11 High-Dimensional Data

36-721 Statistical Graphics and Visualization

Jerzy Wieczorek

10/8/15

Last time

- Principles and R code for geographic maps
- Critiques due

Today

- High-dimensional data graphics
- GGobi demo
- GGobi exercise

Figures in slides are from GGobi website and book:
Cook & Swayne, *Interactive and Dynamic Graphics for Data Analysis, With R and GGobi* (Springer, Amazon).

High-dimensional data graphics

A single scatterplot can only show a few variables:
x, y, color, shape, size, and 2 facets $\approx$ 7D.

Scatterplot matrix like `pairs()` can show every pair of variables, but each is only 2D.

How to explore or present data with more variables than this?

- Brushing and linking
- Projections and tours
- Parallel coordinates plot

Brushing and linking

Brushing selected points identifies them & gives more detail.
Linking highlights the brushed points on other plots too.

Can do basic brushing in base R with `locator()` function;
or you can build a custom Shiny app:

```
library(shiny)
runGitHub("36721-F15", "civilstat",
          subdir = "08_ShinyLab/UScereal_Brushing")
runGitHub("36721-F15", "civilstat",
          subdir = "08_ShinyLab/UScereal_Linking")
```

. . . or use a general tool like GGobi or Tableau.

Projections and tours

Low-dimensional projection: linear mapping from high-dim space into 1D or 2D. Plot the result as a histogram or scatterplot.

Tour: animated sequence of projections, interpolated smoothly.

- Grand tour: projections chosen randomly, but designed to cover the space of all possibly projections efficiently
- Projection pursuit: sequence chosen to seek "interesting" projections; user specifies a function to define what is interesting (clusters, outliers, etc.)
- Manual manipulation: control rotation yourself, e.g. to tweak a projection found by random/guided tour

In GGobi, save interesting projections to R with `rggobi`,
or use Tour2D > Show Projection Vals.

Parallel coordinates plot

Lay out each variable on a 1D axis, then place in parallel:
x-axis shows variables, y-axis shows their (scaled) values.

Order matters. Try ordering by correlations between variables.

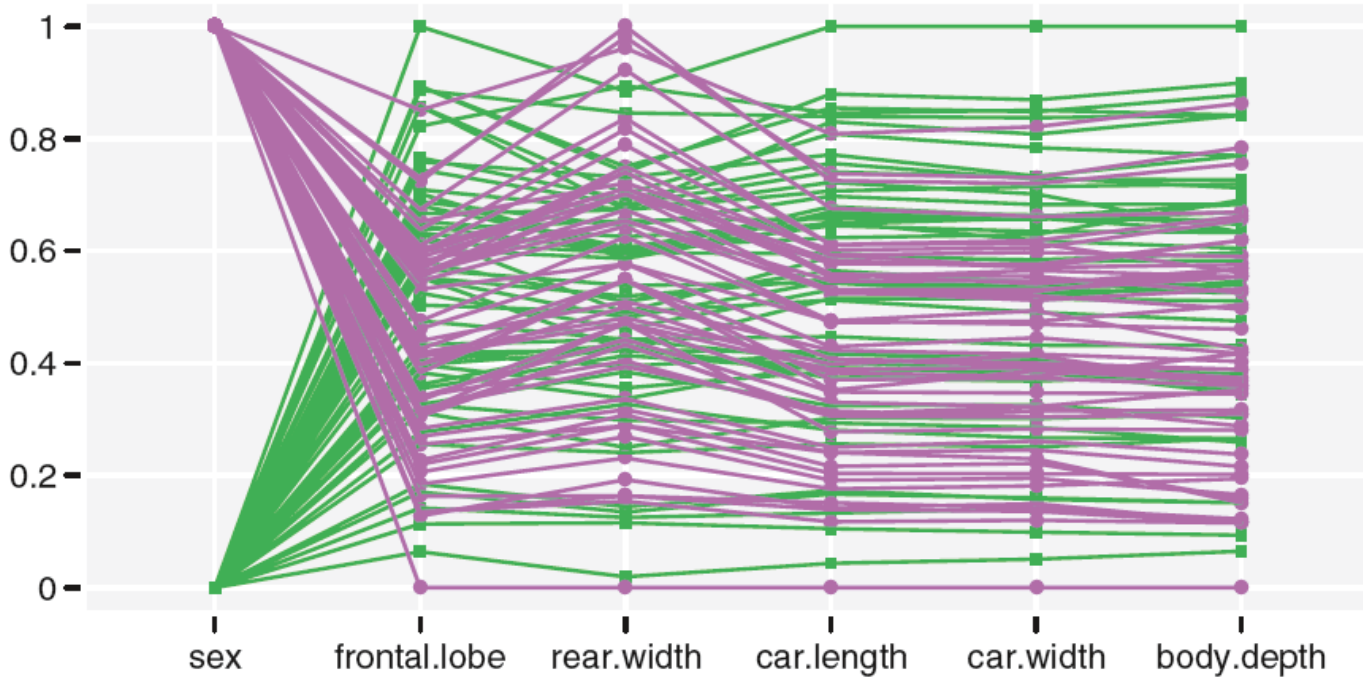

Fig. 2.5. Parallel coordinate plot of six of the seven variables of the Australian Crabs data, with males and females identified as green rectangles and purple circles, respectively. The relatively flat traces indicate strong correlation between variables.

GGobi demo

Example

Italian olive oils

- Number of samples: 572
- Number of variables: 10

Super-classes, 3 regions, and 8 classes, areas within region.

Explanatory variables are % fatty acids in the sample: palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic

How do we distinguish the oils from different regions and areas in Italy based on their combinations of the fatty acids?

GGobi demo

Follow along with Tutorial, Section 2 of GGobi manual.
We explore the Italian olive oils dataset. Can we use the 8 fatty acid composition variables to cluster into Italy's 3 geographic regions (North, South, & Sardinia)?

- Choose variables for 2D scatterplot
- Identify & brush
- Barchart linked with scatterplot
- 2D Tour, show axes, manual manipulation
- Parallel coordinates

GGobi exercise

Shadow out and exclude Regions 2 and 3, keeping only Region 1.
(Tools > Color & Glyph Groups > Shadow > Exclude shadows)

This Region has four Areas:
Calabria, Sicily, North Apulia, & South Apulia.

Explore the data.
Can you find projections that separate these four Areas?
Is one Area harder to classify than the others?

Further reading

GGobi website and book *Interactive and Dynamic Graphics for Data Analysis* have great advice on graphics for:

- Missing values
- Supervised classification
- Cluster analysis
- Visual inference
- Longitudinal data
- Network data
- Multidimensional scaling

For next time

- Sat 10/10: Project 2 (Interaction Design) due 5pm
- Tues 10/13: networks and trees
- Thurs 10/15 (optional): a few bonus topics; office hours
- Sat 10/17: Project 3 (Research) due 5pm
- Sat 10/24: final resubmissions due 5pm