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.

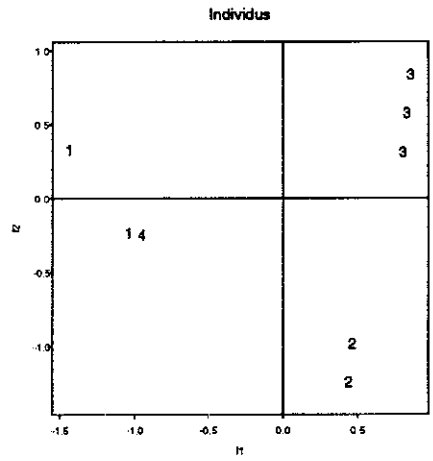

Figure 1: ACM des données discrétisées : plan 1-2.

Approches adaptées à des objectifs clairement définis

La question centrale de l'article de Celeux et Robert est celle du traitement simultané de variables qualitatives et quantitatives. Comme le soulignent ces auteurs, cette situation, assez fréquente dans les applications médicales notamment, motive une pratique contestable consistant à discrétiser toutes les variables quantitatives. Cependant que faire ? Mais d'abord quels sont les objectifs du traitement ? Si l'objectif est la classification, il existe de nombreuses méthodes spécifiques de ces situations mixtes. Dans cette perspective, la méthodologie du biplot (Gower, 1992) offre une approche stimulante. Si l'objectif est la détection de liaisons non-linéaires, des méthodes d'ACP non métrique (voir par exemple PRINQUAL de SAS) peuvent être utilisées (Gifi, 1990). Si l'objectif est la modélisation, la classe générale des *modèles graphiques d'associations* recouvre un ensemble de méthodes prometteuses (voir à ce sujet les bibliographies des exposés présentés par N. Wermuth et S. Lauritzen lors des Journées de Statistique de Vannes, 1993). Rappelons que ces modèles s'appliquent à des variables quantitatives (gaussiennes), qualitatives (multinomiales) ou mixtes (conditionnellement gaussiennes). Dans ces modèles les notions d'indépendance conditionnelle sont traduites par un graphe dont les sommets représentent les variables. Ce graphe est orienté ou non selon que l'on désire ou non mettre en évidence le rôle causal de certaines variables. L'absence d'arête entre deux variables traduit une indépendance conditionnelle de ces variables. Pour une introduction à ces modèles on pourra se reporter à l'exposé introductif de Fine (1992) ou à l'ouvrage de Whittaker (1990). La mise en oeuvre d'un modèle simple pour les papillons de Celeux et Robert est considérée ci-après.

Un modèle graphique gaussien pour des papillons

Parce qu'analysant aussi la structure de la matrice de corrélation, l'approche *modèle de sélection de covariance* (Dempster, 1972) est susceptible de fournir des renseignements pertinents dans une analyse exploratoire de variables quantitatives. On sait que ces modèles sont aussi appelés *modèles graphiques gaus-*

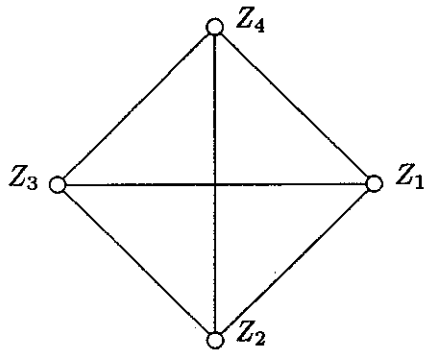

Figure 2: Graphe du modèle saturé.

siens pour souligner leur appartenance à la large classe des *modèles graphiques d'associations* déja évoqués. Les données sont ici analysés avec le logiciel MIM (version 2.1) de Edwards (1993).

Le calcul des corrélations et des corrélations conditionnelles aux autres variables est donné Tableau 1. Ces valeurs empiriques usuelles correspondent au modèle saturé dans lequel on suppose qu'aucune paire de variables n'est indépendante conditionnellement aux autres variables (voir Figure 2).

	Z_1	Z_2	Z_3	Z_4
Z_1	1.000			
Z_2	0.693	1.000		
Z_3	0.636	0.975	1.000	
Z_4	-0.175	-0.229	-0.341	1.000

Table 1: Matrice de corrélation du modèle saturé.

	Z_1	Z_2	Z_3	Z_4
Z_1	1.000			
Z_2	0.464	1.000		
Z_3	-0.320	0.972	1.000	
Z_4	-0.202	0.529	-0.575	1.000

Table 2: Matrice des corrélations partielles du modèle saturé.

La procédure de recherche automatique de modèle donne un modèle dans lequel les variables z_1 et (z_3, z_4) sont indépendantes conditionnellement à z_2 (voir Figure 3). Ce graphe indique en particulier les meilleurs prédicteurs linéaires de chaque variable : il suffit de considérer les sommets adjacents de chaque

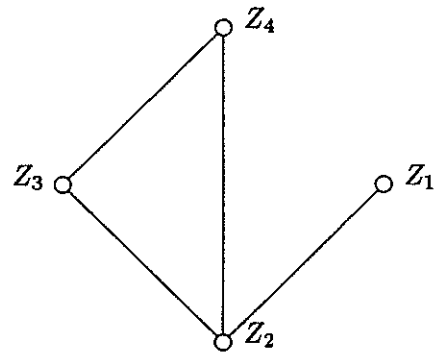

Figure 3: Graphe du modèle retenu.

variable.

	Z_1	Z_2	Z_3	Z_4
Z_1	1.000			
Z_2	0.693	1.000		
Z_3	0.676	0.975	1.000	
Z_4	-0.159	-0.229	-0.341	1.000

Table 3: Estimation de la matrice de corrélation du modèle retenu.

	Z_1	Z_2	Z_3	Z_4
Z_1	1.000			
Z_2	0.180	1.000		
Z_3	0.000	0.965	1.000	
Z_4	0.000	0.493	-0.550	1.000

Table 4: Estimation de la matrice des corrélations partielles du modèle retenu.

Le graphe de ce modèle ne contredit pas les résultats de l'ACP des variables transformées : l'axe 1 est structuré par la clique z_2, z_3 et z_4, l'axe 2 est déterminé par z_1. Enfin on constate que les corrélations estimées sous ce modèle sont assez voisines de celles obtenues pour le modèle saturé (voir Table 1 et Table 3). Les corrélations partielles estimées sont données en Table 4.

Signalons enfin que des procédures de sélection de variables en ACP reposent sur la recherche de modèles graphiques gaussiens particuliers (Falguerolles et Jmel, 1992). Appliquées à ces données, ces procédures font ressortir l'importance des variables z_1, z_2 et z_4 (pour une sélection de trois variables), des variables

z_2 et z_3 (pour une sélection de deux variables), et de la variable z_2 (pour une sélection d'une seule variable).

Références

Actes des $XXV^{\text{ième}}$ Journées de Statistique (1993). Institut Universitaire de Technologie de Vannes, 8 rue Montaigne, F-56014 Vannes Cedex.

Dempster A.P. (1972) : *Elements of Continuous Multivariate Analysis*. Addison-Wesley: Reading, Mass.

Edwards, D. (1993) : *Graphical Modelling with MIM 2.1*. HyperGraph Software, Bymarken 38, DK-4000 Roskilde, Denmark.

Falguerolles, A. de, et Jmel, S. (1992) : *Modèle graphiques gaussiens et analyse en composantes principales. Complémentarité et choix de variables*. Publication du Laboratoire de Statistique et Probabilités, 03-92, Université Paul Sabatier, Toulouse.

Fine, J. (1992) : *Modèles graphiques d'associations*, in Modèles pour l'analyse des données multidimensionnelles, Droesbeke, J.-J. *et al.* (ed.). Economica : Paris, 267-313.

Gifi, A. (1990) : *Non Linear Multivariate analysis*. John Wiley & Sons.

Gower, J. (1992) : Generalized Biplots. *Biometrika*, 79, 3, 475-493.

Whittaker, J. (1990) : *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons, Chichester.