Les Cahiers de l'Analyse des Données
Vol. XXI - 1996 - n°1 - pp.53-74.

APPLICATION DE L'ANALYSE FACTORIELLE ET DE L'ANALYSE DISCRIMINANTE À DES DONNÉES COLLIGÉES POUR ÊTRE SOUMISES À DES RÉSEAUX DE CELLULES

[DONNÉES RÉSEAUX]

*F. MURTAGH**

0 Origine des données

Il y a quarante ans déjà que, pour servir à la reconnaissance des formes, ROSENBLATT proposait, sous le nom de *Perceptron*, une ingénieuse machine. En bref, le Perceptron devait recevoir, sur une couche de cellules sensibles, ou *rétine d'entrée*, des images simples, appartenant à plusieurs classes; et, en modifiant, suivant un programme, ses connexions internes, au cours d'un apprentissage où seraient présentées des images avec l'indication de leur classe, parvenir à fournir, dans la suite, sur un tableau de sortie, les réponses exactes afférentes à des images anonymes.

Il nous suffira de dire que les réseaux neuronaux, dont on prône aujourd'hui l'application, sont les héritiers du perceptron; dont ils conservent le principe, avec une structure plus complexe, compatible avec nos moyens de calculs, incomparablement plus puissants que ceux que considérait ROSENBLATT.

Du point de vue de l'analyse des données, l'apprentissage sur ces réseaux relève de l'approximation stochastique; sur laquelle on a pu fonder un algorithme d'analyse factorielle requérant très peu d'espace en mémoire centrale, et dont le jeu offre, de la synthèse mentale, une analogie suggestive; mais qui ne peut rivaliser présentement avec les algorithmes usuels de diagonalisation de matrice.

Il était donc souhaitable de comparer approximation sur réseaux et analyse multidimensionnelle quant au volume des calculs et à la précision des classements ou discriminations obtenus.

Récemment, a été compilé, par Lutz PRECHELT, sous le titre de [PROBEN1], un recueil de jeux de données destinés à mettre l'épreuve les algorithmes de réseaux dans leurs diverses variantes. Ayant accès à ce recueil, nous voulons, par le présent article, ouvrir la comparaison entre réseaux et

(*) Informaticien à l'observatoire de l'ESO;
Karl-Schwarzschild-Straße ; D-85748 Garching bei München.

analyse multidimensionnelle explicite. Quant au temps de calcul, il ne fait aucun doute que tout l'avantage soit à celle-ci. Reste donc l'exactitude des discriminations.

Dans un domaine où l'automatisme n'est que partiel, il faudrait aussi comparer la compétence requise de l'opérateur humain.

Certains vantent l'universalité des réseaux et prétendraient s'en servir sans égard au contenu des données. La plupart des articles de *CAD* offrent, au contraire, une introduction relevant, selon les cas, de la médecine, des belles letres ou de l'économie. Dans le présent article il s'agira de deux jeux de données; choisis le premier pour sa facilité manifeste, le second pour sa difficulté apparente; mais dont ni l'un l'autre n'ajoute aux chiffres d'utile commentaire. Le débat philosophique sur le format et le contenu sera donc différé.

1 Classification des tumeurs mammaires quant à leur bénignité ou leur malignité

1.1 Format des données et analyse préliminaire

Sur plusieurs centaines de prélèvements de tumeur, ont été relevées 9 variables cytologiques, codées chacune suivant 10 modalités de 1 (normalité) à 10 (malignité maxima). On dispose dans chaque cas d'un diagnostic: suivant les deux modalités, Δs (sain) et Δm (malignité). Les données sont complètes pour 683 cas (444s + 239m); la variable nk manque pour 16 cas (14s + 2m).

```
découpage en 10 des variables : 444s + 239m
Δs = sain (bénignité) ; Δm = malignité
éπ épaisseur ; ut uniformité des cellules en taille ;
uf id en forme ; ad adhérence marginale ; πt taille
d'une cellule épithéliale ; nk noyeaux isolés ;
nx chromatine ; nn normalité des nucléoles ; mt mitoses;
en colonnes : dix niveaux des variables cytologiques
18010  01   02   03   04   05   06   07   08   09   10
éπΔs  136   46   92   67   83   15    1    4    0    0
éπΔm    3    4   12   12   45   18   22   40   14   69
utΔs  369   37   27    8    0    0    1    1    1    0
utΔm    4    8   25   30   30   25   18   27    5   67
ufΔs  344   51   30   12    2    2    2    1    0    0
ufΔm    2    7   23   31   30   27   28   26    7   58
adΔs  363   37   31    5    4    3    0    0    0    1
adΔm   30   21   27   28   19   18   13   25    4   54
πtΔs   43  355   28    7    5    1    2    2    0    1
πtΔm    1   21   43   41   34   39    9   19    2   30
nkΔs  387   21   14    6   10    0    1    2    0    3
nkΔm   15    9   14   13   20    4    7   19    9  129
nxΔs  148  153  125    7    4    1    6    0    0    0
nxΔm    2    7   36   32   30    8   65   28   11   20
nnΔs  391   30   11    1    2    4    2    3    0    0
nnΔm   41    6   31   17   17   18   14   20   15   60
mtΔs  431    8    2    0    1    0    1    1    0    0
mtΔm  132   27   31   12    5    3    8    7    0   14
```

Un tableau de contingence (2×10) donne, d'après les 683 cas, le nombre d'échantillons sains et pathologiques associés aux 10 niveaux de

chacune des 9 variables. Par exemple, on lit, à l'intersection de la ligne utΔm et de la colonne 04 que la variable ut a été relevée au niveau de gravité 4 dans 30 cas de tumeur maligne.

Il suffit d'accorder quelque attention à ce tableau pour apprécier les données cytologiques relevées: il s'en faut de peu que chacune n'offre, à elle seule, la base d'un diagnostic certain. C'est pourquoi, on a cru opportun d'analyser, d'abord, les données telles quelles.

De façon précise, sans groupement de modalité ni codage barycentrique, chaque variable cytologique est découpée en 10 modalités, e.g., pour πt, {πt1, πt2,... , πt9, πtX}; et la variable Δ, diagnostic, en 2: {Δs, Δm}; soit 92 modalités. Le tableau de BURT est analysé avec, en principal, les seules modalités cytologiques; {Δs, Δm} étant en supplément comme lignes et colonnes: ce qui revient à tenter une reconnaissance de forme, sans prendre en compte le diagnostic. Il vaut la peine de noter que la ligne Δs du tableau de BURT, n'est autre que la suite des 9 lignes {éπΔs, ..., mtΔs} du tableau ci-dessus; et de même pour Δm avec les lignes {éπΔm, ..., mtΔm}. En fait, le tableau 18×10 a été construit à partir du tableau à deux colonnes {Δs, Δm}, extrait du BURT, au moyen d'un programme faisant passer un tableau ternaire I×J×T de la forme rectangulaire (I×J)×T à la forme rectangulaire (I×T)×J. Et les individus sont adjoints en supplément, comme des vecteurs en (0, 1).

découpage en 10 des variables cytologiques; 90×90; Δ en supplément

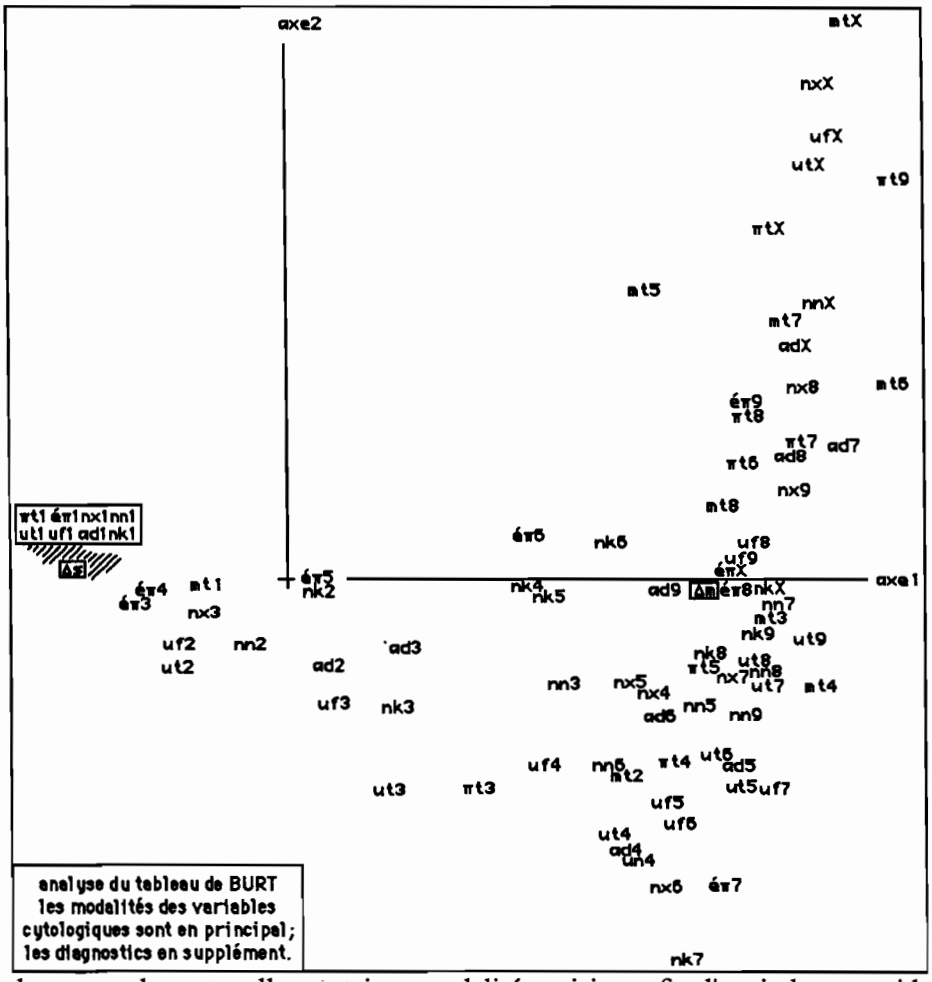
trace	:	1.624e+0									
rang	:	1	2	3	4	5	6	7	8	9	10	
lambda	:	5197	971	542	442	419	398	367	349	338	322	e-4
taux	:	3201	598	334	272	258	245	226	215	208	198	e-4
cumul	:	3201	3799	4132	4405	4663	4908	5134	5349	5557	5755	e-4

Bien que l'analyse fournisse 81 facteurs non triviaux, le 1-er, nettement séparé des suivants, rend compte de 32% de l'inertie.

	SIGJ	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR	F 3	CO2	CTR	F 4	CO2	CTR
ci-dessous élément(s) supplémentaire(s)																
	Δs	993	72	15	-577	992	46	20	1	0	-9	0	0	-1	0	0
	Δm	993	39	28	1071	992	86	-37	1	1	17	0	0	1	0	0

Fait plus remarquable encore, la qualité de représentation des modalités {Δs, Δm} est sur cet axe de 992‰: ce qui signifie, en termes géométriques, que les points {Δs, Δm}, centres de gravités respectifs des sous-nuages des individus sains et malades, s'opposent quasi rigoureusement suivant cet axe. En sorte qu'en se restreignant à l'axe 1, on a une analyse discriminante ordinaire, le F1 de chaque sujet n'étant guère que l'abscisse de sa projection sur l'axe joignant les centres des deux sous-populations s et m.

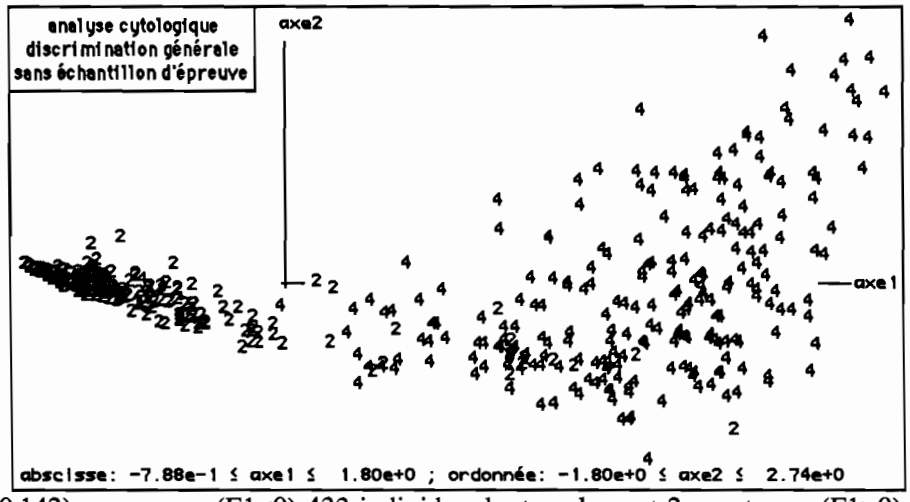
Quant à l'ensemble des modalités, le plan (1, 2) (reproduit avec même échelle sur les deux axes) présente une parabole d'effet GUTTMAN; sur laquelle les modalités de chacune des variables dessinent, approximativement, un chapelet ordonné. Il faut ici prendre garde que certaines modalités sont très légères (peu représentées dans notre corpus); la modalité mt9 manque même totalement. Pour l'application au diagnostic sur des cas nouveaux, il faudrait



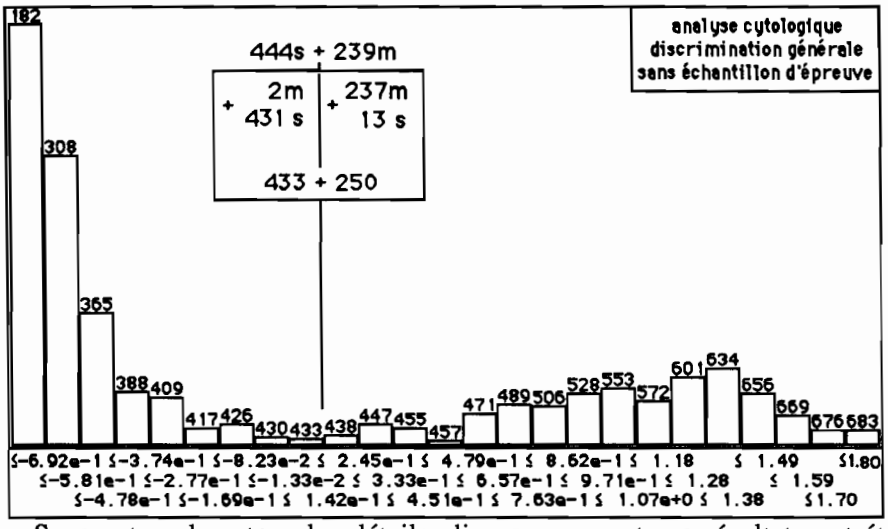
donc cumuler entre elles certaines modalités voisines afin d'atteindre un poids assurant un profil bien défini.

Pour l'ensemble I des individus, le plan (1, 2) est publié en restreignant l'échelle de l'axe 2. Chaque cas est figuré par l'un des chiffres {2, 4}, lesquels, dans le codage numérique des données originelles, servent pour {s, m}. Ici encore, il y a un net effet GUTTMAN; les chiffres 2 s'accumulent vers (F1<0); les chiffres 4 se dispersent dans le quadrant (F1>0; F2>0). Entre les deux diagnostics, la séparation est presque parfaite.

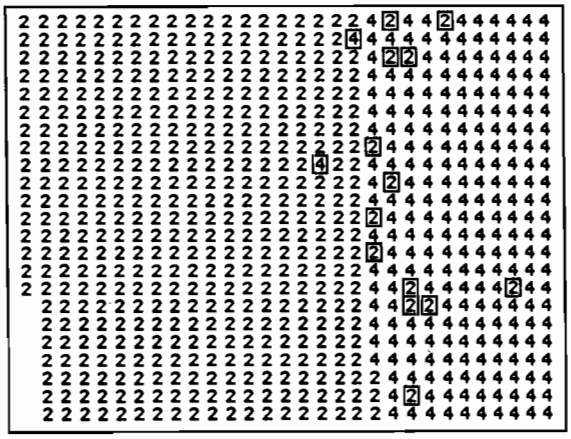
Pour plus de précision, considérons l'histogramme de F1 pour l'ensemble des cas. On y distingue deux modes: si l'on prend pour valeur frontière



(0,142), on a, vers (F1<0) 433 individus dont seulement 2 m; et vers (F1>0), 250 individus, dont 13 s. Discrimination obtenue sur un axe issu d'une analyse où l'opposition {s, m} n'est pas prise en compte explicitement.



Sans entrer dans tous les détails, disons comment ces résultats ont été obtenus. Un programme permet de construire, pour une colonne j d'un tableau I×J considérée comme fonction sur l'ensemble I, des histogrammes couvrant un sous-intervalle choisi (e.g.: du rang 50 au rang 250), avec un nombre spécifié de créneaux; où s'inscrivent, si ceux-ci sont assez larges, les sigles des individus i (plus précisément, les sigles utilisés ici ne sont qu'un numéro de



série, précédé du chiffre du diagnostic; 2 ou 4). En particulier, en demandant un histogramme à un seul créneau, on obtient une liste des sigles des individus rangés suivant les valeurs de j; avec le maximum en haut à droite; et, dans chaque colonne, décroissance de haut en bas.

Une telle liste est publiée ici.

Les sigles sont réduits à leur premier caractère: 2 ou 4, pour s ou m. On peut distiguer un bloc des 2 et un bloc des 4; les exceptions étant encadrées.

1.2 Analyse discriminante avec échantillon d'épreuve

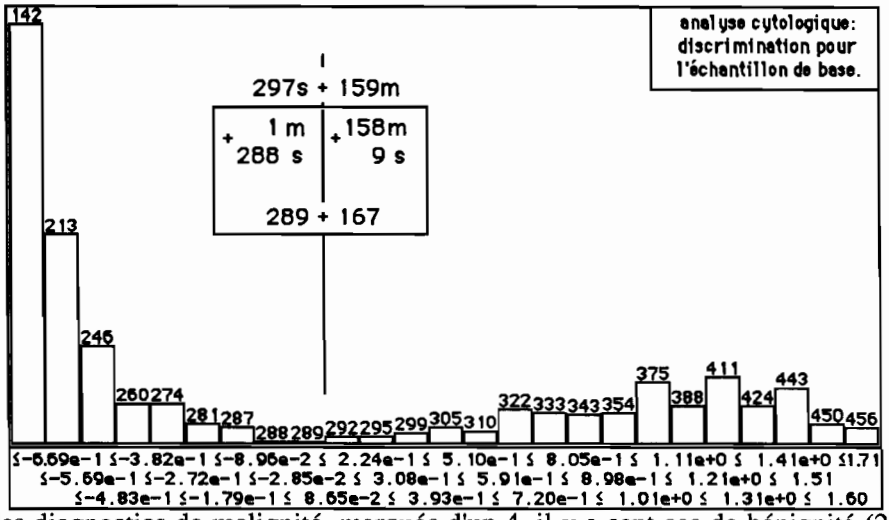
Ayant ainsi constaté que les données offraient matière à un excellent diagnostic, on en a repris l'étude en suivant plus strictement l'ordre de l'analyse discriminante. L'ensemble I des 683 sujets (pour lesquels les données sont complètes) a été divisé en trois tiers, I1, I2, I3, d'après le reste mod 3 du numéro d'ordre dans le fichier de base; I1-I2 devant constituer l'échantillon de base, Ib; et I3 l'échantillon d'épreuve. Un tableau de BURT a été construit suivant le même découpage des variables que précédemment; mais en se restreignant au seul ensemble Ib. Dans l'analyse figure en principal un tableau 90×2, croisant l'ensemble des modalités des 9 variables cytologiques avec les deux diagnostics {Δs, Δm}. À l'analyse, ont été adjoints, comme des colonnes supplémentaires (en 0, 1), les deux ensemble Ib et I3; ainsi que l'ensemble Is des 16 individus pour lesquels manque la variable nk (les 10 modalités de celle-ci étant, pour ces individus, affectées de la valeur 0).

```
90 modalités cytologiques × {Δs, Δm}
trace  :   6.206e-1
rang   :     1
lambda :   6206 e-4
```

Ici, avec deux colonnes principales seulement, il n'y a qu'un seul facteur; l'unique valeur propre dépasse de peu la première de celles trouvées dans l'analyse précédente, ce qui confirme l'interprétation de celle-ci.

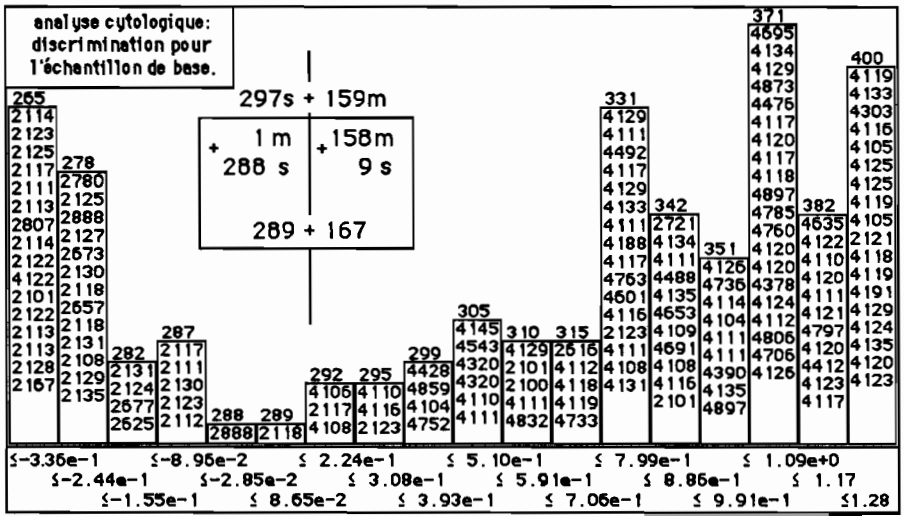
Quant aux cas, considérons d'abord l'ensemble Ib de base.

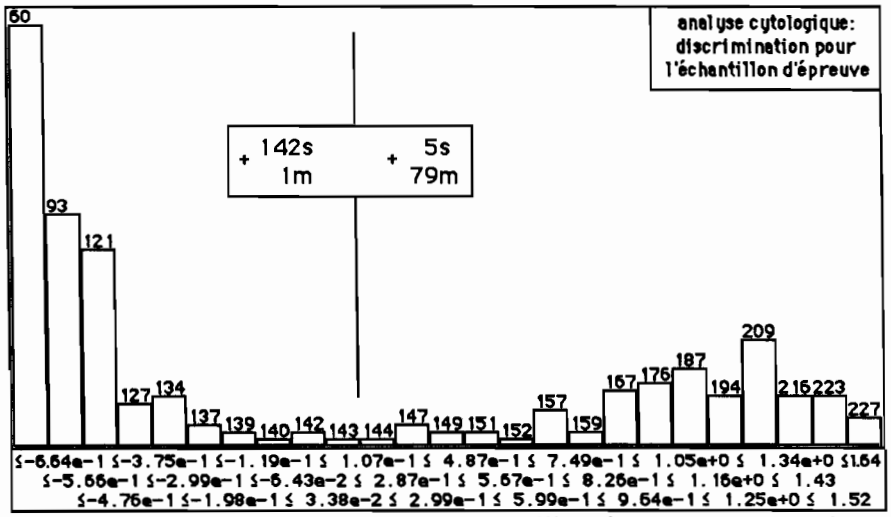
Comme précédemment, on a, pour F1, un histogramme bimodal. Sur la liste ordonnée des cas, on voit une séparation quasi parfaite. Au sein du bloc



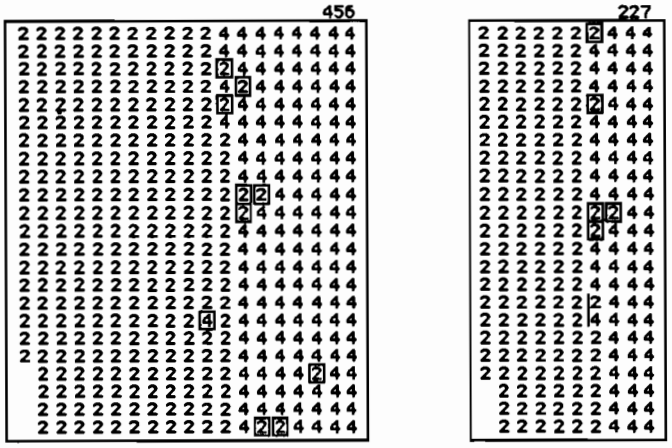
des diagnostics de malignité, marqués d'un 4, il y a sept cas de bénignité (2); un seul cas de malignité se trouve étiqueté comme bénin. Au contact entre les deux blocs, 4 cas sont imbriqués (dont les numéros se lisent sur l'histogramme détaillé du centre de la distribution); si, afin d'éviter de rassurer indûment une patiente, on place le seuil le plus à gauche possible (i.e. vers le bloc des 2), le bilan des 456 cas est: d'une part, (288s + 1m); d'autre part, (9s + 158m).

De façon précise, d'après les abscisses des cas individuels, le seuil peut être placé entre 0,087 et 0,152; la moyenne étant 0,119.

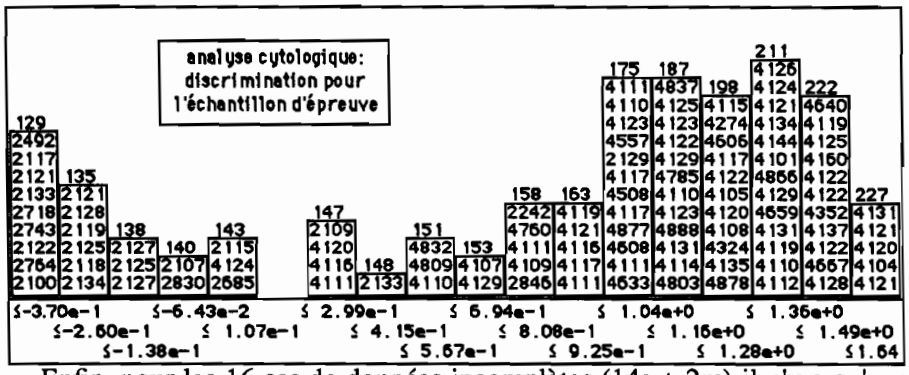




Pour l'ensemble I3, retenu comme échantillon d'épreuve, le bloc des cas de cancer renferme seulement 5 cas bénins; aucun cancer ne se place à l'intérieur du bloc de la bénignité; mais, à la frontière, deux cas (marqués 2115 et 4124 sur l'histogramme du centre de la distribution), l'un bénin, l'autre malin, sont douteux. Avec seuil à (0,119), seul le cas m est indûment étiqueté bénin; si pour sûreté maxima, on adopte le seuil (0,087), le cas s est étiqueté m. Il est clair que, du point de vue de la pratique médicale, l'intervalle central, qui s'étend entre les deux modes de l'histogramme de Ib, doit être considéré comme douteux.



Histogrammes à un seul créneau montrant l'imbrication des diagnostics :
à gauche: échantillon de base à droite: échantillon d'épreuve



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.

En somme, la discrimination est quasi parfaite; et nullement inférieure sur l'échantillon d'épreuve, I3, à ce qu'elle est pour l'échantillon de base, Ib. Mais, *a posteriori*, nous avons constaté qu'une discrimination non moins bonne s'obtient d'après le simple total des 9 notes initiales (cf. §3, conclusion).

2 Appréciation de l'activité thyroïdienne d'après 21 variables

2.1 Les données disponibles: principe du codage

Pour quelque 7000 patients, examinés dans un centre clinique, on dispose d'un diagnostic, K, en trois modalités:

{k1:subnormalité 1; k2:hyperactivité 2; k3:normalité 3};

avec, pour chaque cas, un ensemble de 21 variables, dont 15 sont d'un format logique simple, en (0, 1); et six sont des variables numériques continues. À ceci près, les documents directement accessibles ne disent rien de la signification des données.

Nous basant sur leur présentation nous avons distingué, une première variable continue, V, des cinq autres, notées {W1, ..., W5}; les variables binaires étant désignées par "Q" suivi d'une des minuscules de "a" à "o", soit: {Qa, ..., Qo}. D'où, en somme, une suite de 21 sigles:

{V Qa Qb Qc Qd Qe Qf Qg Qh Qi Qj Qk Ql Qm Qn Qo W1 W2 W3 W4 W5}.

À la différence de l'étude cytologique des tumeurs mammaires, objet du §1, le dossier relatif à la fonction thyroïdienne ne semble pas offrir matière à des diagnostics faciles.

D'abord, les modaltés du diagnostic K, sont de poids très inégal:

k1 ≈ 2,3% ; k2 ≈ 5,1% ; k3 ≈ 92,6%.

Aussi, ayant noté qu'il suffit de répondre toujours "normal" (k3) pour tomber juste dans 92,6% des cas, un utilisateur du fichier, doutant de pouvoir

mieux faire avec un réseau de calcul à trois issues, ajoute-t-il ce commentaire laconique: *But don't think that the problem is boring...* Ne croyez pas que le problème est fâcheux!

Pour un statisticien accoutumé à l'analyse discriminante, il ne fait pas de doute qu'il faut, de quelque manière, réduire les modalités du diagnostic à avoir des poids du même ordre de grandeur.

Quant à nous, ayant (comme il sera précisé dans la suite, au §2.2) découpé en 49 modalités l'ensemble des 22 variables (diagnostic y compris) nous avons obtenu un tableau de BURT; dont, comme il est classique, on a analysé le rectangle croisant l'ensemble I des lignes afférentes aux 46 modalités descriptives, avec l'ensemble J des trois colonnes de diagnostic {k1, k2, k3}; mais en multipliant respectivement ces trois colonnes par {20, 10, 1}, ce qui leur donne des poids quasi égaux: {339, 342, 319}.

Notons, au passage que cette méthode de pondération pourrait s'étendre à l'approximation stochastique sur réseau, en convenant que selon que se présente un individu étiqueté, 1, 2 ou 3, les renforcements des liens doivent être multipliés par ces coefficients {20, 10, 1} (;ou plutôt, relativement aux valeurs usuelles, divisés par 1, 2, 20; ce, afin d'éviter que des renforcements trop brusques n'écartent, sans retour, de l'optimum cherché).

Restent les variables elles-mêmes. D'abord, pour presque toutes les variables logiques, Qx, la modalité Qx1 est de poids très faible relativement à Qx0 (avant ou après pondération). Ainsi en est-il de {Qf1, Qk1, Ql1, Qn1} associées à la normalité, k3; de {Qb1, Qg1}, opposées à k2; de {Qc1, Qd1, Qe1, Qo1}, opposées à k1. Cette particularité est d'autant plus fâcheuse qu'on ne peut y remédier par le codage: le format en {0, 1}, sur deux colonnes, s'imposant absolument. [Anticipant sur l'exposé du §2.4, nous pouvons annoncer que la suppression de ces 15 variables Qx, ne diminue guères le taux de discrimination exacte obtenu].

Quant aux variables continues, le choix s'offre entre découpage de l'intervalle par des bornes délimitant des modalités logiques codées en {0, 1}; et codage barycentrique, continu, relativement à des pivots. En tout cas, bornes ou pivots sont fixés au vu du croisement d'une partition fine (e.g., en 16 classes) de la variable, avec l'ensemble des trois diagnostics. Afin de montrer l'efficacité de la méthode d'analyse discriminante, sous sa forme la plus automatique, nous publions ici les résultats obtenus avec un codage non retouché, fixé avant toute analyse factorielle.

2.2 Méthode d'analyse discriminante: bilan des résultats

Dans le tableau de BURT usuel, associé à un codage binaire (en {0, 1}), on trouve, à l'intersection de la ligne et de la colonne affectées respectivement

aux modalités m et m', le nombre, kB(m, m') des individus, ou sujets s, rentrant, à la fois dans ces deux modalités; en particulier, kB(m, m) est le nombre des individus rentrant dans la modalité m. Dans la cas présent avec un codage barycentrique, chaque sujet, s, contribue avec à kB(m, m') par le produit des poids k(s, m) et k(s, m') avec lesquels ces modalités rentrent dans sa description. On peut encore dire que, dans le tableau de BURT généralisé, chaque ligne m est la somme des lignes décrivant les sujets s (par leurs pondérations sur l'ensemble des modalités), chaque ligne s étant affectée du poids k(s, m); et de même, *mutatis mutandis*, pour les colonnes.

Dans la présente étude, l'appartenance k(s, kd), d'un sujet à l'une des trois modalités d (1, 2 ou 3) du diagnostic, est notée par 0 ou 1 (sans valeur intermédiare pour des cas douteux). Une colonne kd, du tableau I×J, 46×3, analysé, est donc la somme (affectée du coefficient afférent à kd, 20, 10 ou 1, cf. *supra*) des colonnes (formées de nombres compris entre 0 et 1) par les quelles sont décrits les sujets compris dans le diagnostic d. À quoi l'on peut ajouter trois dernières lignes (que nous noterons K1, K2, K3; afin de les distinguer des colonnes kd); avec kB(Kd, kd') nul si d≠d'; et kB(Kd, kd) égal au nombres des sujets affectés du diagnostic d (toujours avec le coefficient 20, 10 ou 1).

L'analyse factorielle du tableau à 3 colonnes ne produit que deux facteurs non triviaux: l'ensemble des résultats en peut donc être présenté dans un plan. Le nuage {k1, k2, k3} a, globalement, sur chacun des axes, même dispersion que le nuage des 46 modalités descriptives; au contraire, avec leurs profils purs dont chacun n'a qu'une composante non nulle, les éléments supplémentaires {K1, K2, K3} définissent un triangle, à l'intérieur duquel se projette tout le reste (avec des coordonnées barycentriques données par le profil repondéré des modalités descriptives). En bref, dans le plan (1, 2), les points {k1, k2, k3} sont plus proches de l'origine que leurs homologues respectifs Kd, mais s'écartent dans la même direction. De façon précise, la figure du plan (1, 2) illustre le §2.3.

Du point de vue de l'analyse discriminante, on peut affecter, à tout individu , comme diagnostic d, soit le point kd le plus proche de sa projection plane; soit le point Kd. On conçoit facilement que, compte tenu de la disposition rayonnante des diagnostics, les deux affectations sont analogues: effectivement, pour les milliers de sujets de l'échantillon de base, comme pour l'échantillon d'épreuve, il y a moins de 0,5% des diagnostics qui diffèrent suivant la procédure.

De façon précise, les tableaux ci-après donnent les bilans des diagnostics; le diagnostic véritable (ou donné pour tel...) étant noté Dd et les affectations respectives, kd ou Kd.

	D1	D2	D3
k1	83	2	22
k2	9	184	71
k3	0	0	3377

	D1	D2	D3
K1	84	2	32
K2	8	184	66
K3	0	0	3372

	k1	k2	k3
K1	107	9	2
K2	0	255	3
K3	0	0	3372

ci-dessus, bilan des affectations pour l'échantillon de base

affectation aux col. pr. affectation aux lig. suppl. croisement des affectations

ci-dessous, bilan des affectations pour l'échantillon d'épreuve

	D1	D2	D3
k1	64	0	27
k2	9	171	90
k3	0	6	3061

	D1	D2	D3
K1	65	1	34
K2	8	170	91
K3	0	6	3053

	k1	k2	k3
K1	91	4	5
K2	0	266	3
K3	0	0	3059

Il ne sied pas d'apprécier par un seul pourcentage la qualité de la discrimination ainsi obtenue. Donner systématiquement un diagnostic de normalité, D3, peut assurer un taux élevé de succès si la population examinée est saine dans son ensemble (cf. *supra*); mais n'apprend rien quant à la valeur clinique des variables; laquelle fait l'objet ultime de l'analyse.

Que l'affectation soit faite aux {kd} (profils de colonnes principales) ou aux {Kd} (profils de lignes supplémentaires), on note d'abord, dans le cas présent, qu'avec l'algorithme de discrimination utilisé, aucun des sujets anormaux de l'échantillon de base ne reçoit le diagnostic de normalité D3; cette erreur ne survenant que 6 fois sur les 240 cas {D1, D2} de l'échantillon d'épreuve; soit un taux de 2,5%.

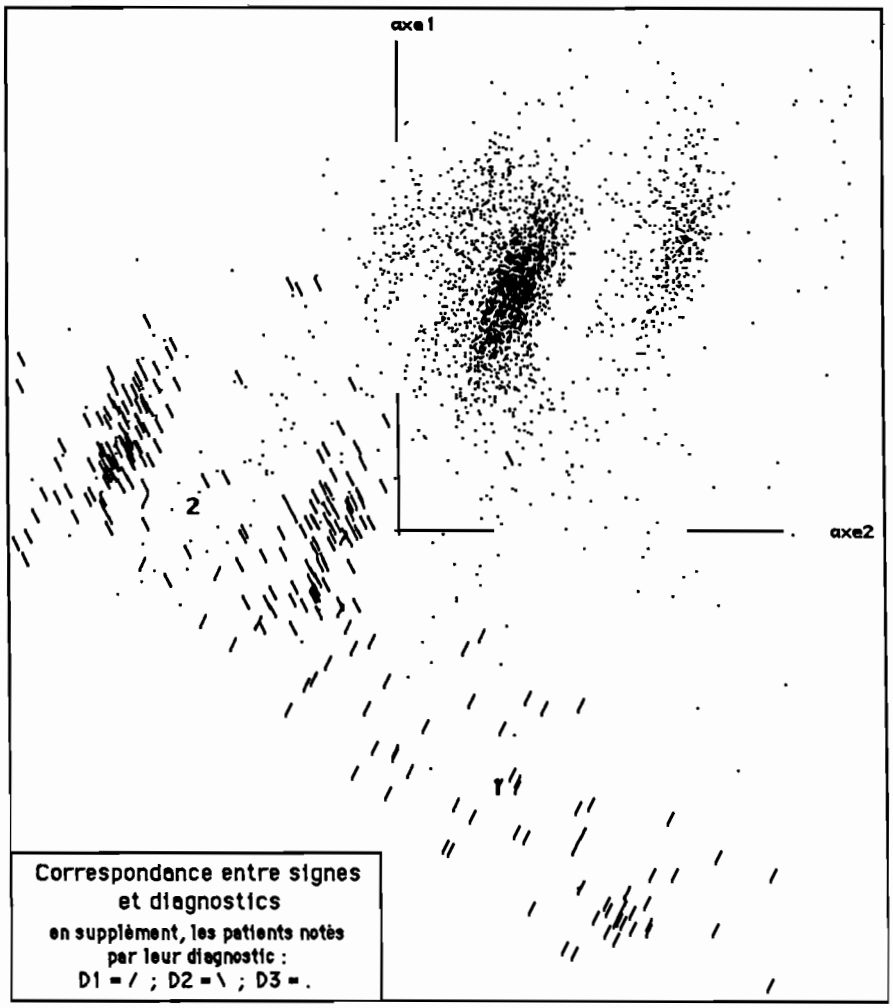
Entre les deux diagnostics d'anormalité, {D1, D2}, les confusions sont exceptionnelles de D2 vers k1 ou K1: 2 cas sur 186 dans l'échantillon de base; 1 ou 0 sur 177 dans l'échantillon d'épreuve. Le taux d'erreur de D1 vers D2 atteint, au maximum, 9/73 ≈ 12,33%: dans l'affectation de l'échantillon d'épreuve aux colonnes principales.

Quant aux sujets normaux, le taux maximum de diagnostic d'anormalité se trouve dans l'affectation de l'échantillon d'épreuve aux lignes suppléméntaires, {Kd}:

$$(91+34) / (91+34+3053) = 3{,}9\% \approx 4\% .$$

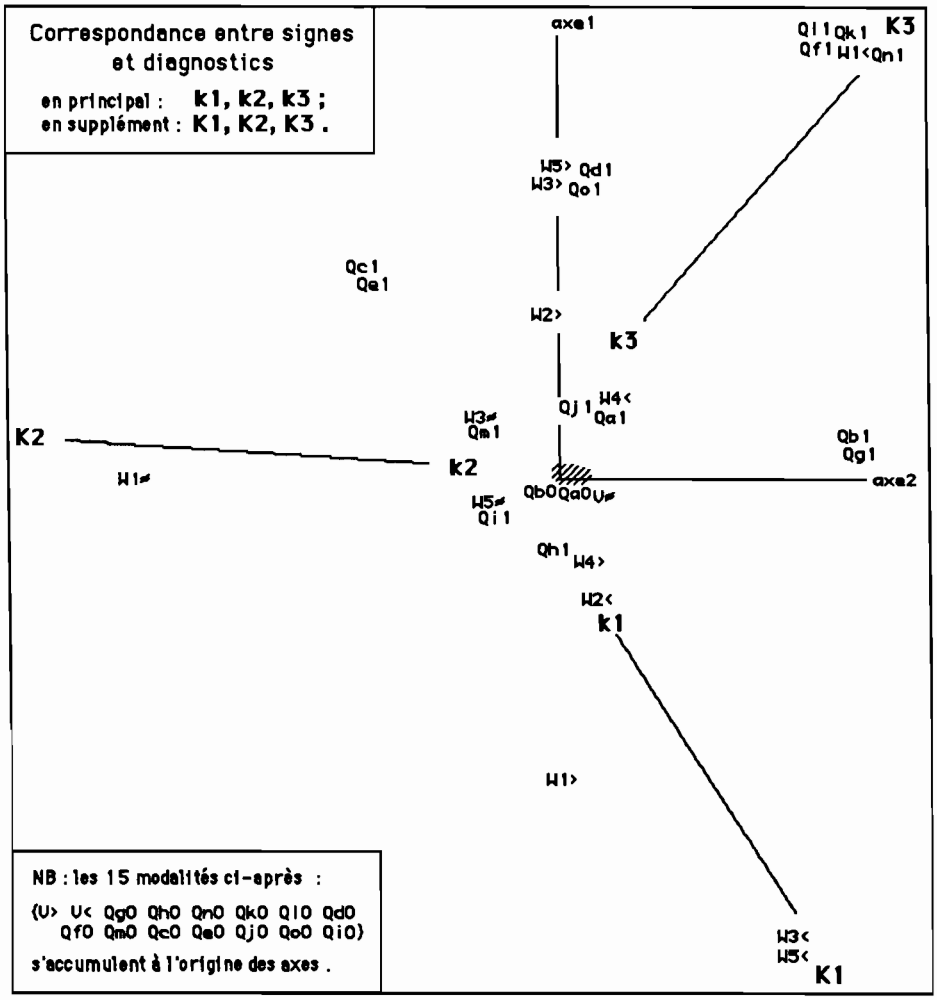
De telles erreurs auront pour conséquence clinique d'inciter à poursuivre l'observation de sujets qui, en définitive, devront être considérés comme sains. Il peut être plus fâcheux de déclarer sain un sujet dont la fonction thyroïdienne est anormale: mais on a vu que cette autre erreur a, ici, pour fréquence maxima 2,5%.

On soulignera, à ce propos, la supériorité des représentations continues issues de l'analyse factorielle sur les réponses en tout ou rien. Sans faire fi du diagnostic issu d'un algorithme quel qu'il soit, il vaut la peine de considérer d'après les valeurs des facteurs si le sujet s affecté à tel point kd n'est pas quasi à égale distance de kd et d'un autre point kd'.



Sur le plan (1, 2), le nuage des individus de l'échantillon d'épreuve est figuré avec des sigles choisis afin que les diagnostics se distinguent au mieux. Les cas normaux, D3, de beaucoup les plus nombreux, sont marqués d'un simple point; pour D1 et D2, on a des barres d'inclinaison différente: / pour D1; \ pour D2.

On voit bien les 6 cas de D2 (\) qui rentrent dans l'aire de D3 (nuage de points); et l'on pourrait d'après ceux-ci, étendre, au dépens de cette aire, celle des cas douteux.



2.3 Examen des variables et représentation de l'ensemble des modalités

D'une part, l'analyse multidimensionnelle a en commun avec les algorithmes d'approximation stochastique sur réseau de recourir au calcul automatique (encore que celle-là en consomme beaucoup moins que ceux-ci!). Mais, d'autre part, comme celle du médecin clinicien, la vision du statisticien se fonde sur la considération attentive des variables; et cela, avant et après analyse. Présentement, l'anonymat des variables en rend l'étude aride: il n'en est pas moins utile de montrer sur des exemples comment on a procédé.

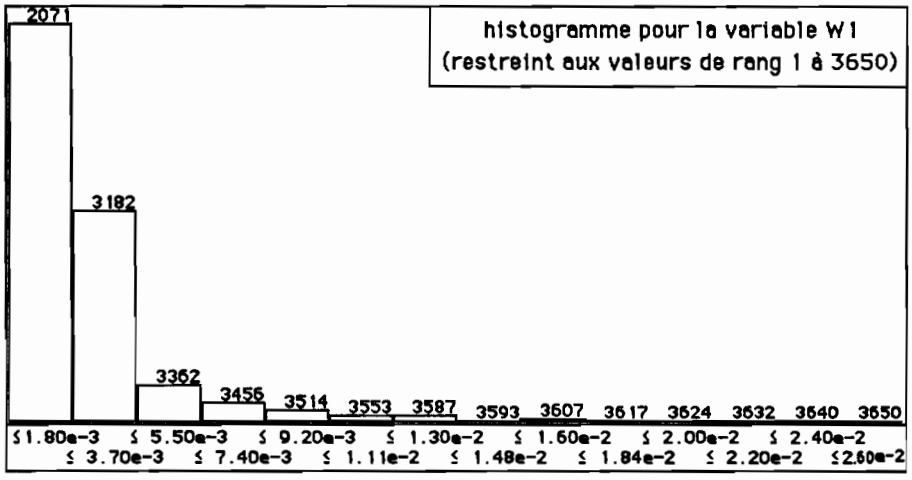
Quant aux variable logiques, il suffira de dire que les particularités notées au §2.1 se voient sur le plan (1, 2) issu de l'analyse factorielle: les modalités {Qf1, Qk1, Ql1, Qn1}, associées à la normalité, s'accumulent sur K3, leur profil de ligne étant concentré sur la 3-ème colonne; les modalités {Qc1, Qd1, Qe1, Qo1}, opposées à K1, se placent sur le côté {K2, K3} du triangle de sustentation du nuage. Au contraire, les modalités Qx0, de profil plat, sont à l'origine.

Nous porterons notre attention sur les variables continues; particuliérement, sur le bloc {W1, ...,W5} (la variable V n'ayant pas, du moins avec notre codage, contribué à la discrimination).

```
découpage des variables
étude de l'activité de la thyroïde
nvar =   6
W1 a 3 µod
W1< W1≈ W1>  0.0056  0.01   0.53
W2 a 2 mod
W2< W2>      0.008   0.027
W3 a 3 mod
W3< W3≈ W3>  0.015   0.07   0.15
W4 a 2 µod
W4< W4>      0.099   10000
W5 a 3 mod
W5< W5≈ W5>  0.0013  0.065  0.14
K  a 3 µod
K1  K2  K3   1       2      3
```

Sans reprendre l'exposé, donné ailleurs, de l'utilisation du programme zrang de découpage, nous donnons ici un fichier de commande; restreint toutefois à 6 variables, les {Wx} et le diagnostic K. De même, le listage d'analyse factorielle est limité aux modalités des variables {Wx}.

Prenons l'exemple de la variable W1, dont les 3 modalités apportent à l'analyse plus de 45% de son inertie. L'histogramme publié ici ne comprend que les valeurs ≤ (0,026), de rang 1 à 3650; au-delà, les 98 valeurs restantes s'étendent jusqu'à (0,53). Sur les premières lignes du tableau de croisement, on lit que les 3 créneaux



sup W1	D1	D2	D3
1.80e-3			2071
3.70e-3			1111
5.50e-3			180
7.40e-3	3	42	49
9.20e-3	1	42	15
1.11e-2	3	29	7
1.30e-2	2	25	7
1.48e-2		3	3
1.60e-2		9	5
1.84e-2	2	6	2
2.00e-2	1	5	1
2.20e-2	2	2	4
2.40e-2	4	4	
2.60e-2	6	2	2
5.30e-1	68	17	13

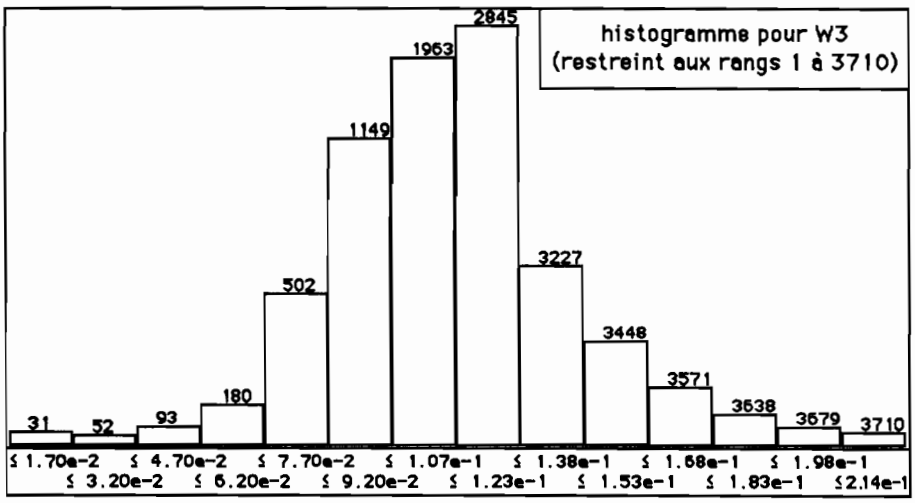
sup W3	D1	D2	D3
1.70e-2	31		
3.20e-2	14		7
4.70e-2	22	1	18
6.20e-2	13	7	67
7.70e-2	10	48	264
9.20e-2	2	57	588
1.07e-1		32	782
1.23e-1		24	858
1.38e-1		10	372
1.53e-1		7	214
1.68e-1			123
1.83e-1			67
1.98e-1			41
2.14e-1			31
4.30e-1			38

sup W5	D1	D2	D3
1.60e-2	31		
2.98e-2	11		6
4.54e-2	15		7
6.00e-2	29		22
7.46e-2	6	39	113
8.90e-2		48	402
1.04e-1		55	802
1.18e-1		28	1037
1.33e-1		13	498
1.47e-1		2	244
1.62e-1		1	145
1.76e-1			79
1.91e-1			37
2.06e-1			30
6.12e-1			48

NB: dans les croisements ci-dessus, la colonne supWx contient les bornes supérieures des intervalles successifs distingués pour la variable Wx.

inférieurs ne contiennent que des valeurs provenant de sujets normaux (diagnostic D3). Au contraire, D1 prédomine dans l'étalement supérieur (au-delà de notre histogramme); le bilan (porté sur la dernière ligne) étant: 68.D1 + 17.D2 + 13.D3. Entre ces deux extrêmes, D2 fournit la moitié du poids des autres crénaux. C'est pourquoi, après avoir considéré divers histogrammes et croisements, on a codé W1 suivant trois modalités logiques {W1<, W1≈, W1>} dont le tableau Dcodx donne les bornes; l'initiale 'μ', de μod, servant à spécifier qu'il n'y a pas de codage barycentrique; mais un simple découpage en classes consécutives.

Les 3 modalités de W3 apportent à l'analyse plus de 21% de son inertie. Comme W1, mais à un moindre degré, W3 s'étale vers les fortes valeurs.



L'histogramme publié est quasi symétrique; mais, au-delà de l'intervalle représenté, (0 ... 0,214), qui s'étend jusqu'au rang 3710, il reste 38 valeurs dont le maximum est (0,42). Les colonnes du tableau de croisement offrent chacune l'apparence de l'histogramme d'une distribution unimodale. D'abord, en haut, couvrant les faibles valeurs de W3, le diagnostic D1; puis D2; Puis la normalité, D3. Du fait du recouvrement des intervalles successifs, on a cru bon de prendre un codage barycentrique. En bref, avant le premier pivot (0,015), D1 est seul présent. Entre ce 1-er pivot et le second (0,07), on passe du maximum de D1 à celui de D2, cependant que s'introduisent les cas normaux, D3. Entre le deuxième pivot et le 3-ème, (0,15), D2 disparaît; et, au-delà de ce dernier pivot, il ne reste plus que D3.

activité thyroïdienne: tableau principal 46 × 3

trace : 1.278e-1

rang : 1 2

lambda : 918 360 e-4

taux : 7180 2820 e-4

cumul : 7180 10000 e-4

SIGI	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR
W1<	1000	15	246	1207	683	235	822	317	277
W1≈	1000	10	99	-1	0	0	-11181	1000	353
W1>	1000	23	110	-785	998	152	-34	2	1
W2<	1000	27	23	-322	970	31	57	30	2
W2>	1000	20	31	431	970	41	-76	30	3
W3<	1000	11	145	-1179	811	164	570	189	98
W3≈	1000	25	15	142	253	5	-243	747	41
W3>	1000	12	56	7771	000	78	-10	0	0
W4<	1000	22	9	221	977	12	34	23	1
W4>	1000	25	8	-194	977	10	-30	23	1
W5<	1000	8	124	-1236	797	137	623	203	89
W5≈	1000	25	7	-54	80	1	-184	920	23
W5>	1000	15	70	779	997	98	-40	3	1

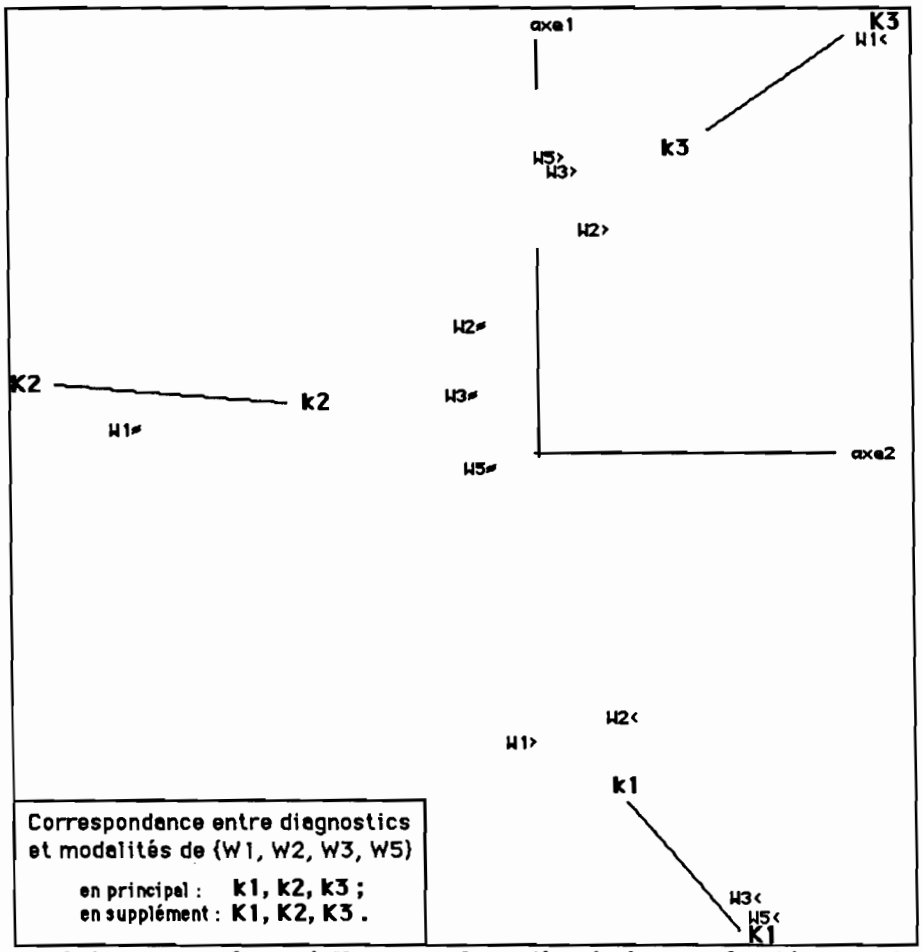
Après W3, vient la variable W5, apportant à l'analyse 20%, i.e. (124+7+70)‰, de son inertie.

Ainsi que l'atteste le tableau de croisement, le cas de W5 est analogue à celui de W3; aussi, là comme ici, a-t-on pris un codage barycentrique.

Avec le codage adopté ici, les contributions des modalités de W2 sont moindres que celles afférentes à {W1, W3, W5}; mais elles se voient sur l'axe 1, où les diagnostics s'ordonnent de D1 à D3.

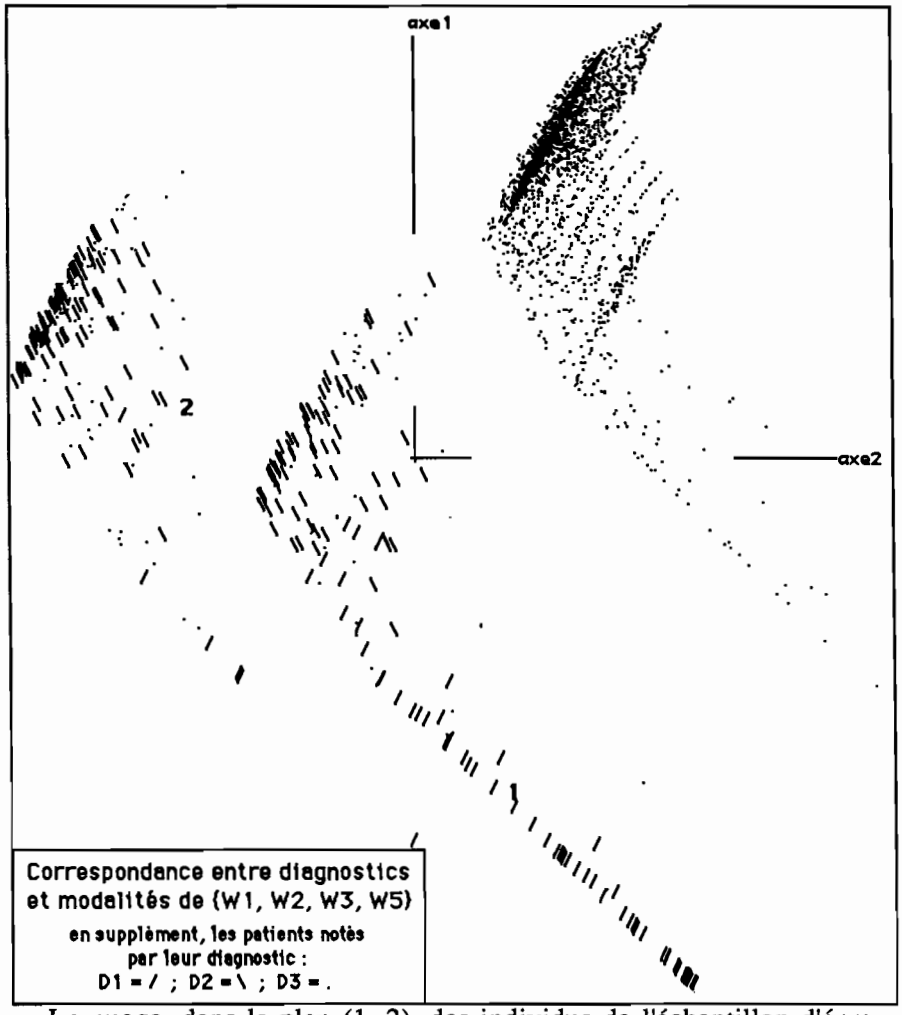
Nous n'avons mis que deux pivots: au-delà du second, (0,027), c'est la normalité, D3; avant le 1-er, (0,008), D1 prédomine; entre les deux, est le maximum de D2. Ici, on se demandera s'il n'eût pas mieux valu placer là un pivot intermédiaire; comme on l'a fait pour W3 et W5: or cette expérience, que nous avons tentée, *a posteriori*, n'a que peu modifié la discrimination, sans l'améliorer en rien. Mais elle nous a préparé à un autre essai qui fait l'objet du §2.4.

sup W2	D1	D2	D3
3.00e-3	15	3	25
6.00e-3	19	4	32
9.60e-3	22	11	85
1.20e-2	9	15	144
1.50e-2	9	24	309
1.80e-2	6	59	494
2.08e-2	7	32	1287
2.40e-2	3	22	493
2.70e-2	2	10	264
3.00e-2		1	132
3.30e-2		1	55
3.60e-2		1	43
3.90e-2		2	32
4.30e-2		1	25
1.06e-1			48



2.4 Analyse factorielle et analyse discriminante fondées sur les modalités de quatre variables principales

Dans l'analyse du §3, fondée sur l'ensemble des 21 variables descriptives, les modalités des 15 variables binaires {Qa, ..., Qo} et des deux variable continues {V, W4}, apportent moins de 7% de l'inertie du nuage. Ceci engage à reprendre l'analyse en se bornant aux modalités des 4 autres variables {W1, W2, W3, W5}. Pour {W1, W3, W5}, on conserve le codage présenté au §2.3; pour W2, selon ce qui a été noté ci-dessus, on a introduit un pivot intermédiaire. Comme précédemment, on construit un tableau de BURT généralisé, dont les colonnes afférentes aux trois diagnostics {k1, k2, k3} sont multipliées repectivement par {20, 10, 1}.



Le nuage, dans le plan (1, 2), des individus de l'échantillon d'épreuve marqué chacun {/, \, .}, montre d'emblée une bonne séparation des trois diagnostics. Les traînées de points, plus nettes encore qu'au §2.3, s'expliquent par le fait que des individus ayant en commun une même modalité de W1 (variable codée en 0,1), et des modalités extrêmes de deux autres variables, ne diffèrent entre eux que par le codage continu d'une dernière variable s'étalant entre deux pivots.

Quant au bilan de l'affectation des individus de l'échantillon d'épreuve, la présente analyse diffère peu de celle du §3.

Affectation d'après les 12 modalités des variables {W1, W2, W3, W5}
affectation aux col. pr. affectation aux lig. suppl. croisement des affectations
ci-dessous, bilan des affectations pour l'échantillon d'épreuve

	D1	D2	D3		D1	D2	D3		k1	k2	k3
k1	62	2	25	K1	64	6	28	K1	89	8	1
k2	11	174	111	K2	9	168	104	K2	0	281	0
k3	0	1	3042	K3	0	3	3046	K3	0	7	3042

De façon précise, le nombre d'erreurs dans les affectations des {D1, D2} (cas pathologiques) était, au §2, de 15, dont 6 diagnostics erronés de normalité; il est ici de 14 ou 18 selon qu'on affecte aux colonnes principales, {kd} ou aux lignes supplémentaires {Kd}, mais avec seulement 1 ou 3 cas anormaux donnés pour normaux. Quant aux 3178 cas véritablement normaux (compris dans D3), le nombre de ceux donnés pour anormaux était, au §2, de 117 ou 135; il est ici de 136 ou 132 (selon que l'affectation se fait aux {kd} ou aux {Kd}).

activité thyroïdienne: tableau principal 12 × 3
trace : 6.719e-1
rang : 1 2
lambda : 4955 1764 e-4
taux : 7374 2626 e-4
cumul : 7374 10000 e-4

SIGI	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR
W1<	1000	78	246	1160	630	210	888	370	346
W1≈	1000	53	99	61	3	0	-1117	997	377
W1>	1000	119	110	-782	990	147	-78	10	4
W2<	1000	92	75	-718	937	96	187	63	18
W2≈	1000	109	25	336	717	25	-211	283	27
W2>	1000	50	27	595	964	35	116	36	4
W3<	1000	57	145	-1209	852	168	503	148	82
W3≈	1000	131	15	155	304	6	-235	696	41
W3>	1000	62	56	776	998	76	33	2	0
W5<	1000	43	123	-1269	840	141	553	160	75
W5≈	1000	129	7	-44	52	1	-187	948	26
W5>	1000	78	70	7801000		95	3	0	0

SIGJ	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR
k1	1000	339	439	-904	938	559	232	62	103
k2	1000	342	179	133	51	12	-577	949	646
k3	1000	319	383	816	827	429	373	173	252

En un temps où il n'est guères de publication médicale qui s'abstienne de calculer des coûts, on appréciera que le nombre des variables à observer puisse être réduit de 21 à 4. Mais il faut avoir égard à la valeur nosologique des variables {Qx}.

Ci-dessous: pivots du codage barycentreique de W2
(le découpage des variables {W1, W3, W5} est fait comme au §2.3)

W2 a 3 mod
W2< W2≈ W2> 0.0095 0.0175 0.027

3 Conclusions et perspectives

Sur deux exemples, relevant chacun de la biologie, l'un d'abord facile, l'autre plus ardu, la méthode usuelle de discrimination après analyse de correspondance a fourni des résultats qui nous paraissent devoir satisfaire le praticien.

Pour l'exemple du §1, il faut redire en conclusion ce qu'on a déjà noté: 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. En fait, comme le suggère HASSAN HAMOUD Anwar, le véritable intérêt des données ne réside pas dans le diagnostic pur et simple de cancer, mais dans une typologie de la malignité, dont dépendent thérapeutique et pronostic.

Pour l'exemple du §2, le succès de la discrimination est essentiellement fondé sur le choix du codage; donc sur l'expérience acquise en analysant de multiples jeux de données. Mais, d'une part, on ne peut, dans la pratique, imaginer, qu'une étude vitale soit laissée absolument à la discrétion d'un algorithme; qu'il s'agisse de réseau ou de diagonalisation de matrice. D'autre part, il y a, dans le codage, une assez grande latitude; comme l'atteste le succès obtenu par HASSAN HAMOUD Anwar, sur les données du §1 (cf. [CODAGE DISCRI.], dans ce même cahier).

Reste à poursuivre la comparaison entreprise; d'une part, en prenant dans tous les domaines des données de tout format; d'autre part, en suivant, jusque dans le détail des liens et des erreurs, les processus de calcul des deux types de méthodes. Va dans ce sens la suggestion, faite au §2.1, d'introduire, dans les renforcements des liens des réseaux, des coefficients de pondération inversement proportionnels aux poids respectifs des diagnostics.

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;

A. El OUADRANI : "Généralisation du tableau de BURT et de l'analyse de ses sous-tableaux, dans le cas d'un codage barycentrique"; in *CAD*, Vol.XIX, n°5, pp.229-246; (1994);

J.-P. BENZÉCRI : "Approximation stochastique dans une algèbre normée non commutative"; in *Bull. Soc. Math. France.*; T.97, pp.225-241; (1969);

HASSAN HAMOUD Anwar : "Diversité des codages permis en analyse discriminante: exemple de données cytologiques"; [CODAGE DISCRI.], ce même cahier;

Références sur l' histoire des données

Origine générale de la compilation de nos données:

[PROBEN1] : A set of Neural Network Benchmark Problems and Benchmarking rules; Sept. 1994; par :

Lutz PRECHELT (prechelt@ira.uka.de), Fakultät für Informatik, Universität Karlsruhe; 78128 Karlsruhe, Allemagne.

Les données de [PROBEN1] renvoient à des collections bien connues:

University of California at Irvine machine learning databases archive: (`ics.uci.edu`, directory: `/pub/machine-learning-databases`);

Carnegie Mellon University *Neural Bench* collection: (`ftp.cs.cmu.edu`), directory: `/afs/cs/project/connect/bench`;

Exemple du §1 : Classification des tumeurs mammaires

Origine: William H. WOLBERG, médecin des hopitaux de l'Université de Wisconsin; Madison, Wisconsin, U.S.A. Transmis par:

Olvi MANGASARIAN, (`mangasarian@cs.wisc.edu`).

Publications antérieures:

O. L. MANGASARIAN & W. H. WOLBERG : "Cancer diagnosis via linear programming", *SIAM news*, Vol.23, n°5, pp.1-10; sept. 1990;

William H. WOLBERG & O. L. MANGASARIAN : "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Vol.87, pp. 9193-9196, Dec 1990;

Exemple du §2 : Appréciation de l'activité thyroïdienne

Transmis par:

Randolf WERNER, (`evol@infko.uni-koblenz.de`); (Reçu via Daimler-Benz).

Publications antérieures:

"Optimization of the Backpropagation Algorithm for Training Multilayer Perceptrons": `ftp archive.cis.ohio-state.edu` or `ftp 128.146.8.52 cd pub/neuroprose`, (application de quinze algorithmes différents à un "big, very hard to solve, practical data set);

"Synthesis and Performance Analysis of Multilayer Neural Network Architectures": ftp archive.cis.ohio-state.edu or ftp 128.146.8.52.