Multivariate Data Analysis and Data Mining

Outline

1. Multivariate Data
2. Data Visualization for Multivariate Data.
3. A basic multivariate example: Crime data.
4. Geometric intuition of Multivariate data.
5. Dimension Reduction Principal Components
6. Biplots
7. Clustering
8. Software

Data visualization of Multivariate Data

Most datasets contain multiple variables.
- Variables maybe correlated.

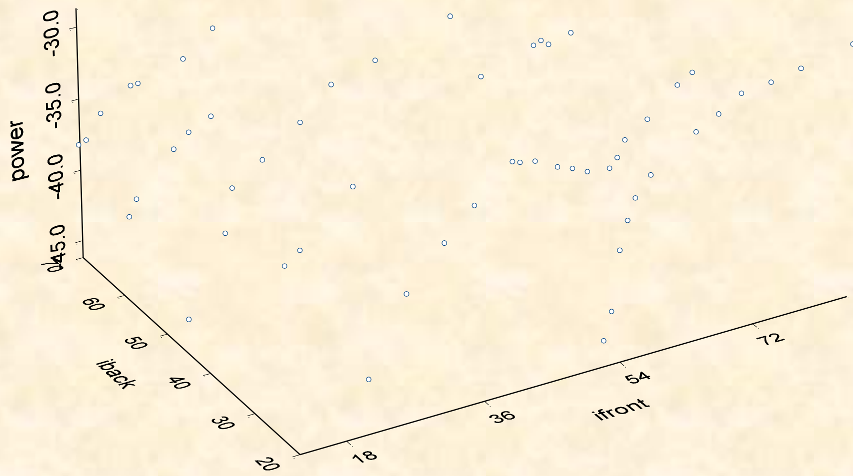
Objectives:

1. Explore, Summarize , reduce dimensionality
2. Find
   - Interesting patterns linear or nonlinear
   - Clusters
   - Outliers.

DATA VISUALIZATION OF MULTIVARIATE DATA

2D Plots: **Masking with color.**

3D Plots: **Are sometimes useful but may need animation (This example is from Splus)**



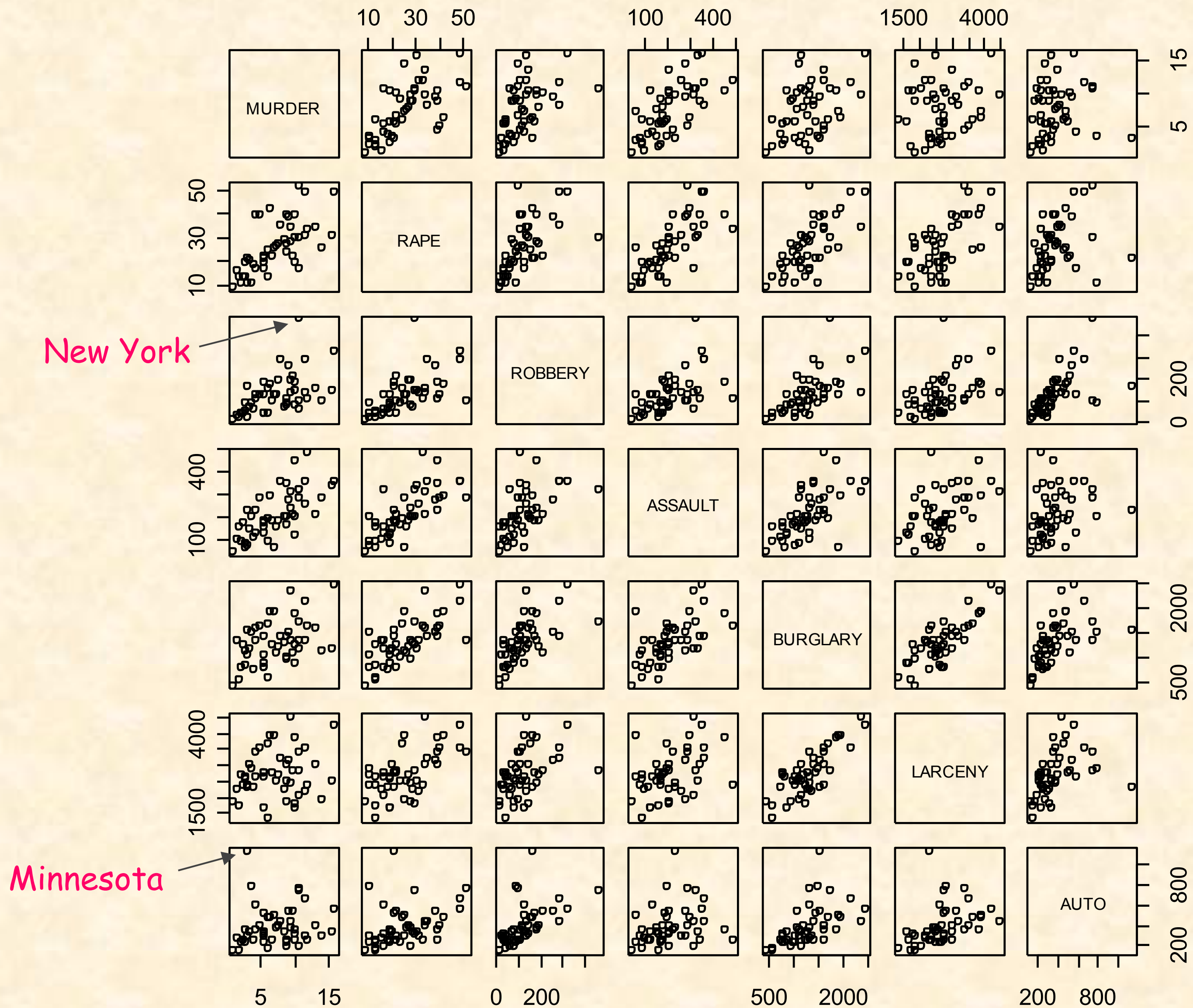
Scatter Matrices:

EXAMPLE: CRIME RATES
(PER 100,000 POPULATION BY STATE)

STATE	MURDE	RAPE	ROBBER	ASSAULT	BURGLARY	LARCENY	AUTO
ALABAMA	14.20	25.20	96.80	278.30	1135.50	1881.90	280.70
ALASKA	10.80	51.60	96.80	284.00	1331.70	3369.80	753.30
ARIZONA	9.50	34.20	138.20	312.30	2346.10	4467.40	439.50
ARKANSAS	8.80	27.60	83.20	203.40	972.60	1862.10	183.40
CALIFORN	11.50	49.40	287.00	358.00	2139.40	3499.80	663.50
COLORADO	6.30	42.00	170.70	292.90	1935.20	3903.20	477.10
CONNECTI	4.20	16.80	129.50	131.80	1346.00	2620.70	593.20
DELAWARE	6.00	24.90	157.00	194.20	1682.60	3678.40	467.00
FLORIDA	10.20	39.60	187.90	449.10	1859.90	3840.50	351.40
GEORGIA	11.70	31.10	140.50	256.50	1351.10	2170.20	297.90
HAWAII	7.20	25.50	128.00	64.10	1911.50	3920.40	489.40
IDAHO	5.50	19.40	39.60	172.50	1050.80	2599.60	237.60
ILLINOIS	9.90	21.80	211.30	209.00	1085.00	2828.50	528.60
INDIANA	7.40	26.50	123.20	153.50	1086.20	2498.70	377.40
IOWA	2.30	10.60	41.20	89.80	812.50	2685.10	219.90
KANSAS	6.60	22.00	100.70	180.50	1270.40	2739.30	244.30
KENTUCKY	10.10	19.10	81.10	123.30	872.20	1662.10	245.40
LOUISIAN	15.50	30.90	142.90	335.50	1165.50	2469.90	337.70
MAINE	2.40	13.50	38.70	170.00	1253.10	2350.70	246.90
MARYLAND	8.00	34.80	292.10	358.90	1400.00	3177.70	428.50
MASSACHU	3.10	20.80	169.10	231.60	1532.20	2311.30	1140.10
MICHIGAN	9.30	38.90	261.90	274.60	1522.70	3159.00	545.50
MINNESOT	2.70	19.50	85.90	85.80	1134.70	2559.30	343.10
MISSOURI	9.60	28.30	189.00	233.50	1318.30	2424.20	378.40
MONTANA	5.40	16.70	39.20	156.80	804.90	2773.20	309.20

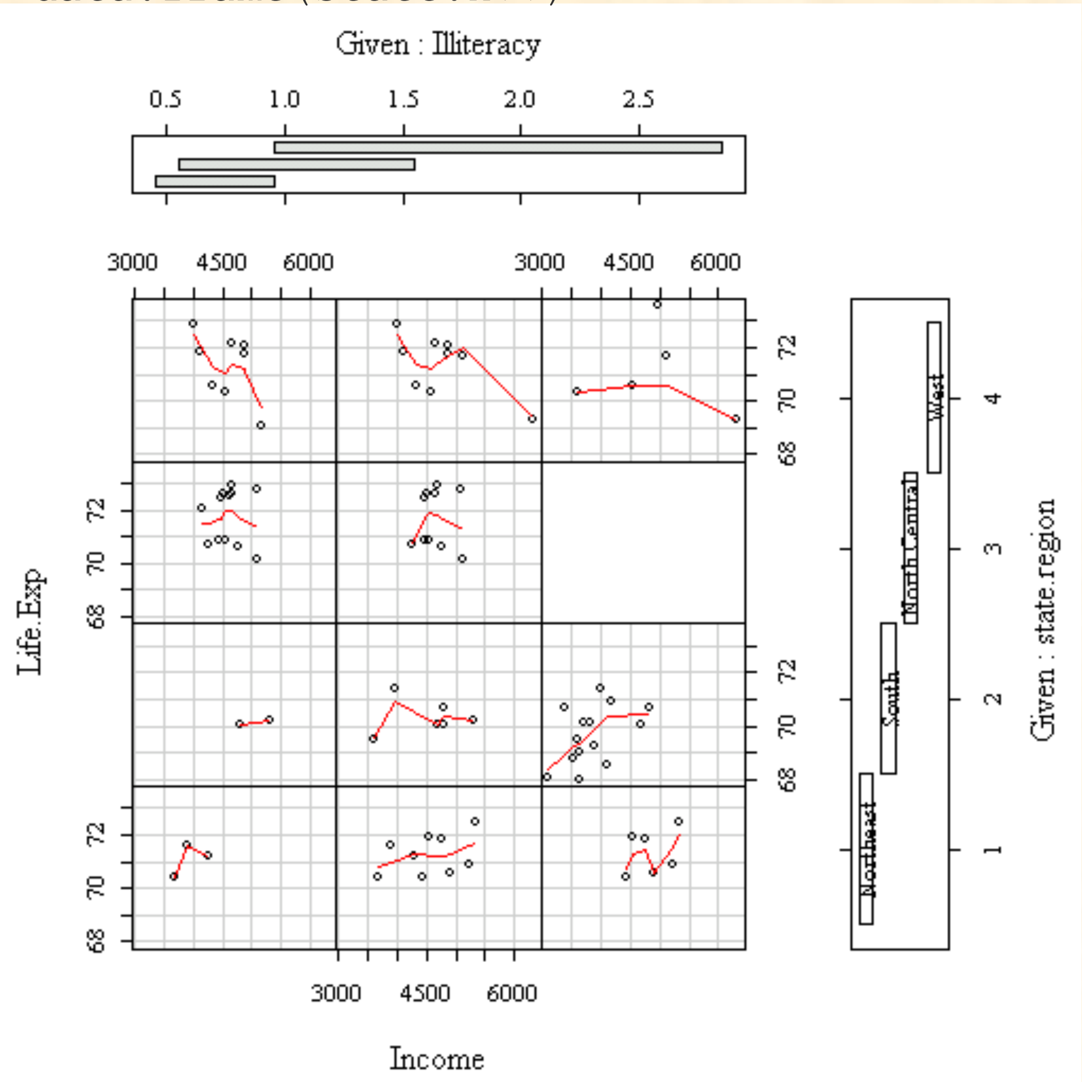
STATE	MURDE	RAPE	ROBBER	ASSAULT	BURGLARY	LARCENY	AUTO
NEBRASKA	3.90	18.10	64.70	112.70	760.00	2316.10	249.10
NEVADA	15.80	49.10	323.10	355.00	2453.10	4212.60	559.20
NEW HAMP	3.20	10.70	23.20	76.00	1041.70	2343.90	293.40
NEW JERS	5.60	21.00	180.40	185.10	1435.80	2774.50	511.50
NEW MEXI	8.80	39.10	109.60	343.40	1418.70	3008.60	259.50
NEW YORK	10.70	29.40	472.60	319.10	1728.00	2782.00	745.80
NORTH CA	10.60	17.00	61.30	318.30	1154.10	2037.80	192.10
NORTH DA	0.90	9.00	13.30	43.80	446.10	1843.00	144.70
OHIO	7.80	27.30	190.50	181.10	1216.00	2696.80	400.40
OKLAHOMA	8.60	29.20	73.80	205.00	1288.20	2228.10	326.80
OREGON	4.90	39.90	124.10	286.90	1636.40	3506.10	388.90
PENNSYLV	5.60	19.00	130.30	128.00	877.50	1624.10	333.20
RHODE IS	3.60	10.50	86.50	201.00	1489.50	2844.10	791.40
SOUTH CA	11.90	33.00	105.90	485.30	1613.60	2342.40	245.10
SOUTH DA	2.00	13.50	17.90	155.70	570.50	1704.40	147.50
TENNESSE	10.10	29.70	145.80	203.90	1259.70	1776.50	314.00
TEXAS	13.30	33.80	152.40	208.20	1603.10	2988.70	397.60
UTAH	3.50	20.30	68.80	147.30	1171.60	3004.60	334.50
VERMONT	1.40	15.90	30.80	101.20	1348.20	2201.00	265.20
VIRGINIA	9.00	23.30	92.10	165.70	986.20	2521.20	226.70
WASHINGT	4.30	39.60	106.20	224.80	1605.60	3386.90	360.30
WEST VIR	6.00	13.20	42.20	90.90	597.40	1341.70	163.30
WISCONSI	2.80	12.90	52.20	63.70	846.90	2614.20	220.70
WYOMING	5.40	21.90	39.70	173.90	811.60	2772.20	282.00

CRIME Data: Scatterplot Matrix



Conditional plots

```
(In R) data(state)
    attach(data.frame(state.x77))#> don't need `data' arg. below
    coplot(Life.Exp ~ Income | Illiteracy * state.region, number = 3,
     panel = function(x, y, ...) panel.smooth(x, y, span = .8, ...))
    detach() # data.frame(state.x77)
```



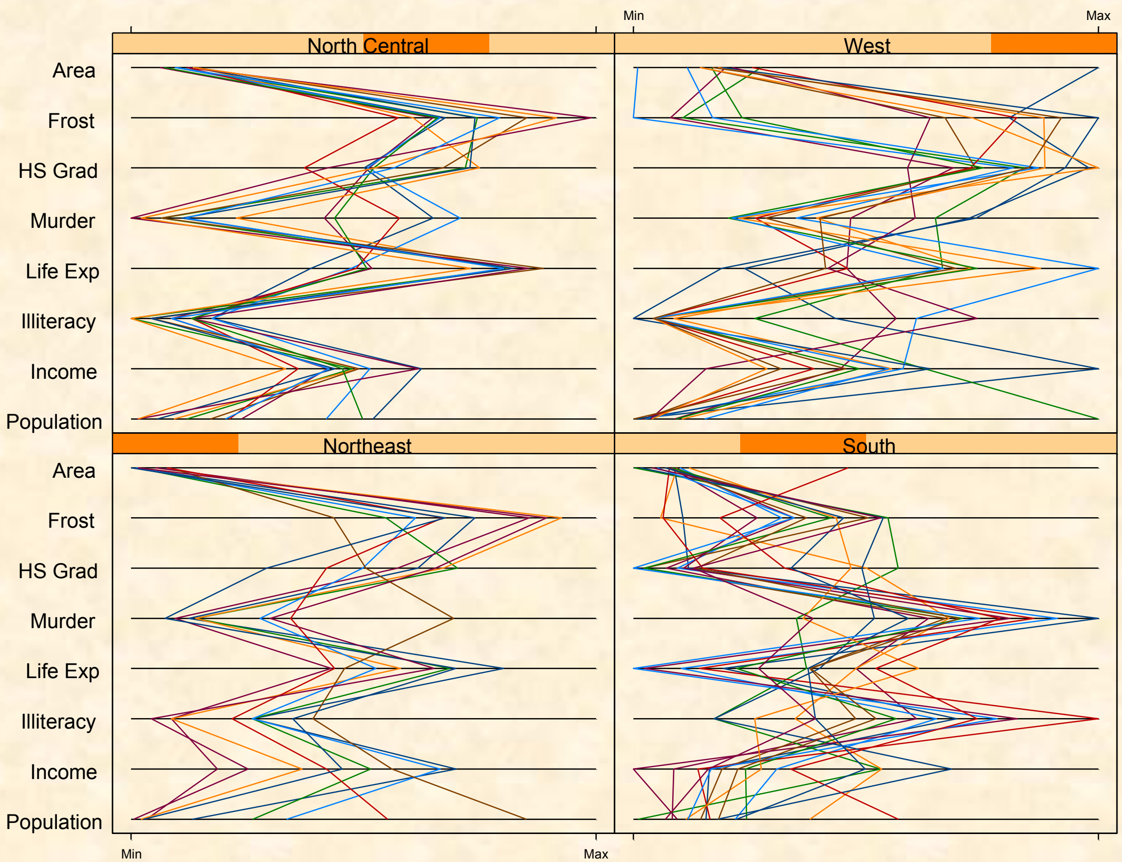
Parallel Plot:

Graph of a multivariate dataset where the observations are represented by lines.

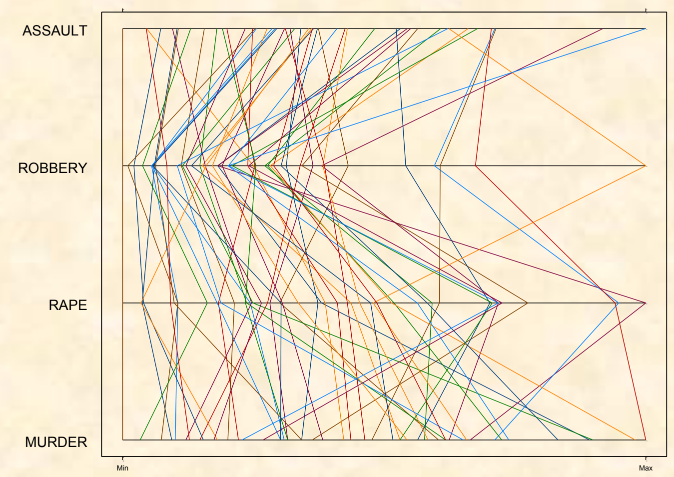
Objectives:

1. To visualize comparisons between multivariate data groups.
2. Help asses the quality of classification tools
3. To find data clusters and outliers.

```
parallel( ~ state.x77 | state.region )
```



Using the Crime dataset:
parallel(~X[,1:4])



DIMENSION REDUCTION: (PRINCIPAL COMPONENTS)

Principal components analysis is a method for dimension reduction.

Applications:

- Data Mining: Reducing the number of variables.
- Regression Analysis: The number of predictors q is comparable to the error df's ν_E. We need $q \ll \nu_E$.
- MANOVA: The number of responses p is comparable to the error df's ν_E. We need $p \ll \nu_E$.

Data: $\mathbf{y}_i=(y_{i1},\ldots, y_{ip})$ $i=1,..,n$, we assume that the $\{\mathbf{y}_i\}$ are centered.

Let A be an orthogonal transformation such that the $\mathbf{z}_i = A\mathbf{y}_i$ are uncorrelated.

Dimension Reduction

Examples:

1. DNA MICROARRAYS:

Khan *et al* (2001): 4 types of small round blue cell tumors (SRBCT)
Neuroblastoma (NB) Rhabdomyosarcoma (RMS)
Ewing family of tumors (EWS) Burkitt lymphomas (BL)

Arrays: Training set= 63 arrays(23 EWS, 20 RMS, 12 NB, 8 BL)
Testing set= 25 arrays(6 EWS, 5 RMS, 6 NB, 3 BL, 5 other)

Genes: 2308 genes were selected because they showed minimal expression levels.

2. PLASTIC EXPLOSIVES: The data comes from a study for the detection of plastic explosives in suitcases using X-ray signals. The 23 variables are the discrete x-components of the xray absorption spectrum. The objective is to detect the suitcases with explosives. 2993 suitcases were use for training and 60 testing. (see web page for dataset).

Covariance Vs Correlation Matrix

1. Use covariance or correlation matrix? If variables are not in the same units $\Rightarrow$ Use Correlations

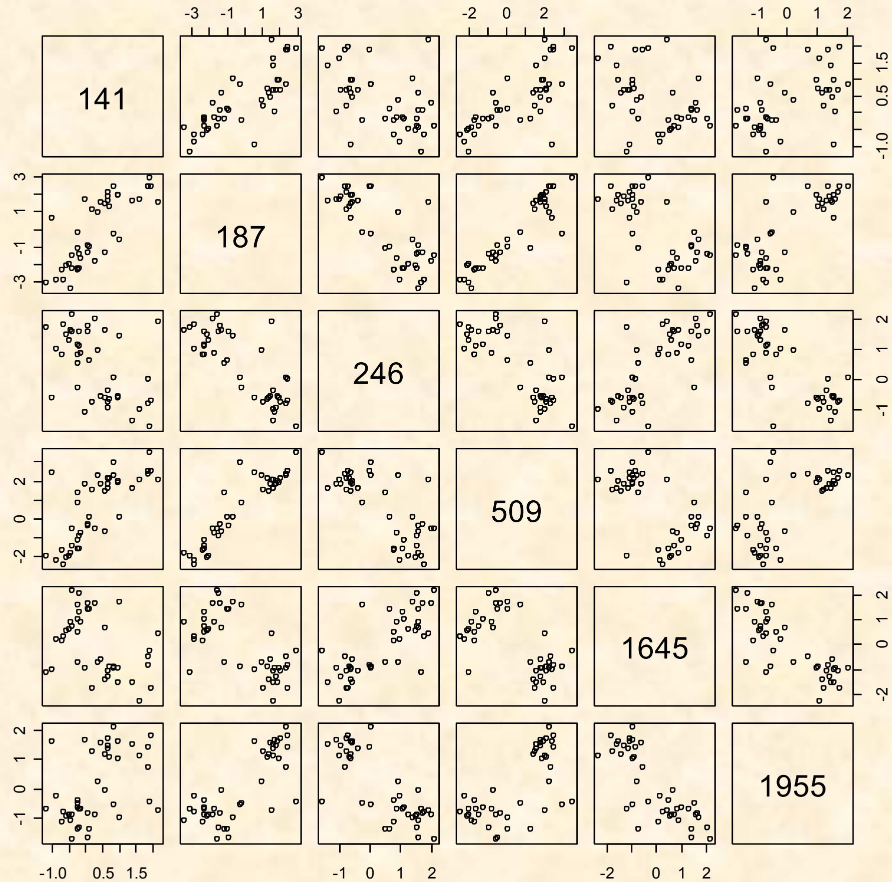
$$S = \begin{pmatrix} s_1^2, s_{12}, \ldots\ s_{1p} \\ s_{21}, s_2^2, \ldots\ , s_{2p} \\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots \\ s_{p1}, s_{p2}, \ldots\ , s_p^2 \end{pmatrix} R = \begin{pmatrix} 1, r_{12}, \ldots\ , r_{1p} \\ r_{21}, 1, \ldots\ , r_{2p} \\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots \\ r_{p1}, r_{p2}, \ldots\ , 1 \end{pmatrix}; r_{ij} = \frac{s_{ij}}{s_i s_j}$$

2. Dim(V) =Dim(R) = pxp and if p is large $\Rightarrow$ Dimension reduction.

Sample Correlation Matrix

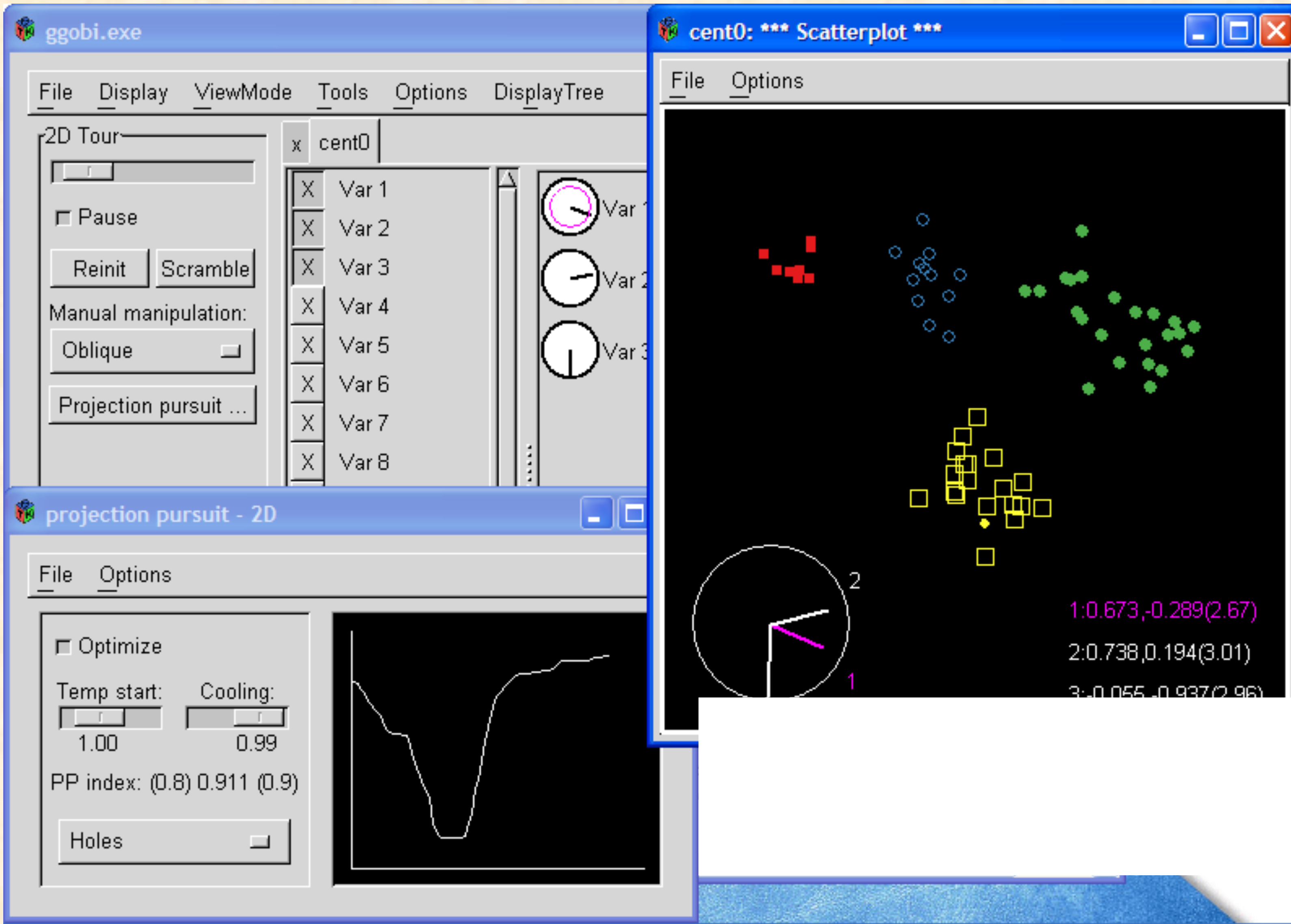
	Gene 141	Gene 187	Gene 246	Gene 509	Gene 1645	Gene 1955
Gene 141	1.0000	0.7983 (0.000)	-0.5058 (0.001)	0.7463 (0.000)	-0.4049 (0.007)	0.4676 (0.002)
Gene 187	0.7983 (0.000)	1.0000	-0.8111 (0.000)	0.9357 (0.000	-0.6621 (0.000)	0.7891 (0.000)
Gene 246	-0.5058 (0.001)	-0.8111 (0.000)	1.0000	-0.7717 (0.000)	0.7624 (0.000)	-0.7977 (0.000)
Gene 509	0.7463 (0.000)	0.9357 (0.000)	-0.7717 (0.000)	1.000	-0.6388 (0.000)	0.6827 (0.000)
Gene 1645	-0.4049 (0.007)	-0.6621 (0.000)	0.7624 (0.000)	-0.6388 (0.000)	1.000	-0.8143 (0.000)
Gene 1955	0.4676 (0.002)	0.7891 (0.000)	-0.7977 (0.000)	0.6827 (0.000)	-0.8143 (0.000)	1.000

Scatterplot Matrix



Data visualization of Multivariate Data

Ggobi display finding four clusters of tumors using the PP index on the set of 63 cases. The main panel shows the two dimensional projection selected by the PP index with the four clusters in different colors and glyphs. The top left panel shows the main controls and the left bottom panel displays the controls and the graph of the PP index that is been optimized. The graph shows the index value for a sequence of projection ending at the current one.



H_2O software

Environment for fast for machine learning implementations

Go to web: http://www.h2o.ai/download/h2o/r

Install H_2O

Install R packages so you use it from R

EXAMPLE: K-MEANS CLUSTERING

```
###  demo  FOR h2o.kmeans
library(h2o)
h2o.init()
prostate.hex = h2o.uploadFile(path=system.file("extdata”,"prostate.csv”,package="h2o"),
                  destination_frame="prostate")
summary(prostate.hex)
prostate.km = h2o.kmeans(prostate.hex,k=10,x=c("AGE","RACE","GLEASON","CAPSULE","DCAPS"))
print(prostate.km)
prostate.data = as.data.frame(prostate.hex)
par(mfrow=c(1,2))
prostate.ctrs=as.data.frame(prostate.km@model$centers)
plot(prostate.ctrs[,1:2])
plot(prostate.ctrs[,3:4])
title("K-Means Centers for k = 10”,outer=TRUE,line=-2.0)
```