

Commentaires sur l'article de G. Celeux et C. Robert
Gérard Govaert
CNRS URA 817 - Université de Technologie de Compiègne

Le texte de Celeux et Robert sur les effets parfois néfastes du recodage des variables continues en variables qualitatives appellent deux remarques de nature différente: la première concerne l'analyse des variables initiales qui aurait pu être un peu simplifiée et la seconde porte sur une analogie que l'on peut faire entre la discrétisation des images et le codage de variables continues en variables qualitatives.

0.1 L'analyse des variables continues

L'analyse préalable des variables continues, telle qu'elle est préconisée dans la conclusion du texte de Celeux et Robert peut être effectuée plus simplement et plus rapidement, sans faire appel notamment à l'analyse en composantes principales et en utilisant uniquement l'analyse des statistiques descriptives élémentaires (histogrammes, tableau de corrélations, ...):

- l'histogramme de la variable z_4 met en évidence la présence du papillon aberrant,
- la corrélation très forte de 0.97 entre les variables z_2 et z_3 permet d'éliminer sans une grande perte d'information l'une des deux variables,
- enfin, la prise en compte de l'effet taille, assez classique pour ce type de données, conduit à ne retenir que les deux nouvelles variables $t_1 = z_1/z_4$ et $t_2 = z_2/z_4$.

Il suffit alors de représenter le nuage des papillons dans le plan (t_1, t_2) (figure 1) pour mettre en évidence très clairement la structure en trois classes de l'ensemble des papillons.

Remarquons que même une simple analyse des différents plans (z_i, z_j) des variables initiales aurait pu suffire: en effet, le plan (z_3, z_4) (figure 2) montre assez nettement une structure en trois classes ainsi que le papillon aberrant.

0.2 Discrétisation des images et recodage des variables continues

En ce qui concerne la perte d'information due au recodage des variables continues en variables qualitatives, il est intéressant de faire un parallèle avec le traitement d'images: si nous nous limitons, par exemple aux deux variables z_1 et z_2 , le nuage des papillons peut être considéré comme une image dans le plan (z_1, z_2) . Comme la variable z_1 codée en entier de 21 à 31 peut prendre 11 valeurs différentes et z_2 codée en entier de 30 à 42 peut prendre 13 valeurs différentes, l'image ainsi définie est constituée de $11 \times 13 = 143$ pixels ou unités élémentaires d'image.

Avec cette analogie, le recodage des variables z_1, z_2 en variables qualitatives à 3 modalités s'interprète alors comme une discrétisation de l'image qui passe de 143 pixels à 9 pixels.