Classification de variables : le package ClustOfVar

Marie Chavent, Vanessa Kuentz, Benoit Liquet, Jérôme Saracco

Classification de variables : le package ClustOfVar

Marie Chavent[1,2] & Vanessa Kuentz[3] & Benoît Liquet[4] & Jérôme Saracco[1,2]

1. IMB, Université de Bordeaux, France
2. Equipe CQFD, INRIA Bordeaux Sud-Ouest, France
3. CEMAGREF, UR ADBX, France
4. ISPED, Université de Bordeaux, France

Résumé

Le package R "ClustOfVar" a été développé spécifiquement pour répondre au problème de la classification de variables. Les variables considérées peuvent être toutes quantitatives, toutes qualitatives ou un mélange des deux. Le critère d'homogénéité d'une classe est la somme des carrés des corrélations (pour les variables quantitatives) et des rapports de corrélations (pour les variables qualitatives) à une variable synthétique (quantitative) résumant au mieux les variables de la classe. La variable synthétique qui maximise ce critère est la première composante principale calculée par la méthode PCAMIX. Deux algorithmes sont proposés : un algorithme de classification ascendante hiérarchique et un algorithme de partitionnement de type k-means. Une approche de type bootstrap est proposée pour l'aide au choix du nombre de classes. Les différentes méthodologies sont illustrées sur des données réelles.
Mots-clés : Données mixtes, hiérarchie de variables, partition de variables, réduction de dimension.

Abstract

The R package "ClustOfVar" was developed specifically for clustering of variables. Variables can be either all quantitative, all qualitative or a mixture of both. The homogeneity criterion of a cluster is the sum of correlation ratios (for qualitative variables) and squared correlations (for quantitative variables) to a synthetic variable, summarizing "as good as possible" the variables in the cluster. This synthetic variable is the first principal component obtained with the PCAMIX method. Two algorithms for the clustering of variables are proposed: iterative relocation algorithm, ascendant hierarchical clustering. We also propose a bootstrap approach in order to determine suitable numbers of clusters. The proposed methodologies are illustrated on real datasets.

Keywords: Mixture of quantitative and qualitative variables, hierarchical clustering of variables, k-means clustering of variables, dimension reduction.

Introduction.

L'objectif de la classification de variables est d'obtenir des classes de variables liées et redondantes. Des algorithmes spécifiques ont ainsi été développés pour la classification de variables. Une méthode divisive de classification de variables quantitatives est implémentée dans la procédure VARCLUS du logiciel SAS. Les méthodes CLV de Vigneau et Qannari (2003) et Diametrical Clustering de Dhillon et al. (2003) sont des algorithmes de partitionnement de variables quantitatives de type k-means où la variable synthétique d'une classe est la première composante principale d'une ACP et où l'homogéneité d'une classe est la somme des carrés des corrélations linéaires des variables de la classe à cette variable synthétique. La méthode de la vraisemblance du lien (Lerman, 1993) est un algorithme hiérarchique de classification de variables qualitatives. Abdallah et Saporta (1998) ont proposé de nouveaux indices de similarité pour la classification de variables qualitatives. Chavent et al. (2009) proposent une adaptation de l'algorithme de partitionnement CLV au cas qualitatif où la variable synthétique d'une classe est la première composante principale de l'ACM et où l'homogéneité d'une classe est la somme des rapports des corrélations des variables de la classe à cette variable synthétique.

Nous proposons dans le package R "ClustOfVar" deux algorithmes de classification de variables (un algorithme ascendant hiérarchique et un algorithme de partitionnement) qui ont été développés pour gérer le cas de données quantitatives et/ou qualitatives. Une approche de type bootstrap est également proposée pour l'aide au choix du nombre de classes.

Le critère d'homogénéité d'une classe.

L'homogénéité d'une classe est la somme des carrés des corrélations linéaires (pour les variables quantitatives) et des rapports de corrélations (pour les variables qualitatives) à une variable synthétique résumant au mieux les variables de la classe. La variable synthétique qui optimise ce critère d'homogénéité est la première composante principale de la méthode PCAMIX. Cette méthode d'analyse factorielle de données mixtes (Kiers, 1991), aussi appellée AFDM (Pagès, 2004), inclut l'ACP et l'ACM comme cas particuliers. Une approche par décomposition en valeurs singulières a été implémentée pour le calcul de la première composante principale et de sa variance (qui mesure l'homogénéité de la classe).

Les deux algorithmes de classification de variables.

Un *algorithme de classification ascendante hiérarchique* a été implémenté dans la fonction `hclustvar` du package. A chaque étape les deux classes qui minimisent la perte d'homogénéité sont agrégées. La fonction `plot.hclustvar` permet de visualiser le dendrogramme de la hiérarchie et le graphique des indices de niveau de cette hiérarchie. La fonction `cutreevar` permet d'extraire une partition de cette hiérarchie.

Un *algorithme itératif de partitionnement de type k-means* a été implémenté dans la fonction `kmeansvar` du package. A chaque étape les représentants des classes sont les premières composantes principales de PCAMIX de chaque classe. Chaque variable est ensuite réaffectée à la classe qui lui est la plus proche, c'est à dire celle dont le carré de la corrélation (si la variable est quantitative) ou le rapport de corrélation (si la variable est qualitative) entre la variable et le représentant de la classe est le plus grand. La partition initiale peut-être choisie par l'utilisateur ou définie à partir de k variables tirées aléatoirement et utilisées comme "centres" initiaux. Une première étape d'affectation est alors réalisée avec une mesure de similarité qui s'interprète comme le carré d'une corrélation canonique. Cette mesure de similarité entre deux variables de type quelconque a été implémentée dans la fonction `mixedVarSim`.

Afin d'évaluer la stabilité de toutes les partitions en 2 à p-1 classes issues de la classification hiérarchique, la fonction `stability` a été développée dans le package. Elle est fondée sur une approche "bootstrap" décrite ci-après.
La fonction `hclustvar` est appliquée à B réplications bootstrap de l'échantillon des n observations initiales. Pour chaque réplication, les partitions de 2 à p-1 classes obtenues sont comparées aux partitions de la hiérarchie initiale par le biais des calculs du Rand corrigé [1]. Les moyennes (sur les B r{eplications) de ces indices de Rand corrigé sont alors calculées et représentées graphiquement en fonction du nombre de classes. Cette représentation graphique est alors un outil pouvant aider au choix du nombre de classes : l'utilisateur pourra choisir le nombre k de classes correspondant à la valeur moyenne du Rand corrigé la plus forte, ce qui sous-entend une structure stable pour cette partition en k classes.

Les sorties.

Outre les représentations graphiques des dendrogrammes, des indices de niveaux et des stabilités des partitions, des sorties numériques sont fournies pour chaque partition de variables. Par exemple, on obtient la liste des variables de chaque classe et leurs "squared loadings", le "squared loadings" d'une variable étant le carré de sa corrélation linéaire (si elle est quantitative) ou son rapport de corrélation (si elle est qualitative) avec la variable synthétique de la classe. On obtient également la matrice des "scores", c'est-à-dire la matrice des k nouvelles variables synthétiques résumant les k classes. On peut alors considérer que ces k nouvelles variables résument bien les variables initiales tout en supprimant la redondance d'information des données initiales. La matrice des scores peut alors être utilisée pour des traitements ultérieurs tout comme la matrice des scores en ACP, en ACM ou encore en PCAMIX.

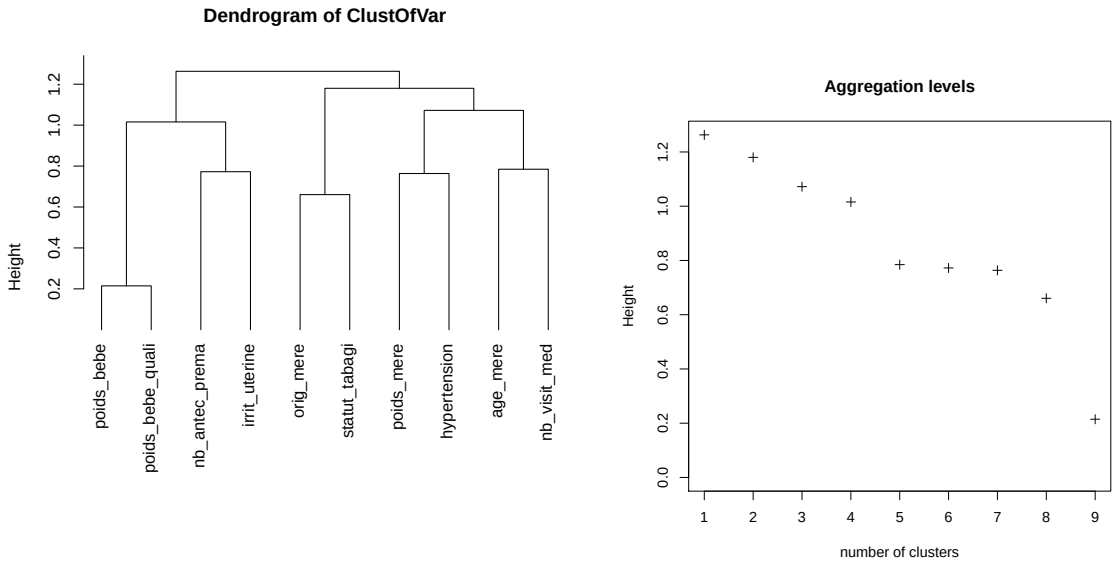
1. L'indice de Rand corrigé ainsi que sa version non corrigée et sa version asymétrique ont été implémentés dans la fonction `Rand` du package.

Un exemple illustratif.

Nous utilisons le jeu de données "Poids de naissance" contenant la description de 189 femmes qui sont venues consulter dans un centre médical. Ces données sont disponibles à l'adresse suivante : www.biostatisticien.eu/springeR/donnees.html. Les données sont mixtes : 5 variables sont quantitatives ("âge de la mère", "poids de la mère", "nombre d'antécédents de prématuré", "nombre de visites médicales durant le premier trimestre de grossesse", "poids du bébé à la naissance") et 5 variables sont qualitatives ("origine de la mère", "statut tabagique de la mère", "antécédents d'hypertension", "présence d'irritabilité utérine", "poids de naissance du bébé inférieur à 2500g"). Voici un exemple de code R suivi de quelques sorties graphiques et numériques obtenues avec le package "ClustOfVar". Les lignes de code,

```
require(ClustOfVar)
tree<-hclustvar(X.quanti=X1,X.quali=X2)
plot(tree,main="Dendrogram of ClustOfVar")
```

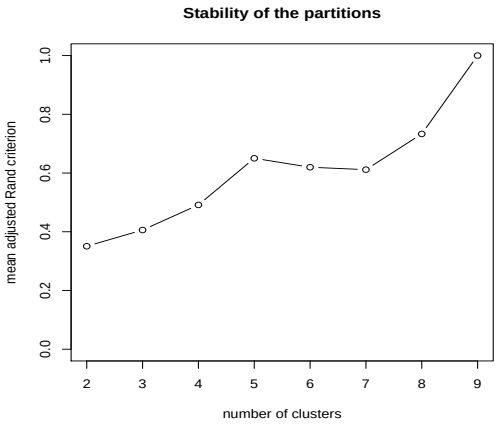
donnent les graphiques suivant :



Les lignes de code,

```
stab<-stability(tree,B=100)
plot(stab,main="Stability of the partitions")
```

calculent la stabilité des toutes les partitions en 2 à 7 classes de cette hiérarchie. Sur le graphique ci-dessous, on note que la stabilité augmente jusqu'à 5 classes. On choisit donc la partitition en 5 classes.



Les lignes de code,

```
P5 <- cutreevar(tree,5)
summary(P5)
```

fournissent une description de la partition en 5 classes :

```
Cluster  1 :
             squared loading
age_mere                0.61
nb_visit_med            0.61

Cluster  2 :
             squared loading
poids_mere              0.62
hypertension            0.62

Cluster  3 :
               squared loading
nb_antec_prema            0.61
irrit_uterine             0.61

Cluster  4 :
                 squared loading
poids_bebe                   0.9
poids_bebe_quali             0.9

Cluster  5 :
              squared loading
orig_mere                0.67
statut_tabagi            0.67

Gain in cohesion (in %):  58.64
```

La ligne de code,

```
round(P5$scores,digit=2)
```

donne les variables synthétiques quantitatives des 5 classes résumant les 10 variables :

```
   cluster1 cluster2 cluster3 cluster4 cluster5
1    -1.10    -1.03     1.41     0.07    -0.60
2     2.78    -0.40    -0.58     0.09    -1.47
3    -0.30     0.76    -0.58     0.10     1.52
...
187    -0.56     1.01    -0.58    -1.49    -0.03
188    -1.37    -3.00    -0.58    -1.49    -0.60
189     1.18    -2.72    -0.58    -1.49     1.52
```

Bibliographie

Abdallah, H. and Saporta, G., (1998), Classification d'un ensemble de variables qualitatives, *Revue de Statistique Appliquée*, **46**(4), 5-26.

Chavent, M, Kuentz, V. and Saracco J. (2009). A Partitioning Method for the clustering of Categorical variables. In *Classification as a Tool for Research*, Hermann Locarek-Junge, Claus Weihs (Eds), Springer, Proceedings of the IFCS'2009, Dresden.

Dhillon, I.S., Marcotte, E.M. and Roshan, U. (3003). Diametrical clustering for identifying anti-correlated gene clusters, *Bioinformatics*, **19**(13), 1612-1619.

Pagès, J. (2004). Analyse factorielle de données mixtes, *Revue de Statistique Appliquée*, **52**(4), 93-111.

Kiers, H.A.L., (1991). Simple structure in Component Analysis Techniques for mixtures of qualitative and quantitative variables, *Psychometrika*, **56**, 197-212.

Lerman, I.C., (1993). Likelihood linkage analysis (LLA) classification method : An example treated by hand, *Biochimie*, **75**,(5) 379-397.

Vigneau, E. and Qannari, E.M., (2003). Clustering of Variables Around Latent Components, *Communications in Statistics - Simulation and Computation*, **32**(4), 1131–1150.