Les Cahiers de l'Analyse des Données
Vol. XXI - 1996 - n°1 - pp.75-83.

DIVERSITÉ DES CODAGES PERMIS EN ANALYSE DISCRIMINANTE: EXEMPLE DE DONNÉES CYTOLOGIQUES

[CODAGE DISCRI.]

*HASSAN HAMOUD Anwar**

0 Choix et préparation des données

Dans l'article [DONNÉES RÉSEAUX], publié dans ce même cahier, F. MURTAGH entreprend d'appliquer l'analyse statistique multidimensionnelle à des jeux de données d'un recueil compilé, par Lutz PRECHELT, sous le titre de [PROBEN1], afin de mettre à l'épreuve les algorithmes de réseaux dans leurs diverses variantes. Dans le présent article, nous reprenons l'un des deux exemples traités par F. MURTAGH, afin de montrer que le succès de la discrimination est compatible avec une certaine diversité dans le codage des variables.

Le format des données est rigoureusement homogène. Sur plusieurs centaines de prélèvements de tumeur mammaire, ont été relevées 9 variables cytologiques, notées chacune de 1 (normalité) à 10 (malignité maxima). La liste des variables est donnée dans l'article cité. On dispose dans chaque cas d'un diagnostic Δ: suivant les deux modalités, Δs (sain) et Δm (malignité). Les données sont complètes pour 683 cas (444s + 239m).

Dans [DONNÉES RÉSEAUX], F. MURTAGH adopte le codage le plus direct possible: chacune des 9 variables est éclatée suivant 10 modalités qui ne sont autres que les notes utilisées, de 1 à 10: la discrimination obtenue dans cette voie est très satisfaisante. Dans le présent article, nous montrons qu'on peut, avec succès, appliquer un codage beaucoup plus réduit. Premièrement, chaque variable Vx est codée barycentriquement suivant deux modalités, Vx0 et Vx+: la note 1, normalité, est codée {1, 0} (i.e. 1 sur Vx0, et 0 sur Vx+); au contraire, la note 10, risque maximum, est notée {0, 1}; entre les deux, le codage varie linéairement: e.g., pour la note 8, {2/9, 7/9}.

Certes, comme le note F. MURTAGH lui-même, on peut faire une excellente discrimination d'après le simple total des 9 variables de base: ce qui ne laisse pas de surprendre, alors que le jeu de données a fait l'objet de publications qui semblent postuler la difficulté du problème pratique. Mais nous poursuivons l'étude de ces données afin de montrer que, dans d'autres cas offrant une véritable difficulté, on pourra recourir à de multiples codages.

(*) Docteur de l'Université Pierre et Marie CURIE.

Les cahiers de l'analyse des données 0339-3097/96/01/$ 4.00.

Ainsi, pour 683 cas, on obtient, par un tel codage barycentrique, un tableau 683 × 18, dont l'ensemble des colonnes est (en reprenant les sigles adoptés dans l'article cité):

{éπ0, éπ+, ut0, ut+, uf0, uf+, ad0, ad+, πt0, πt+, nk0, nk+, nx0, nx+, nn0, nn+, mt0, mt+}.

Dans ce tableau, les 9 modalités Vx0, dont chacune n'apporte pas de signification particulière mais seulement une présomption générale de normalité, sont cumulées en une seule, notée tt0; les variables Vx+ étant conservées comme propres à caractériser la tumeur. D'où un ensemble de 10 colonnes principales:

{tt0, éπ+, ut+, uf+, ad+, πt+, nk+, nx+, nn+, mt+} ;

le principe de ce codage est expliqué dans T. K. GOPALAN & F. MURTAGH, (*op. laud.*).

Les 683 cas sont répartis en 4 blocs {sb:297cas, mb:159cas, sé:147cas, mé:80cas}, selon qu'il s'agit de sujet sain ou de tumeur maligne (s ou m); et de l'échantillon de base ou de l'échantillon d'épreuve (b ou é). À chacun des blocs est associé une ligne de cumul, ajoutée au tableau 683 × 10; d'où un tableau 687 × 10. Ont été ajoutées deux colonnes de diagnostic, en (0,1): {Δs, Δm}.

1 Première analyse: les individus de l'échantillon de base sont en principal

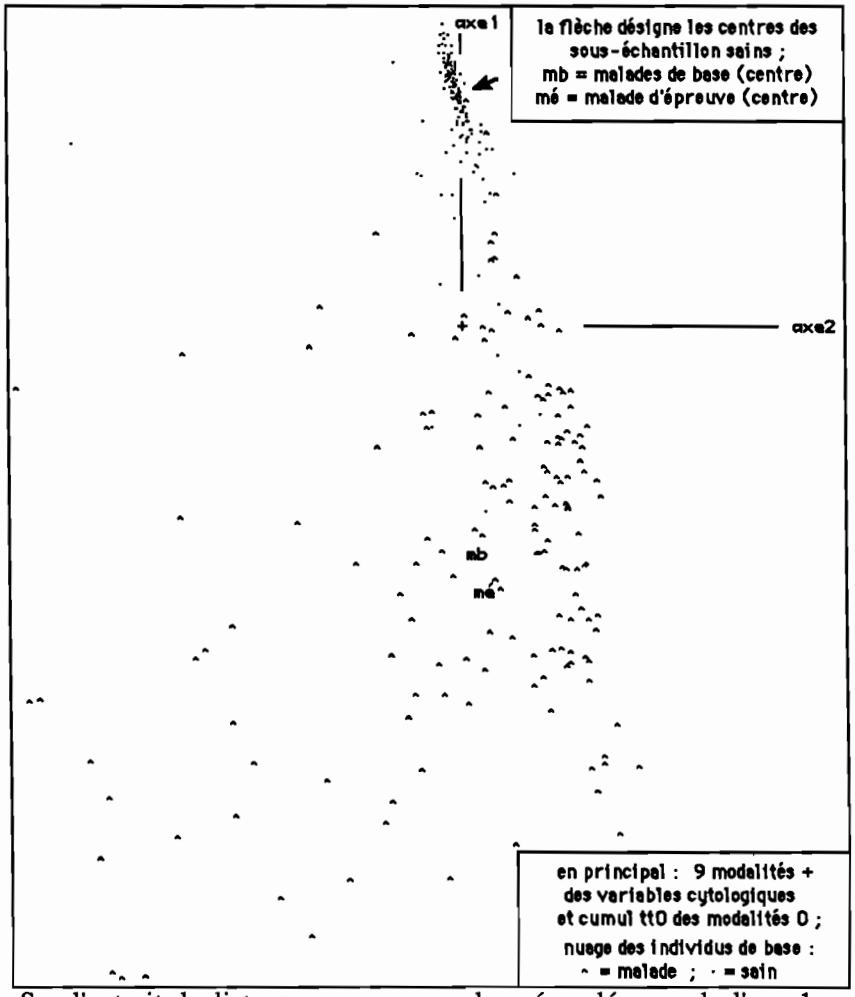
Wisconsin cancer

trace :	3.830e-1								
rang :	1	2	3	4	5	6	7	8	9	
lambda :	3029	234	142	130	89	77	57	51	22	e-4
taux :	7908	612	372	338	232	201	148	133	56	e-4
cumul :	7908	8520	8892	9230	9462	9663	9811	9944	10000	e-4

SIGI	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR	F 3	CO2	CTR	F 4	CO2	CTR
mb	1000	259	163	-491	998	205	16	1	3	9	0	2	-5	0	1
ci-dessous élément(s) supplémentaire(s)															
sb	1000	483	348	525	998	440	-17	1	6	-10	0	3	6	0	1
sé	1000	239	163	511	998	206	-20	2	4	0	0	0	-6	0	1
mé	1000	130	113	-572	981	140	30	3	5	-67	13	41	12	0	1

SIGJ	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR	F 3	CO2	CTR	F 4	CO2	CTR
tt0	1000	690	238	364	999	301	-6	0	1	-3	0	0	6	0	2
ci-dessous élément(s) supplémentaire(s)															
Δs	877	54	150	954	851	161	-112	12	29	-84	7	27	51	2	11
Δm	877	57	140	-891	851	151	104	12	27	79	7	25	-48	2	10

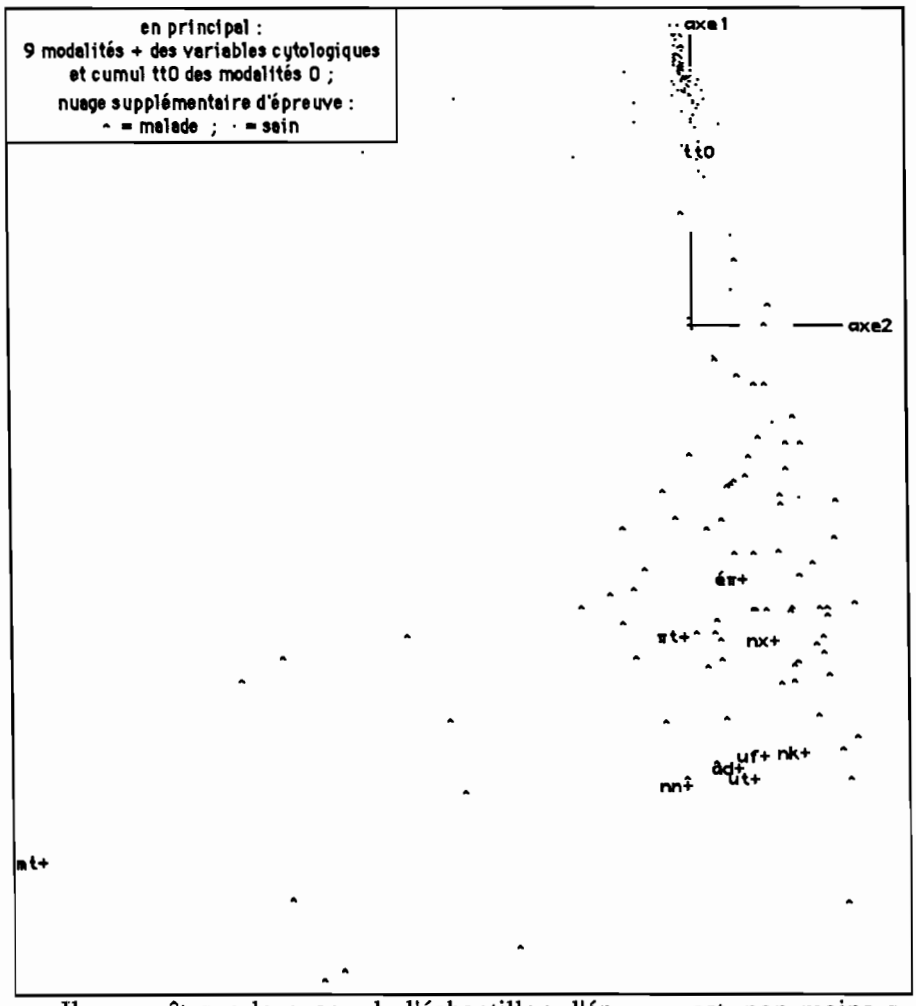
En lignes principales sont les 456 cas individuels de l'échantillon de base (297s+159m); ainsi que la ligne de cumul mb; celle-ci étant ajoutée afin que s et m s'équivalent en poids (on pourrait aussi bien multiplier par 2 les lignes des malades: pour un tel jeu de coefficients, cf. F. MURTAGH, *op. laud.*, §2.1).



Sur l'extrait du listage, on remarque la prépondérance de l'axe 1, avec 80% de l'inertie. À cet axe, les lignes de cumul des 4 blocs de cas sont quasi parfaitement corrélées; ainsi que la variable, tt0, de cumul des 9 modalités de normalité, Vx0.

Dans le plan (1, 2), on trouve une excellente discrimination des deux fractions, saine et maligne, de l'échantillon de base; mais, à la différence de celle-là, celle-ci se disperse notablement suivant la direction de l'axe 2.

Quant aux modalité, Vx+, elles sont très inégalement corrélées à l'axe1: la corrélation étant maxima pour ut+ (889‰); et minima pour mt+ (392‰).



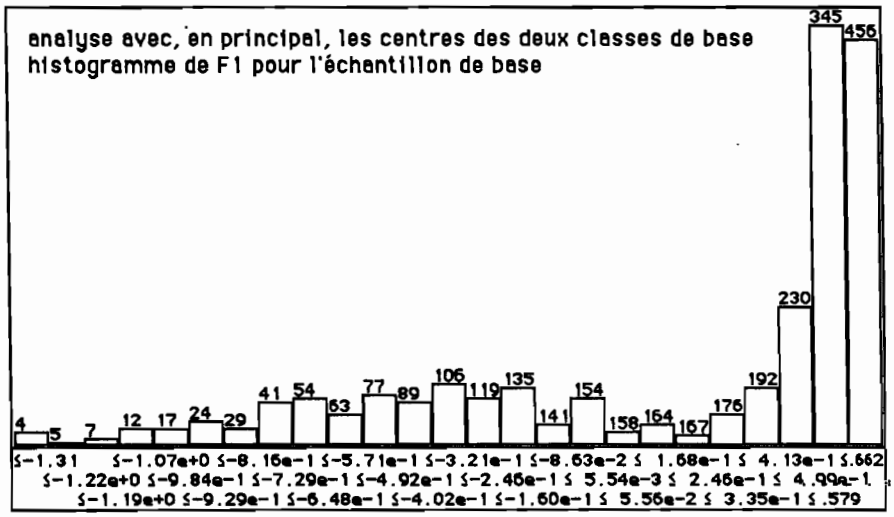
Il apparaît que le nuage de l'échantillon d'épreuve est, non moins que celui de l'échantillon de base, séparé en deux blocs assez bien distincts. Toutefois, comme dans [DONNÉES RÉSEAUX] §1, la discrimination sera faite d'après une deuxième analyse où figurent seules en principal des lignes de cumul afférentes à l'échantillon de base.

2 Deuxième analyse: seuls sont en principal les centres de classe d'individus de l'échantillon de base

Les lignes principales sont sb et mb; celle-ci affectée d'un coefficient 2, afin que soient quasi égaux les poids des deux diagnostics.

```
Wisconsin cancer
trace  :   2.581e-1
rang   :      1
lambda :   2581   e-4
taux   :  10000   e-4
cumul  :  10000   e-4
+----+-----------+-----------+
|SIGI| QLT PDS INR|  F 1 CO2 CTR|
+----+-----------+-----------+
ci-dessous éléments supplémentaires
|  sé| 999 239 242|  511 999 242|
|  mé| 977 130 168| -570 977 164|
+----+-----------+-----------+
```

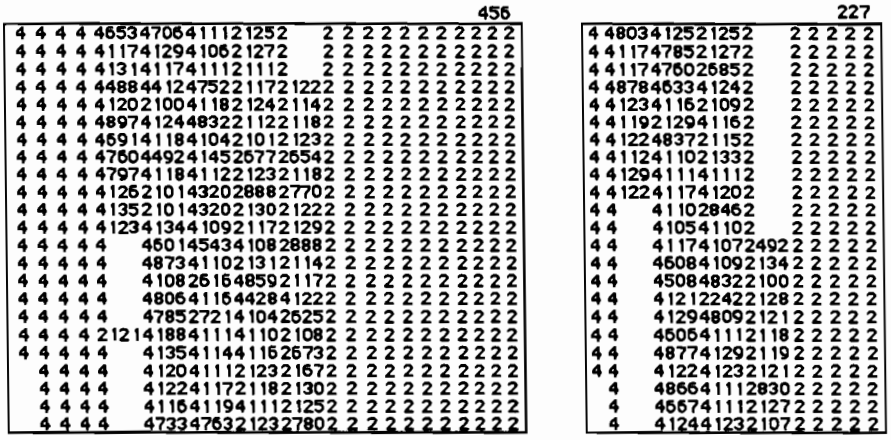
Avec un tel tableau principal, il n'y a qu'un seul facteur non trivial, l'unique valeur propre est les 85% de la 1–ère du §1. Dans l'espace des profils sur J (cardJ = 10, comme au §1), les profils, mis en supplément, des centres, sé, mé, (des deux blocs de l'échantillon d'épreuve) ont, sur l'axe unique, une qualité de représentation parfaite: QLT(sé) = 999‰; QLT(mé) = 977‰.



Comme dans toutes les analyses effectuées sur nos données cytologiques, au §1 comme dans [DONNÉES RÉSEAUX], l'histogramme de l'ensemble de base, comprenant les spécimens sains et les cas de malignité, apparaît bimodal; ces derniers cas étant, ici, vers (F1<0).

Afin de placer au mieux un seuil, il convient d'observer la distribution exacte des cas; ce qu'on fera, comme F. MURTAGH, en considérant, d'une part, l'ordre de tous les cas sur l'axe 1, d'après un histogramme général à un seul créneau; et, d'autre part, les abscisses précises des individus de transition, d'après un histogramme, en plusieurs créneaux, restreint à la partie centrale de l'échantillon de sorte qu'y puissent être inscrits les sigles de tous les individus.

Dans l'histogramme de la partie centrale, on a les sigles de 4 chiffres; dont le 1-er donne le diagnostic, 4=m, 2=s; et les autres font un numéro de série qui peut être commun à plusieurs cas.

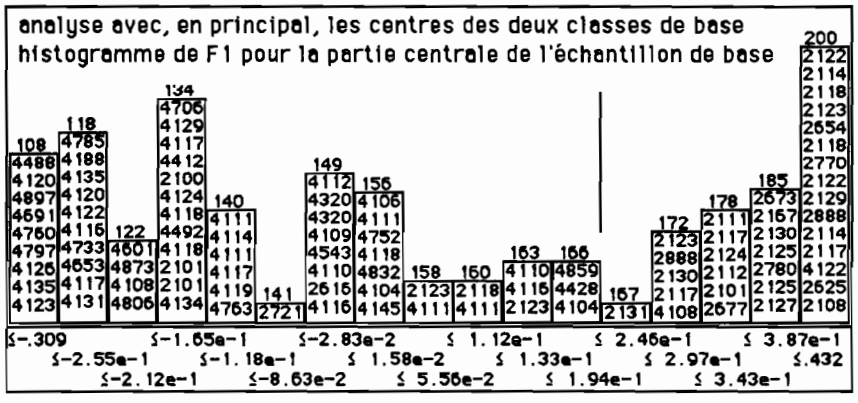


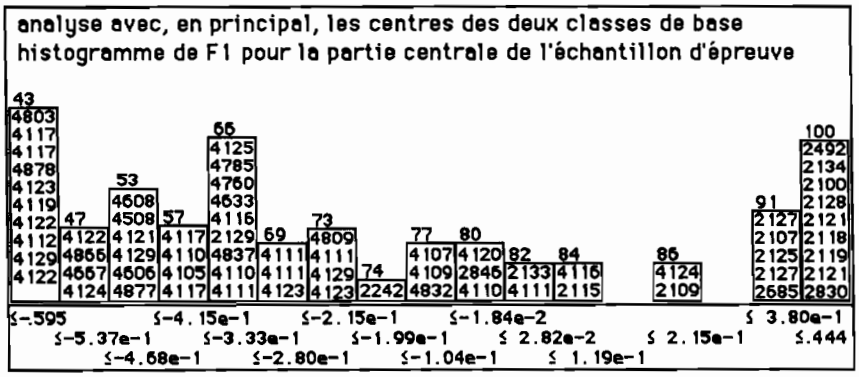
tableaux des sigles des individus rangés suivant les valeurs de F1; avec le maximum en haut à droite; et, dans chaque colonne, décroissance de haut en bas

à gauche, tableau de l'échantillon de base à droite, échantillon d'épreuve

Dans un tableau à un seul créneau, les sigles complets ne sont conservés que pour les cas rentrant dans l'histogramme central, ou pour les cas exceptionnels (s=2 isolé au sein de m=4; ou réciproquement); les autres sigles sont réduits au chiffre initial.

Pour l'échantillon de base, il y a, en dehors de l'histogramme central, des cas de malignité à gauche et des cas sains à droite; un seul individu, marqué 2121, faisant exception à gauche. Le seuil de malignité est donc certainement à chercher sur l'histogramme central. En prenant, la valeur de (0,194), suggérée, sur le graphique ci-dessous, par un trait vertical avant le créneau réduit à l'élément '2131', de rang 167, on a, au total, sur l'échantillon de base, 11 erreurs; dont seulement deux cas de malignité déclarés bénins.





En prenant ce seuil pour l'échantillon d'épreuve, on prononce le diagnostic de malignité pour 5 cas de bénignité; et un seul cas de malignité est donné pour bénin; un autre choix, plus prudent, du seuil de malignité aurait permis d'éviter cette erreur (sigle 4124) en déclarant malin un cas bénin (dont le sigle est 2109). Toutes ces erreurs se voient sur l'histogramme de la partie centrale de l'échantillon d'épreuve.

Quant au taux de succès, ces résultats ne diffèrent aucunement de ceux obtenus au §1.2 de [DONNÉES RÉSEAUX], avec un tout autre codage: soit, sur l'échantillon de base, 9 erreurs (s $\rightarrow$ m) et une erreur (m $\rightarrow$ s); et, sur l'échantillon d'épreuve, 5 erreurs (s $\rightarrow$ m) et une erreur (m $\rightarrow$ s). Plus précisément, il apparaît que, excepté au voisinage immédiat du seuil, dont le choix ne s'impose pas rigoureusement, les erreurs portent sur les mêmes individus avec les deux codages.

Enfin, pour les 16 cas de données incomplètes (14s + 2m) il n'y a qu'une seule affectation erronée, un cas bénin étant considéré comme malin; résultat identique à celui obtenu dans [DONNÉES RÉSEAUX].

Le codage adopté dans le présent article a le mérite d'être compact (10 colonnes; au lieu de 90 pour la forme disjonctive complète sans cumul de modalités voisines); ce qui, comme on l'a noté, serait un avantage pour des applications pratiques ultérieures.

Les données de la présente étude, quant à elles, s'accomodent, on le sait, du codage le plus radical: la somme des 9 variables de base, ou, ce qui revient au même, notre seule variable tt0, suffisant à décider de la normalité ou de la malignité d'une tumeur; tout aussi bien que le fait l'analyse discriminante et mieux que l'approximation stochastique.

En général, en analyse factorielle, la valeur d'un facteur se calcule aisément, pour tout nouvel individu, par combinaison linéaire des variables

(après leur codage éventuel). De ce point de vue on peut regarder l'analyse factorielle comme une méthode qui, d'une part, fournit les coefficients de ces notes globales que les cliniciens, particulièrement les psychiatres, utilisent systématiquement aujourd'hui; et, d'autre part, invite à distinguer plusieurs dimensions au sein d'un domaine de la pathologie.

En particulier, les données considérées ici prendraient leur véritable intérêt si, au lieu du seul diagnostic de malignité, on cherchait à déterminer le type du cancer; type qui suggère une thérapeutique et dont dépend le pronostic. Et, sans faire montre indûment de compétence en cancérologie, nous dirons que la variable mt+, "mitoses", qui, au §1, s'écarte le plus sur l'axe 1 et crée essentiellement l'axe 2, est un indice manifeste de l'activité d'une tumeur. Cette interprétation étant à mettre au crédit de l'analyse du §1, fondée sur un codage compact.

Références bibliographiques relatives à l'Analyse des Données

T. K. GOPALAN & F. MURTAGH : "The Role of Input Data Coding in Multivariate Data Analysis: The Example of Correspondance Analysis"; à paraître;

F. MURTAGH : "Application de l'analyse factorielle et de l'analyse discriminante à des données colligées pour être soumises à des réseaux de cellules"; [DONNÉES RÉSEAUX], ce même cahier;

On trouvera dans [DONNÉES RÉSEAUX] des références complètes sur l'histoire des données traitées dans le présent article.