Commentaires sur une histoire de discrétisation.

L. Lebart (CNRS-ENST)

Si le contre-exemple présenté de façon plaisante par Gilles Celeux et Claudine Robert peut décourager définitivement les statisticiens d'application de procéder à une discrétisation aveugle de leurs données, il remplit une mission extrêmement importante.
Comme l'on pressenti les auteurs, qui ont fait un "appel à commentaires", cet exemple peut susciter d'innombrables remarques touchant aussi bien la théorie, la méthodologie, la pratique de la statistique.

Compte tenu du volume de commentaires attendu, nous avons dû opérer un choix, en nous restreignant aux trois points suivants :

1) Quelle "structure fondamentale" ?
2) La structure cachée.
3) Comment ne pas discrétiser.

1 - Quelle structure fondamentale ?

Quand dit-on en statistique qu'une structure fondamentale existe ?
La question n'est pas résolue en général, ni même toujours pertinente, mais dans le cadre de ce type d' exemple aux dimensions modestes, une concentration anormale sur un sous-espace, ou une partition en classes clairement disjointes sont habituellement considérées par les analystes de données comme des faits de structure.
Les termes *anormalement* et *clairement* font nécessairement appel à des modèles statistiques complexes. Ni les termes, ni les modèles ne sont évoqués par les auteurs qui se cantonnent (volontairement) à l'appréciation "structure fondamentale du nuage de 23 points, à savoir l'existence très apparente d'une partition en 3 groupes et un point isolé".

Sur cette appréciation qualitative d'expert, on peut emettre un avis différent : la structure n'est pas "très apparente", elle est cachée.

- L'analyse en composantes principales originale ne montre pas de structure aussi évidente que le disent les auteurs: le pattern observé sur la figure 3 montre incontestablement deux groupes; il est surtout remarquable par l'alignement des points sur la partie gauche. Mais les réalisations de processus spatiaux de type poissoniens font souvent apparaître des patterns aussi surprenant (surtout avec seulement 23 points).
Quand à la figure 4, il faut remarquer en lisant son échelle, que son axe vertical (axe 3) est considérablement dilaté (conséquence des sorties graphiques standards). Tous calculs refaits, les valeurs propres expliquent respectivement 60%, 32%, et 8% de la trace, ce qui confirme que le groupe du bas de la figure 3 est dans la réalité beaucoup plus proche des autres groupes que ne le suggère cette figure. L'analyse en composantes principale *normée au sens classique* (variables réduites) ne donne pas de meilleure structure apparente.
Ceci est confirmé par une classification hiérarchique des 23 points (opérée soit sur données brutes, soit sur données réduites, avec le critère de Ward) qui ne produit jamais la partition "très apparente", même si l'on réaffecte itérativement les individus après coupure de l'arbre en 4 classes. On retrouve en revanche les deux groupes visibles sur la figure 1, (cette figure correspond à 92% de la variance !).(voir figure (a) ci-dessous).

Il n'est pas étonnant qu'une structure qui repose en partie sur un troisième facteur expliquant 8% de la trace (pour 4 variables!), soit altérée par des perturbations du tableau de donnée, et la discrétisation en est une.

Le dendrogramme qui suit montre qu'au niveau des distances dans tout l'espace, sans recourir à la dissection trompeuse des plans factoriels, la partition fondamentale ne s'impose pas.

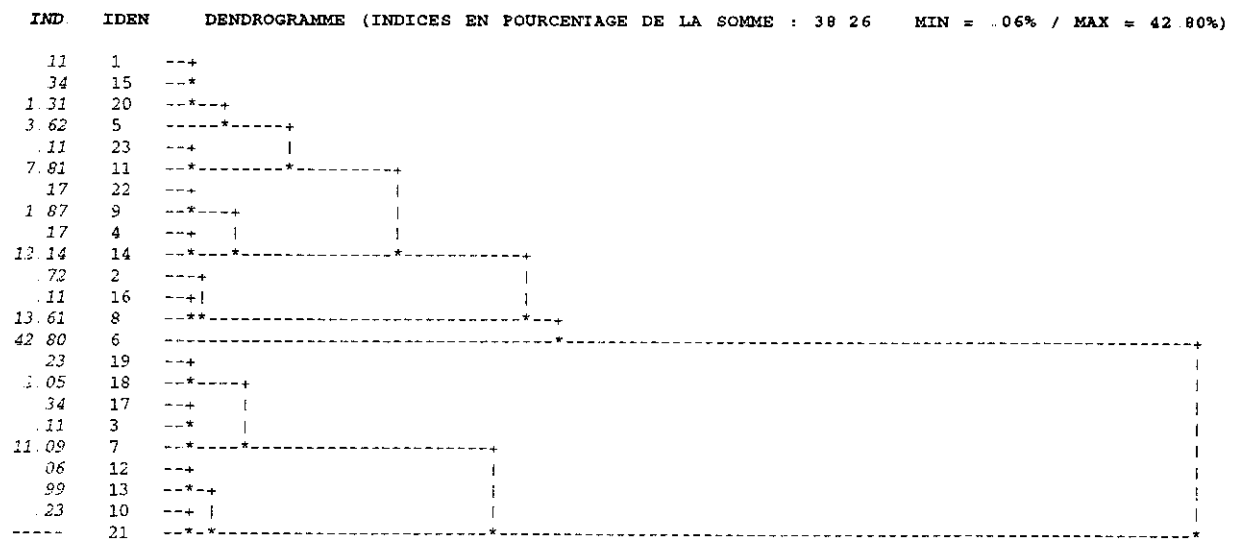

Fig. (a) : dendrogramme sur les distances réelles des figures 3 et 4.
(Aucune coupure ne restitue la "partition fondamentale")

2- La structure cachée

Un changement de variable simple mais non évident, proposé par les auteurs sans justification heuristique, fait apparaître une vraie structure, lourde : l'ACP du tableau des 3 variables z_1/z_4, z_2/z_4, z_3/z_4 génère une première valeur propre qui représente 98% de la trace, et la classification hiérarchique suivant le critère de Ward redonne bien cette fois-ci les 4 classes annoncées, avec cependant des niveaux d'indices très différents.

Tableau (a) : Observations "normées", classées suivant les valeurs de Z_2/Z_4

n°	Z_1/Z_4	Z_2/Z_4	Z_3/Z_4
6	0.813	1.094	.719
2	1.091	1.409	.955
8	1.100	1.500	.950
9	1.136	1.500	1.000
14	1.167	1.500	.958
16	1.150	1.500	1.000
4	1.174	1.565	1.043
22	1.238	1.619	1.048
15	1.100	1.800	1.200
23	1.143	1.810	1.238
5	1.167	1.833	1.278
1	1.158	1.842	1.263
20	1.150	1.850	1.250
11	1.143	1.857	1.286
12	1.706	2.294	1.588
19	1.733	2.333	1.600
21	1.722	2.333	1.611
13	1.706	2.353	1.588
17	1.750	2.375	1.625
3	1.800	2.400	1.667
10	1.765	2.412	1.647
18	1.786	2.429	1.643
7	1.800	2.467	1.733

Lisons le tableau des variables transformées, après l'avoir ordonné suivant les valeurs de z_2/z_4 (tableau (a) ci-dessus). Il est clair que les trois groupes et l'éléments isolés apparaissent sur les secondes et troisièmes colonnes (nous avons sauté des lignes entre les groupes pour que cela soit encore plus clair). Sur la colonne 2, les groupes sont concentrés près des valeurs (1.0, 1.5, 1.8, 2.4) et sur la colonne 3, autour des valeurs (0.7, 1.0, 1.2, 1.6) et les groupes sont séparés par des plages importantes sans observations.

Autrement dit, des histogrammes faits sur ces nouvelles variables aurait permis au moniteur lépidoptériste de retrouver ses espèces sans analyse multidimensionnelle.

Si de plus, la discrétisation avait eu lieu *après* ce changement de variable, après consultation des histogrammes, les limites de classes aurait été choisies sans problème, si ce n'est celui posé par l'outlier n°6. Autrement dit, la discrétisation n'aurait entrainé aucune perte d'information en ce qui concerne la structure maintenant très apparente.

Comment les auteurs ont-ils découvert ce changement de variable illuminant, qui n'a rien d'une normalisation standard ? A partir d'information a priori d'ordre biométrique ? A partir d'une *analyse logarithmique*, comme celle préconisée par J.B. Kazmierczak (Revue de Statistique Appliquée, 1985, vol. 23, n°1) , qui a l'avantage de linéariser quotients et produits ? Les trois premières variables z_1 , z_2 , z_3 sont très corrélées entre elles, et donc une suppression de l'effet taille au sens usuel aurait consisté à diviser les observations par $(z_1 + z_2 + z_3)$. Pour un lecteur peu averti, cette "normalisation" par z_4 reste un peu mystérieuse.

C'est un des mérite de cet exemple de montrer qu'une structure relativement forte et simple peut être dissimulée et assez mal détectée par les outils usuels.

3- Comment ne pas discrétiser

Les auteurs ont choisi d'utiliser l'algorithme de partitionnement de W.D.Fisher, pour que leur démarche reste transparente et reproductible, laisse le moins de place possible à la subjectivité.
Il n'est pas besoin de faire des analyses multidimensionnelle pour vérifier que cet algorithme optimal au sens de la variance interne des classes, ne permet pas de retrouver certaines partitions sur un seul axe.

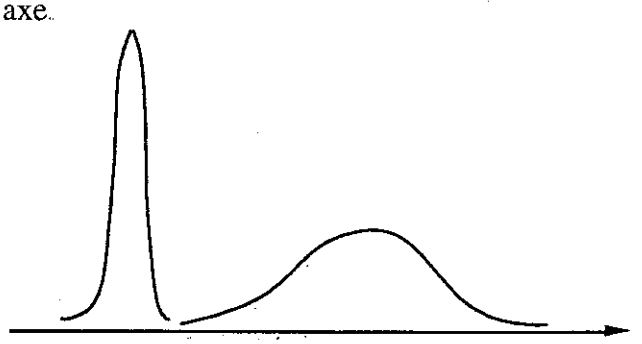

Fig. (b) : Distributions mal séparées par W.D.Fisher

Ainsi, sur la figure (b), il attribuera à la distribution de gauche une partie de la queue gauche de la distribution de droite. En fait, il exige une certaine homoscédasticité des distributions pour fonctionner convenablement.
Cette incompétence de l'algorithme qui n'a aucune conséquence grave dans le cas des papillons, doit cependant être soulignée à l'usage de ceux qui penseraient que la discrétisation n'est qu'une formalité.

Terminons sur une note d'espoir : dans les mornes traités de statistique, les iris de Fisher (qui sont, on le sait, les iris d'Anderson) vont-ils enfin être remplacés par les papillons de Claudine et Gilles ?