36-315: Statistical Graphics and Visualization

Handout 24
Date: April 21, 2003

Parallel-coordinate plots for visualizing high-dimensional data

Nomogram—A graphical lookup table where variable axes are parallel lines (not perpendicular). Pick values along any two variables and draw a line through these two points. The intersection of the line with the other axes is the predicted value of those variables.

Parallel-coordinate plot—A data display based on the nomogram. The variable axes are parallel lines, with data points plotted along each. Points on neighboring axes corresponding to the same individual are connected by a line, similar to a line chart. The line connecting all values for a single individual is its **profile**.

Duality:

- Points in a scatterplot become lines in a parallel-coordinate plot.
- A scatterplot can show many individuals (1000s) in a few dimensions (2). This makes it good for finding outliers and clusters, and judging the correlation between dimensions.
- A parallel-coordinate plot can show many dimensions (10) for a few individuals (10). This makes it good for finding unusual variables and variable groups, and judging the similarity between individuals.

Scaling and sorting makes parallel-coordinates different from a line chart. Variables need to be put onto a comparable scale, and sorted so that similar variables are together. As in projection, this leads to a search for the 'optimal' parameters.

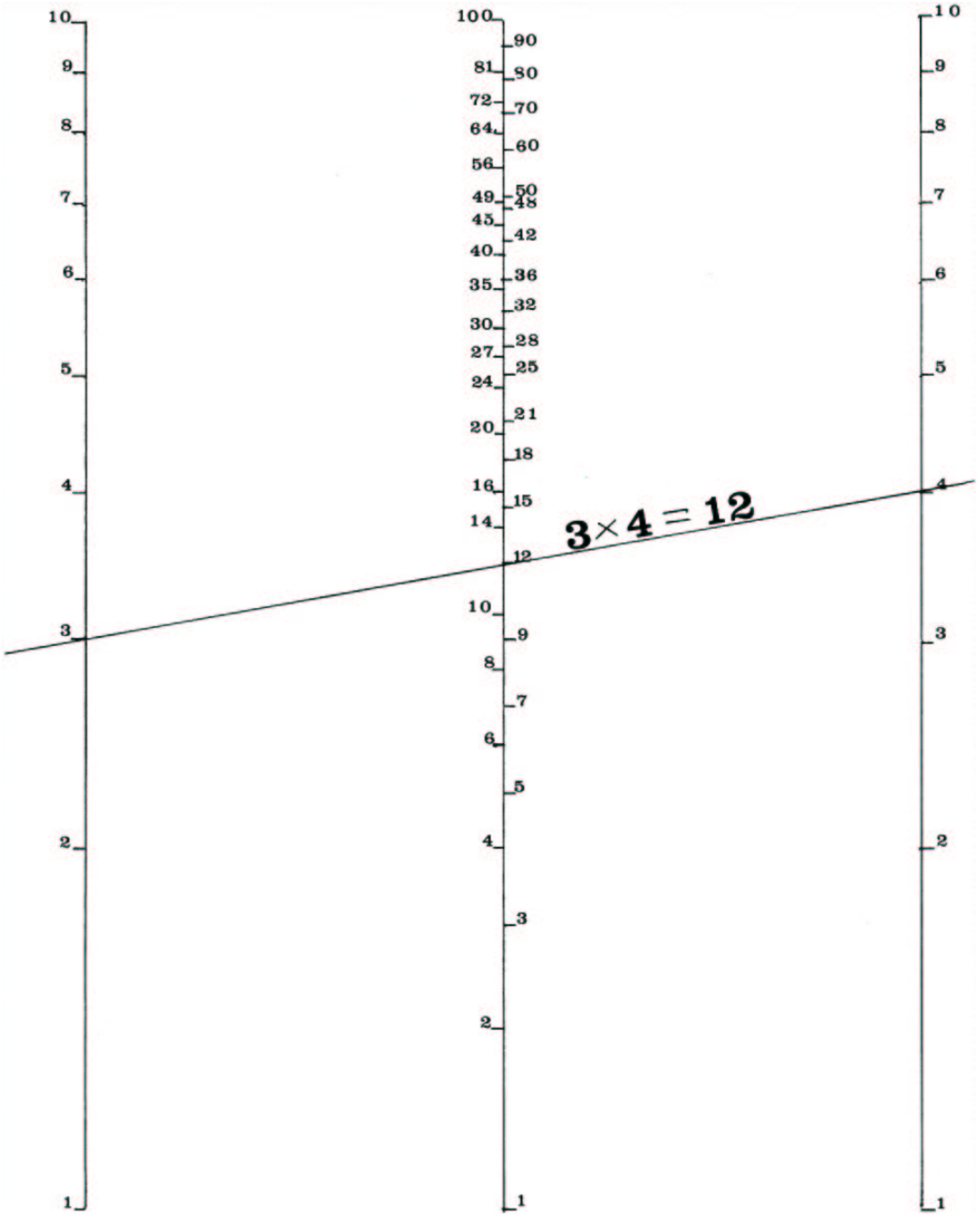
Linear profiles method—Variables are scaled and sorted to make the profiles turn out as straight as possible (minimize wiggle). Sometimes this involves reversing (negating) an axis. It also helps to make the axes unevenly spaced. In this way, you can construct your own nomograms.

Shakespeare data: Too many plays to show all at once, so the median value in each genre is plotted. Comedy and History are "opposites". Tragedy and Romance are in the middle. Six dimensions separate Romance from the other genres.

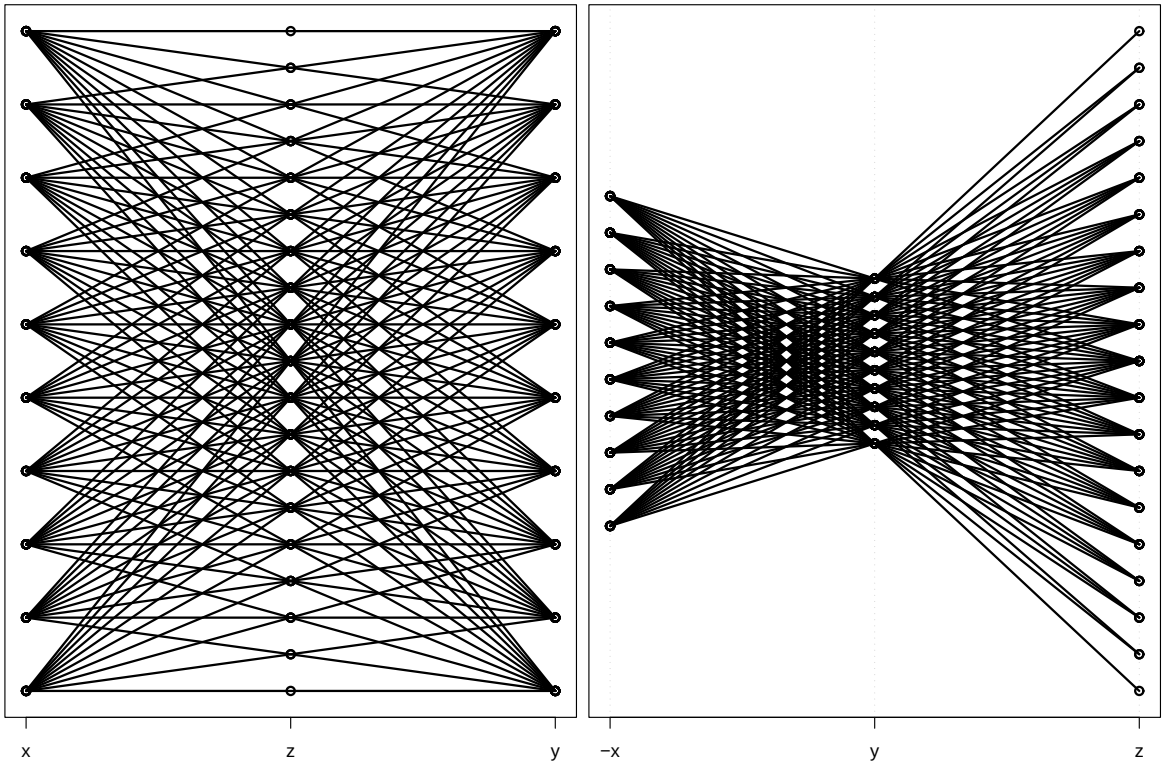
Census data: Too many tracts to show all at once, so they are grouped into 4 clusters. Outliers can be compared to the clusters on many variables simultaneously (instead of one at a time on a color plot).

References

[1] T.L. Hankins. "Blood, dirt, and nomograms: A particular history of graphs", *Isis* 90:50-80, 1999.

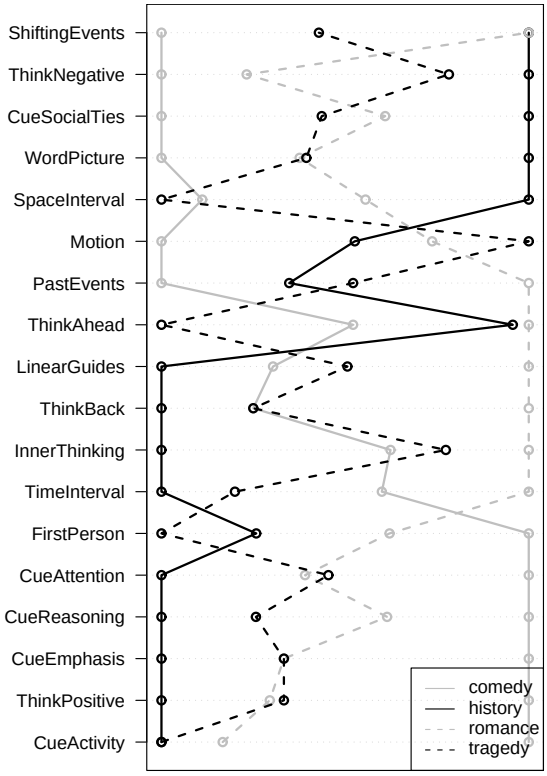


Parallel-coordinate plots of a multiplication table
with different axis orderings—right plot requires a reversal
Remove the lines to get a multiplication nomogram

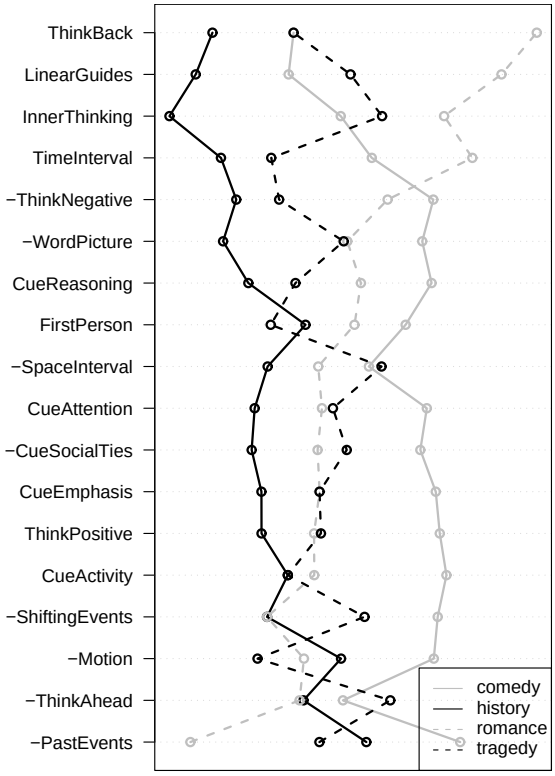


Parallel-coordinate plots of Shakespeare data

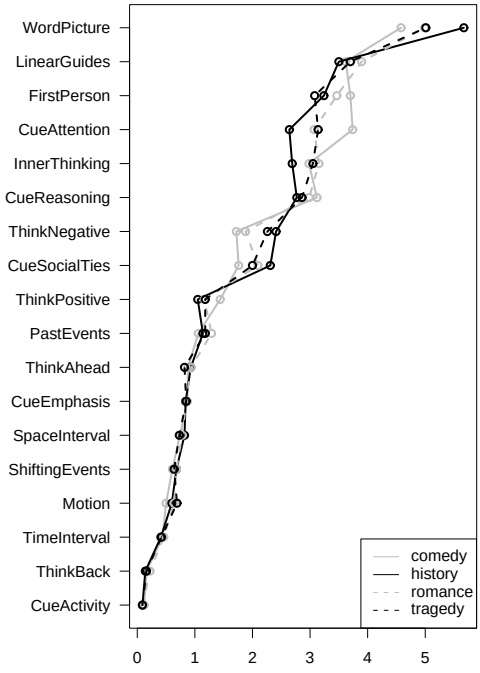
(Left) Each variable scaled to $[0, 1]$



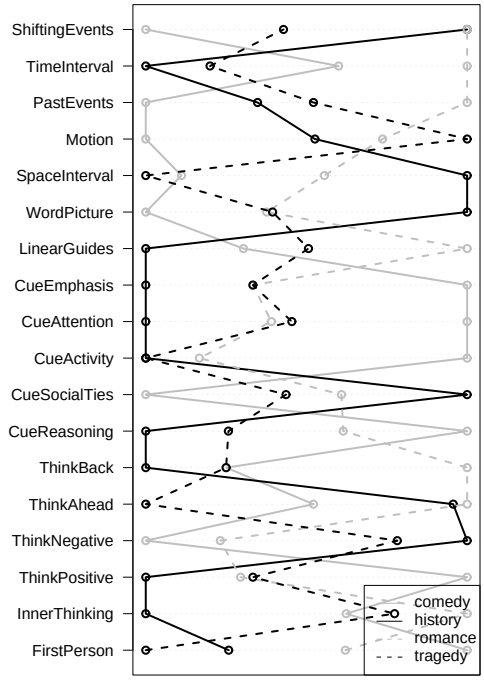
(Right) Linear profiles method



(Left) Without scaling



(Right) Without sorting



(Left) Clustering of Pennsylvania tracts into 4 groups plus 2 outliers.
(Right) Parallel-coordinate plot compares them on many dimensions simultaneously (outliers are dotted).