ESMA 4016
Data Visualization

Dr. Edgar Acuna

Departamento de Ciencias Matematicas

Universidad de Puerto Rico Recinto de Mayaguez

Website: academic.uprm.edu/eacuna

Contenido

- Uso de Visualizacion
- Representando datos en 1,2, y 3-D
- Representando datos en mas de 4 dimensiones
  - Scatterplot Matrix
  - Survey plots
  - Parallel coordinates
  - Radviz, Starcoord

El uso de Visualizacion

- Visualizacion es el proceso de transformar la informacionen una forma visual de tal manera que el usuario pueda observar toda la informacion.
- El uso de una buena tecnica de visualizacion en data mining puede reducir el tiempo que toma entender los datos, encontra relaciones entre las variables y descubir informacion.
- Uno de los objetivos de la visualizacion es hacer analisis exploratorio de los datos.

El uso de visualization (cont)

- En analisis exporatorio, se usan tecnicas de visualizacion antes de aplicar un algoritmo de data mining para obtener informacion de las caracteristicas del dataset. El resultado de la exploracion piuede conducer a formular hipotesis acerca de los datos.
- El uso de visualizacion permite al usuario mejorar el entendimiento de sus datos y evitar que pueda cometer errores en sus conclusions.
- Ayuda en la presentacion de los resultados
- Desventaja: Requiere de los ojos del humano y puede ser mal interpretada

Los principios de Tufte de una Buena grafica

- Dar al observador
  - El mayor numero de ideas
  - En el tiempo mas corto
  - Con el minimo de Tinta en el espacio mas pequeno.

- Decir la verdad acerca de los datos!

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

Metodos de visualizacion

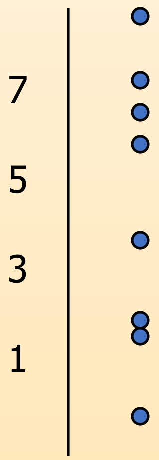
- Visualizando en 1-D, 2-D y 3-D
  - Hay bastantes metodos conocidos
- Visualizando en mas dimensiones
  - Scatterplot matriicial
  - Survey plots
  - Parallel Coordinates
  - Radviz
  - Star Coordinates

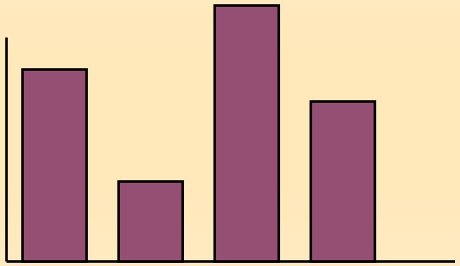
1-D (Univariate) data (R)

- stripchart(x,vertical=T,col=2) #Dotplot
- hist(x,col=3) #Histogram
- boxplot(x,horizontal=T,col="blue") #Boxplot

1-D (Univariate) Data

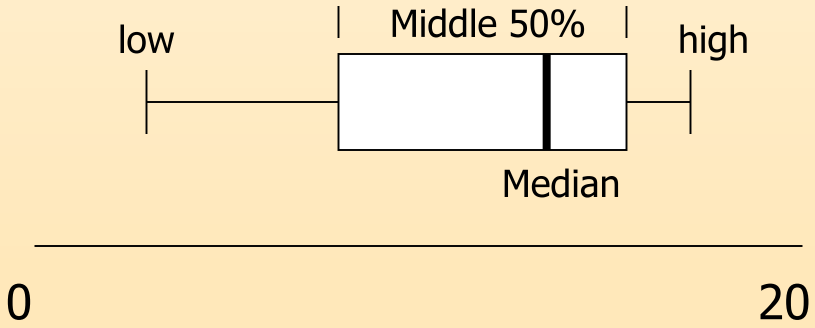
- Representations





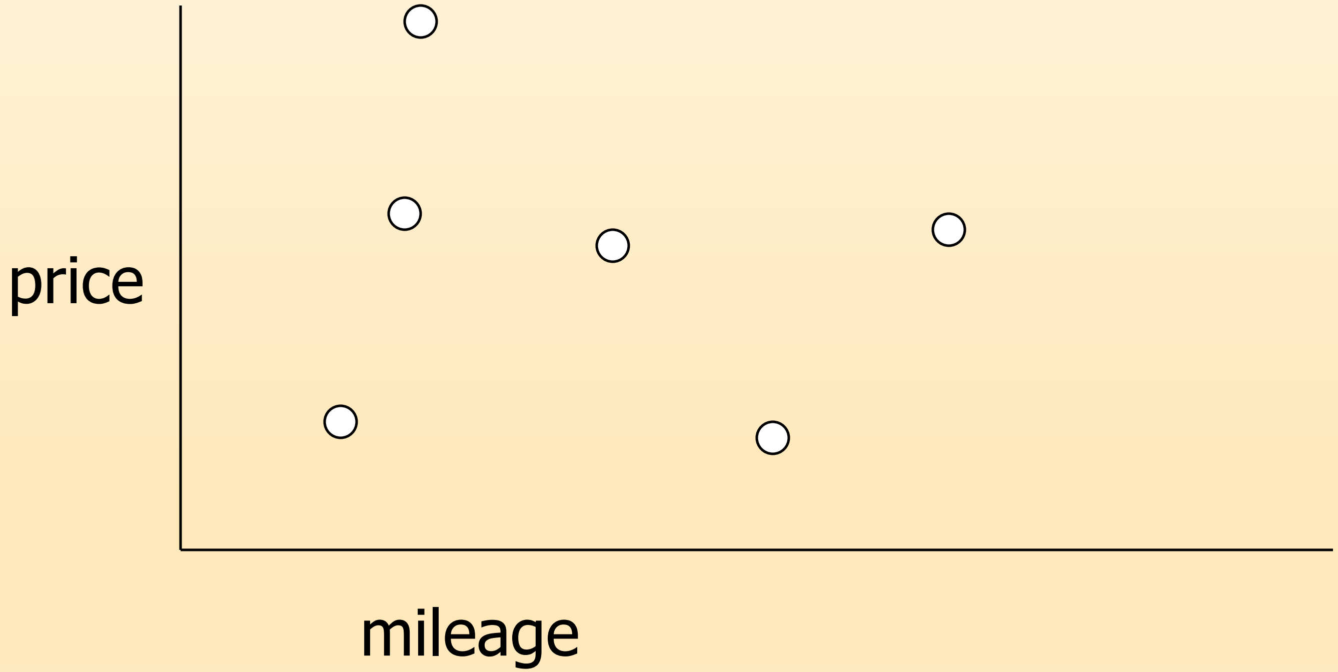
Histogram

Tukey box plot



2-D (Bivariate) Data

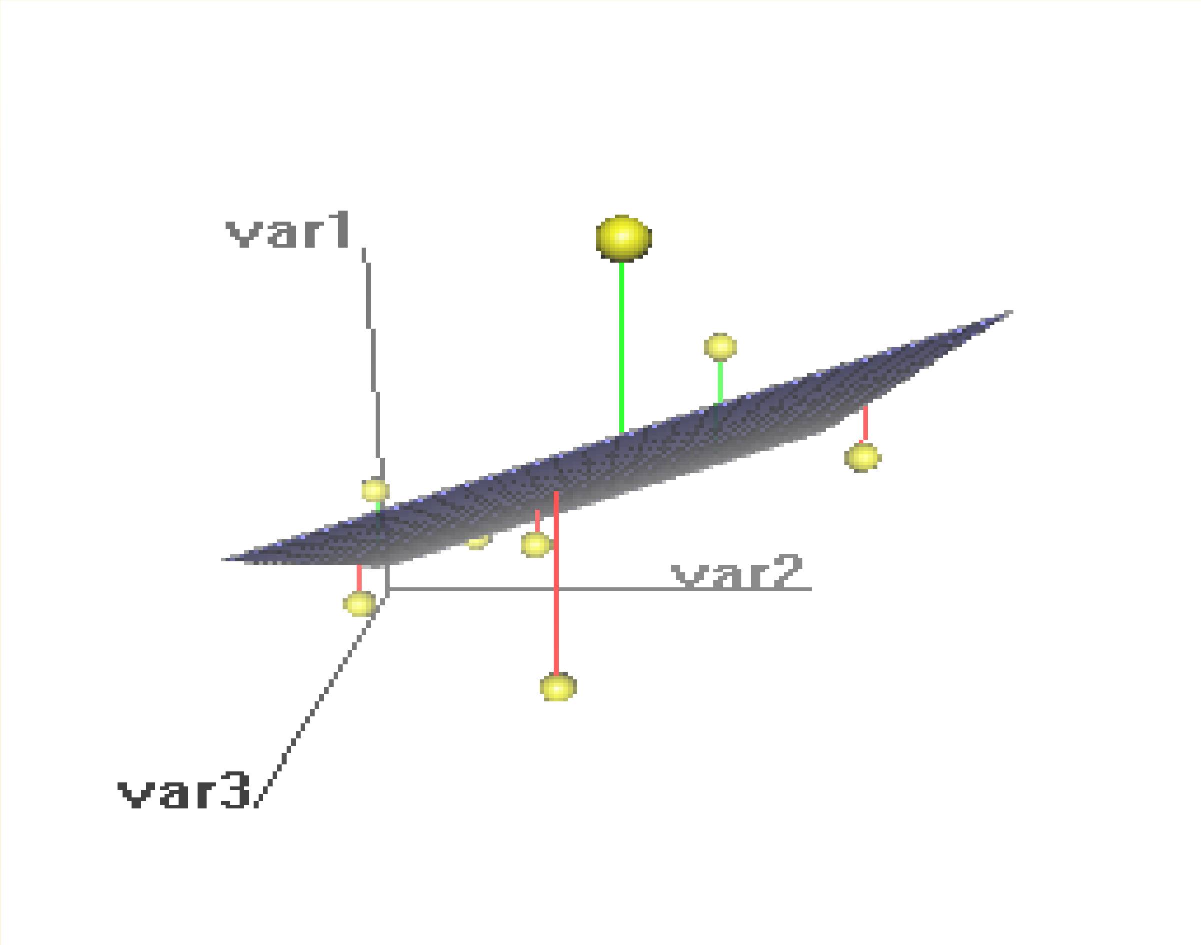
- Scatter plot, …



3-D Data

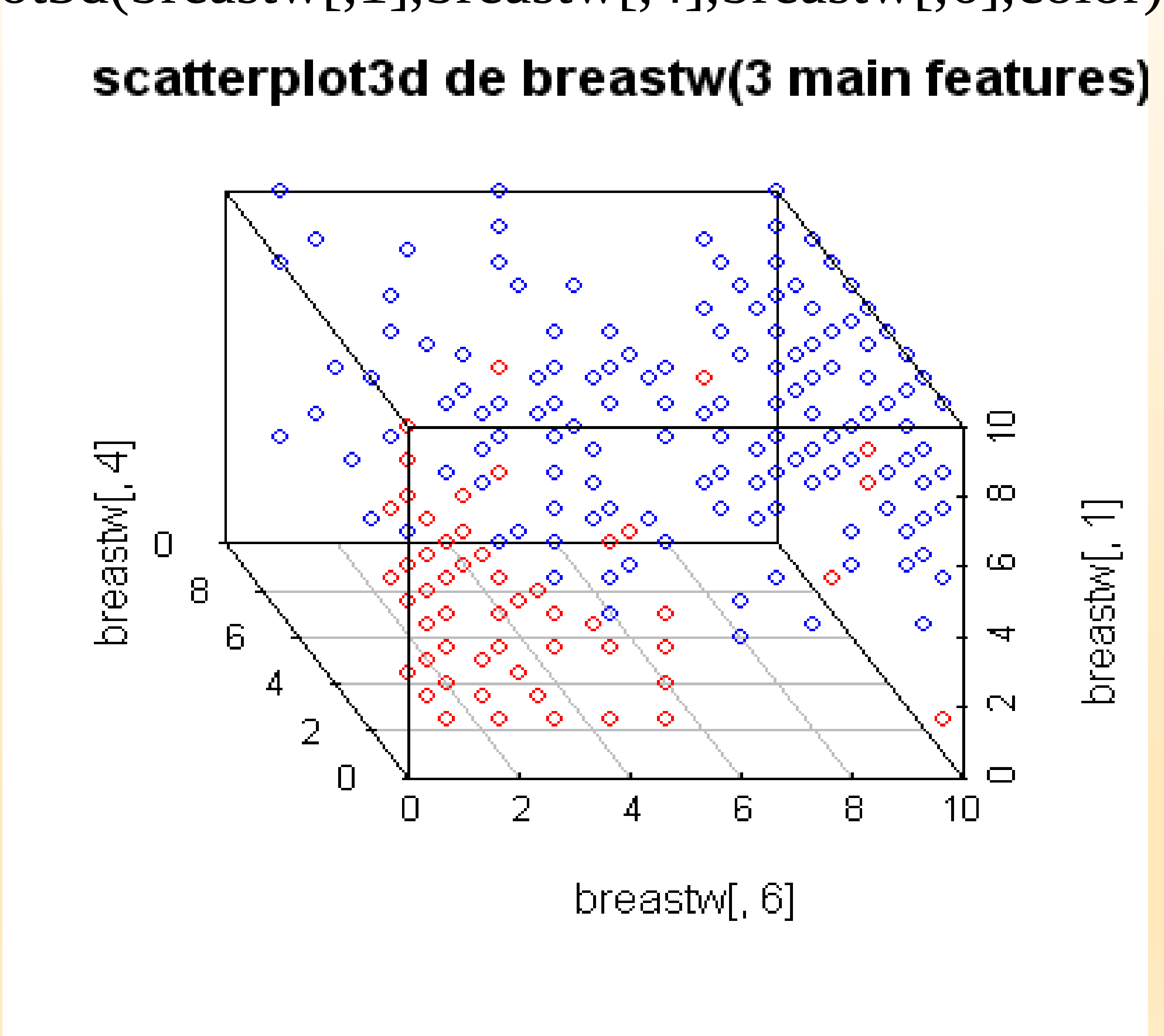
- Scatter3d in Rcmdr library (R)
- Scatterplot3 in the scatterplot3d library (R)
- Cloud() in lattice library. Lattice is a free version of trellis. (R)

Scatter3d from Rcmdr library

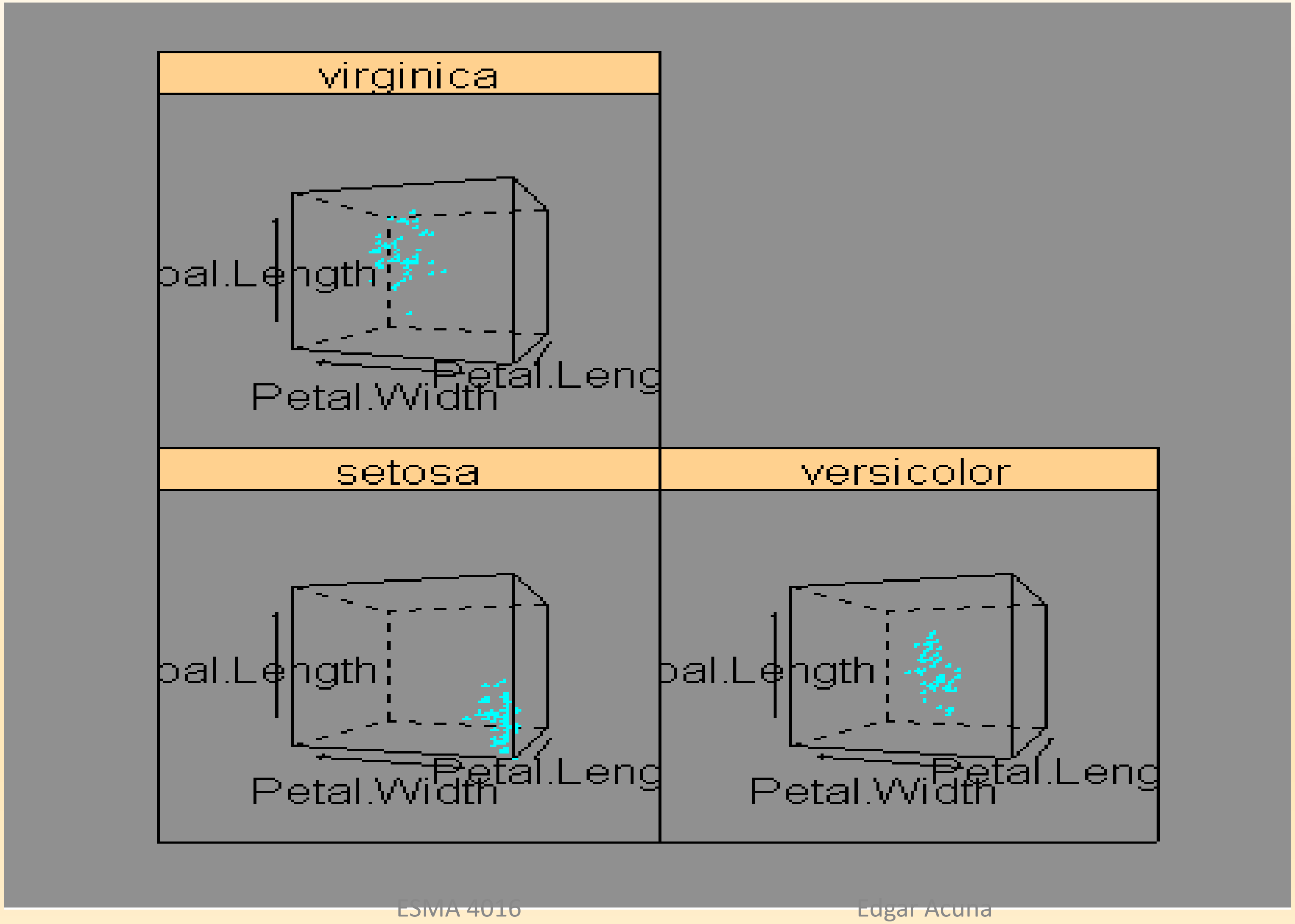


Scatterplot3d from scatterplot3d library

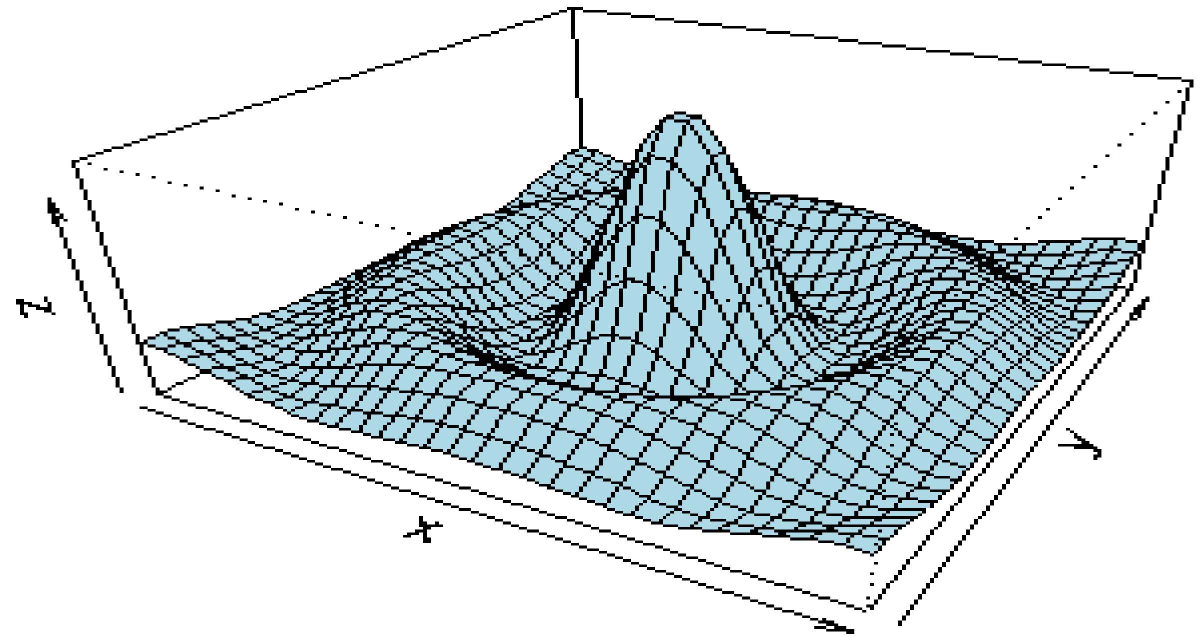
```
scatterplot3d(breastw[,1],breastw[,4],breastw[,6],color)
```



Cloud() from the lattice library

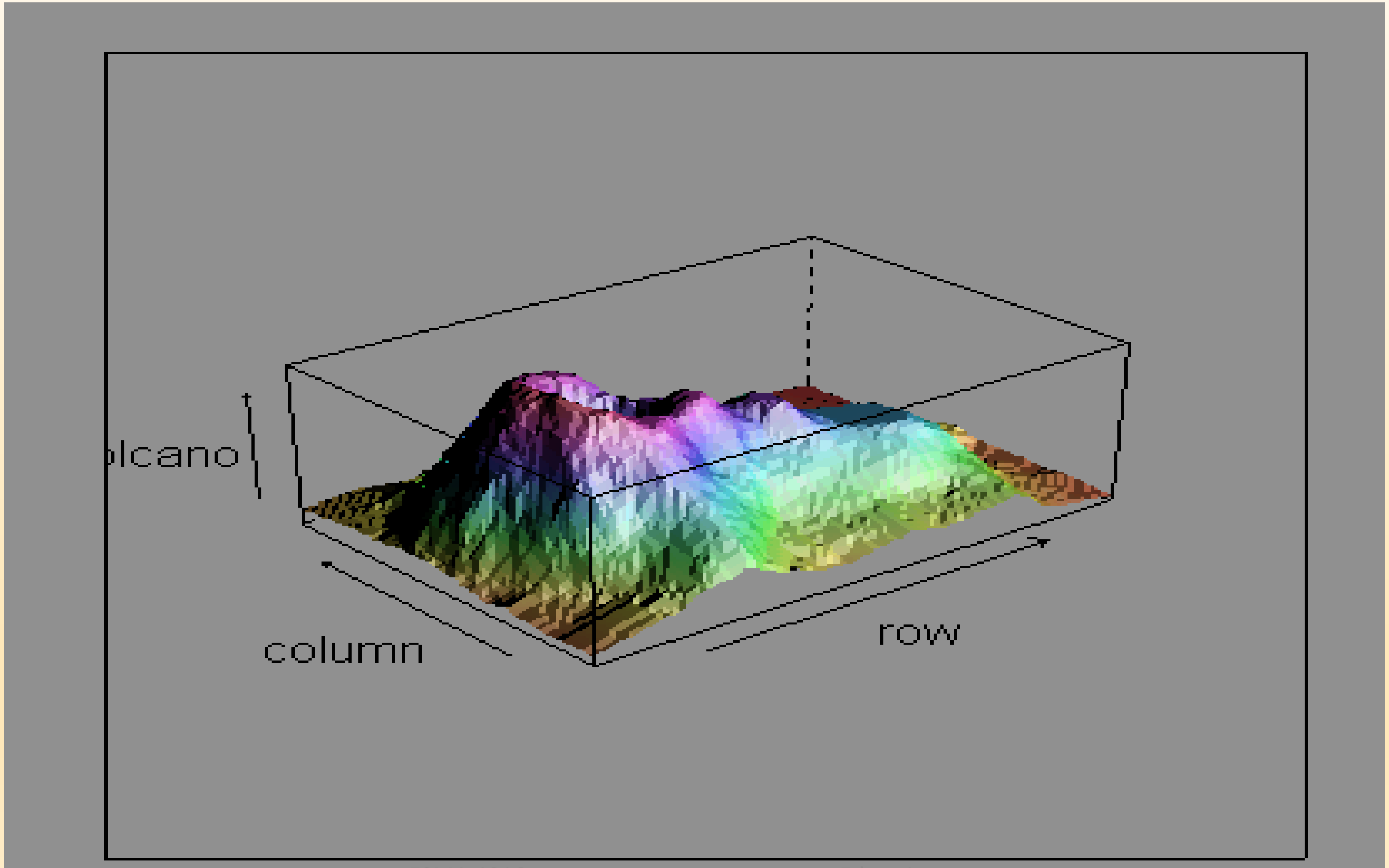


3-D Data (persp)



```
> 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)



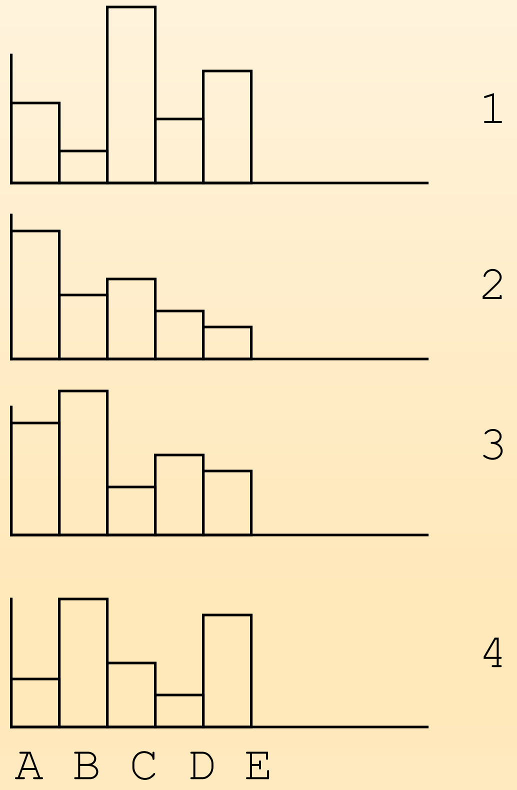
Visualizing in 4+ Dimensions

- 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

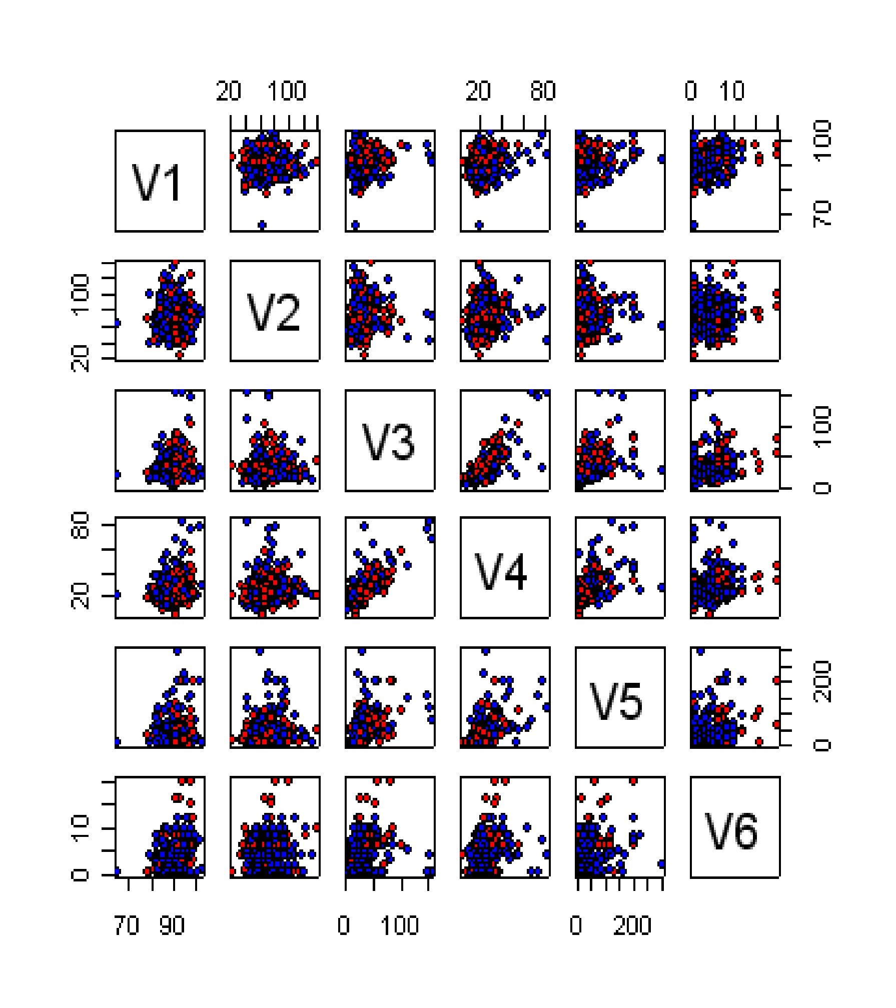
Represent each possible pair of variables in their own 2-D scatterplot (car data)

Useful for detecting

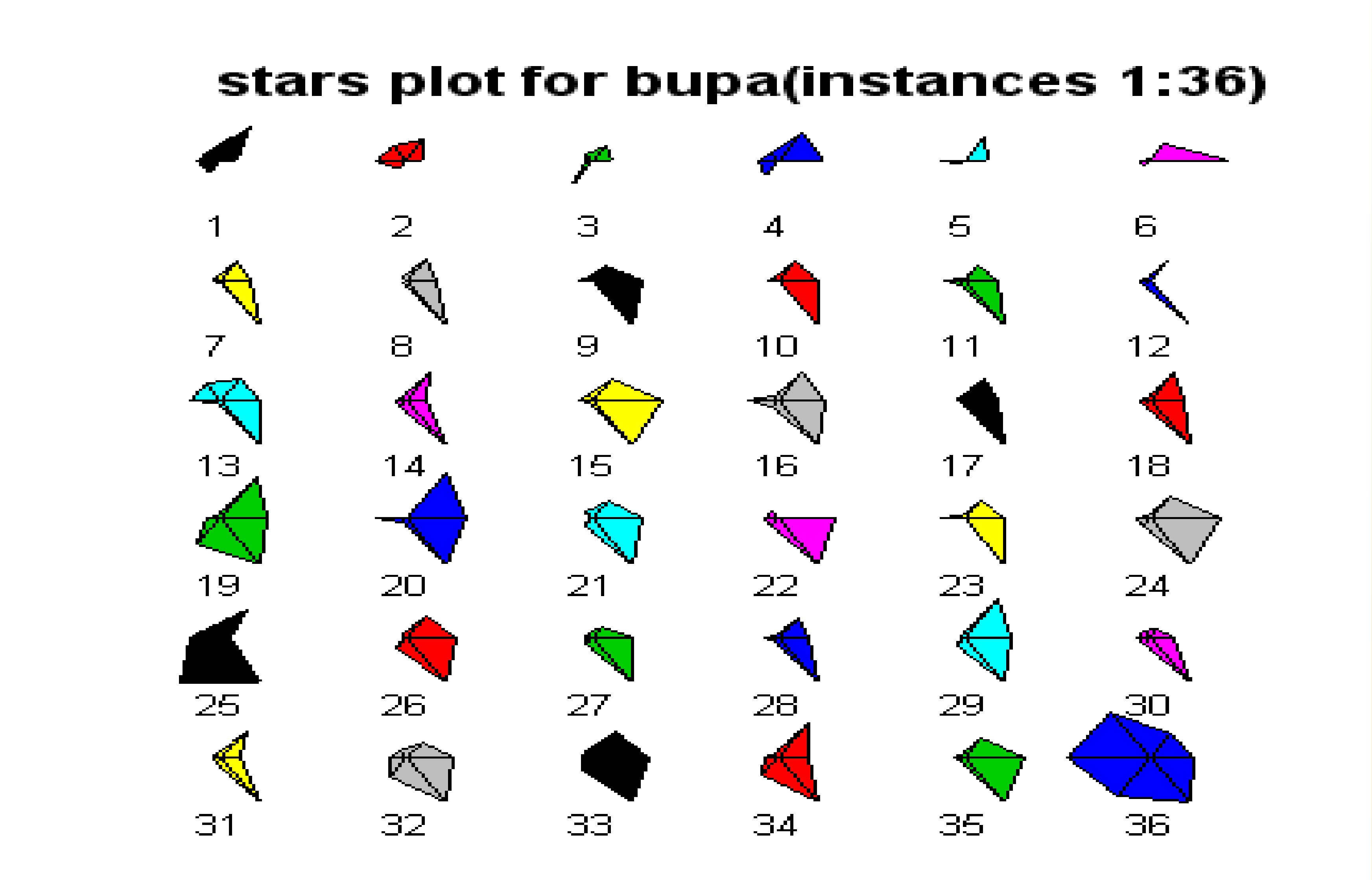
linear correlations (e.g. V3 & V4)

But misses

multivariate effects



Star Plots (Chambers et al., 1983)



Visualization function in *dprep*:

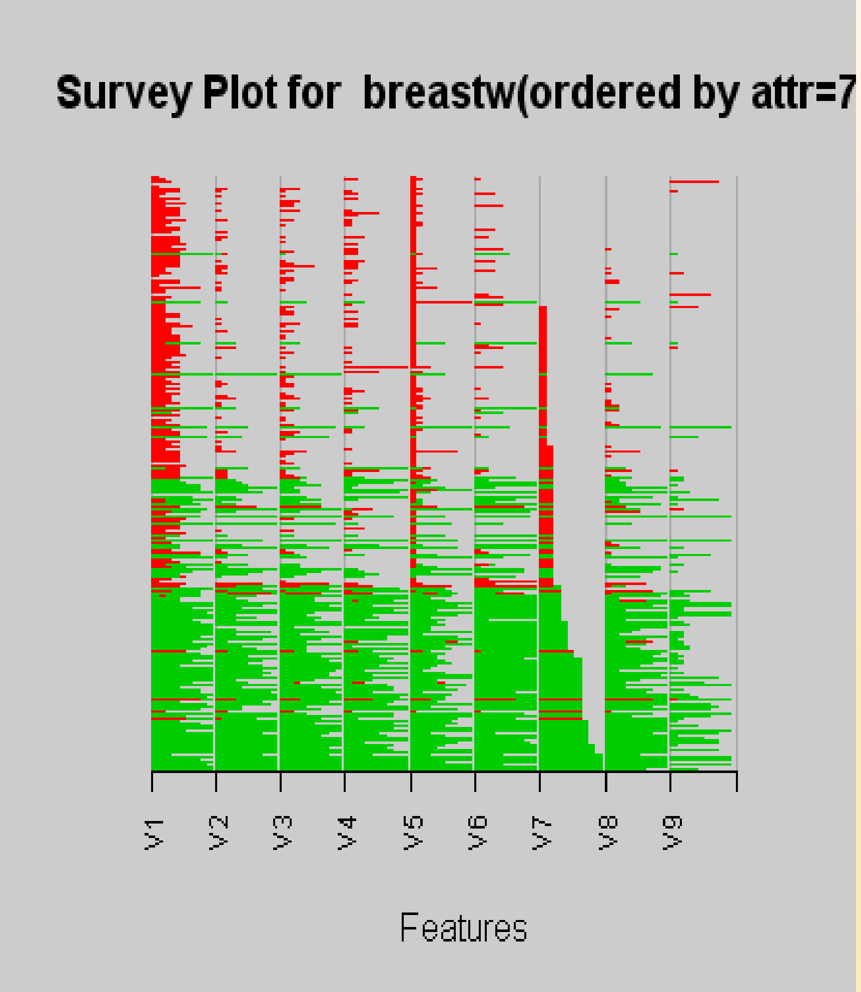
- *imagmiss( )* determine the existence of missing values in the dataset, identify their location and quantity.
- *surveyplot( )* constructs a survey plot of the dataset
- *parallelplot( )* constructs a parallel coordinate plot of the data
- Starcoord(), Starcoor3d()
- Radviz()

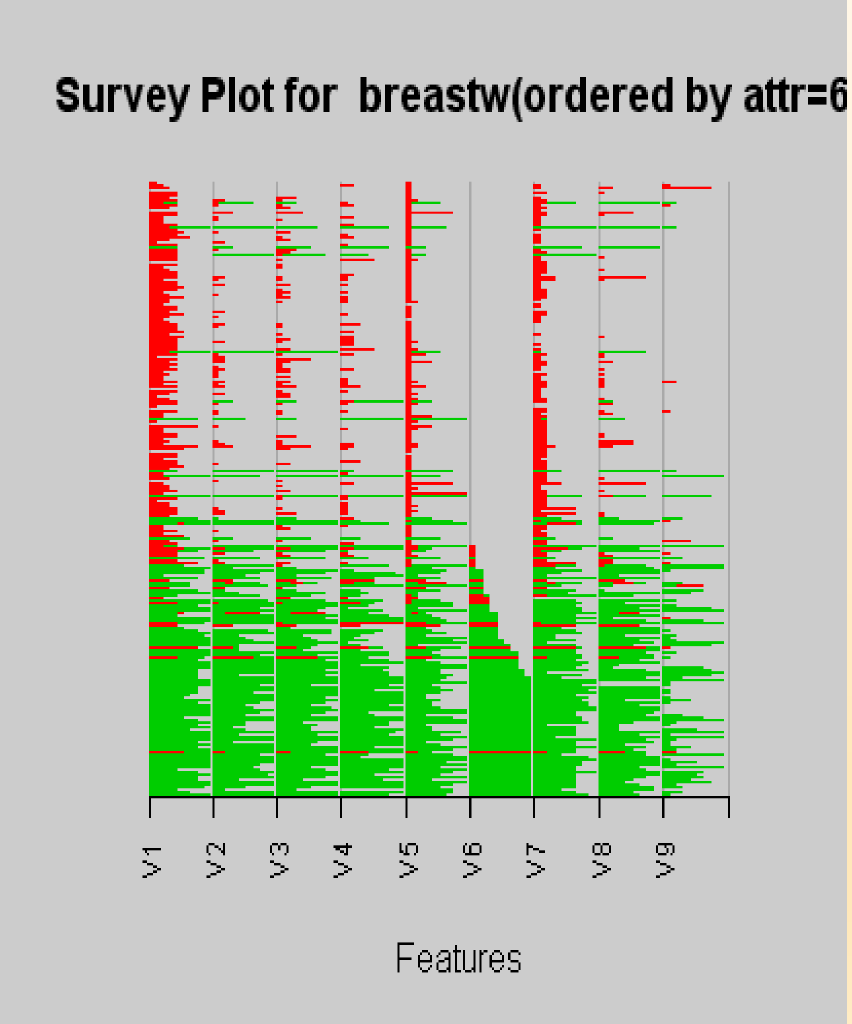
The survey plot (Lohninger, 1994)

- 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.

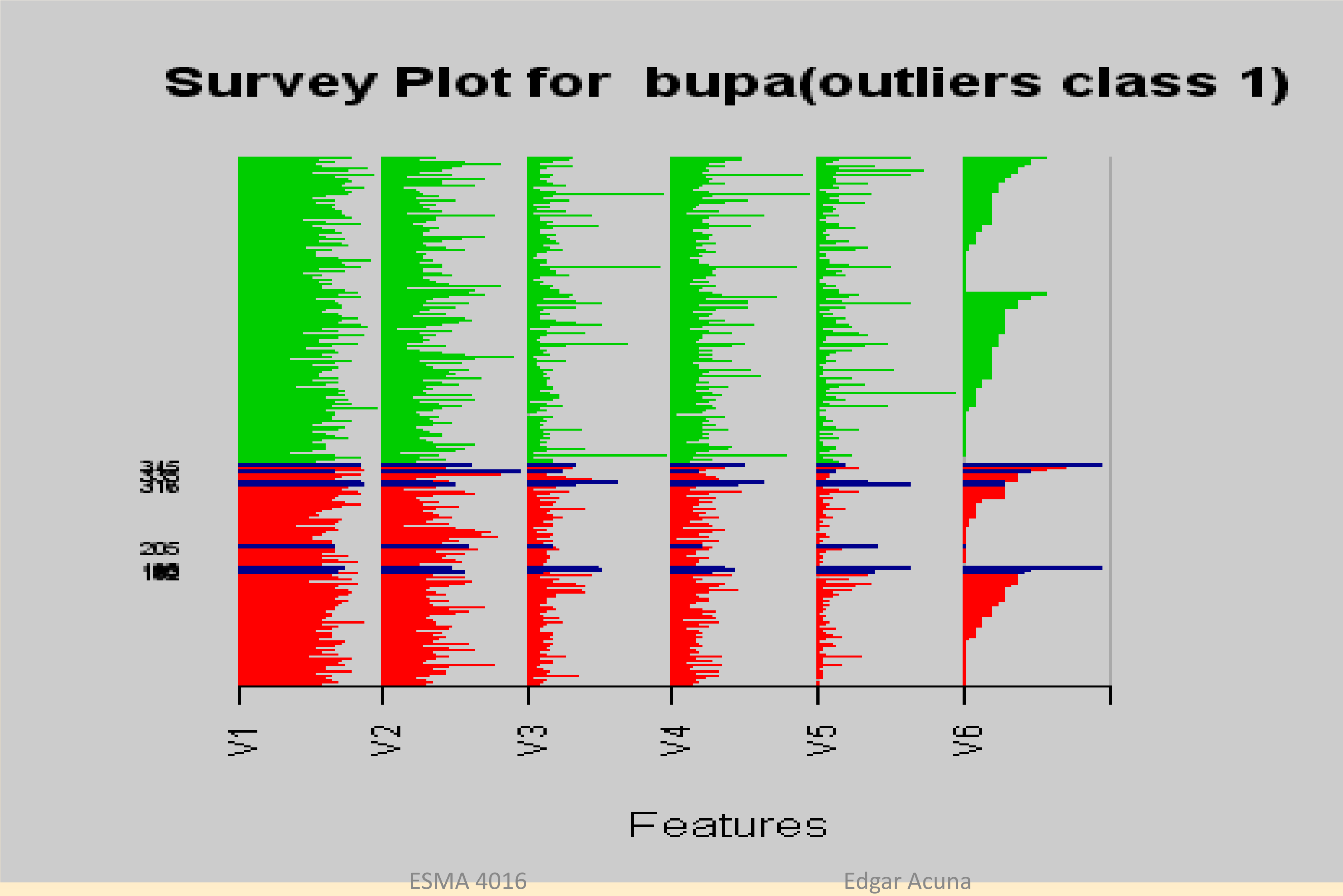
The survey plot:

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



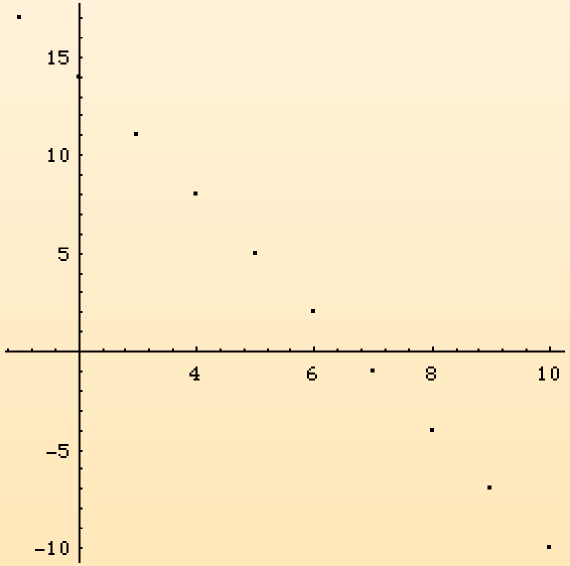


Surveyplot as a tool to detect outliers

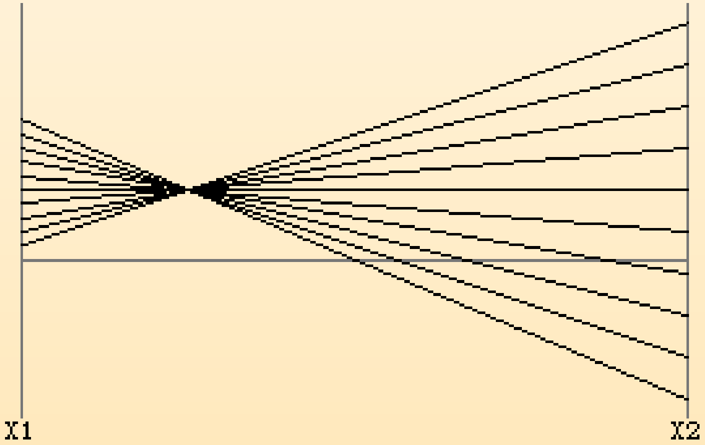


Parallel Coordinates

- 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

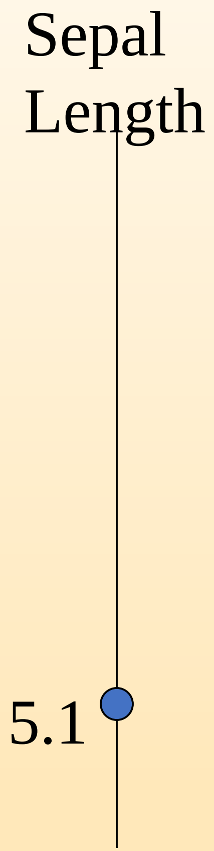
Iris setosa

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

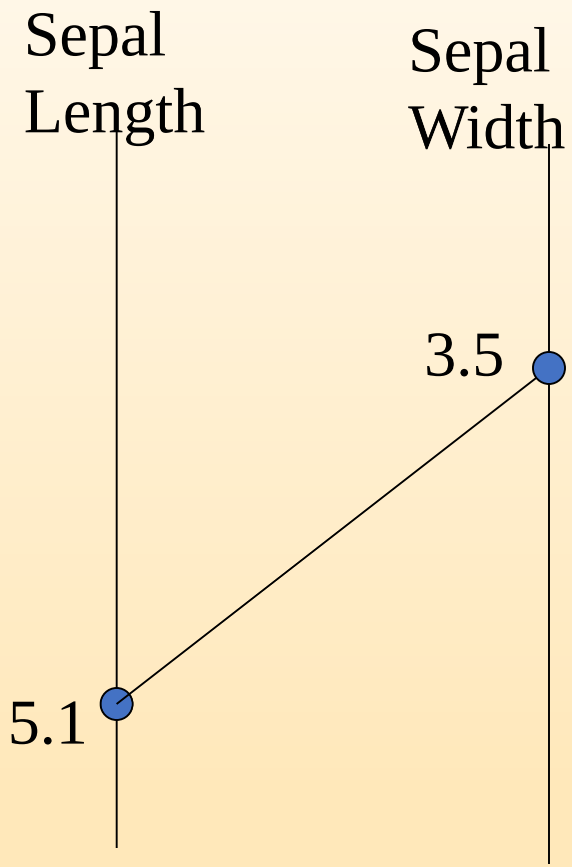
Iris virginica

Parallel Coordinates



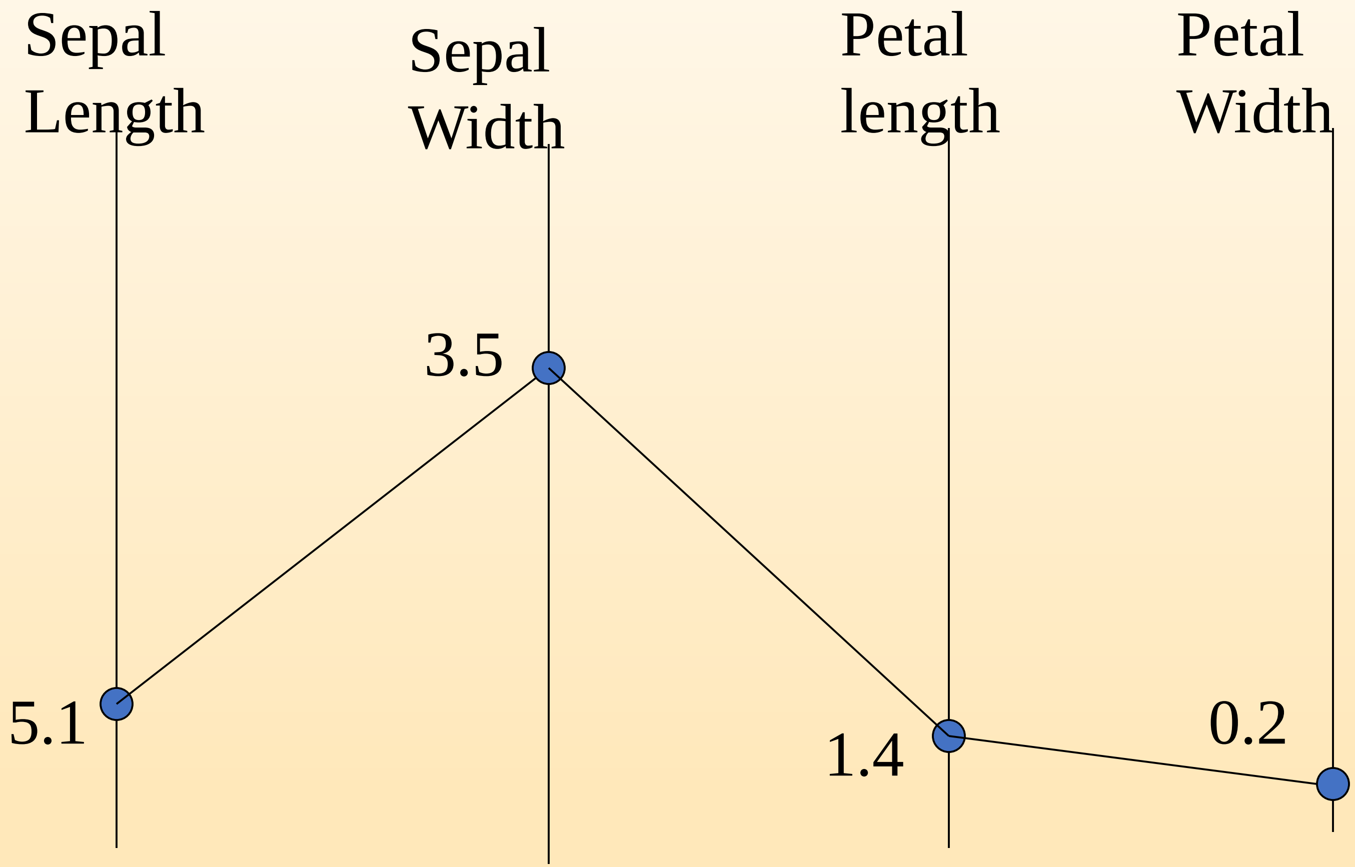
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Parallel Coordinates: 2 D



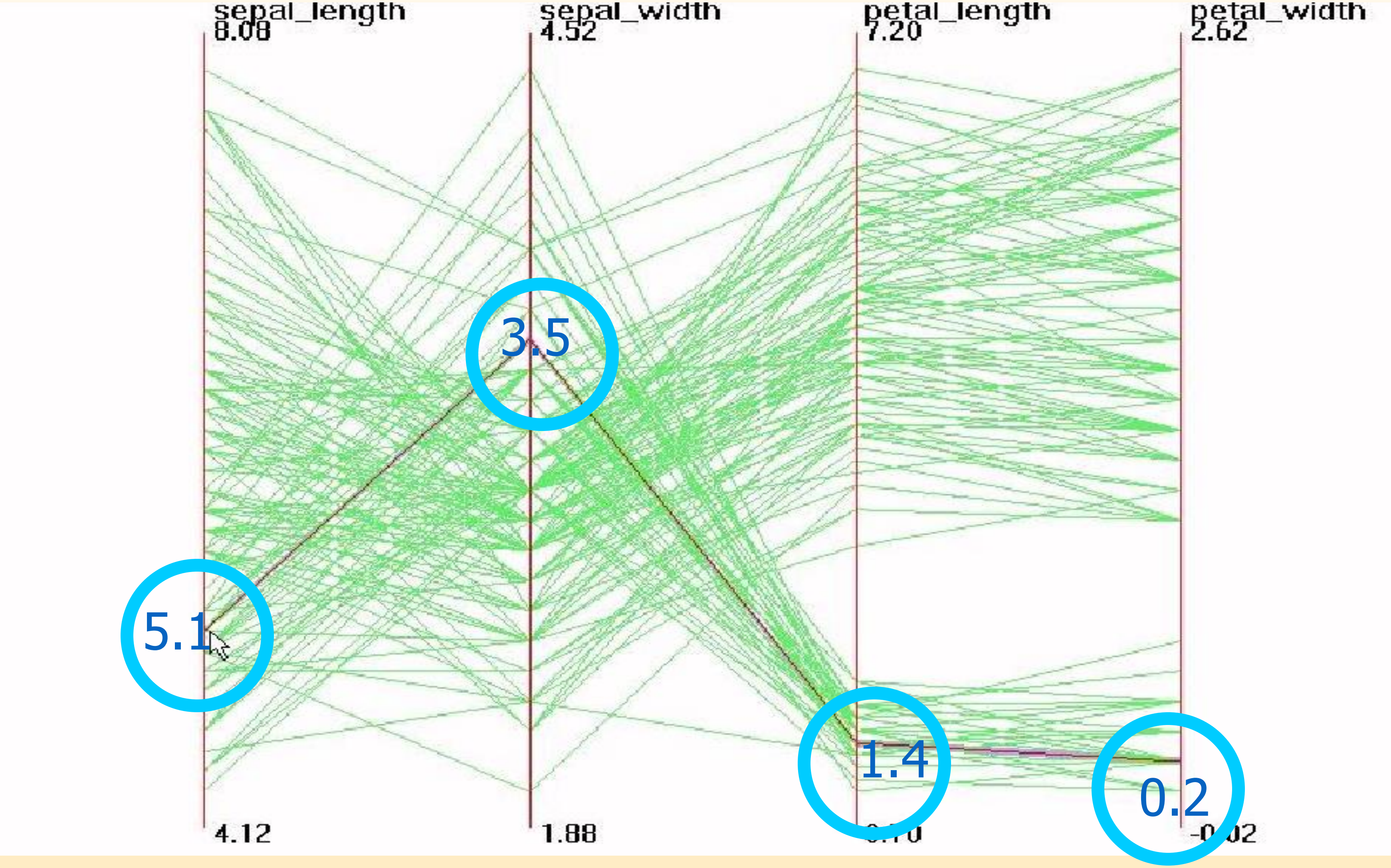
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

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)

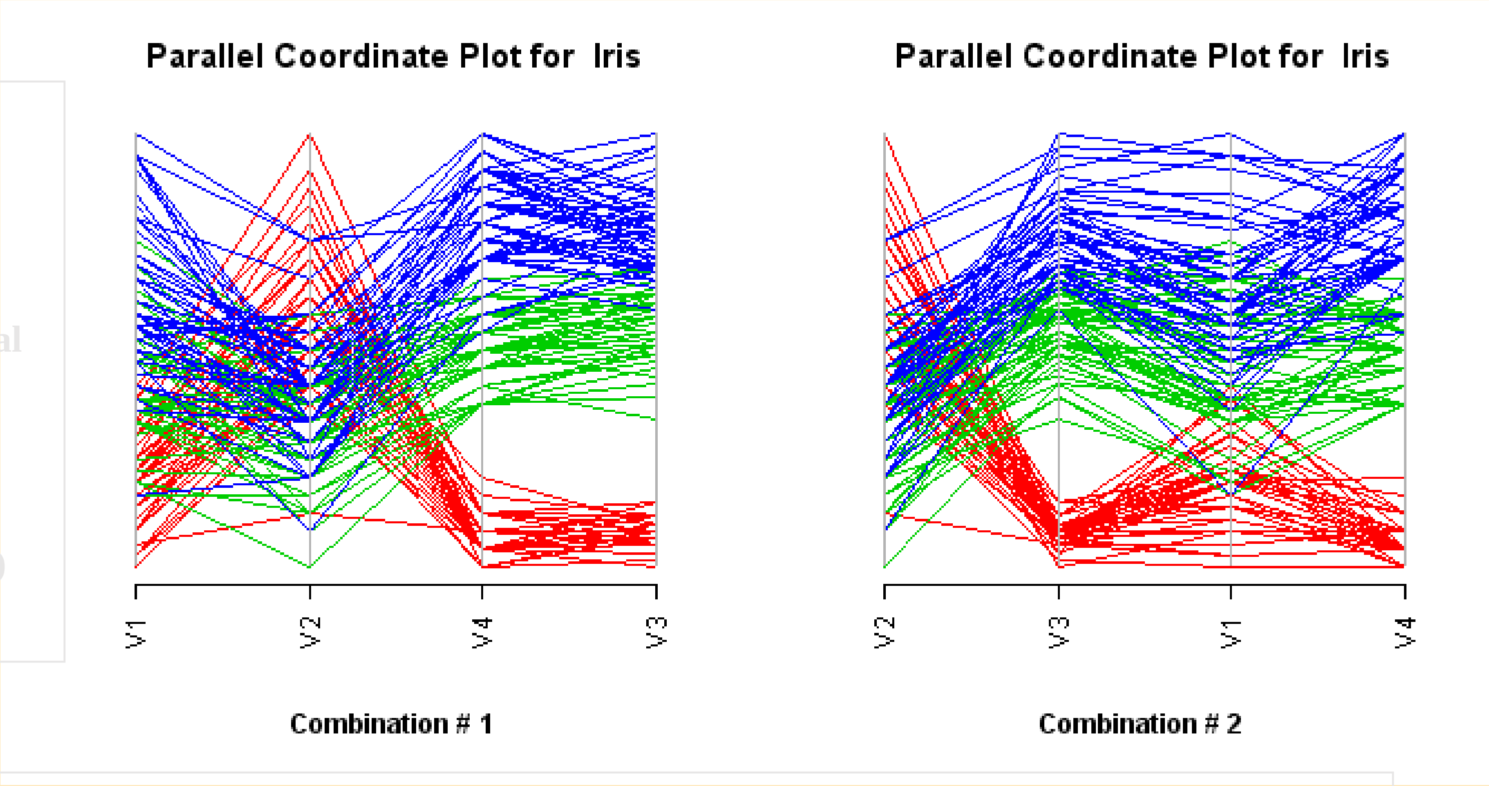
- Pairwise comparison is limited to those axis that are adjacent.
- For a dataset with p attributes there are p! permutations of the attributes so each of them is adjacent to every attribute in some permutation.
- Wegman (1990) determined that only [(p+1)/2] permutations are needed.([.] is the greatest 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.



Interpretation:

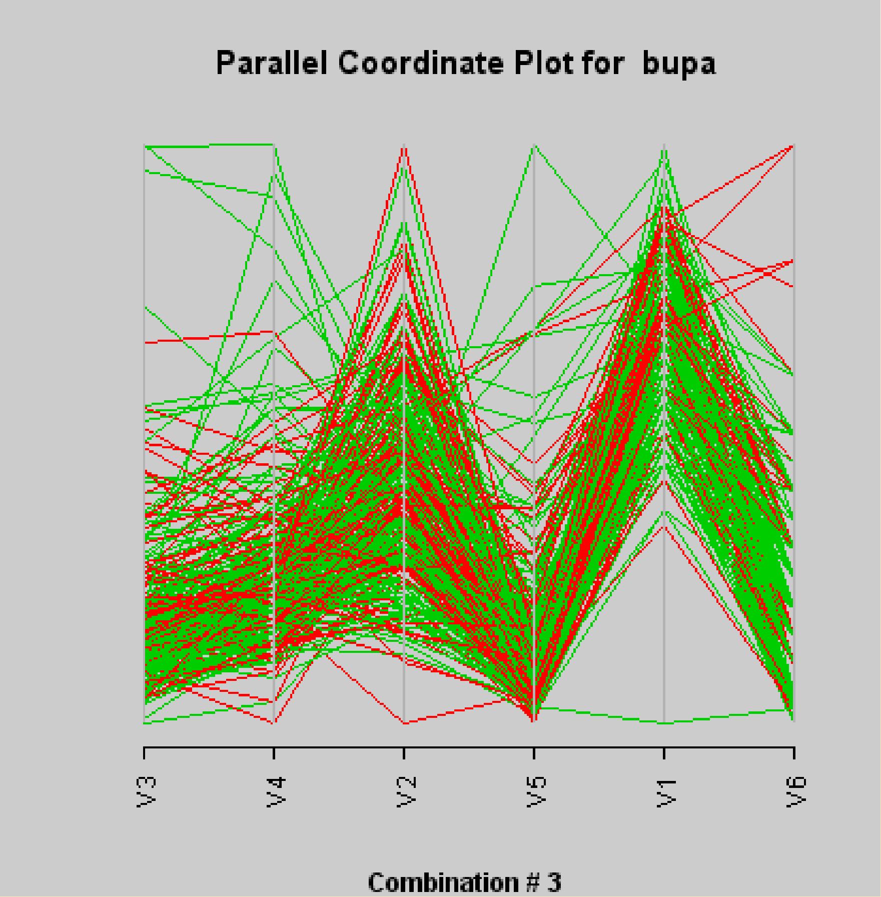
- Each different color represents a different class.
- If two attributes are highly positively correlated, lines passing from one feature to another tend not to intersect between the parallel coordinate axes.

The parallel coordinate plot

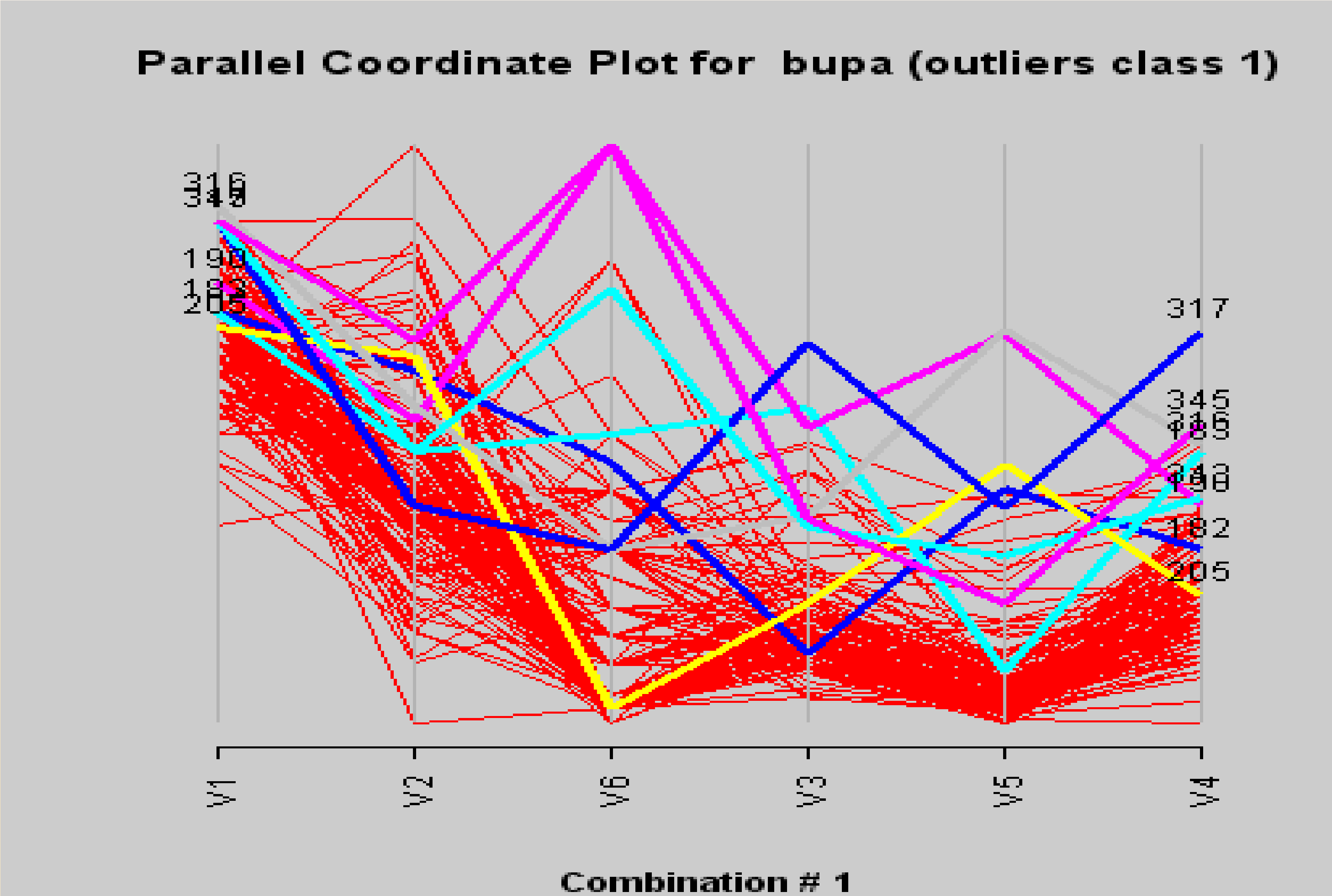
- For highly negatively correlated attributes, the line segments tend to cross near a single point between the two parallel coordinate axes.
- The presences of outliers is suggested by poly-lines that do not follow the pattern for their class.

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.



Parallelplot as a tool to detect outliers



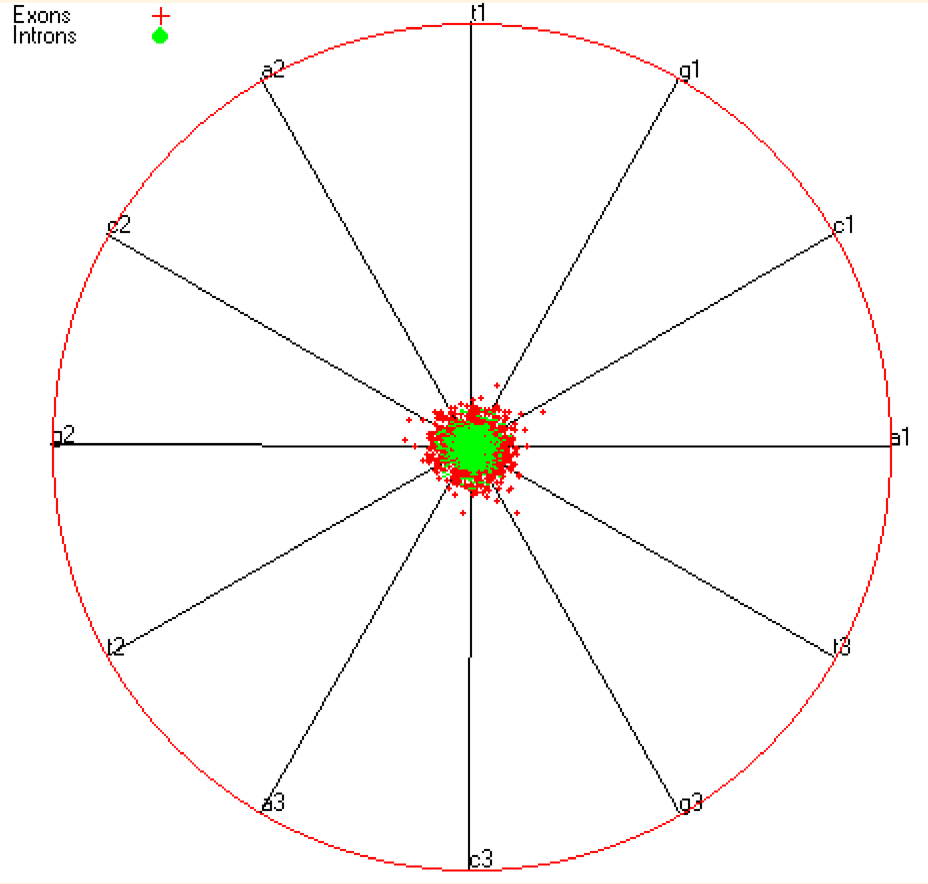
Parallel Visualization Summary

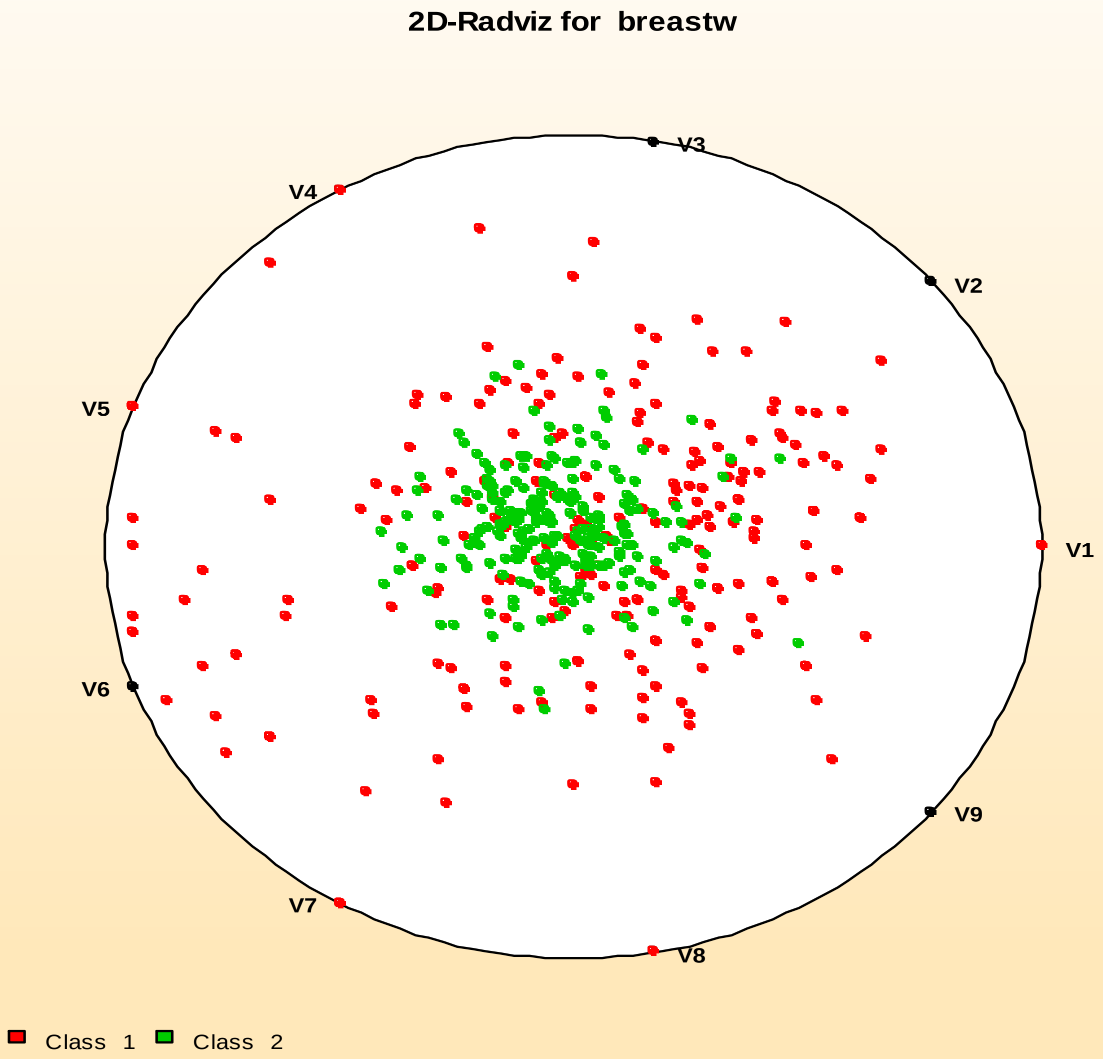
- Each data point is a line
- Similar points correspond to similar lines
- Lines crossing over correspond to negatively correlated attributes
- Interactive exploration and clustering

- 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.



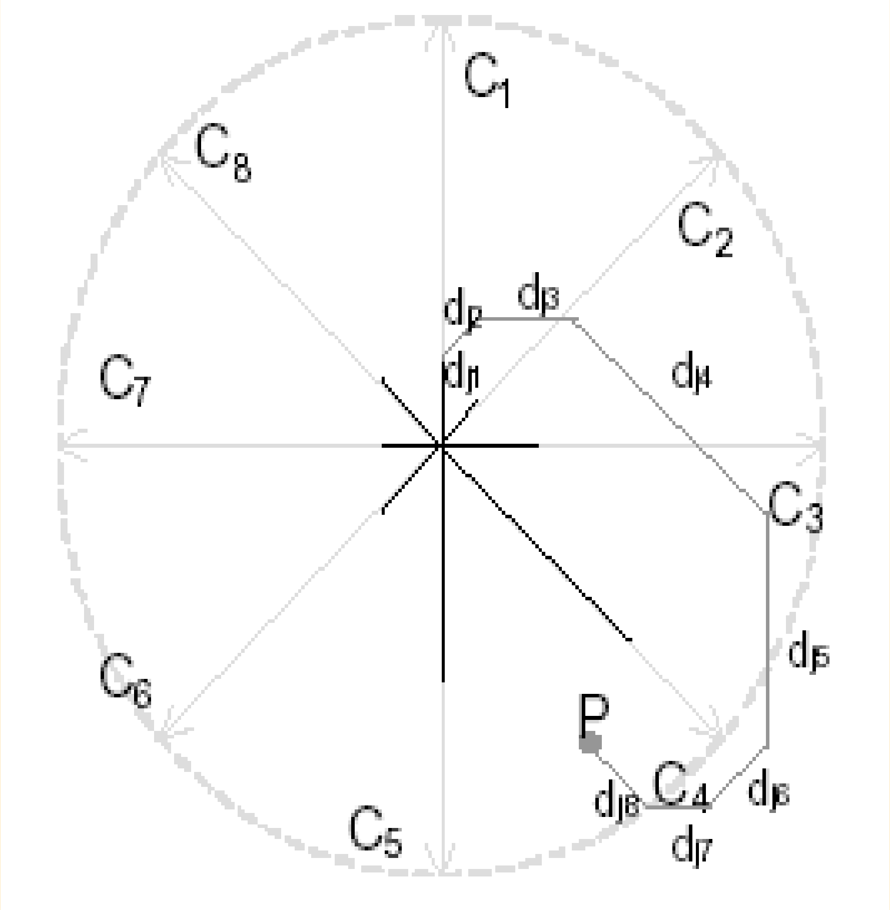

2D-Radviz for breastw
V1
V2
V3
V4
V5
V6
V7
V8
V9
🟥 Class 1
🟩 Class 2

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

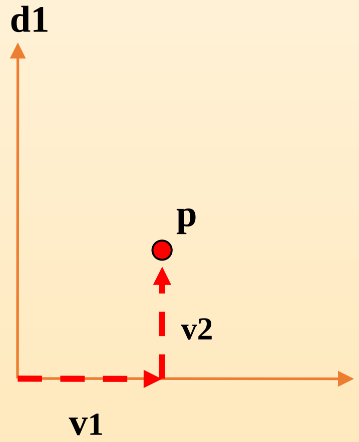
- max mapping to the end



Star Coordinates Contd

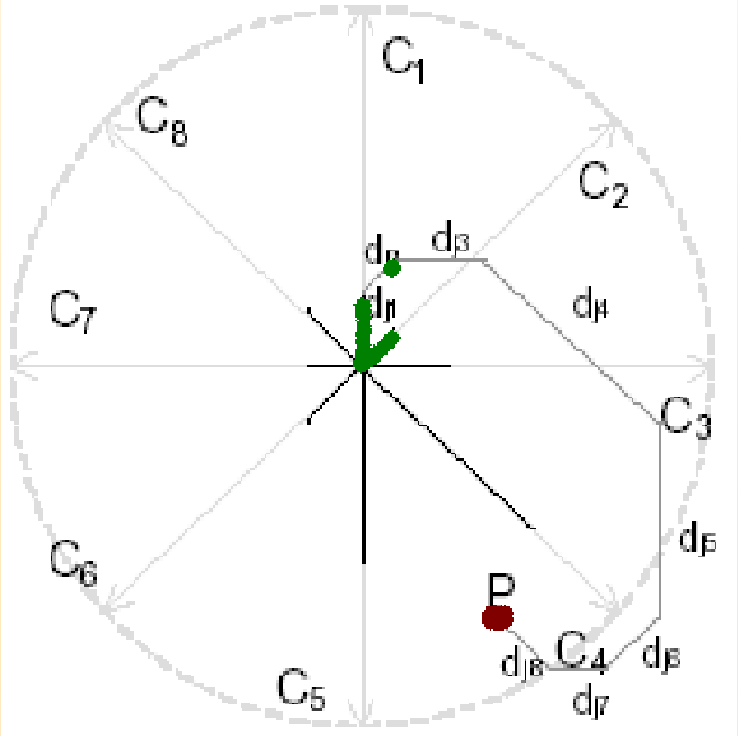
Cartesian

P=(v1, v2)



Star Coordinates

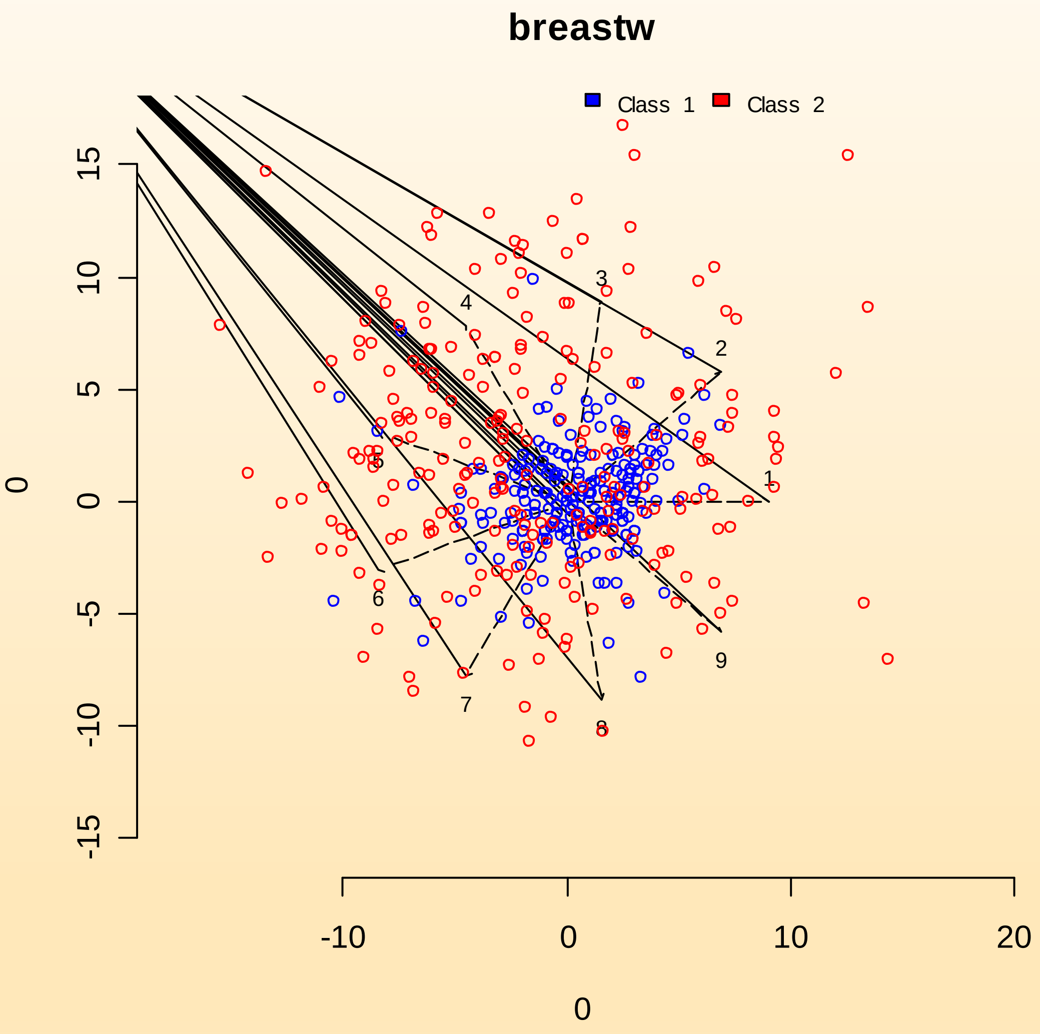
P=(v1,v2,v3,v4,v5,v6,v7,v8)



Mapping:

- Items → dots
- Σ attribute vectors → position

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



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/
- XmdvTool. 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