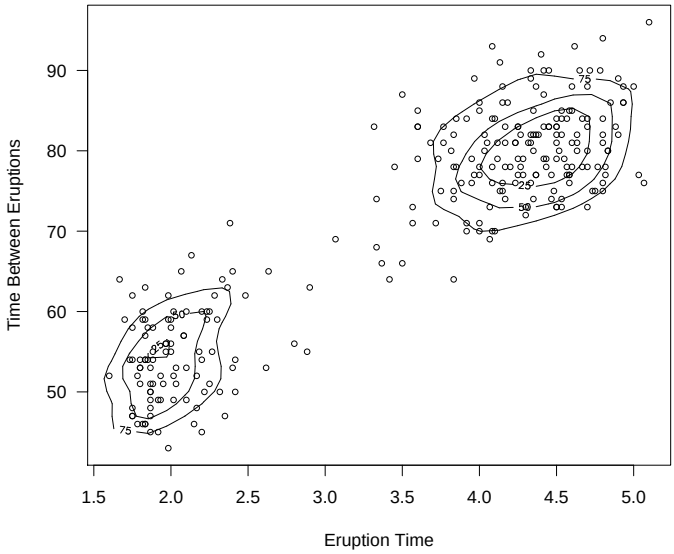
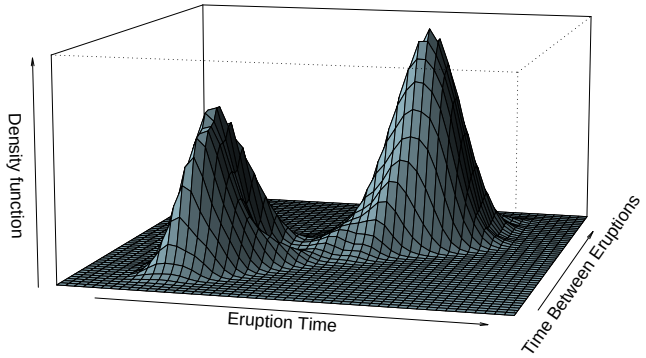
Graphing Data in High Dimensions

Low Dimensions

- There are many techniques which can be used for examining a set of observations: E.g. one-dimensional scatterplots, boxplots, stem-and-leaf plots, histograms, density traces, ...
- In two dimensions the workhorse plot is the scatterplot, by we can also look at contour plots of density estimates. One dimensional plots can be used to examine marginal distributions.
- In three dimensions, we can use animation to convey the impression of a rotating cloud of points. It is also possible to estimate and plot the isosurfaces of a three dimensional density estimate.


Time Between Eruptions
90
80
70
60
50
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Eruption Time
75
25
50
75

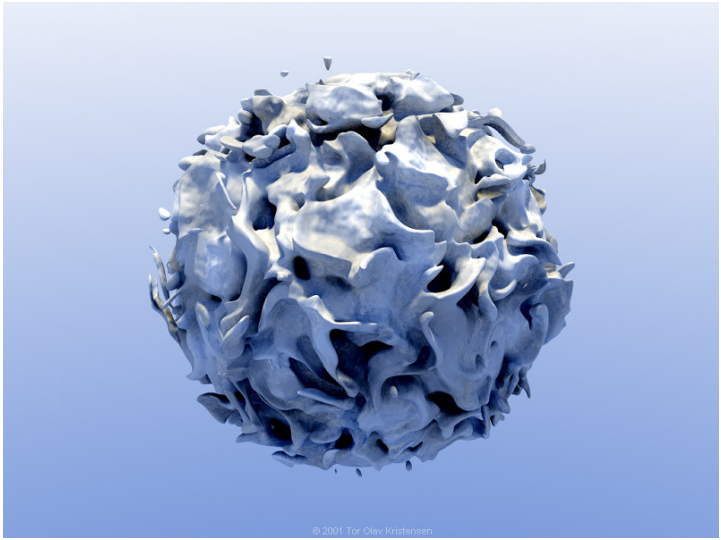

Density function
Eruption Time
Time Between Eruptions

Two Dimensional Density Estimation

Here is a simple example of two dimensional density estimation using the `ks` library.

```
> library(ks)
> H = Hpi(faithful)
> d = kde(faithful, H)
> plot(d, xlab = "Eruption Time",
       ylab = "Time Between Eruptions")
> plot(d, display = "persp",
       theta = 15, phi = 10, expand = .5,
       ltheta = -100, shade = .5,
       col = "lightblue",
       xlab = "Eruption Time",
       ylab = "Time Between Eruptions")
```

A Density Isosurface



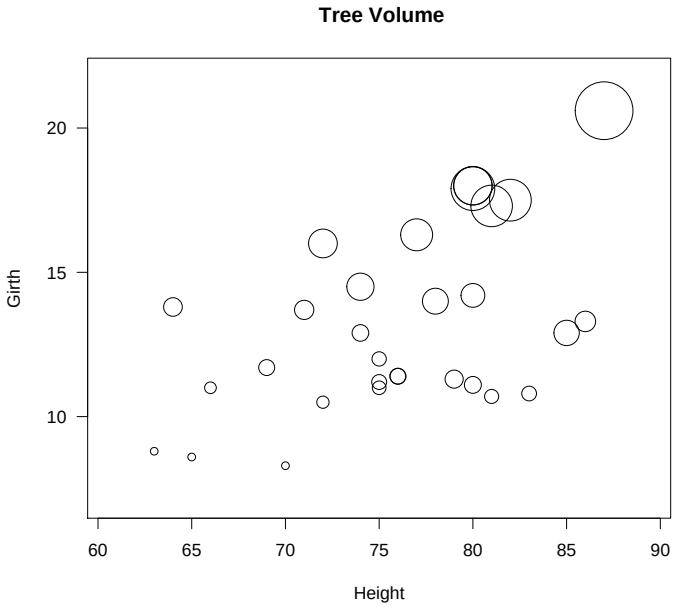
Special Techniques for Three Dimensions

- The first two variables are represented in the standard way in a scatterplot.
- The third variable is represented by using different symbols at the locations in the scatterplot.

Bubble Plots

- The symbols used in the scatterplot are circles whose size is proportional to the value of the third variable.
- There are many possible measures of *size*: The most common are *area* and *radius*
- The following code uses radius.

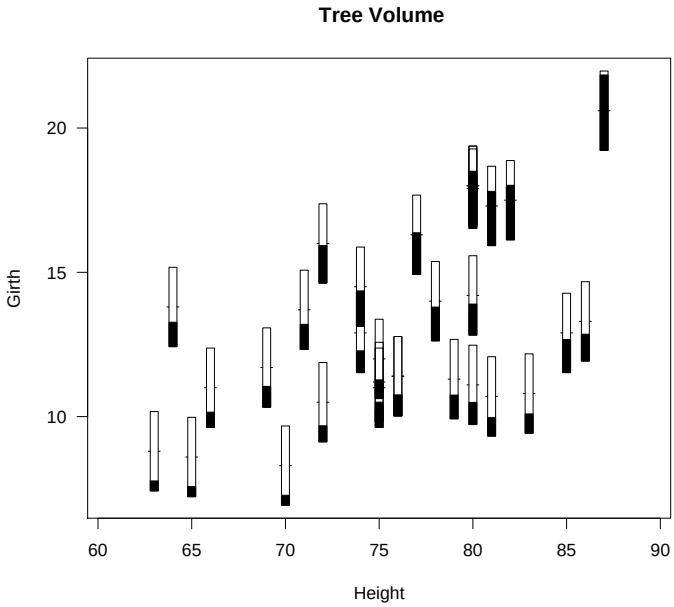
```
> symbols(trees[,2], trees[,1],
          circles = trees[,3]/50, inches = FALSE,
          xlab = "Height", ylab = "Girth",
          main = "Tree Volume")
```


Tree Volume
Girth
Height
10
15
20
60
65
70
75
80
85
90

Thermometer Plots

- When the points are well spaced in the x-y plane, other symbols work well.
- On the basis of his work on graphical perception, Bill Cleveland has suggested using a symbol which looks like a thermometer.

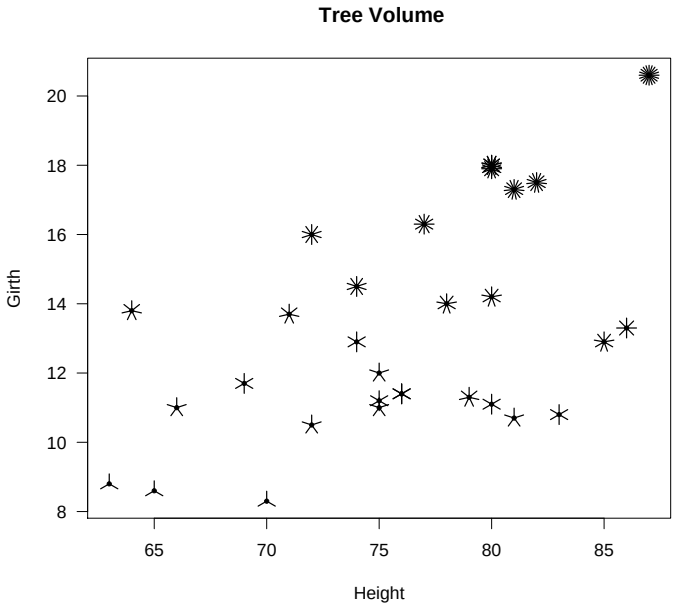
```
> p = 0.95 * trees[,3]/max(trees[,3])
> symbols(trees[,2], trees[,1],
          thermometers = cbind(.05, .5, p),
          xlab = "Height", ylab = "Girth",
          main = "Tree Volume")
```


Tree Volume
Girth
20
15
10
60
65
70
75
80
85
90
Height

Sunflower Plots

- Another representation of three dimensional data is the sunflower plot.
- Here the symbols look like flowers, with more petals representing higher values.

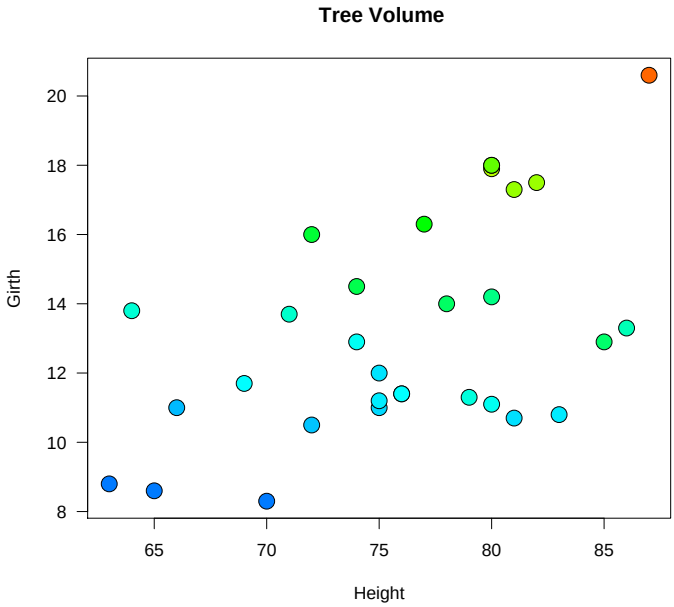
```
> n = ceiling(20 *  trees[,3]/max(trees[,3]))
> sunflowerplot(trees[,2], trees[,1], n,
                seg.col = "black",
                xlab = "Height", ylab = "Girth",
          main = "Tree Volume")
```


Tree Volume
Girth
Height
20
18
16
14
12
10
8
65
70
75
80
85

Encoding Using Colour

- Bearing in mind that colour is not a good way of encoding numerical values, it is also possible to encode the third value as a colour.

```
> hue = (1 - .9 * trees[,3]/
                  max(trees[,3])) * 2 / 3
> plot(trees[,2], trees[,1],
       pch = 21, cex = 2, bg = hsv(hue),
       xlab = "Height", ylab = "Girth",
       main = "Tree Volume")
```


Tree Volume
Girth
Height
8
10
12
14
16
18
20
65
70
75
80
85

Four or More Directions

- In four dimensions our perceptual abilities fail us.
- We would not able to understand a four-dimensional display, even if one were available.
- A large number of indirect ways of examining high dimensional data have been developed.

The Iris Data

- This set of data was collected by a botanist - Edgar Anderson.
- It gives the widths and lengths of the petals and sepals of three species of Iris:
  - *Iris Setosa*
  - *Iris Versicola*
  - *Iris Virginica*
- The dataset is often used to test statistical techniques which attempt to distinguish different groupings on the basis of measurements.

The Iris Data

Sepal Length	Sepal Width	Petal Length	Petal Width
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
5.4	3.9	1.7	0.4
4.6	3.4	1.4	0.3
5.0	3.4	1.5	0.2
4.4	2.9	1.4	0.2
4.9	3.1	1.5	0.1
⋮	⋮	⋮	⋮

Neglected Iris Data

Iris Setosa

Iris Versicolor

Iris Virginica

Unfortunately, the data set doesn't contain the most important information about the Iris flowers.

Scatterplot Matrices (Draughtsman's Displays)

- A simple way to examine high dimensional datasets is to plot all possible pairs of variables.
- There are $p \times (p-1)$ scatter plots to be viewed.
  - There are p choices for the x variable.
  - For each x variable there are $p-1$ possible choices for the y variable.
- One way to display the plots is to lay them out a $p \times p$ matrix.
- This kind of display is called a *scatterplot matrix* or a *draughtsman's display*.

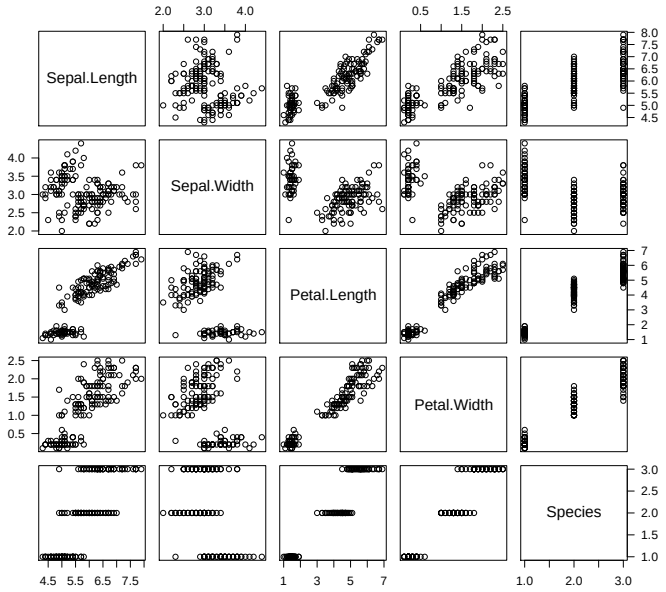
Scatterplot Matrices in R

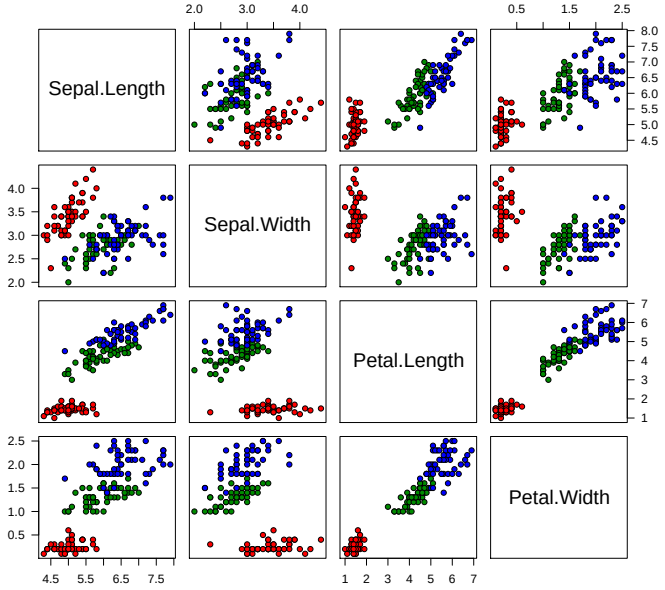
- The R function `pairs` produces a scatterplot matrix.

```
> pairs(iris)
```

- The function allows a degree of customisation – plotting symbol and default colour can be easily changed.

```
> fill = rep(c("red", "green4", "blue"),
             c(50, 50, 50))
> pairs(iris[,1:4], pch = 21, bg = fill)
```

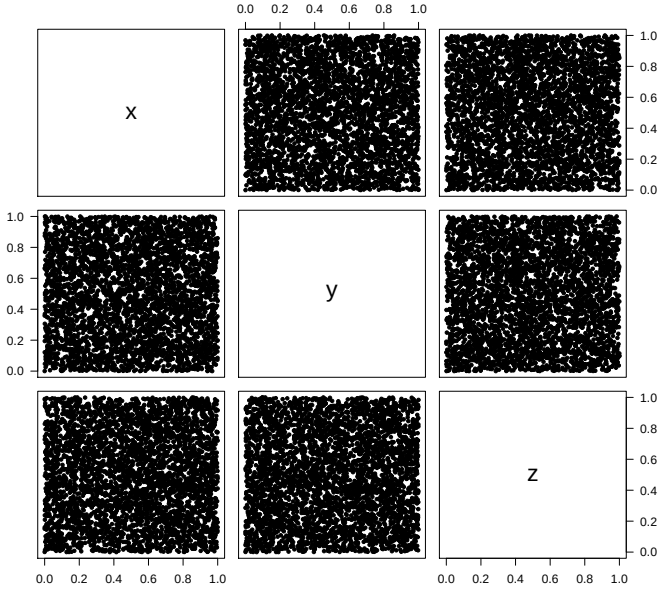

Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species

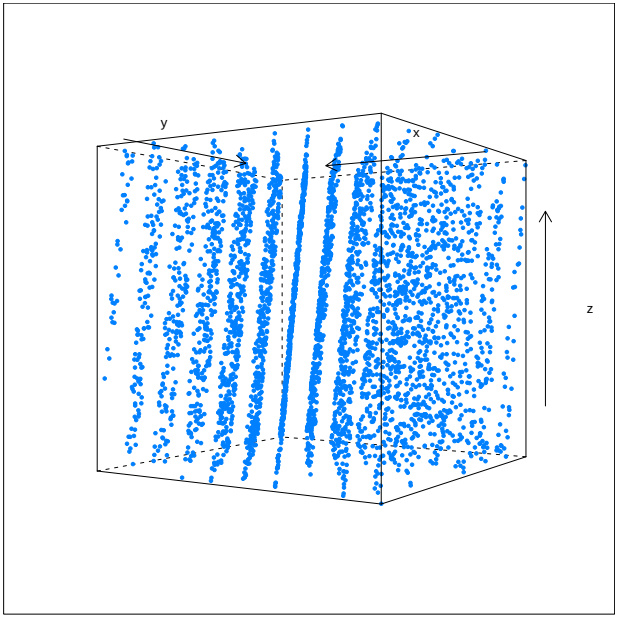

Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
2.0
3.0
4.0
0.5
1.5
2.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
7
6
5
4
3
2
1
2.5
2.0
1.5
1.0
0.5
4.5
5.5
6.5
7.5
1
2
3
4
5
6
7

Limitations of Scatterplot Matrices

- Scatterplot matrices can give a good overall view of a set of data values, but they also be misleading.
- This is because they only show a very limited view of the data. To illustrate the problem, we will look at the "randu" dataset.
- This data set consists of consecutive triples produced by the randu random number generator.

```
> pairs(randu)
> cloud(y ~ x * z, data=randu,
        screen = list(y = 147),
        perspective = FALSE)
```


x
y
z
0.0
0.2
0.4
0.6
0.8
1.0


y
x
z

Comments

- The consecutive triples produced by randu are constrained to lie on a series of parallel planes which cut through the unit cube.
- The paper which pointed this fact out was titled “The Random Numbers Fall Mainly on the Planes.”
- The planes are not aligned with the sides of the unit cube and so do not show up in any of the panels of the scatter plot.
- This problem can be even worse in higher dimensions.

Clustering

- One of the ways we seek to make sense of the world around us is by grouping the things we see about us into classes of similar objects.
- If the objects in a group are sufficiently similar and sufficiently distinct from other objects we may give them a common name — person, dog, chair, etc.
- In a further step, we may begin to create theories about the relationships between groups.
- In statistics, forming groups of similar objects is known as *cluster analysis* or *clustering*.

Clustering and Graphics

- There are a number of graphical techniques which aim to help users establish the degree to which observations are similar or different.
- All these techniques work by encoding each observations as a symbol or *glyph*.
- The visual system is very good at letting us detect visual simularity.
- This can form the basis for informally clustering observations.

Example – United States Voting

Percentage of Republican Votes
in Presidential Elections in Six Southern States
in the Years 1932–1940 and 1960–1968.

	1932	1936	1940	1960	1964	1968
Missouri	35	38	48	50	36	45
Maryland	36	37	41	46	35	42
Kentucky	40	40	42	54	36	44
Louisiana	7	11	14	29	57	23
Mississippi	4	3	4	25	87	14
South Carolina	2	1	4	49	59	39

Stars — A Simple Glyph

- One simple way of encoding the vote data is to draw a star with one arm for each voting year.
- The lengths of the arms will be proportional to the vote for the corresponding year.
- Each State will be encoded as a six-pointed star.

Creating a Star Plot

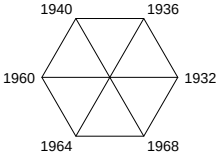
```
> votes =
  data.frame(yr1932 = c(35, 36, 40, 7, 4, 2),
             yr1936 = c(38, 37, 40, 11, 3, 1),
             yr1940 = c(48, 41, 42, 14, 4, 4),
             yr1960 = c(50, 46, 54, 29, 25, 49),
             yr1964 = c(36, 35, 36, 57, 87, 59),
             yr1968 = c(45, 42, 44, 23, 14, 39),
             row.names = c("Missouri", "Maryland",
               "Kentucky", "Louisiana", "Mississippi",
               "South Carolina"))
> stars(votes/max(votes), scale = FALSE,
        xlim =c(0, 8), ylim = c(0, 8),
        locations = cbind(c(1, 4, 1, 4, 1, 4),
                          c(7, 7, 4, 4, 1, 1)),
        col.stars = rep(hcl(180), 6),
        key.loc = c(7, 4),
        key.labels = c(1932, 1936, 1940, 1960, 1964, 1968))
```

Missouri

Maryland

Kentucky

Louisiana



Mississippi

South Carolina

Interpretation

- Clearly Missouri, Maryland and Kentucky exhibit very similar voting patterns.
- They can be regarded as forming a cluster.
- Louisiana, Mississippi and South Carolina are different from each other and the other cluster.
- Many other glyphs have been proposed.

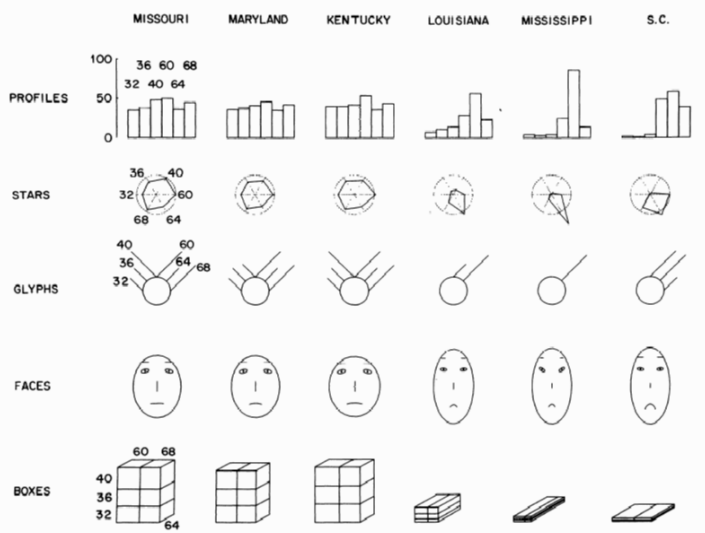


Figure 1. Profiles, Stars, Glyphs, Faces, and Boxes of Percentage of Republican Votes in Six Presidential Elections in Six Southern States. The Circles in the Stars Are Drawn at 50%. The Assignment of Variables to Facial Features in the Faces Is: 1932—Shape of Face; 1936—Length of Nose; 1940—Curvature of Mouth; 1960—Width of Mouth; 1964—Slant of Eyes; 1968—Length of Eyebrows

Critique

- Glyphs work well when there are just a few observations.
- With even moderate numbers of observations the ability of the brain to group the observations is overwhelmed.
- Little is known about how well our interpretation of the similarity of glyphs corresponds to the true similarity between the observations.
- In the case of faces, there are likely to be strong cultural and gender biases in an individuals groupings.

Representation as Functions

- Since glyphs only work well for a small number of observations, attempts have been made to look at other techniques for representing multivariate data.
- One of the more interesting is the idea that observations can be represented and plotted as functions.

Andrews Plots

- Andrews plots represent an observation $(x_1, \ldots, x_n)$ in the form:

$$\begin{aligned} f(t) =& x_1/\sqrt{(2)} + x_2 \sin t + x_3 \cos t \\ &+ x_4 \sin 2t + x_5 \cos 2t + \cdots \end{aligned}$$

- This function is graphed over the interval $[-\pi, \pi]$.
- It is possible to superimpose the functions associated with many observations on the same graph.

Properties of Andrews Plots I

- The function mapping preserves means.

$$f_{\bar{\mathbf{x}}}(t) = \frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{x}_i}(t)$$

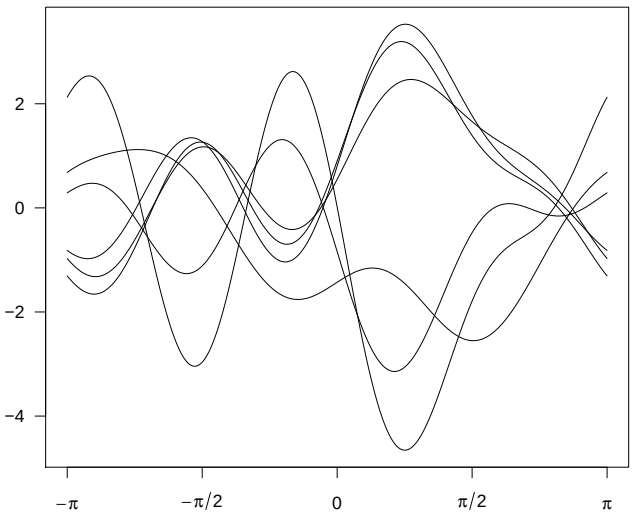
- The function mapping preserves distances.

$$\frac{1}{\pi}\int_{-\pi}^{\pi} |f_{\mathbf{x}}(t) - f_{\mathbf{y}}(t)|^2 dt = \|\mathbf{x} - \mathbf{y}\|^2$$

- The function mapping preserves linear relationships. If $\mathbf{y}$ lies on the line joining $\mathbf{x}$ and $\mathbf{z}$ then $f_{\mathbf{y}}(t)$ lies between $f_{\mathbf{x}}(t)$ and $f_{\mathbf{z}}(t)$ for all t.

Properties of Andrews Plots II

- For each t, the function mapping produces a projection of the data onto a one-dimensional subspace. Thus, each t tells us about a particular aspect of the data set. If two functions take on different values for some t there are important differences between the observations.
- Andrews plots are a useful tool for looking for clusters and outliers.


An Andrews Plot for the Votes Data
2
0
−2
−4
−π
−π/2
0
π/2
π

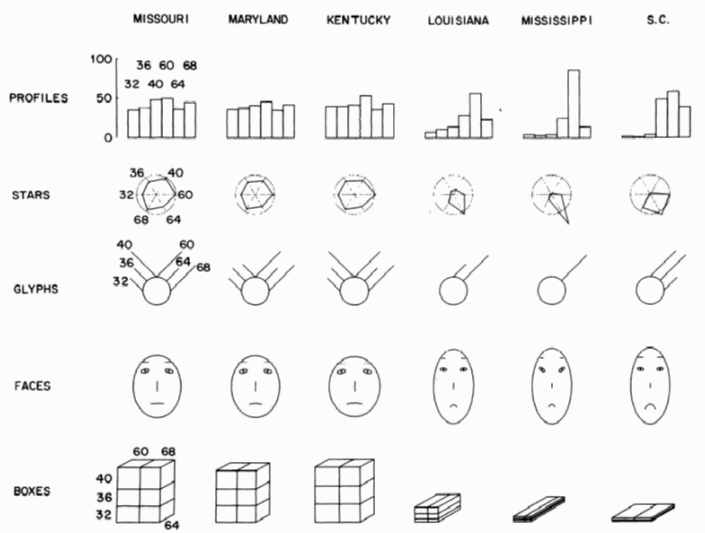
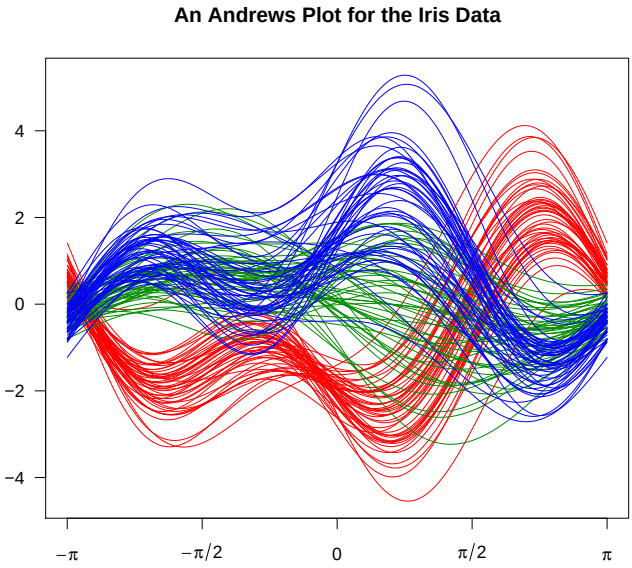
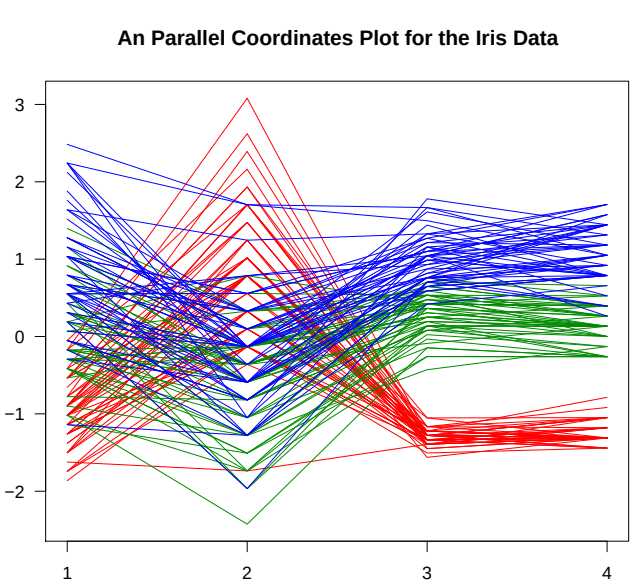


Figure 1. Profiles, Stars, Glyphs, Faces, and Boxes of Percentage of Republican Votes in Six Presidential Elections in Six Southern States. The Circles in the Stars Are Drawn at 50%. The Assignment of Variables to Facial Features in the Faces Is: 1932—Shape of Face; 1936—Length of Nose; 1940—Curvature of Mouth; 1960—Width of Mouth; 1964—Slant of Eyes; 1968—Length of Eyebrows


An Andrews Plot for the Iris Data
4
2
0
−2
−4
−π
−π/2
0
π/2
π

Parallel Coordinate Plots

- Parallel coordinate plots are quite similar in nature to Andrews plots.
- The jth variable is assigned position j on the x axis and the points for that variable are plotted against the y axis at that position.
- The coordinates for all the variables of the ith observation are joined by straight-line segments.
- The plots reveal clusters and outliers in the same way that andrews plots do.

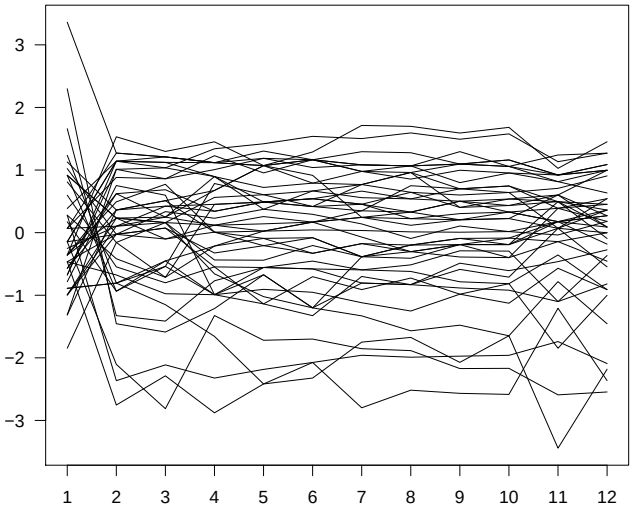

An Parallel Coordinates Plot for the Iris Data
3
2
1
0
−1
−2
1
2
3
4

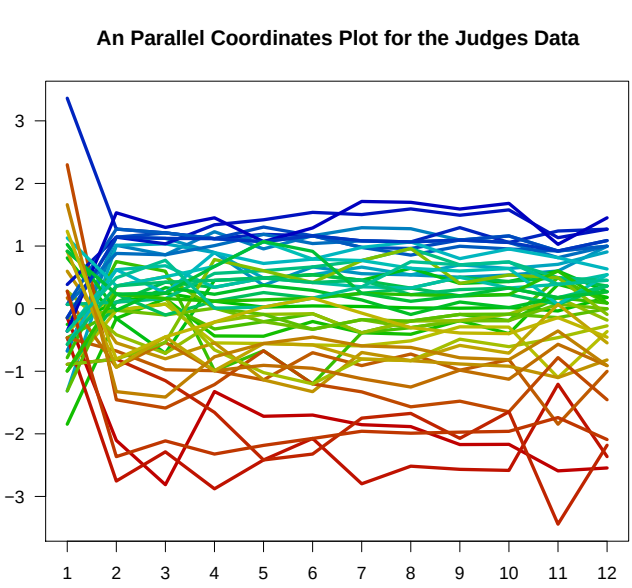
Judge Rating Data

Lawyers' ratings of state judges in the Connecticut State Court. The variables are:

CONT	Number of contacts of lawyer with judge.
INTG	Judicial integrity.
DMNR	Demeanor.
DILG	Diligence.
CFMG	Case flow managing.
DECI	Prompt decisions.
PREP	Preparation for trial.
FAMI	Familiarity with law.
ORAL	Sound oral rulings.
WRIT	Sound written rulings.
PHYS	Physical ability.
RTEN	Worthy of retention.

An Parallel Coordinates Plot for the Judges Data




An Parallel Coordinates Plot for the Judges Data
3
2
1
0
−1
−2
−3
1
2
3
4
5
6
7
8
9
10
11
12

A Parallel Coordinates Function

```
pcoord =
    function(x, scale.data = TRUE,
             col = "black", lty = "solid", lwd = 1)
    {
        if (scale.data)
            x = scale(x)
        nobs = nrow(x)
        col = rep(col, length = nobs)
        lty = rep(lty, length = nobs)
        lwd = rep(lwd, length = nobs)
        matplot(1:ncol(x), t(x),
                type = "l",
                col = col,
                lty = lty,
                lwd = lwd, axes = FALSE, ann = FALSE)
        axis(1, at = 1:ncol(x)); axis(2); box()
    }
```