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 relation ship 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 2

Measures Selected as Classification Variables and Alternatives by Factor and Test

Factor		Test	
Number	Description	Subtyping	Reliability & Validity
1.	Phonological awareness	Auditory Analysis Test[a]	Embedded Phonemes—Neutral Foils[b]
2.	Verbal Short-Term Memory	Word String Recall—nonrhyme ordered[c]	WISC–R Digit Span[d]
3.	Rapid Naming	Rapid Naming Subtest 1[e]	Rapid Naming Subtest 3[e]
4.	Lexical/Vocabulary	WISC–R Similarities[d]	WISC–R Vocabulary[d]
5.	Speech Production	Speed of Articulation[f]	Tongue Twisters—Easy[g]
6.	Visual–Spatial	Judgement of Line Orientation[h]	WISC–R Block Design[d]
7.	Visual Attention	Underlining—subtest 3[i]	Underlining—subtest 2[i]
8.	Nonverbal Short-Term Memory	Corsi Blocks—ordered[j]	Corsi Blocks—Total[j]

Note. WISC–R = Wechsler Intelligence Scale for Children—Revised.
[a]Rosner & Simon (1971). [b]Fowler (1990). [c]Shankweiler, Liberman, Mark, Fowler, & Fisher (1979). [d]Wechsler (1974). [e]Katz & Shankweiler (1985). [f]Hulme, Thomson, Muir, & Lawrence (1984). [g]Kupin (1979); Rapala & Brady (1990). [h]Lindgren & Benton (1980). [i]Doehring (1968). [j]Milner (1971).

Additional Measures Used To Externally Validate Result

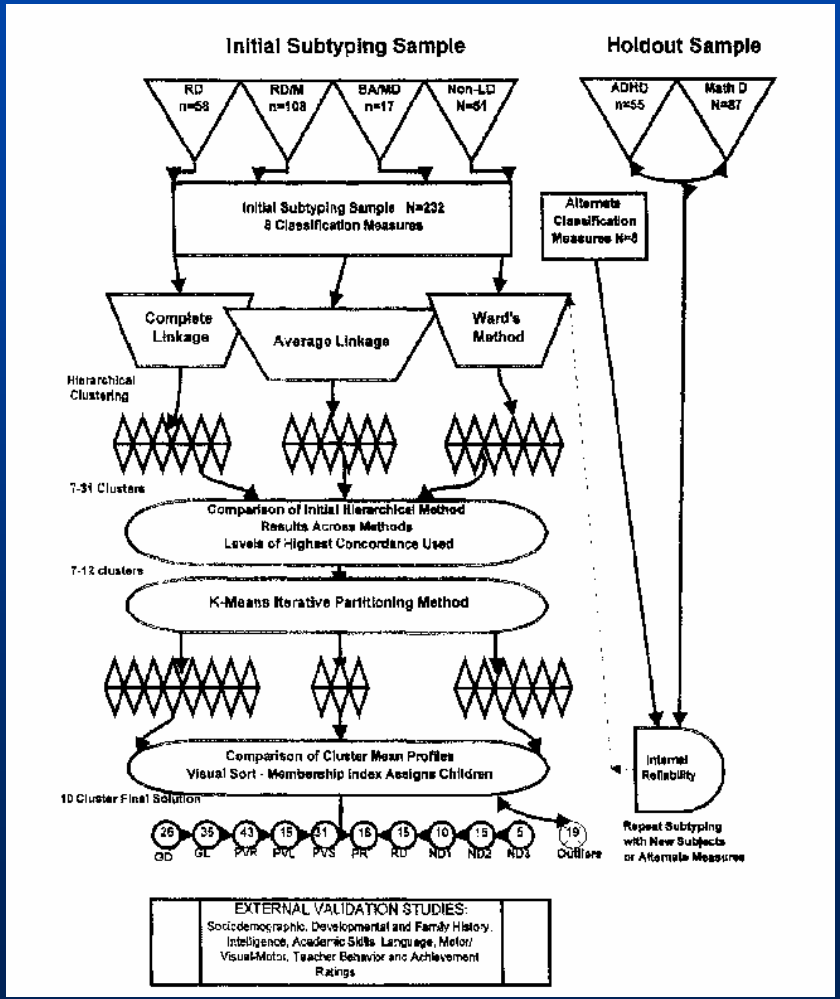
- Additionally, six measures from different domains of information was used to validate result.

Table 4
Measures Used for External Validity Analyses

Domain	Test	Variables
Sociodemographic	Yale Children's Inventory[a]	Age, gender, socioeconomic status[b]
Developmental history	Yale Children's Inventory[a]	
	Early Development	Prenatal problems, perinatal problems, early reading problems, clumsy
	Identified Problems	Learning disability, language disability, hyperactivity, attention deficit/hyperactivity disorder
	Family History	Academic problems (father, mother)
		reading (father, mother, siblings)
		language (father, mother)
		math (father, mother)
		hyperactivity (parents, siblings)
Intelligence	Wechsler Intelligence Scale for Children—Revised[c]	Verbal IQ, performance IQ
Academic	Wide Range Achievement Test—Revised[d]	Reading, spelling, arithmetic
	Gray Oral Reading Test—Revised[e]	Reading rate, accuracy, comprehension
	Formal Reading Inventory[f]	Silent reading comprehension
Language	Morphological Awareness and Syntactic Comprehension Test[g]	Morphological awareness
		Syntax comprehension
	Peabody Picture Vocabulary Test—Revised[h]	Receptive vocabulary
	Boston Naming Test[i]	Confrontation naming
Motor visuomotor	Finger Tapping, Synkinesis, Finger Dexterity—Fine and Gross, Developmental Test of Visual-Motor Integration[j]	
Teacher ratings	Multi-Grade Inventory for Teachers[k]	Academic scale
		Language scale
		Attention scale
		Activity scale
		Behavior scale
		Dexterity scale

[a]Shaywitz, Holahan, Marchione, Sadler, and Shaywitz (1992). [b]Hollingshead and Redlich (1958). [c]Wechsler (1974). [d]Jastak & Wilkinson (1984). [e]Wiederholt and Bryant (1988). [f]Wiederholt (1986). [g]Shankweiler et al. (1995). [h]Dunn and Dunn (1981). [i]Kaplan, Goodglass, and Weintraub (1983). [j]Beery (1982). [k]Shaywitz, Escobar, Shaywitz, Fletcher, and Makuch (1992).

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 varaibles.
  - 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:

$$P(\mathbf{X}) = \Sigma\, \pi_g\, f(\mathbf{X}|g) = \pi_1\, f(\mathbf{X}|g=1) + \ldots + \pi_G\, f(\mathbf{X}|g=G)$$

Finite Mixture Models

$$P(\mathbf{X}) = \Sigma\, \pi_g\, f(\mathbf{X}|g) = \pi_1\, f(\mathbf{X}|g=1) + \ldots + \pi_G\, f(\mathbf{X}|g=G)$$

- Here, **X** is the data matrix.
- g is the distribution (g=1,…,G).
- f(**X**|g) is the statistical distribution of **X** given g.
  - This can be, literally, anything.
- π_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

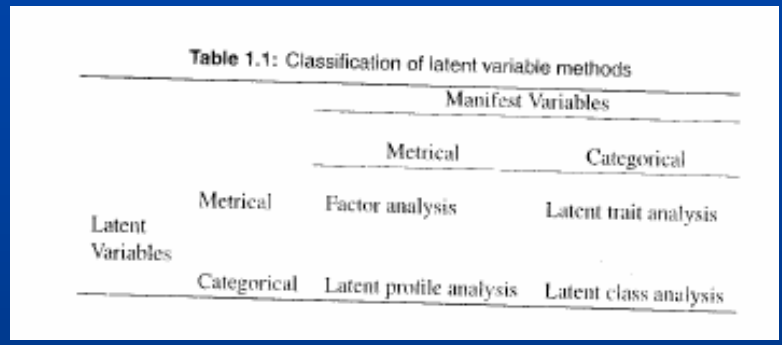


Table 1.1: Classification of latent variable methods

		Manifest Variables	
		Metrical	Categorical
Latent Variables	Metrical	Factor analysis	Latent trait analysis
	Categorical	Latent profile analysis	Latent class analysis

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

	Heavy Smoker	Not a Heavy Smoker	
Lung Cancer	350	200	550
No Lung Cancer	150	300	450
	500	500	1000

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.

cancer * smoker Crosstabulation

Count

		smoker		
		.00	1.00	Total
cancer	.00	300	150	450
	1.00	200	350	550
Total		500	500	1000

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	90.909[b]	1	.000		
Continuity Correction[a]	89.701	1	.000		
Likelihood Ratio	92.402	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	90.818	1	.000		
N of Valid Cases	1000				

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

Urban Environment			
	Heavy Smoker	Not a Heavy Smoker	
Lung Cancer	320	80	400
No Lung Cancer	80	20	100
	400	100	500

Non-Urban Environment			
	Heavy Smoker	Not a Heavy Smoker	
Lung Cancer	30	120	150
No Lung Cancer	70	280	350
	100	400	500

- 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.

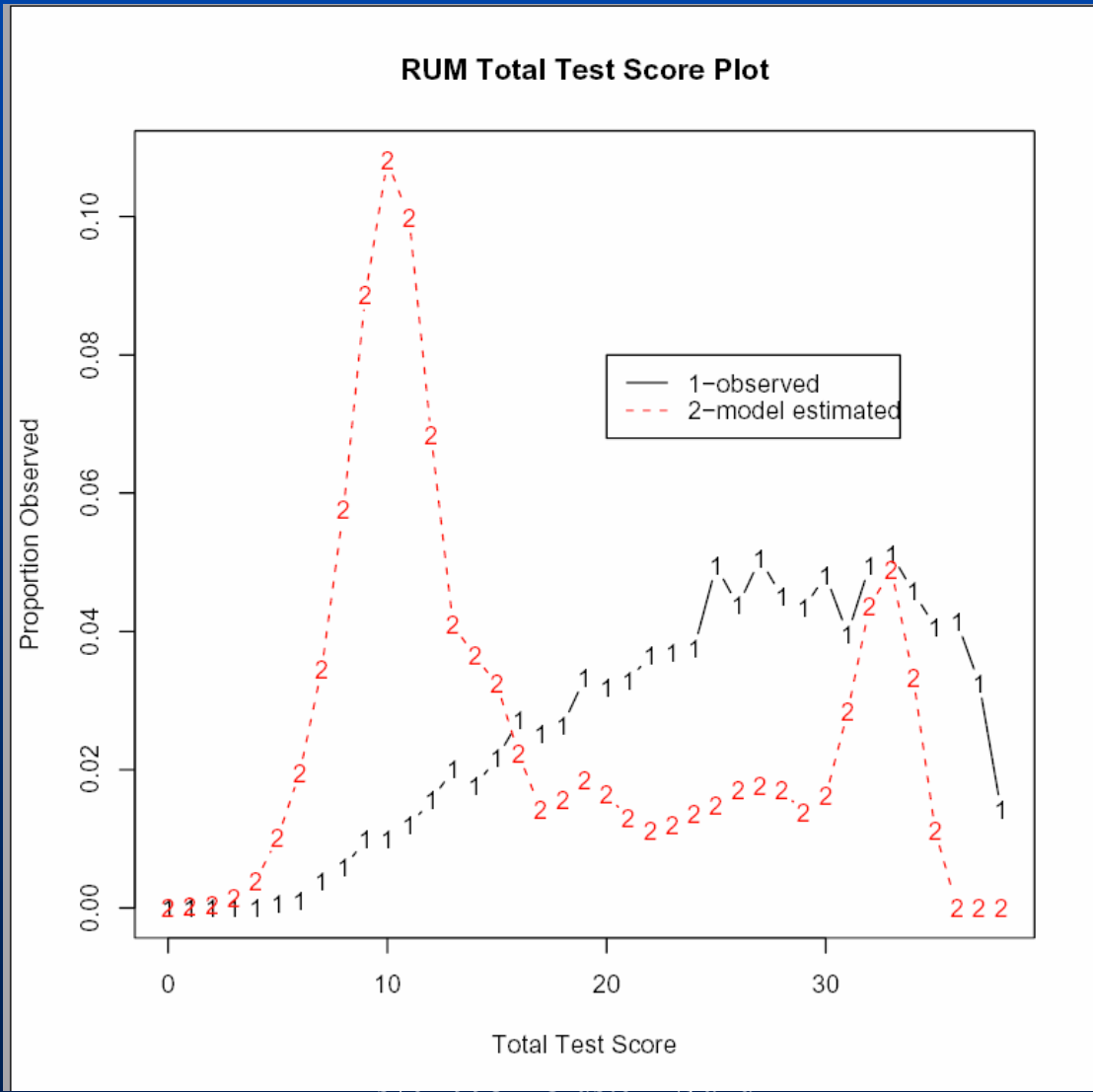
What Can FMM Do For You?

- 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



Next Time

- Specifics of latent class analysis.
- How to do LCA.