

*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

# *Multivariate Data*

## Multivariate Data.

Most datasets contain multiple variables.

Variables maybe correlated.

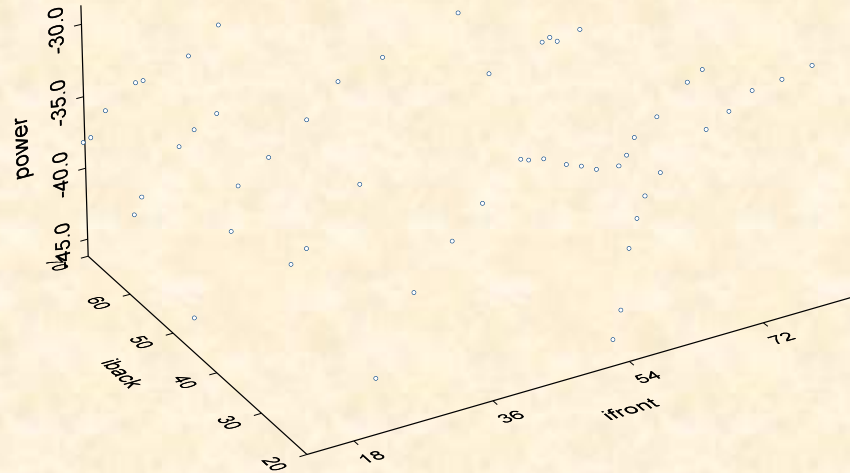
Objectives are:

1. Explore, Summarize , reduce dimensionality
2. Find interesting patterns, clusters, outliers.
3. Find classification rule that assign each observation to a class.

# 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:** We saw many examples of this already

**Conditional plots**

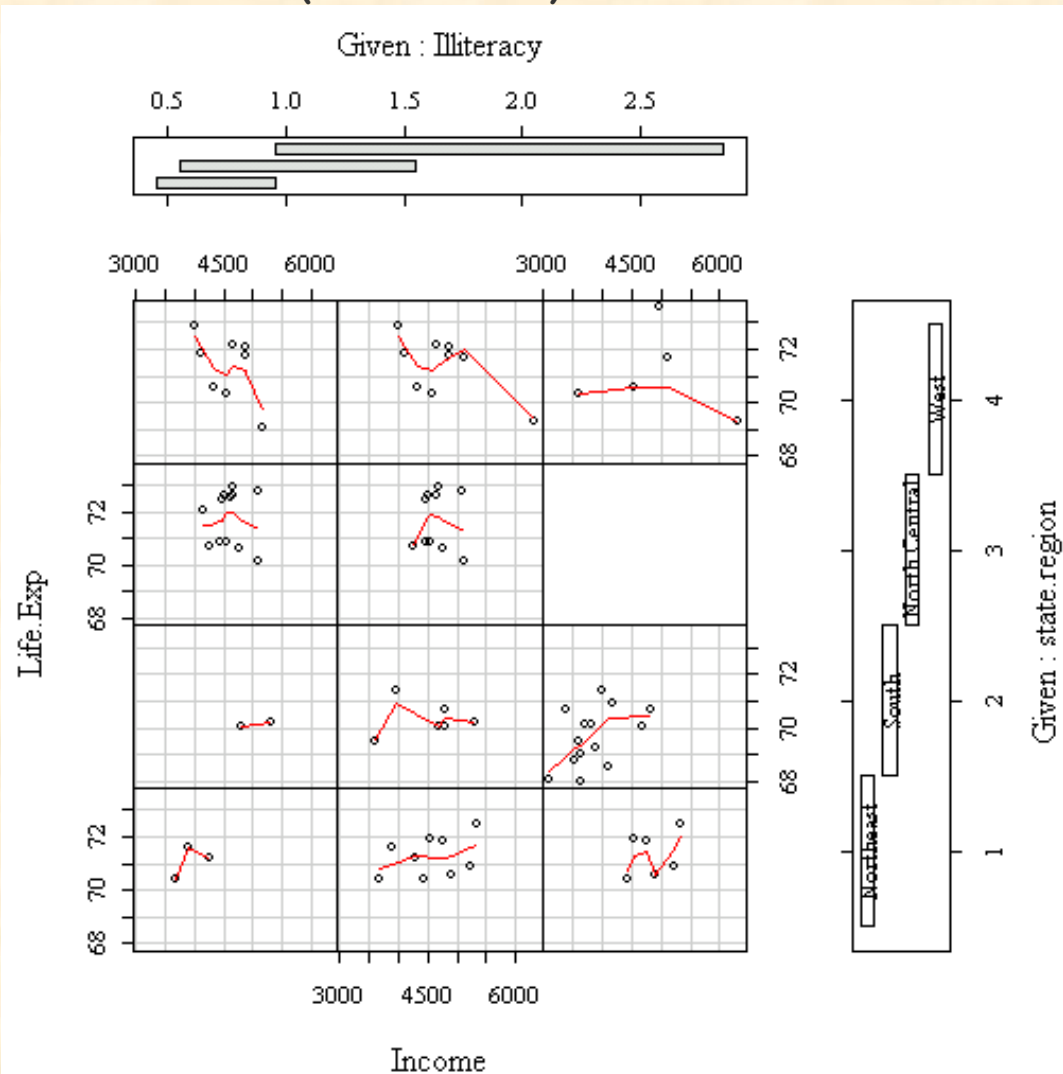
(In R) `data(state)`

```
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)
```

# 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)
```



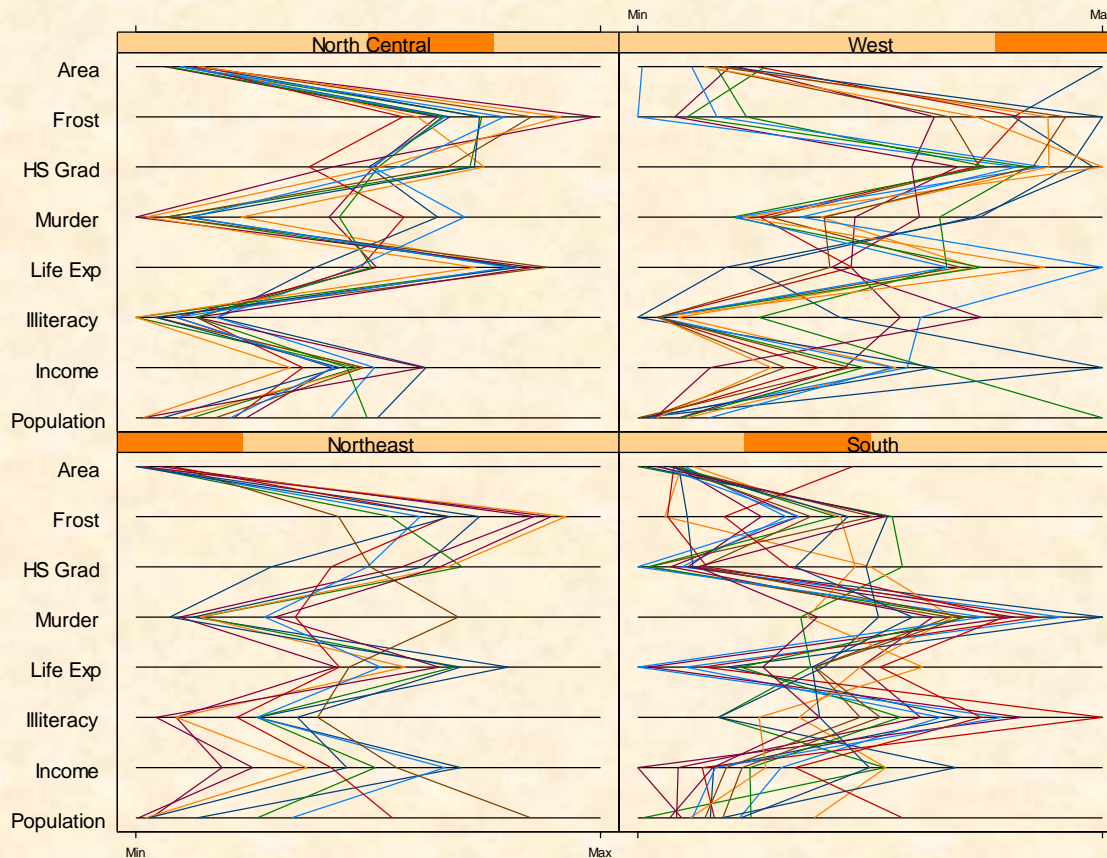
## Parallel Plot:

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

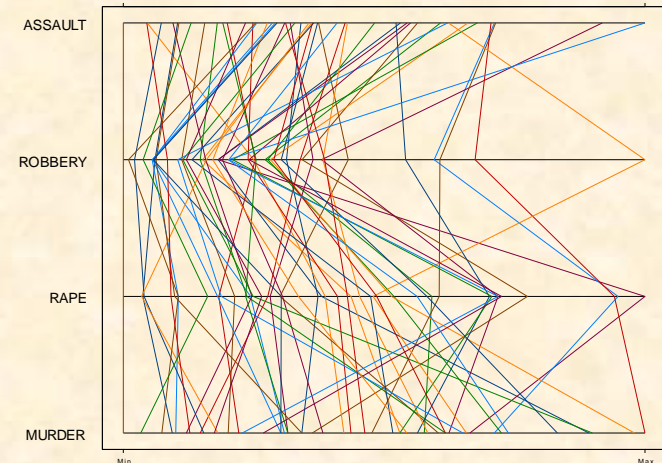
Objectives:

1. To visualize comparisons between multivariate data groups.
2. Help assess 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  $\nu_E$ . We need  $q \ll \nu_E$ .
- MANOVA: The number of responses  $p$  is comparable to the error df's  $\nu_E$ . We need  $p \ll \nu_E$ .

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

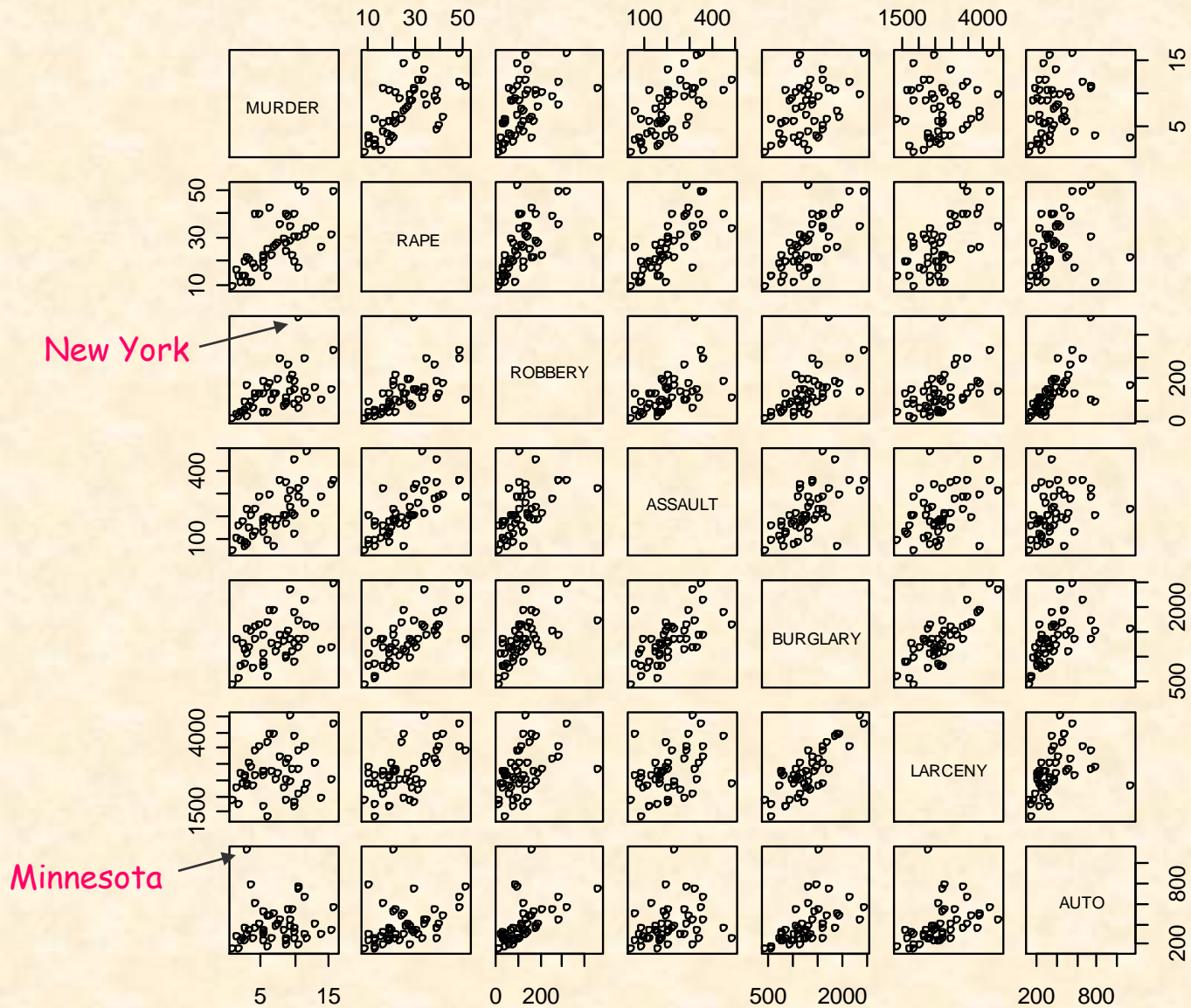
Let  $A$  be an orthogonal transformation such that the  $\mathbf{z}_i = A\mathbf{y}_i$  are uncorrelated.

# 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
CALIFORNIA	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
CONNECTICUT	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
LOUISIANA	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
MASSACHUSETTS	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
MINNESOTA	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

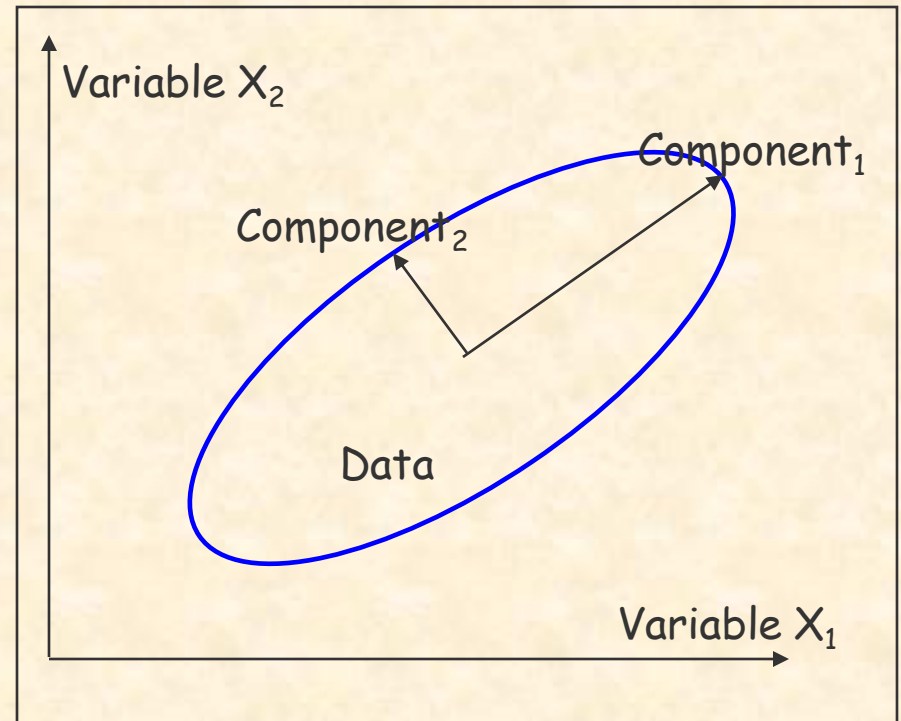
STATE	MURDE	RAPE	ROBBER	ASSAULT	BURGLARY	LARCENY	AUTO
NEBRASKA	3.90	18.10	64.70	112.70	760.00	2316.10	249.10
NEVADA	15.80	49.10	323.10	355.00	2453.10	4212.60	559.20
NEW HAMPSHIRE	3.20	10.70	23.20	76.00	1041.70	2343.90	293.40
NEW JERSEY	5.60	21.00	180.40	185.10	1435.80	2774.50	511.50
NEW MEXICO	8.80	39.10	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 CAROLINA	10.60	17.00	61.30	318.30	1154.10	2037.80	192.10
NORTH DAKOTA	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
PENNSYLVANIA	5.60	19.00	130.30	128.00	877.50	1624.10	333.20
RHODE ISLAND	3.60	10.50	86.50	201.00	1489.50	2844.10	791.40
SOUTH CAROLINA	11.90	33.00	105.90	485.30	1613.60	2342.40	245.10
SOUTH DAKOTA	2.00	13.50	17.90	155.70	570.50	1704.40	147.50
TENNESSEE	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
WASHINGTON	4.30	39.60	106.20	224.80	1605.60	3386.90	360.30
WEST VIRGINIA	6.00	13.20	42.20	90.90	597.40	1341.70	163.30
WISCONSIN	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



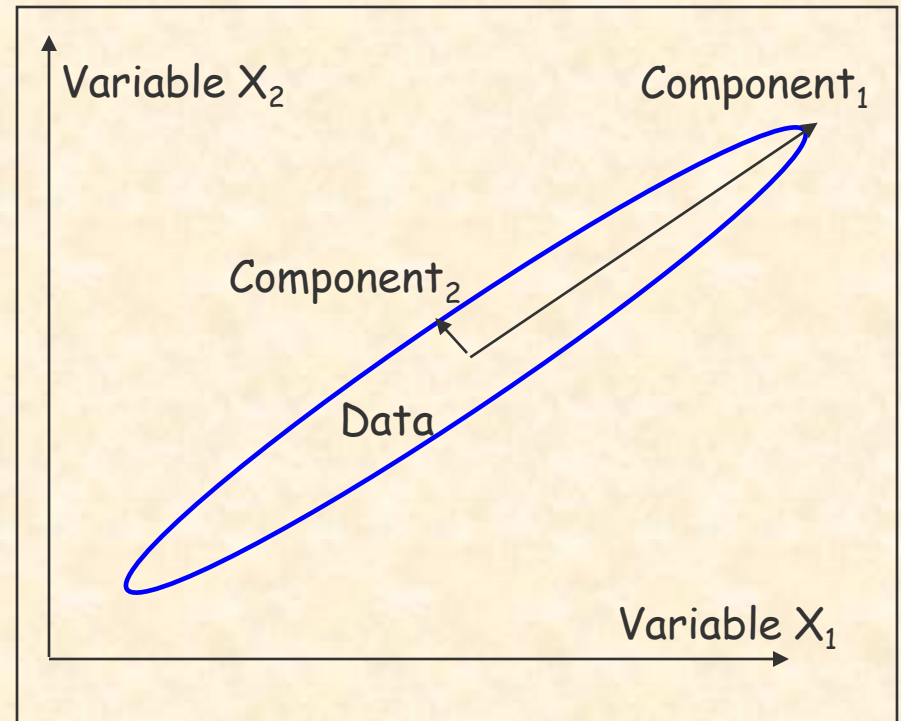
# Geometrical Intuition

- *The data cloud is approximated by an ellipsoid*
- *The axes of the ellipsoid represent the natural components of the data*
- *The length of the semi-axis represent the variability of the component.*



# DIMENSION REDUCTION

- *When some of the components show a very small variability they can be omitted.*
- *The graphs shows that Component 2 has low variability so it can be removed.*
- *The dimension is reduced from  $\text{dim}=2$  to  $\text{dim}=1$*



# Covariance and Correlation Matrices

1. The Variance covariance matrix estimates the shape of the ellipsoid that approximates the data.

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

2. Use covariance or correlation matrix? If variables are not in the same units  $\Rightarrow$  Use Correlation
3.  $\text{Dim}(V) = \text{Dim}(R) = p \times p$  and if  $p$  is large  $\Rightarrow$  Dimension reduction.

## PRINCIPAL COMPONENTS TABLE

### Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
MURDER	0.329	0.588	0.190	-0.217	0.521	-0.377	0.223
RAPE	0.429	0.182	-0.221		0.299	0.746	-0.285
ROBBERY	0.392		0.489	-0.590	-0.467	0.190	
ASSAULT	0.395	0.355		0.606	-0.543		0.217
BURGLARY	0.435	-0.219	-0.228			-0.505	-0.673
LARCENY	0.355	-0.380	-0.572	-0.227			0.589
AUTO	0.287	-0.546	0.543	0.424	0.352		0.145

### Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	2.0436891	1.0763811	0.8621946	0.5664485	0.50353374
Proportion of Variance	0.5966664	0.1655138	0.1061971	0.0458377	0.03622089
Cumulative Proportion	0.5966664	0.7621802	0.8683773	0.9142150	0.95043587

### *Analysis:*

***Dimension Reduction: 2 components explain 76.2% of variability***

***First component: represents the sum or average of all crimes because the loadings are very similar .***

***PC1 = violent crimes + non-violent crimes***

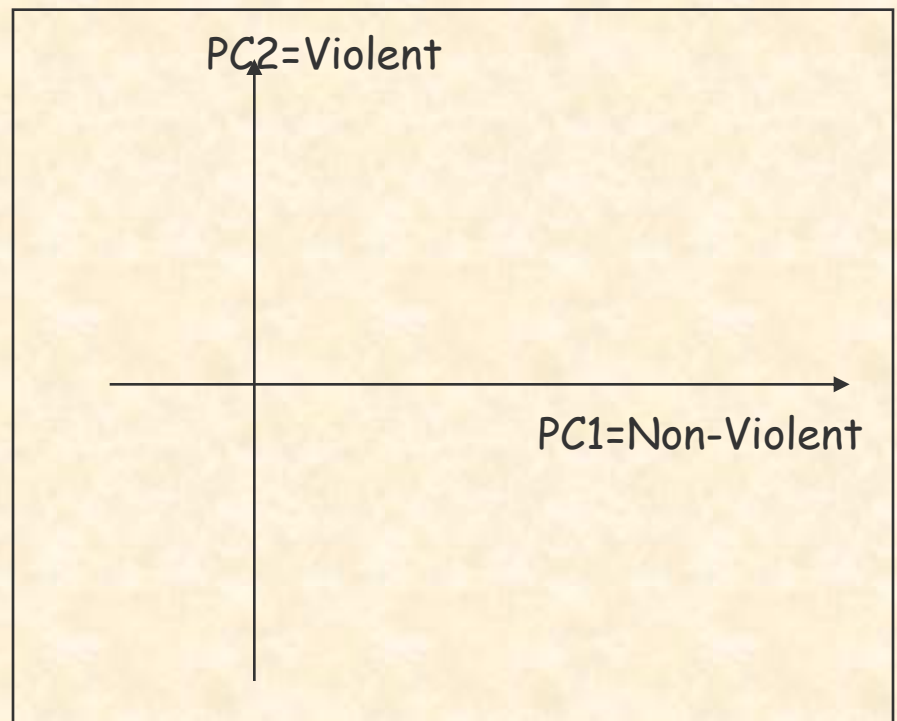
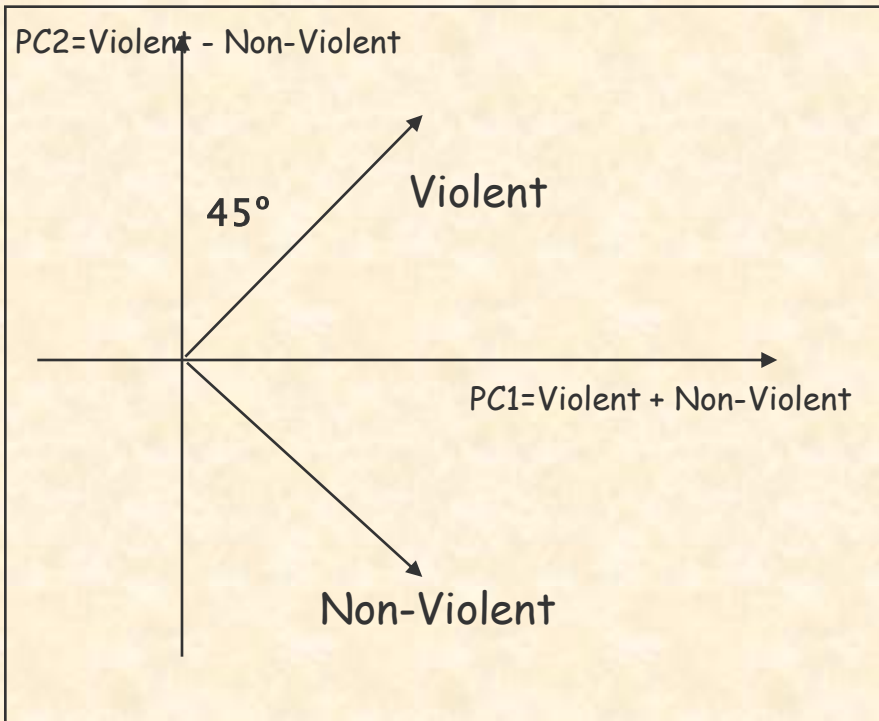
***Second component:***

***Violent crimes: MURDER RAPE ROBBERY ASSAULT  
all have positive coefficients.***

***Non-violent crimes: BURGLARY LARCENY AUTO  
all have negative coefficients.***

***PC2 = violent crimes - non-violent crimes***

# Geometrical Intuition



PC1= Violent + NonViolent  
PC2= Violent - NonViolent

45° rotation →

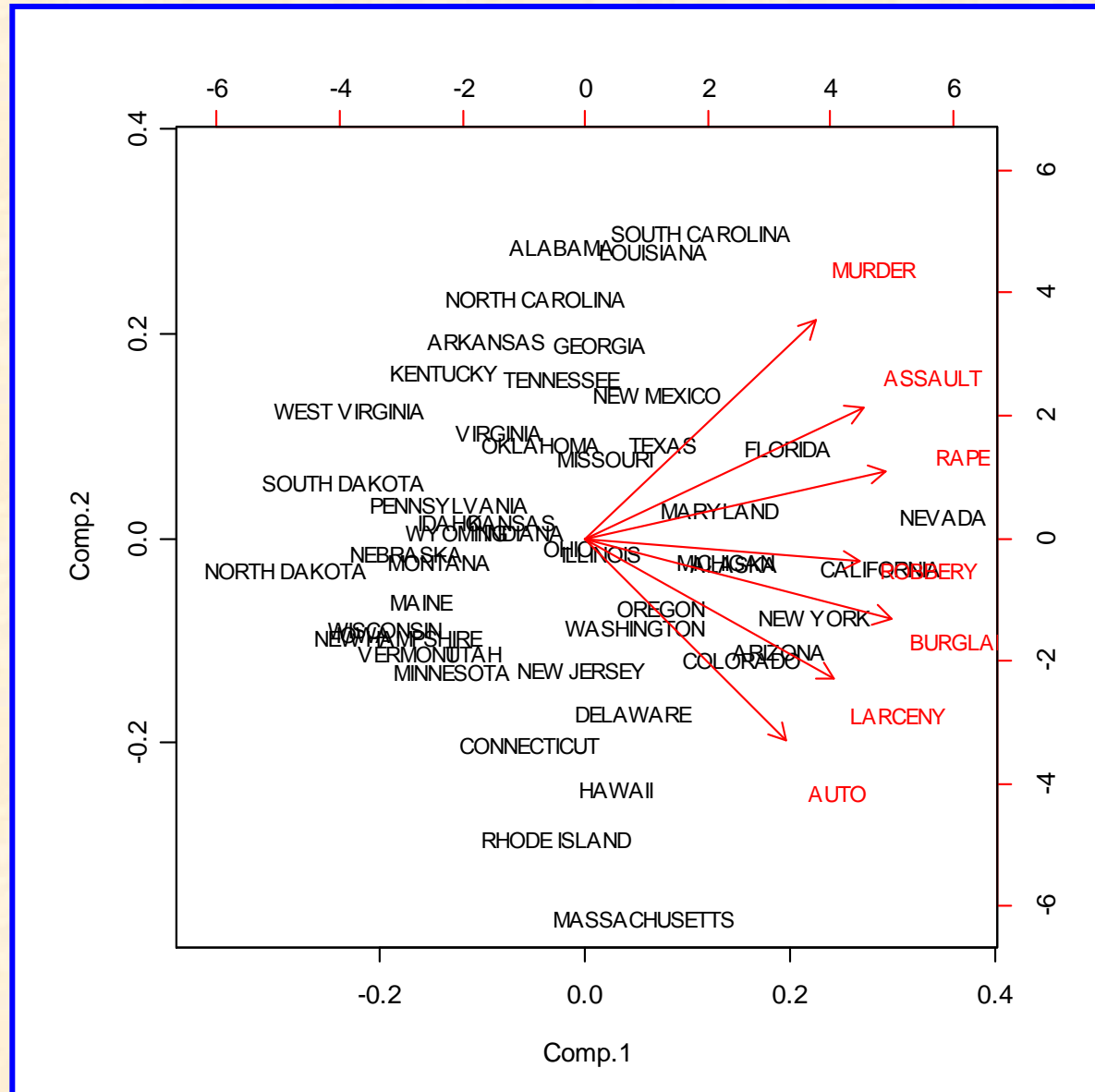
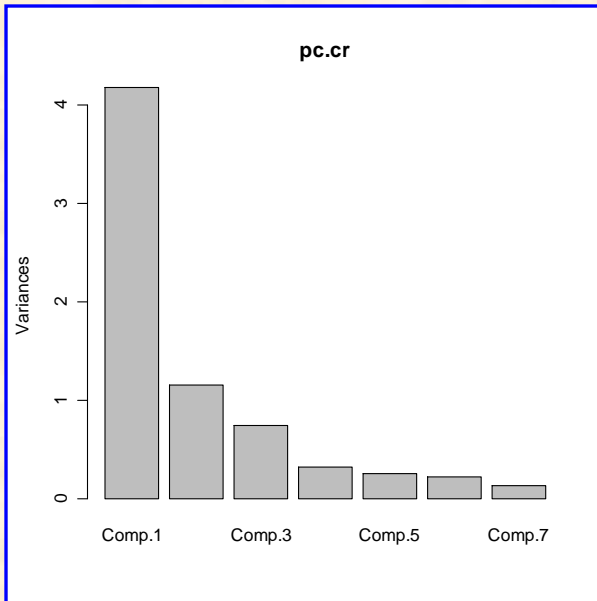
PC1= NonViolent  
PC2= Violent

# *Biplot*

*Combination of two graphs into one:*

- 1. Graph of the observations in the coordinates of the two principal components.*
- 2. Graph of the Variables projected into the plane of the two principal components.*
- 3. The variables are represented as arrows, the observations as points or labels.*

# Variations and Biplot





# IN SAS

```
options ls=64 ps=50;
DATA CRIME;
  TITLE 'CRIME RATES PER 100,000 POPULATION BY STATE';
  INPUT STATE $1-15 MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY AUTO;
  CARDS;
```

ALABAMA	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
ALASKA	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3
ARIZONA	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
ARKANSAS	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
CALIFORNIA	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5
COLORADO	6.3	42.0	170.7	292.9	1935.2	3903.2	477.1
CONNECTICUT	4.2	16.8	129.5	131.8	1346.0	2620.7	593.2
DELAWARE	6.0	24.9	157.0	194.2	1682.6	3678.4	467.0
FLORIDA	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
GEORGIA	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9
HAWAII	7.2	25.5	128.0	64.1	1911.5	3920.4	489.4
IDAHO	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6
ILLINOIS	9.9	21.8	211.3	209.0	1085.0	2828.5	528.6
INDIANA	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4
IOWA	2.3	10.6	41.2	89.8	812.5	2685.1	219.9
KANSAS	6.6	22.0	100.7	180.5	1270.4	2739.3	244.3
KENTUCKY	10.1	19.1	81.1	123.3	872.2	1662.1	245.4
LOUISIANA	15.5	30.9	142.9	335.5	1165.5	2469.9	337.7
MAINE	2.4	13.5	38.7	170.0	1253.1	2350.7	246.9
MARYLAND	8.0	34.8	292.1	358.9	1400.0	3177.7	428.5
MASSACHUSETTS	3.1	20.8	169.1	231.6	1532.2	2311.3	1140.1
MICHIGAN	9.3	38.9	261.9	274.6	1522.7	3159.0	545.5
MINNESOTA	2.7	19.5	85.9	85.8	1134.7	2559.3	343.1
MISSISSIPPI	14.3	19.6	65.7	189.1	915.6	1239.9	144.4
MISSOURI	9.6	28.3	189.0	233.5	1318.3	2424.2	378.4
MONTANA	5.4	16.7	39.2	156.8	804.9	2773.2	309.2
NEBRASKA	3.9	18.1	64.7	112.7	760.0	2316.1	249.1
NEVADA	15.8	49.1	323.1	355.0	2453.1	4212.6	559.2
NEW HAMPSHIRE	3.2	10.7	23.2	76.0	1041.7	2343.9	293.4
NEW JERSEY	5.6	21.0	180.4	185.1	1435.8	2774.5	511.5
NEW MEXICO	8.8	39.1	109.6	343.4	1418.7	3008.6	259.5
NEW YORK	10.7	29.4	472.6	319.1	1728.0	2782.0	745.8
NORTH CAROLINA	10.6	17.0	61.3	318.3	1154.1	2037.8	192.1
NORTH DAKOTA	0.9	9.0	13.3	43.8	446.1	1843.0	144.7
OHIO	7.8	27.3	190.5	181.1	1216.0	2696.8	400.4
OKLAHOMA	8.6	29.2	73.8	205.0	1288.2	2228.1	326.8
OREGON	4.9	39.9	124.1	286.9	1636.4	3506.1	388.9
PENNSYLVANIA	5.6	19.0	130.3	128.0	877.5	1624.1	333.2
RHODE ISLAND	3.6	10.5	86.5	201.0	1489.5	2844.1	791.4
SOUTH CAROLINA	11.9	33.0	105.9	485.3	1613.6	2342.4	245.1
SOUTH DAKOTA	2.0	13.5	17.9	155.7	570.5	1704.4	147.5
TENNESSEE	10.1	29.7	145.8	203.9	1259.7	1776.5	314.0
TEXAS	13.3	33.8	152.4	208.2	1603.1	2988.7	397.6
UTAH	3.5	20.3	68.8	147.3	1171.6	3004.6	334.5
VERMONT	1.4	15.9	30.8	101.2	1348.2	2201.0	265.2
VIRGINIA	9.0	23.3	92.1	165.7	986.2	2521.2	226.7
WASHINGTON	4.3	39.6	106.2	224.8	1605.6	3386.9	360.3
WEST VIRGINIA	6.0	13.2	42.2	90.9	597.4	1341.7	163.3
WISCONSIN	2.8	12.9	52.2	63.7	846.9	2614.2	220.7
WYOMING	5.4	21.9	39.7	173.9	811.6	2772.2	282.0

```
PROC PRINCOMP OUT=CRIMCOMP;
```

```
PROC SORT;
```

```
  BY PRIN1;
```

```
PROC PRINT;
```

```
  ID STATE;
```

```
  VAR PRIN1 PRIN2 MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY AUTO;
```

```
  TITLE2 'STATES LISTED IN ORDER OF OVERALL CRIME RATE';
```

```
  TITLE3 'AS DETERMINED BY THE FIRST PRINCIPAL COMPONENT';
```

```
PROC SORT;
```

```
  BY PRIN2;
```

```
PROC PRINT;
```

```
  ID STATE;
```

```
  VAR PRIN1 PRIN2 MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY AUTO;
```

```
  TITLE2 'STATES LISTED IN ORDER OF PROPERTY VS. VIOLENT CRIME';
```

```
  TITLE3 'AS DETERMINED BY THE SECOND PRINCIPAL COMPONENT';
```

```
PROC PLOT;
```

```
  PLOT PRIN2*PRIN1=STATE;
```

```
  TITLE2 'PLOT OF THE FIRST TWO PRINCIPAL COMPONENTS';
```

```
PROC PLOT;
```

```
  PLOT PRIN3*PRIN1=STATE;
```

**CRIME RATES PER 100,000 POPULATION BY STATE**

**Principal Component Analysis**

50 Observations      7 Variables      Simple Statistics

	MURDER	RAPE	ROBBERY	ASSAULT
Mean	7.444000000	25.73400000	124.0920000	211.3000000
Std	3.866768941	10.75962995	88.3485672	100.2530492

	BURGLARY	LARCENY	AUTO
Mean	1291.904000	2671.288000	377.5260000
Std	432.455711	725.908707	193.3944175

**Correlation Matrix**

	MURDER	RAPE	ROBBERY	ASSAULT
MURDER	1.0000	0.6012	0.4837	0.6486
RAPE	0.6012	1.0000	0.5919	0.7403
ROBBERY	0.4837	0.5919	1.0000	0.5571
ASSAULT	0.6486	0.7403	0.5571	1.0000

	BURGLARY	LARCENY	AUTO
BURGLARY	0.3858	0.7121	0.6229
LARCENY	0.1019	0.6140	0.4044
AUTO	0.0688	0.3489	0.2758

	BURGLARY	LARCENY	AUTO
MURDER	0.3858	0.1019	0.0688
RAPE	0.7121	0.6140	0.3489
ROBBERY	0.6372	0.4467	0.5907
ASSAULT	0.6229	0.4044	0.2758
BURGLARY	1.0000	0.7921	0.5580
LARCENY	0.7921	1.0000	0.4442
AUTO	0.5580	0.4442	1.0000

**Eigenvalues of the Correlation Matrix**

	Eigenvalue	Differen	Proportion	Cumulative
PRIN1	4.11496	2.87624	0.587851	0.58785
PRIN2	1.23872	0.51291	0.176960	0.76481
PRIN3	0.72582	0.40938	0.103688	0.86850
PRIN4	0.31643	0.05846	0.045205	0.91370
PRIN5	0.25797	0.03593	0.036853	0.95056
PRIN6	0.22204	0.09798	0.031720	0.98228
PRIN7	0.12406	.	0.017722	1.00000

**Eigenvectors**

	PRIN1	PRIN2	PRIN3	PRIN4
MURDER	0.300279	-.629174	0.178245	-.232114
RAPE	0.431759	-.169435	-.244198	0.062216
ROBBERY	0.396875	0.042247	0.495861	-.557989
ASSAULT	0.396652	-.343528	-.069510	0.629804
BURGLARY	0.440157	0.203341	-.209895	-.057555
LARCENY	0.357360	0.402319	-.539231	-.234890
AUTO	0.295177	0.502421	0.568384	0.419238

	PRIN5	PRIN6	PRIN7
MURDER	0.538123	0.259117	0.267593
RAPE	0.188471	-.773271	-.296485
ROBBERY	-.519977	-.114385	-.003903
ASSAULT	-.506651	0.172363	0.191745
BURGLARY	0.101033	0.535987	-.648117
LARCENY	0.030099	0.039406	0.601690
AUTO	0.369753	-.057298	0.147046

How many components?

- Explain some fix % of the variance (70%, 80%...)
- Exclude eigenvalues less than the average.  
(For the correlation matrix the average is 1)
- Graph of eigenvalues (In R)

Test the null hypothesis that the last k eigenvalues are equal

Let 
$$\bar{\lambda} = \sum_{i=p-k+1}^p \frac{\lambda_i}{k}$$

$$u = (n - (2p+11)/6)(k + \log \bar{\lambda} - \sum_{i=p-k+1}^p \log \lambda_i)$$

The test statistic is

The test statistic u is approximately  $\chi^2$  with  $df = (k-1)(k+2)/2$ .

In the example dataset: The last four eigenvalues are small

$$> (50 - (2*7+11)/6)*(4*\log(\text{mei}) - \text{sum}(\log(\text{ei})))$$

[1] 10.12649

> qchisq(0.95,9)

[1] 16.91898

Now with the last 5 eigenvalues:

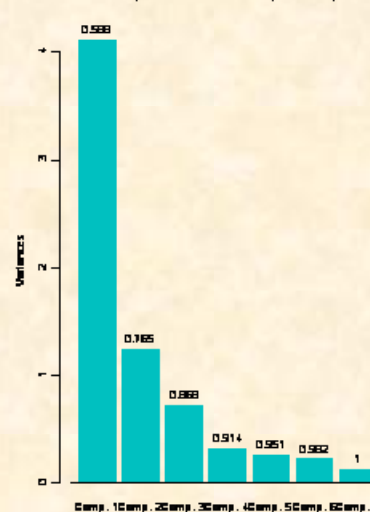
$$> (50 - (2*7+11)/6)*(5*\log(\text{mei}) - \text{sum}(\log(\text{ei})))$$

$$> \text{qchisq}(0.95,14)$$

[1] 39.57434

[1] 23.68475

Relative Importance of Principal Components



# *Cluster Analysis:*

Group the samples into  $k$  distinct natural groups.

Hierarchical clustering: Build a hierarchical tree

Inter point distance is normally the Euclidean distance (some times we may use Manhattan distance).

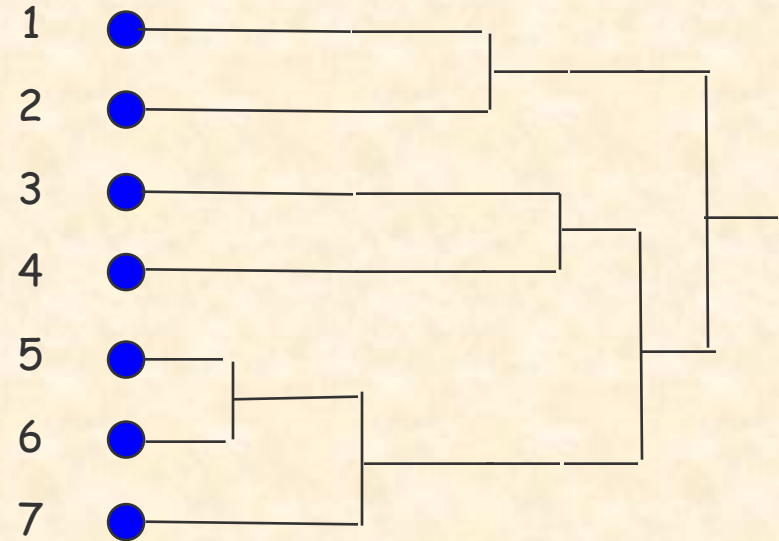
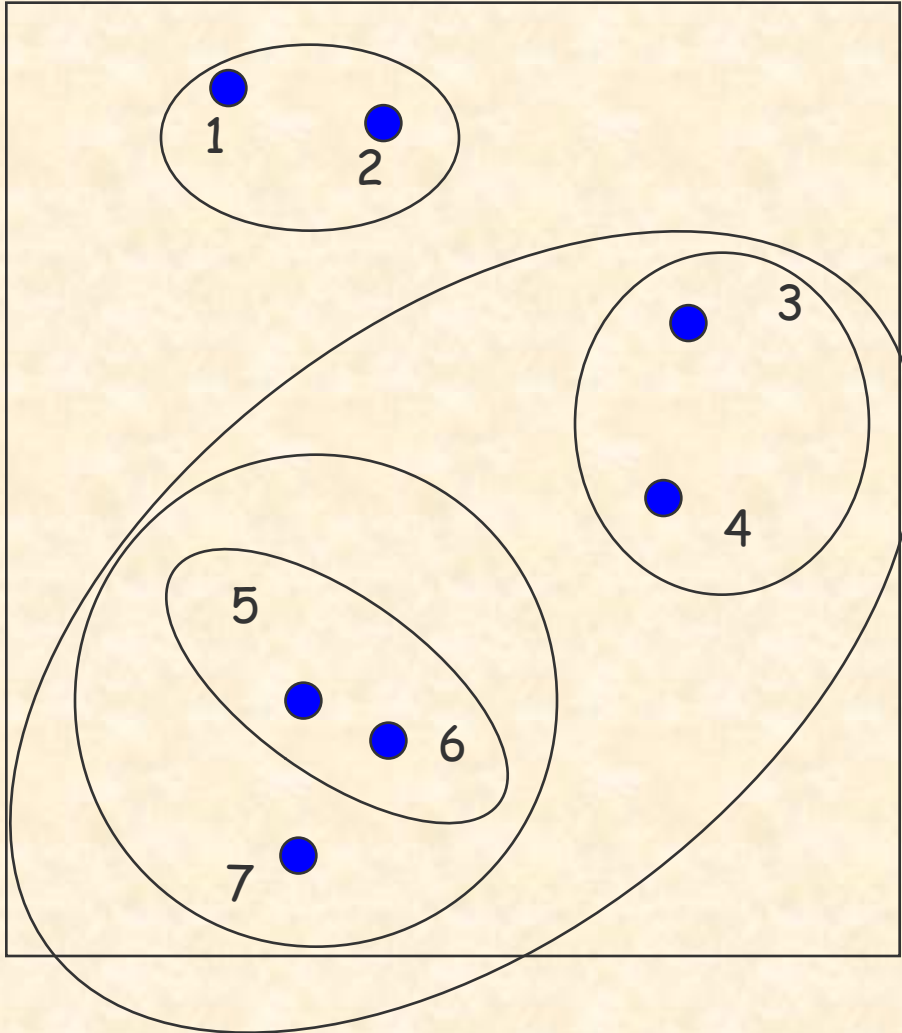
Inter cluster distance:

- Single Linkage: distance between the closes two points
- Complete Linkage: distance between the furthest two points
- Average Linkage: Average distance between every pair of points
- Ward:  $R^2$  change.

Build a hierarchical tree:

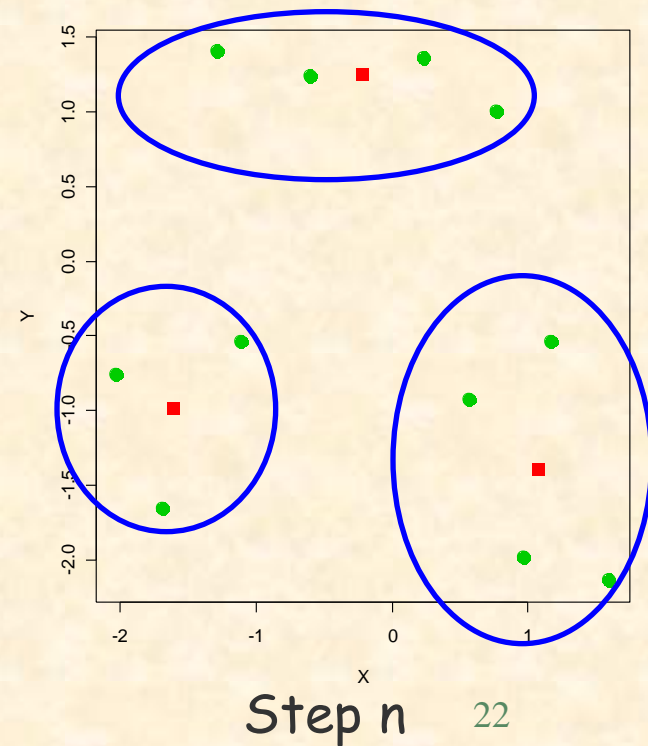
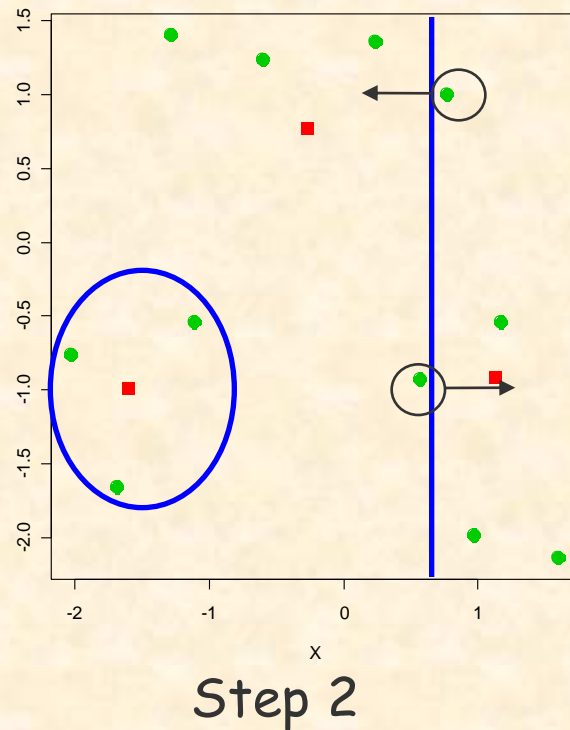
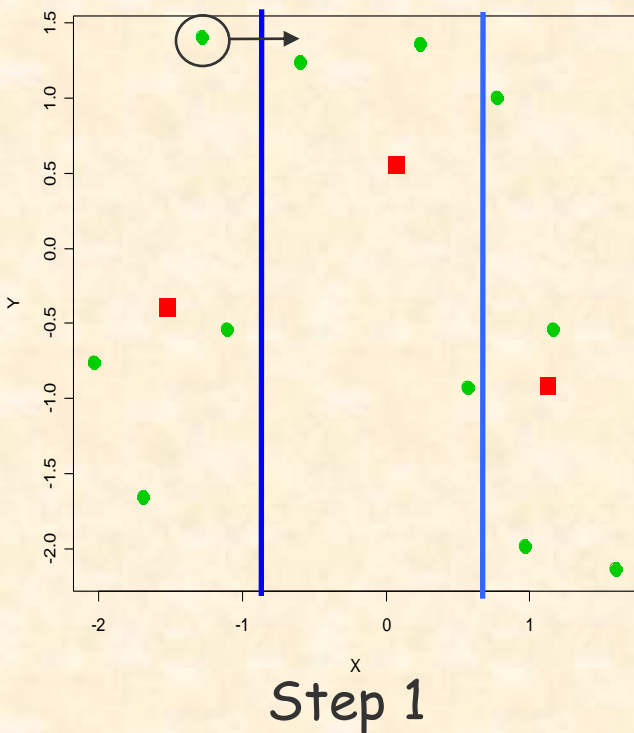
1. Start with a cluster at each sample point
2. At each stage of building the tree the two closest clusters joint to form a new cluster.

# Hierarchical Cluster Example

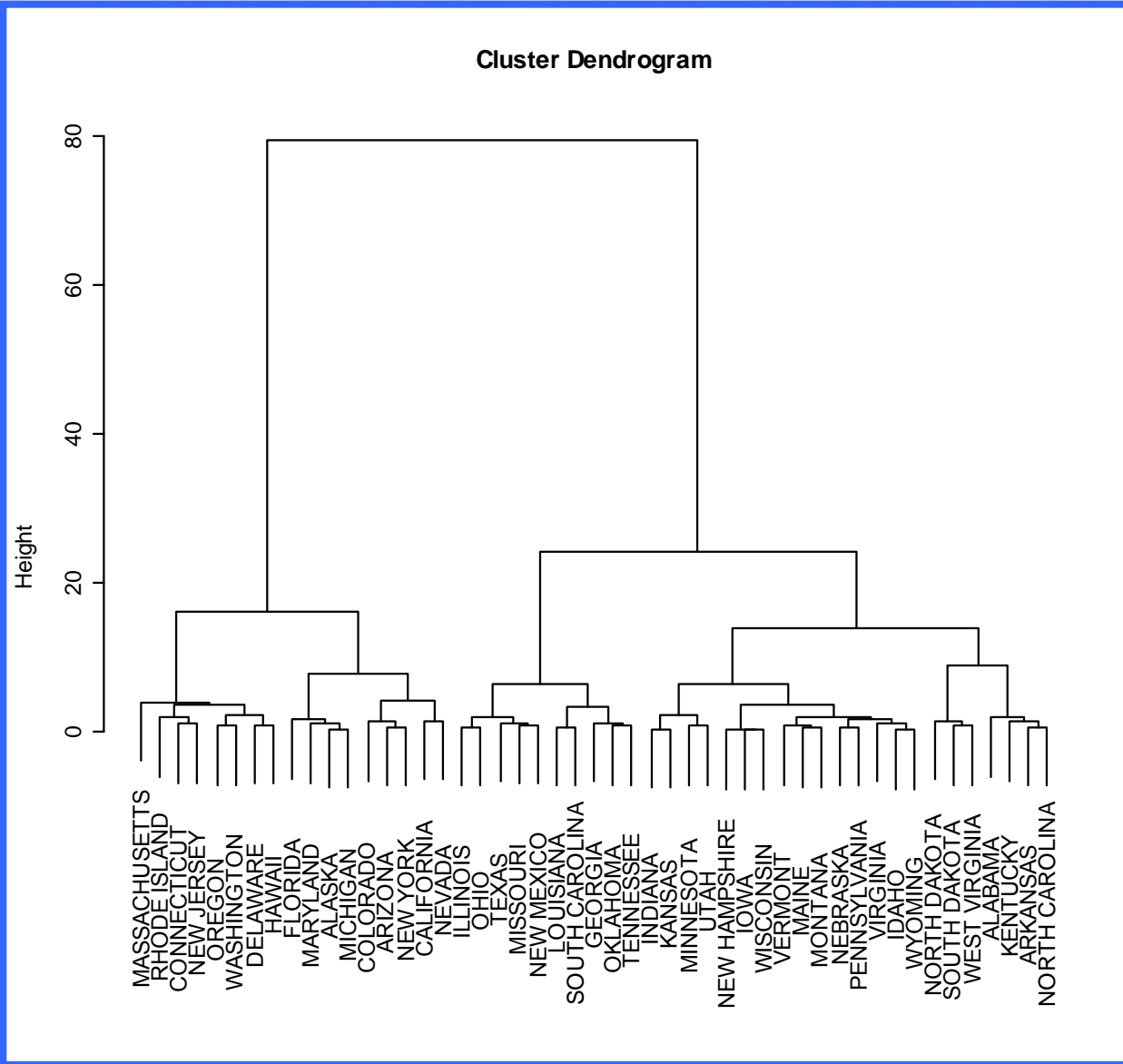


## *Centroid methods: K-means algorithm.*

1. K seed points are chosen and the data is distributed among k clusters.
2. At each step we switch a point from one cluster to another if the  $R^2$  is increased.
3. Then the clusters are slowly optimized by switching points until no improvement of the  $R^2$  is possible.



# Cluster Analysis : Dendrogram using Ward's method



# Cluster Analysis : 6 clusters selected using Ward's method

