

Commentaire sur la note de Gilles Celeux et Claudine Robert

Jean-Christophe Turlot
STID Pau, URA CNRS 1104

Un exemple à l'appui, les auteurs présentent avec acuité les conséquences "malheureuses" d'une discrétisation de variables numériques "continues". Une telle transformation est probablement nécessaire dès lors que nous sommes en présence de variables de nature mixte (quantitatives et qualitatives) ; elle a pour effet de dénaturer les données, que l'on se place d'un point de vue probabiliste (le passage d'une densité sur la droite réelle à une densité discrète) ou d'un point de vue géométrique (les points d'une droite réelle sont transformés en sommets d'un simplexe). C'est le point de vue géométrique qui est abordé ici : l'analyse en composantes principales est fondée sur la matrice des covariances entre variables continues ; elle vise à décrire l'essentiel des liaisons entre variables de manière linéaire au moyen d'un petit nombre de facteurs. Cette technique peut être "étendue" à des variables de nature mixte par discrétisation des variables numériques, le tableau "disjonctif-complet" étant alors traité par l'analyse des correspondances multiples (ACM). L'essentiel des liaisons entre les variables qualitatives naturelles ou construites est décrit par les premières composantes principales de l'ACM. La liaison entre deux variables qualitatives est mesurée ici par la statistique du F2 de Pearson.

Le point que je souhaite aborder est le suivant : cette technique d'homogénéisation des données engendre-t-elle une perte de précision ou une perte d'information ? Une réponse sybilline serait la suivante : oui, il y a perte de précision (l'utilisateur est souvent réticent de prime abord à une opération de discrétisation) ; quant à la seconde question, on ne peut y répondre. Cependant, les applications montrent que les choses ne sont pas si simples.

i) la perte de précision :

lorsque le médecin établit un diagnostic, celui-ci dispose d'un ensemble de paramètres dont certains sont de nature quantitative. Or, le même niveau d'un paramètre sera décrété comme normal ou au contraire anormal selon le niveau d'autres critères quantitatifs ou qualitatifs.

Une telle démarche montre que les mesures sont interprétées de manière essentiellement qualitative, mais aussi que cette interprétation peut être fortement dépendante du niveau de variables concomitantes. La présence d'interactions nécessite une discrétisation adaptée des données et l'on peut penser que, pour bien des problèmes, l'adéquation entre la structure des données et le modèle qualitatif retenu rend évanescence la perte de précision liée à la discrétisation.

ii) la perte d'information :

tout statisticien jugera, comme les auteurs, que l'ACP normée (non pas au sens usuel, mais en considérant les ratios z_1/z_4 , z_2/z_4 , z_3/z_4) conduit à une discrimination éclatante sur le plan principal (fig. 5). A contrario, l'analyse en composantes principales (fig. 3) fondée sur une information exhaustive (en regard des mesures effectuées) donne une image beaucoup plus confuse des dissemblances entre espèces.

On peut aisément imaginer des situations où la discrétisation de variables numériques conduirait à une représentation plus pertinente du point de vue géométrique que celle fournie par l'ACP. Aussi ne peut-on parler formellement de perte d'information liée à un codage nécessaire de variables numériques. Dans la situation présente, l'information utile est attachée à la forme des papillons : l'ACP non normée (fig. 3) exhibe à l'évidence un facteur "taille" faiblement discriminant associé à l'axe 1, la variable z_4 engendrant par sa variabilité une grande dispersion intra-classe comme semble le montrer la figure 5 en regard de la figure 3. On doit reconnaître que le plan principal de l'ACP normée par z_4 contient "l'essentiel" de l'information discriminante.

L'analyse des correspondances effectuée ne peut être pertinente au sens de la discrimination entre espèces : le découpage en classes effectué conduit à un effet taille