Empirical Article on Clustering
Introduction to “Model” Based Methods

Clustering and Classification
Lecture 10
Today’s Class

• Review of Morris et al. (1998).

• Introduction to clustering with statistical models.
  – Background of Latent Class Analysis
    • One type of Finite Mixture Model.
Subtypes of Reading Disability: Variability Around a Phonological Core

Morris et al. (1998)
Background Issues

- Researchers have believed that children with reading disability are a heterogeneous population.
  - Because of heterogeneity, research is splintered
    - Many hypotheses tested.
    - Many inconsistent findings.

- Article attempts to define homogeneous groups of children with reading disabilities.
Previous Attempts to Identify Subtypes

- Classification based on IQ discrepancies has been widely questioned due to failure to demonstrate the ecological validity of the result.

- Multivariate methods have not lead to reliable results.
  - Consistency in groupings a problem.
Author-identified Problems With Most Classification Based Studies in Field

• The authors state that a “successful classification study” comes from a theoretical framework leading to:
  – A priori hypotheses about classification.
  – Selection of attributes that best represent these hypotheses.
  – Specification of analyses to evaluate how the groups differ from one another.
Study Conceptualization

• Three subtypes of phonological reading disability:
  – Phonological awareness
  – Phonology-verbal short term memory
  – General cognitive
Study Design

- “Cognitive measures selected according to contemporary hypotheses addressing relationship of language and reading skills” (p. 350), with measures of:
  - Phonological awareness
  - Naming skills
  - Vocabulary-lexical skills
  - Morphosyntactic ability
  - Speech production and perception
  - Verbal memory

- Nonverbal measures (thought to be weakly associated to reading ability):
  - Nonverbal memory
  - Visuospatial skills
  - Visual attention

- Additionally, a systematic assessment of the consistency and reliability of the identified subtypes was used.
  - Validation was thought of and demonstrated!!!
Participant Selection

• A heterogeneous sample of children was selected:
  – disability in reading
  – disability in math
  – disability in math and reading

• Contrast groups for all three (without disability).

• Sample ranged broadly in achievement and intellectual levels.
  – Was this way to minimize any a priori beliefs about learning disability.
Hold-out Sample

- To check the stability of the clustering solution, a hold out sample was created.

- This sample was not used in the original analysis, only used once groups were formed.

- The hold out sample consisted of children in the reading disability and nondisabled groups.
  - Math disability and ADHD were held out.
Measures Used To Classify

• Eight measures were used to classify children
  – The eight were selected on the basis of a CFA onto characteristics of important factors.
  – Measures were age-adjusted and standardized.

• Eight measures were then used to validate the classification.
  – Matched the factors of the original eight measures.
Measures Used To Classify

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Measures Selected as Classification Variables and Alternatives by Factor and Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
<td>Number</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>1.</td>
</tr>
<tr>
<td>2.</td>
<td>2.</td>
</tr>
<tr>
<td>3.</td>
<td>3.</td>
</tr>
<tr>
<td>4.</td>
<td>4.</td>
</tr>
<tr>
<td>5.</td>
<td>5.</td>
</tr>
<tr>
<td>6.</td>
<td>6.</td>
</tr>
<tr>
<td>7.</td>
<td>7.</td>
</tr>
<tr>
<td>8.</td>
<td>8.</td>
</tr>
</tbody>
</table>

*Note. WISC–R = Wechsler Intelligence Scale for Children—Revised.
Additional Measures Used To Externally Validate Result

- Additionally, six measures from different domains of information was used to validate result.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Test</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socio-demographic</td>
<td>Yale Children's Inventory</td>
<td>Age, gender, socioeconomic status</td>
</tr>
<tr>
<td>Developmental history</td>
<td>Yale Children's Inventory</td>
<td>Prenatal problems, perinatal problems, early reading problems, clumsy</td>
</tr>
<tr>
<td></td>
<td>Early Development</td>
<td>Learning disability, language disability, hyperactivity, attention deficit/hyperactivity disorder</td>
</tr>
<tr>
<td></td>
<td>Identifed Problems</td>
<td>Academic problems (father, mother)</td>
</tr>
<tr>
<td></td>
<td>Family History</td>
<td>Research problems (father, mother, siblings)</td>
</tr>
<tr>
<td>Intellligence</td>
<td>Webster Intelligence Scale for Children—Revised</td>
<td>Verbal IQ, performance IQ</td>
</tr>
<tr>
<td>Academic</td>
<td>Wide Range Achievement Test—Revised</td>
<td>Reading, spelling, arithmetic</td>
</tr>
<tr>
<td></td>
<td>Gray Oral Reading Test—Revised</td>
<td>Reading rate, accuracy, comprehension</td>
</tr>
<tr>
<td></td>
<td>Formal Reading Inventory</td>
<td>Silent reading comprehension</td>
</tr>
<tr>
<td>Language</td>
<td>Morphological Awareness and Syntactic Comprehension Test</td>
<td>Morphological awareness, Syntax comprehension</td>
</tr>
<tr>
<td></td>
<td>Peabody Picture Vocabulary Test—Revised</td>
<td>Receptive vocabulary</td>
</tr>
<tr>
<td></td>
<td>Boston Naming Test</td>
<td>Confrontation naming</td>
</tr>
</tbody>
</table>
| Motor visuo-motor           | Finger Tapping, Synkinesis, Finger Dexterity—Fine and粗
| Teacher ratings             | Gross, Development Test of Visual-Motor Integration                 | Academic scale                                                            |
|                             | Mult-Grade Inventory for Teachers                                   | Language scale                                                             |
|                             |                                                                       | Attention scale                                                            |
|                             |                                                                       | Activity scale                                                             |
|                             |                                                                       | Behavior scale                                                             |
|                             |                                                                       | Dexterity scale                                                            |

Overall Methods Used
Clustering Methods Used

- Ultrametric hierarchical clustering procedures (all agglomerative):
  - Ward’s method
  - Single link
  - Complete link

- K-means.
  - Used to “clarify and refine the initial solutions produced by the three hierarchical methods.”

- Used multiple starting points.

- Funny quote about clustering procedures (p. 354):
  - “These methods, although descriptive in nature and historically not founded in any significant mathematical theory, do have heuristic value and have been used in many scientific areas.”
Distance Measures Used

• The authors tried:
  – Squared Euclidean distance
  – Pearson correlation

• Both measures quantified the distance between each child in the sample

• The Pearson correlation technique did not yield consistent results – so they went with squared Euclidean distance.
Determination of Number of Clusters

• To decide the number of clusters, the authors examined several different measures:
  – Review of changes in between/within variability
  – Visual inspection of dendrogram
  – Inspection of cluster profiles as clusters were merged (averages of variables)
  – Visual inspection of individual child profiles within and across clusters.
Results of Hierarchical Clustering

• Looked at concordance of results across methods:
  – Total of 7-31 clusters examined.
  – Highest level of concordance between 7-12 cluster solutions
    • Concordance being greater than 80% agreement
  – Used concordance to indicate optimal number of clusters.
Applying K-means

• K-means clustering was applied to the solutions of each of the hierarchical procedures used.
  – Six procedures for each hierarchical method
    • ???
  – “Iterated down to a five-cluster solution”
    • ???

• Relocation methods resulted in 17 different solutions with 151 clusters.
  – I am not sure why or what was done here.
Reducing Clusters

• Because there were 17 different solutions and 151 different clusters, something had to be done to identify consistent clusters.

• Three raters sorted the mean profiles of the attributes based on visual similarity.

• Ten profiles (subtypes) were selected – occurred repeatedly across most of the 17 solutions.
Subtypes Identified

1. GD – Global Deficit
2. GL – Global Language.
3. PVR – Phonology – Verbal short-term Memory
4. PVL – Phonology – VSTM lexical
5. PVS – Phonology VSTM spatial
6. PR – Phonology – rate
7. RD – Rate – disabled
8. ND1 – Nondisabled
9. ND2 – Nondisabled
10. ND3 – Nondisabled
Classifying Children

• To classify each child, an index of group membership was formed:
  – For each of the 10 subtypes.
  – For each of the 17 solutions.

• Index was percentage of times child got classified into a subtype.

• Child was assigned to subtype with highest index, if value was greater than 0.7.
Those Not In Subtypes

• Of the 40 children with index values below 0.7:
  – 19 had low membership indices across multiple subtypes
    • Identified as outliers
  – 21 were placed within best matching subtype based on their index and profile of scores.
Analyses of Internal Validity

• Concordance was checked.
  – Not sure what was used.

• Holdout sample was added – reclustered using same procedures.
  – 73% – 88% of original children were in same cluster.
Conclusions about Internal Validity

• Final 10-subtype solution classified 92% of children from original sample

• When hold-out sample was added, 80% of children were classified
  – 20% were “outliers”
External Validity Checks

• To check external validity comparison of groups was made on second set of classification variables.

• Used discriminant analysis to do this.
  – 97% of children were correctly put into same clustering group with second set of variables.

• This is not a strong test – high correlation between sets of variables.
  – Used other variables to detect differences.
External Validity Checks

• Did a series of MANOVAs to detect differences between groups on alternative classification variables.
  – Found differences.

• Looked at external domain variables – found differences there, too.
Summary

• Methods described by Morris et al. (1998) present a cluster analysis that sought both internal and external verification of results.

• The analyses provided a wonderful description of the types of children with reading disabilities.

• What did you think?
Introduction to “Model” Based Clustering Techniques
Finite Mixture Models

• Finite mixture models are models that express a set of observable variables as a mixture (sum) of a set of distributions.

• The typical equation for such a mixture looks like:

\[ P(X) = \sum \pi_g f(X|g) = \pi_1 f(X|g=1) + \ldots + \pi_G f(X|g=G) \]
Finite Mixture Models

\[ P(X) = \sum \pi_g f(X|g) = \pi_1 f(X|g=1) + \ldots + \pi_G f(X|g=G) \]

- Here, \( X \) is the data matrix.
- \( g \) is the distribution (\( g=1,\ldots,G \)).
- \( f(X|g) \) is the statistical distribution of \( X \) given \( g \).
  - This can be, literally, anything.
- \( \pi_g \) is the so-called “mixing proportion” for group \( g \).
  - This represents the probability that any observation from the population represented by the sample comes from group \( g \).
What is this $g$ of which you speak?

• $g$ – is the group/class/distribution a population may come from.

• Bartholomew and Knott develop a nice way of looking at $g$ as a categorical latent variable.
  – They give a table (p. 3) that is a bit misleading for general FMM approaches, but works for the topics covered in their book.

• We will discuss these terms in the following weeks.
  – For now, consider the table complete
Example of a Mixture Model

• Imagine you were interested in the effects of heavy smoking on lung cancer.

• You are able to tell:
  – who is a heavy smoker (>1 pack per day)
  – who has lung cancer

• Now imagine you get your study approved by the human subjects committee, and you go out and collect the data on the next page.
### Smoking and Cancer Contingency Table

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Heavy Smoker</strong></td>
<td><strong>Not a Heavy Smoker</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>350</td>
<td>200</td>
<td>550</td>
</tr>
<tr>
<td>No Lung Cancer</td>
<td>150</td>
<td>300</td>
<td>450</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>500</td>
<td>1000</td>
</tr>
</tbody>
</table>
What About This Association?

• There appears to be a significant association between smoking and lung cancer.

• However, if there was a third variable lurking out there, this effect might be considered spurious.

<table>
<thead>
<tr>
<th>Cancer</th>
<th>Smoker 00</th>
<th>Smoker 100</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer 00</td>
<td>300</td>
<td>150</td>
<td>450</td>
</tr>
<tr>
<td>Cancer 100</td>
<td>200</td>
<td>350</td>
<td>550</td>
</tr>
<tr>
<td>Total</td>
<td>500</td>
<td>500</td>
<td>1000</td>
</tr>
</tbody>
</table>

Chi-Square Tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig (2-sided)</th>
<th>Exact Sig (2-sided)</th>
<th>Exact Sig (1-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>90.000</td>
<td>1</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuity Correction*</td>
<td>89.701</td>
<td>1</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>92.402</td>
<td>1</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fisher's Exact Test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>98.018</td>
<td>1</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Computed only for a 2x2 table
b. 0 cells (0%) have expected count less than 5. The minimum expected count is 225.00
The Hidden Third Variable

<table>
<thead>
<tr>
<th>Urban Environment</th>
<th>Non-Urban Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heavy Smoker</td>
<td>Heavy Smoker</td>
</tr>
<tr>
<td>Not a Heavy Smoker</td>
<td>Not a Heavy Smoker</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>Lung Cancer</td>
</tr>
<tr>
<td>320</td>
<td>30</td>
</tr>
<tr>
<td>80</td>
<td>120</td>
</tr>
<tr>
<td>400</td>
<td>150</td>
</tr>
<tr>
<td>No Lung Cancer</td>
<td>No Lung Cancer</td>
</tr>
<tr>
<td>80</td>
<td>70</td>
</tr>
<tr>
<td>20</td>
<td>280</td>
</tr>
<tr>
<td>100</td>
<td>350</td>
</tr>
<tr>
<td>400</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>400</td>
</tr>
<tr>
<td>500</td>
<td>500</td>
</tr>
</tbody>
</table>

- Notice how the original association has now changed (or vanished)?
What The Example Means

• What we are trying to demonstrate is the idea that we can try to parse out groups from our data.
  – Just like all of our clustering methods.

• Only here, we will say that certain groups have distributions for the variables that differentiate themselves from other groups.
  – Here the non-urban group’s distribution of the two variables was different from the urban group’s distribution.
  – The exact form of the distribution may differ, too (although here it did not).
Where We Are Going

- Over the course of the next few weeks, we will learn about FMM, using differing distributions.

- Perhaps the easiest case to learn is that of Latent Class Analysis (LCA).
  - LCA works with categorical manifest variables.
  - Here the variables are assumed to be independent within group.
After LCA

- After LCA, we will switch to Latent Profile Analysis (LPA):
  - In LPA, the manifest variables are now assumed to be “metrical”
  - Each distribution within group is considered MVN.
  - Independence within group, however, still holds.

- After LCA and LPA, we will then move to more general mixture models.
  - Differing distributions
  - Differing assumptions about covariance within group.
What Can FMM Do For You?

• FMM can be used to:
  – Identify groups of people differing on sets of variables.
    • Similar to our clustering methods.
  – Identify outliers in your data.
  – Provide goodness of fit of some (possibly none-mixture method) to your data.
    • What proportion of cases would you have to throw away to fit perfectly?
Additional FMM Fun

• Because distributional assumptions are involved in FMM we can:
  – Use likelihood-based methods to fit models.
    • EM
    • MCMC
    • Method of Moments (like SEM)
    • Minimization-optimization of Log Likelihood
  – Attempt to validate our results by generating data assuming our model is true.
    • See picture on next page for fun result of a mixture model.
When FMM Go Bad

RUM Total Test Score Plot

Proportion Observed

Total Test Score

1-observed
2-model estimated
Next Time

• Specifics of latent class analysis.

• How to do LCA.