Multivariate continuous — features

- Features to look for
  - outliers
  - unusual groupings
  - association between variables
  - dependencies
  - groupings/clusters
  - …

Multivariate continuous — displays

- SPLOM (scatterplot matrices)
- Rotating plots
- Comparisons
  - Density estimates, QQ plots, Distribution functions, boxplots
- Contour plots
- Parallel coordinate plots
- Glyphs (starplots, profile plots, …)
- Matrix visualization (heatmaps)
- Lattice (using shingles)

Scatterplots (1)

- Possible features
  - association or dependence
  - triangular structures
  - 2-d boundaries
  - 2-d outliers
  - groupings
  - low density areas
  - modes
  - …

Scatterplot examples?

- Example from the *car* package
  - Angell
  - Prestige (Canadian occupational data)
- ShotScale (Film shots)

Scatterplots (2)

- One variable is plotted against another.
- If there is a dependency, the dependent variable should be on the Y axis.
- Point symbol, size, and colour can be used.
- Aspect ratio is important.
- Guidelines may help (e.g., Y=X).
- Adding functions may help (e.g., smooth, density contours).
- Alphablending can be useful for large datasets.
- Density estimation with colour coding may also work.

Scatterplot matrices

- Each variable is plotted against every other one.
- On the diagonal you can have
  - variable name
  - histogram
  - density estimate
- Can restrict to only upper of lower triangle.
- Scale labelling can be tricky.
- Not good for many variables.

Rotating plots

- Three variables are plotted with (x,y) intially on screen and z perpendicular to the screen.
- The plot is rotated interactively.
- For more than 3 variables, the plot shows a 3-d projection and projection pursuit indices are used to drive the direction of rotation through the m-d space.
- ggobi software
- Difficult to interpret in higher dimensions and the projection pursuit indices do not necessarily achieve their goals.

Distribution comparisons

- Density estimates
  - easy to overlay on one another for comparisons
- QQ plots
  - plot quantiles of one empirical distribution against another
- Distribution functions
  - good for stochastic dominance
- Boxplots
  - efficient comparison of many distributions (as long as boxplots are appropriate)

Contour plots

- 2-d density estimates

Parallel coordinate plots

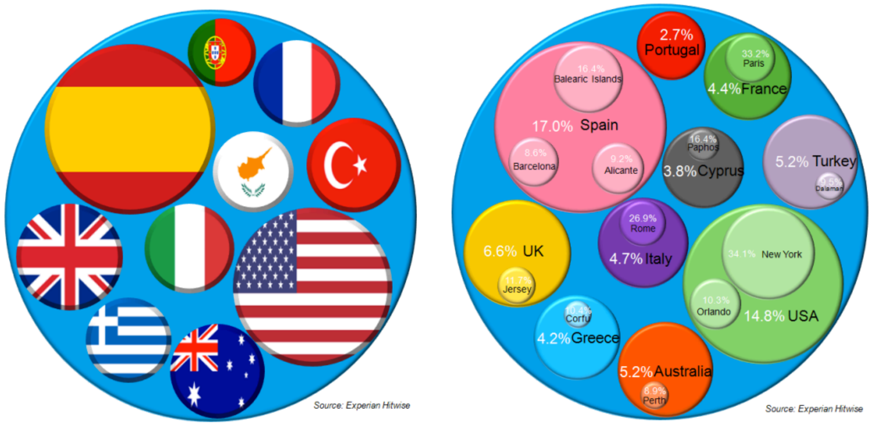
- Each variable has its own vertical axis.
- Each case is represented by a set of line segments joining its points on the axes.
- Scaling and axes order affect the display a great deal.
- Interactive tools are important.
- Rescale axes
  - inversion, common scaling
- Display as boxplots
- Reorder variables
  - by hand
  - sorting by statistics (max, median, IQ-range ...)

Glyphs

- Each case is represented by a symbol with dimensions proportional to the individual variable values.
- Each variable is standardised individually.
- Symbols can be
  - Chernoff faces, stars (with or without axes, filled), profile plots, barcharts, …
- Choice of symbol, assignment/ordering of variables, layout of glyphs, ordering of cases all influence interpretation.
- cf. GAUGUIN software

Matrix Visualization

- Each row is a case and each column is a variable. Cell (ij) represents the value of case i on variable j.
- Each variable is standardised (though common scaling would be possible where appropriate).
- Values are usually represented on a colour scale and the use of heat colour scales gives the name heatmap.
- Choice of colour scales is important.
- Associated correlation matrices for rows (cases) and columns (variables) are sometimes drawn.



Flight searches by the UK Internet population

weblogs.hitwise.com/james-murray/2011/09/flight_search_infographic_new.html

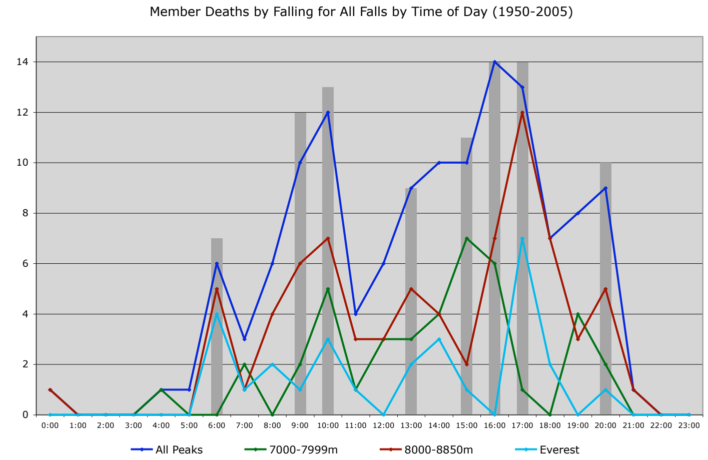


Chart D-30a: Member deaths by falling for all falls by time of day from 1950-2006

The Himalaya by the Numbers
A Statistical Analysis of Mountaineering in the Nepal Himalaya
Richard Salsibury, Elizabeth Hawley www.himalayandatabase.com/hbn.html