COMP 6315
Visualizacion de datos

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 informacion en una forma visual de tal manera que el usuario pueda observar graficamente 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 matricial
  - 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

Python:

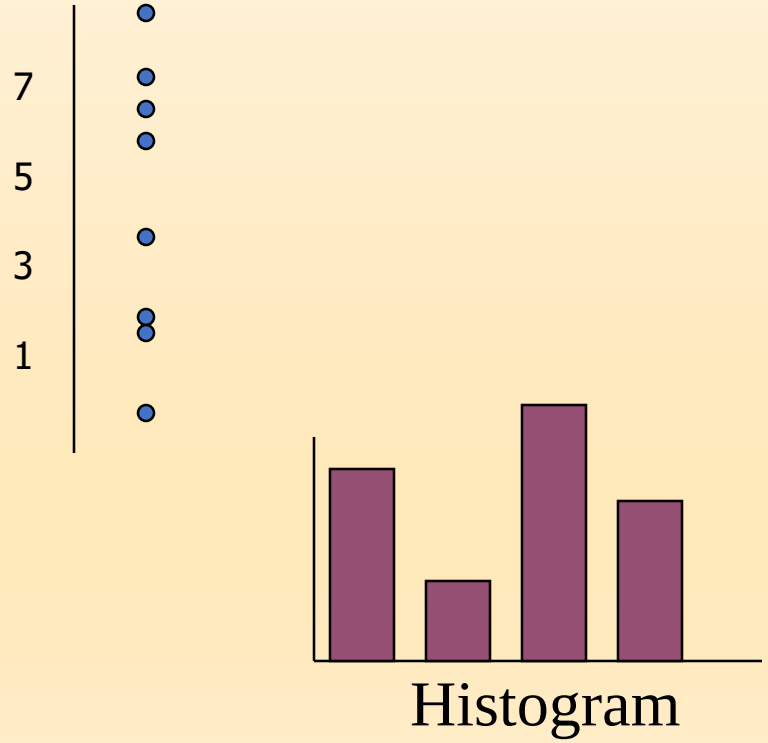
Dotplot esta en la libreria plotnine

Histograma en varias librerias: plotnine, matplotlib, seaborn, plotly, bokeh

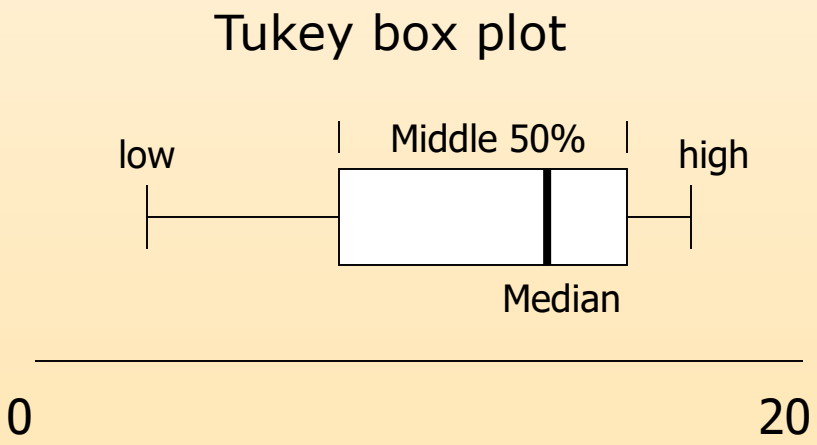
Boxplot en plotnine, matplotlib, seaborn, plotly y bokeh

1-D (Univariate) Data

- Representations

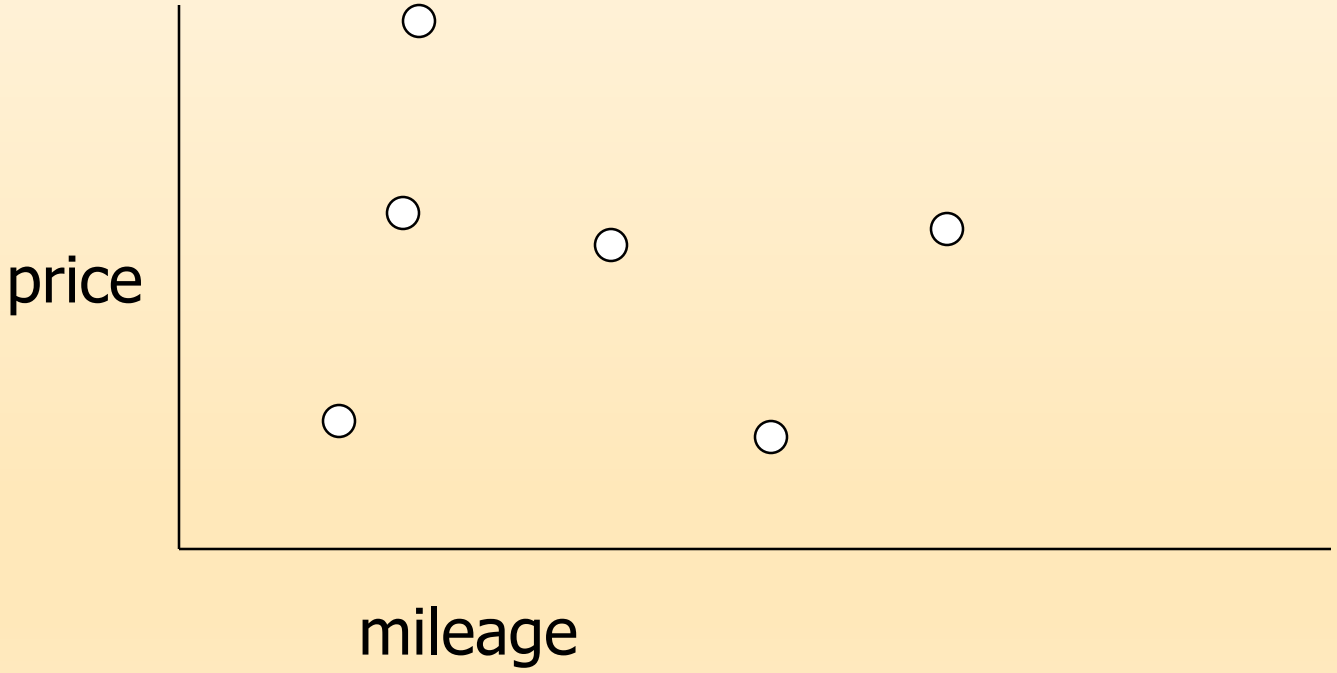


Histogram



2-D (Bivariate) Data

- Scatter plot, …

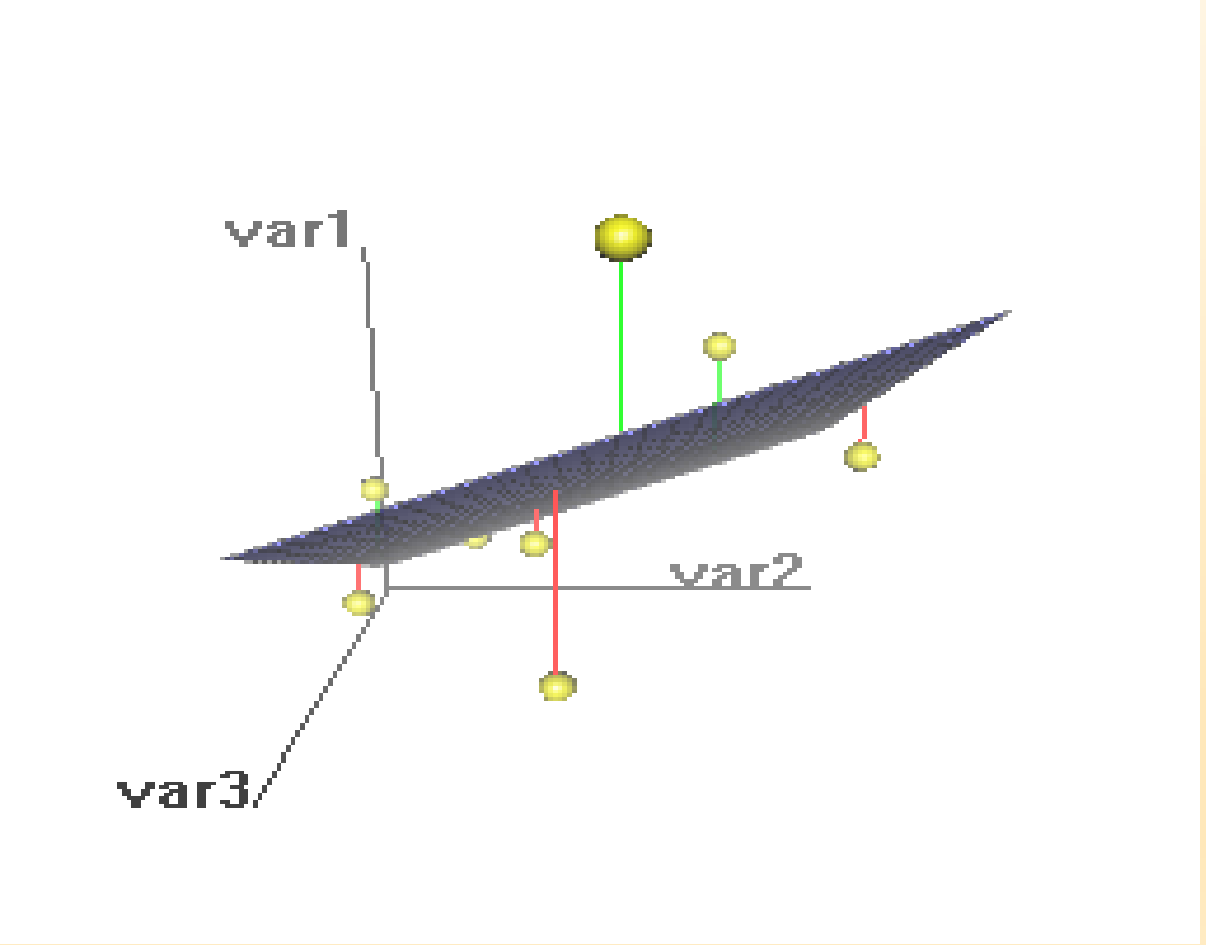


3-D Data

- Scatter3d en la libreria Rcmdr (R)
- Scatterplot3 en la libreria scatterplot3d (R)
- Cloud() en la libreria lattice que es la version free de trellis. (R)

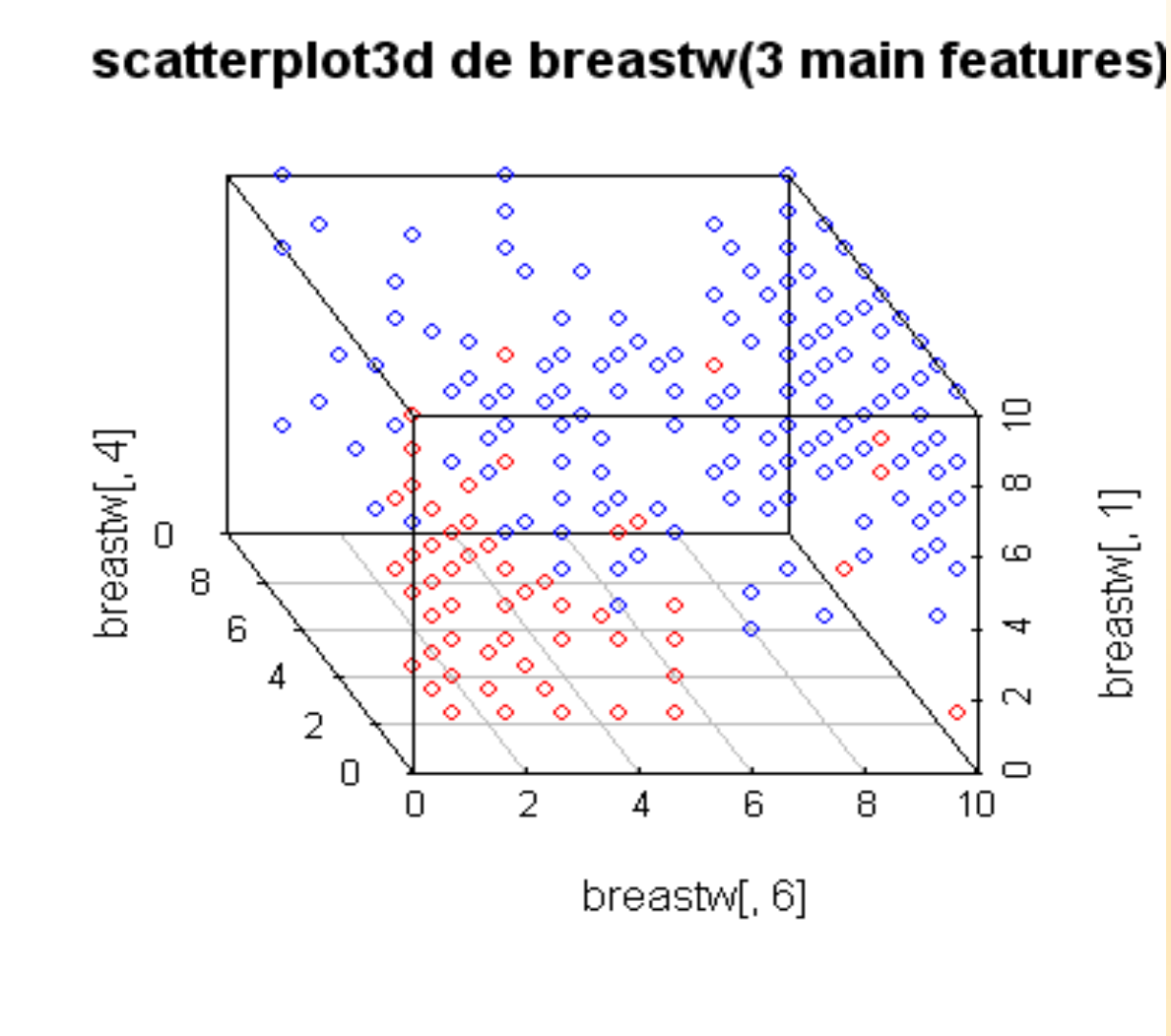
En Python: Las libreria matplotlib, bokeh, plotly hacen scatterplot 3D

Scatter3d de la libreria Rcmdr

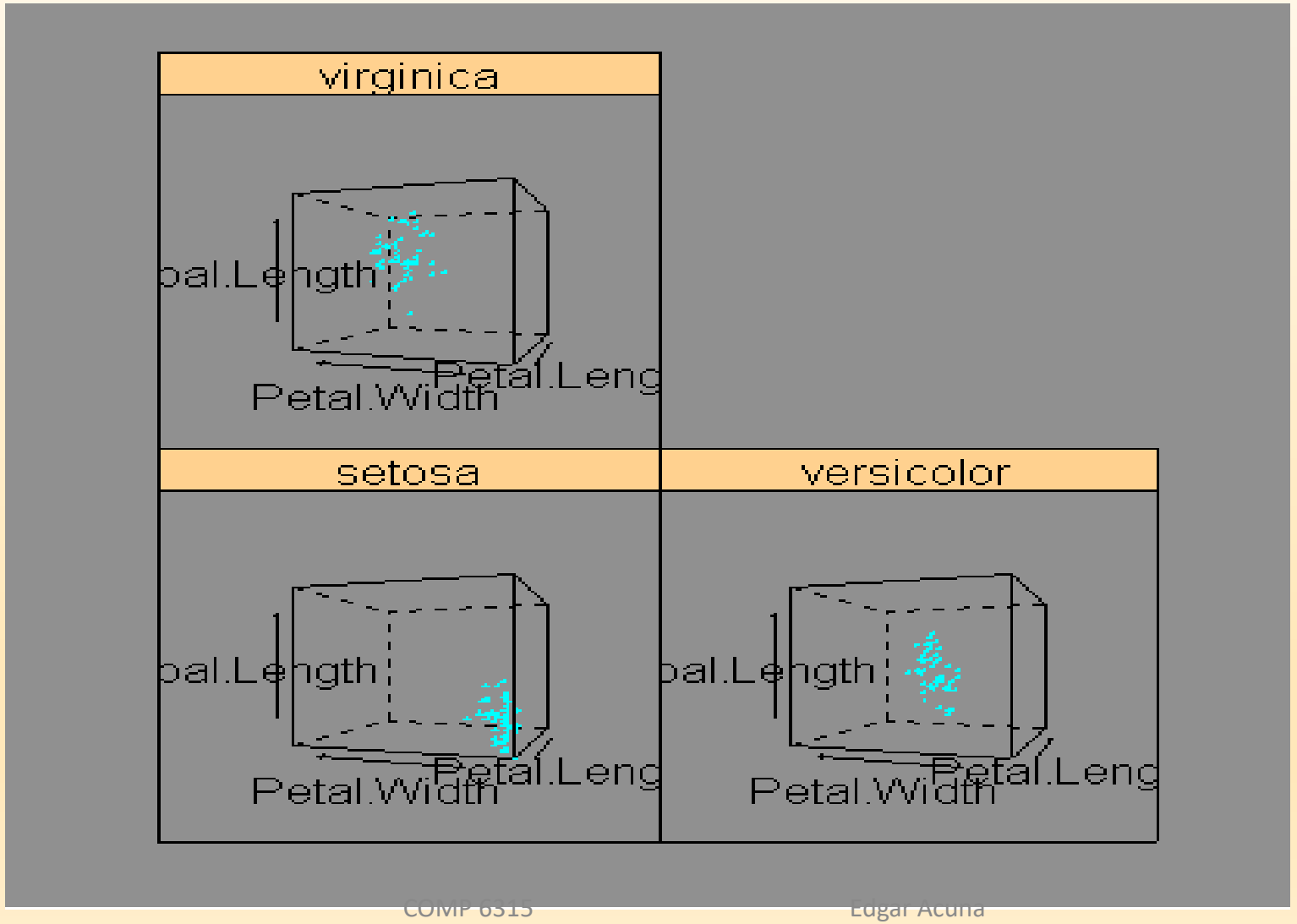


Scatterplot3d de la libreria scatterplot3d

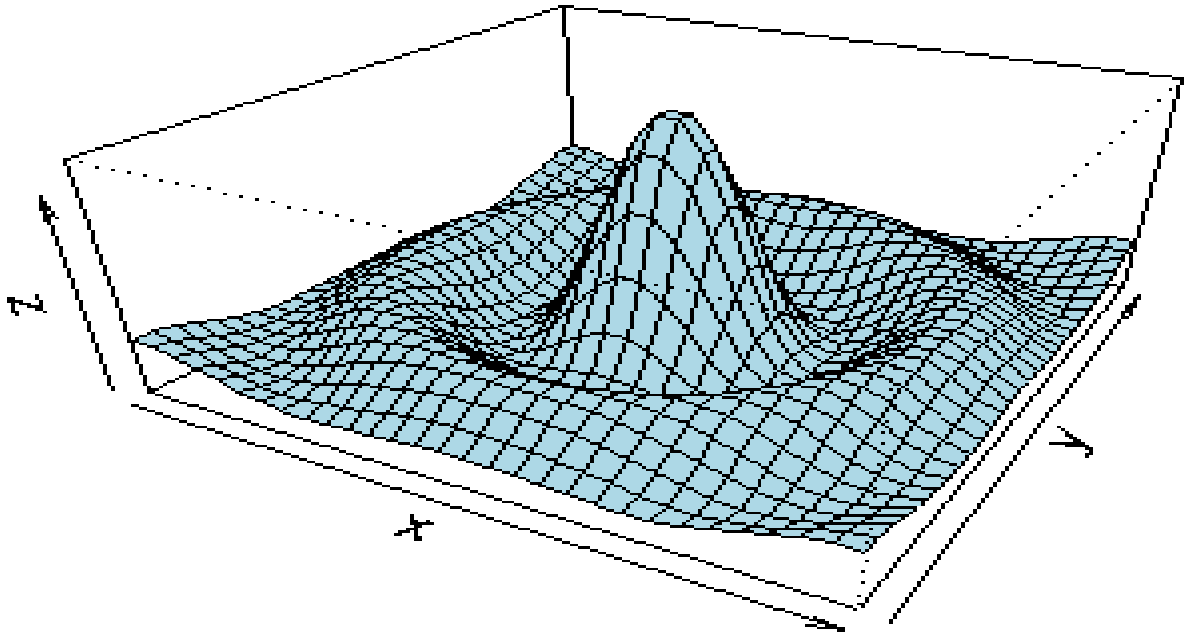
scatterplot3d(breastw[,1],breastw[,4],breastw[,6],color)



Cloud() de la libreria lattice

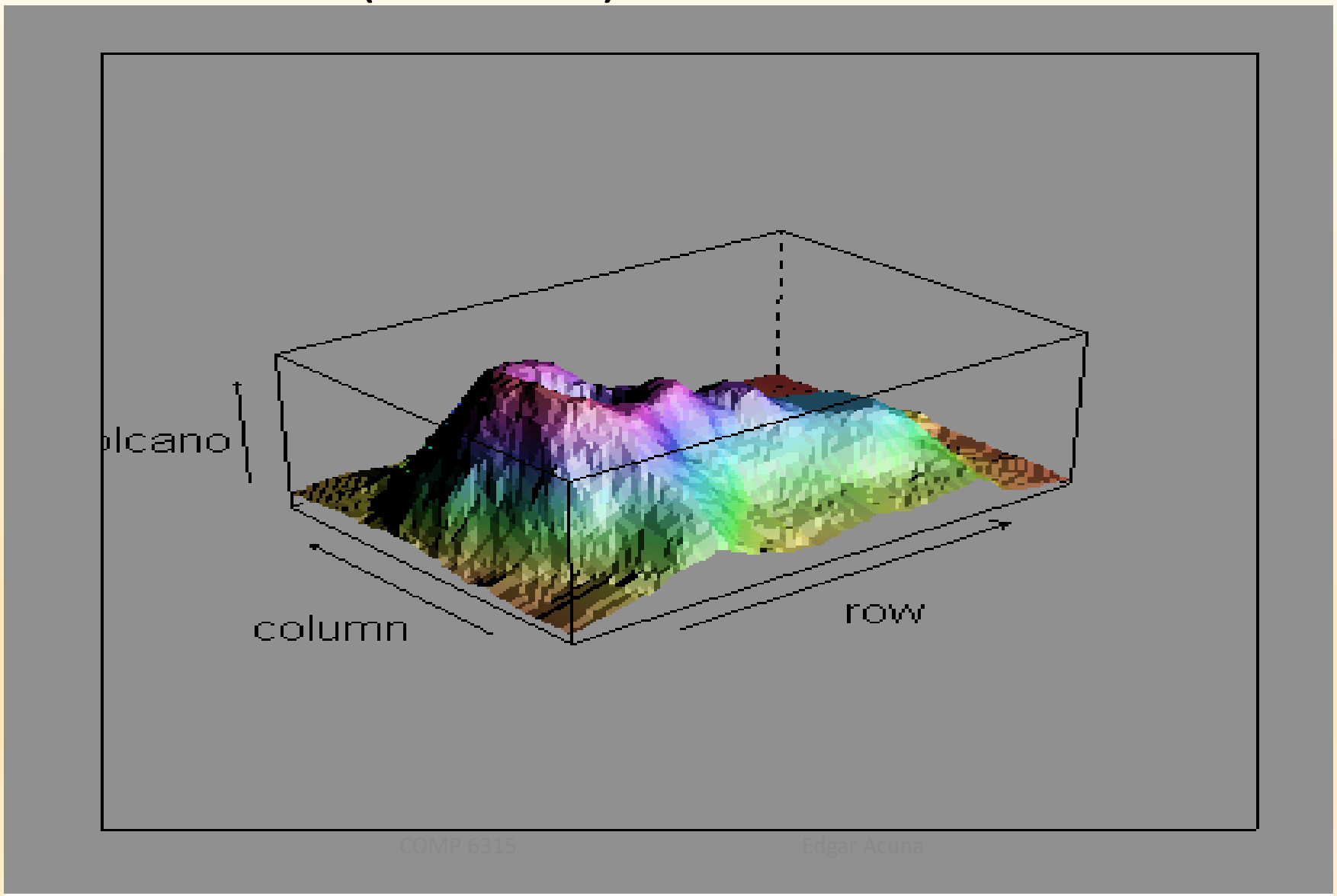


3-D Data (persp)




```
> 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)



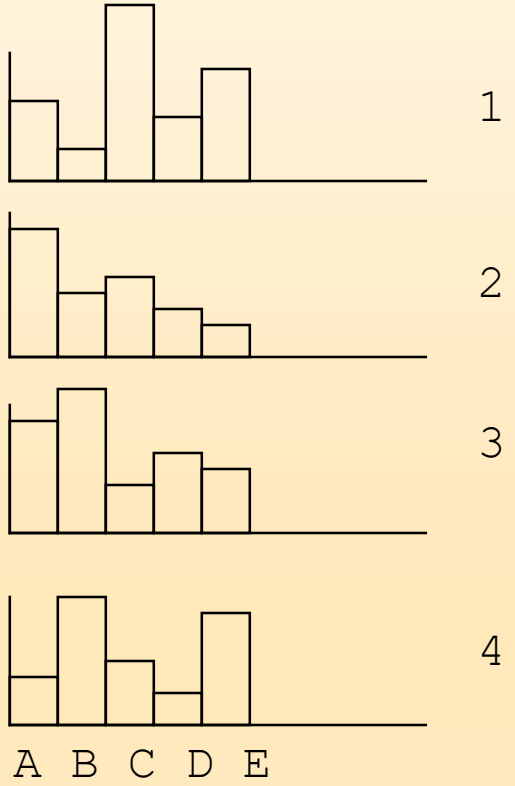
Visualizando en 4 o mas Dimensiones

- Scatterplot Matrix
- Survey Plot
- Parallel coordinate plot
- Radviz
- Star Coordinates

Multiple Vistas

Cada variable es graficada separadamente

	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



Problema: No se muestran las correlaciones

pairs() Scatterplot Matrix

Presenta plots 2-D para cada possible par de variables.

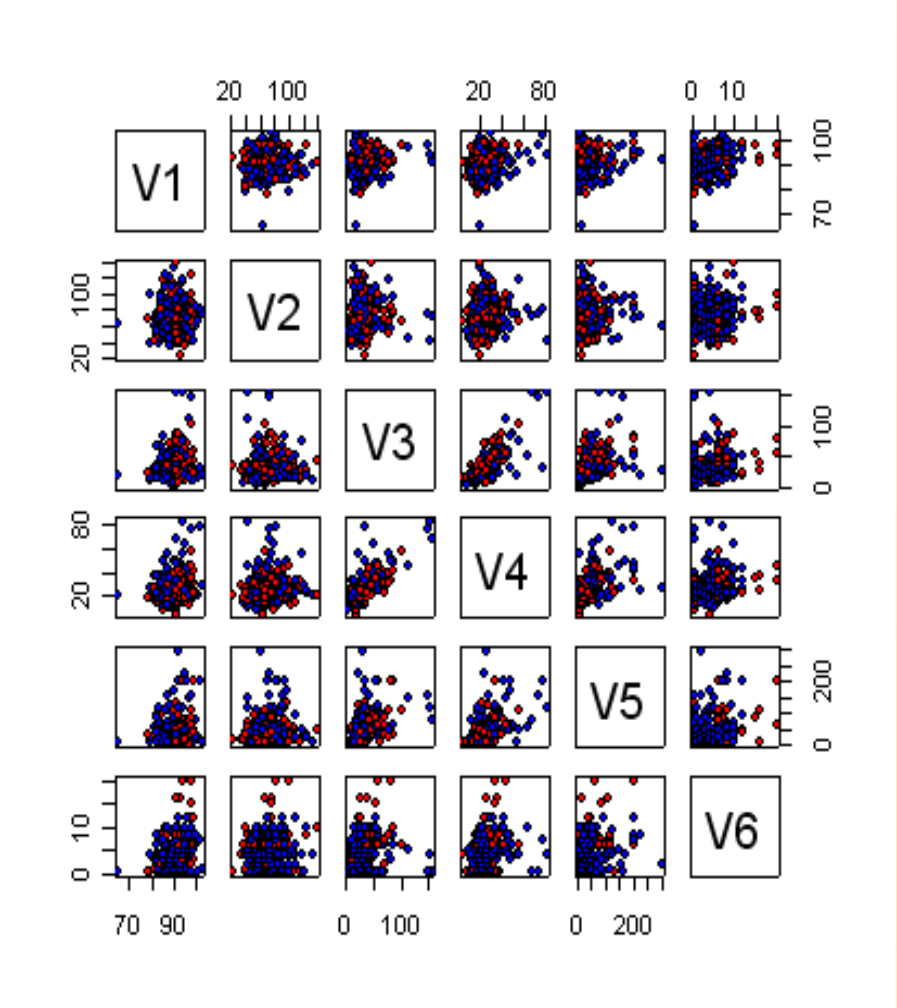
Util para detectar

Correlaciones Lineales (e.g. V3 & V4)

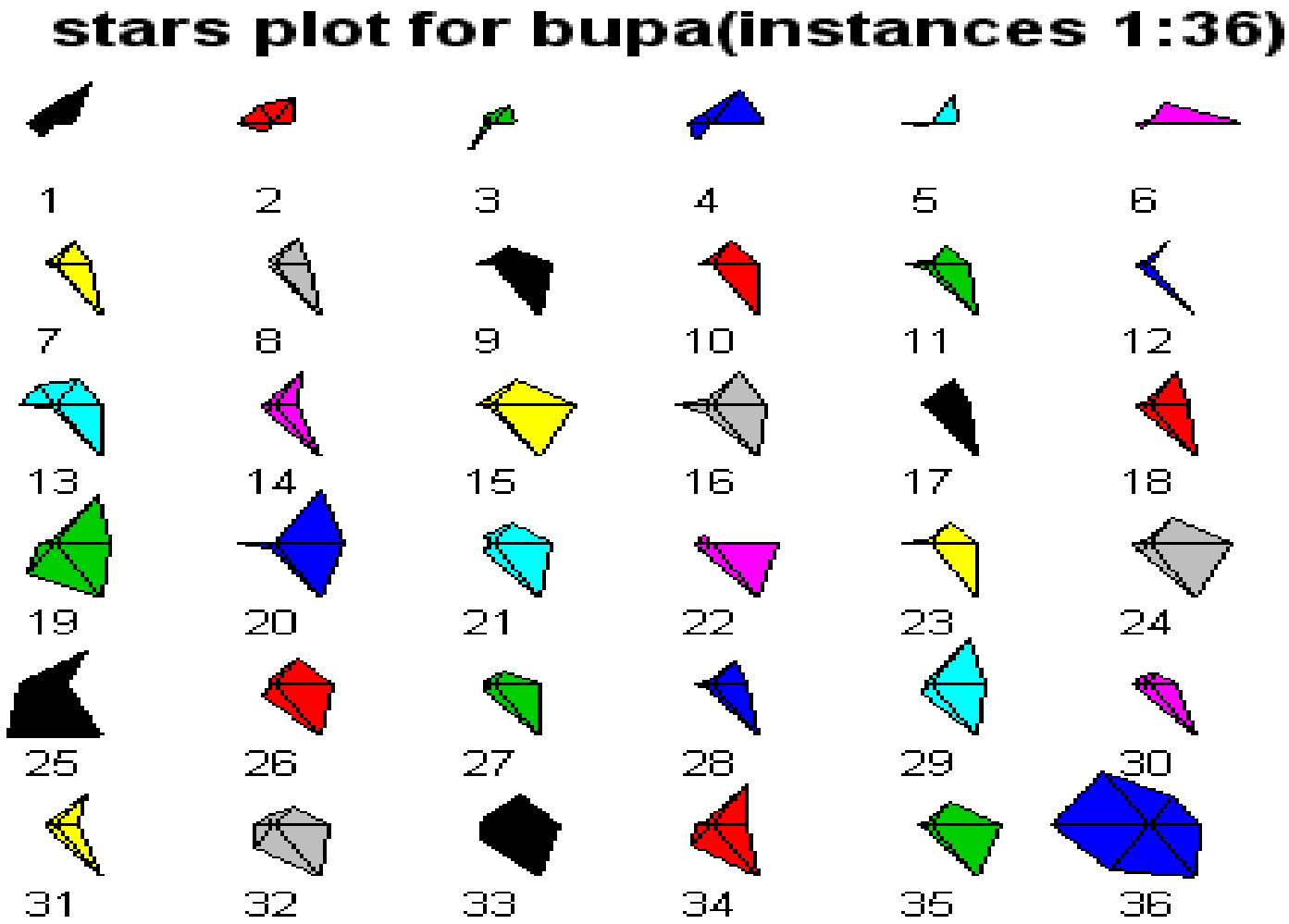
Pero no detecta

efectos multivariados

Esta disponible en Pandas, seaborn y en plotly (algo limitado)



Star Plots (Chambers et al., 1983)



Funciones de dprep para Visualization:

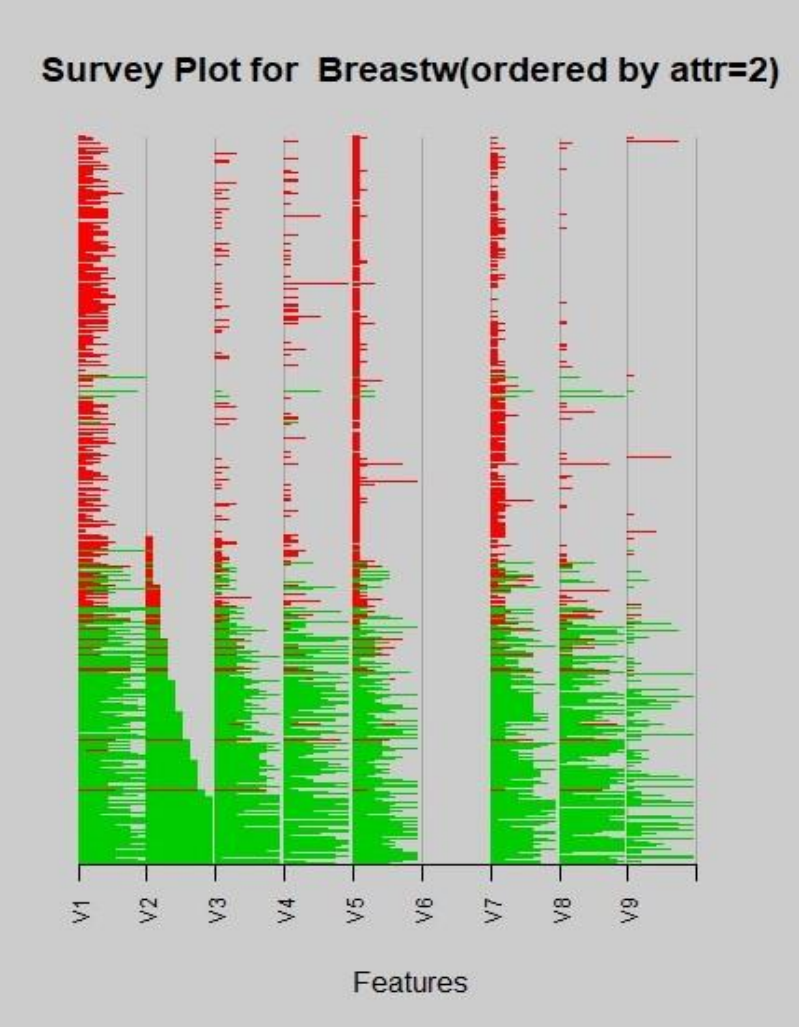
- *imagmiss( )* determina la existencia de missing values en un conjunto de datos.
- *surveyplot( )* construye un survey plot del conjunto de datos
- *parallelplot( )* construye un plot de coordenadas paralelas del conjunto de datos
- Starcoord(), Starcoor3d() contruye un plot de coordenadas star.
- Radviz() construye un plot de vizualicion radial

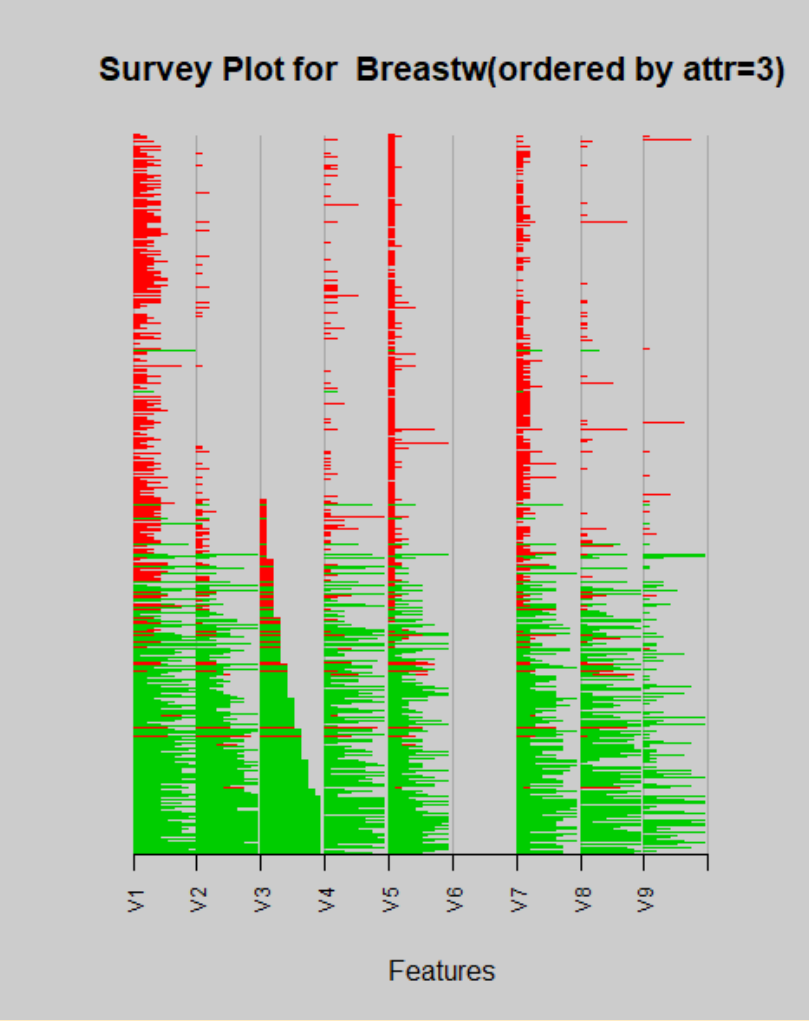
The survey plot (Lohninger, 1994)

- Es una tecnica de visualizacion inventada por el cartografo, Jacques Bertin, que se relaciona bastante con las tecnicas de visualizacion: Graficas de barras y Grafica de matriz de permutacion
- Consiste de n areas rectangulares, una por cada dimension del conjunto de datos, las cuales son arregladas verticalmente.
- Cada valor de un atributo es mapeado a un punto en la barra vertical y su valor es extendido a una linea con longitud proporcional al valor correspondiente.
- La fortaleza de esta tecnica de visualizacion esta en reconocer las relaciones y dependencias entre dos atributos cualesquiera. Cuando los datos son ordenados de acuerdo a un cierto atributo y aparecen atributos que muestran el mismo orden, entonces esos atributos estan correlacionados entre si
- El program Orange tambien tiene esta grafica.

The survey plot:

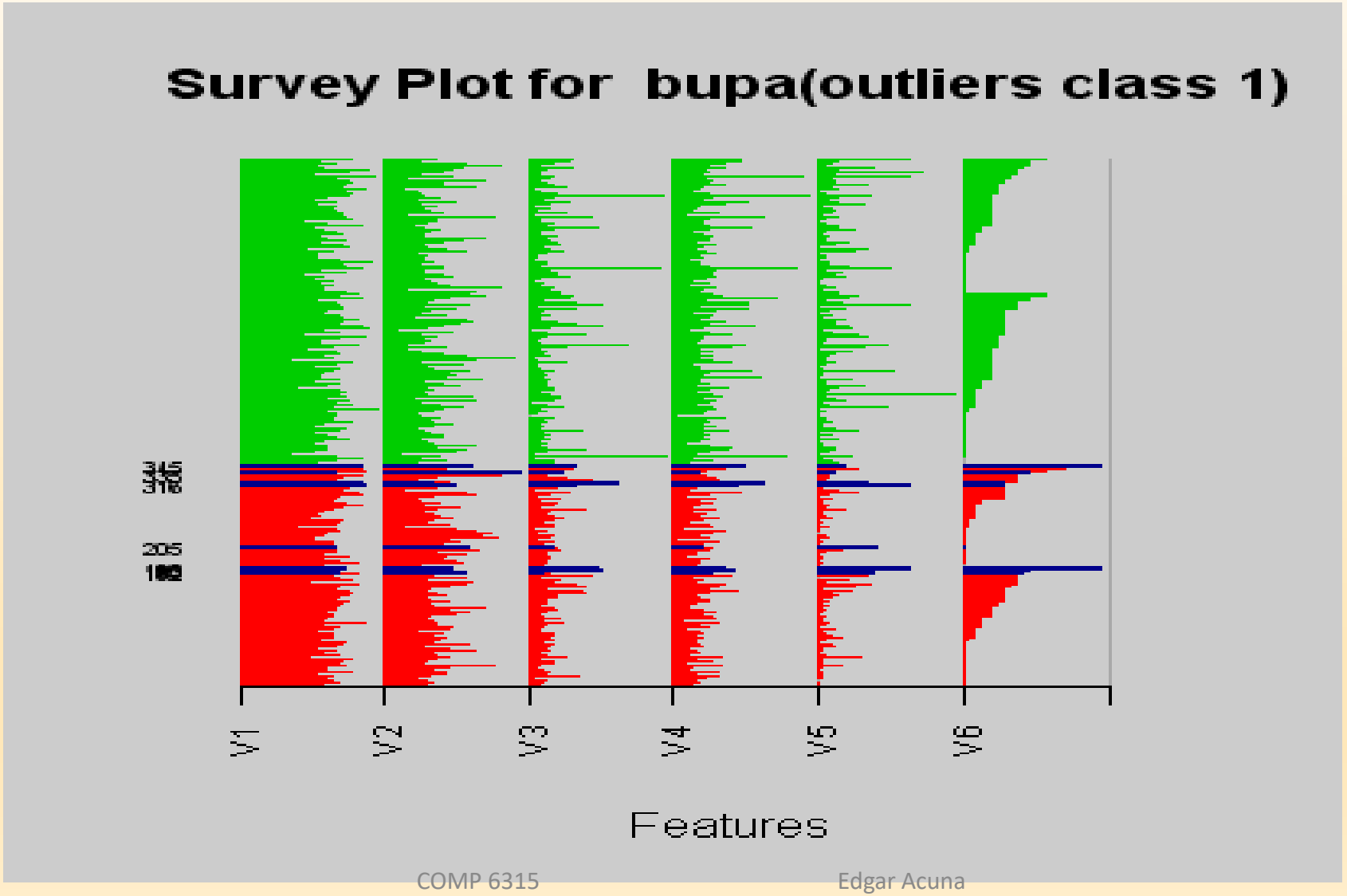
surveyplot(dataset: matrix , name: string, class: integer, orderon: integer, obs: list of integer)





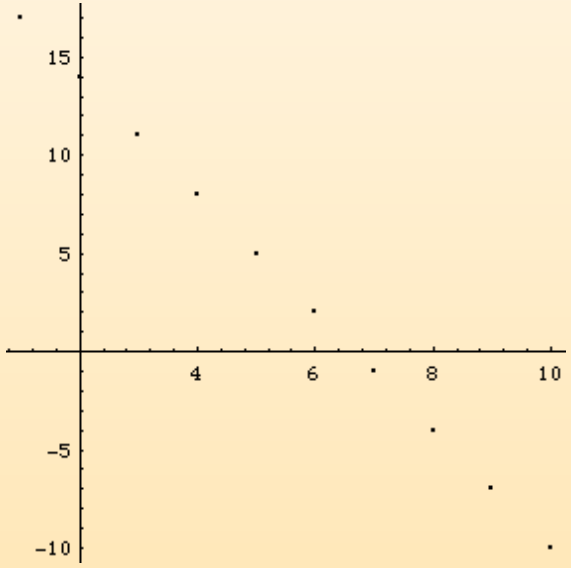
Las variables V2 y V3 tienen mas o menos el mismo comportamiento, esto sugieere que estan correlacionadas

Surveyplot as a tool to detect outliers

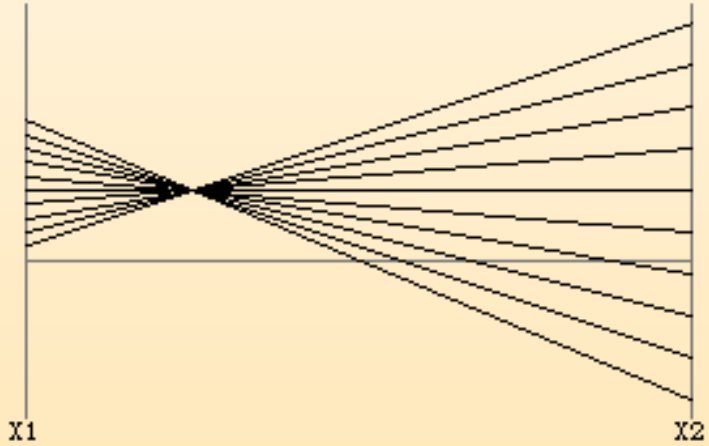


Coordenadas Paralelas

- Cada variable es representada por un linea vertical
- En cada linea vertical se especifica el valor de cada variable



Datos en el coordenadas cartesianas



El mismo conjunto de datos en coordenadas paralelas

Invented by Alfred Inselberg while at IBM, 1985

El plot de coordenadas paralelas

- Fue introducido por Al Inselberg (1985) y representa datos multidimensionales usando lineas.
- A diferencia de las coordenadas cartesianas tradicionales donde todos los ejes son perpendiculares, en el plot de coordenadas paralelas todos los ejes son paralelos e igulamente espaciados.
- En este metodo, un punto en el espacio *m*-dimensional es representado como una serie de m-1 segmentos de linea en un espacio bi-dimensional. Asi si la observacion origina es escrita como $(x_1, x_2, \ldots x_m,)$, entonces su representacion en coordendas paralela es la union de *m*-1 segmentos lineales que conecta a los puntos $(1,x_1)$, $(2,x_2)$, . . . (m,x_m).
- Se recomienda que los atributos sean normalizados antes de trazar el plot de coordendas paralelas.

Ejemplo: Visualizando el conjunto de datos Iris

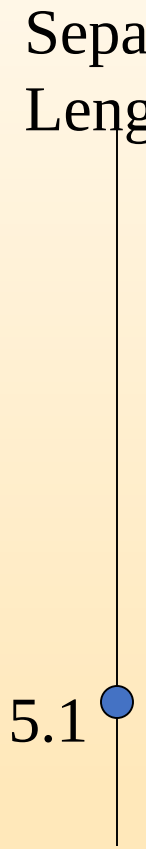
Iris setosa

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

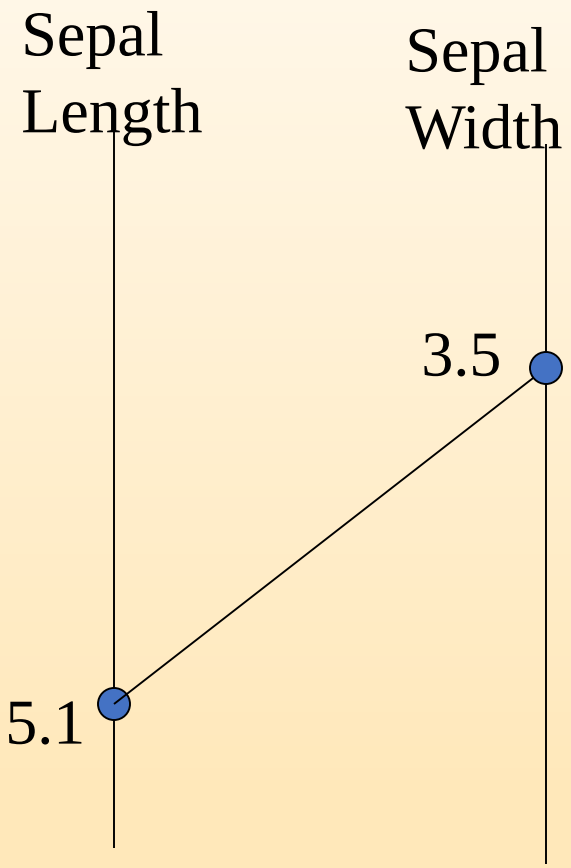
Iris virginica

Coordenadas Paralelas



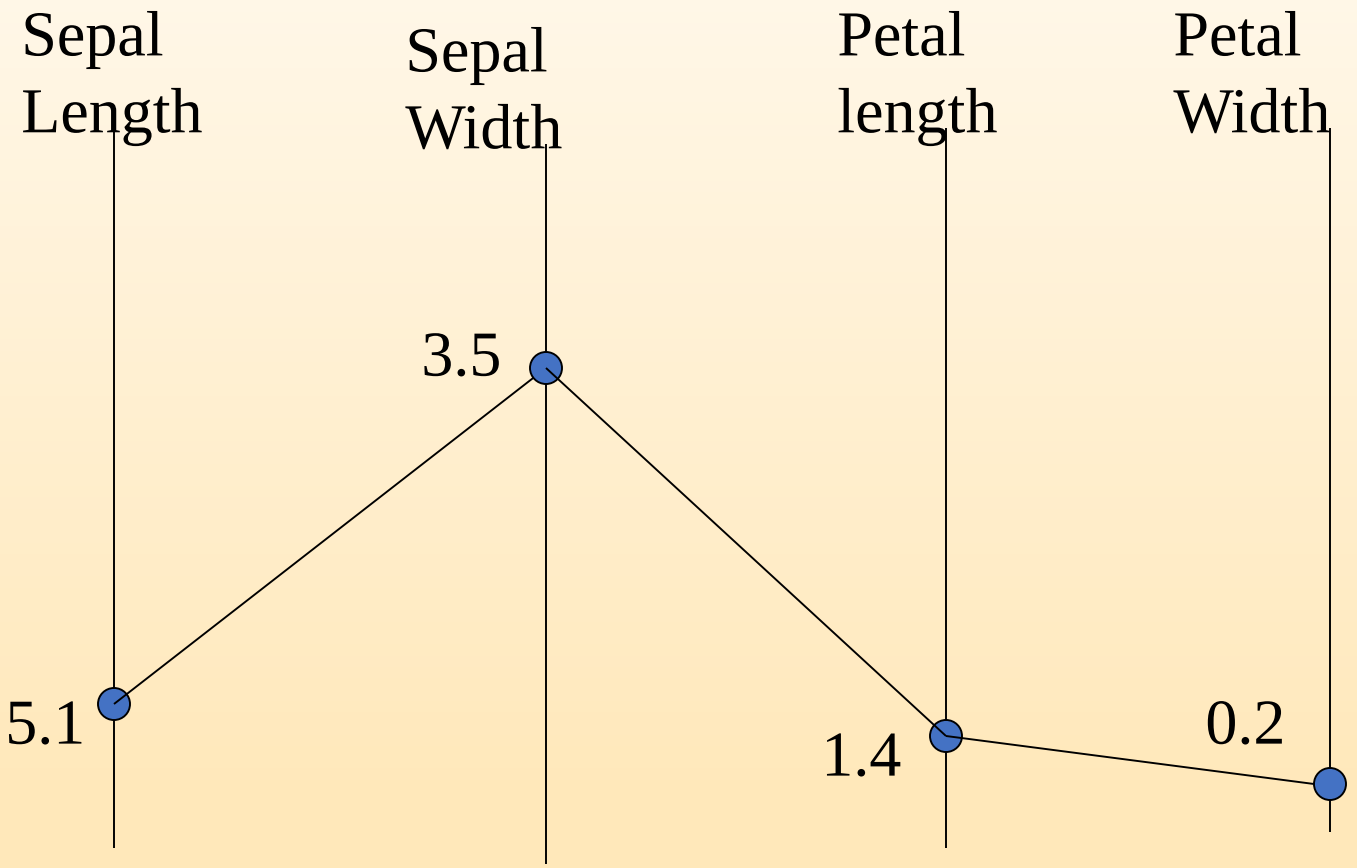
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Coordenadas Paralelas: 2 D



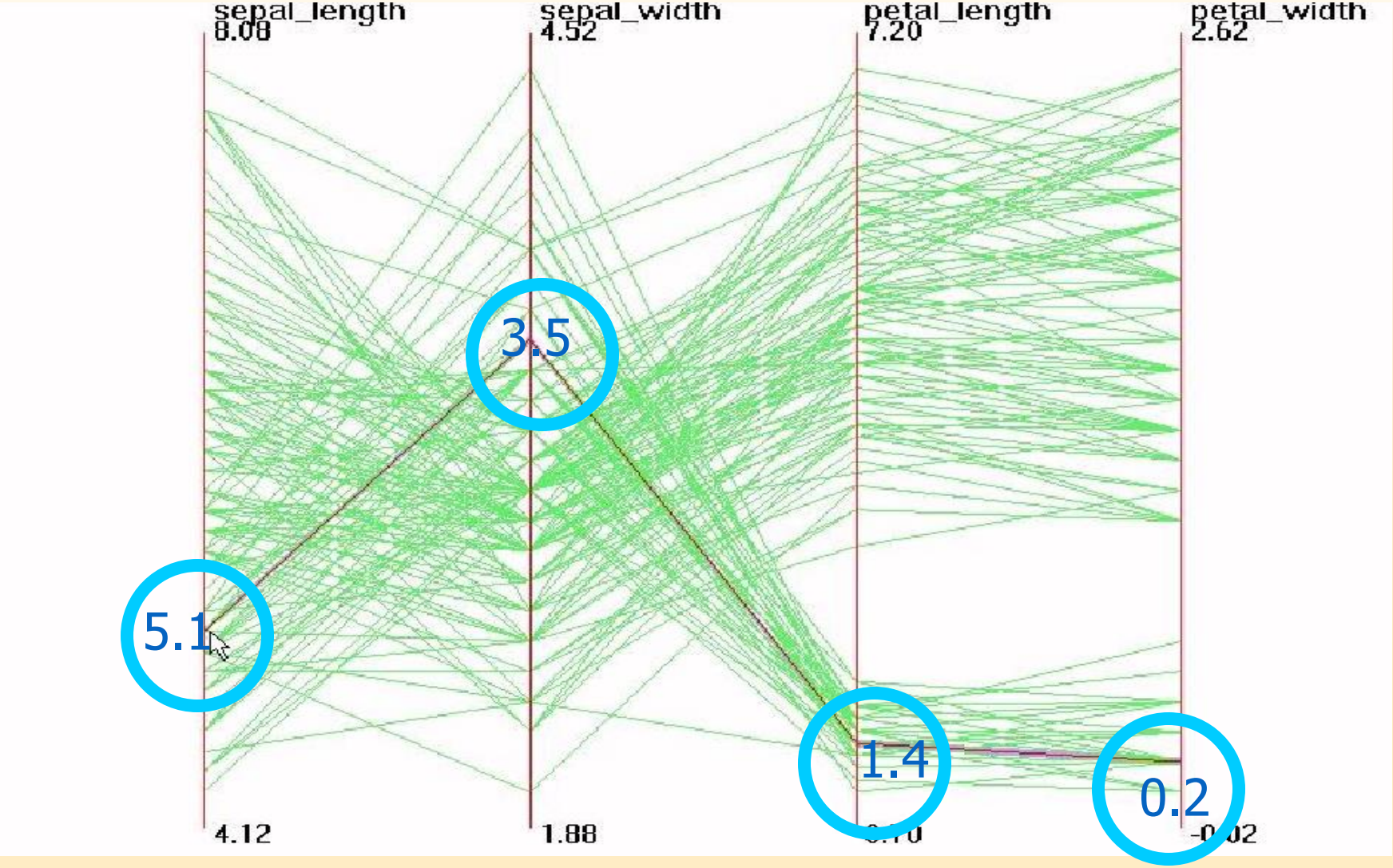
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Coordenadas Paralelas: 4 D



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

Visualizacion Paralela de los datos de Iris



Plot de Coordenadas Paralelas (cont)

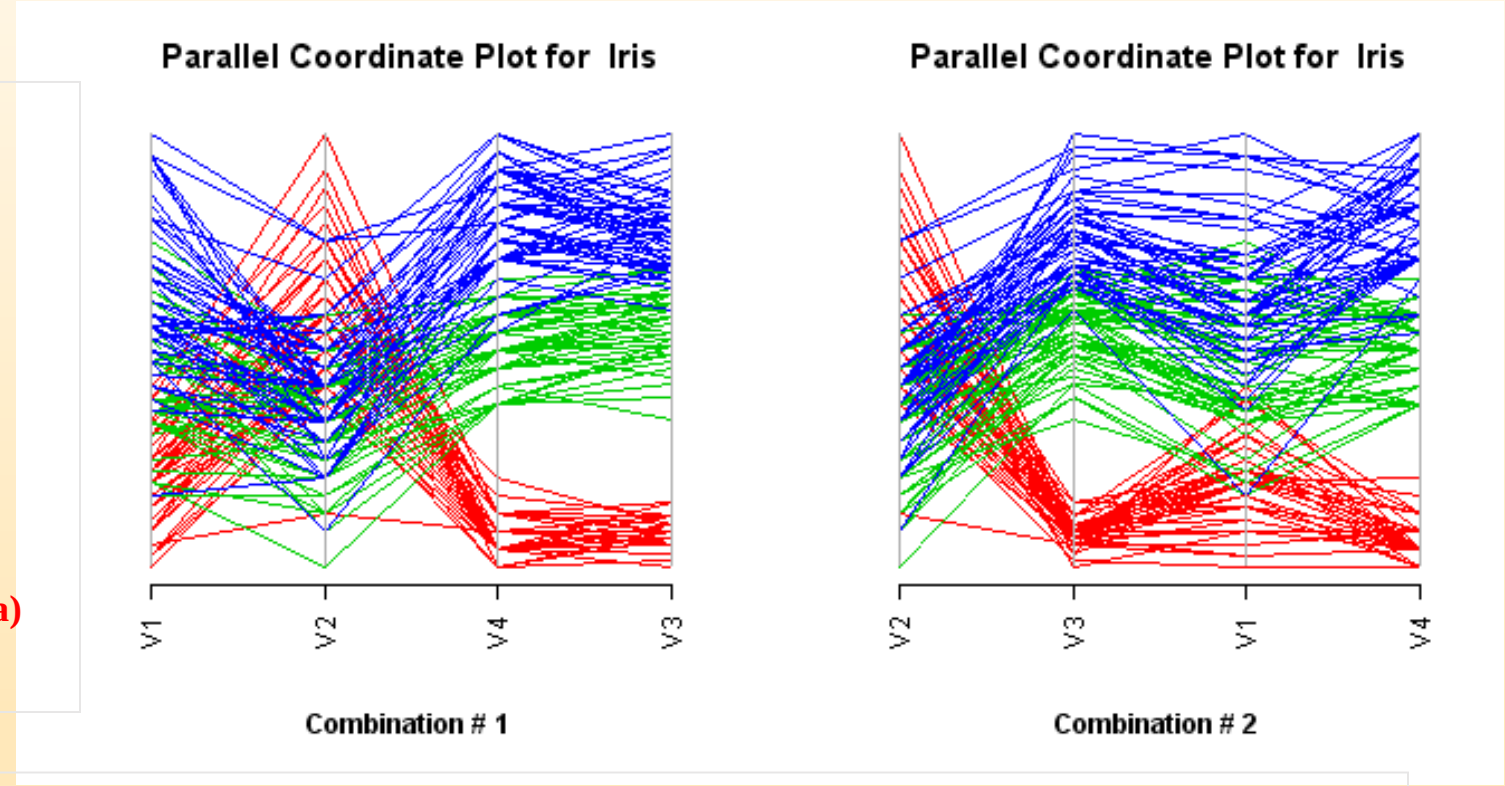
- La comparacion en parejas es limitada solo a los atributos que estan en ekes adyacentes.
- Para un conjunto de datos con p atributos, se tienen p! permutaciones de los atributos de tal manera que cada uno de ellos sea adyacente a cada atrbuto en alguna permutacion.
- Wegman (1990) determine que son necesarias solo [(p+1)/2] permutaciones.([.] is function entero mayor).

El plot de coordenadas paralelas

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

Iris dataset:

- **Datos acerca de Flores.**
- **4 atributos (sepal length, sepal width petal length, y petal width,)**
- **150 instancias**
- **3 classes (Setosa, Versicolor, Virginica)**
- **No missing values.**



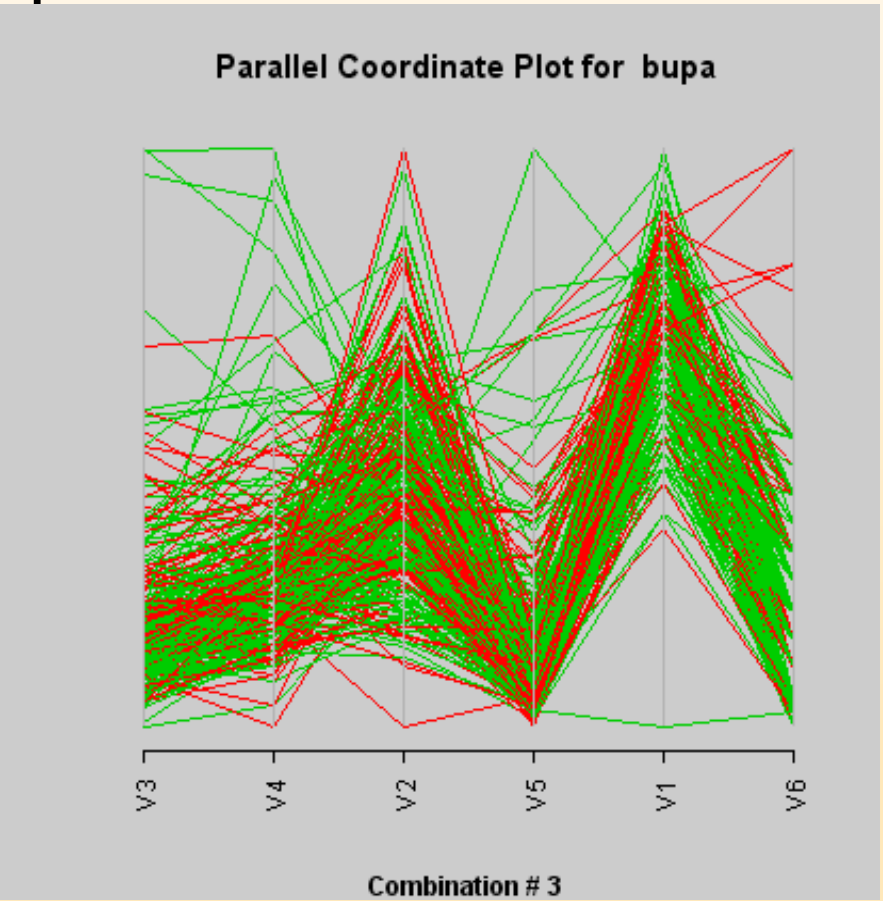
Interpretacion:

- **Cada color distinto representa una clase.**
- **Si dos atributos tienen correlacion positive alta, lines que pasan de un atributo a otro tienden ano intersectarse entre los ejes de coordenadas paralelas. (V3 y V4)**

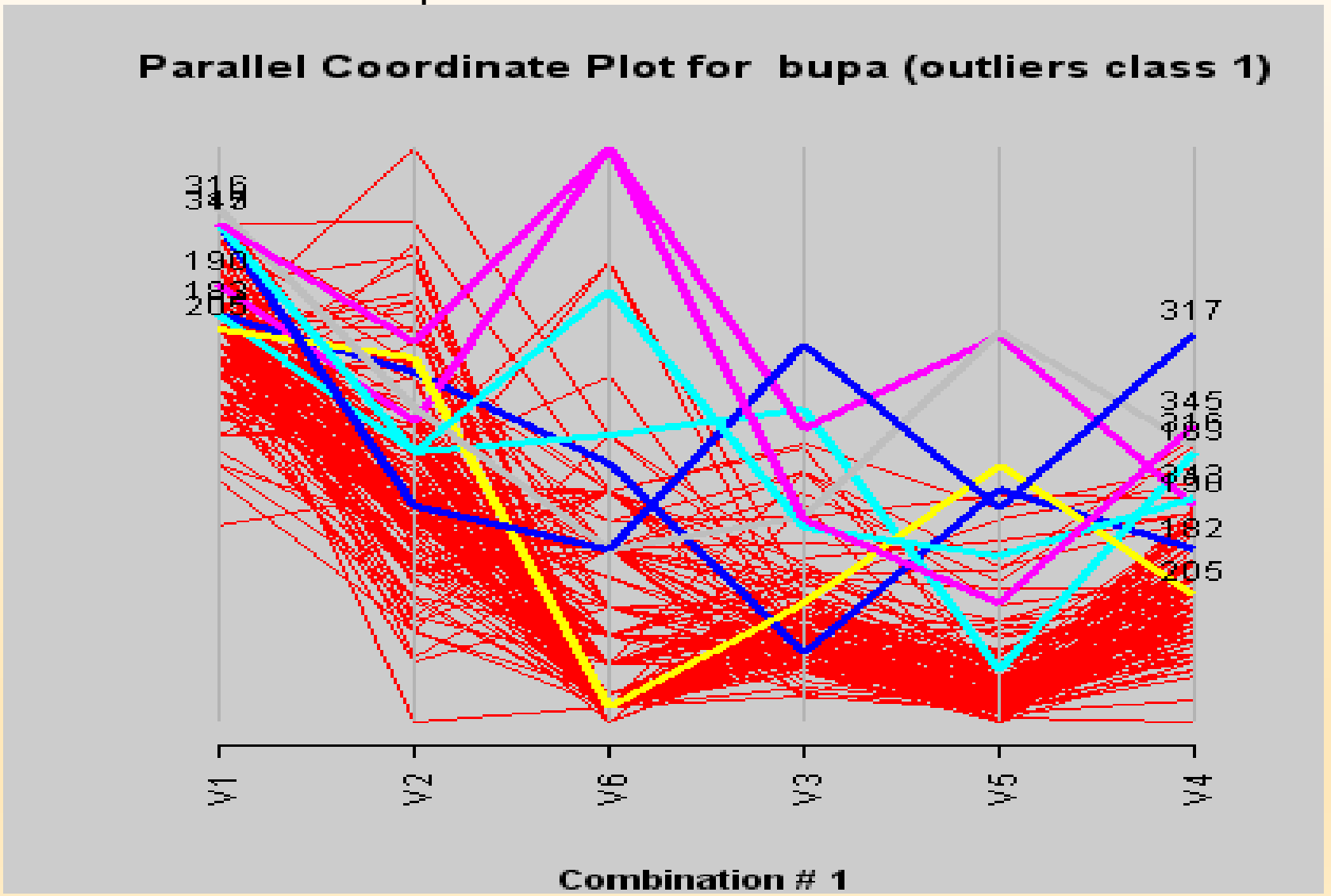
The parallel coordinate plot

- **Cuando dos atributos que tiene correlacion negativa entonces los segmentos de lineas tienden a a cruzarse en un solo punto ubicado entre dos ejes de coordenadas paralelas.**
- **La presencia de outliers es sugerido por observaciones con segmentos de lineas que no siguen el patron de su clases.**

Una limitacion de esta grafica es la perdida de informacion que dan las lineas entre atributos discretos.



Plot de Cordenadas Paralelas como una herramienta para detectar outliers

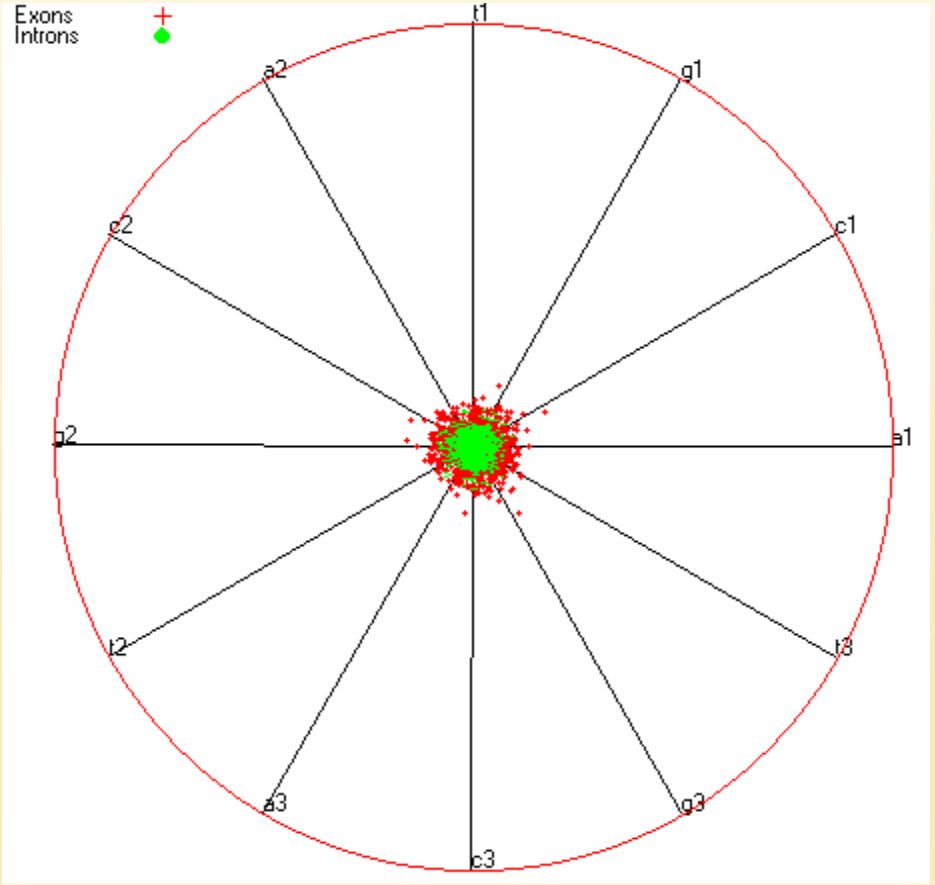


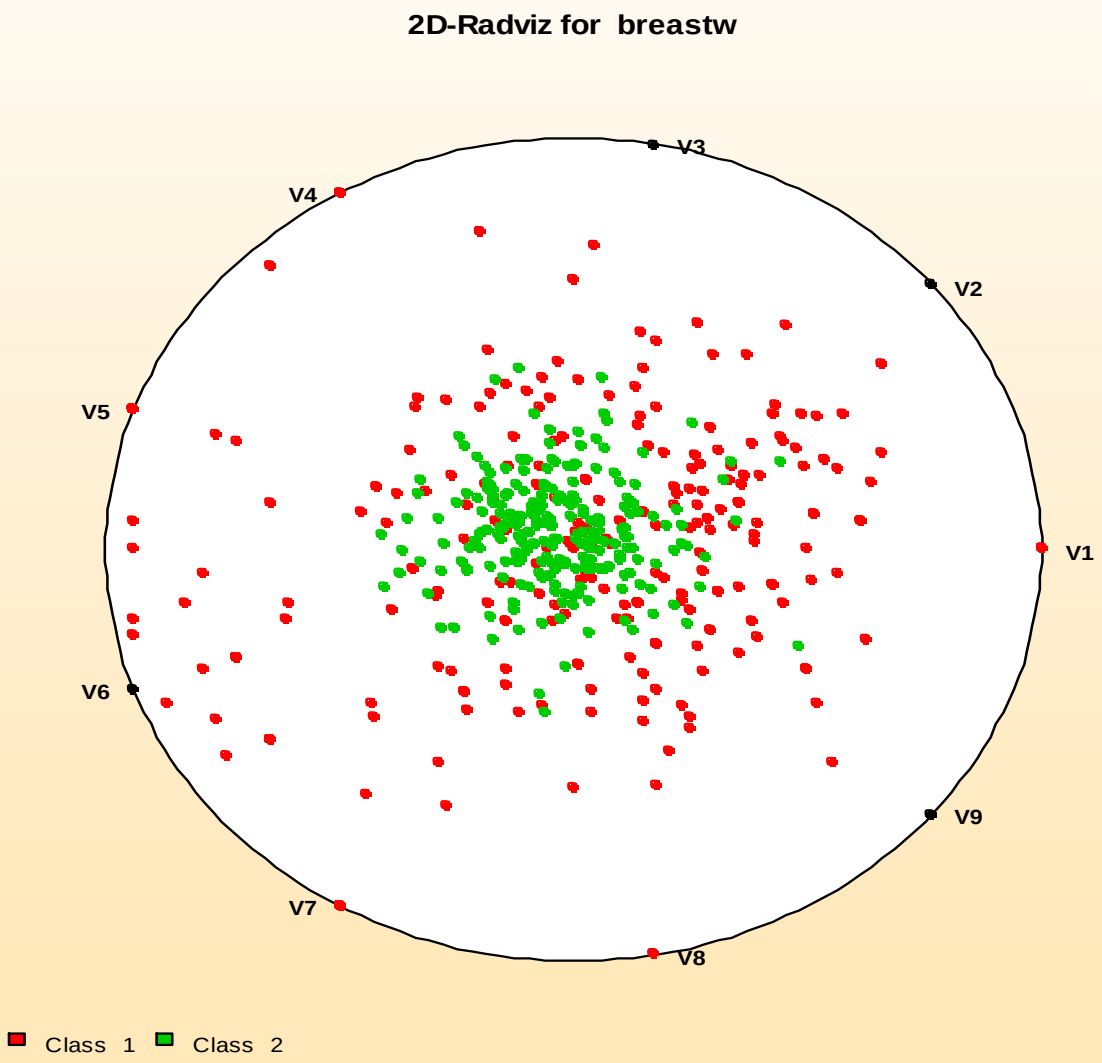
Resumen de Coordenadas Paralelas

- Cada dato observado es una linea.
- Puntos similares corresponden a lineas similares.
- Lineas que se cruzan corresponden a atributos con correlacion negative.
- Se puede usar para explorer la data interactivamente y para clustering.
- Problemas: El ordenamiento de los ejes, limite de aproximadamente 20 dmensiones.
- Esta disponible en Pandas (Python)

RadViz (Ankerst, et al., 1996)

- Es un metodo de visualizacion radial.
- Un spring le corresponde a cada atributo.
- Un extremo del spring esta conectado al borde del circulo donde esta ubicada la posicion del atributo y el otro estermo esta conectado al dato observado.
- Cada punto del conjunto de datos es mostrado dentro del circulo en el punto donde la suma de las fuerzas de los springs da CERO.
- Es adecuado para deteccion de outliers
- Esta disponible en Pandas y en Orange. En R Tambien hay una libreria Radviz



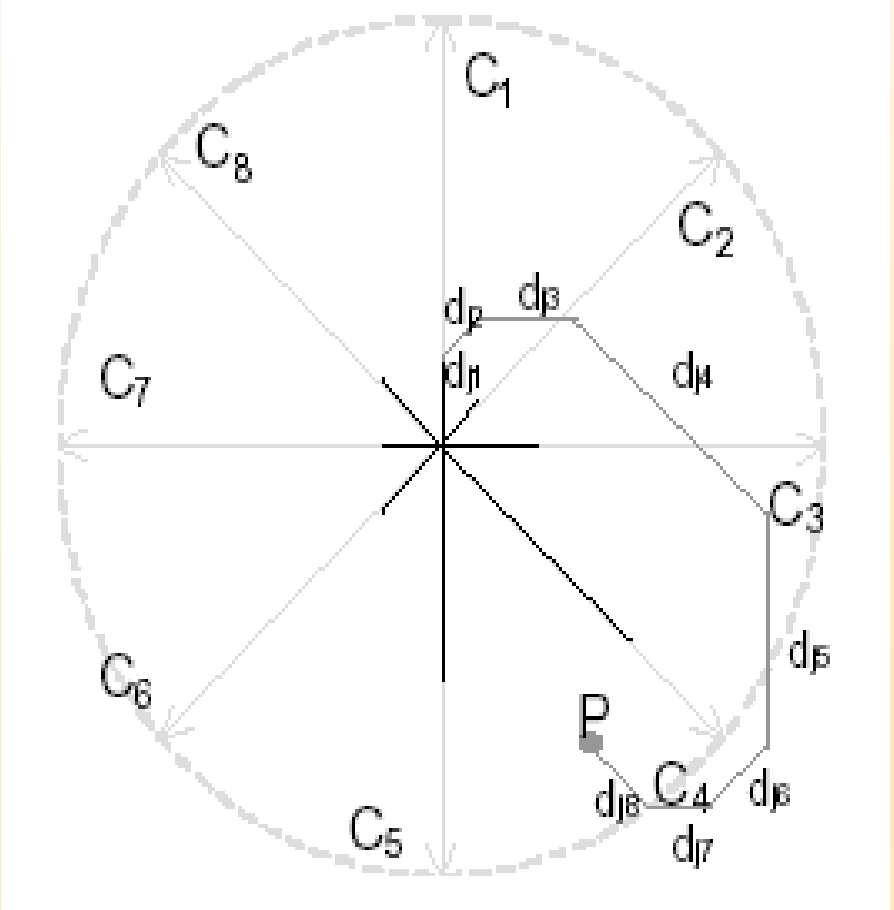

2D-Radviz for breastw
V3
V4
V2
V5
V1
V6
V9
V7
V8
Class 1
Class 2

Star Coordinates (Kandogan, 2001)

- A cada atributo le corresponde un eje dimensional.
- El valor de un dato en cada dimension es representado como un vector.
- Los valores observados observados de cada atributo son escalados con respect a la longitude del eje.

  - el valor minimo se mapea al origen.

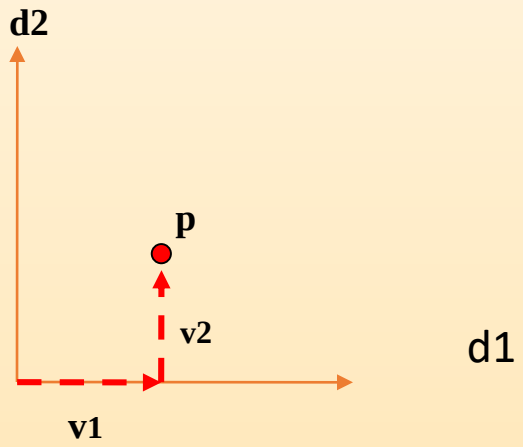
  - el valor maximo se mapea al extremo final



Star Coordinates Contd

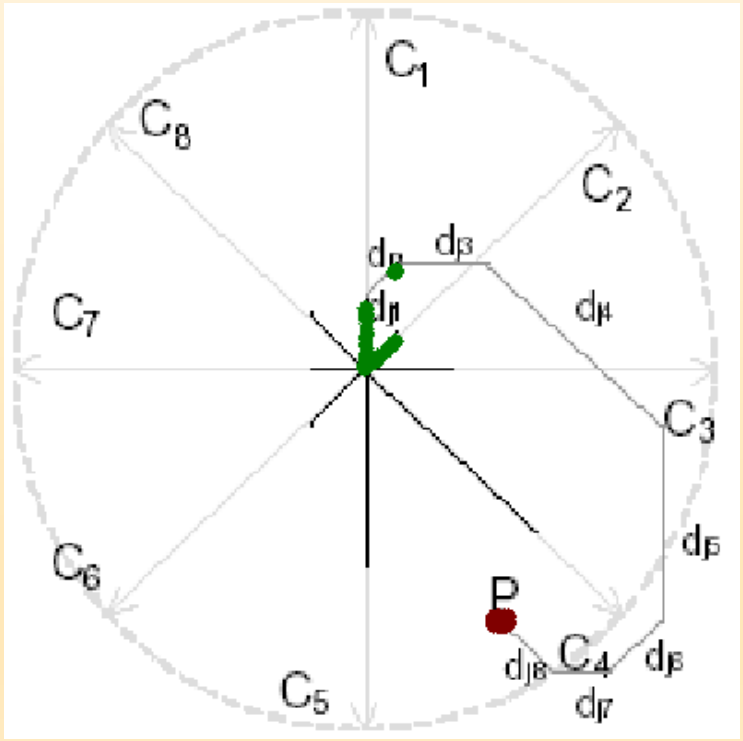
Cartesian

P=(v1, v2)



Star Coordinates

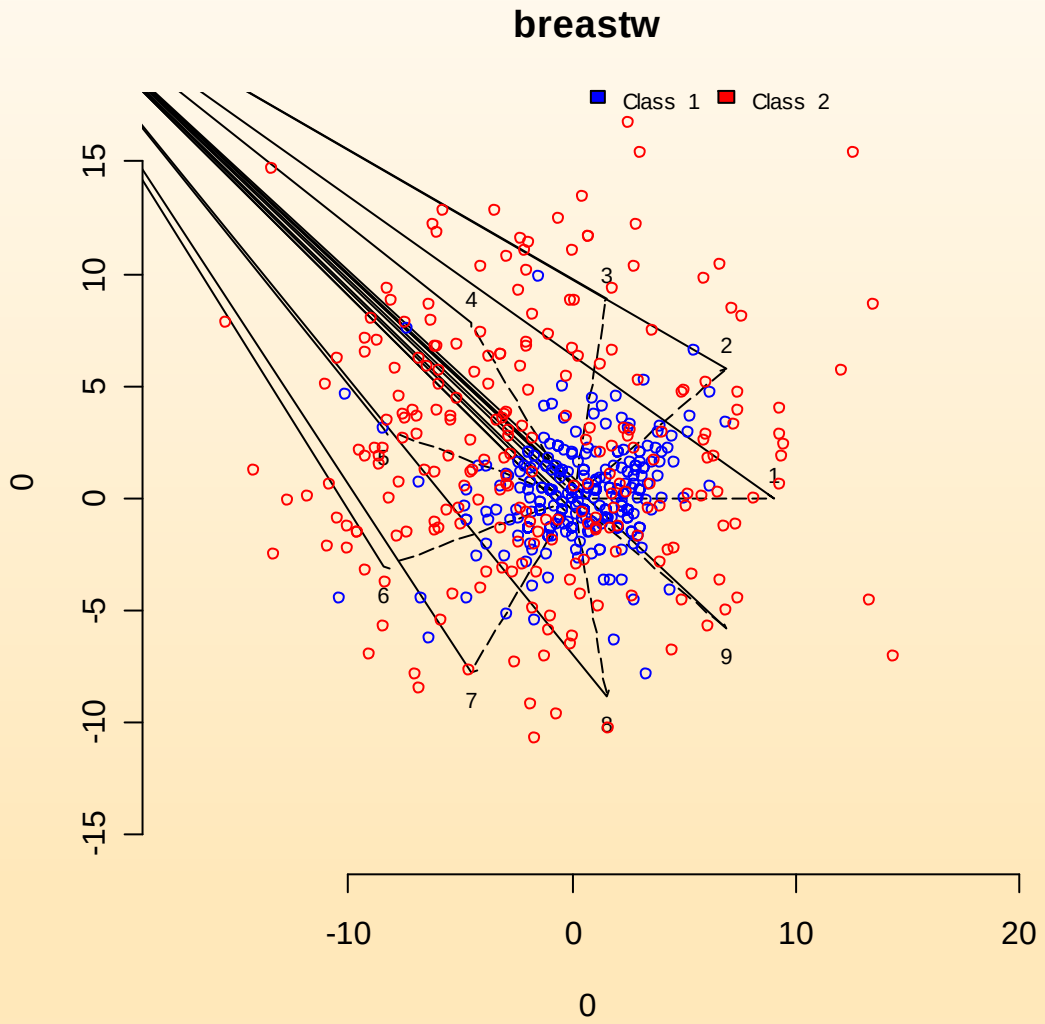
P=(v1,v2,v3,v4,v5,v6,v7,v8)



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/

- Muchos mas - ver www.kdnuggets.com/software/visualization.html