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.