HAL open science

HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data

Laurent Bergé, Charles Bouveyron, Stéphane Girard

To cite this version:

Laurent Bergé, Charles Bouveyron, Stéphane Girard. HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data. 1ères Rencontres R, Jul 2012, Bordeaux, France. hal-00717506

HAL Id: hal-00717506

https://hal.science/hal-00717506v1

Submitted on 13 Jul 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data

L. Bergé[a] and **C. Bouveyron**[b] and **S. Girard**[c]

[a]Laboratoire GREthA
Université Bordeaux IV
laurent.berge@u-bordeaux4.fr

[b]Laboratoire SAMM, EA 4543
Université Paris 1 Panthéon-Sorbonne
charles.bouveyron@univ-paris1.fr

[b]Equipe Mistis
INRIA Rhône-Alpes & LJK
stephane.girard@inrialpes.fr

Mots clefs : Model-based classification and clustering, high-dimensional data, subspaces.

This paper presents the **R** package *HDclassif* which is devoted to the clustering and the discriminant analysis of high-dimensional data. The classification methods proposed in the package result from a new parametrization of the Gaussian mixture model which combines the idea of dimension reduction and model constraints on the covariance matrices. The supervised classification method using this parametrization is called high dimensional discriminant analysis (HDDA). In a similar manner, the associated clustering method is called high dimensional data clustering (HDDC) and uses the expectation-maximization algorithm for inference. In order to correctly fit the data, both methods estimate the specific subspace and the intrinsic dimension of the groups. Due to the constraints on the covariance matrices, the number of parameters to estimate is significantly lower than other model-based methods and this allows the methods to be stable and efficient in high dimensions. Two introductory examples illustrated with **R** codes allow the user to discover the *hdda* and *hddc* functions. Experiments on simulated and real datasets also compare HDDC and HDDA with existing classification methods on high-dimensional datasets. *HDclassif* is a free software and distributed under the general public license, as part of the **R** software project.

The **R** package *HDclassif* (currently in version 1.2) implements these two classification methods for the clustering and the discriminant analysis of high-dimensional data. The package is available from the CRAN at *http://CRAN.R-projects.org/package=HDclassif*.

Références

[1] L. Bergé, C. Bouveyron and S. Girard, *HDclassif : an R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data*, Journal of Statistical Software, vol. 42 (6), pp. 1-29, 2012.