Multivariate analysis and classification I

Eric Feigelson

IMAV 2013

Multivariate problems in astronomy

Any tabular dataset is a multivariate dataset:

Rows are stars, planets, galaxies, clouds, AGNs, GRBs, protoplanetary disks, interstellar clouds, radio/X-ray/gamma-ray sources, photons, and so forth

Columns are RA, Dec, radial velocities, proper motions, fluxes at multiple wavebands, polarization fraction, color indices, spectral indices, interstellar absorption, masses, densities, surface temperatures, emission/absorption line strengths, elemental abundances, variability/periodicity timescales, ellipticities, star/galaxy classifications, spectral types, morphological types, variability classes, SED class, evolutionary stages, mineralogical types, flare types, orbital classes, and so forth

Thousands of astronomical studies involve multivariate data,
yet astronomers use only a narrow scope of multivariate statistical techniques

Goals of multivariate analysis

- ✧ Central location and density of data in p-space
  - Multivariate mean & variance, multivariable normal, density estimation, outliers
- ✧ Structural simplification
  - Linear combinations of related variables, removal of noise variables
- ✧ Dependencies between variables
  - Correlation, regression, principal components, parametric fits
- ✧ Testing hypotheses
  - k-sample tests, goodness-of-fit tests
- ✧ Clustering and classification into distinct groups
  - Thursday tutorial
- ✧ Visualization

Classification in astronomy

Astronomers have constructed classifications of celestial objects for centuries:

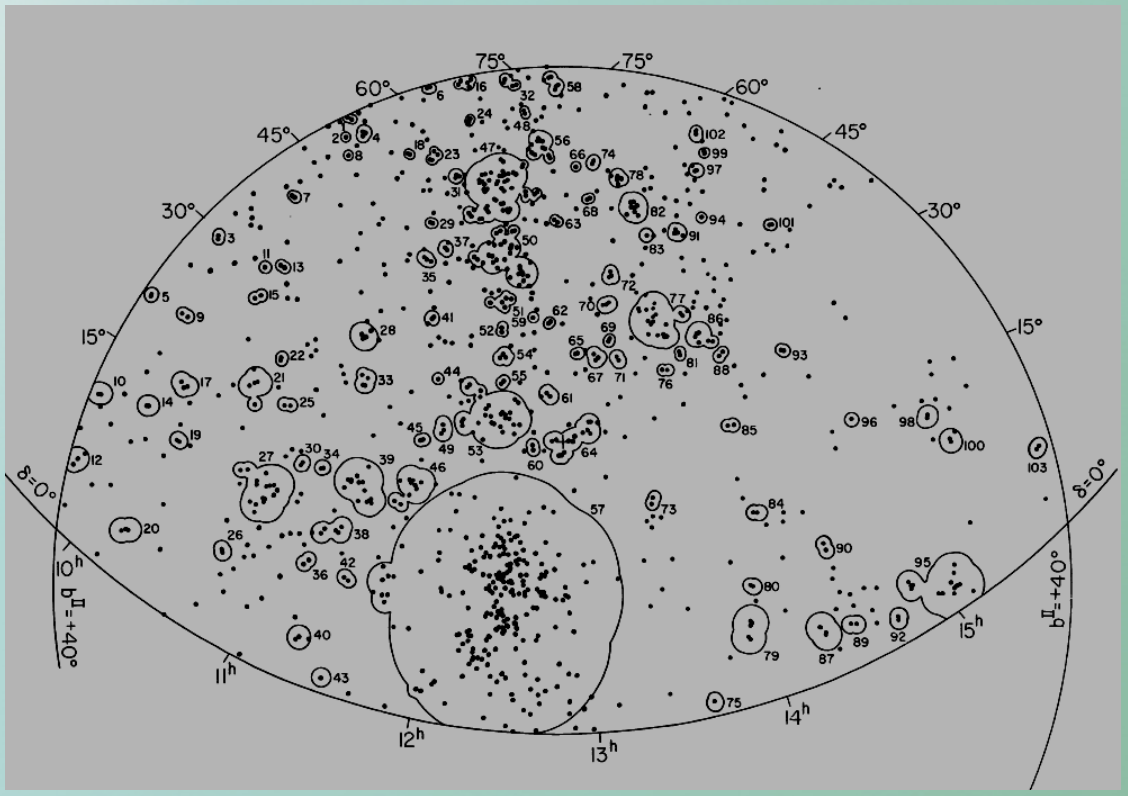
- Asteros (fixed stars) vs. planetos (roving stars) [Greece, >2Kyr ago]
- Luminosae, Nebulosae, Occultae [Hodierna, mid-17th c]
- Comet orbits & morphology [Alexander 1850, Lardner 1853, Barnard 1891]
- Stellar spectra: 6 classes (Secchi 1860s), 7 classes (Pickering/Cannon 1900-20s), 10 classes w/ brown 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*

Percolation or `friends-of-friends' algorithm

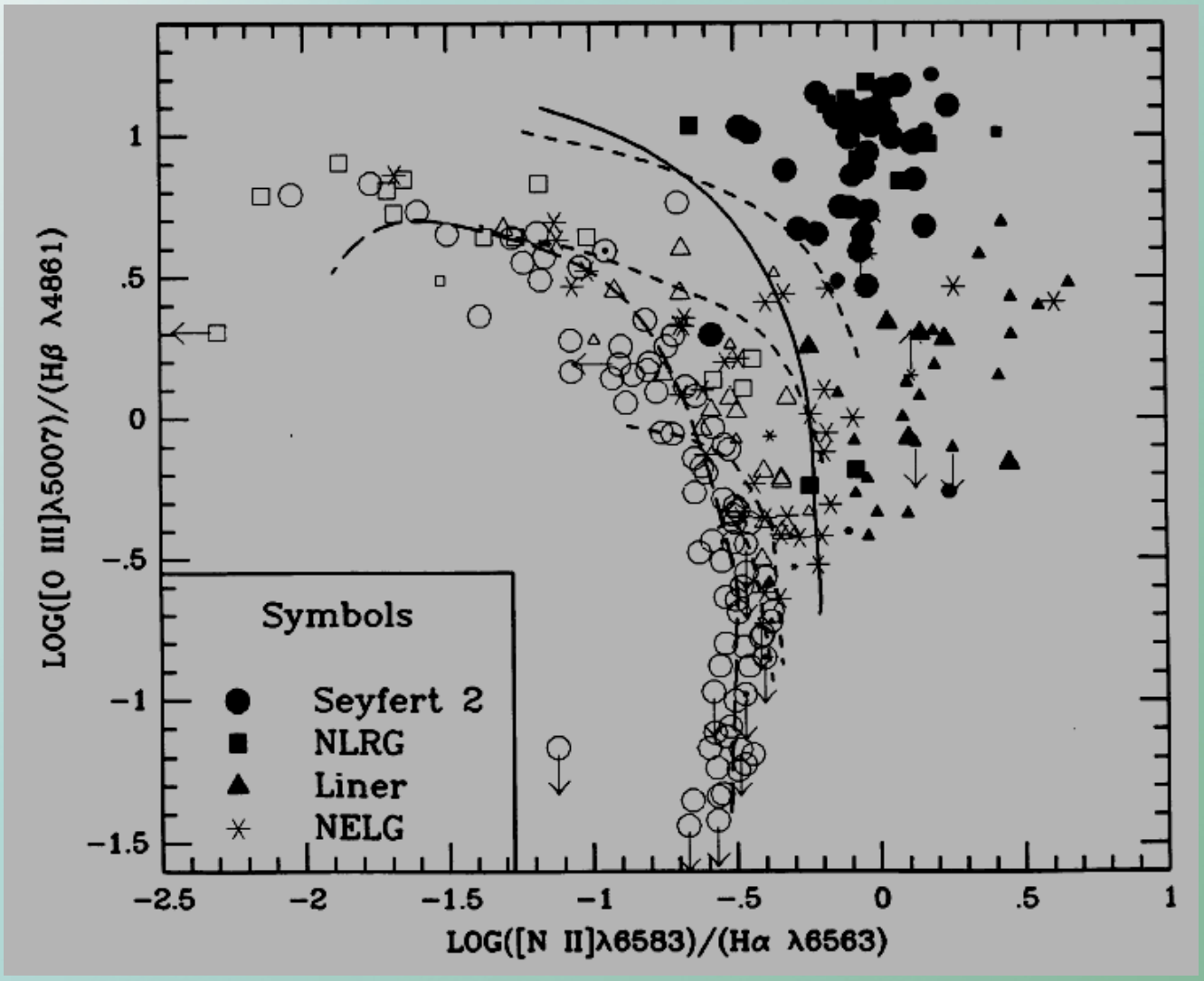
1. Plot data points in a 2-dimensional diagram
2. Find the closest pair, and call the merged object a `cluster'
3. Repeat step 2 until some chosen threshold is reached. Some objects will lie in rich clusters, others have one companion, and others are isolated.



Turner & Gott
Groups of Galaxies I
A Catalog ApJS 1976

In statistics, this is `single linkage hierarchical clustering'

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



Spectral classification of emission-line galaxies

Veilleux & Osterbrock 1987

1300 citations

Basic mathematics of multivariate analysis

X_{ij} is the measured value of the j-th variable (j=1,2,…,p) for the i-th object (i=1,2, …,n).

X is the n x p matrix whose elements are X_{ij}.

The mean and variance of the j-th variable are

$$\bar{X}_j = n^{-1} \sum_{i=1}^{n} X_{ij}$$

$$S_{jj} = n^{-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2.$$

The elements of the **variance-covariance matrix** measure the linear association between the j-th and k-th variables:

$$s_{jk} = \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k).$$

For large-n and i.i.d. variables, the sample covariance matrix **S** converges to the population covariance matrix **Σ**.

Choice of distance metric

Many methods of multivariate analysis require a measure of the `distance' between two point in p-space, d_{ij}. When all of the variables have the same physical unit (e.g. parsecs, km/s, magnitude, keV) then the choice of a Euclidean distance is clear:

$$d_{ij} = \sum_{k=1}^{p} \sqrt{(X_{ik} - X_{jk})^2}$$

A more general formulation is the m-norm distance

$$d_{ij} = d(\mathbf{X}_i, \mathbf{X}_j) = \left(\sum_{k=1}^{p} |X_{ik} - X_{jk}|^m \right)^{1/m}$$

Note however that these distances change with arbitrary choice of units. A common solution is to use standardized variables which, in the multivariate context, is the important Mahalanobis distance (Euclidean distance scaled by the covariance matrix)

$$D_M(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)' \mathbf{S}^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}.$$

Whatever metric is chosen, the matrix **D** with elements d_{ij} is commonly called the distance, dissimilarity or proximity matrix.

Choice of a distance metric is often not obvious and can critically affect the results of a multivariate analysis

Multiple linear regression

The dataset consists of a vector of response variable measurements y_i and a matrix of independent variable measurement x_{ij} where i=1,2,…,n data points and j=1,2,…,p variables. The model has a p-vector of parameters β plus a variance of the noise term ε:

$$Y = \mathbf{X}\beta + \epsilon \qquad Y \sim \mathrm{N}_p(\mathbf{X}\beta, \sigma^2\mathbf{I})$$

The MLE best-fit parameters and their distribution are:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y. \qquad \hat{\beta} \sim \mathrm{N}_p(\beta, (\mathbf{X}'\mathbf{X})^{-1})$$

Problems arise when many variables are collinear and the X'X matrix becomes singular. Various solution are used: ridge regression `regularizes' the matrix as $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}$, and removal of variables that weakly contribute to the

Model selection for multivariate regression is more complicated than for bivariate regression, as one must both optimize functional complexity and the choice of variables:

- Variable selection can be attempted through **stepwise regression** where *F* tests or information criteria are used to test which variables are to be included or excluded.

- **LASSO** regression and **Least angle regression (LARS)** are newer techniques analogous to stepwise regression, with constraints based on L_1 (absolute deviations) in addition to L_2 (least squares).

- The **BIC (Bayesian), AIC (Akaike), FIC (Focused)** and **DIC (Deviance) Information Criteria** are penalized measures of likelihood used to compare models. A more traditional measure for least-squares methods is **Mallow's C_p**, a penalized sum-of-squared residuals.

Principal components analysis (PCA)

Here the covariance structure of a multivariate dataset is modeled with linear combinations of the variables, assuming MVN distributions, without choosing one as the response variable.

PCA both allows study of linear relationships between variables, and gives structural simplification through reduction of dimensionality. PCA is the most commonly used multivariate method in astronomy.

PCA can be remarkably effective in combining colinear variables, reducing the importance of noise variables, and simplifying a p-dimensional problem to a k<<p dimensional problem. Sometimes, but not always, the PCA coefficients (**loadings**) can be interpreted astrophysically.

Algebraic viewpoint

The first PC is the linear combination $PC_1 = \sum_{i=1}^{n} a_{1i}\mathbf{X}_i$ with the largest variance of any linear combination under the constraint $a_1' a_1 = 1$. The second component is similar assuming $a_2' a_1 = 0$ so the components are uncorrelated. This method is also known as the **Karhunen-Loeve transform.**

Matrix viewpoint

The first PC is the eigenvector associated with the largest eigenvalue of the sample covariance matrix. The k-th PC accounts for the fraction of the total variance equal to the k-th eigenvalue normalized by the trace of ***S***.

Geometric viewpoint

PCA is thus the iterative selection of axis rotations in p-space to maximize the variance along the new axes.

Regression viewpoint

PCA is a generalization of Pearson's orthogonal regression to multivariate data. It thus gives a relationship symmetric with respect to all of the variables.

Limitations of PCA

- Solution depends on variable scalings or standardization
- No mathematical guidance to number of significant components. Heuristic graphical procedures are used.
- Traditional PCA is limited to linear relationships among MVN variables. Robust, nonlinear and other variants are available.

Nonlinear multivariate methods

This has been an active area of mathematical research since the 1970s with various approaches involving manifold topology, geodesic distances, and spectral embedding.

Methods include:

- Multivariate Adaptive Regression Splines (MARS)
- Projection pursuit and independent component analysis (ICA)
- Kohonen's self-organizing map (SOM) algorithm
- Principal curves and principal manifolds
- Neural networks
- Isomap, diffusion maps, elastic maps, etc

Principal curves: a generalization of PCA using a chosen functional family of curves applied to kernel-smoothed data. Also used for classification.

Projection pursuit: p-space data are projected onto lower-dimensional spaces designed to highlight `interesting' departures from normality.

MARS: fits local polynomials around chosen knot locations with cross-validation to select model complexity

Neural networks: Establishes ≥ 2 layered relationships between variables to predict outcomes