

# Cluster Analysis

We have a dataset with  $n$  observations and we want to group the observations into  $k$  distinct natural groups of similar observations.

We distinguish three stages of cluster analysis:

- (i) Input Stage
- (ii) Algorithm stage
- (iii) Output stage

## I. Input Stage

### 1. Scaling:

- a. Divide variables by the standard deviation.
- b. Spherize the data: Invariance under affine transformations.

$$Z = A Y ; A = \text{Chol} ( S )^{-1} \text{ or the symmetric square root } S^{-1/2};$$

- c. Spherize the data with the within variance.

$$T = W + B$$

To obtain  $W$  use iteration. Find clusters then find  $W$  then find clusters and so on...

### 2. Similarity measures.

Clustering methods require the definition of a similarity or dissimilarity measure.

For example an inter-point distance  $d(x_1, x_2)$  and an inter-cluster distance  $d^*(C_1, C_2)$  are examples of dissimilarity.

The inter point distance is often taken to be the Euclidean distance

$$d_e(x_1, x_2) = \sqrt{\sum_{j=1}^p (x_{1j} - x_{2j})^2}$$

Or Mahalanobis distance  $d_A(x_1, x_2) = \sqrt{(x_1 - x_2)' A (x_1 - x_2)}$

Some times we may use the Manhattan distance:

$$d_M(x_1, x_2) = \sum_{j=1}^p |x_{1j} - x_{2j}|$$

When the data is not metric we may define any distance or similarity measure from characteristics of the problem. For example for binary data given any two vector observations we construct the table

	1	0	Total
1	a	b	a+b
0	c	d	c+d
Total	A+c	b+d	P

Then we define distance as the square root of the  $\chi^2$  statistic.

Also  $d = 1 - (a+d)/p$  or  $d = 1 - a/(a+b+c)$

## II. Algorithm Stage

There are many approaches to cluster analysis and this very noticeable on the software implementations of cluster analysis on the usual statistical packages, which implement 10 or 15 different methods. Although it is very intuitive the ideas

of cluster analysis are difficult to formalize in a general unique sense. We present here a limited review. Two general methods for doing cluster Analysis:

### 1. Hierarchical clustering.

The inter cluster distance between two clusters is defined as a function of the inter point distances between pairs of points where each point comes from a different cluster. The popular definitions of inter cluster distances are:

- Single Linkage or Minimum method: distance between the closes two points
- Complete Linkage or Maximum method: distance between the furthest two points
- Average Linkage: Average distance between every pair of points
- Ward:  $R^2$  change. Define  $R^2$  like in a linear model where the  $k$  clusters correspond to to the factor variable and calculate the  $R^2$ . The clusters to be joined at each step are such that the  $R^2$  is reduced the least.

We build a hierarchical tree starting with a cluster at each sample point, and at each stage of the tree the two closest clusters joint to form a new cluster. Once we finish building the tree the question becomes:

"how many clusters do we chose?"

One way of making this determination is by inspecting the hierarchical tree and finding a reasonable point to break the clusters. We can also plot the criteria function for the different number of cluster and visually look for unusually large jumps.

### 2. Non Hierarchical procedures:

Centroid methods. K-means algorithm.

We start with a choice of  $k$  clusters and a choice of distance.

- a. Determine the initial set of  $k$  clusters.  $k$  seed points are chosen and the data is distributed among  $k$  clusters.
- b. Calculate the centroids of the  $k$  clusters and move each point to the cluster whose centroid is closest.
- c. Repeat step b. until no change is observed.

This is the same as optimizing the  $R^2$  criteria. At each stage of the algorithm one point is moved to the cluster that will optimize the criteria function. This is iterated until convergence occurs. The final configuration has some

dependence on the initial configuration so it is important to take a good start.

One possibility is to run WARD's method and use the outcome as initial configuration for k-means.

## EXAMPLE OF WARD'S

SAS code: For the Florida & Georgia subset of the hospital database.

```
PROC CLUSTER METHOD=WARD;
VAR BEDS RBEDS OUTV ADM SIR
    HIP95 KNEE95 TH TRAUMA REHAB HIP96
    KNEE96 FEMUR96;
COPY BEDS RBEDS OUTV ADM SIR
    HIP95 KNEE95 TH TRAUMA REHAB HIP96
    KNEE96 FEMUR96 SALES12 SALESY ;

PROC TREE NOPRINT NCL=7 OUT=TXCLUST;
COPY BEDS RBEDS OUTV ADM SIR
    HIP95 KNEE95 TH TRAUMA REHAB HIP96
    KNEE96 FEMUR96 SALES12 SALESY ;

RUN;
```

## OUTPUT:

Ward's Minimum Variance Cluster Analysis					
Number of Clusters	-Clusters	Joined--	Frequency of New Cluster	Semipartial R-Squared	R-Squared
349	OB283	OB341	2	0.000003	0.999997
348	OB209	OB244	2	0.000004	0.999993
347	OB249	OB255	2	0.000006	0.999987
346	OB50	OB290	2	0.000006	0.999981
.....					
31	CL48	CL44	15	0.001934	0.922789
30	CL64	CL81	22	0.002040	0.920750
29	CL115	OB132	4	0.002124	0.918626
28	OB21	OB105	2	0.002133	0.916493
27	CL37	CL43	18	0.002503	0.913991
26	CL49	CL42	50	0.002944	0.911047
25	CL41	CL36	37	0.003055	0.907991
24	CL32	CL51	11	0.003103	0.904888
23	CL47	CL38	23	0.003314	0.901575
22	CL31	CL56	18	0.003447	0.898127
21	CL59	OB192	27	0.003504	0.894624
20	CL24	CL34	16	0.003826	0.890798
19	CL40	CL136	5	0.004258	0.886541
18	CL27	OB53	19	0.004396	0.882144
17	CL33	CL35	66	0.004814	0.877331
16	CL21	CL55	47	0.005127	0.872204
15	CL46	CL19	16	0.006717	0.865486
14	CL28	CL20	18	0.007246	0.858240
13	CL30	CL29	26	0.007860	0.850379
12	CL18	CL39	23	0.008437	0.841942
11	CL26	CL53	68	0.008496	0.833447
10	CL45	CL15	24	0.008555	0.824891
9	CL25	CL16	84	0.009749	0.815142
8	CL23	CL13	49	0.009836	0.805306
7	CL8	CL22	67	0.009713	0.795593
6	CL17	CL11	134	0.037362	0.758231

5	CL9	CL14	102	0.037383	0.720848
4	CL7	CL12	90	0.049615	0.671232
3	CL6	CL10	158	0.063334	0.607898
2	CL4	CL5	192	0.114961	0.492937
1	CL2	CL3	350	0.492937	0.000000

SUMMARY TABLE OF CLUSTER MEANS

OBS	CLUSTER	MBEDS	MRBEDS	MOUTV	MADM	MSIR	MHIP95	MKNEE95
1	1	9.3313	6.48534	3.2865	6.37741	0.41479	0.6544	0.0833
2	2	8.6549	0.18475	9.7564	7.63686	7.09047	1.7569	0.8568
3	3	10.3784	0.13552	9.4711	8.01049	7.77997	4.7082	3.3104
4	4	13.9877	0.17487	10.1449	8.63182	8.34591	7.8699	6.3164
5	5	18.8639	1.32716	10.2084	9.40176	9.08937	10.6524	9.6837
6	6	23.9667	0.86351	11.7035	9.48459	8.87018	5.9921	3.9397
7	7	20.5378	0.82940	10.6222	9.58004	9.31534	15.8222	16.5184

OBS	MTH	MTRAUMA	MREHAB	MHIP96	MKNEE96	MFEMUR96	MSALES12	MSALESY
1	0.16667	0.00000	0.83333	0.5681	0.1006	0.8287	1.25634	1.21142
2	0.13235	0.04412	0.05882	1.7226	0.4283	2.0169	0.42665	0.38974
3	0.04545	0.06061	0.03030	4.9527	3.3867	5.5761	1.70466	1.49698
4	0.01190	0.08333	0.03571	7.8181	6.3195	8.5522	2.57560	2.32965
5	0.23881	0.25373	0.25373	10.7619	9.4927	10.7842	3.84354	3.44889
6	0.83333	0.22222	0.22222	5.1608	3.4116	7.0432	1.32706	1.09774
7	0.30435	0.21739	0.17391	15.6797	16.1569	13.1042	3.92407	3.40702