Subjective decisions in hierarchical clustering

① Which variables are used?

② How is the distance or similarity defined?

③ Which linkage is used?

- nearest neighbor
- furthest neighbor
- average linkage
- Centroid linkage

K-means clustering (Quick cluster)

is preferred if the # of clusters is known.

Suggestion: Use hierarchical clustering first, decide how many clusters are appropriate, then run k-means

Factor analysis can be used before clustering, to reduce the # variables

Multidimensional scaling (ALSCAL) (3)

- method of compressing a cloud of points in a higher-dimensional space into 2 or 3 dimensions.
- the ordering of distances is maintained

Classification methods

- logistic regression
- discriminant analysis

Require that you have a known set of items from each of several populations.

Logistic regression (4)

Binary: 2 populations

Left-hand side of the equation is

$\ln\left(\frac{p}{1-p}\right)$, where p is the prob. that the item came from the 1st population

$$\underbrace{\ln\left(\frac{p}{1-p}\right)}_{\text{logit function}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

Logistic regression with more than 2 populations: there must be a natural ordering of the populations.

left-hand side is $\ln\left(\frac{F(x)}{1-F(x)}\right)$

where $F(x)$ is the cumulative probability of the item coming from population x or any population $< x$.

cluster.sav

	pres	south	elected	party	congress	vp
1	reagan	0	1	0	0	0
2	carter	1	1	1	0	0
3	ford	0	0	0	1	1
4	nixon	0	1	0	1	1
5	johnson	1	0	1	1	1
6	kennedy	0	1	1	1	0

cluster.sav

	QCL_1	QCL_2
1	2	1
2	2	2
3	1	3
4	1	3
5	1	3
6	2	1

```
CLUSTER   south elected party congress vp
  /METHOD BAVERAGE
  /MEASURE=SEUCLID
  /ID=pres
  /PRINT SCHEDULE
  /PLOT DENDROGRAM HICICLE.
```

Cluster

```
[DataSet1] F:\cluster.sav
```

Case Processing Summary[a,b]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
6	100.0	0	.0	6	100.0

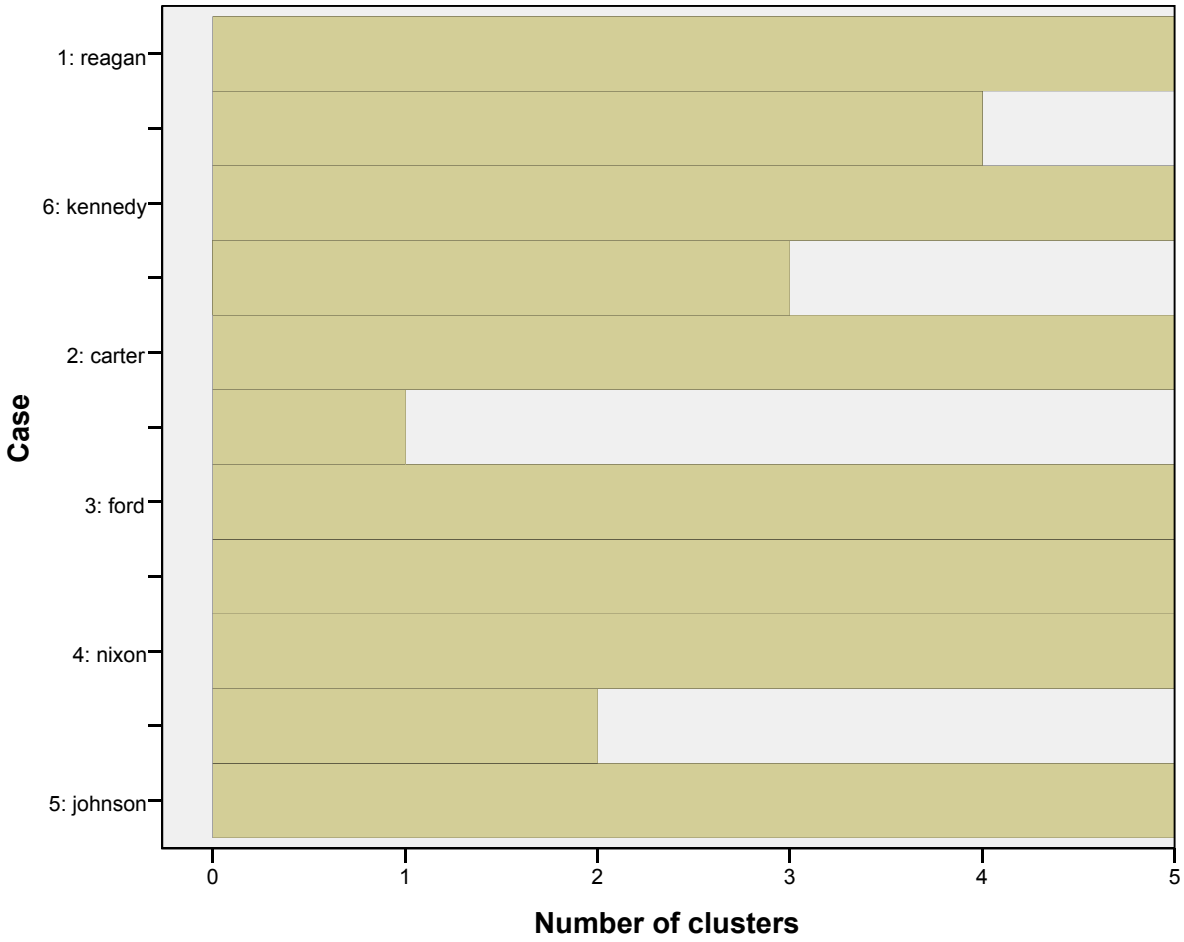
a. Squared Euclidean Distance used

b. Average Linkage (Between Groups)

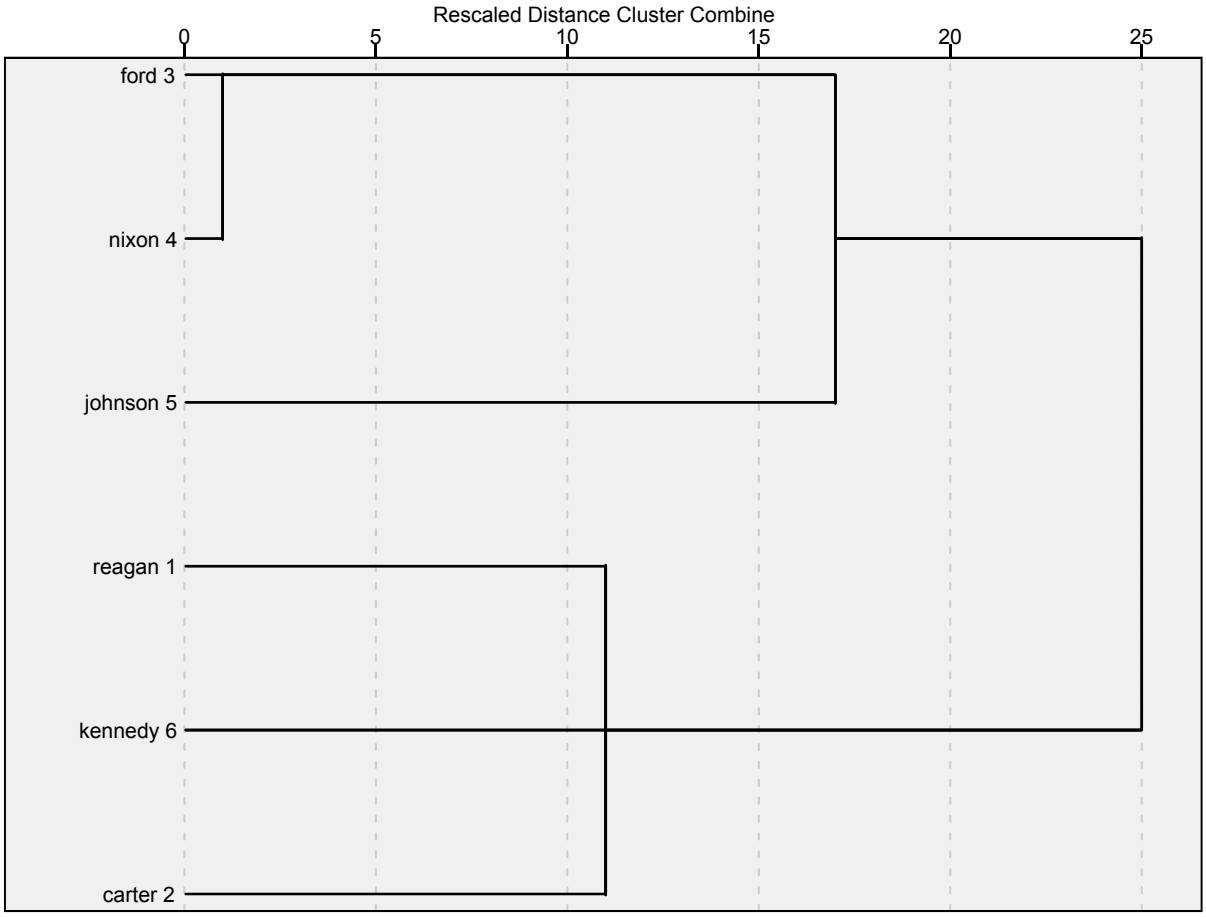
Average Linkage (Between Groups)

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	4	1.000	0	0	4
2	1	6	2.000	0	0	3
3	1	2	2.000	2	0	5
4	3	5	2.500	1	0	5
5	1	3	3.333	3	4	0


1: reagan
6: kennedy
2: carter
3: ford
4: nixon
5: johnson
Case
0
1
2
3
4
5
Number of clusters

Dendrogram using Average Linkage (Between Groups)




```
QUICK CLUSTER south elected party congress vp
  /MISSING=LISTWISE
  /CRITERIA=CLUSTER(2) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER
  /PRINT ID(pres) INITIAL.
```

Quick Cluster

```
[DataSet1] F:\cluster.sav
```

Initial Cluster Centers

	Cluster	
	1	2
south	0	1
elected	0	1
party	0	1
congress	1	0
vp	1	0

Iteration History[a]

	Change in Cluster…	
Iteration	1	2
1	.577	.816
2	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 2.236.

Final Cluster Centers

	Cluster	
	1	2
south	0	0
elected	0	1
party	0	1
congress	1	0
vp	1	0

Number of Cases in each Cluster

Cluster	1	3.000
	2	3.000
Valid		6.000
Missing		.000

```
QUICK CLUSTER south elected party congress vp
  /MISSING=LISTWISE
  /CRITERIA=CLUSTER(3) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER
  /PRINT ID(pres) INITIAL.
```

Quick Cluster

```
[DataSet1] F:\cluster.sav
```

Initial Cluster Centers

	Cluster		
	1	2	3
south	0	1	0
elected	1	1	0
party	0	1	0
congress	0	0	1
vp	0	0	1

Iteration History[a]

	Change in Cluster Centers		
Iteration	1	2	3
1	.707	.000	.577
2	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 1.414.

Final Cluster Centers

	Cluster		
	1	2	3
south	0	1	0
elected	1	1	0
party	0	1	0
congress	0	0	1
vp	0	0	1

Number of Cases in each Cluster

Cluster	1	2.000
	2	1.000
	3	3.000
Valid		6.000
Missing		.000

```
SAVE OUTFILE='F:\cluster.sav'
 /COMPRESSED.
```

Table 1 Flying Mileages Between 10 American Cities

Atlanta	Chicago	Denver	Houston	Los Angeles	Miami	New York	San Francisco	Seattle	Washington, DC	
0	587	1212	701	1936	604	748	2139	2182	543	Atlanta
587	0	920	940	1745	1188	713	1858	1737	597	Chicago
1212	920	0	879	831	1726	1631	949	1021	1494	Denver
701	940	879	0	1374	968	1420	1645	1891	1220	Houston
1936	1745	831	1374	0	2339	2451	347	959	2300	Los Angeles
604	1188	1726	968	2339	0	1092	2594	2734	923	Miami
748	713	1631	1420	2451	1092	0	2571	2408	205	New York
2139	1858	949	1645	347	2594	2571	0	678	2442	San Francisco
2182	1737	1021	1891	959	2734	2408	678	0	2329	Seattle
543	597	1494	1220	2300	923	205	2442	2329	0	Washington, DC

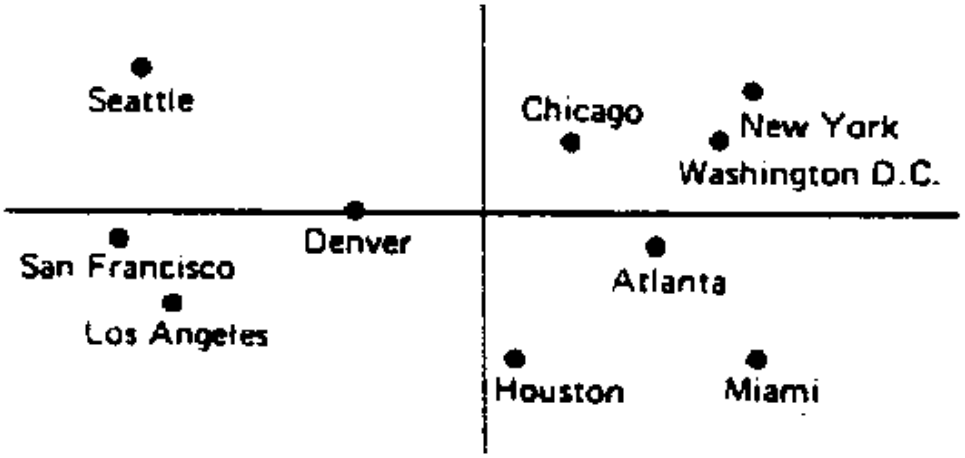
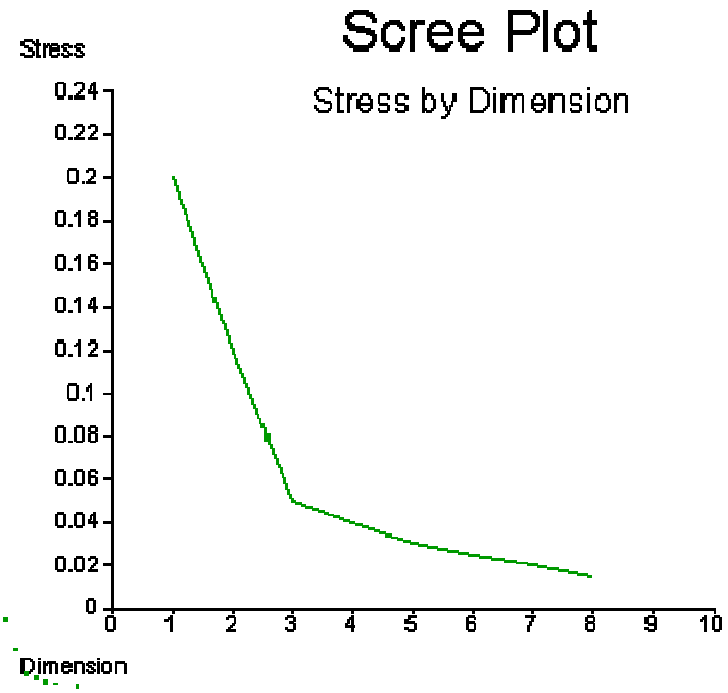


Figure 1 CMDS of flying mileages between 10 American cities.



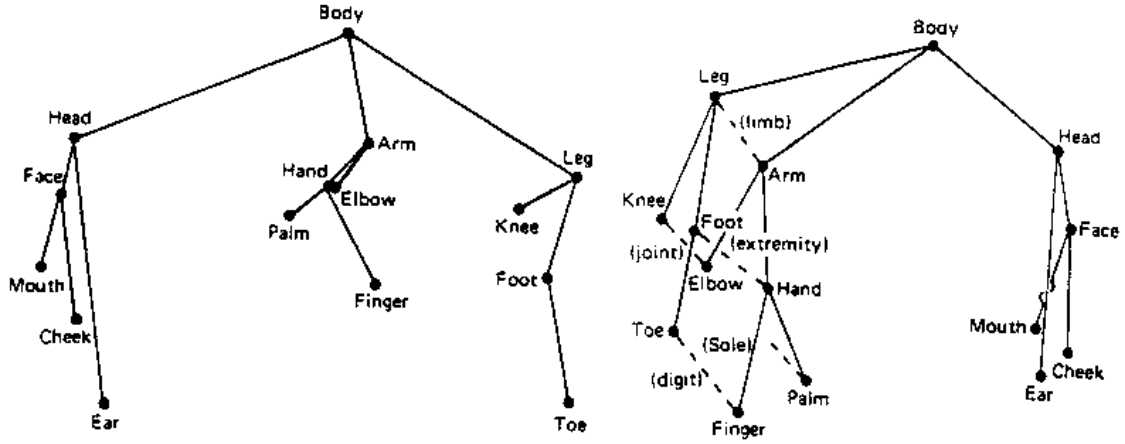
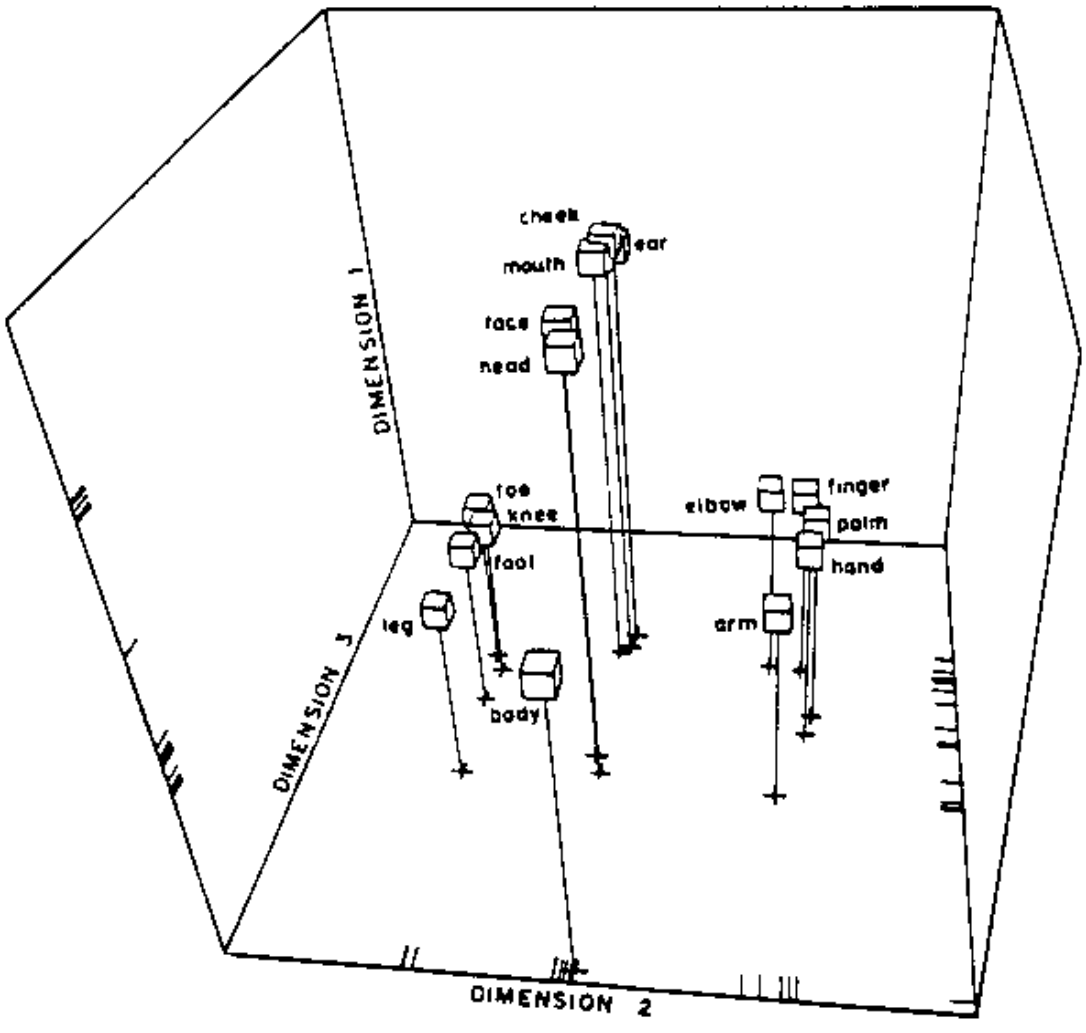


Figure 2 (a) RMDS of children's similarity judgments about 15 body parts; (b) RMDS of adults' similarity judgments about 15 body parts.



(a)

Figure 3 (a) WMDS of children's and adults' similarity judgments about 15 body parts.