Unsupervised classification in high dimension
Didier Fraix-Burnet, Charles Bouveyron, Stéphane Girard, Julyan Arbel

To cite this version:

HAL Id: hal-01569733
https://hal.archives-ouvertes.fr/hal-01569733
Submitted on 27 Jul 2017

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.

Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License
Abstract

Dealing with large databases of galaxy spectra is a good example of a new problematic task in astrophysics. Current and forthcoming big surveys provide millions of spectra each containing thousands of wavelengths. These spectra must be confronted with physical and chemical models. This requires an unsupervised classification which is a dimensionality reduction in both the number of observations and parameters. In this poster, we present some approaches that we are implementing.

The SDSS galaxy and quasar spectra

The spectra of 702 248 galaxies and quasars with redshift smaller than 0.25 were retrieved from the Sloan Digital Sky Survey (SDSS) database, release 7 (http://www.sdss.org/dr7/). There are 5740 wavelength points within the useful range of wavelengths between 3806 and 7371 Å after redshift correction. Spectra of galaxies reveal their composition and history. In the figure to the right are a few typical spectra corresponding to different categories of galaxies, devised mainly by eye.

Canopy Technique
The Canopy technique first divides the data set into overlapping subsets termed as “canopies” based on a “cheap distance measure”. In a second stage, clustering is performed by measuring exact distances only between points which belong to a common canopy. Under reasonable assumptions, appropriate selection of cheap distance metric reduces computational cost without any loss in clustering accuracy.

Fisher-EM Algorithm
The Fisher-EM algorithm (Bouveyron & Brunet 2012) estimates both the discriminative subspace and the parameters of the mixture model. It is based on the Expectation-Maximization (EM) algorithm from which an additional step, named F-step is introduced, between the E- and the M-step. This F-step uses the maximization of the Fisher’s criterion under orthonormality constraints and conditionally to the posterior probabilities.

Mars hyperspectral data
We have applied the High-Dimensional Data Clustering (HDDC) algorithm (Bouveyron et al 2007) to segment hyperspectral images of the Martian surface (each image is 300 × 128 pixels, each pixel having 255-dimensional spectral points), with a specific model and for the expected number of groups. Note that it is also possible to let the algorithm determine which model and number of groups are the most adapted for the data at hand. Regarding model parameters, HDDC estimates that the intrinsic dimensions of groups are all around 10 whereas, for recall, the original dimension is 255. The figure below shows the associated segmentation of the image and allows to compare it with an expert segmentation. Both segmentations look very similar, confirming the interest of such model-based clustering techniques in this context.

References
Bouveyron C., Girard S. & Schmid C. 2007, High-Dimensional Data Clustering, Computational Statistics & Data Analysis 52(1), 503-519
De T., Fraix-Burnet D. & Chattopadhyay A.K. 2016, Clustering large number of extragalactic spectra of galaxies and quasars through canopies, Communication in Statistics - Theory and Methods 45(9), 2638-2653