

**dad : package pour l'analyse de données multi-groupes  
via les densités de probabilité associées**

**R. Boumaza<sup>a</sup>, P. Santagostini<sup>a</sup>, S. Yousfi<sup>b</sup> and S. Demotes-Mainard<sup>c</sup>**

<sup>a</sup>IRHS, Agrocampus Ouest, INRA, Université d'Angers, SFR 4207 QuaSaV  
49045, Angers, France  
rachid.boumaza@agrocampus-ouest.fr  
pierre.santagostini@agrocampus-ouest.fr

<sup>b</sup>Département de Mathématiques  
Université Mouloud Mammeri  
Tizi-Ouzou, Algérie  
smail\_yousfi@ymail.com

<sup>c</sup>IRHS, INRA, Agrocampus Ouest, Université d'Angers, SFR 4207 QuaSaV  
49070, Beaucouzé, France  
sabine.demotes-mainard@inra.fr

**Mots clefs** : données multi-groupes, analyse en composantes principales, classification hiérarchique, multidimensional scaling, analyse discriminante

En vue d'explorer des données multivariées, observations de plusieurs variables sur un nombre fini d'individus, l'analyste de données utilise des techniques telles que l'analyse en composantes principales, la classification automatique, le positionnement multidimensionnel (multidimensional scaling), l'analyse discriminante. Ces techniques traitent des données du type "individus x variables", généralement stockées dans des objets R de type `data frame`. Ces techniques sont disponibles à partir des packages comme `MASS`, `ade4`, `FactoMineR`, `cluster`... Dans le cas où les individus sont organisés en groupes, l'analyste peut être intéressé de tenir compte de cette organisation. Le package `dad` ([1]) propose un moyen d'y parvenir.

Il consiste à associer à chaque groupe une fonction de densité de probabilité supposée de carré intégrable (les densités gaussiennes en sont un cas particulier) puis on soumet l'ensemble de ces densités à l'une des analyses fonctionnelles suivantes :

- L'analyse en composantes principales (ACP) des fonctions de densité de probabilité est une technique de réduction d'un nombre fini de densités qui est utilisé pour identifier les principaux modes de variation des données numériques multi-groupes ([2]). Dans ce contexte d'analyse en composantes principales, Eslami et al. ([3]) ont fait un survol général des méthodes d'analyse de données multi-groupes. Leur travail est à la base du package `multigroup` implémentant ces méthodes ;
- Le positionnement multidimensionnel (MDS) des fonctions de densité de probabilité conduit à visualiser l'ensemble des densités de sorte que les distances entre densités soient le mieux préservées possible. Dans certaines conditions, cette visualisation est semblable à celle obtenue par l'ACP précédente, à la différence près que MDS offre une plus grande flexibilité dans le choix des distances entre densités ;
- La classification hiérarchique des fonctions de densité de probabilité est utilisée pour répartir un ensemble de densités (ou groupes d'individus) en sous-ensembles de sorte que les densités du même sous-ensemble soient aussi semblables que possible, et sont dissemblables de ceux des autres sous-ensembles ;

- L'analyse discriminante des fonctions de densité de probabilité est une autre méthode qui traite le même type de données, connaissant une partition a priori des densités (ou groupes d'individus) en classes. Son objectif est de classer une nouvelle densité dans une classe définie a priori ([4]). Cela équivaut à classer collectivement un groupe d'individus dans une classe.

Les techniques précédentes, implémentées dans le package `dad`, feront l'objet d'une présentation sur la base d'exemples de données sensorielles, archéologiques ou agronomiques.

- [1] Boumaza, R., Santagostini, P., Yousfi, S., Hunault, G., Bourbeillon, J., Pumo, B. and Demotes-Mainard, S. (2018). `dad` : Three-Way Data Analysis Through Densities. R package version 3.1.0. <https://CRAN.R-project.org/package=dad>
- [2] Boumaza, R., Yousfi, S. and Demotes-Mainard, S. (2015). Interpreting the principal component analysis of multivariate density functions. *Theory and Methods*, 44(16), 3321-3339.
- [3] Eslami, A., Qannari, E.M., Kohler, A. and Bougeard, S (2013). General overview of methods of analysis of multi-group datasets. *Revue des Nouvelles Technologies de l'Information*, 25 : 108–123.
- [4] Boumaza, R. (2004). Discriminant analysis with independently repeated multivariate measurements : An  $L^2$  approach. *Computational Statistics and Data Analysis*, 47(4), 823-843.