Chapter 7
Cluster Analysis (2)

Rung-Hung, Su

Fu Jen Catholic University
Department of Statistics and Information Science
Mail: 141637@mail.fju.edu.tw
Phone: 2905-3194 (SL474)

7.6.1 Interpreting the SAS output

Evaluating the Cluster Solution and Determining the Number of Clusters

1) The most widely used statistics are:

- *Root-mean-square standard deviation (RMSSTD) of the new cluster.*
- *R-squared (RS)*
- *Semipartial R-SQUARED (SPR)*
- *Distance between two cluster*

2) The following Table will be used in computing the various statistics (includes Error sum of square, ESS=SSw and degree of freedom)

Table 7.8 **Within-Group Sum of Squares and Degrees of Freedom for Clusters Formed in Steps 1, 2, 3, 4, and 5**

		Within-Group Sum of Squares			Degrees of Freedom		
Step Number	Cluster	Income	Education	Pooled	Income	Education	Pooled
1	CL5	0.500	0.500	1.000	1	1	2
2	CL4	0.500	0.500	1.000	1	1	2
3	CL3	12.500	0.500	13.000	1	1	2
4	CL2	157.000	26.000	183.000	3	3	6
5	CL1	498.333	202.833	701.166	5	5	10

7.6.1 Interpreting the SAS output

Evaluating the Cluster Solution and Determining the Number of Clusters

RMSSTD of the new cluster (3d)

1) *RMSSTD* is the pooled standard deviation of all the variable forming the cluster.

2) Since the objective of cluster analysis is to form homogeneous group, the *PMSSTD* of a cluster should be small as possible, vice versa.

3) Note that there are no way to decide what is "small"

$$Pooled\ SD: s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}} \qquad s_{p,CL2} = \sqrt{\frac{157+26}{3+3}} = 5.523$$

Table 7.8 Within-Group Sum of Squares and Degrees of Freedom for Clusters Formed in Steps 1, 2, 3, 4, and 5

		Within-Group Sum of Squares			Degrees of Freedom		
Step Number	Cluster	Income	Education	Pooled	Income	Education	Pooled
1	CL5	0.500	0.500	1.000	1	1	2
2	CL4	0.500	0.500	1.000	1	1	2
3	CL3	12.500	0.500	13.000	1	1	2
4	CL2	157.000	26.000	183.000	3	3	6
5	CL1	498.333	202.833	701.166	5	5	10

$(n_1-1)s_1^2$

$(n_2-1)s_2^2$

7.6.1 Interpreting the SAS output

Evaluating the Cluster Solution and Determining the Number of Clusters

R-SQUARED (RS) 3f

1) R^2 is the ration of SS_b to SS_t , *in which* $SS_t = SS_{b+}\ SS_w$ and SS_w=*ESS*

$$R^2 = \frac{SS_b}{SS_t} = \frac{SS_b}{SS_w + SS_b}, \qquad 0 \le R^2 \le 1.$$

2) The value of R^2 is as large as possible (SS_b is large, and SS_w is small)

3) For example, in the step 4, the result has two group CL2={S3, S4, S5, S6} and CL5={S1, S2}.

$$SS_w = SS_{w,Inc} + SS_{w,Edu} = (157 + 0.5) + (26 + 0.5) = 184$$

$$SS_t = (n-1)s_1^2 + (n-1)s_2^2 = 701.166$$

$$SS_b = SS_t - SS_w = 701.166 - 184 = 517.666$$

$$R^2 = \frac{SS_b}{SS_t} = \frac{517.666}{701.166} = 0.783.$$

7.6.1 Interpreting the SAS output

Evaluating the Cluster Solution and Determining the Number of Clusters

Semipartial R-SQUARED (SPR) 3e

1) The new cluster formed at any given step is obtained by merging two cluster formed in previous steps.

2) The difference between the pooled SS_w of the new cluster and the sum of pooled SS_w's of cluster joined to obtain the new cluster is called ***loss of homogeneity***

3) The value of SPR is loss of homogeneity to divided by SS_t, therefore, the value of SPR is as small as possible.

4) For example, step 4 is combine CL4 and CL3

$$SPR = \frac{\textbf{loss of homogeneity}}{SS_t} = \frac{SS_{w,after} - SS_{w,before}}{SS_t}$$

$$= \frac{SS_{w,CL2} - (SS_{w,CL4} + SS_{w,CL3})}{SS_t} = \frac{184 - (1+13)}{701.166} = 0.241$$

7.6.1 Interpreting the SAS output

Evaluating the Cluster Solution and Determining the Number of Clusters

Distance between two cluster 3g

1) In the centroid method, it is simply the Euclidean distance between the centroids of the two clusters that are to be joined or merged and it is termed the centroid distance (CD).

2) For example, step 4 is combine CL4 and CL3, see Table 7.5

Table 7.1 **Hypothetical Data**

Subject Id	Income ($ thous.)	Education (years)
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

$$CD = \sqrt{(27.5-15.5)^2 + (19.5-14.5)^2} = 13.00$$

Table 7.5 **Centroid Method: Three Clusters**

Data for Three Clusters

Cluster	Cluster Members	Income ($ thous.)	Education (years)
1	S1&S2	5.5	5.5
2	S3&S4	15.5	14.5
3	S5&S6	27.5	19.5

Similarity Matrix

	S1&S2	S3&S4	S5&S6
S1&S2	0.00	181.00	680.00
S3&S4	181.00	0.00	169.00
S5&S6	680.00	169.00	0.00

7.6.1 Interpreting the SAS output

Table 7.9 **Summary of the Statistics for Evaluating Cluster Solution**

Statistic	Concept Measured	Comments
RMSSTD	Homogeneity of new cluster	Value should be small
SPR	Homogeneity of merged clusters	Value should be small
RS	Heterogeneity of clusters	Value should be high
CD	Homogeneity of merged clusters	Value should be small

7.6.1 Interpreting the SAS output

1) The 4 statistics can be used for determining the number of clusters.
2) Essentially, one looks for a big jump in the value of a given statistics.
3) It is clear that there is a "big" change in the values when going from a three-cluster to a two-cluster solution, it appears that there are three clusters.

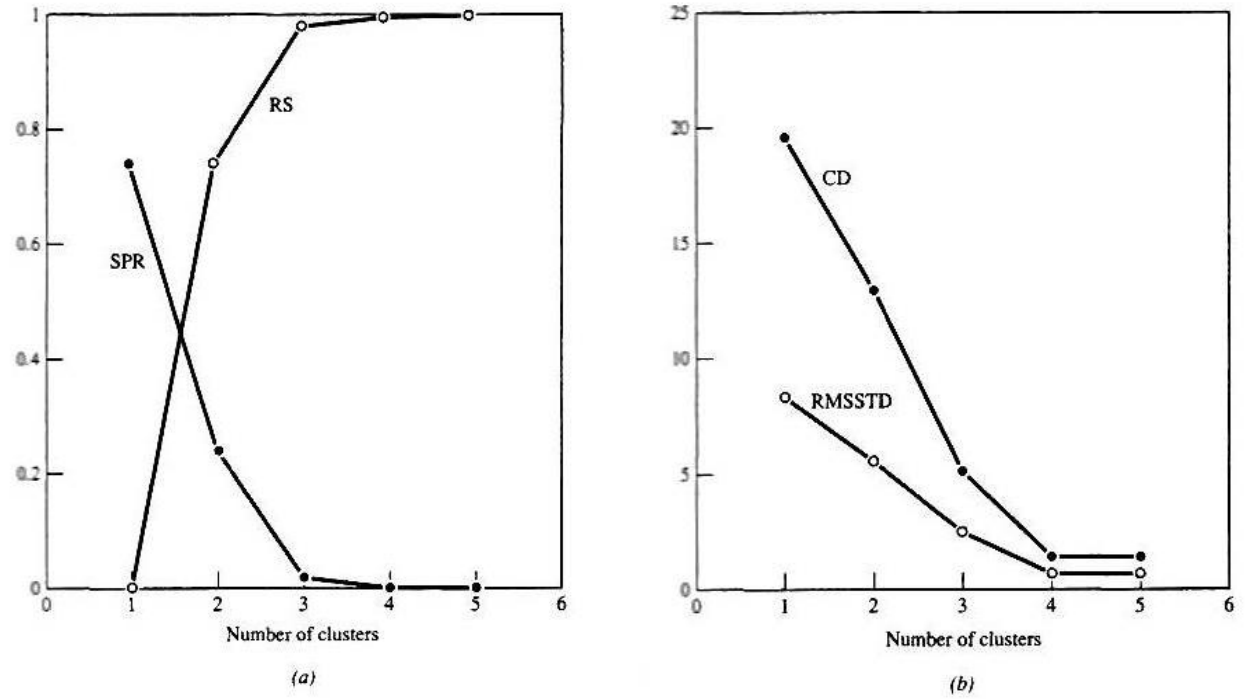

Figure 7.3 Plots of (*a*) SPR and RS and (*b*) RMSSTD and CD.