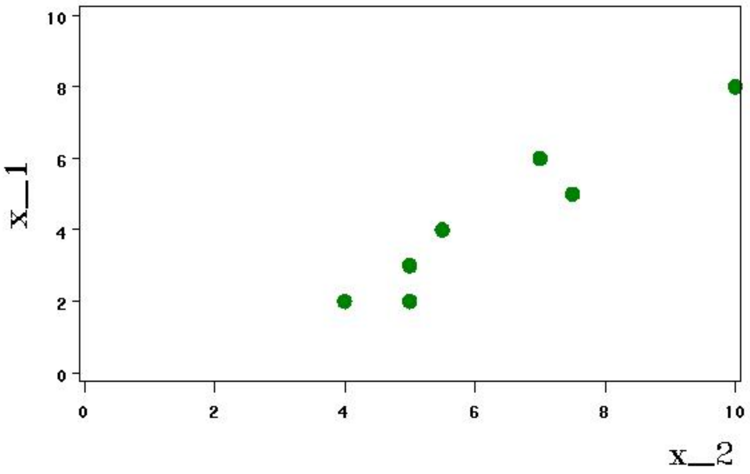
Graphical Representation of Multivariate Data

- One difficulty with multivariate data is their visualization, in particular when $p > 3$.

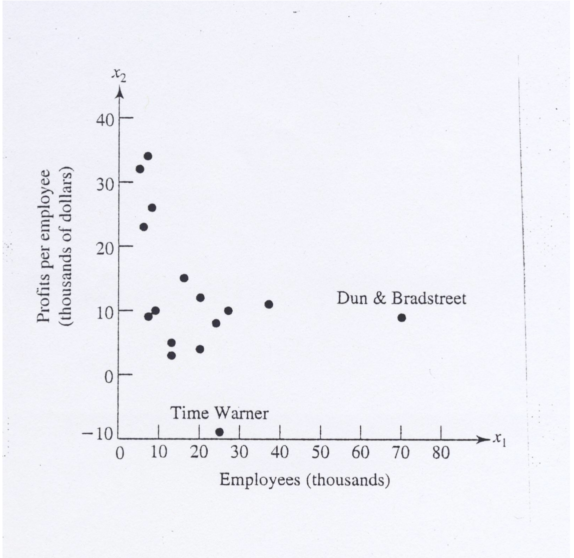
- At the very least, we can construct pairwise scatter plots of variables. Data from exercise 1.1 (transpose of Figure 1.1)



Graphs and Visualization (cont'd)

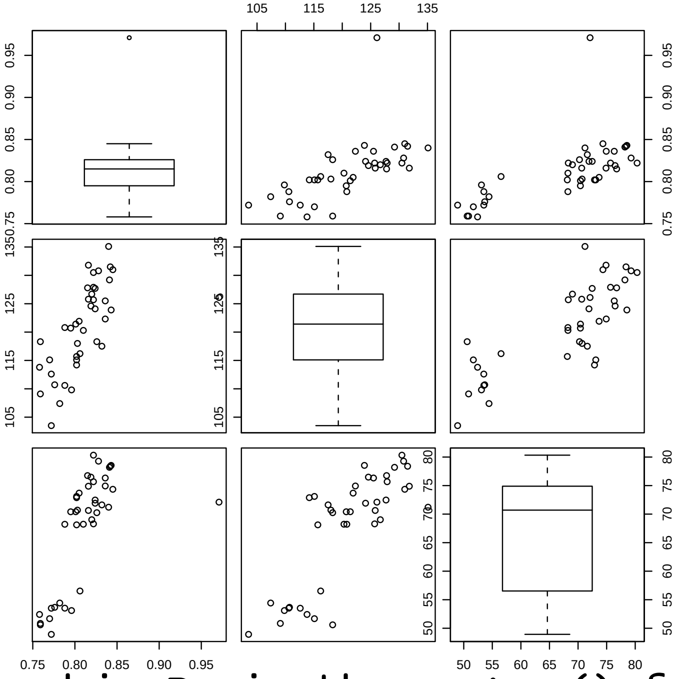
- Graphs convey information about associations between variables and also about unusual observations.

- Example 1.3: correlation between number of employees and productivity at 16 publishing firms is -0.39 for all firms and -0.56 if Dun & Bradstreet is not included.



Multiple Scatter Plots

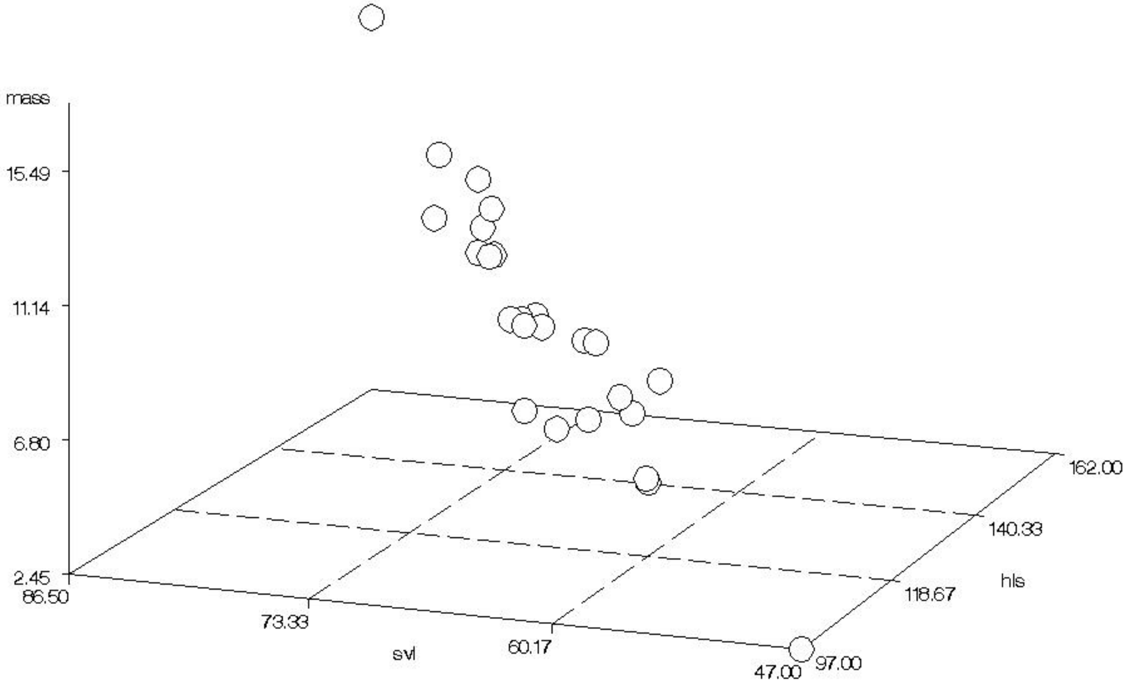
Example 1.5: Paper quality data: X_1=density(g/cc),
X_2=strength(pounds) in machine direction,
X_3=strength (pounds) in the cross direction



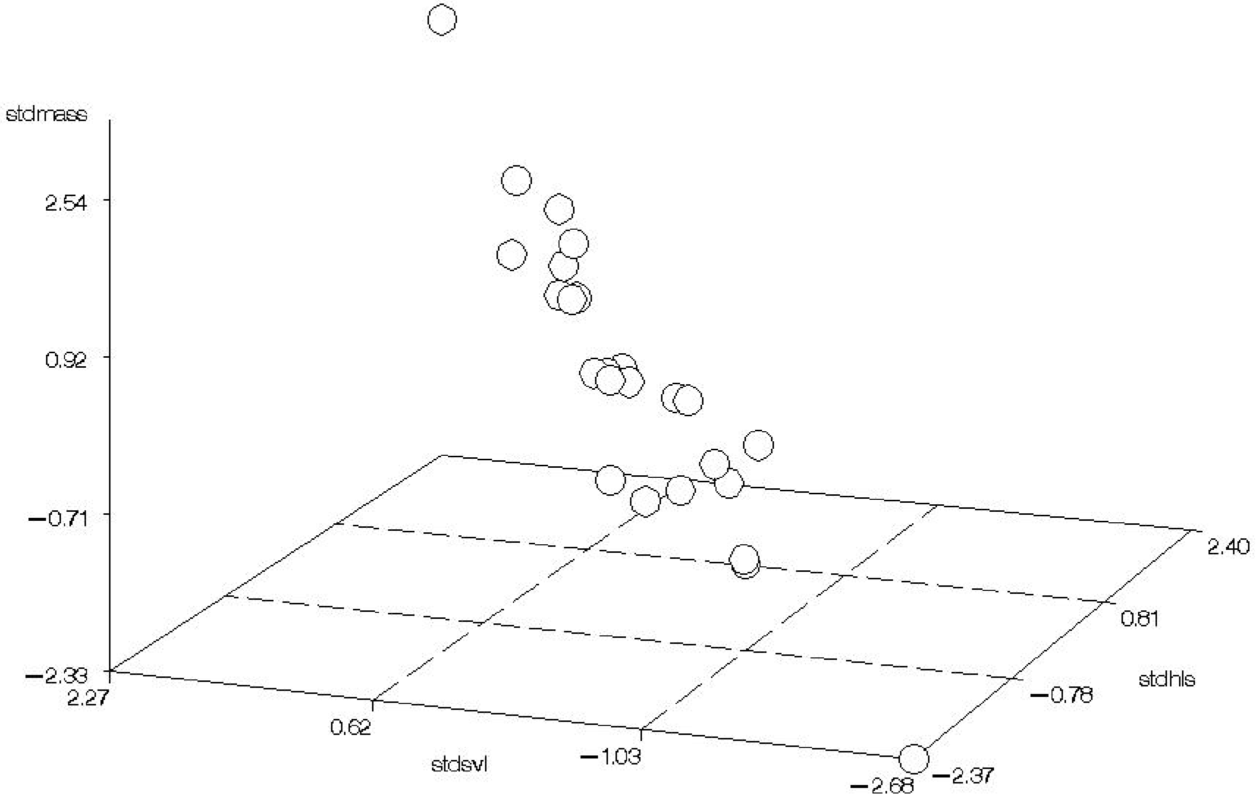
displayed in R via the `pairs()` function

Three-dimensional Plots

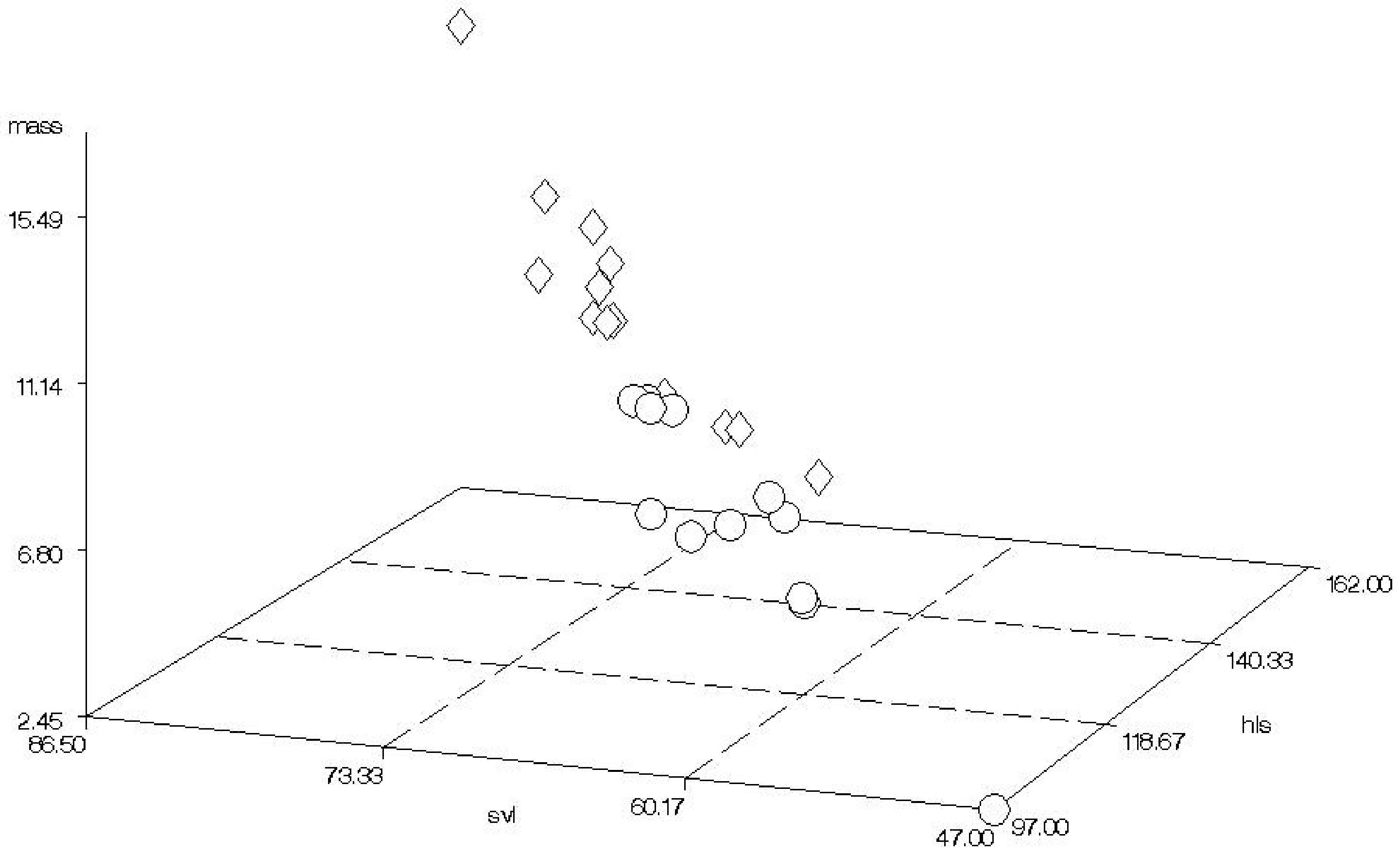
(Example 1.6: Lizard mass (g), snout-vent length (mm), hind limb span (mm))



Three-dim Plot: after standardizing



Four-dimensional Plots?



fourth dimension is plotting character (denotes gender)

Other Graphical Methods

- Specialized software to visualize data in high dimensions is now available.

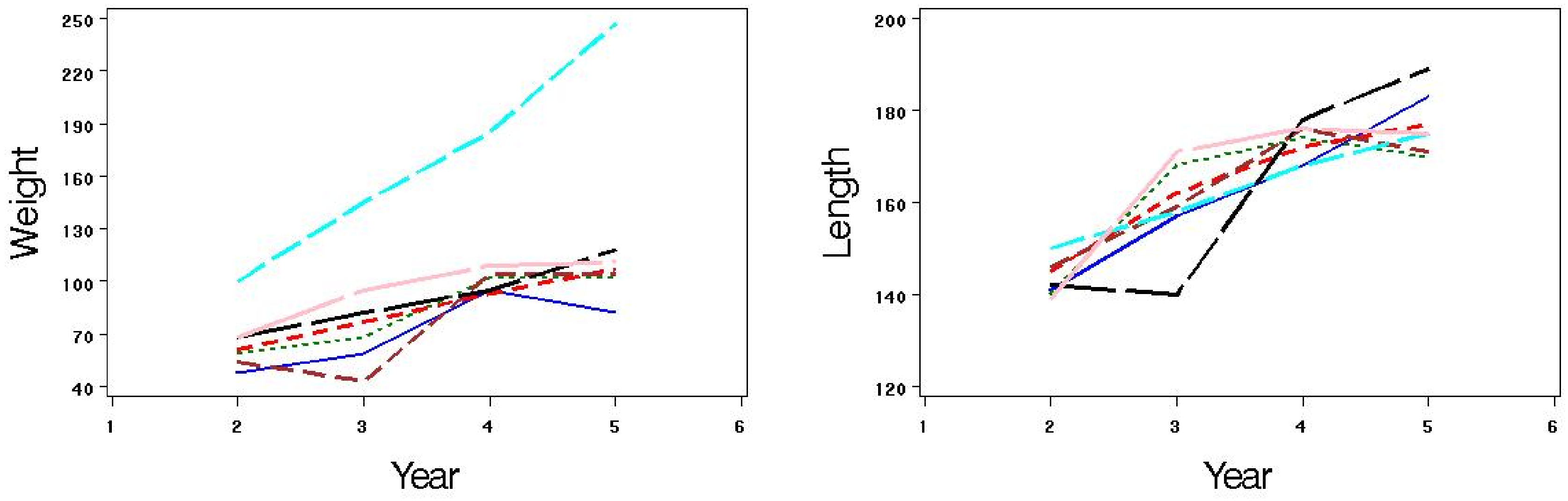
- One example is GGobi, downloadable free of charge from the ISU Statistics Department web site.

- Typically, these programs provide dynamic two-dimensional projections of high dimensional data that reveal
  - Associations among variables
  - Groupings of the units in the sample
  - Other data attributes.

- Graphical methods to display data are very context-dependent

Other Graphical Methods

- *Growth curves*: appropriate when multivariate observations consist of repeated measurements on a set of units. Ex. 1.10: weight and length of female bears over three years.



Even More Graphical Methods

- *Stars*:
  - Suitable for non-negative measurements (standardize the data for each variable and let the center of the circle correspond to the minimum values).
  - Circle of fixed radius with p equally spaced rays (one ray for each variable).
  - Value of each variable is represented by the length of a ray (distance from minimum value for all sample units).
  - Construct one circle for each unit in the sample
  - Group units with similar patterns

Example 1.11: 22 Public Utility Companies

X_1 : Fixed-charge coverage

X_2 : Rate of return on capital

X_3 : Cost per kW capacity

X_4 : Annual load factor

X_5 : Peak kW hours demand

X_6 : Annual sales (kW hours)

X_7 : Percent nuclear

X_8 : Fuel costs (cents per kWh)

- The variables are standardized

- Each variable receives equal weight in the visual impression

- Focus on variables X_6 and X_8 in the next display

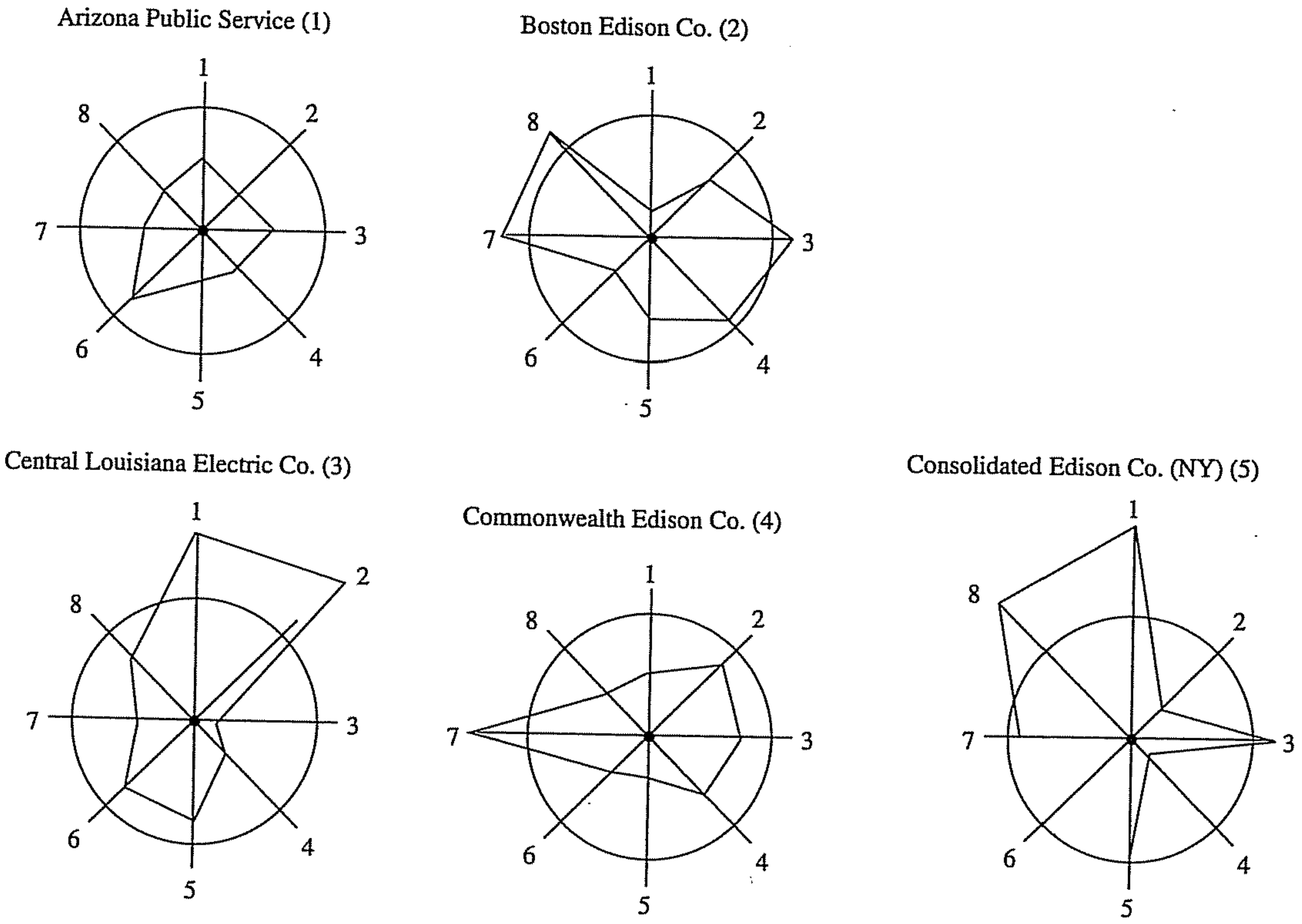


Figure 1.16 Stars for the first five public utilities.

Even More Graphical Methods

- *Chernoff faces*:
  - Same idea as stars, but now each facial characteristic (nose, mouth, etc.) represents a variable.
  - Some facial features can have greater impact than others on how the faces are compared.
  - Group units with similar faces

Example 1.12: 22 Public Utility Companies

X_1 : Fixed-charge coverage	$\longleftrightarrow$	Half-height of face
X_2 : Rate of return on capital	$\longleftrightarrow$	Face width
X_3 : Cost per kW capacity	$\longleftrightarrow$	Position of mouth
X_4 : Annual load factor	$\longleftrightarrow$	Slant of eyes
X_5 : Peak kW hours demand	$\longleftrightarrow$	Eccentricity (ht/w) of eyes
X_6 : Sales (kW hours per year)	$\longleftrightarrow$	Half length of eye
X_7 : Percent nuclear	$\longleftrightarrow$	Curvature of mouth
X_8 : Fuel costs (cents per kWh)	$\longleftrightarrow$	Length of nose

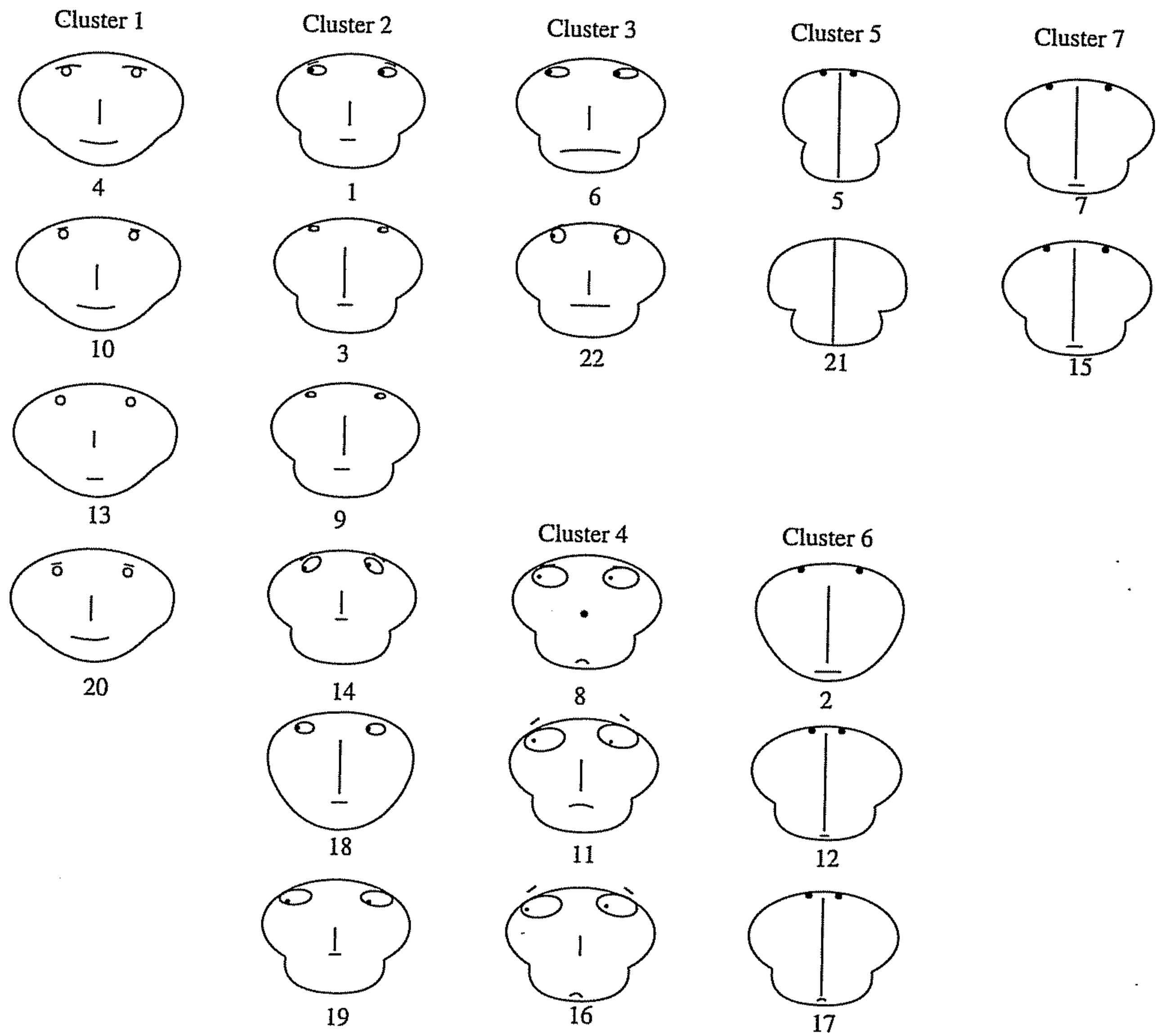


Figure 1.17 Chernoff faces for 22 public utilities.

Additional Techniques for Visualizing High-Dimensional Datasets

- Data in more than two dimensions are quite problematic to represent

- However, there are quite a few suggestions in the area of information visualization in computer science

- Here, we study some static display options beyond the ones studied so far

Survey Plots

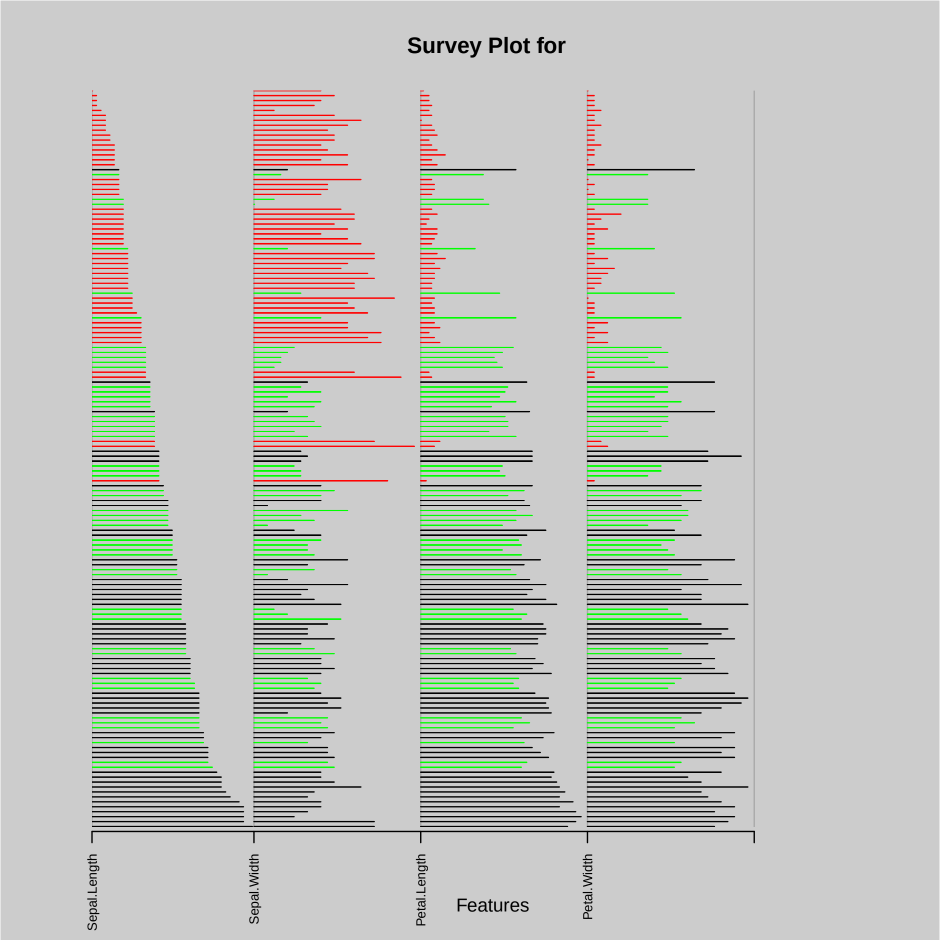
- A simple technique of extending a point in a line graph (like a bar graph) down to an axis has been used in many systems.

- A simple variation of this extends a line from a center point, where the line length corresponds to the dimensional value.

- This particular visualization of n-dimensional data allows one to see correlations between any two variables especially when the data are sorted according to a particular dimension.

- Using color for different classifications may help (after sorting perhaps) determine best coordinates for classifying data.

Survey Plot of Iris Data



Parallel Coordinates Plots

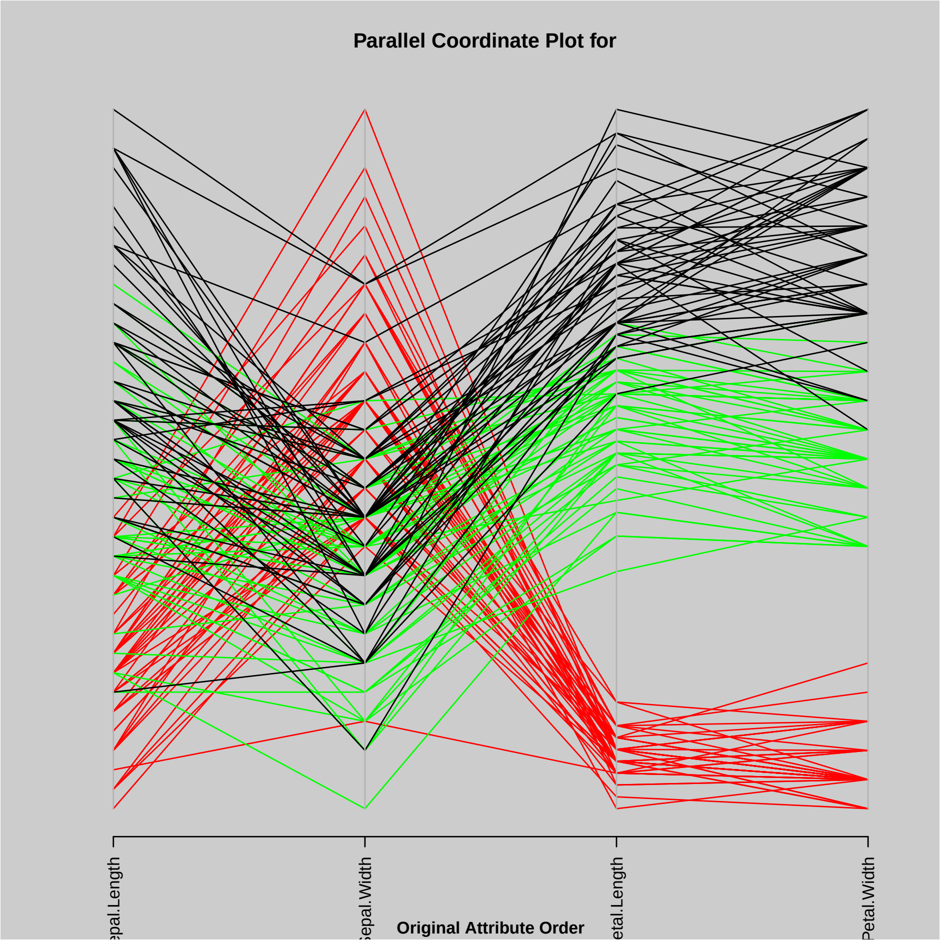
- Parallel coordinates plots represent multidimensional data using lines.

- A vertical line represents each dimension or attribute.

- The maximum and minimum values of that dimension are usually scaled to the upper and lower points on these vertical lines.

- $n - 1$ lines connected to each vertical line at the appropriate dimensional value represent an n-dimensional point.

Parallel Coordinates Plots – Iris Data



Andrews' curves

- Andrews' curves plot each n-dimensional point $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p})'$ as a curved line using the function

  $$f(t) = x_1 + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \ldots.$$

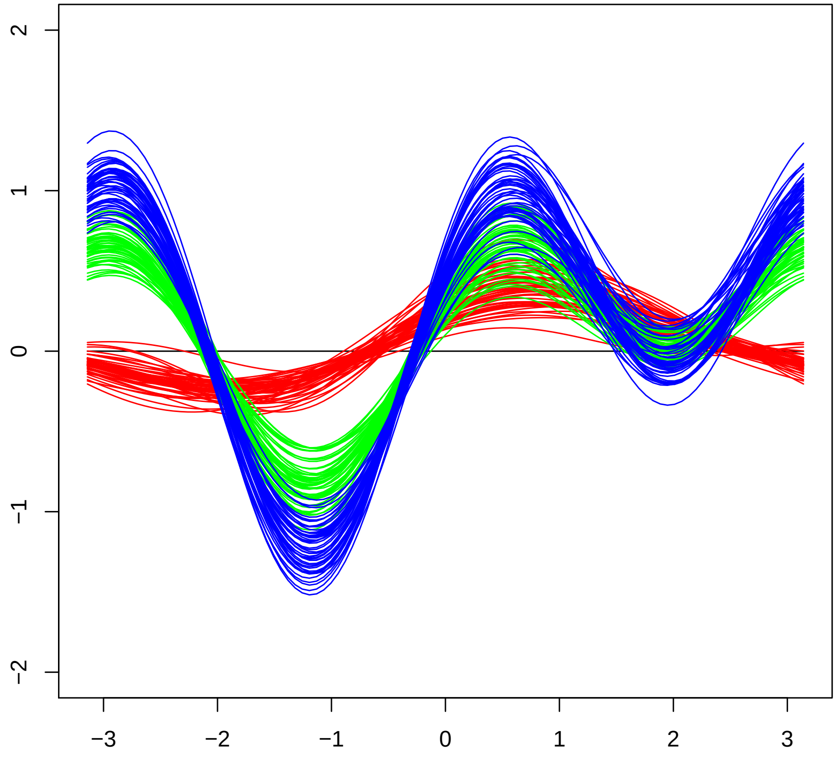
  The function is usually plotted in the interval $-\pi < t < \pi$.

- This is similar to a Fourier transform of a data point.

- Advantage: can represent many dimensions.

- Disadvantage: computational time to display each n-dimensional point for large datasets. Also, order matters.

Andrews' Curves – Iris Data



Radviz or Radial Visualization Plots

- The idea behind radial visualization plots or $Radviz$ is that spring constants can be used to represent relational values between points.

- Like parallel coordinates this is also a lossless visualization method

- Here p-dimensional data points are laid out as points equally spaced around the perimeter of a circle.

- The ends of each of p springs are attached to these p perimeter points. The other ends of the springs are attached to a data point.

- The spring constant Ki equals the values of the ith coordinate of the fixed point.

- Each data point is then displayed where the sum of the spring forces equals 0.

- All the data point values are usually normalized to have values between 0 and 1. For example if all p coordinates have the same value, the data point will lie exactly in the center of the circle.

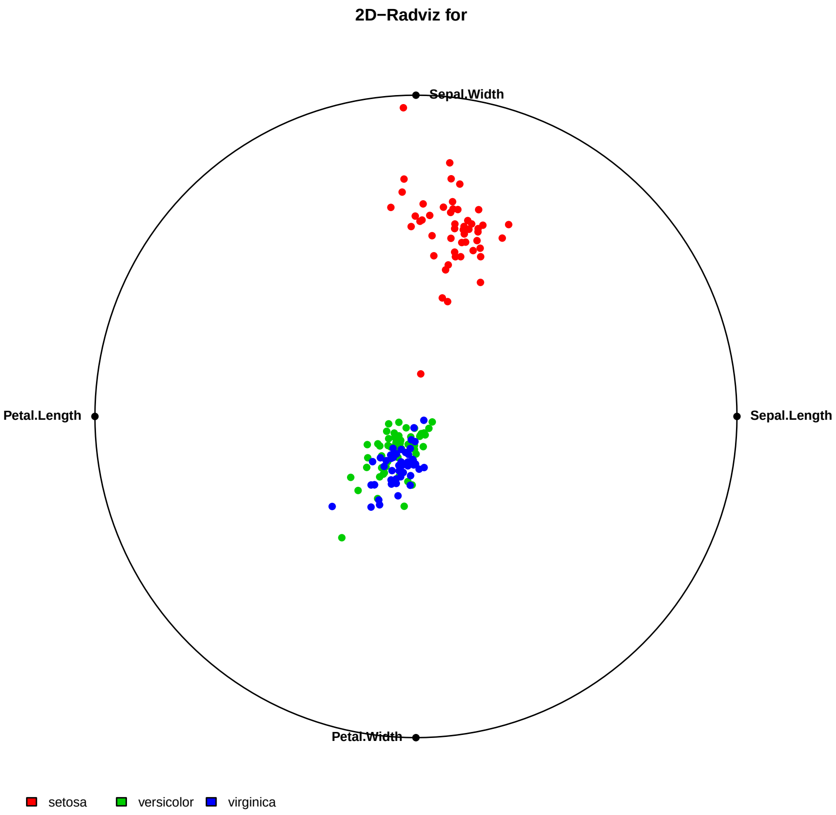
- If the point is a unit vector, then that point will lie exactly at the fixed point on the edge of the circle (where the spring for that dimension is fixed). Many points can map to the same position.

- This represents a non-linear transformation of the data, which preserves certain symmetries and which produces an intuitive display. Some features of this visualization are:
  - Points with approximately equal coordinate values will lie close to the center
  - Points with similar values whose dimensions are opposite each other on the circle will lie near the center

- Points which have one or two coordinate values greater than the others lie closer to those dimensions
- A p-dimensional line will map to a line
- A sphere will map to an ellipse
- A p-dimensional plane maps to a bounded polygon

- Can not really handle severely high-dimensional datasets

Radial Visualization Plot – Iris Data


2D−Radviz for
Sepal.Width
Petal.Length
Sepal.Length
Petal.Width
setosa
versicolor
virginica