Multivariate clustering

Astro 585 Spring 2013

Astronomical context

Astronomers have constructed classifications of celestial objects for centuries:

- Asteros (fixed stars) vs. planetos (roving stars) [Greece, >2Kyr ago]
- Luminosae, Nebulosae, Occultae [Hodierna, mid-17th c]
- Comet orbits & morphology [Alexander 1850, Lardner 1853, Barnard 1891]
- Stellar spectra: 6 classes (Secchi 1860s), 7 classes (Pickering/Cannon 1900-20s), 10 classes w/ brown dwarfs (Kirkpatrick 2005)
- Variable stars: 6+5 classes (Townley 1913), ~80 classes (Samus 2009)
- Galaxy morphology: 6+3 classes (Hubble 1926)
- SupernovaeL Ia, Ib, Ic, Iib, IIP, Iin (Turatto 2003)
- Active galactic nuclei: Seyfert gal, radio gal, LINERs, quasars, QSOs, BL Lac, blazars
- Gamma ray bursts: short, long, intermediate (Kouveliotou 1993; Mukherjee 1998)
- Protostars/PMS stars: Class 0, 0/I, I, II, III (Lada 1992, Andre 1993)

In nearly every case, these classes were created by well-argued, but subjective assessment of source properties.

In statistical parlance, the problem is called *unsupervised clustering of multivariate data*

Traditional clustering methods in astronomy

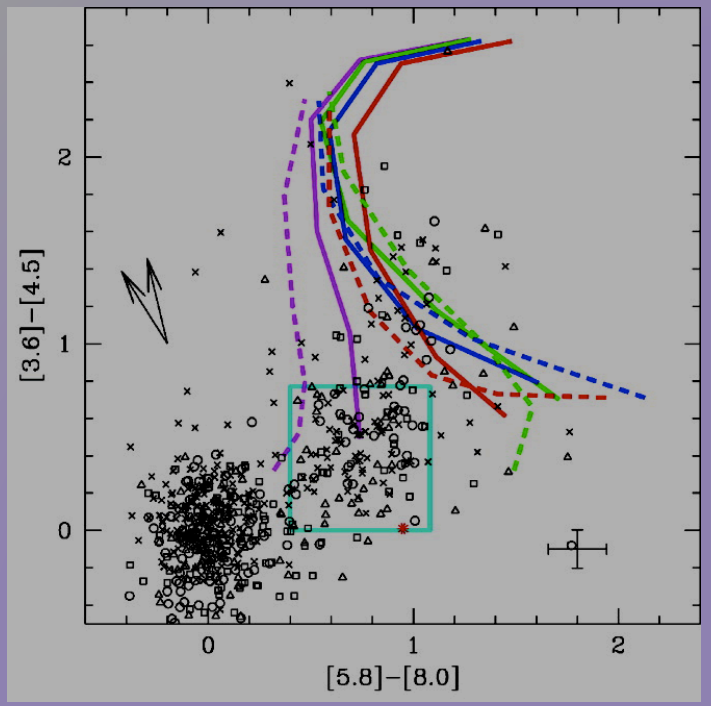
Histogram-like univariate criteria

- Abell galaxies with 80-129 galaxies within 2 magnitude of their third brightness member are classified as *Richness class 2* (Abell 1958)
- Young stellar objects with infrared colors 0.4<[5.8]-[8.0]<1.1 and 0.0<[3.6]-[4.5]<0.8 are classified as *Class II* (Allen 2004)
- Seyfert galaxies are distinguished from starburst emission line galaxies by a curve on a 4-emission-line diagram (Veilleux & Osterbrock 1987)

TABLE 4

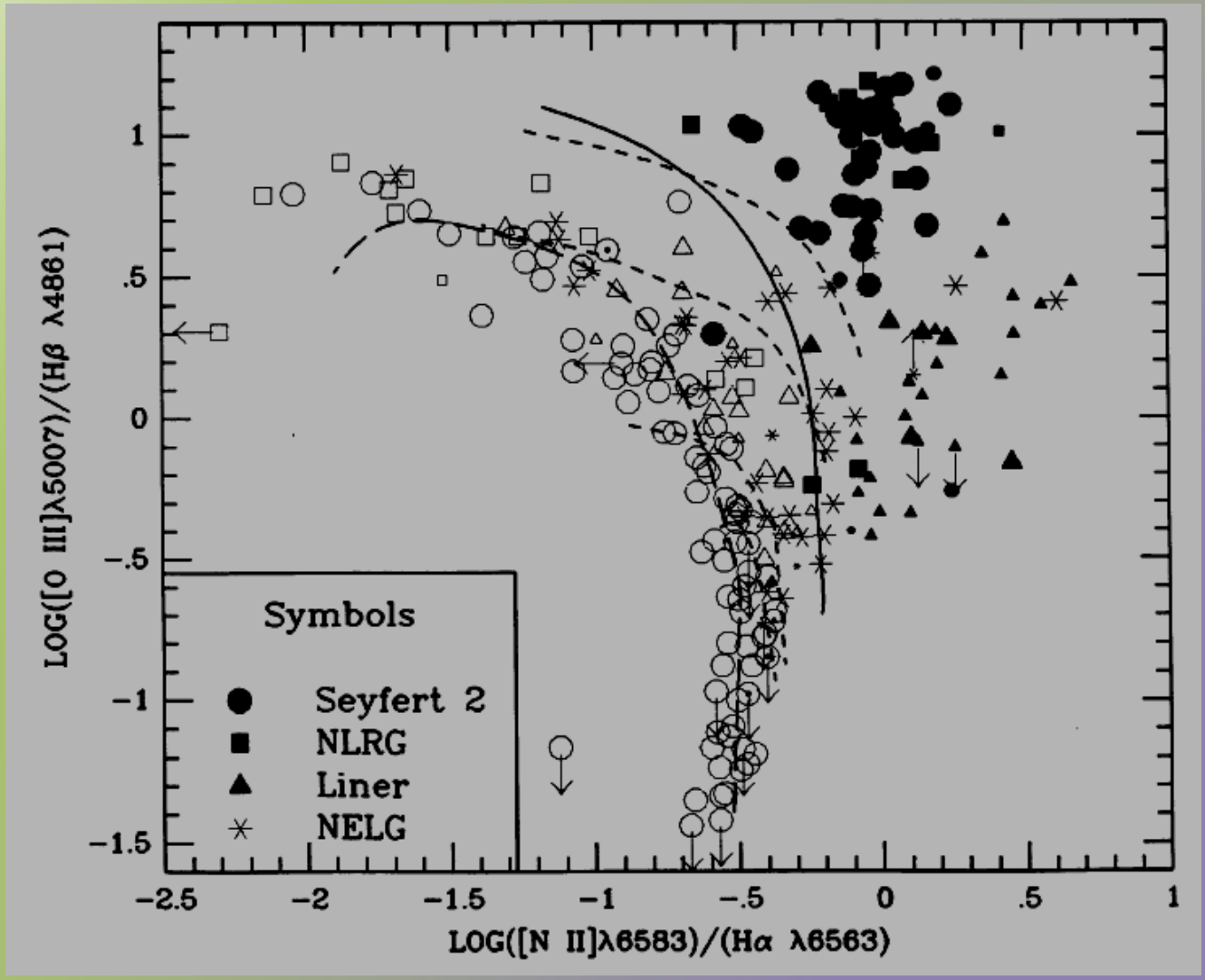
RICHNESS-GROUP INTERVALS

Richness Group	Counts	Richness Group	Counts	Richness Group	Counts
0	30–49	2	80–129	4	200–299
1	50–79	3	130–199	5	300 or over



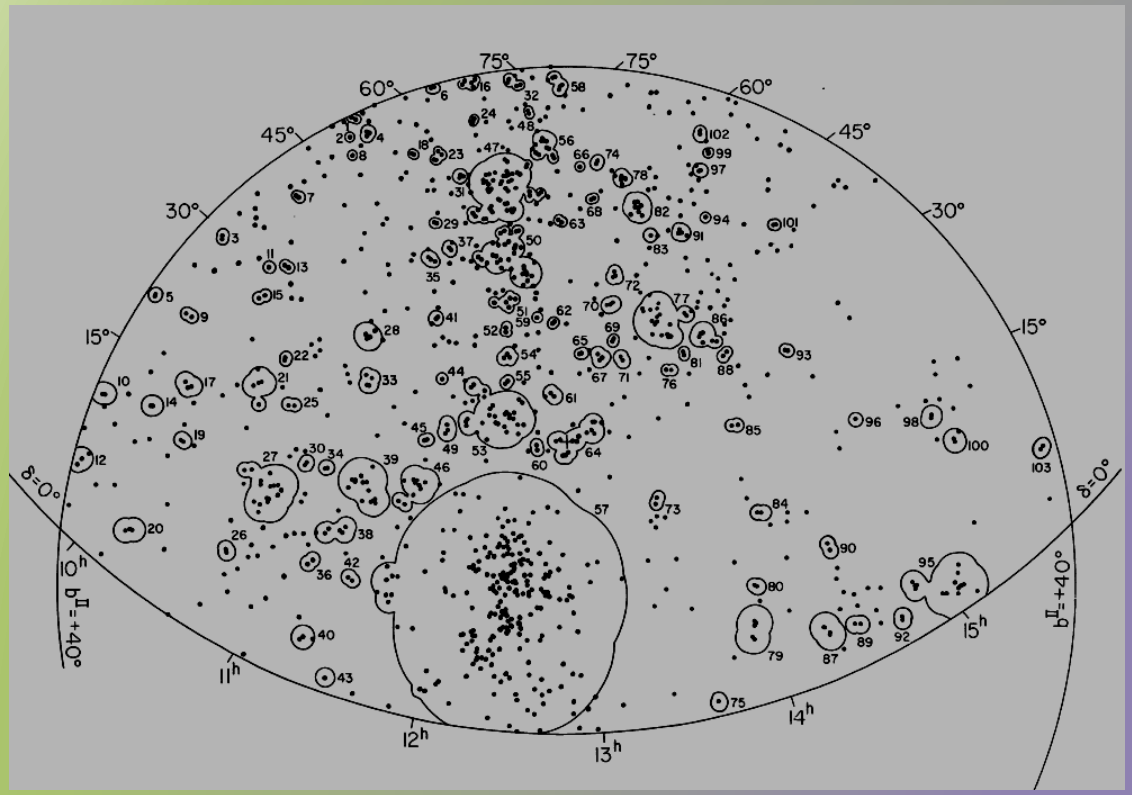
In statistics and computer science, these are `deterministic decision trees'

As with astronomical histograms, the choice of class boundaries are typically chosen heuristically without mathematical calculation



Percolation or `friends-of-friends' algorithm

1. Plot data points in a 2-dimensional diagram
2. Find the closest pair, and call the merged object a `cluster'
3. Repeat step 2 until some chosen threshold is reached. Some objects will lie in rich clusters, others have one companion, and others are isolated.



Turner & Gott
Groups of Galaxies I
A Catalog ApJS 1976

In statistics, this is `single linkage hierarchical clustering'

Statistical approach to unsupervised clustering

In **unsupervised clustering** of a multivariate *nxp* dataset, the number, location, size and morphology of the data groupings is unknown. There is no `prior knowledge' of classes.

Nonparametric clustering algorithms:

- Agglomerative hierarchical clustering
- K-means partitioning
- Density-based clustering

Parametric clustering algorithms:

- Normal mixture models
- Special class on Friday: Isothermal ellipsoid mixture models for young star clusters (Mike Kuhn, PSU)

Nonparametric unsupervised clustering is a very uncertain enterprise, and different algorithms give different outcomes without mathematical guidance (e.g. there is no likelihood to maximize or stopping criterion to choose number of clusters). Results should be viewed with great caution for scientific inference.

Parametric unsupervised clustering lies on a stronger foundation (e.g. there is a likelihood to maximize, and BIC/AIC for model selection). But it assumes the clusters in fact follow the chosen parametric form.

Agglomerative hierarchical clustering

1. Construct the distance matrix $d(\mathbf{x}_i, \mathbf{x}_j)$ for the dataset, assuming a distance metric (e.g. Euclidean with standardized variables). Call each point a `cluster'.
2. Merge two clusters with the smallest `distance'. Several common choices for measuring the `distance' between a cluster and a new data point:

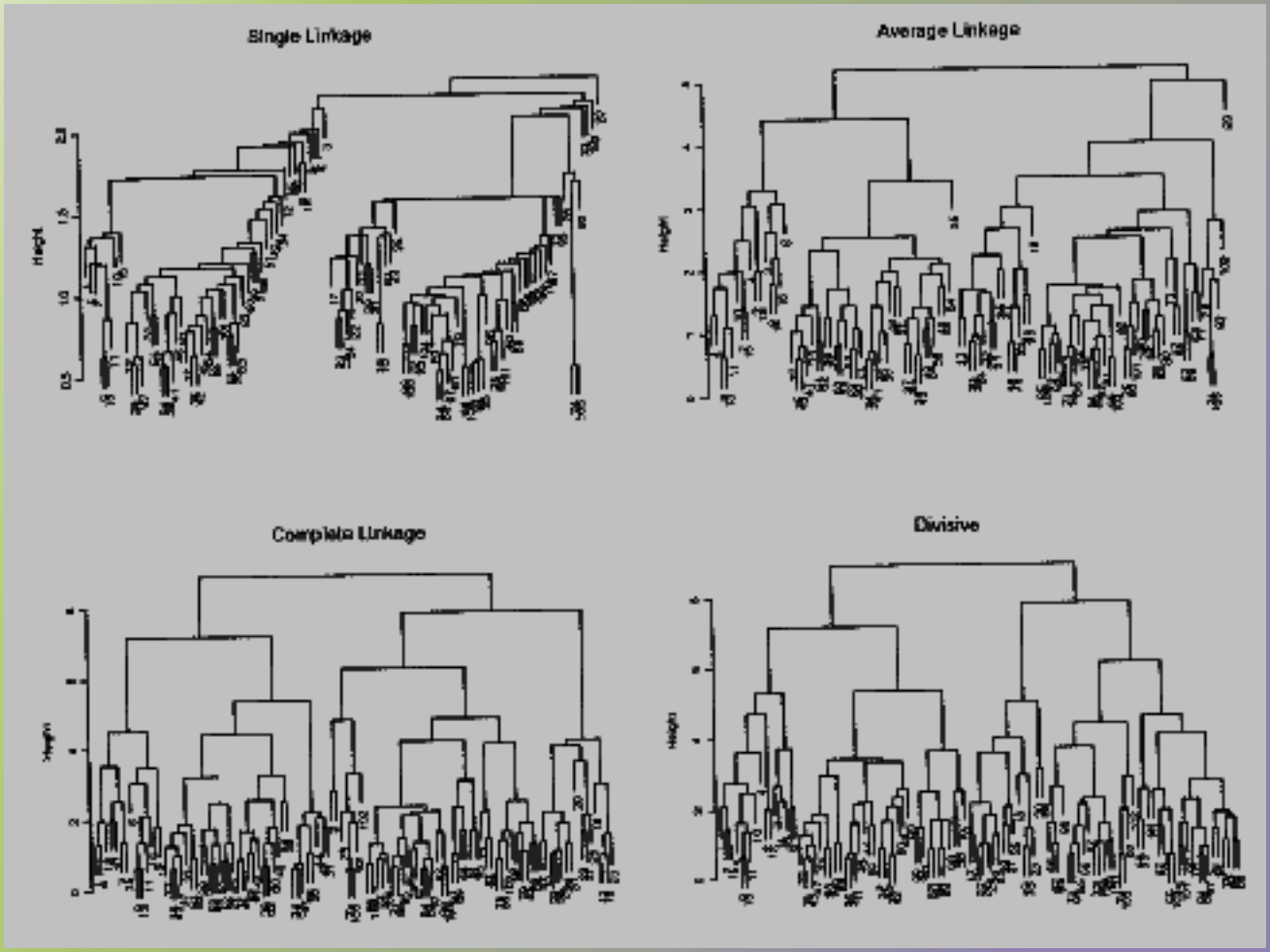
   - Minimum distance between any constituent point of the cluster and the new point = **single linkage clustering**. This procedure is equilvalent to `pruning' the **minimal spanning tree** of the multivariate dataset. This is the astronomers' friends-of-friends or percolation algorithm. This method is vulnerable to spurious `chaining' of distinct clusters into elongated superclusters, and is ***not recommended*** by statisticians.

   - Average distance between the constituent points of the cluster and the new point = **average linkage clustering.** This often gives an intermediate outcome but is scale-depenent.

   - Maximum distance between any constituent point of the cluster and the new point = **complete linkage clustering**. This is a conservative procedure that tends to give hyperspherical clusters.

   - Minimize the intra-cluster variances (**W** matrix) = **Ward's minimum variance clustering**

The result of an agglomerative (or divisive) clustering procedure is a dendrogram, or tree, showing the membership of each cluster at each stage of the clustering. ***There is no mathematical basis for choosing where to cut the tree, and thereby establishing the true number of clusters present.*** Qualitatively, objects combined at greater `heights' in the dendrogram are more dissimilar.



Comparison of hierarchical clustering methods

Primate scapular shapes
N=105, p=7

A. J. Izenman
Modern Multivariate Statistical Techniques
(2008)

k-means partitioning

An important non-hierarchical approach to multivariate clustering …

- Choose k, the number of cluster suspected to be present
- Select k seed locations in p-space
- Iteratively (re)assign each object to nearest cluster to minimize **W**
- Recalculate cluster centroid after objects are added or subtracted
- Stop when no reassignments are made

✓Often run with different k and different seeds
✓Pairwise distance matrix not calculated or stored … good for large datasets
✓An NP-hard problem, but computationally efficient approximations
✓Several computational procedures: Lloyd's algorithm (Voronoi iteration), EM algorithm, kd-trees, Linde-Buzo-Grey algorithm, …
✓Important variant: **k-medoid** partitioning (to reduce outlier effects)

Very important in data mining, computer vision, and other computationally intensive problems

Density-based clustering

Computer scientists and statisticians have developed methods for cases familiar to astronomers: where several clusters may exist in a background of unclustered objects. The methods assign objects either to clusters or to an unclustered background. Unlike k-means, these methods do not need a specified number of clusters.

Friedman & Fisher (1998) **bump hunting (PRIM)**: progressively shrink hyperrectangles to increase the enclosed density of points with a preset minimum population. Remove box from the dataset, and repeat. User interaction permitted. Related to CART.

DBSCAN (density-based spatial clusters of applications with noise) by Ester et al. (1996). User specifies minimum population and maximum extent (`reach') of a cluster. A cluster is expanded based on a single-linkage criterion.

Others include **BIRCH, DENCLUE, CHAMAELEON, OPTICS, …**

Normal mixture models

These are parametric regression models where the multivariate dataset is assumed to consist of *k* multivariate normal (MVN) clusters.

Each cluster has a hyperellipsoidal morphology extending over the entire space with mean vector μ_j and covariance matrix Σ_j where j=1, …, p.

The model has *2kp+k+1* parameters: *k* means and *k* variances in *p* dimensions, *k* mixture weights, and *k* itself.

Parameters are estimated by MLE using the EM Algorithm:

- Seed values of k, μ_j , and lower bound to Σ_j must be provided
- E step: Calculate likelihood of each object lying in each cluster
- M step: Cluster μ_j and Σ_j are updated with weighted objects
- Iterate EM until likelihood (or MAP for Bayesian) is maximized
- Run for different *k* with model selection from BIC or bootstrap

Other techniques include use of **W** and **B** matrices, robust procedures, and penalties for roughness

Codes include EMMIX, MCLUST, and AutoClass
(http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass, Cheesman & Stutz 1995)