# Mineria de Datos

Data Visualization

Dr. Edgar Acuna

Departmento de Matematicas

Universidad de Puerto Rico- Mayaguez

academic.uprm.edu/eacuna

# Outline

- Visualization role

- Representing data in 1,2, and 3-D

- Representing data in 4+ dimensions

  - Scatterplot Matrix

  - Survey plots

  - Parallel coordinates

  - Radviz, Starcoord

# The visualization role

- Visualization is the process of transforming information into a visual form enabling the user to observe the information.

- Using successful visualizations for data mining and knowledge discovery tasks can reduce the time it takes to understand the underlying data, find relationships, and discover information.

- One of the goals of visualization is *explorative analysis.*

# Visualization role (cont)

- In *explorative analysis*, visualization techniques are applied prior to the application of classification techniques to obtain insight into the characteristics of the dataset. The process involves a search for structures and the result is a visualization of the data which provides a hypothesis about the data.

- This use of visualization may improve the understanding that users have of their data, thereby, increasing the likelihood that new and useful information will be gained from the data.

# Visualization Role

- Support interactive exploration

- Help in result presentation

- Disadvantages:

  - requires human eyes

  - Can be misleading

# Tufte's Principles of Graphical Excellence

- Give the viewer

  - the greatest number of ideas

  - in the shortest time

  - with the least ink in the smallest space.

- Tell the truth about the data!

**(E.R. Tufte, "The Visual Display of Quantitative Information", 2nd edition)**

# Visualization Methods

- Visualizing in 1-D, 2-D and 3-D
  - well-known visualization methods

- Visualizing more dimensions
  - Scatterplot matrix
  - Survey plots
  - Parallel Coordinates
  - Radviz
  - Star Coordinates

# 1-D (Univariate) data ( R)

- stripchart(x,vertical=T,col=2) #Dotplot

- hist(x,col=3) #Histogram
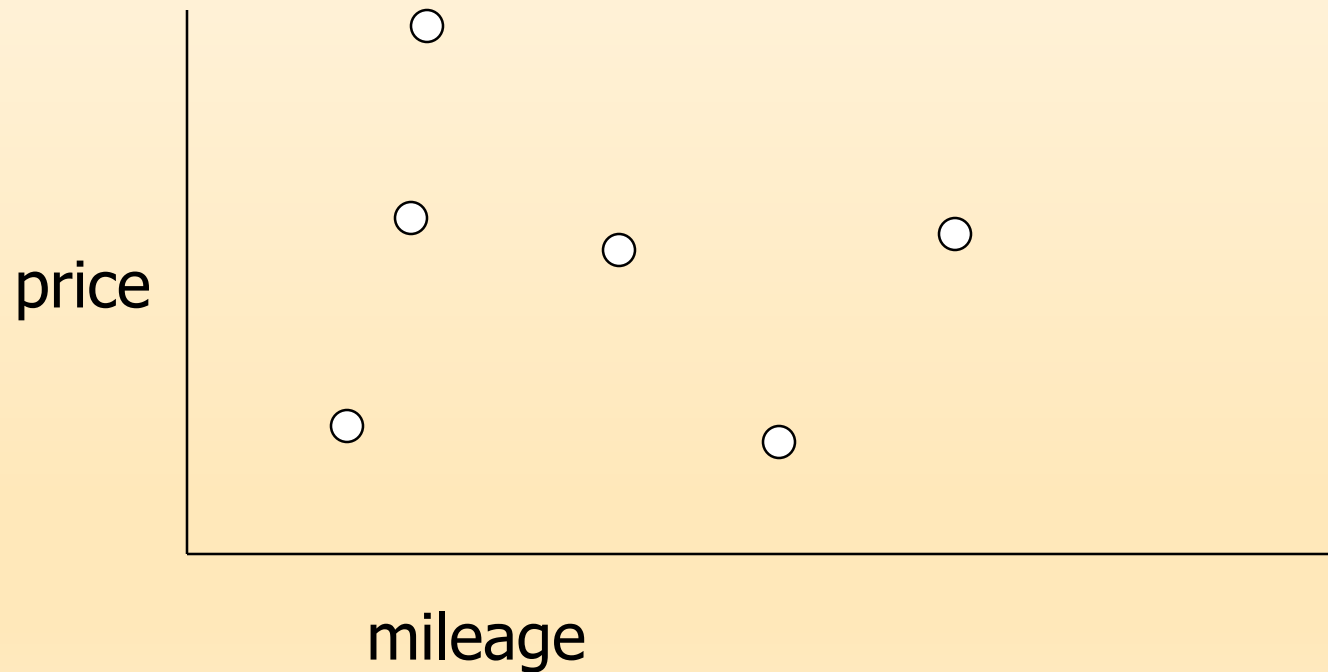
- boxplot(x,horizontal=T,col="blue") #Boxplot

# 1-D (Univariate) Data

- Representations



Histogram

Tukey box plot

low    Middle 50%    high
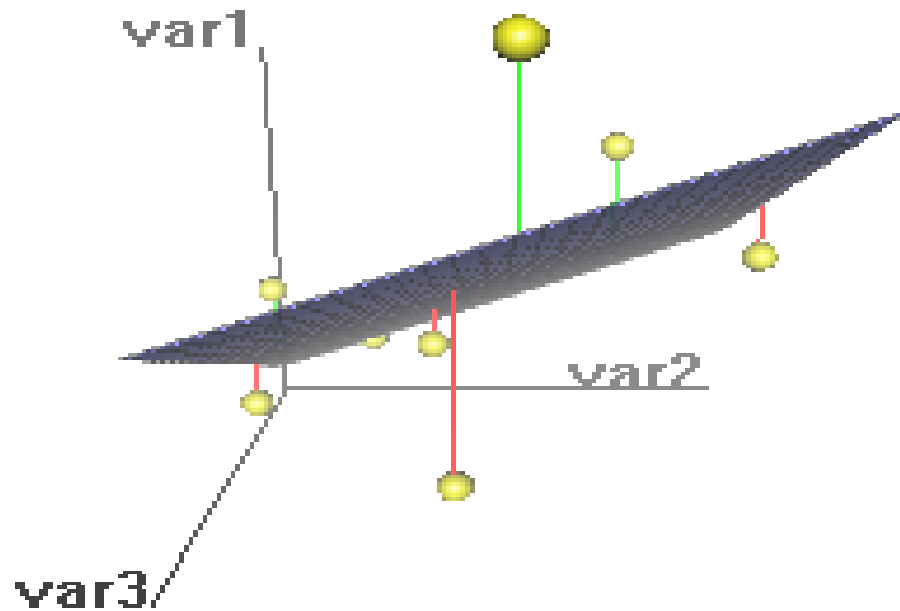
Median

0    20
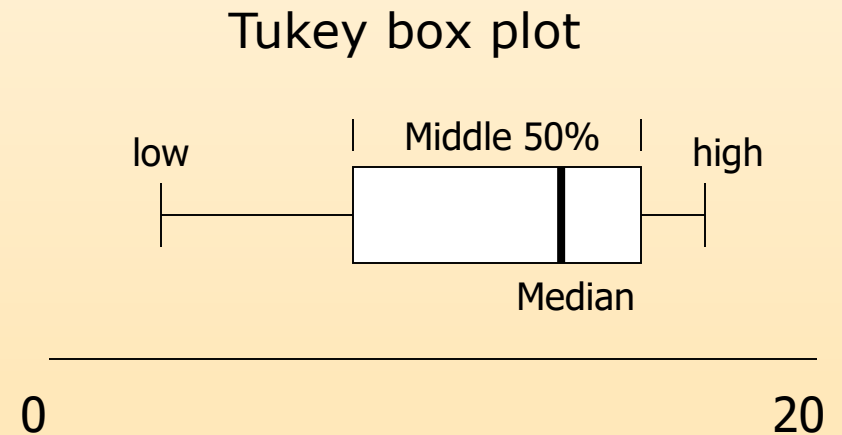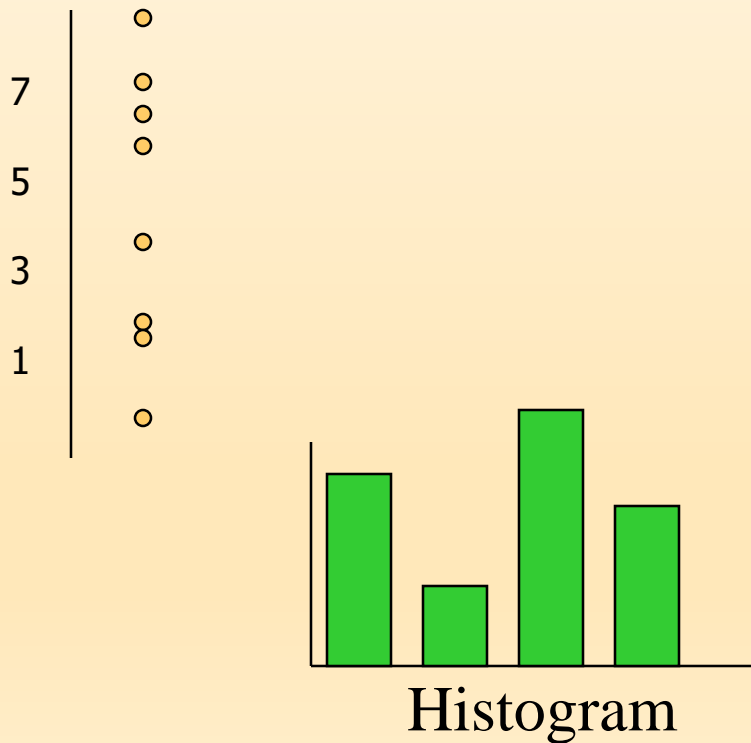
# 2-D (Bivariate) Data

- Scatter plot, …

# 3-D Data

- Scatter3d in Rcmdr library

- Scatterplot3 in the scatterplot3d library

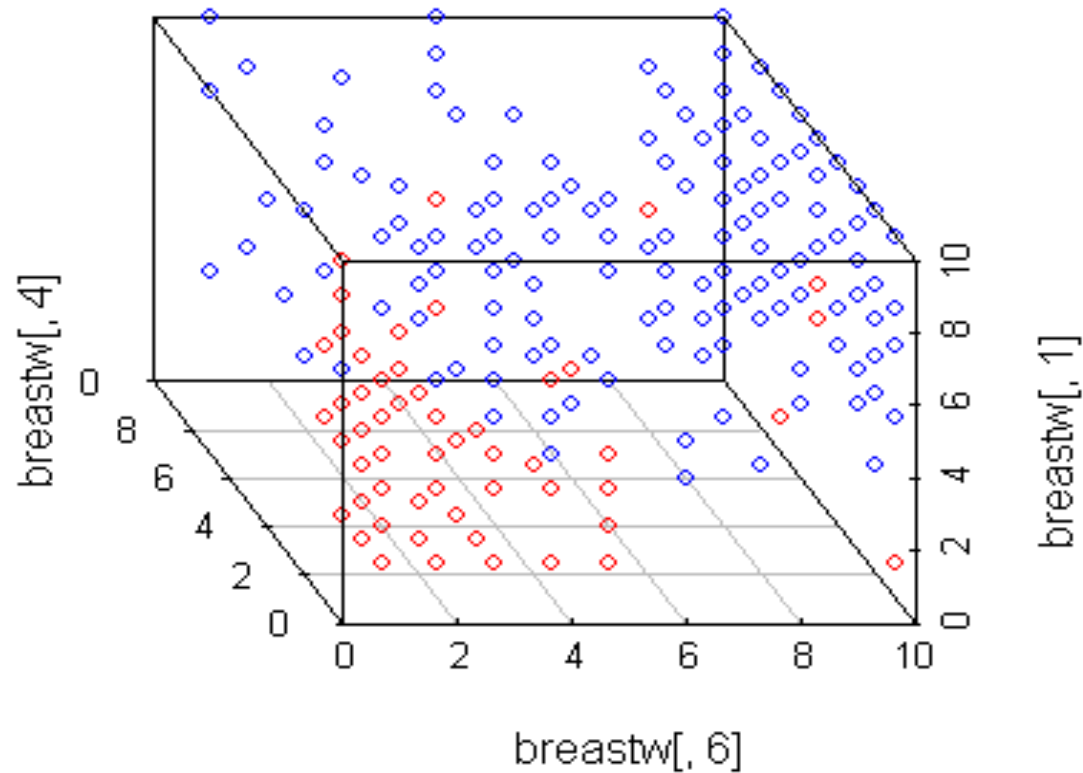- Cloud() in lattice library. Lattice is a free version of trellis.
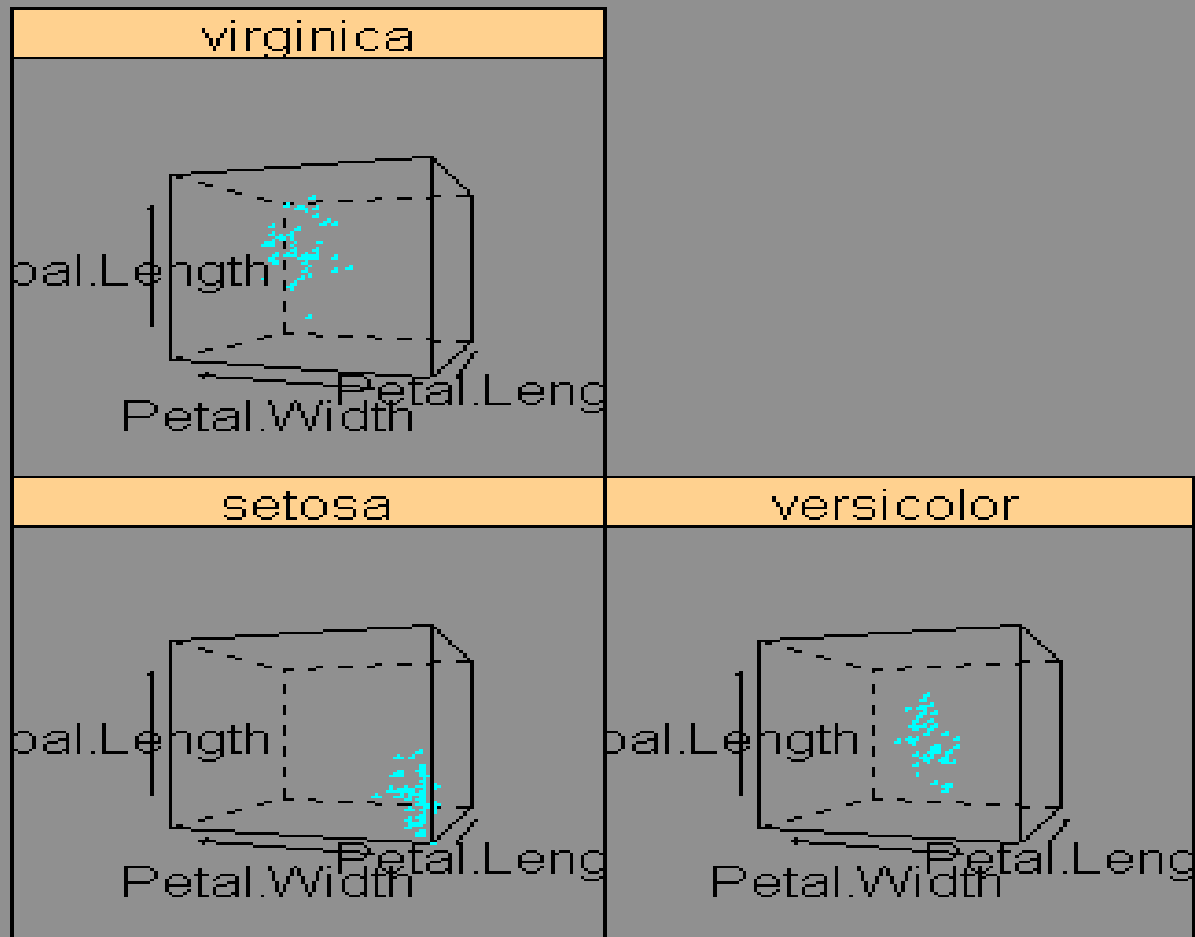
# Scatter3d from Rcmdr library

# Scatterplot3d from scatterplot3d library

scatterplot3d(breastw[,1],breastw[,4],breastw[,6],color)


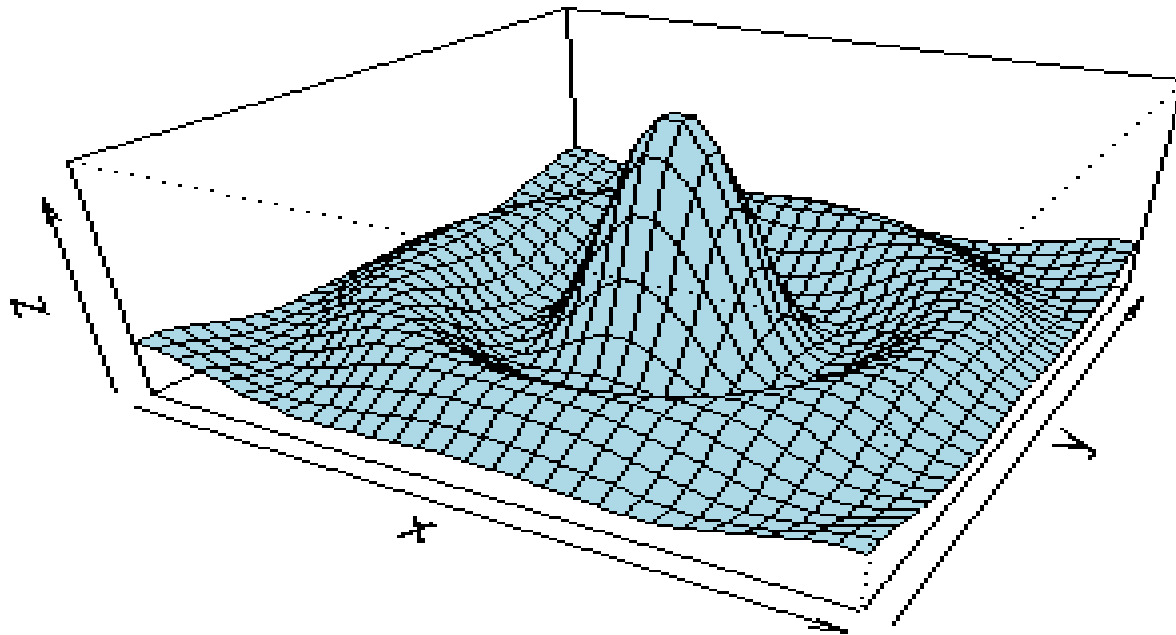
scatterplot3d de breastw(3 main features)

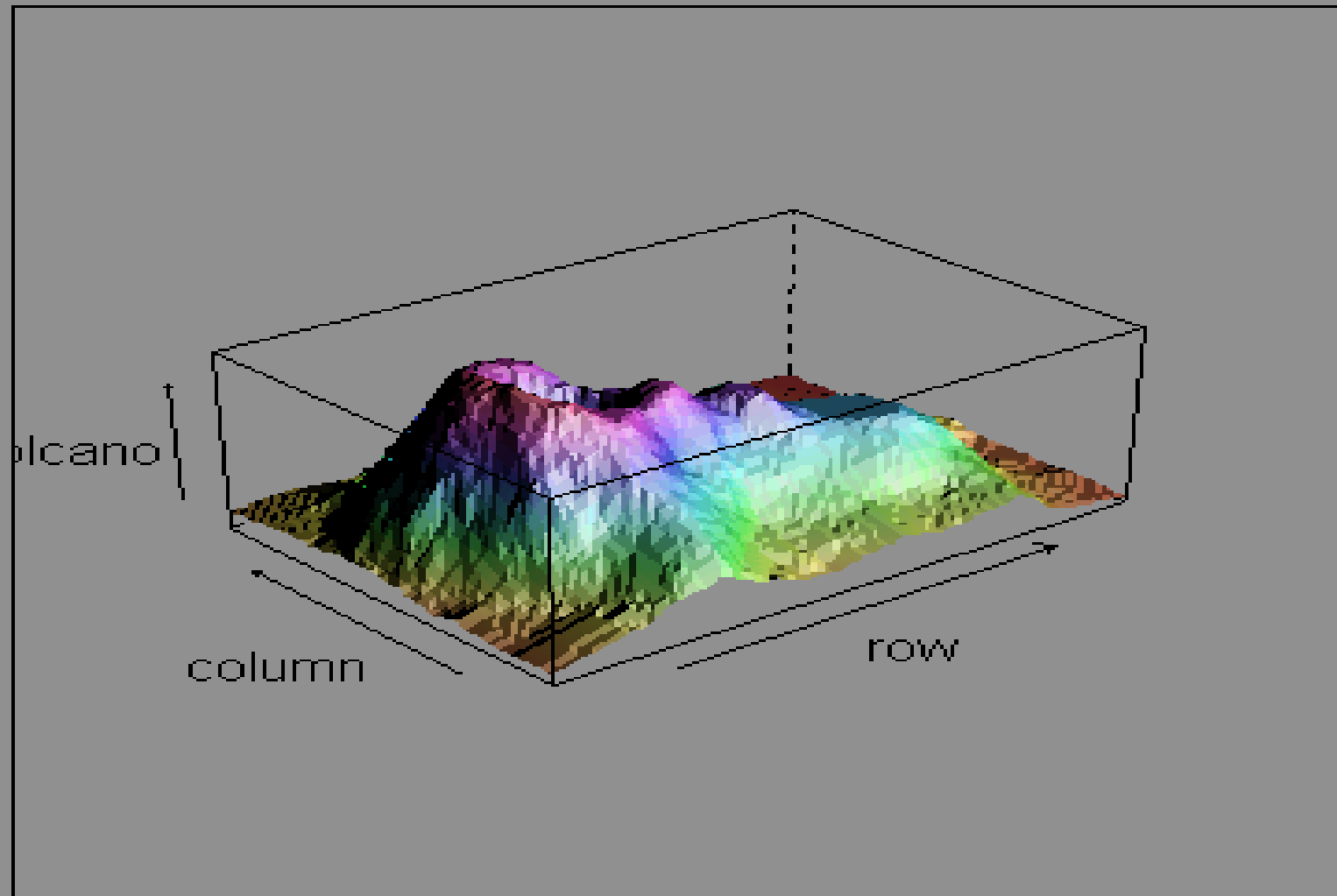# Cloud() from the lattice library

# 3-D Data (persp)



```
> x =seq(-10, 10, length= 30); y=x
> f =function(x,y) { r <- sqrt(x^2+y^2); 10 * sin(r)/r }
> z =outer(x, y, f); z[is.na(z)] = 1; op = par(bg = "white")
> persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue")
```

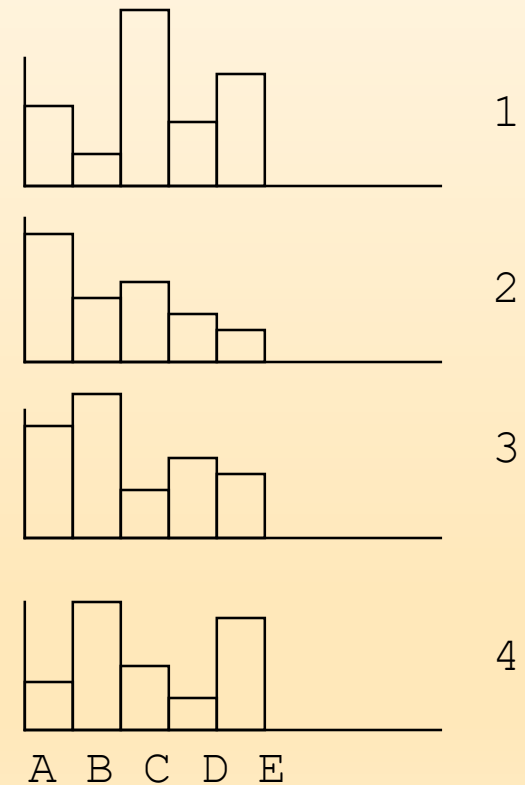# 3-D wireframe(lattice)

# Visualizing in 4+ Dimensions

- Scatterplot Matrix

- Survey Plot

- Parallel coordinate plot

- Radviz

- Star Coordinates

# Multiple Views

Give each variable its own display



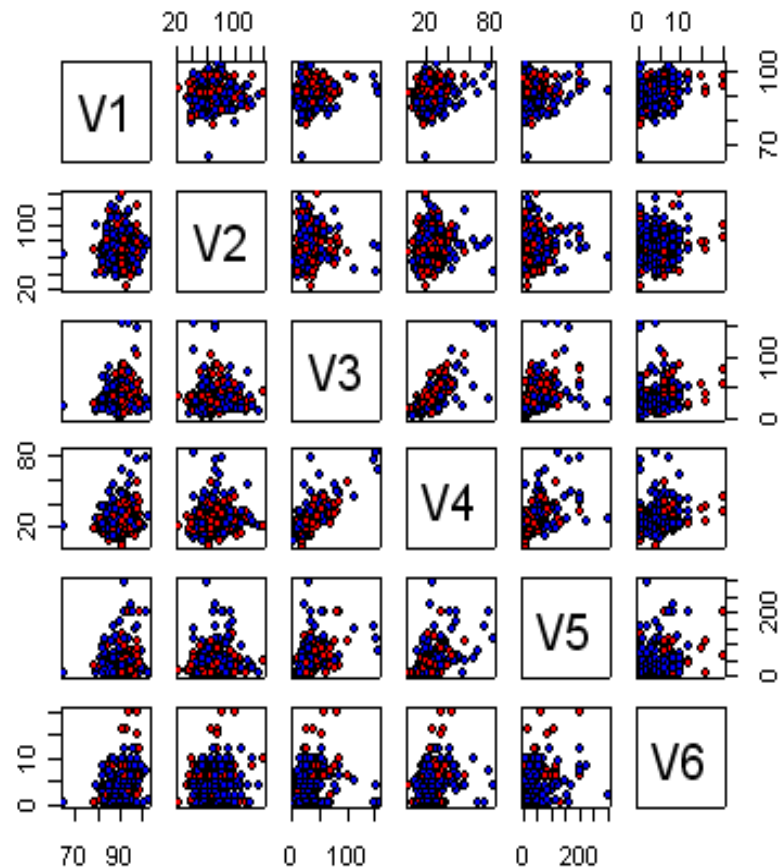|   | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 4 | 1 | 8 | 3 | 5 |
| 2 | 6 | 3 | 4 | 2 | 1 |
| 3 | 5 | 7 | 2 | 4 | 3 |
| 4 | 2 | 6 | 3 | 1 | 5 |

Problem: does not show correlations

# pairs() Scatterplot Matrix

Represent each possible
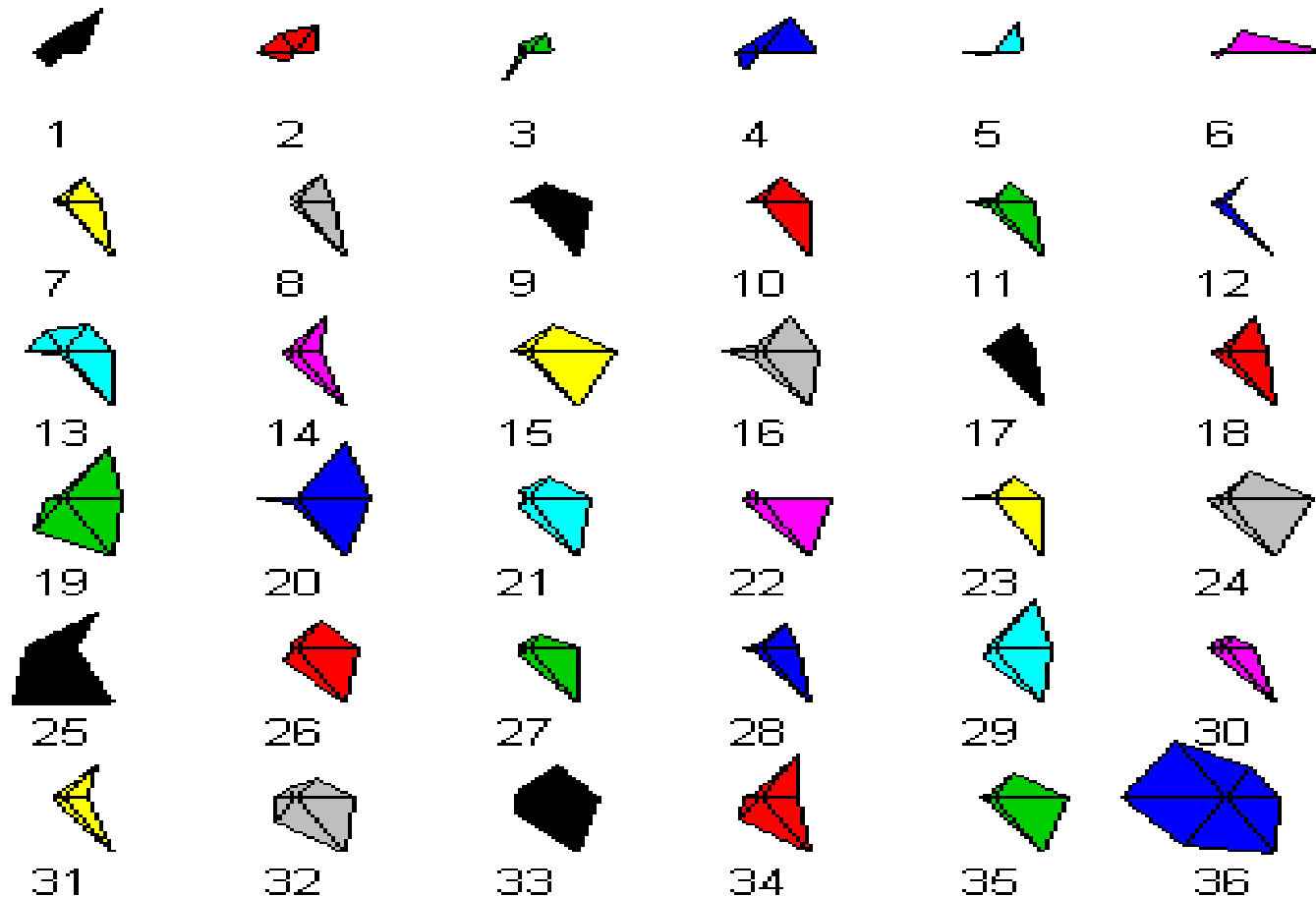pair of variables in their
own 2-D scatterplot
(car data)
*Useful for detecting*
linear correlations
(e.g. V3 & V4)

*But misses*
multivariate effects

# Star Plots (Chambers et al., 1983)



stars plot for bupa(instances 1:36)

# Visualization function in *dprep*:

- *imagmiss( )* determine the existence of missing values in the dataset, identify their location and quantity.

- *surveyplot( )* constructs a survey plot of the dataset

- *parallelplot( )* constructs a parallel coordinate plot of the data

- Starcoord(), Starcoor3d()

- Radviz()

# The survey plot (Lohninger, 1994)

- A visualization invented by a French cartographer, Jacques Bertin, that is closely related to the visualization techniques: bar graph and permutation matrix.

- Consists of $n$ rectangular areas or lines, one for each dimension of the dataset, that are vertically arranged.

- Each data value of an attribute is mapped to a point on the vertical line and the point is extended to a line with length proportional to the corresponding value.

- The strength of this visualization lays in its ability to show the relations and dependencies between any two attributes, especially when the data is sorted on a particular dimension.
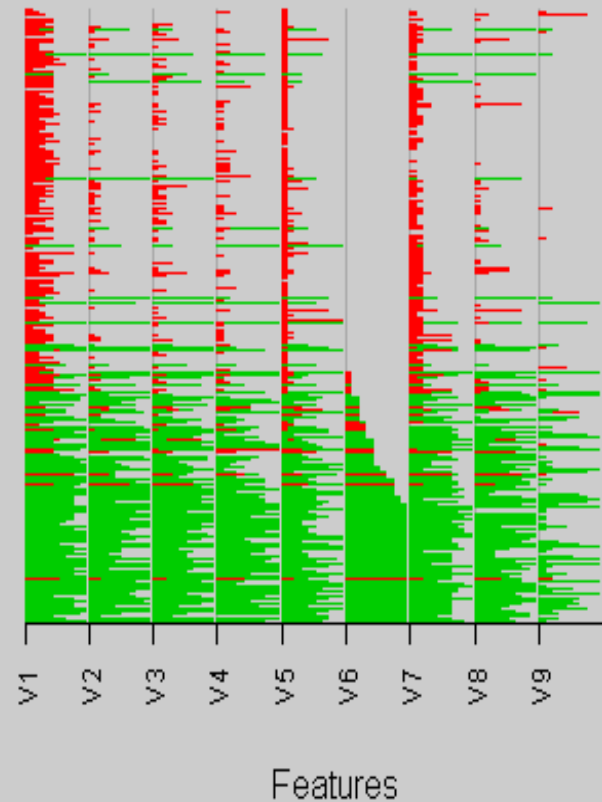
# The survey plot:

surveyplot(dataset: matrix , name: string, class: integer, orderon: integer, obs: list of integer )
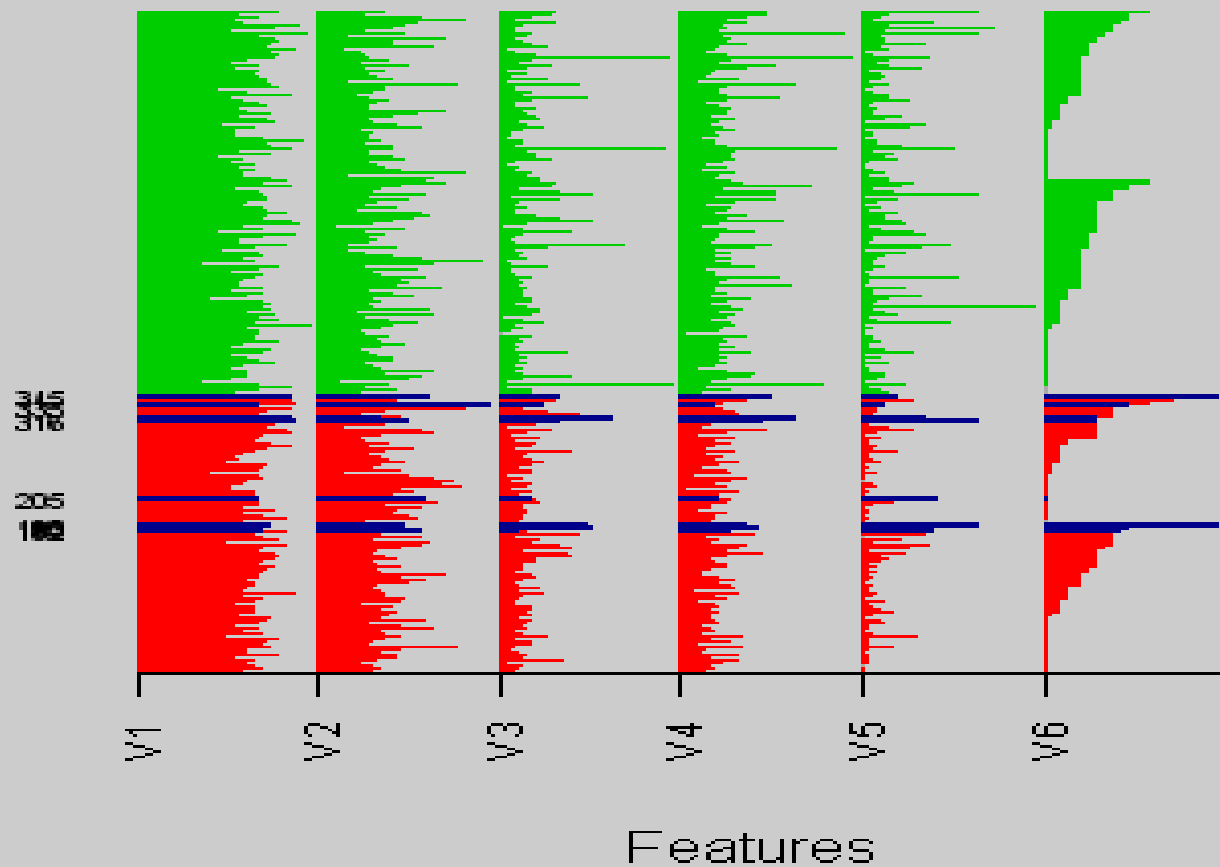


Survey Plot for breastw(ordered by attr=7
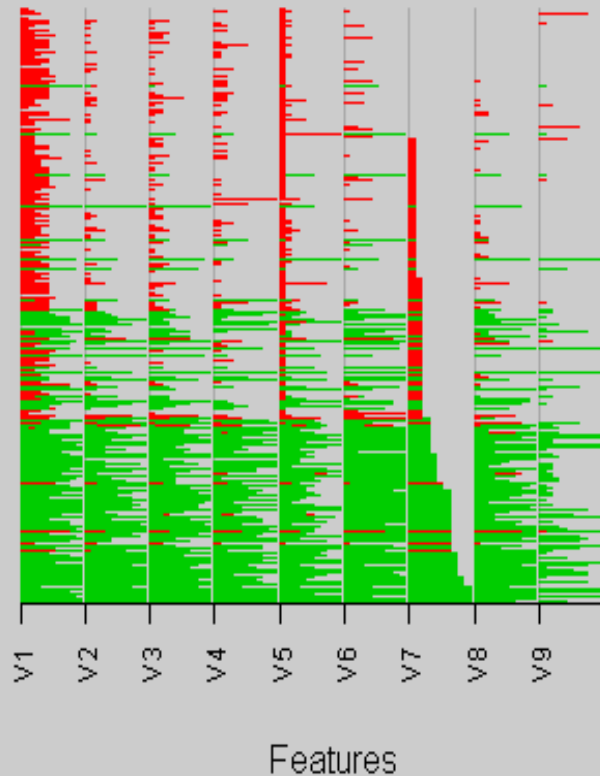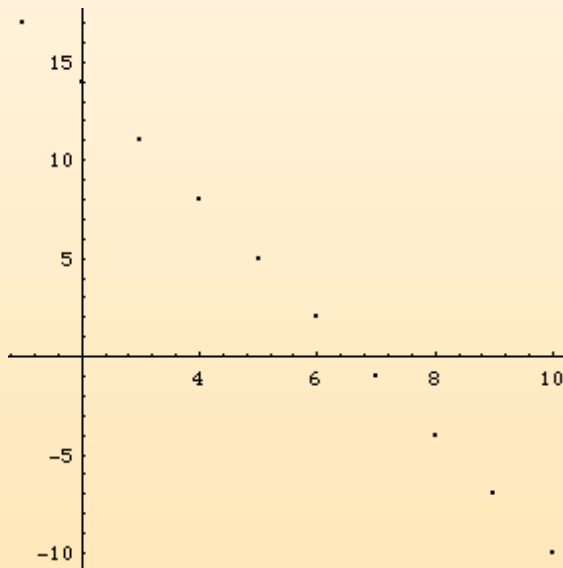


Survey Plot for breastw(ordered by attr=6

# Surveyplot as a tool to detect outliers



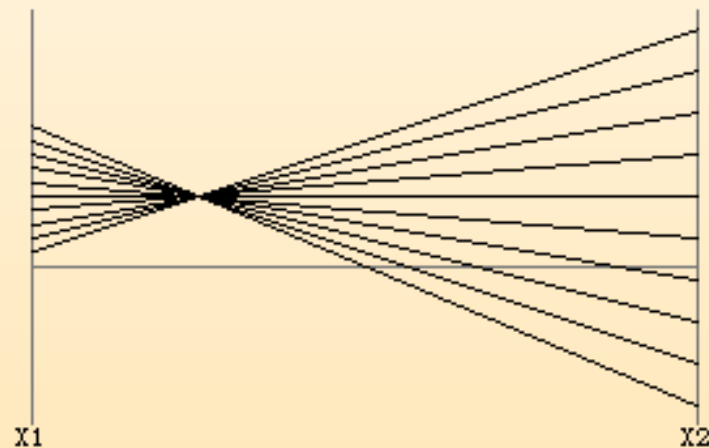Survey Plot for bupa(outliers class 1)

# Parallel Coordinates

- Encode variables along a horizontal row
- Vertical line specifies values



Dataset in a Cartesian coordinates



Same dataset in parallel coordinates

Invented by Alfred Inselberg
while at IBM, 1985

# The parallel coordinate plot:

- The parallel coordinate plot, described by Al Inselberg (1985), represents multidimensional data using lines.

- Whereas in traditional Cartesian coordinates all axes are mutually perpendicular, in parallel coordinate plots, all axes are parallel to one another and equally spaced.

- In this approach, a point in $m$-dimensional space is represented as a series of $m$-1 line segments in 2-dimensional space. Thus, if the original data observation is written as $(x_1, x_2, \ldots x_m,)$, then its parallel coordinate representation is the $m$-1 line segments connecting points $(1,x_1)$, $(2,x_2)$, . . . $(m,x_m)$.

- Typically, features will be standardized before a parallel coordinate plot is drawn.

# Example: Visualizing Iris Data



Iris setosa

| sepal length | sepal width | petal length | petal width |
|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 |
| 4.9 | 3 | 1.4 | 0.2 |
| ... | ... | ... | ... |
| 5.9 | 3 | 5.1 | 1.8 |



Iris versicolor



Iris virginica

# Parallel Coordinates

Sepal
Length

5.1 ⚬

| sepal length | sepal width | petal length | petal width |
|--------------|-------------|--------------|-------------|
| 5.1 | 3.5 | 1.4 | 0.2 |

# Parallel Coordinates: 2 D

Sepal
Length

Sepal
Width

3.5

5.1
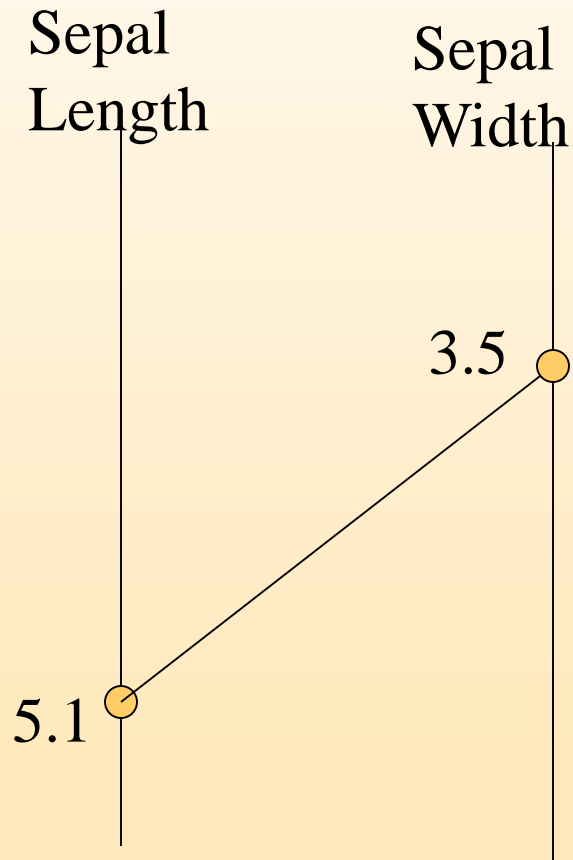
| sepal length | sepal width | petal length | petal width |
|:---:|:---:|:---:|:---:|
| 5.1 | 3.5 | 1.4 | 0.2 |

# Parallel Coordinates: 4 D
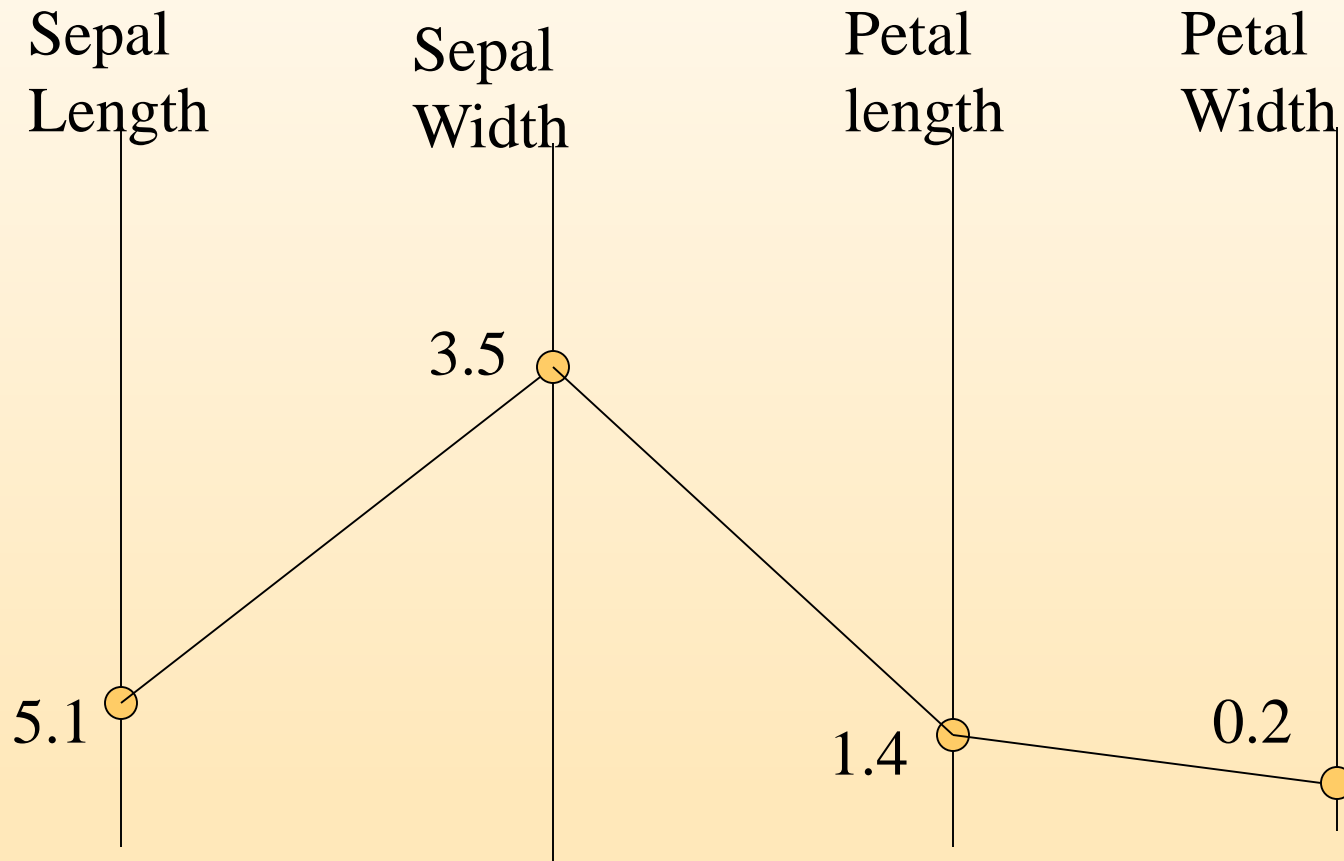


| sepal length | sepal width | petal length | petal width |
|:---:|:---:|:---:|:---:|
| 5.1 | 3.5 | 1.4 | 0.2 |

# Parallel Visualization of Iris data
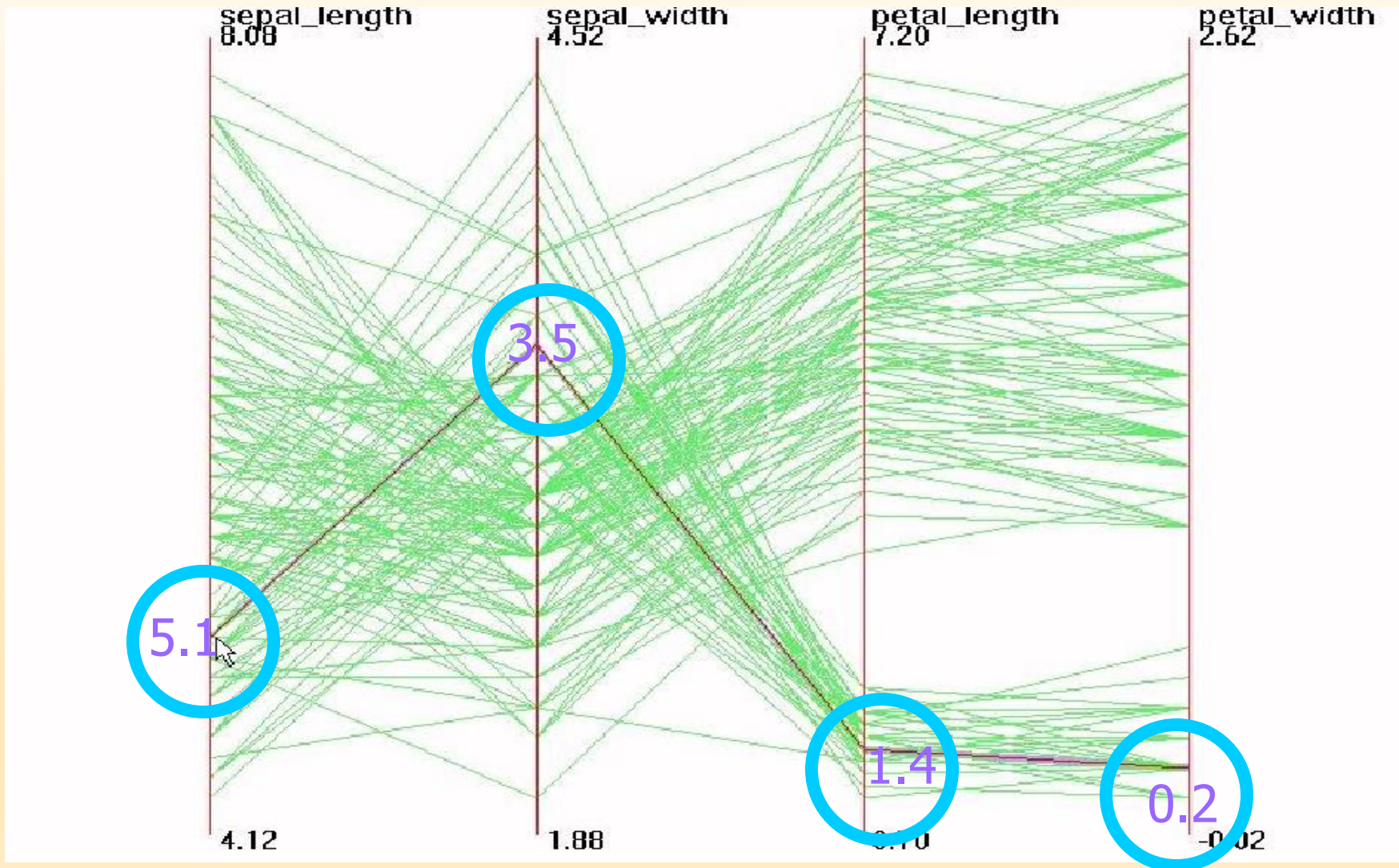
# Parallelplot (cont)

- Pairwise comparison is limited to those axis that are adjacent.

- For a dataset with p attributes there are p! permutations of the attributes so each of them is adjacent to every attribute in some permutation.

- Wegman (1990) determined that only [(p+1)/2] permutations are needed.([.] is the greatest integer function).

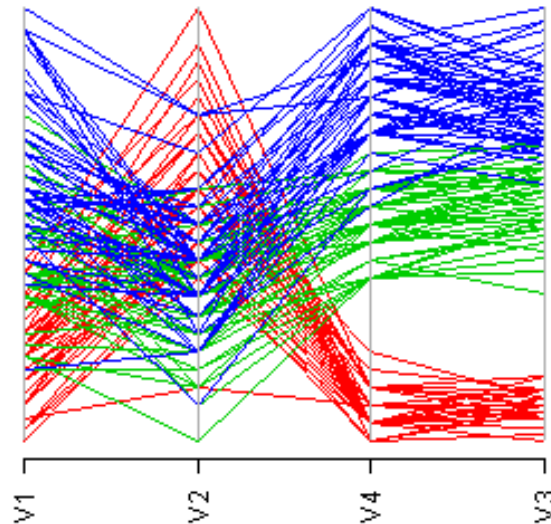# The parallel coordinate plot

parallelplot(dataset: matrix , name: string, class: integer, comb: integer, obs: list of integer )
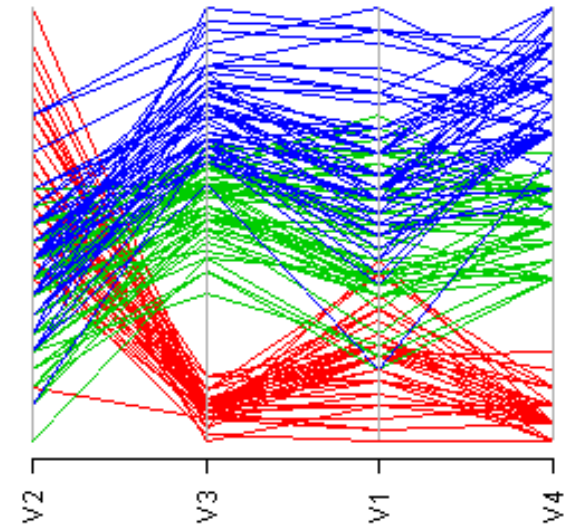
**Iris dataset:**

•**Data on the flowers.**

•**4 attributes (sepal length, sepal width petal length, and petal width,)**

•**150 instances**

•**3 classes (Setosa, Versicolor, Virginica)**

•**No missing values.**



Parallel Coordinate Plot for Iris — Combination # 1
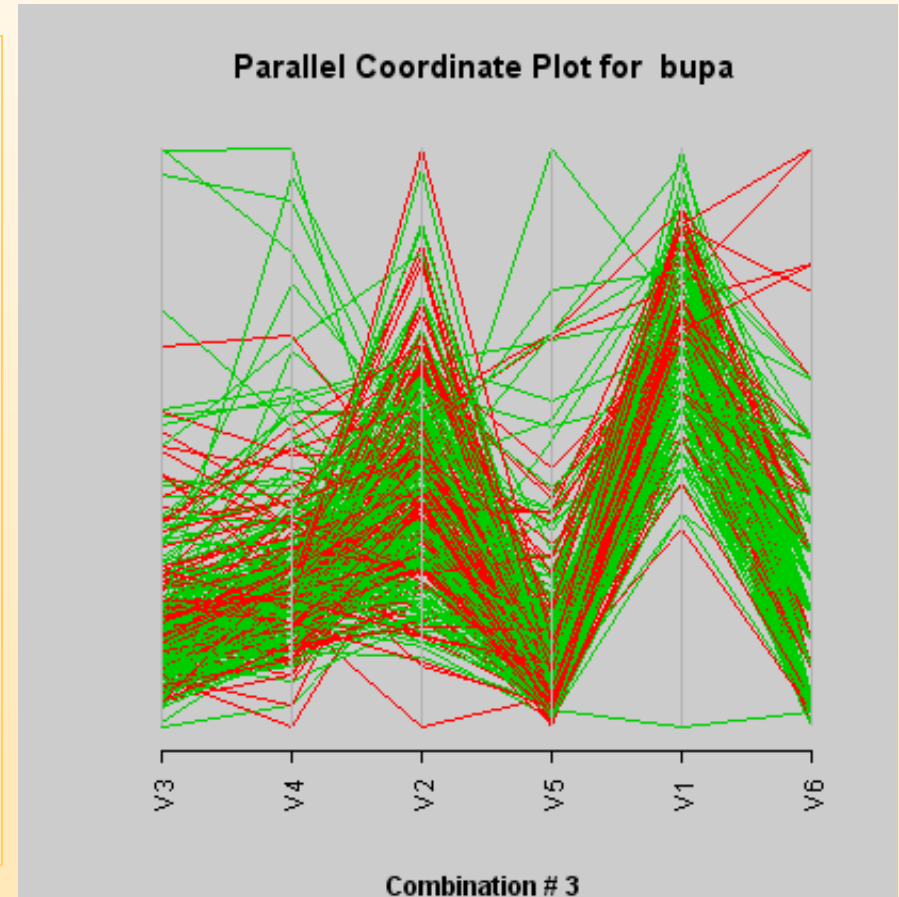
Parallel Coordinate Plot for Iris — Combination # 2

**Interpretation:**
•**Each different color represents a different class.**
•**If two attributes are highly positively correlated, lines passing from one feature to another tend not to intersect between the parallel coordinate axes.**
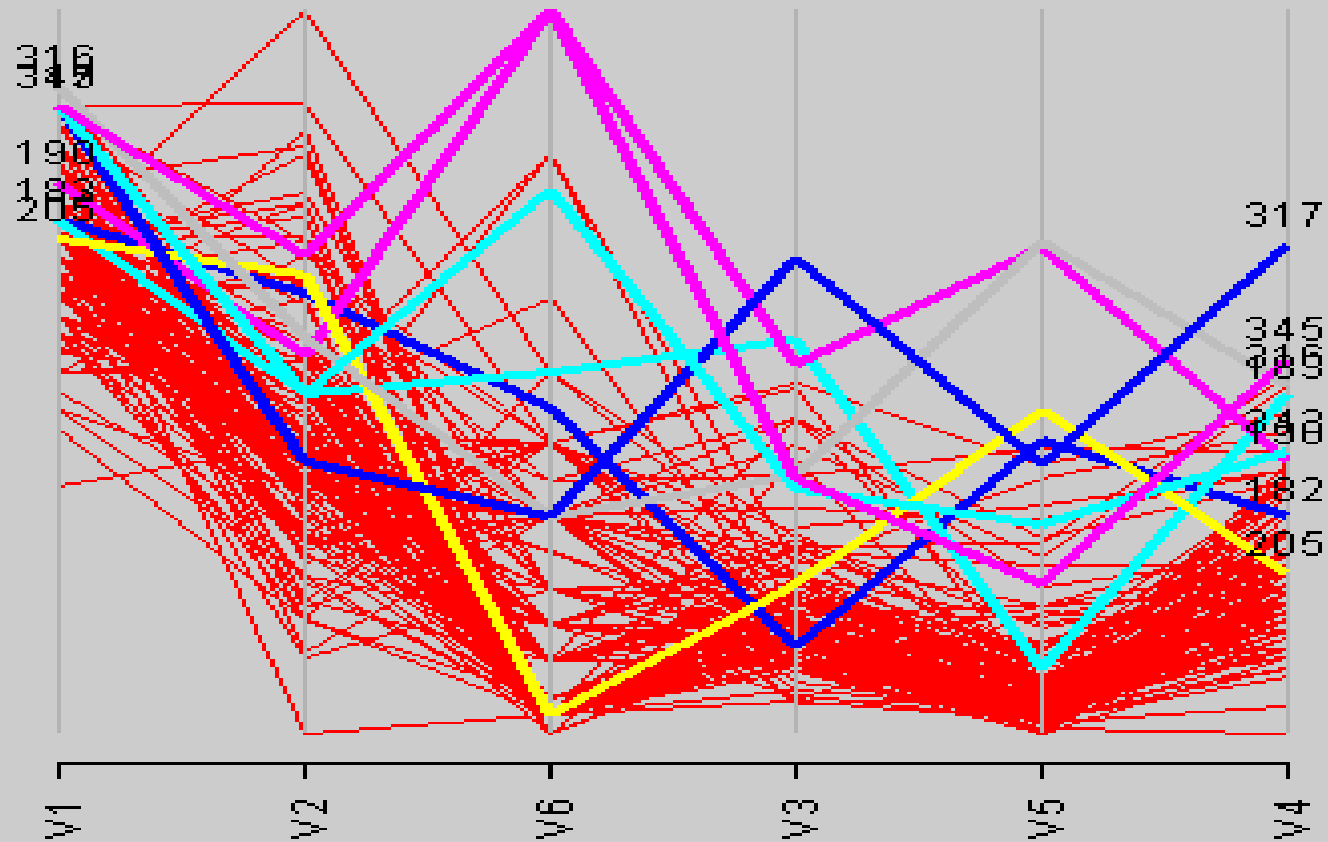
# The parallel coordinate plot

•**For highly negatively correlated attributes, the line segments tend to cross near a single point between the two parallel coordinate axes.**
•**The presences of outliers is suggested by poly-lines that do not follow the pattern for their class.**
**Some discrimination can be observed for several features.**
**One limitation of this displays is the loss of the information that is encoded into the lines between the axes for discrete, heterogeneous data attributes.**



Parallel Coordinate Plot for bupa

Combination # 3

# Parallelplot as a tool to detect outliers



Parallel Coordinate Plot for bupa (outliers class 1)

# Parallel Visualization Summary

- Each data point is a line

- Similar points correspond to similar lines

- Lines crossing over correspond to negatively correlated attributes

- Interactive exploration and clustering


- Problems: order of axes, limit to ~20 dimensions

# RadViz (Ankerst, et al., 1996)

- a radial visualization

- One spring for each feature .

- One end attached to perimeter point where the feature position is located. The other end attached to a data point.

- Each data point is displayed inside the circle where the sum of the spring forces equals 0.
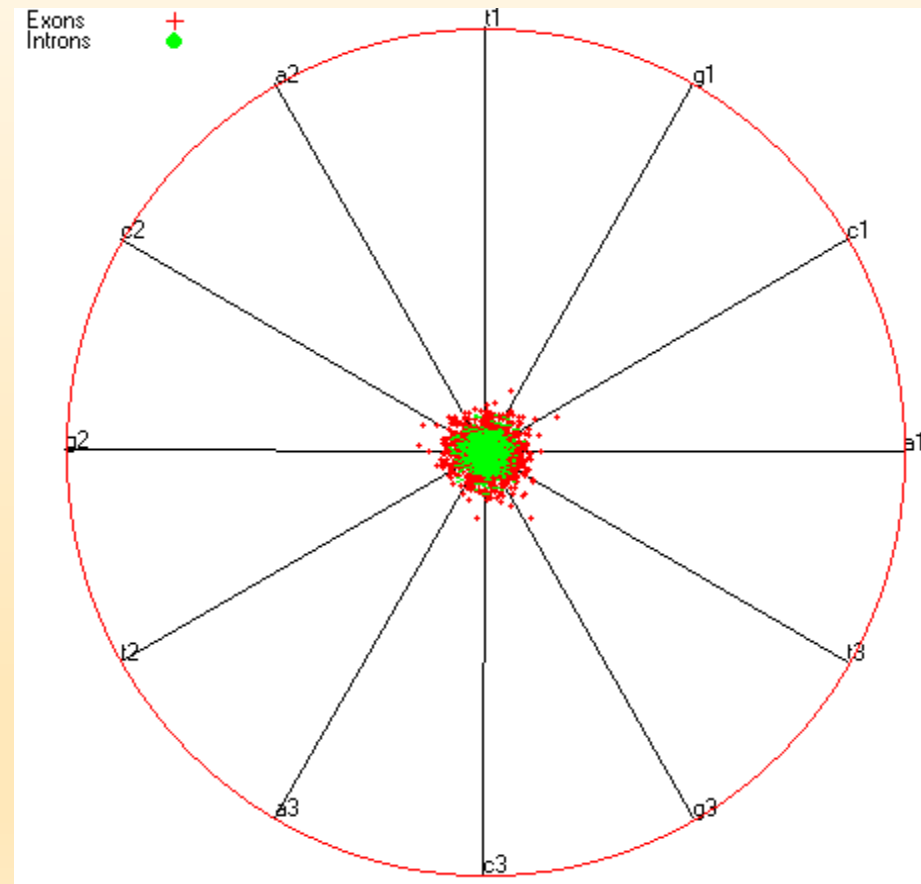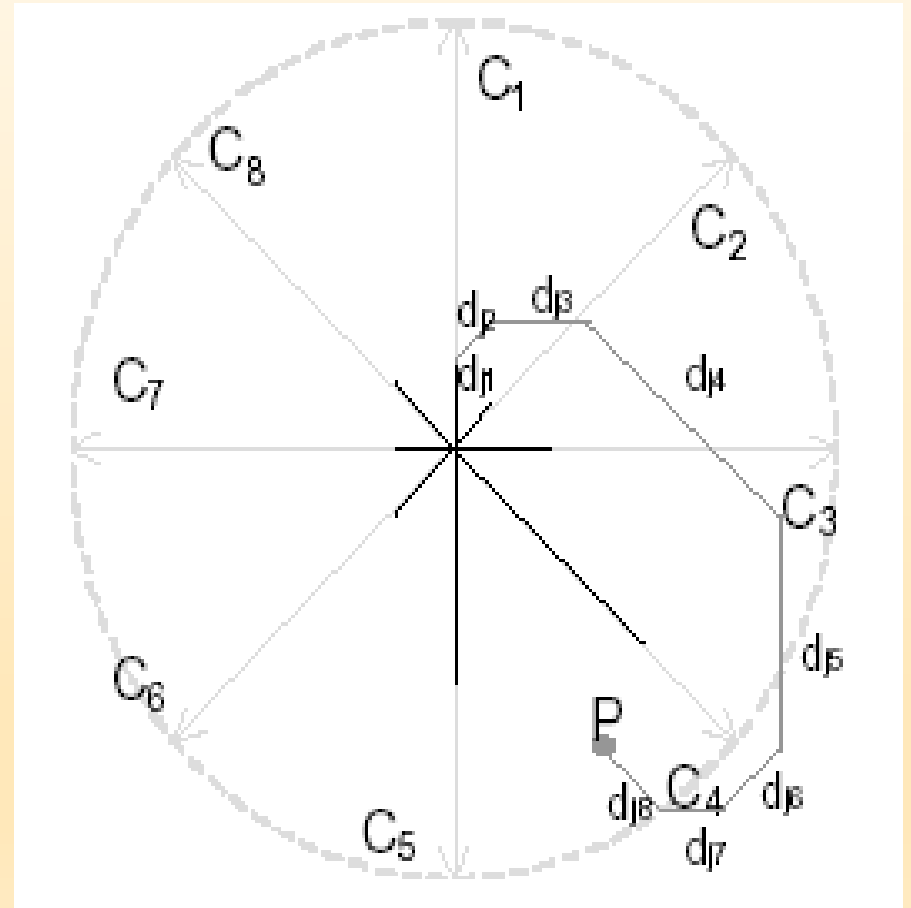
**2D-Radviz for breastw**

Class 1   Class 2
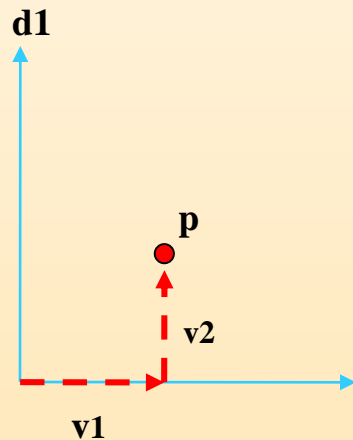
38
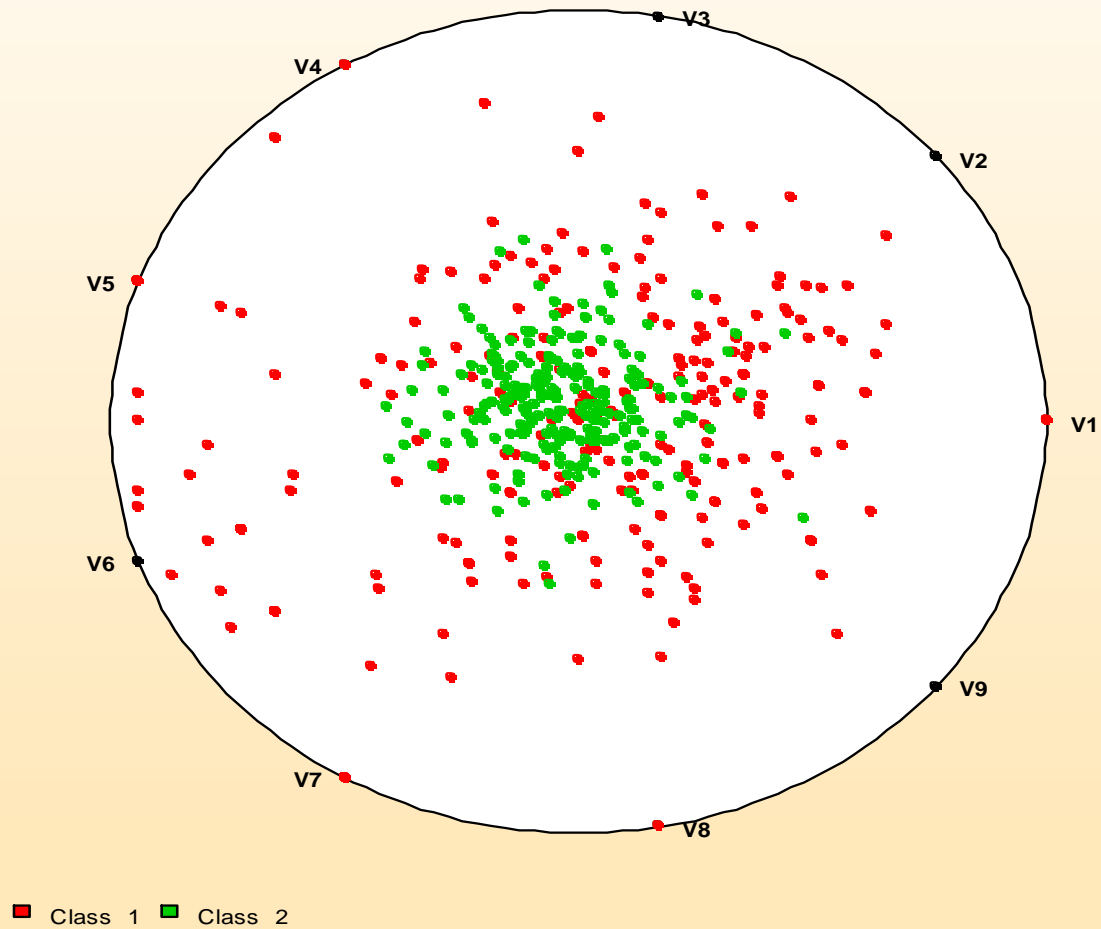
# Star Coordinates (Kandogan, 2001)

- Each dimension shown as an axis

- Data value in each dimension is represented as a vector.

- Data points are scaled to the length of the axis

  - min mapping to origin

  - max mapping to the end
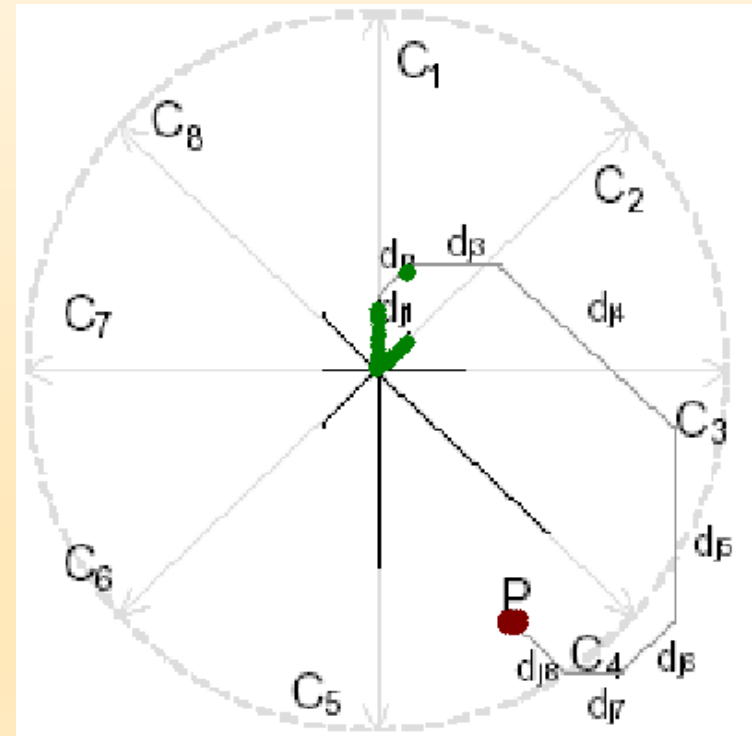
# Star Coordinates Contd

## Cartesian

$P=(v_1, v_2)$
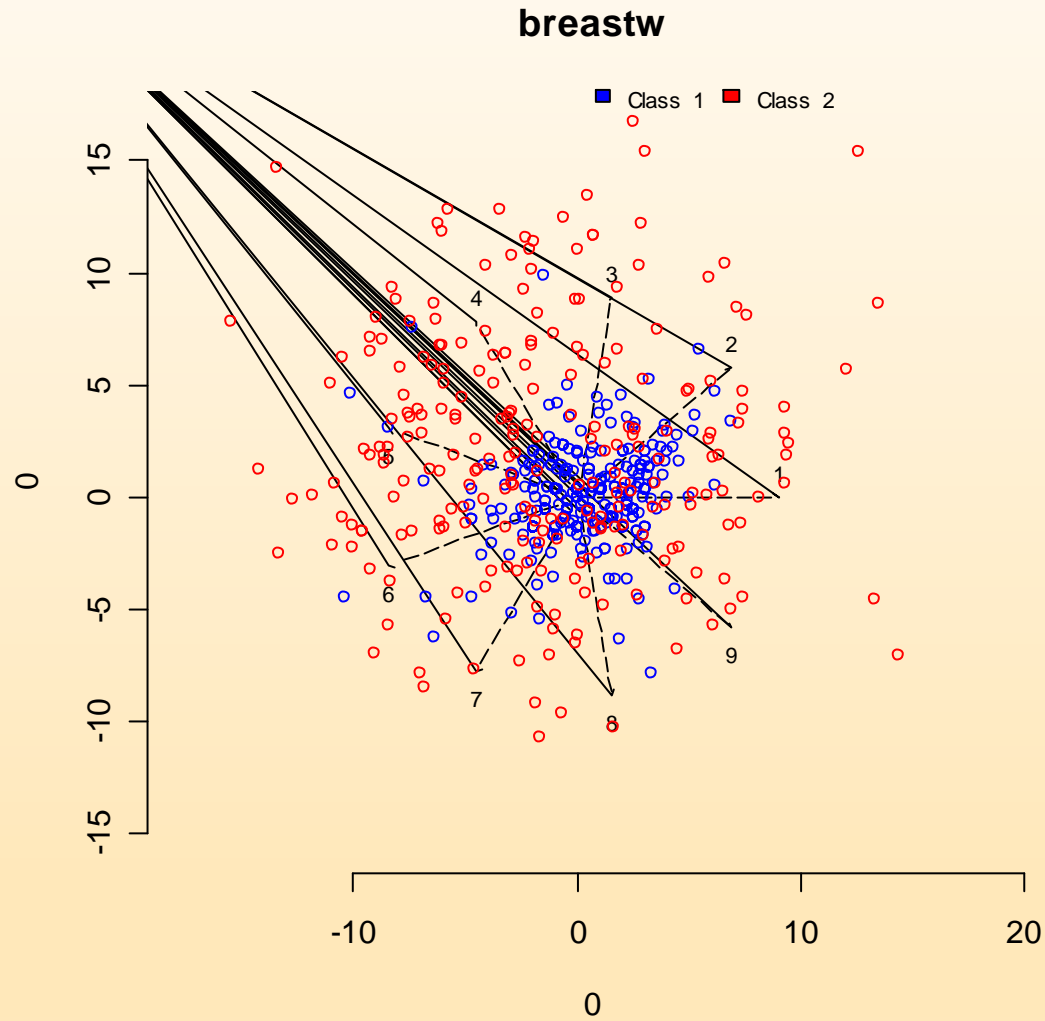


## Star Coordinates

$P=(v_1,v_2,v_3,v_4,v_5,v_6,v_7,v_8)$



Mapping:
- Items → dots
- Σ attribute vectors → position

# starcoord(breastw,main="breastw",class=T)



**breastw**

# Visualization software

Free and Open-source

- Ggobi (before was xgobi). Built using Gtk. Interface with databases systems. Runs on Windows and Linux. http://www.ggobi.org/

- Xmdv. The multivariate data visualization tool. Available for Linux and Windows. Built using OpenGL and Tcl/Tk. See http://davis.wpi.edu/~xmdv/

- Many more - see www.kdnuggets.com/software/visualization.html