \sum_{e \in E} \hat{Y}^1(e) = \frac{1}{n} \sum_{e \in E} Y^1(e) = \frac{80}{100} = 0{,}8 \ ;$$

semblablement, $\frac{1}{n} \sum_{e \in E} \hat{Y}^2(e) = 0{,}2$.

On constate donc que le point moyen des $\hat{Y}(e)$ a pour coordonnées $(0{,}8 \ ; \ 0{,}2)$. Par construction, le modèle a donc tendance à attirer les $\hat{Y}(e)$ vers C^h lorsque E^h est la classe la plus nombreuse. Autrement dit, les points $\hat{Y}(e)$ seront, dans l'ensemble, plus proches de ce point C^h que des autres points C^l $(l \neq h)$ et du seul fait que $|E^h| > |E^l|$, $l \neq h$.

IV.2 Régression pondérée

On peut expliquer facilement l'inconvénient évoqué ci-dessus à partir du critère (5 bis). On y constate en effet que toutes les classes y sont considérées sur le même plan sans distinction suivant leur effectif et le poids des classes à fort effectif est tel que celles-ci attirent les $\hat{Y}(e)$ vers les points C^h correspondants. Afin d'y remédier, on peut envisager de modifier le critère (5 bis) et de minimiser $\sum_{h=1}^{k} \frac{1}{|E^h|} \sum_{e \in E^h} \| C^h - X(e)\beta \|_{I_k}^2$ (9)

au lieu de $\sum_{h=1}^{k} \sum_{e \in E^h} \|C^h - X(e)\beta\|^2_{I_k}$, ce qui permet de donner à chaque classe le même poids. Or $\sum_{h=1}^{k} \frac{1}{|E^h|} \sum_{e \in E^h} \|C^h - X(e)\beta\|^2_{I_k} = \sum_{h=1}^{k} \sum_{e \in E^h} \frac{1}{|E^h|} \|C^h - X(e)\beta\|^2_{I_k}$, soit encore, si on note $p(e)$ le poids de l'individu e tel que $p(e) = \frac{1}{|E^h|}$ si $e \in E^h$, alors :

$$(9) = \sum_{e \in E} \frac{1}{p(e)} \|Y(e) - X(e)\beta\|^2_{I_k}$$

$$= \sum_{h=1}^{k} \sum_{e \in E} \frac{1}{p(e)} (Y^h(e) - X(e)\,\beta^h)^2 = \sum_{h=1}^{k} \|Y^h - X\,\beta^h\|^2_{D^{-1}}$$

où D^{-1} est la matrice diagonale d'ordre n dont les éléments diagonaux sont les quantités $1/p(e)$. Autrement dit :

$$D^{-1} = \begin{bmatrix} 1/|E^1| & & & 0 & \\ & \ddots & & & \\ & & 1/|E^1| & & \\ \hline & 0 & & 1/|E^k| & \\ & & & & \ddots \\ & & & & & 1/|E^k| \end{bmatrix} \begin{matrix} \\ \\ \\ \} |E^1| \text{ lignes} \\ \\ \\ \} |E^k| \text{ lignes} \end{matrix}$$

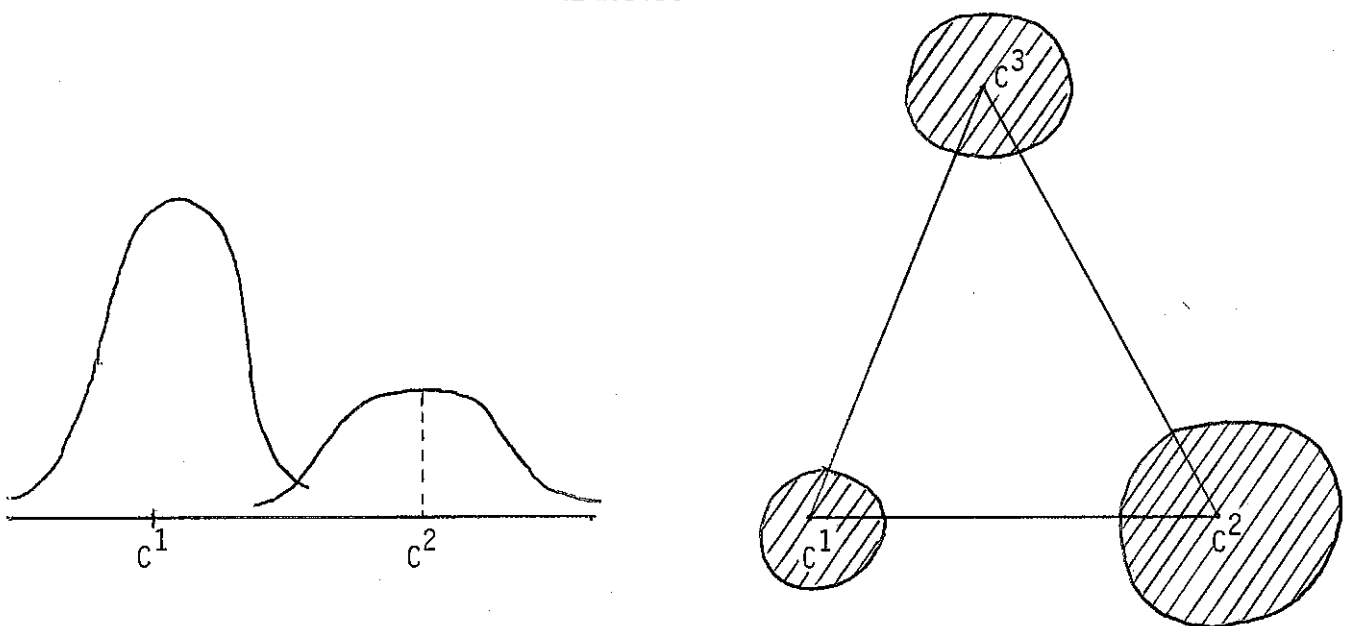
On en déduit que, minimiser le critère (9), revient à opérer k régressions unidimensionnelles mais en munissant $\mathbb{R}^n$ non plus de la norme euclidienne "canonique" mais de celle associée à D^{-1}. On effectuera donc k régressions pondérées où chaque individu aura pour poids l'inverse de l'effectif de sa classe.

On peut montrer (voir DEVAUD (1982)) que les propriétés du II.2 sont conservées et que des représentations dans $\mathcal{L}_k$ sont toujours possibles.

IV.3 Régression sous contraintes

Une seconde approche, également envisageable, consiste à imposer aux $\hat{Y}(e)$, pour les e appartenant à la classe E^h, d'admettre C^h pour centre de gravité. En particulier, lorsque les différents individus seront bien discriminés, on peut espérer avoir des représentations de la forme suivante :

Figure 6



Imposer, pour chaque sous-population E^h, aux $\hat{Y}(e)$, $e \in E^h$, d'avoir leur point moyen en C^h peut s'écrire :

$$\begin{cases} \sum_{e \in E^h} \hat{Y}^h(e) = |E_h| \\ \sum_{e \in E^h} \hat{Y}^i(e) = 0 \qquad i \neq h \\ h = 1, \ldots, k \end{cases}$$

On a donc alors :

$$\begin{cases} \sum_{e \in E^h} \hat{Y}^h(e) = |E_h| \\ \sum_{\substack{e \in E^i \\ i \neq h}} \hat{Y}^h(e) = 0 \\ h = 1, \ldots, k \end{cases}$$

ce que l'on peut écrire matriciellement $Y' X \beta = N$ où N est une matrice diagonale d'ordre k dont le h^e élément diagonal est $|E^h|$.

Nous allons donc maintenant chercher la matrice β^* de format $(p + 1, k)$ permettant d'ajuster au mieux Y mais sous les contraintes que nous venons de décrire. En se rappelant que $\|Y - X\beta\|^2_{I_{nk}} = \text{Trace } [(Y - X\beta)' (Y - X\beta)]$, on devra résoudre :

$$\begin{cases} \underset{\beta}{\text{Min}} \ [\text{Trace } (Y - X\beta)' (Y - X\beta)] \\ \text{avec } Y' X\beta = N \end{cases}$$

Intéressons-nous aux conditions nécessaires d'optimum. Soit Λ la matrice carrée d'ordre k contenant les k^2 multiplicateurs de Lagrange du problème. Le lagrangien de notre problème s'écrit alors :

$$\mathcal{L}(\beta, \Lambda) = [\text{Trace } (Y - X\beta)' (Y - X\beta)] - [\text{Trace } \Lambda(Y' X\beta - N)].$$

En annulant les dérivées partielles de $\mathcal{L}$, il vient (cf. RAO (1973)) :

$$\begin{cases} 2 X' X\beta - 2 X' Y - X' Y \Lambda' = 0 & (10.1) \\ Y' X\beta - N = 0 & (10.2) \end{cases}$$

Après avoir prémultiplié les deux membres de 10.1 par $Y' X (X' X)^{-1}$, on a :

$$2 Y' X\beta - 2 Y' X(X' X)^{-1} X' Y - Y' X(X' X)^{-1} X' Y \Lambda' = 0,$$

d'où $\Lambda' = 2 [Y' X(X' X)^{-1} X' Y]^{-1} Y' X\beta - 2 I_k$ (10.3)

à condition que $Y' X(X' X)^{-1} X' Y$ sont inversible. Or, $(X' X)^{-1}$ est une matrice symétrique et définie positive de rand $p + 1$, on peut donc l'écrire sous la forme $C' C$ où C est une matrice carrée régulière d'ordre $p + 1$. $Y' X(X' X)^{-1} X' Y$ apparaît donc comme une matrice de Gram dont la composante gramienne est $C X' Y$. On doit donc s'intéresser au rang de $C X' Y$. C étant inversible, son endomorphisme associé (une base étant choisie dans $\mathbb{R}^{p+1}$) est bijectif ; le rang de $C X' Y$ est donc égal au rang de $X' Y$. Pour que $Y' X(X' X)^{-1} X' Y$ soit inversible, on doit donc s'assurer que $X' Y$ est de rang k et que $p + 1 \geq k$.

On imposera donc $p + 1 \geq k$ et rang $(X' Y) = k$. Il est facile de constater que les k colonnes de $X' Y$ sont constituées des coordonnées des centres de gravité des k sous-nuages multipliées chacune par l'effectif de la sous-population correspondante en se plaçant dans l'espace affine F de dimension $p + 1$ naturellement associé au problème (à chacune des p variables explicatives est associée une dimension, la $(p + 1)^e$ variable étant identiquement égale à 1). Ainsi, la h^e colonne de $X' Y$ s'écrira :

$$\begin{bmatrix} |E^h| \\ |E^h| \bar{x}_1^h \\ \vdots \\ |E^h| \bar{x}_p^h \end{bmatrix}$$

où $\bar{x}_i^h$ est la moyenne de X_i calculée sur la sous-population E^h.

En général, $X' Y$ sera donc de rang k. Les cas où cette propriété ne sera pas vérifiée correspondent à des configurations particulières des centres de gravité dans l'espace affine F cité plus haut. On peut montrer qu'il en est ainsi lorsque certains de ces centres de gravité sont confondus ou alignés ou, plus généralement, dans une variété linéaire de F.

En tenant compte du fait que $Y' X \beta = N$ et en reportant 10.3 dans 10.1, il vient :

$$2 X' X \beta = 2 X' Y + 2 X' Y([Y' X(X' X)^{-1} X' Y]^{-1} N - I_k).$$

En notant $\hat{\beta}$ l'optimum du problème sans contrainte ($\hat{\beta} = (X' X)^{-1} X' Y$), on aboutit à :

$$\beta^* = \hat{\beta} \, (Y' \, X \, \hat{\beta})^{-1} \, N.$$

La fonction objectif du problème apparaît clairement comme une fonction convexe et les contraintes sont linéaires. β^* est donc bien l'optimum recherché.

Là encore (cf. DEVAUD (1982)), on peut montrer que les propriétés du II.2 sont conservées et que les visualisations dans $\mathcal{L}_k$ sont encore valables. On remarquera également que, sous les contraintes imposées, les quantités $\sum_{e \in E^h} \|C^h - X(e)\beta\|^2_{I_k}$ introduites dans le critère (5 bis) s'interprètent comme les inerties de chacune des classes par rapport à leur centre de gravité dans $\mathcal{L}_k$.

V - UN EXEMPLE D'APPLICATION

Cet exemple d'application porte sur des données d'accidents de travail s'étant produits en 1978 dans une grande entreprise nationalisée française. Plus précisément, on disposait de 3 000 accidents s'étant produits dans cette entreprise cette année-là et, pour chacun d'entre eux, on connaissait le type (accident de trajet ou accident de travail pendant le service) ainsi que l'âge, le sexe et le niveau hiérarchique de l'accidenté. Notre problème consistait à expliquer le type d'accident par les variables âge, sexe et niveau hiérarchique. La variable âge comprenait 4 modalités (20-30 ans, 30-40 ans, 40-50 ans, 50-60 ans) et la variable niveau hiérarchique 3 modalités (agent d'exécution, agent de maîtrise, cadre). On a effectué les traitements sur 4 fichiers (notés F1, ..., F4) de 1 500 accidents (tirés au hasard parmi les 3 000) différant les uns des autres par 10 % de leurs enregistrements, cela afin d'avoir une idée de la robustesse des résultats.

Les données étant déséquilibrées (environ 1 accident de trajet pour 4 accidents en service), on a utilisé l'approche consistant à utiliser des régressions pondérées (cf. IV.2).

Les estimations des différents paramètres sur les 4 fichiers figurent dans le tableau 1. Ces résultats correspondent à la régression de l'indicatrice associée aux accidents de trajet ; les paramètres de la régression de l'indicatrice associée à l'autre classe s'en déduisent à l'aide des résultats vus au II.2.

On constate sur ce tableau que c'est toujours le paramètre correspondant à l'indicatrice des cadres qui est le plus fort, ce qui tendrait à prouver que c'est cette population qui a la plus forte propension à avoir des accidents de trajet, les femmes venant immédiatement après. A l'opposé, on trouve l'indicatrice "hommes" et, juste avant, l'indicatrice

"exécution", les deux catégories correspondantes étant donc plutôt atteintes par des accidents en service. On remarque également que les coefficients associés aux indicatrices de l'âge sont les plus faibles en valeur absolue ; le rôle de cette variable dans la discrimination apparaît donc comme relativement faible.

Comme on discrimine entre deux classes, la règle de réaffectation est invariante quelle que soit la métrique choisie et on doit comparer les $\hat{Y}^1(e)$ à $\frac{1}{2}$ (cf. III.2). Il est clair que, dans le cas qui nous préoccupe, tous les individus d'une même sous-population constituée par un croisement des variables âge, sexe et niveau hiérarchique auront le même score $\hat{Y}^1(e)$, donc tous les individus d'une même sous-population auront la même réaffectation. Les valeurs des $\hat{Y}^1(e)$ associés à chaque sous-population figurent dans le tableau 2.

Globalement, pour les quatre fichiers, le pourcentage d'individus mal réaffectés est de l'ordre de 21 %. Sur les fichiers F1, F2, F3, on affectera, dans la classe des accidents de trajet, tous les cadres, les "femmes-maîtrise", les "femmes-exécution" ainsi que les hommes-maîtrise de 50 à 60 ans ; les autres sont affectés dans la classe accidents en service. Sur le fichier F4, la règle est modifiée : les "hommes-maîtrise" sont tous affectés aux accidents en service. Il est possible, à cet égard, que la présence de la variable âge, qui semble avoir un pouvoir discriminnant assez faible, soit à l'origine de ces quelques "perturbations" dans la règle de réaffectation.

Sur la figure 7, on a représenté les modalités des variables (variables explicatives et variable expliquée) dans l'espace $\mathcal{L}_2$ pour les 4 fichiers comme indiqué au III.3. Les valeurs des abscisses des points représentatifs des modalités sont données dans le tableau 3 (l'origine de l'axe est le point C^2 associé aux accidents en service). On constate que les points "cadres" et "femmes" ont des abscisses supérieures à celle du point "trajet" alors que celles des points "hommes" et "exécution"

sont inférieures à celle du point "service", ce qui met en évidence les propensions des cadres et des femmes aux accidents de trajet et celles des hommes et des agents d'exécution aux accidents en service. Toutefois, on constate également que les femmes sont plus nettement au-dessus du point représentatif "trajet" que les hommes au-dessous du point "service", ce qui traduit que la propension de femmes pour les accidents de trajet est plus forte que celle des hommes pour les accidents en service. Enfin, les modalités de la variable âge sont représentées par des points situés entre les points "trajet" et "service", ce qui tend à mettre en évidence le caractère peu discriminant de cette variable.

Tableau 1

	F1	F2	F3	F5
Constante	0,774	0,772	0,764	0,758
Age : 20-30 ans	0,004	0,005	0,006	0,008
30-40 ans	- 0,007	- 0,009	- 0,016	- 0,007
40-50 ans	- 0,036	- 0,033	- 0,019	- 0,013
50-60 ans	0,039	0,037	0,029	0,012
Sexe : Hommes	- 0,235	- 0,223	- 0,220	- 0,225
Femmes	0,235	0,223	0,220	0,225
Catégorie : Exécution	- 0,167	- 0,168	- 0,164	- 0,149
Maîtrise	- 0,076	- 0,064	- 0,068	- 0,076
Cadres	0,243	0,232	0,232	0,225

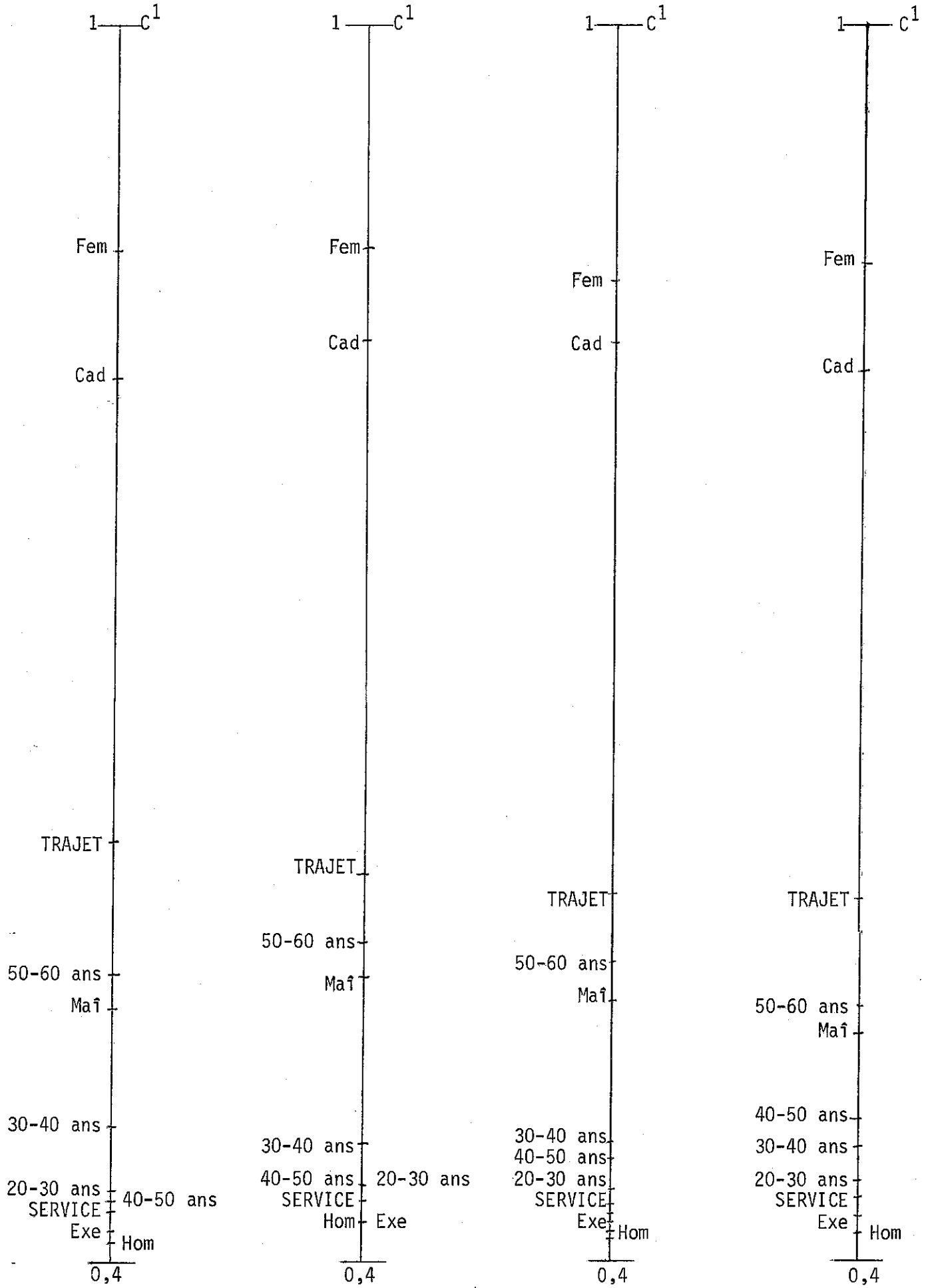
Tableau 3

Modalités	F1	F2	F3	F4
Hommes	0,406	0,418	0,416	0,413
Exécution	0,409	0,417	0,419	0,426
Service	0,418	0,431	0,430	0,429
40-50 ans	0,423	0,435	0,453	0,466
20-30 ans	0,427	0,434	0,431	0,435
30-40 ans	0,460	0,455	0,455	0,459
Maîtrise	0,515	0,532	0,524	0,507
50-60 ans	0,532	0,551	0,540	0,519
Trajet	0,595	0,582	0,577	0,573
Cadres	0,825	0,839	0,841	0,825
Femmes	0,887	0,860	0,869	0,873

Tableau 2

Sous-populations			F1	F2	F3	F4
Exécution	Hommes	20-30 ans	0,376	0,386	0,386	0,392
		30-40 ans	0,365	0,372	0,364	0,377
		40-50 ans	0,336	0,348	0,361	0,371
		50-60 ans	0,411	0,418	0,409	0,396
	Femmes	20-30 ans	0,846	0,832	0,826	0,842
		30-40 ans	0,835	0,818	0,804	0,827
		40-50 ans	0,806	0,794	0,801	0,821
		50-60 ans	0,881	0,864	0,849	0,846
Maîtrise	Hommes	20-30 ans	0,467	0,490	0,482	0,465
		30-40 ans	0,456	0,476	0,460	0,450
		40-50 ans	0,427	0,452	0,457	0,444
		50-60 ans	0,502	0,522	0,505	0,469
	Femmes	20-30 ans	0,937	0,936	0,922	0,915
		30-40 ans	0,926	0,922	0,900	0,900
		40-50 ans	0,897	0,898	0,897	0,894
		50-60 ans	0,972	0,968	0,945	0,919
Cadres	Hommes	20-30 ans	0,786	0,786	0,782	0,766
		30-40 ans	0,775	0,772	0,760	0,751
		40-50 ans	0,746	0,748	0,757	0,745
		50-60 ans	0,821	0,818	0,805	0,770
	Femmes	20-30 ans	1,256	1,232	1,222	1,216
		30-40 ans	1,245	1,218	1,200	1,201
		40-50 ans	1,216	1,194	1,197	1,195
		50-60 ans	1,291	1,264	1,245	1,220

Figure 7


1
C1
Fem
Cad
TRAJET
50-60 ans
Maî
30-40 ans
20-30 ans
40-50 ans
SERVICE
Exe
Hom
0,4
1
C1
Fem
Cad
TRAJET
50-60 ans
Maî
30-40 ans
40-50 ans
20-30 ans
SERVICE
Hom
Exe
0,4
1
C1
Fem
Cad
TRAJET
50-60 ans
Maî
30-40 ans
40-50 ans
20-30 ans
SERVICE
Exe
Hom
0,4
1
C1
Fem
Cad
TRAJET
50-60 ans
Maî
40-50 ans
30-40 ans
20-30 ans
SERVICE
Exe
Hom
0,4

CONCLUSION

L'approche qui vient d'être exposée a donc permis de donner une représentation des individus et des modalités des variables qualitatives dans un espace de dimension k - 1, dans le cas d'un problème à k classes. On a également proposé deux façons d'éliminer les inconvénients liés aux données déséquilibrées.

Afin d'améliorer encore cette méthode, plusieurs voies de recherche sont possibles. En particulier, il conviendrait sans doute d'intégrer une procédure permettant de juger de l'importance de l'apport d'un prédicteur au modèle et, par la même occasion, un moyen de sélectionner les variables explicatives pas à pas. Une difficulté résiderait sans doute dans l'absence de normalité évoquée au I. On pourrait alors envisager d'utiliser des méthodes de type "Monte-Carlo" pour donner une réponse à cette question. Il serait sans doute également intéressant de trouver un mode de représentation des prédicteurs quantitatifs dans $\mathcal{E}_k$, ce qui permettrait de les visualiser simultanément avec les autres prédicteurs.

ANNEXE : REMARQUES SUR LE CHOIX D'UNE DISTANCE DANS $\mathcal{L}_k$

Supposons que $\mathbb{R}^k$ soit muni de la norme euclidienne associée à une matrice Q, $Q = (q_{ij})$, $(i, j) \in \{1, \ldots, k\}^2$. On a :

$$\begin{aligned} d^2(\hat{Y}(e), C^1) &= \|\hat{Y}(e) - C^1\|_Q^2 \\ &= \|\hat{Y}(e)\|_Q^2 + \|C^1\|_Q^2 - 2\, \hat{Y}(e)\, Q\, C^{1\prime} \\ &= \|\hat{Y}(e)\|_Q^2 + q_{11} - 2 \sum_{i=1}^{k} \hat{Y}^i(e)\, q_{i1} \end{aligned}$$

Minimiser $d(\hat{Y}(e), C^1)$ reviendra donc à minimiser $q_{11} - 2 \sum_{i=1}^{k} \hat{Y}^i(e)\, q_{i1}$.

- Si Q est diagonale, $q_{i1} = 0$ pour $i \neq 1$; donc il suffira de minimiser sur 1 la quantité $q_{11}[1 - 2\, \hat{Y}^1(e)]$.

- Si Q est scalaire et, en particulier, si $Q = I_k$ (cas de la norme euclidienne canonique), tous les q_{11} sont égaux ; donc trouver le minimum des $d(\hat{Y}(e), C^1)$ reviendra à trouver le maximum sur 1 des $\hat{Y}^1(e)$.

Plaçons-nous dans le cas où $k = 2$ et munissons $\mathbb{R}^2$ de la norme euclidienne associée à Q, $Q = \begin{bmatrix} q_1 & q_{12} \\ q_{12} & q_2 \end{bmatrix}$. On a vu, au paragraphe précédent, qu'il suffisait alors de comparer $(q_1 - 2[\hat{Y}^1(e)\, q_1 + \hat{Y}^2(e)\, q_{12}]$ à $(q_2 - 2[\hat{Y}^1(e)\, q_{12} + \hat{Y}^2(e)\, q_2])$. Or, comme $\hat{Y}^1(e) + \hat{Y}^2(e) = 1$, on a :

$$\|\hat{Y}(e) - C^1\|_Q^2 - \|\hat{Y}(e) - C^2\|_Q^2 = (1 - 2\, \hat{Y}^1(e))\, (q_1 + q_2 - 2\, q_{12}).$$

mais $q_1 + q_2 - 2\, q_{12} = \|C^1\, C^2\|_Q^2 > 0$. Donc $(\|\hat{Y}(e) - C^1\|_Q^2 - \|\hat{Y}(e) - C^2\|_Q^2)$ est du signe de $1 - 2\, \hat{Y}^1(e)$. La recherche du minimum sur 1 de $d(\hat{Y}(e), C^1)$ se fera donc, pour n'importe quelle matrice Q, suivant la valeur de $\hat{Y}^1(e)$ $(\hat{Y}^1(e) > \frac{1}{2}$ ou $\hat{Y}^1(e) < \frac{1}{2})$.

BIBLIOGRAPHIE

ANDREWS F.M. et MESSENGER R.C. (1973) : "Multivariate Nominal Scale Analysis : A report on a new analysis technique and a computer program", Survey Research Letter, Institute for Social Research, University of Michigan, Ann Arbor, Michigan.

COX D.R. (1970) : The Analysis of Binary Data, Methuen, Londres.

DAUDIN J.J. (1978) : "Etude de la liaison entre variables aléatoires - Régression sur variables qualitatives", Thèse de 3e cycle, Université Paris XI.

DAY N.E. et KERRIDGE D.F. (1967) : "A general maximum likelihood discriminant", Biometrics, Vol. 23, pp. 313-323.

DEVAUD J.M. (1982) : "Discrimination et description sur variables qualitatives - Application à des données d'accidents de travail", Thèse de 3e cycle, Université de Paris-Dauphine.

GOLDSTEIN M. et DILLON W.R. (1978) : Discrete discriminant analysis, John Wiley and Sons, New York.

LEBART L., MORINEAU A. et TABARD N. (1977) : Techniques de la description statistique - Méthodes et logiciels pour l'analyse des grands tableaux, Dunod, Paris.

MORGAN J.N. et SONQUIST J.A. (1963) : "Problems in the analysis of survey data, and a proposal", JASA, Vol. 58, n° 302.

NAKACHE J.P. (1980) : "Méthodes de discrimination sur variables de nature quelconque - Théorie et pratique", Thèse d'Etat, Université Paris VI.

RAO C.R. (1973) : Linear statistical inference and its applications, John Wiley and Sons, New York, 2nd edition.

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

SAPORTA G. (1977) : "Une méthode et un programme d'analyse discriminante pas à pas sur variables qualitatives", Colloque IRIA : Analyse des données et informatique, Vol. 1, pp. 201-210.

TENENHAUS M. et BOUROCHE J.M. (1970) : "Quelques méthodes de segmentation", RIRO, Vol. 5, n° 2, pp. 29-42.

ULMO J. (1973) : "Différents aspects de l'analyse discriminante", Rev. Stat. Appl., Vol. 21, n° 2, pp. 17-55.