Commentaire sur une histoire de discrétisation

P. Besse, A. Carlier et A. de Falguerolles

Laboratoire de Statistique et Probabilités
Université Paul Sabatier, Toulouse

Introduction

Celeux et Robert affirment que *"... il est souvent avantageux de garder les variables dans leur forme originelle le plus longtemps possible ..."*. Nous souscrivons volontiers à cette recommandation trop fréquemment ignorée. Cependant, dans des situations exploratoires et en absence de problématique précise, la démarche contestée de discrétisation reste assez efficace. Dans une première partie nous montrons comment mener la discrétisation pour éviter d'observer le phénomène contrariant rapporté par les auteurs. Dans une seconde partie nous évoquons quelques méthodes d'analyse permettant de considérer conjointement variables qualitatives et quantitatives. Dans une troisième partie, nous analysons de nouveau les données initiales en cherchant à leur ajuster un modèle graphique gaussien.

Discrétisation

Ce travail est mené dans l'environnement Splus. Pour éliminer le facteur d'échelle, nous avons choisi de diviser chaque ligne du tableau par la moyenne de la ligne plutôt que par z_4. Les variables ainsi transformées ont été centrées et réduites. On note alors que l'ACP de ces variables met encore nettement en évidence les quatre groupes observés par les auteurs. La discrétisation proposée est déterminée comme suit. Les variables transformées étant homogènes, elles sont concaténées en une série unique de 4×23 observations. L'histogramme de cette série indique que les valeurs $-0,7$ et $+0,7$ correspondent approximativement aux fractiles d'ordre $\frac{1}{3}$ et $\frac{2}{3}$. Ces valeurs sont alors utilisées pour discrétiser chaque variable quantitative en une variable qualitative à trois modalités. Cette discrétisation revient à construire une table de contingence multiple. Dans cet exemple, il se trouve que tous les individus d'une même cellule appartiennent à un même groupe. Chaque individu (cellule) peut donc être représenté par le numéro du groupe auquel il appartient. On donne, Figure 1, la représentation de l'ACM effectuée sur les données ainsi discrétisées. On retrouve trois des groupes mis en évidence par l'ACP.

Si, compte tenu de son caractère grossier, la méthode n'identifie pas la singularité de l'observation 6 constituant le quatrième groupe, ce dernier est cependant affecté au groupe dont il est le plus proche. Mais faut-il considérer que cette observation constitue un groupe significatif, dans la mesure où elle présente des caractéristiques semblables à celle du groupe auquel elle est rattachée, mais sous forme accentuée ? En conclusion, une discrétisation bien menée peut rester assez efficace.