STAT 530/J530
October 18th, 2005

Instructor: Brian Habing
Department of Statistics
LeConte 203
Telephone: 803-777-3578
E-mail: habing@stat.sc.edu

Classical Multidimensional Scaling

The goal of multidimensional scaling is to construct a map from a distance matrix.

Classical Multidimensional Scaling with Euclidean Distance produces the same solution as Principal Components Analysis on the raw data.

A Property of Classical MDS

- It gives the smallest values of

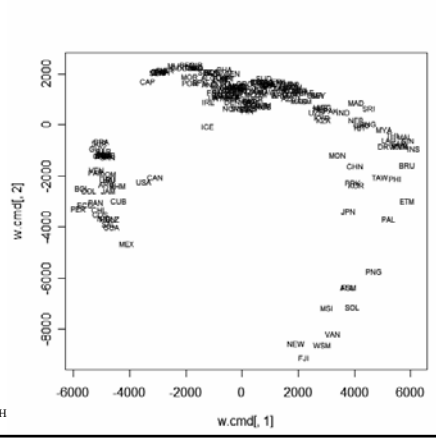
$$\sum (d_{ij}^2 - \hat{d}_{ij}^2)$$

among all the possible projections into lower dimensional Euclidean space.

Example

```
source("http://www.stat.sc.edu/
~habing/courses/530/worlddist.txt")
w.cmd<-cmdscale(world,k=3)
plot(w.cmd[,1],w.cmd[,2],type="n")
text(w.cmd[,1],w.cmd[,2],
   labels=wnames,cex=0.6)
```

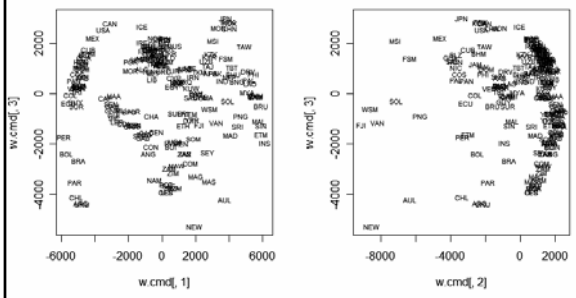

In 2-D



Example Continued

```
par(mfrow=c(1,2))
plot(w.cmd[,1],w.cmd[,3],type="n")
text(w.cmd[,1],w.cmd[,3],
   labels=wnames,cex=0.6)
plot(w.cmd[,2],w.cmd[,3],type="n")
text(w.cmd[,2],w.cmd[,3],
   labels=wnames,cex=0.6)
```

Other Two Views



What About the Census Data?

Karl Pearson distance:

$$d_{ij} = \sqrt{\sum_{k=1}^{q} \left(\frac{x_{ik} - x_{jk}}{s_k} \right)^2}$$

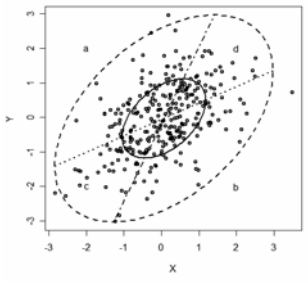
Mahalanobis distance:

$$d_{ij} = \sqrt{(\mathrm{x}_i - \mathrm{x}_j)^T S^{-1} (\mathrm{x}_i - \mathrm{x}_j)}$$

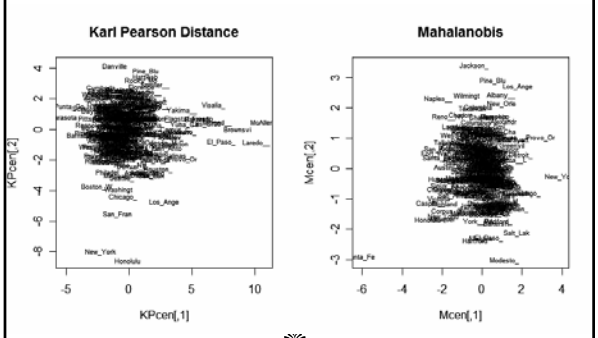
Mahalanobis Distance in 2D

Euclidean distance from a to b or c to d is 5.66

Mahalanobis from a to b is 8 from c to d is 4.62



Census Data



Honolulu and New York?

```
    State Population PopChange PopDens Under5 Over65
110    HI     869857       4.0  1449.3    7.6   12.6
176    NY   19876488       2.0  1955.3    7.2   13.4

    Asian Black Hispanic Birth InfantD CarD HeartD Income
110  64.3   3.8      7.3  17.0     6.7  8.2  241.1  25329
176   6.2  19.3     16.8  15.8     7.9  9.9  403.0  29021

    Poverty Unemploy Grants  BankpP HousepP
110     8.9      5.3   7281 13070.9    30.2
176    14.7      6.5   4831 19988.5    11.9
```


Standardized

```
    Population PopChange PopDens Under5 Over65 Asian
110       0.04     -0.55     4.95   0.43  -0.04 14.09
176      10.20     -0.81     7.07   0.05   0.18  0.90

    Black Hispanic Birth InfantD  CarD HeartD Income
110 -0.67    -0.04  0.82   -0.59 -1.60  -1.18   1.78
176  0.75     0.63  0.38   -0.08 -1.28   0.58   2.98

    Poverty Unemploy Grants BankpP HousepP
110   -1.29     0.01   1.24   1.05   -0.31
176   -0.11     0.50  -0.08   3.02   -1.20
```

Differences on the Dimensions

Values on the first 5 axes

```
            1         2         3         4         5
110 -0.1122433 -8.544552 -3.309014 -1.7871408 -1.3955465
176 -2.2234098 -7.965195 -8.104034 -2.9662571  3.3732595
```

Distances between the observations

```
         110      176      225
110  0.00000 17.24987 16.64626
176 17.24987  0.00000 15.47732
```


How Many Dimensions?

Classical Multidimensional Scaling Tries to Minimize:

$$Stress = \sqrt{\frac{\sum_{i<j}\left(d_{ij}^2 - \hat{d}_{ij}^2\right)}{\sum_{i<j} d_{ij}^2}}$$

And aim for between under 0.1.

Dim	2	3	4	15
Stress	0.64	0.56	0.48	0.09

Other Distances?

For interval valued data -

Maximum distance: $d_{ij} = \max_k |x_{ik} - x_{jk}|$

Manhattan Distance: $d_{ij} = \sum_k |x_{ik} - x_{jk}|$

Other Distances?

For presence absence data -

Matching Index: $d = (a+d)/n$

Ochiai Index: $d = a/\sqrt{(a+b)(a+c)}$

Dice Index: $d = 2a/(2a+b+c)$

Jaccard Index: $d = a/(a+b+c)$

Non-metric Multidimensional Scaling

Sammon Mapping minimizes:

$$\frac{1}{\sum_{i<j} d_{ij}} \sum_{i<j} \frac{\left(d_{ij} - \hat{d}_{ij}\right)^2}{d_{ij}}$$

Isometric Scaling

$$\sqrt{\sum_{i<j} \frac{\left(f(d_{ij}) - \hat{d}_{ij}\right)^2}{\hat{d}_{ij}^2}}$$

Example