Learning from Data: Visualisation

Amos Storkey, School of Informatics

October 13, 2005

http://www.anc.ed.ac.uk/~amos/lfd/

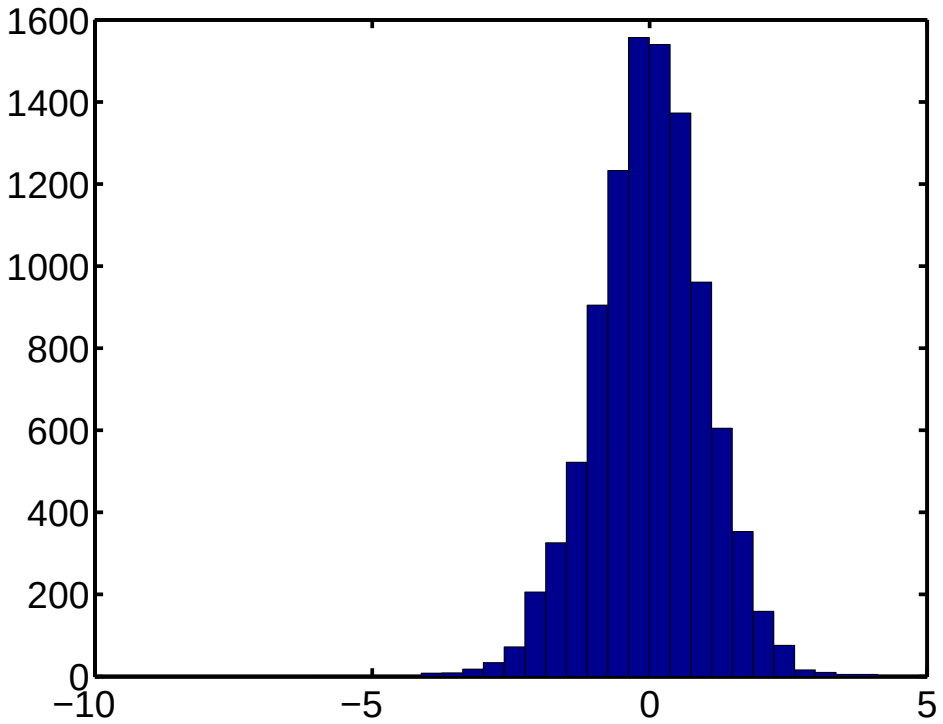
Visualisation

- Presented with new data you should
  - Try to acquire knowledge of how it was created.
  - Visualize the data to see what is in it.
- Visualization is important for understanding what issues there might be with the data, and what forms of assumptions are going to be invalid.
- Visualization is an informal assessment of the data using high level modelling concepts.

Histogram

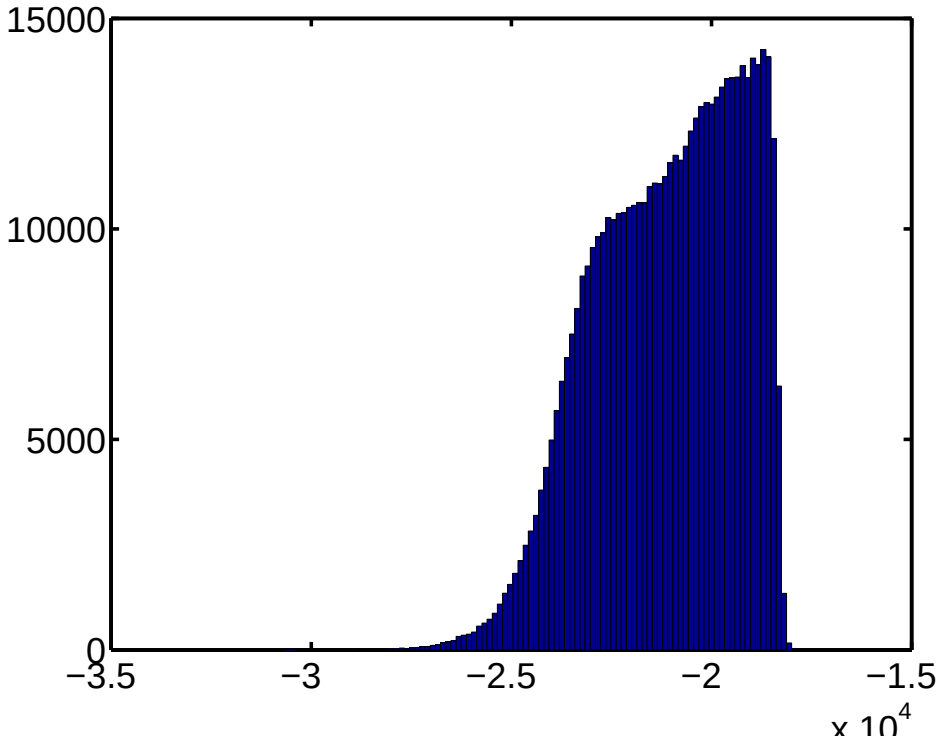
- What sort of distribution does each attribute of the data have?
- Is it normal?
- Does it have heavy tails?
- Is it skewed?
- Does it cluster?

Histogram



- Pretty much Gaussian.
- Tails a bit heavy?

Histogram

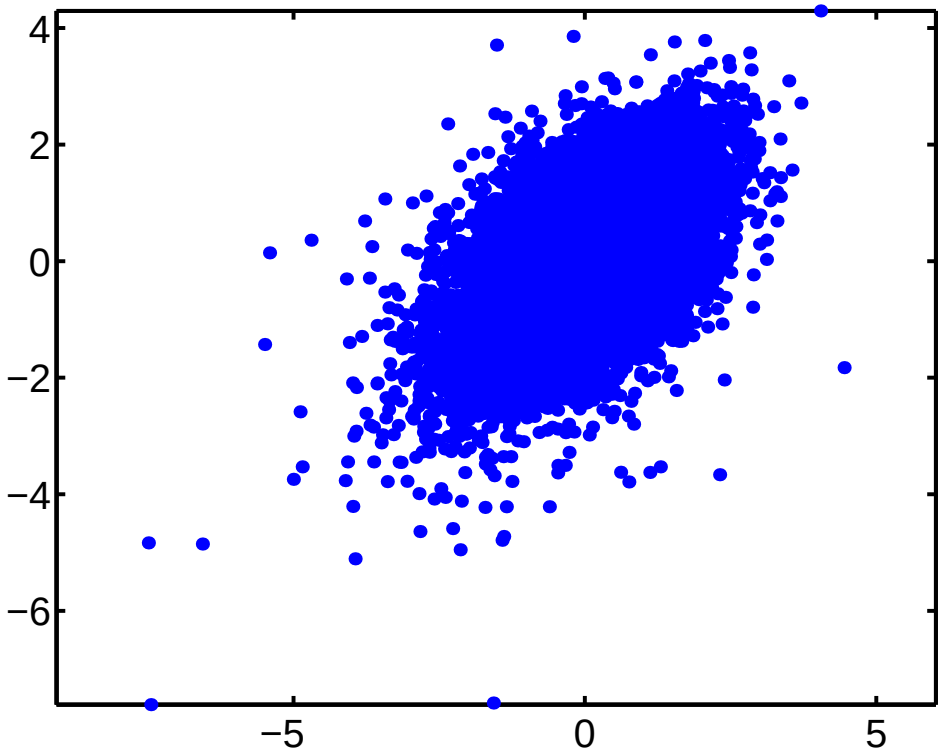


- Very skewed.
- Some sort of cutoff.
- This is the magnitude of objects in a particular region of sky - more faint objects than very bright ones, but there is a detection limit.

Plots

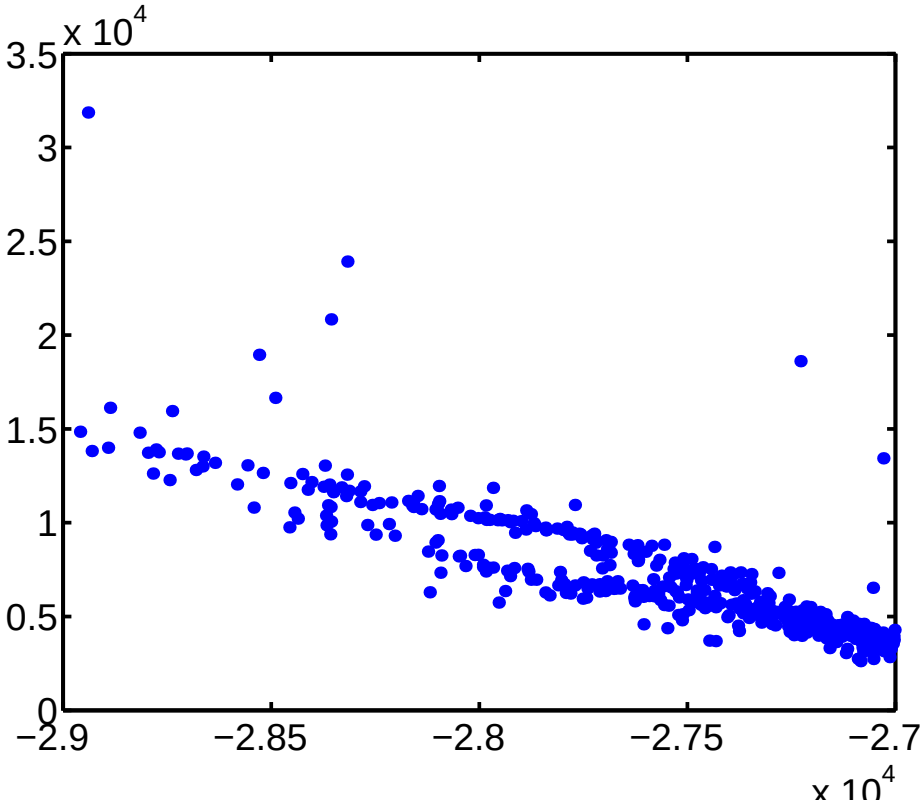
- Plot one attribute against another
- Look at the joint distribution
- Do they appear dependent or independent?
- Are they uniform, peaked, clustered?

Plots



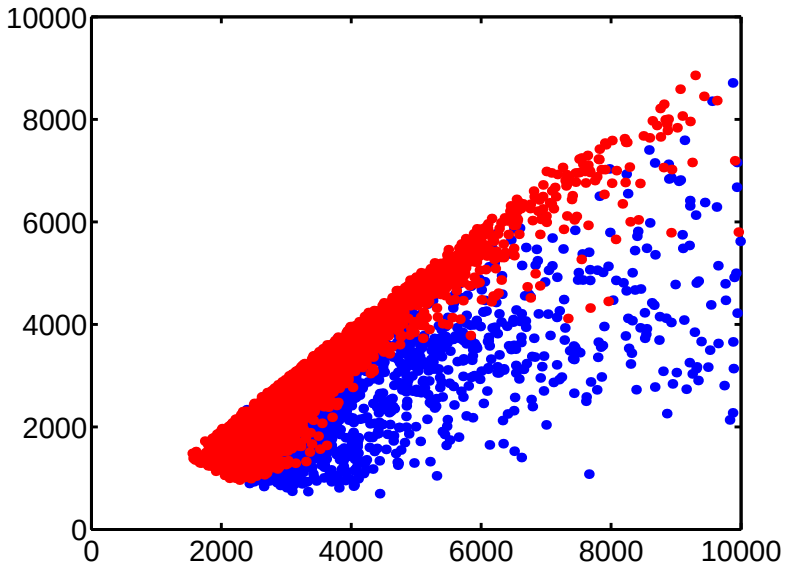
- Some dependence
- Positive correlation
- Centred around mean, cannot rule out Gaussianity.

Plots



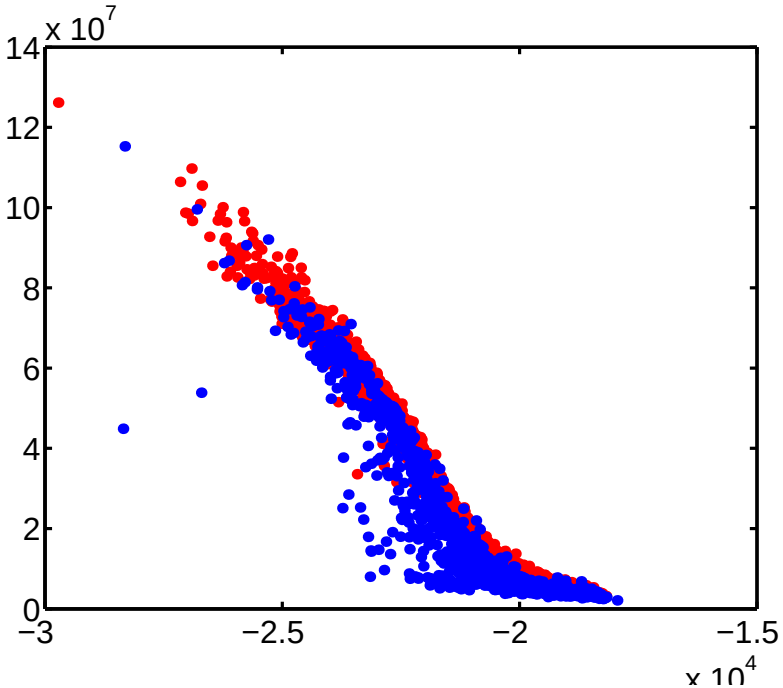
- Plot of magnitude (x axis) against area on sky (y axis).
- Definite dependence
- Two clusters, Some outliers

Class Conditional Plots



- Major elliptical axis (x) versus minor elliptical axis (y) for stars (red) and galaxies (blue).
- Galaxies are more likely to be highly elliptical.
- Two axes definitely related.

Class Conditional Plots

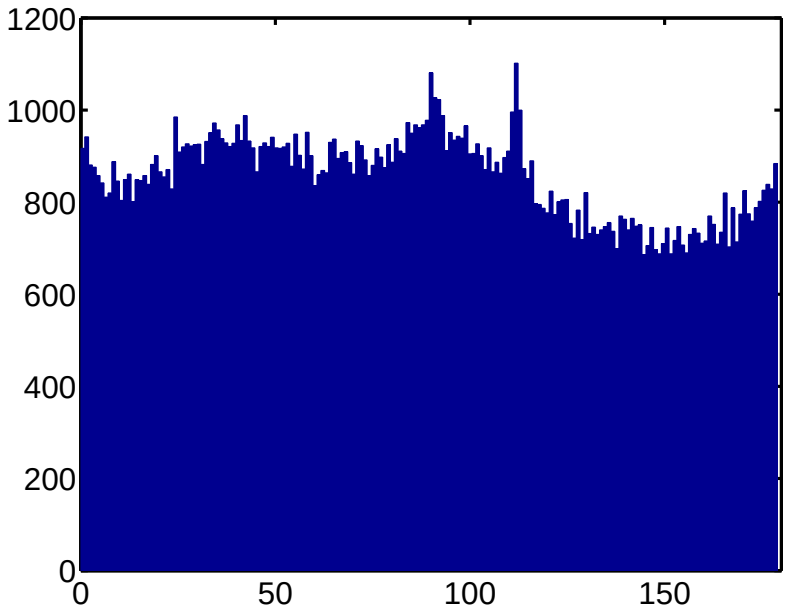


- Magnitude (x) versus peak pixel brightness (y)- stars (red), galaxies (blue).
- Stars are more “point like” than galaxies. Higher peak brightness for given magnitude.
- Already have two relevant measures for classifying stars and galaxies.

Anomaly Detection

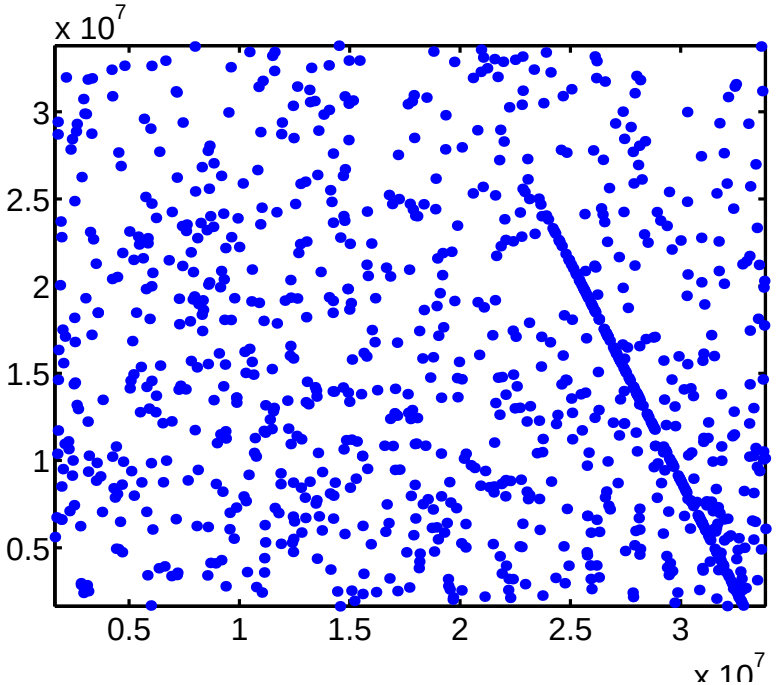
- Pay attention to outliers, or unusually high peaks in histograms.
- Restrict the data set to those outliers or peaks.
- Attempt to see if there is a potential explanation for them
- You might need to remove outliers to do other analysis

Anomaly Detection



- Histogram of galaxy orientation
- A few peculiar peaks.
- Restrict data to those within the peak regions.

Anomaly Detection

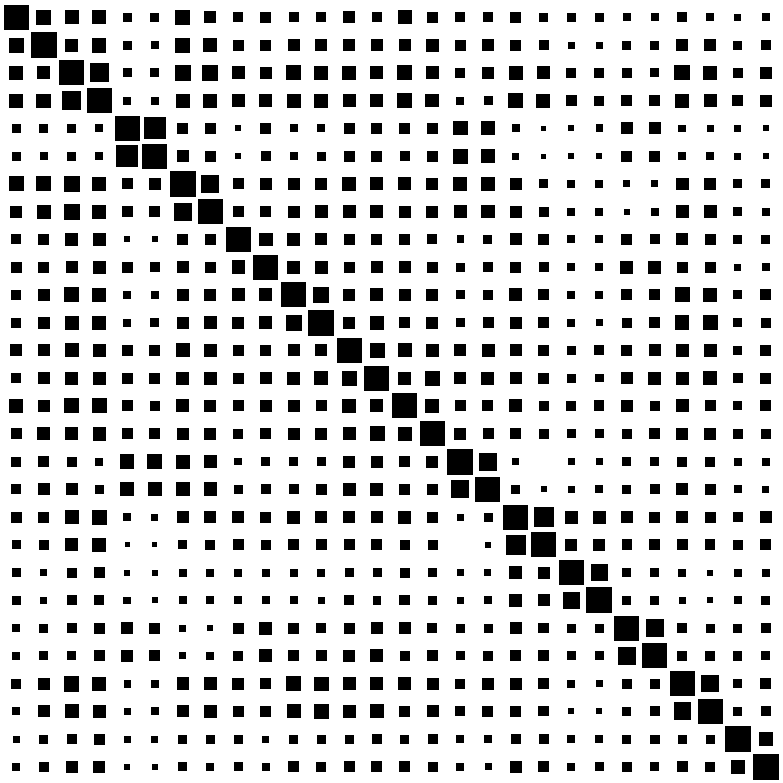


- Plot of x-position versus y-position on photograph.
- A long line of points. Wouldn't expect that of galaxies.
- In fact a satellite track which the detection program has mis-recognised.

Covariance Visualisation

- Try to get some idea regarding what variables are dependent.
- Especially appropriate with approximately Gaussian data.
- Do we have positive or negative correlations?
- Is there some structure to the correlations?

Covariance Visualisation



- Correlation visualisation via Hinton Diagram
- Consecutive pairs correlated.
- Some blockiness.

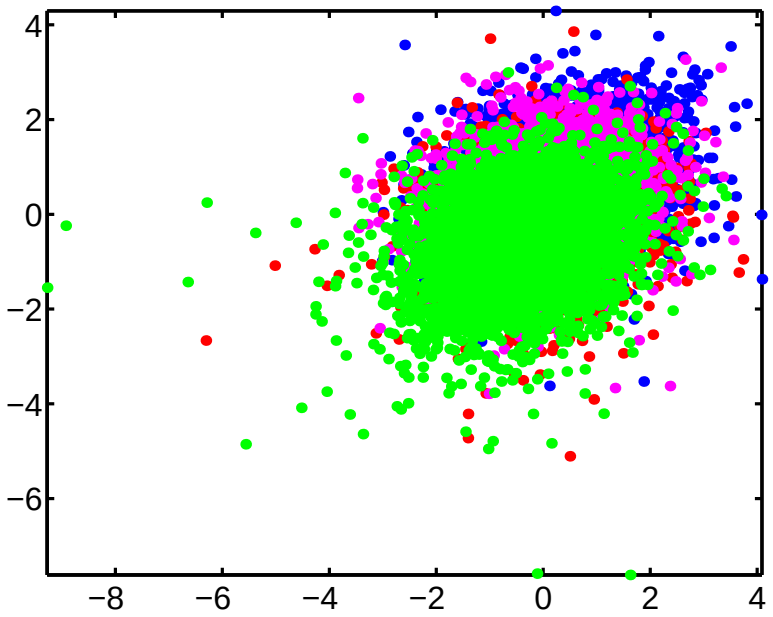
PCA

- What are the main linear components in the data set. Do they capture any intrinsic concepts?
- Are there any clusters if the data are plotted onto the principal components?

Multi-Attribute Plots

- Turn continuous atrributes into classes. Plot using different colours.
- Eg C1 ($x > 0, y > 0$), C2 ($x < 0, y < 0$), C3 ($x > 0, y < 0$), C4 ($x < 0, y > 0$).
- Plot other attributes conditioned on class.
- Look for class conditional differences.

Multi-Attribute Plots



- Some variation with colour change.

Summary

- Do not fail to look at your data.
- Real data is always messy. There are going to be things which mess it up.
- Try to identify dependencies.
- How can you reduce the data size: PCA? Ignore components? Ignore data points?