STAT 530/J530
October 11th, 2005

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

So Far...

The Fundamentals

- Introduction to the R, SAS, and SPSS
- Mutlivariate Data
- Multivariate Normal Distribution
- Statistical Graphics

So Far...

Methods Based for Simplifying or Modeling Data based on Correlations or Covariances

- Principal Components
- Factor Analysis

Next

Methods Based for Representing of Finding Patterns in Data Based on Distances

- Multidimensional Scaling and Correspondence Analysis
- Cluster Analysis

Distances

A *distance* (or dissimilarity) d_{ij} between two points i and j must satisfy:

1) Symmetry: $d_{ij} = d_{ji}$

2) Non-negativity: $d_{ij} \geq 0$

3) Identification: $d_{ii} = 0$

Metric

A distance d_{ij} is called a *metric* if it also satisfies:

4) Definiteness: $d_{ij} = 0$ if and only if $i = j$.

5) Triangle Inequality:

$$d_{ij} \leq d_{ik} + d_{jk}$$

Euclidean Distance

The *Euclidean distance* between

$x_i = (x_{i1}, x_{i2}, \dots x_{iq})$

and

$x_j = (x_{j1}, x_{j2}, \dots x_{jq})$

is

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots (x_{iq} - x_{jq})^2}$$
$$= \sqrt{\sum_{k=1}^{q} (x_{ik} - x_{jk})^2}$$

Example

APPROXIMATE MILEAGE BETWEEN SELECTED CITIES

Mileage may vary according to actual route traveled.	AIKEN	ALLENDALE	ANDERSON	BENNETTSVILLE	CHARLESTON	CHESTERFIELD	CLINTON	COLUMBIA	DARLINGTON	DILLON	FLORENCE	GAFFNEY	GEORGETOWN	GREENVILLE	GREENWOOD	HILTON HEAD ISLAND	LANCASTER	MCCORMICK	MYRTLE BEACH	ORANGEBURG	ROCK HILL	SPARTANBURG	SUMTER	WALHALLA
AIKEN		52	97	155	126	137	70	56	134	167	137	119	158	108	55	131	112	49	192	53	124	107	100	132
ALLENDALE	52		149	162	90	157	122	76	141	164	136	153	140	160	107	80	132	99	174	43	148	158	102	184
ANDERSON	97	149		213	229	171	51	117	189	222	194	81	240	32	43	270	128	53	262	142	120	59	168	35
BENNETTSVILLE	155	162	213		168	28	146	100	28	26	39	140	101	191	173	209	71	175	85	119	91	146	67	225
CHARLESTON	126	90	229	168		161	168	116	120	133	133	209	60	215	175	102	162	174	96	75	188	206	102	252
CHESTERFIELD	137	157	171	28	161		118	81	41	54	51	112	123	151	147	223	43	157	112	114	66	120	67	197
CLINTON	70	122	51	146	168	118		60	137	170	141	50	183	42	27	220	75	50	201	101	66	36	104	87
COLUMBIA	56	76	117	100	116	81	60		137	111	81	96	123	103	74	156	61	76	149	41	69	94	43	147
DARLINGTON	134	141	189	28	120	41	137	137		50	10	135	82	174	152	189	66	154	80	97	89	143	40	220
DILLON	167	164	222	26	133	54	170	111	50		32	166	84	205	185	209	97	187	64	120	120	174	70	251
FLORENCE	137	136	194	39	133	51	141	81	10	32		145	72	180	155	174	76	157	68	89	99	171	39	228
GAFFNEY	119	163	81	140	209	112	50	96	135	166	145		209	51	75	255	67	98	215	137	44	20	131	97
GEORGETOWN	158	140	240	101	60	123	183	123	82	84	72	209		224	198	168	140	199	35	105	163	216	79	270
GREENVILLE	108	160	32	191	215	151	42	103	174	205	180	51	224		53	255	108	75	248	142	93	31	153	46
GREENWOOD	55	107	43	173	175	147	27	74	152	185	155	75	198	53		186	104	23	215	100	95	64	118	77
HILTON HEAD I.	131	80	270	209	102	223	220	156	189	209	174	255	168	255	186		220	172	194	117	229	246	142	257
LANCASTER	112	132	128	71	162	43	75	61	66	97	76	67	140	108	104	220		125	146	97	23	77	68	154
MCCORMICK	49	99	53	175	174	157	50	76	154	187	157	98	199	75	23	172	125		217	99	118	87	120	88
MYRTLE BEACH	192	174	262	85	96	112	201	149	80	64	68	215	35	248	215	194	146	217		139	167	239	95	288
ORANGEBURG	53	43	142	119	75	114	101	41	97	120	89	137	105	142	100	117	97	99	139		113	134	59	178
ROCK HILL	124	148	120	91	188	66	66	69	89	120	99	44	163	93	95	229	23	118	167	113		64	90	141
SPARTANBURG	107	158	59	146	206	120	36	94	143	174	171	20	216	31	64	246	77	87	239	134	64		144	77
SUMTER	100	102	168	67	102	67	104	43	40	70	39	131	79	153	118	142	68	120	95	59	90	144		191
WALHALLA	132	184	35	225	252	197	87	147	220	251	228	97	270	46	77	257	154	88	288	178	141	77	191	

How?

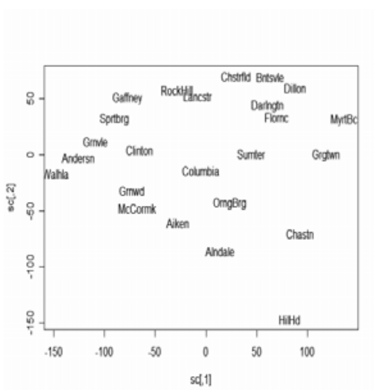
Let X be the *n* by *q* matrix of coordinates we are looking for.

Consider $B = XX^T$…

In R

```
source("http://www.stat.sc.edu/
  ~habing/courses/530/scdist.txt")
sc<-cmdscale(scdist,k=2)
plot(sc,type="n")
text(sc,labels=scname)
```


Result



Properties of Classical MDS

- If we are using Euclidean distance then the classical MDS result is the same as the centered PCA solution.
- It gives the smallest values of

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

among all the possible projections into lower dimensional Euclidean space.