Robust clustering / kernel density estimation / visualization tools

Author: Karen Ng
Institution: UC Davis

Goals

Astronomy data usually show clustering on some scale due to gravity. This project tries to put together some code (maybe Scikit-learn code) for identifying clusters of stars / galaxies and find center of mass and density / luminosity peaks. Furthermore, it would be helpful to have some kernel density estimation tools for plotting accurate number density contours to compare against the clustering results. We want to come up with a clustering method of galaxies / stars to best

- represent the underlying structure(s) of our data
- validate our method or show it is consistent with estimates with other statistical methods

I work from with very specific goals in mind but I imagine other astronomy / astrophysics people may have similar needs for pre-processing their data for inference.

Description of my data

- we have around 20 sets of data, each for one galaxy clusters
- the galaxy clusters show signs of merging, so there exist several subclusters within each data set
- number of observations (galaxies) of each data set is of the order of 60 - 200
- number of useful variables are limited, i.e. ~ 10
- origin of our data can be heterogenous, thus there can be missing variables from observations
- no training set is available, even if we can divide up our data set into a training (observations with less missing variables) and a test set, the training set may be too small to be useful- we may have to come up with simulations to validate our method
- we want to weight the variables (or apply prior) to reflect the underlying physical properties / uncertainties

Action plan

Questions that need to be addressed

- whether we have enough data for clustering and testing hypothesis
- determine the number of subcluster(s)
- assign membership of data points into subcluster(s)
- weight our data points to represent the underlying physical properties / uncertainties
- find best way(s) to represent the spatial distributions of each set of data
- unwanted (foreground / background) data removal

variables to be included in the feature matrix

- spatial coordinates (x and y) in the plane of the sky after correcting for anarmorphic distortions (projection effects)
- redshifts or

- the corresponding velocity dispersions
- color (g-i band)

Proposed methods

proposed non-parametric methods smoothing and for determining bin-widths

- bootstrap the samples
- smooth the map of data points with Gaussian filter
- identify peaks

higher dimensional kernel density estimation (KDE)

- minimize the mean integrated squared error (MISE) to find the band width
- find the peak(s) of the estimated density

proposed method(s) for assigning membership

- bootstrap the samples, assign membership based on number density contours find the smallest band width that would allow the "signal-to-noise" within each subcluster to be higher than 3 or so
- K-means randomly assign means of subclusters, perform K-means for a number of K, find the smallest suitable K from an elbow plot
- normal mixture model, check the BIC to find suitable number of clusters

math of KDE

in 1D

$$f(x) = \frac{1}{nh} \sum_{i}^{n} K_i \left(\frac{x - x_i}{h} \right)$$

where n is the number of observations, h is the bandwidth, K is the kernel function which can be any smooth function but usually people pick a Gaussian or so