Multivariate data visualization

Lecture 4

732A98

Multivariate data visualization

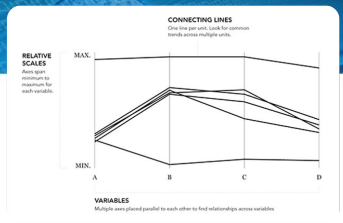
Continious variables involved in the following:

- Parallel coordinate plots
- Heatmaps
- Star charts

Parallel coordinates

Construction:

- Vertical axis: Values
- Horizontal Axis : Variables
- 1 trace line = 1 observation



Analysis: - Clusters - Outliers - Correlated variables

Parallel coordinates

Analysis

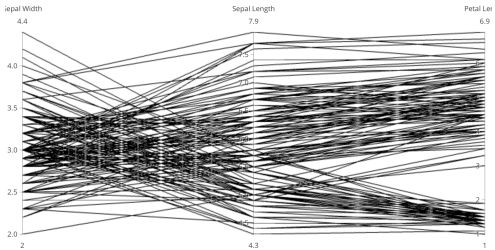
- Positive correlation between two adjacent variables: almost all segments are parallel to each other
- Clusters in some variable space: several trace lines that are near each other and have similar pattern
- Outliers: trace lines that have unusual pattern and/or fall out outside the common plot area

Problems

- Trace lines overlap each other -> difficult to find patterns, difficult to follow a specific trace line
- Analysis depends much on the order of variables (correlation, clusters) -> a proper reordering may improve the analysis

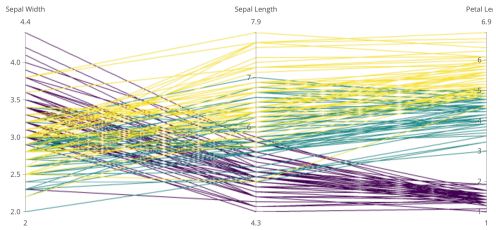
Parallel coordinates

Example: Iris dataset - How many clusters do you see?



Parallel coordinates

- Sometimes clusters overlap with categories given by some variable
  - Non-mixing groups is not the same as clustering!



Ordering problem

- Problem of ordering (variables, observations) is one of the key problems in multidimensional visualization
  - Sometimes has a huge impact on perception (heatmaps)
- A lot of approaches exist

Problem formulation: Given data set $\chi = \left(x_{ij} | i = 1, \dots, n, j = 1, \dots, p\right)$

- Select order $\Psi = i_1, \dots, i_p$ that optimizes visual perception (analysis) -> this defines reordering of data columns $\Psi : \chi \rightarrow \chi'$

Note: $p!$ possible orderings exist...

Ordering problem

Solution

- early approaches (for ex. Ankerst et al. 1998):

1. Choose a distance (proximity) matrix $D = \left\{d_{ij} = d(x^i, x^j)\right\}$ between variables (columns)
   - Euclidian distance on scaled columns
   - 1- correlation
2. This defines graph with vertices $1, \dots, p$ and edge weights d_{ij} -> **Hamiltonian path** (Traveling Salesman Problem)

$\min_\Psi \sum_{j=1}^{p-1} d'_{j,j+1}$

TSP is NP-complete → Approximate solutions are used.

Ordering problem

Solution: modern approaches

- Based on:
  - Decreasing visual clutter
  - Clustering data points/dimensions
  - Outlier detection
  - Dimensionality reduction (for ex. MDS)
  - …
- Note: most of these can be applied both for ordering observations and ordering variables
  - Just transpose the data matrix…

Ordering problem

Objective functions:

- Gradient measures (anti-Robinson)
- Hamiltonian path length
- Least squares
- …

They based on $\min_\Psi L(\Psi(D))$

Optimization algorithms:

- Partial enumeration
- Traveling salesman solvers

· Hierarchial clustering

Gradient measures

Aim: distances should increase from diagonal

$$d_{ik} \leq d_{ij} \text{ for } 1 \leq i < k < j \leq n$$

$$d_{kj} \leq d_{ij} \text{ for } 1 \leq i < k < j \leq n$$

Objective function:

$$L(D) = \sum_{i<k<j} f(d_{ik}, d_{ij}) + f(d_{kj}, d_{ij})$$

where

$$f(z, y) = sign(z - y) \text{ or } f(z, y) = z - y$$

Other objectives

Hamiltonian path length:

$$L(D) = \sum_{i=1}^{n-1} d_{i,i+1}$$

Least squares criterion (PCA)

- Solution is similar to first PCA component

$$L(D) = \sum_i \sum_j (d_{ij} - |i - j|)^2$$

Optimization algorithms

Partial enumeration methods

- Ex: Branch and bounds and dynamic programming
- Constructing solutions by parts

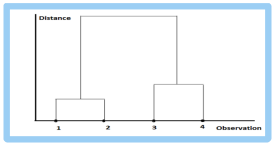
TSP solver

- Suitable for hamiltonian path objective
- Find shortest path by dynamic programming or heuristics

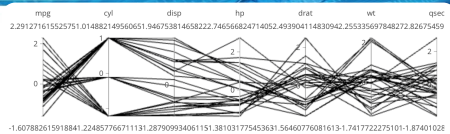
Optimization algorithms

Hierarchical clustering

- Observations are joined into clusters
- Clusters are joined in larger clusters
- Until only one cluster left
- Leaves and branches are permuted to minimize given objective



Effect of ordering

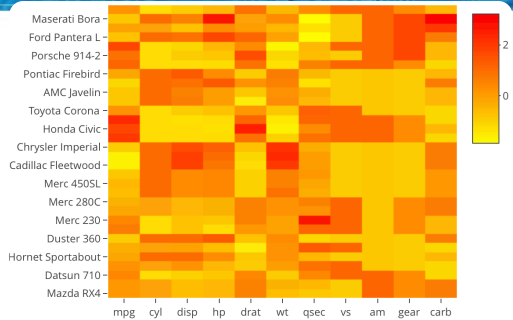


Heatmaps

A heat map visualizes a matrix [n x m]

- Normally rows=observations, columns= parameters
- Heatmap has the corresponding size
- Each cell of the matrix corresponds a cell in the heatmap
- High values correspond intense colors in this map (or visa versa for other color schemes!)
- Names of variables and observations are shown

Heatmaps



Heatmaps

Analysis:

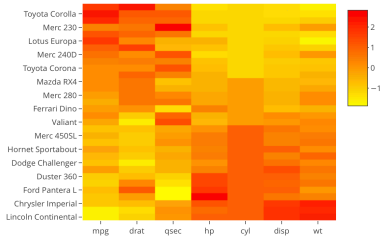
- Compare the values of a parameter for different observations (row)
- Compare the values for a single observation (column)
- Compare the patterns for different rows or columns
- Find similar observations (areas with the same color intensity)
- Find which variables define similarity for a group
- Find correlated variables (similar pattern within a column)

Heatmaps

- Exercise (last picture):
  - How many clusters do you see?
  - Which variables define clusters?
  - Which variables are correlated?

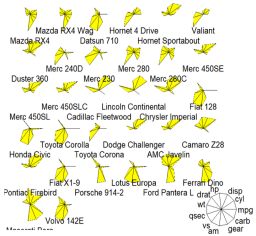
Effect of reordering

- Gradient measure objective used
- See new analysis possibilities



Radar charts

- Use polar coordinate system
- Map column value as a coordinate in certain direction



Radar charts

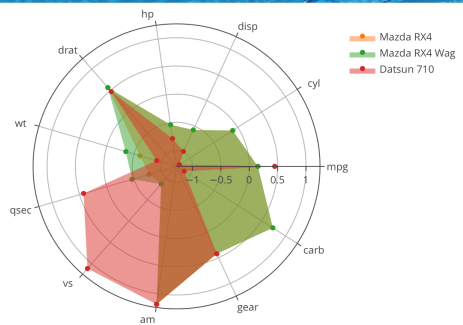
If juxtaposed, analyse:

- Clusters
- Outliers
- Outlying directions

If superimposed,

- Comparing variable length
- Seeing similar and outlying observations

Radar charts



Radar charts

Problems:

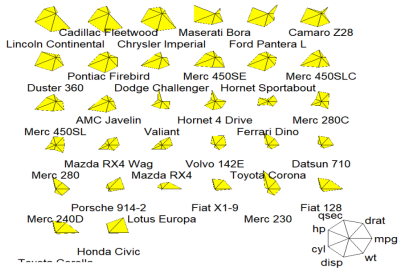
- Difficult to judge orientations
- Number of dimensions are observations is very limited
  - Number of observations is extremely limited if superimposed
- More close radar charts easier to compare
- Perception is much affected by observation ordering

Ordering:

- Same as before plus
- Dimensions can be sorted to promote more symmetric charts

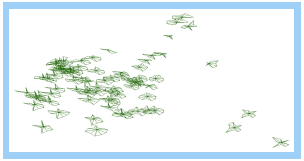
Radar charts

- Now with reordering by Gradient Measures



Radar charts

Other positioning possible - PCA/MDS



Trellis plots / facets

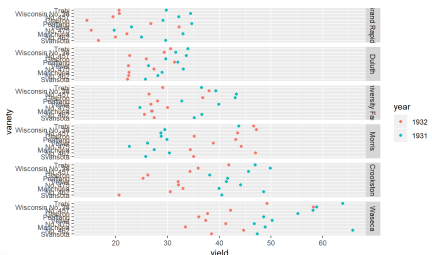
Idea:

1. Make same kind of plot for subsets of data
2. Plot together
3. See patterns/differences

Analogy: cutting a sausage

Trellis plots / facets

- Example: Barley data
  - Anything strange?



Trellis plots / facets

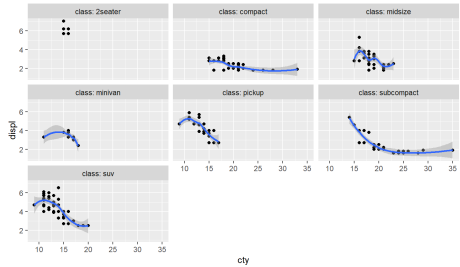
- Faceting = one more aesthetics
- What can be analysed?
  - Patterns within/between plots
  - Conditional dependence $Y \sim X|Z$
  - Variable interaction, additivity

–> Useful tool for modeling!

- Compare : 3D- scatter plots

Trellis plots / facets

- Another car data: is there additivity?



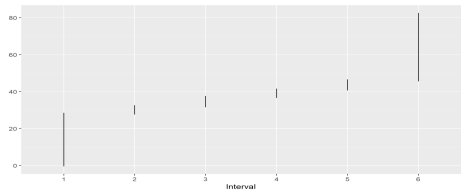
Trellis plots / facets

- Design issues:
  - How to order rows/columns in trellis?
    - A: X=one var, Y0=another var (`facet_grid`)
    - B: independently of aes (`facet_wrap`)
  - How to handle categorical vars?
    - One value/panel
    - Group
    - Ordering? (R: decide factor levels)
  - How to handle real-valued vars?
    - Split equal size/length -Shingles

Shingles

- Creates overlap
- To avoid boundary effects

Example: Aids data (Age, Time of Death, Time of Diag)



Shingles

- Aids data: conclusions?

Read at home

- Chapter 5
- Paper "Hahsler, M., Hornik, K., & Buchta, C. (2008). Getting things in order: an introduction to the R package seriation. Journal of Statistical Software, 25(3), 1-34".
- (Browse through) paper "Ankerst, M., Berchtold, S., & Keim, D. A. (1998, October). Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In Information Visualization, 1998. Proceedings. IEEE Symposium on (pp. 52-60). IEEE."
- Becker, R. A., Cleveland, W. S., & Shyu, M. J. (1996). The visual design and control of trellis display. Journal of computational and Graphical Statistics, 5(2), 123-155.

34/34