Clustering

James Long

November 10, 2015

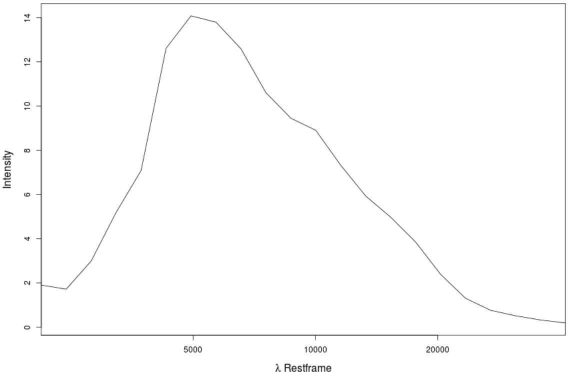
Clustering References

- **Elements of Statistical Learning** (Tibshirani, Hastie, Friedman)
  - Chapter 14.3
  - `http://statweb.stanford.edu/~tibs/ElemStatLearn/`
- **Statistics, Data Mining, and Machine Learning in Astronomy** (Ivezic, et al)
  - Section 6.4
- **Modern Statistical Methods for Astronomy** (Feigelson, Babu)
  - Sections 9.2 – 9.5

What is clustering?

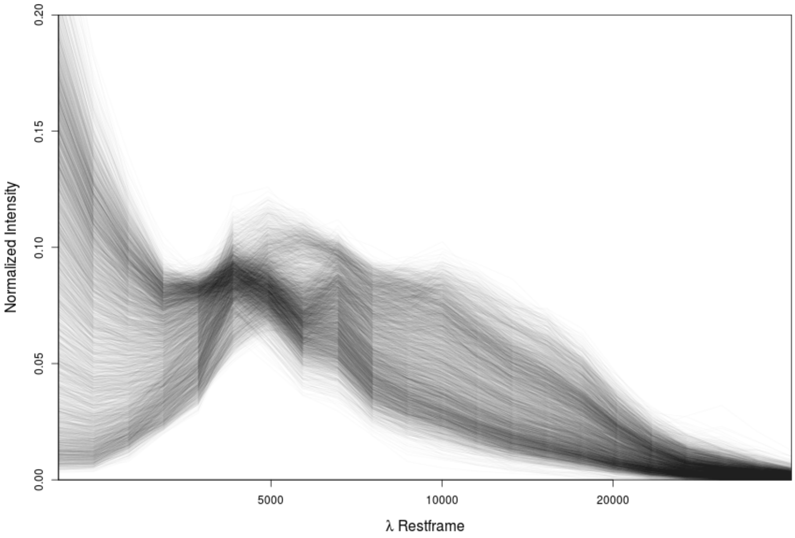
clustering: a partition of the data into sets

- objects in the same cluster (set) are "similar"
- objects in different clusters are "different"



Objects could be light curves, images, galaxy photometry.

Normalized Rest Frame Synthetic Photometry



Notation, Data Dimension, and Clustering

- $X \in \mathbb{R}^{n \times p}$
  - n is number of observations (galaxies)
  - p is number of variables / features
  - $x_i \in \mathbb{R}^p$ is i^{th} observation
- p is called the dimension of the data.
- Clustering methods useful for "high" dimensional ($p > 3$) data where we do not have a priori have idea of structure.

Types of Clustering Methods

- Dissimilarity (distance) based
  - Compute dissimilarity between every pair of objects.
  - Similar objects in same cluster, dissimilar objects in different clusters.
- Model based
  - Construct (mixture) model and estimate parameters.
  - Object belongs to component in mixture.
  - eg mixture of Gaussians
- Centroid based
  - Find cluster centers (centroids).
  - Object belongs to closest centroid.
  - eg. k–means

Generic Dissimilarity (Distance) Measures

Let $x_{i\lambda}$ be the flux at filter λ for observation i.

Squared Euclidean Dissimilarity:

$$d(x_i, x_j) = \sum_{\lambda} (x_{i\lambda} - x_{j\lambda})^2$$

More generally:

$$d(x_i, x_j) = \sum_{\lambda} |x_{i\lambda} - x_{j\lambda}|^p$$

Even more general:

$$d(x_i, x_j) = \sum_{\lambda} w(\lambda) |x_{i\lambda} - x_{j\lambda}|^p$$

Note: The log scale implicitly imposes a weight w.

Building Invariances into Dissimilarity Measures

A galaxy identical to x_i but at a different (physical) distance will have flux ax_i where a is some constant. Therefore we should choose d such that

$$d(x_i, x_j) = d(ax_i, bx_j)\,\forall a, b \tag{1}$$

One possibility is

$$d(x_i, x_j) = \sum_{\lambda} \left(\frac{x_{i\lambda}}{\sum_\lambda x_{i\lambda}} - \frac{x_{j\lambda}}{\sum_\lambda x_{j\lambda}} \right)^2$$

Or simply normalize rest frame SEDs

$$x_i \to \frac{x_i}{\sum_\lambda x_{i\lambda}}$$

Kriek 2011 Dissimilarity

$$d(x_i, x_j) = \sqrt{\frac{\sum_\lambda (x_{i\lambda} - a_{12} x_{j\lambda})^2}{\sum x_{i\lambda}^2}}$$

where

$$a_{12} = \frac{\sum x_{i\lambda} x_{j\lambda}}{\sum x_{j\lambda}^2}$$

- d satisfies invariance relation (1).
- $d(x_i, x_j)$ are contained in `AS689_b.dat`.

Other Ideas for Dissimilarity

- Derivatives (synthetic photometry is functional data)
- Extract “features”, compute distances in feature space
- Dynamic Time Warping (distance in x,y space)
- Invariances to errors in photometric redshift

Dissimilarity Based Clustering Methods

- Kriek 2011
- Hierarchical agglomerative
- Hierarchical divisive
- See references for other methods.

Kriek 2011 Clustering Method Pseudocode

- $N \leftarrow \{1, \ldots, n\}$
- $d_{ij} \leftarrow d(x_i, x_j) \ \forall \ i, j \in N$
- $K \leftarrow 0$
- **repeat:**
  - $A_i \leftarrow \{j : d_{ij} < 0.05, \ j \in N\} \ \forall \ \ i \in N$
  - $c \leftarrow \underset{i}{\operatorname{argmax}} \ \#(A_i)$
  - **if** $\#(A_c) < 19$:
    - **break**
  - $K \leftarrow K + 1$
  - $C_K \leftarrow \{x_j : j \in N \cap A_c\}$
  - $N \leftarrow N \backslash A_c$

$C_1, \ldots, C_K$ are the clusters. Some objects are unclustered.

Hierarchical Agglomerative Clustering Idea

Main Idea:

- Every observation starts as own cluster.
- Iteratively merge "close" clusters together.
- Iterate until one giant cluster left.

This method is

- **Hierarchical:** Each iteration produces a clustering, so do not specify number of clusters in advance.
- **Agglomerative:** Initially every observation in own cluster.

Hierarchical Agglomerative Clustering Pseudocode

- $N \leftarrow \{1, \ldots, n\}$
- $d_{ij} \leftarrow d(x_i, x_j) \; \forall \; i, j \in N$
- $C_{in} \leftarrow \{x_i\} \;\; \forall i \in N$
- **for** $k = n, \ldots, 2$**:**
  - $i, j \leftarrow \underset{\{i,j: i<j, \; i,j \in N\}}{\text{argmin}} \; d_C(C_{ik}, C_{jk})$
  - $C_{i(k-1)} \leftarrow C_{ik} \cup C_{jk}$
  - $C_{l(k-1)} \leftarrow C_{lk} \; \forall l \neq i, j$ and $l \in N$
  - $N \leftarrow N \backslash \{j\}$

The $C_{\cdot k}$ are the k clusters in the k^{th} level of the hierarchy.

How to Merge Clusters (What is d_C?)

- Average Linkage

$$d_C(C_i, C_j) = \frac{1}{\#(C_i)\#(C_j)} \sum_{x \in C_i} \sum_{x' \in C_j} d(x, x')$$

- Complete Linkage

$$d_C(C_i, C_j) = \max_{x \in C_i, x' \in C_j} d(x, x')$$

- Single Linkage

$$d_C(C_i, C_j) = \min_{x \in C_i, x' \in C_j} d(x, x')$$

Constructing a Dendogram

- At iteration k

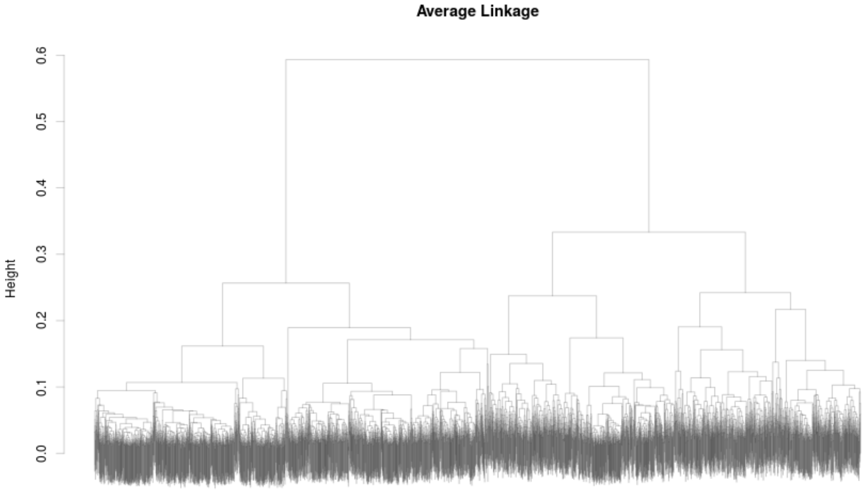
$$i, j \leftarrow \underset{\{i,j: i<j, i,j \in N\}}{\operatorname{argmin}} d_C(C_{ik}, C_{jk}).$$

- The “height” of this cluster merger is

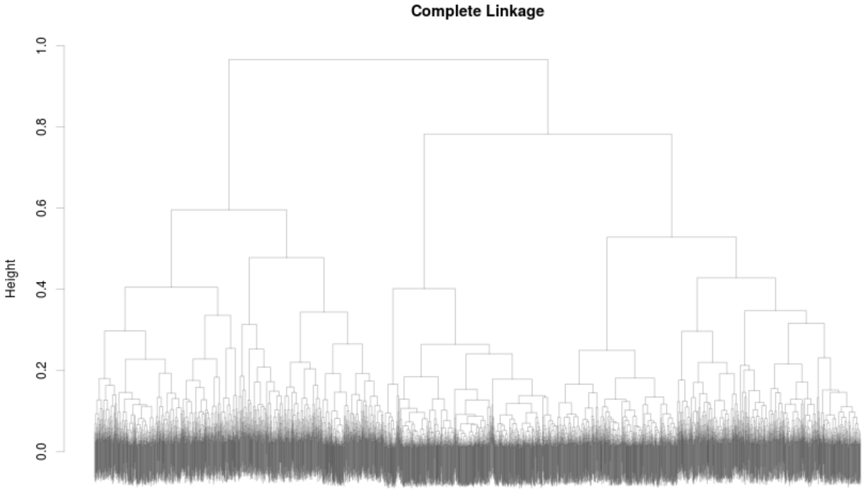
$$h_k = d_C(C_{ik}, C_{jk})$$

- The sequence $h_n, \ldots, h_2$ is monotonically increasing.
- Plot with heights of cluster mergers is a **dendogram**.

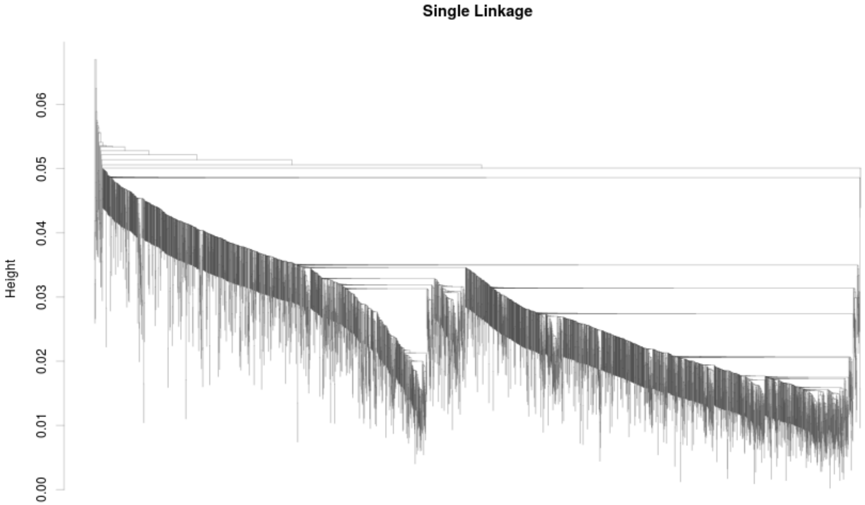
Average Linkage



Complete Linkage



Single Linkage



Number of Clusters, Quality of Clustering

- Quantification of success in classification is (relatively) objective and easy.
- Quantification of success in clustering is more subjective.
  - General measures output by clustering method.
    - Cophenetic distance.
    - Confusion matrix to compare clustering methods.
  - Application specific measures.
    - Scatter in composites.
    - Physical interpretation of clusters.

Cophenetic Distance

- The ordinary distance between x_i and x_j is

$$d_{ij} = d(x_i, x_j)$$

- Suppose x_i and x_j first share cluster C_{lk} ie $x_i, x_j \in C_{lk}$, $x_i \in C_{m(k+1)}$, $x_j \in C_{q(k+1)}$, $C_{m(k+1)} \neq C_{q(k+1)}$. The *cophenetic distance* between x_i and x_j is

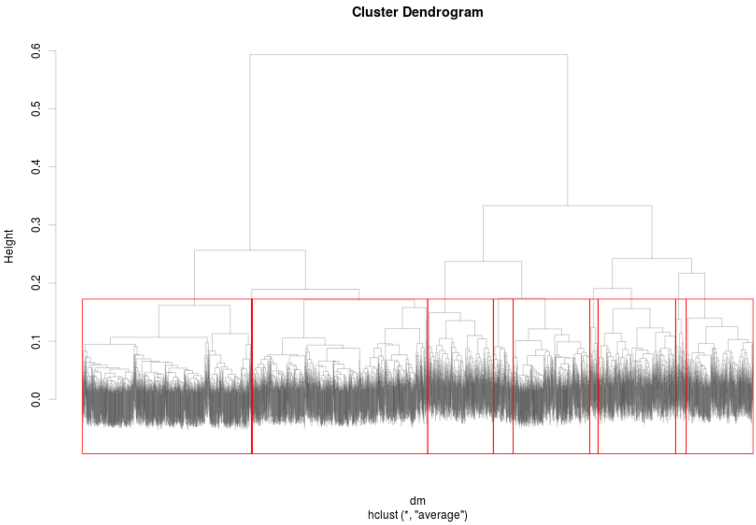
$$d_{ij}^C = d_C(C_{m(k+1)}, C_{q(k+1)})$$

- The *cophenetic correlation coefficient* is

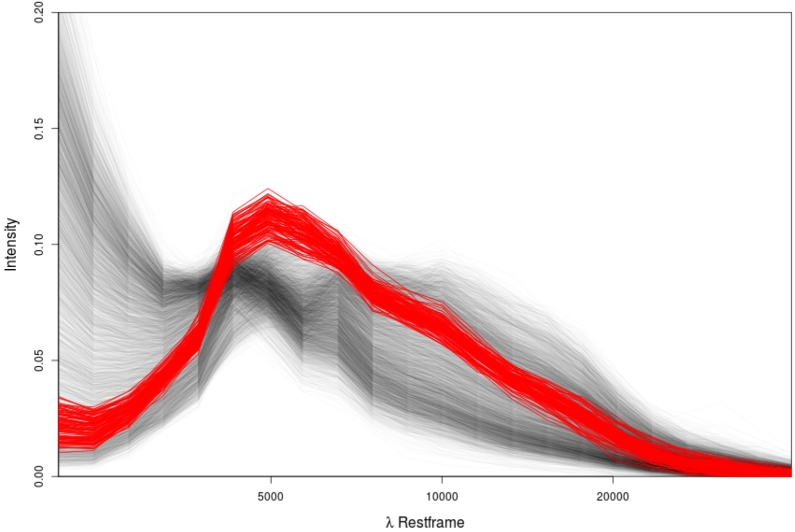
$$corr(d_{ij}, d_{ij}^C)$$

- For average linkage clustering cophenetic correlation is 0.81.

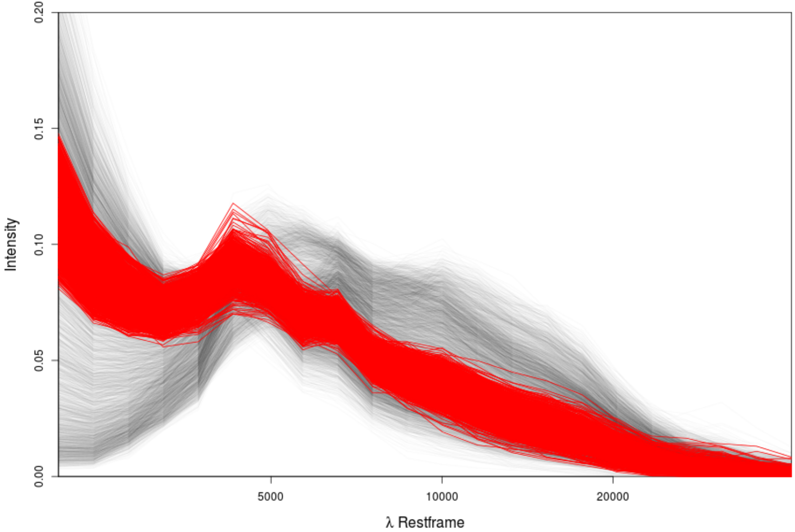
Visualize 10 Clusters for Average Link



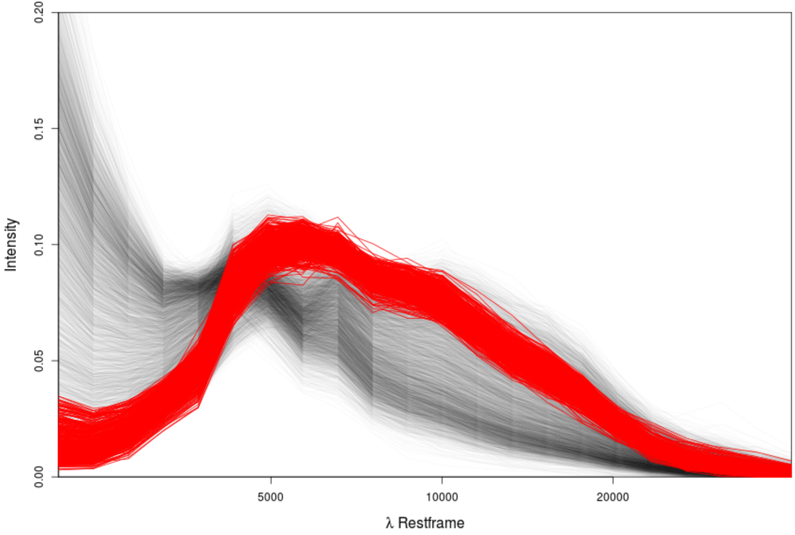
Cluster



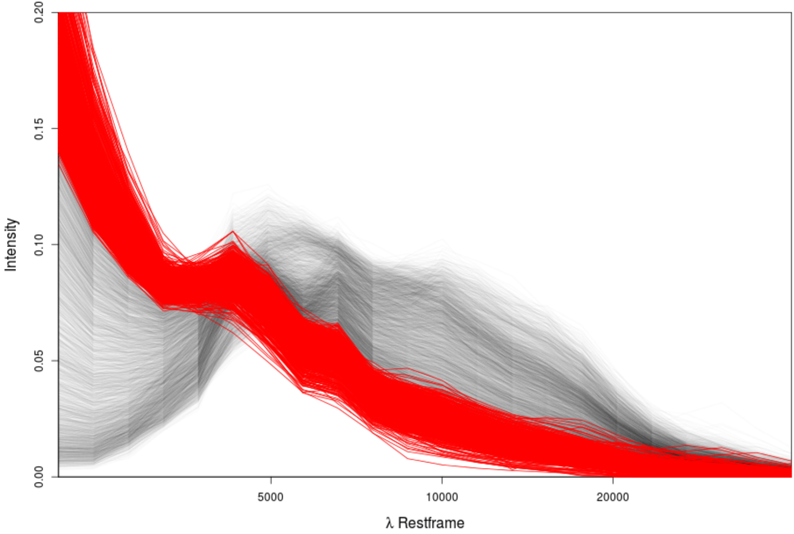
Cluster



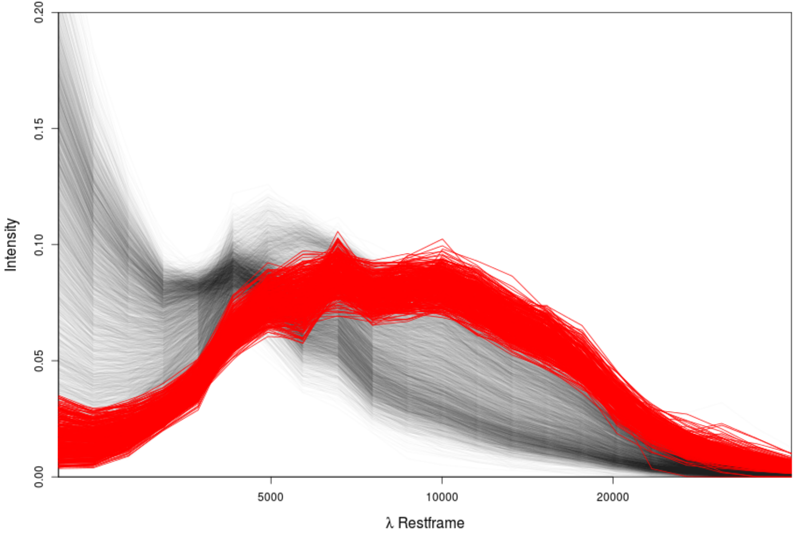
Cluster



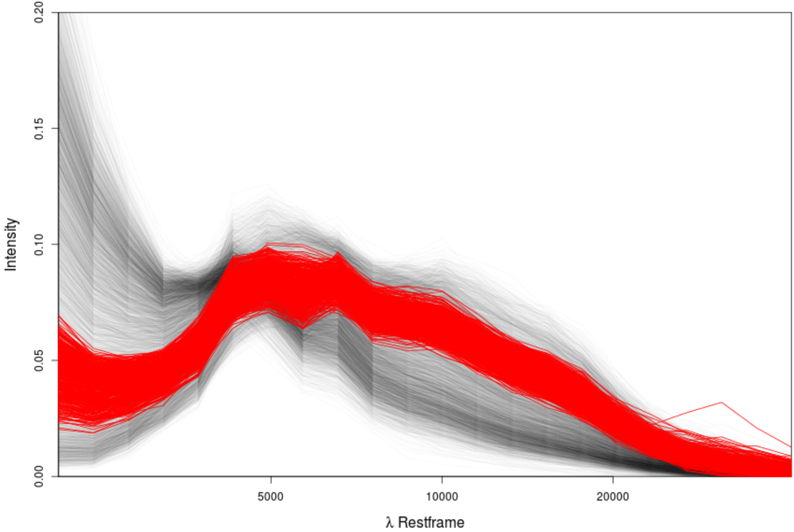
Cluster



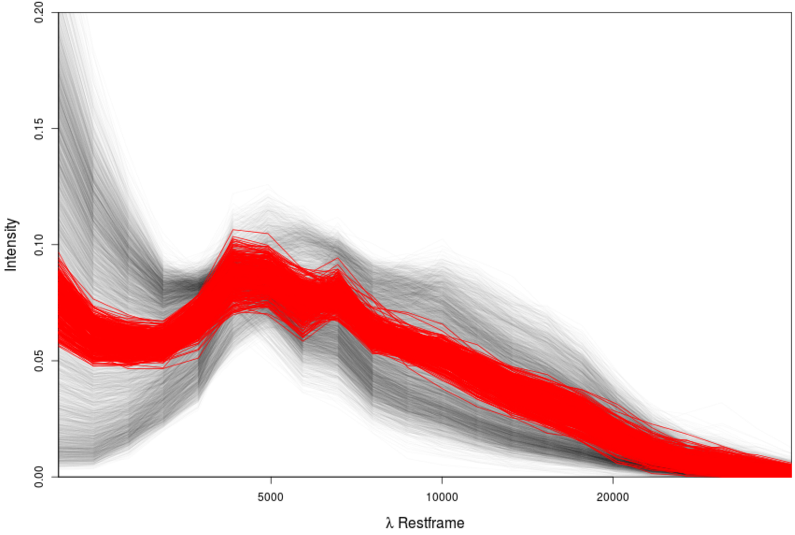
Cluster



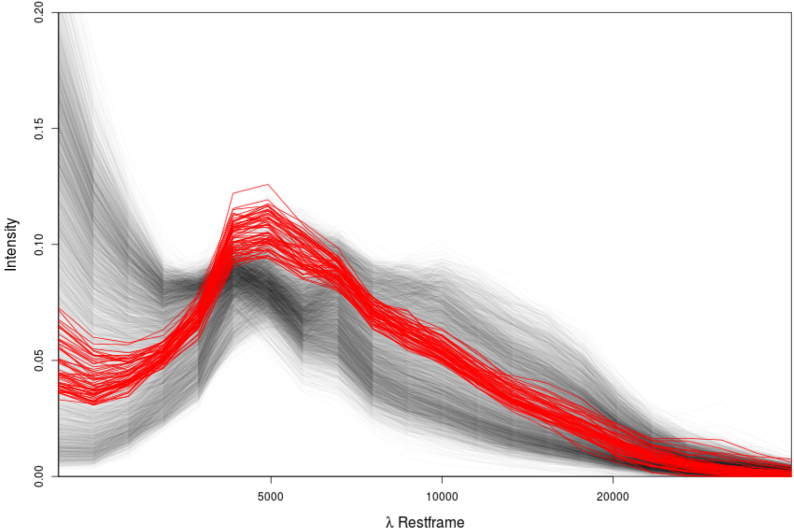
Cluster



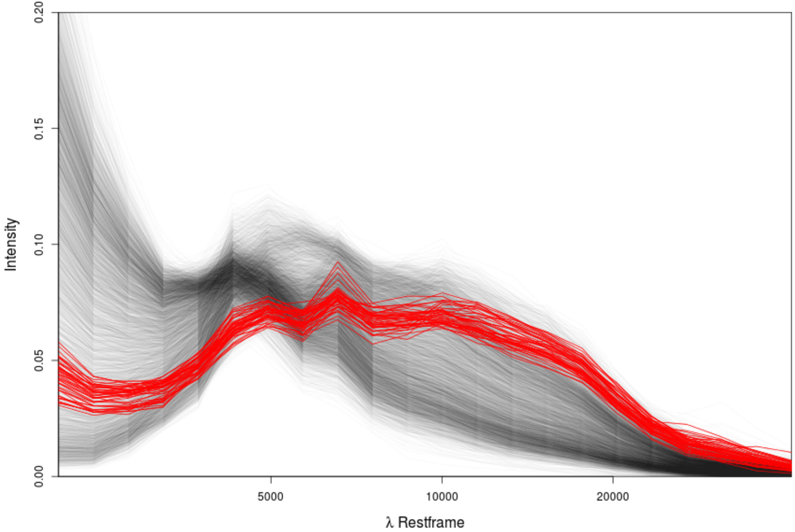
Cluster



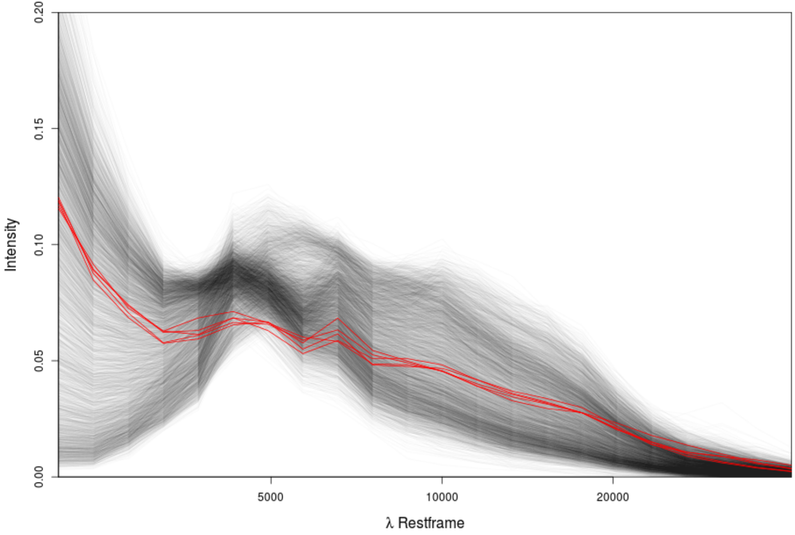
Cluster



Cluster



Cluster



Is Clustering the Right Tool?

- Photometry lies on some low dimension linear subspace:
  - Principal Components Analysis
- Photometry lies on some low dimension non-linear subspace:
  - Principal Curves and Surfaces
  - Local Linear Embedding
  - Self Organizing Maps
- Model the photometry:

$$x_i(\lambda) = g_{\theta_i}(\lambda)$$
$$\theta_i \in \mathbb{R}^d$$
$$\theta_i \sim f_\theta \;\; iid$$