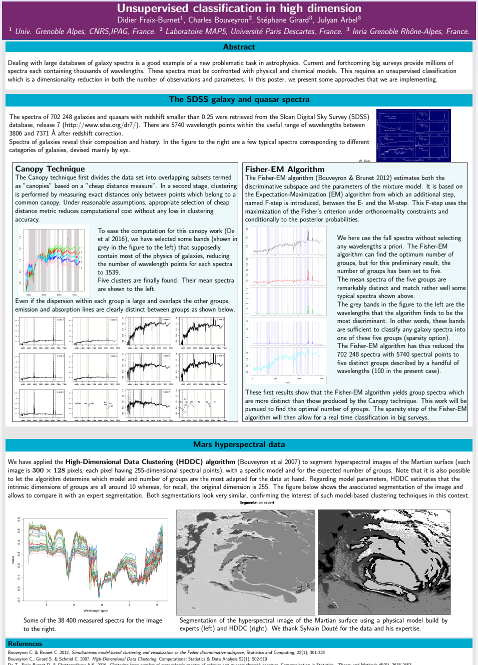
Unsupervised Classification in High Dimension (S14.8)



Unsupervised classification in high dimension

Didier Fraix-Burnet[1], Charles Bouveyron[2], Stéphane Girard[3], Julyan Arbel[3]
[1] *Univ. Grenoble Alpes, CNRS,IPAG, France.* [2] *Laboratoire MAP5, Université Paris Descartes, France.* [3] *Inria Grenoble Rhône-Alpes, France.*

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.

W. Keel

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.

To ease the computation for this canopy work (De et al 2016), we have selected some bands (shown in grey in the figure to the left) that supposedly contain most of the physics of galaxies, reducing the number of wavelength points for each spectra to 1539.

Five clusters are finally found. Their mean spectra are shown to the left.

Even if the dispersion within each group is large and overlaps the other groups, emission and absorption lines are clearly distinct between groups as shown below.

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.

We here use the full spectra without selecting any wavelengths a priori. The Fisher-EM algorithm can find the optimum number of groups, but for this preliminary result, the number of groups has been set to five.

The mean spectra of the five groups are remarkably distinct and match rather well some typical spectra shown above.

The grey bands in the figure to the left are the wavelengths that the algorithm finds to be the most discriminant. In other words, these bands are sufficient to classify any galaxy spectra into one of these five groups (sparsity option).

The Fisher-EM algorithm has thus reduced the 702 248 spectra with 5740 spectral points to five distinct groups described by a handful of wavelengths (100 in the present case).

These first results show that the Fisher-EM algorithm yields group spectra which are more distinct than those produced by the Canopy technique. This work will be pursued to find the optimal number of groups. The sparsity step of the Fisher-EM algorithm will then allow for a real time classification in big surveys.

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.

Some of the 38 400 measured spectra for the image to the right.

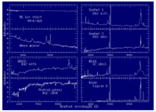
Segmentation of the hyperspectral image of the Martian surface using a physical model build by experts (left) and HDDC (right). We thank Sylvain Douté for the data and his expertise.

References

Bouveyron C. & Brunet C. 2012, *Simultaneous model-based clustering and visualization in the Fisher discriminative subspace*, Statistics and Computing, 22(1), 301-324
Bouveyron C., Girard S. & Schmid C. 2007, *High-Dimensional Data Clustering*, Computational Statistics & Data Analysis 52(1), 502-519
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

702 248 Spectra of Galaxies and Quasars

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.


W. Keel

Canopy Technique

d

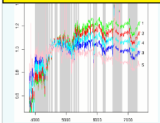
ing

a

Canopy Technique:

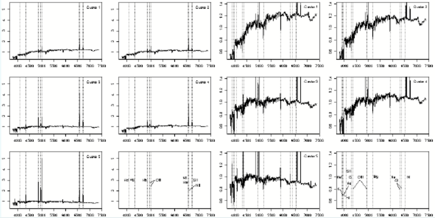
1. cheap distance clustering
2. k-means

De



et al 2010), we have selected some bands (shown in grey in the figure to the left) that supposedly contain most of the physics of galaxies, reducing the number of wavelength points for each spectra to 1539.
Five clusters are finally found. Their mean spectra are shown to the left.

Even if the dispersion within each group is large and overlaps the other groups, emission and absorption lines are clearly distinct between groups as shown below.



Fisher-EM Algorithm

The Fisher

discrimina

the Expect

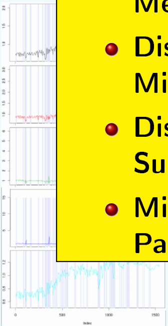
named F-s

maximizat

conditiona

Fisher-EM Algorithm

- Subspace Clustering Method
- Discriminative Latent Mixture Model
- Discriminative Subspace Estimation
- Mixture Model Parameter Estimation



702 248 spectra with 5740 spectral points to five distinct groups described by a handful of wavelengths (100 in the present case).

These first results show that the Fisher-EM algorithm yields group spectra which are more distinct than those produced by the Canopy technique. This work will be pursued to find the optimal number of groups. The sparsity step of the Fisher-EM algorithm will then allow for a real time classification in big surveys.

Hyperspectral Data of Mars Surface

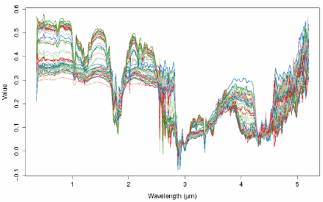
High-Dimensional Data Clustering (HDDC) algorithm

- Subspace Clustering Method
- HD-GMM

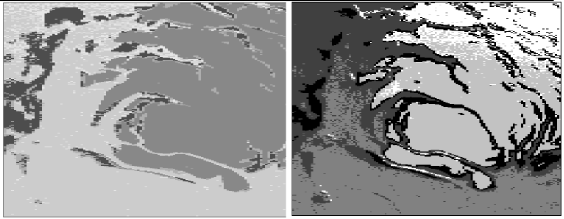
Mars hyperspectral data

We have applied t
image is 300×1
to let the algorith
intrinsic dimensior
allows to compare

ce (each
possible
t the
ge and
context.



Some of the 38 400 measured spectra for the image to the right.



Segmentation of the hyperspectral image of the Martian surface using a physical model build by experts (left) and HDDC (right). We thank Sylvain Douté for the data and his expertise.

Didier Fraix-Burnet,
Charles Bouveyron, Stéphane Girard, Julyan Arbel,
Asis Kumar Chattopadhyay