

DSC 410/510
Multivariate Statistical Methods
Cluster Analysis

1

What Does Cluster Analysis Do?

- **Classifies** cases (e.g. respondents) into clusters (groups) based on their variable values (characteristics)
 - ◆ cases within clusters are similar to each other
 - ◆ clusters are dissimilar from one another
- Example: group the class according to part-worths for the Jobs example
- Set of variables used to classify cases is the **cluster variate**
 - ◆ specified by researcher
 - ◆ not estimated empirically

2

Examples of Cluster Analysis

- Marketing segmentation analyses
- Analyze similarities and differences among new products
- Identify groupings of firms based on strategic orientations, performance, etc.
- Make survey results more manageable by providing summaries within “natural groupings” of the sample
- Profile resulting clusters for demographic similarities and differences

3

Features of Cluster Analysis

- Descriptive, exploratory technique
- Non-inferential
 - ◆ no statistical inference from sample to population
- Wide variety of methods
 - ◆ can produce many different “solutions”
- Clusters always produced, regardless of “true structure”
- Highly dependent on the variables used

4

Illustrative Example

- Determine class segments based on patterns of part-worths for:

| Id | Denver | Eugene |
|--------|--------|--------|
| ID1156 | -1.50 | 1.50 |
| ID5056 | -5.00 | 4.25 |
| ID6275 | -6.00 | 1.00 |
| ID6654 | -2.38 | 2.88 |
| ID7713 | -6.00 | 2.00 |
| ID9285 | -1.00 | 2.25 |
| ID9348 | -0.75 | 0.75 |

5

Example: Scatterplot

See `jobsppw1.xls` file from data page on web-site

6

How Cluster Analysis Works

- Define a measure of (dis)similarity
 - ◆ e.g. (Euclidean) distance
- Form clusters using similarity measure
 - ◆ e.g. iteratively group cases that are closest together (“Hierarchical nearest-neighbor”)
- Decide the number of clusters to end with
 - ◆ e.g. trade-off decreasing # to simplify matters with increasing # to make cases within clusters as similar as possible

7

(Dis)similarity: Euclidean Distance

- In two dimensions, Pythagoras!
- Extend to more than two dimensions:

Euclidean distance between (x_1, x_2, \dots, x_p) and (y_1, y_2, \dots, y_p) is $((x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_p - y_p)^2)^{1/2}$

8

Example: Scatterplot

Example:
Euclidean distance between 1156 and 9285 is:
 $(0.5^2 + 0.75^2)^{1/2} = 0.9014$

9

Proximity Matrix in Excel

| id | Denver | Eugene | ID1156 | ID5056 | ID6275 | ID6654 | ID7713 | ID9285 | ID9348 |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| | | | -1.50 | -5.00 | -6.00 | -2.38 | -6.00 | -1.00 | -0.75 |
| | | | 1.50 | 4.25 | 1.00 | 2.88 | 2.00 | 2.25 | 0.75 |
| ID1156 | -1.50 | 1.50 | 0.0000 | 4.4511 | 4.5277 | 1.6298 | 4.5277 | 0.9014 | 1.0607 |
| ID5056 | -5.00 | 4.25 | 4.4511 | 0.0000 | 3.4004 | 2.9633 | 2.4622 | 4.4721 | 5.5057 |
| ID6275 | -6.00 | 1.00 | 4.5277 | 3.4004 | 0.0000 | 4.0812 | 1.0000 | 5.1539 | 5.2559 |
| ID6654 | -2.38 | 2.88 | 1.6298 | 2.9633 | 4.0812 | 0.0000 | 3.7291 | 1.5104 | 2.6751 |
| ID7713 | -6.00 | 2.00 | 4.5277 | 2.4622 | 1.0000 | 3.7291 | 0.0000 | 5.0062 | 5.3968 |
| ID9285 | -1.00 | 2.25 | 0.9014 | 4.4721 | 5.1539 | 1.5104 | 5.0062 | 0.0000 | 1.5207 |
| ID9348 | -0.75 | 0.75 | 1.0607 | 5.5057 | 5.2559 | 2.6751 | 5.3968 | 1.5207 | 0.0000 |

Excel: e.g. cell I4 = $((\$B4-\$2)^2+(\$C4-\$3)^2)^{0.5}$
See **jobspw1.ed.xls** file from data page on website

10

Cluster Analysis in SAS

- SAS: see **jobspw1.sas** file from data page on web-site
- This is a text file of SAS code that will run a SAS “procedure” on the **jobspw1** data:
 - ◆ Import the **jobspw1** data as usual in SAS
 - ◆ Select **File > Open** to open the **jobspw1.sas** file in the Editor Window
 - ◆ Select **Run > Submit** to run the code

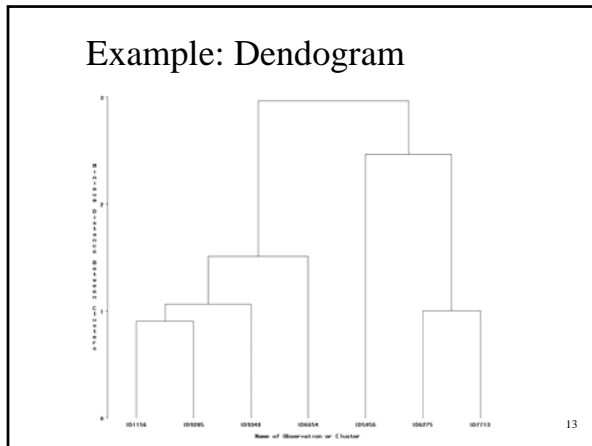
11

Form Clusters: Hierarchical

| Cluster History | NCL | --Clusters Joined-- | FREQ | Min Dist |
|-----------------|-----|---------------------|------|----------|
| | 6 | ID1156 ID9285 | 2 | 0.9014 |
| | 5 | ID6275 ID7713 | 2 | 1 |
| | 4 | CL6 ID9348 | 3 | 1.0607 |
| | 3 | CL4 ID6654 | 4 | 1.5104 |
| | 2 | ID5056 CL5 | 3 | 2.4622 |
| | 1 | CL3 CL2 | 7 | 2.9633 |

CL6: 1156 and 9285 combine since they are closest (distance = 0.9014).
CL5: 6275 and 7713 combine (distance = 1).
CL4: 9348 joins cluster CL6 (distance from 9348 to 1156 = 1.0607).
CL3: 6654 joins cluster CL4 (distance from 6654 to 9285 = 1.5104).
CL2: 5056 joins cluster CL5 (distance from 5056 to 7713 = 2.4622).
CL1: cluster CL3 joins cluster CL2 (distance from 5056 to 6654 = 2.9633).

12

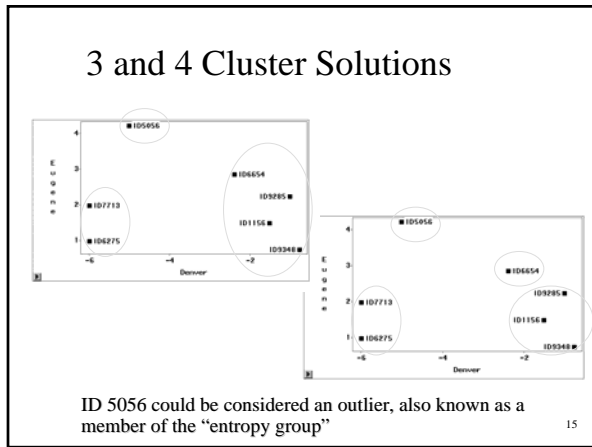


Decide Number of Clusters

- Start from 1 cluster, and look back for large “jumps” in the similarity measure (distance for cases joined at each stage)
- Going from 3 clusters to 2 led to a 63.0% jump
- Going from 4 clusters to 3 led to a 42.4% jump
- Suggests the 3 and 4 cluster solutions are worth pursuing

| Clusters | Coefficient | % change to next level |
|----------|-------------|------------------------|
| 6 | 0.9014 | 10.9% |
| 5 | 1.0000 | 6.1% |
| 4 | 1.0607 | 42.4% |
| 3 | 1.5104 | 63.0% |
| 2 | 2.4622 | 20.4% |
| 1 | 2.9633 | |

14



Cluster Analysis Objectives

- Partition cases into groups based on similarities of characteristics:
 - To form a taxonomy (empirically-based classification)
 - To compare with a typology (theoretically-based classification)
 - To simplify data structure (clusters can be profiled by their general characteristics)
 - To reveal hidden relationships among cases

16

Selecting Variables

- Clusters reflect structure only within selected variables
 - selection can be based on theoretical, conceptual, and practical considerations
- Cluster analysis does not distinguish relevant from irrelevant variables
 - inclusion of irrelevant variables can dramatically affect results

17

Research Design

- Cluster analysis has a fair amount of “art” rather than “science” to it:
 - Outlier detection/deletion – yes or no?
 - Which similarity measure to use?
 - Standardize data or not?
- No definitive answers, just guidelines
- No final correct “solution”, just suggested answers worth pursuing

18

Outliers

- Profile plot helps identify:
 - extreme values
 - unique patterns
- If truly aberrant, delete
- Otherwise, be sure it represents population group

19

Similarity Measures

- “Similarity” measures correspondence (or resemblance) between cases
- “Dissimilarity” measures the opposite
- Selected p variables are combined into a (dis)similarity measure for all pairs of cases
 - ★ ♦ distance (metric)
 - ♦ association (non-metric)
 - ♦ correlational (metric)

20

Distance Measures

- “Distance” between points in p -dimensional space
 - ♦ Resulting clusters contain cases with similar values across the set of variables (although patterns can be different), e.g. 3572 and 8729
- Many possibilities, for example:
 - ★ ♦ (Squared) Euclidean
 - ♦ City block (sum absolute differences)
 - ♦ Chebychev (max absolute difference)
 - ♦ Minkowski (generalization of Euclidean)

21

Association Measures

- Degree of agreement (matching) between each pair of cases for non-metric variables
 - ♦ For example, three yes/no variables
 - ♦ Two cases are similar if they provide the same pattern of yes/no responses (e.g. yes, no, no)
 - ♦ Resulting clusters again contain cases with similar values across the set of variables
- Many possibilities, for example:
 - ♦ % of times there is agreement (both respondents said yes, or both said no) across the set of questions

22

Correlational Measures

- Correspondence of “patterns” across the variables
 - ♦ For example, two cases with an increasing trend over the variables would be similar
 - ♦ But, the cases need not have similar *values* for the variables
- Correlational measures are rarely used in cluster analysis, since we’re usually interested in the magnitudes of variables not their patterns

23

Standardizing Data by Variable

- Similarity measures are very sensitive to differing scales among the variables
 - ♦ e.g. consider clustering cases based on age in years and income in \$ – clusters would be dominated by income differences
- To avoid this problem, cluster analysis is *usually* based on variables standardized to have mean 0 and variance 1 (i.e. standardized variables are given equal weight when clustering)
- *Sometimes* you don’t want to standardize (e.g. for part-worths, where variances indicate relative importance which we *do* want to weight by)

24

Standardizing Data by Case

- Attitudinal data often benefits from standardizing by case (i.e. values for each case have mean 0 and variance 1 across variables)
- Why?
 - ◆ otherwise, “yea-sayers” appear in one cluster, “nay-sayers” in another, and *patterns* of attitudes are lost
- Should not be done if the *magnitude* of the variables is important in deriving clusters

25

Cluster Analysis Assumptions

- On the plus side, Normality, linearity, homoscedasticity are not important, but ...
- If we want results that generalize to a population, our sample better be representative
- Multicollinearity (of variables) can adversely affect results:
 - ◆ suppose population structure depends on x_1 and x_2 only, but a variable (x_3) highly collinear with x_2 is also used to cluster sample
 - ◆ the x_2 “dimension” will now be “double-weighted” in defining the clusters

26

Deriving Clusters

- General principal: form clusters to minimize variation within clusters relative to variation between clusters (*c.f.* ANOVA!)
- Two broad categories of algorithm:
 - ◆ Hierarchical
 - ◆ Nonhierarchical (a.k.a. “K-Means”)
- Can be used in conjunction, e.g. use hierarchical first, then feed results into nonhierarchical

27

Hierarchical Cluster Analysis

- Iteratively group cases using a “cluster method”
 - ◆ single linkage (nearest neighbor)
 - ◆ complete linkage (furthest neighbor)
 - ★ ◆ average linkage (between-groups)
 - ★ ◆ Ward’s method (minimum within-cluster sum of squares)
 - ◆ Centroid method (centroids are within-cluster variable means)
- Dendrogram show tree-like process as number of clusters reduces from n to 1

28

Some Linking Criteria

29

Nonhierarchical Cluster Analysis

- Assign cases to one of k clusters based on (Euclidean) distance to cluster “seed” (center)
 - ◆ k is fixed by the researcher
 - ◆ Initial seeds can be chosen at random, selected empirically, or specified by the researcher
 - ◆ SAS uses the “sequential threshold” method for clustering (cases are assigned one at a time, rather than simultaneously)
 - ◆ Cluster seeds can be updated iteratively, or left fixed

30

Hierarchical or Nonhierarchical?

- Hierarchical (SAS Proc Cluster):
 - ◆ (+) Easy, intuitive
 - ◆ (-) Undesirable early clusters persist
 - ◆ (-) Sensitive to outliers, irrelevant variables
 - ◆ (-) Not suited to very large samples
- Nonhierarchical (SAS Proc Fastclus):
 - ◆ (+) Less sensitive to outliers, irrelevant variables
 - ◆ (+) Can handle a large number of cases
 - ◆ (-) Requires an objective choice of initial cluster seeds (random seeds perform poorly!) 31

Use Both!

- Use hierarchical to establish clusters and identify obvious outliers
 - ◆ save variable means in each cluster
- Eliminate outliers, then use nonhierarchical
 - ◆ use hierarchical cluster means as initial seeds
- Nonhierarchical stage can “fine-tune” the hierarchical results to reconfigure any undesirable clusters 32

How Many Clusters?

- No definitive answer
- “Easy rule”: look for large jumps in the similarity measure at each stage
 - ◆ implies we should maybe stop clustering at the previous stage
- More sophisticated rules offer little improvement
- Use practical judgment, common sense, or theoretical foundations
- Strive for parsimony, but avoid over-simplifying
 - ◆ consider “small” clusters very carefully 33

Interpreting Clusters

- Calculate cluster centroids (variable means)
- Compare differences between centroids
 - ◆ from a practical viewpoint
 - ◆ Statistically using ANOVA F-test (don’t interpret F-values literally since the procedure tries to form clusters that do differ, but their relative size does provide information about each variable’s contribution).
- Use the (important) differences to describe or label the clusters 34

Validating Cluster Solution

- Attempt to assure the solution is generalizable to a population
 - ◆ Split sample into two, run a cluster analysis on each part, and compare solutions
- Establish predictive validity
 - ◆ Select variable(s) *not* used to form clusters, but expected to vary across clusters
 - ◆ Statistically test for differences between clusters (ANOVA F-test with metric variables, Pearson Chi-square with non-metric) 35

Profiling Cluster Solution

- Describe characteristics of each cluster using variables *not* used in the cluster analysis itself
 - ◆ these variables are typically demographics, psychographics, etc.
- Use **discriminant analysis** to understand cluster differences and predict cluster membership
 - ◆ Cluster analysis (on conjoint part-worths!) can establish market segments, then discriminant analysis can predict segment membership based on demographics 36

Cluster Analysis Examples

- Jobs Example
 - ◆ jobspw1 (7 cases, 2 variables)
 - ◆ jobspw2 (22 cases, 8 variables)
 - ◆ jobspw3 (jobspw2 + demographics)
- Illustrative example in textbook (p502-515)
 - ◆ hatcocluster

37