- Data Mining Principles
- MSc in Analytics

- Dr. Anil D Chaturvedi
- Email: anilchaturvedi@uchicago.edu
- Phone: 301-299-2434

THE UNIVERSITY OF CHICAGO

- Good Morning.
- Welcome to Data Mining Principles
- Session 3

THE UNIVERSITY OF CHICAGO

Data Mining Principles
Session 3: Topics

1. Clustering Categorical Data
2. K-modes clustering
3. Latent Class Analysis
4. Hybrid Data - mixtures
5. Conclusion

Examples of Bilinear models: Row Reduction via Cluster Analysis

$$d_{ij} \approx \sum_{r=1}^{R} m_{ir} c_{jr}$$

$$\mathrm{D}_{IxJ} \approx \mathrm{M}_{IxR} \mathrm{C}_{RxJ}$$

- Only categorical Data D_{ij} are known
- Cluster memberships M are unknown
- Cluster centers C are unknown
- M is constrained to be binary: 1 or 0

Examples of Bilinear models:
Row Reduction: K-Modes Clustering

$$d_{ij} \approx \sum_{r=1}^{R} m_{ir} c_{jr}$$

Minimize $$\left| D_{IxJ} - M_{IxR} C_{RxJ} \right|^{p,\, p \to 0}$$

- **Only Categorical Data D_{ij} are known**
- **Cluster memberships M are unknown**
- **Cluster centers C are unknown**
- **M is constrained to be binary: 1 or 0**
- **Each Row of M sums to 1.**

K-Modes: The procedure

1. Randomly assign a number, from 1 to *K (fixed at start), to each of the observations or records. Fix M.* These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   A. For each of the *K clusters, compute the cluster mode. The kth cluster centroid is the vector of the p variable modes for the* observations in the *kth cluster.* ***Find C given M***
   B. Assign each observation to the cluster whose centroids match the most categorical values of the observation. ***Find M given C.***

Find **C** Given **M**

Find **M** Given **C**

Why Means in step 2A?

- Assume you have 6 data points in a variable: Male, Female, Male, Female, Female, Female.
- What is the best summary statistic that describes such categorical data?
- Can you match your intuition with mathematical logic?

K-modes clustering

$$\text{Minimize} |3-x|^{0.0001} + |5-x|^{0.0001} + |3-x|^{0.0001} + |3-x|^{0.0001} + |5-x|^{0.000}$$

K-modes clustering

$$\text{Minimize}\,|3-x|^{0.0001}+|5-x|^{0.0001}+|3-x|^{0.0001}+|3-x|^{0.0001}+|5-x|^{0.0001}$$

- Minimization yields x to be the mode of the numbers (3,5,3,3,5) = 3. Why? Because

- $0^0 = 1$ $\qquad$ $0^{0.0001} = 0$
- $3^0 = 1$ $\qquad$ $3^{0.0001} = 1$
- $5^0 = 1$ $\qquad$ $5^{0.0001} = 1$

K-Modes: The procedure

Minimize $\left| D_{IxJ} - M_{IxR} C_{RxJ} \right|^{p,p \to 0}$

Find **C**
Given **M**

$$\begin{bmatrix} 1 & 67 & 0 \\ 1 & 67 & 0 \\ 2 & 77 & 5 \\ 1 & 67 & 5 \\ 2 & 77 & 3 \end{bmatrix} \approx \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 67 & 0 \\ 1 & 67 & 0 \\ 2 & 77 & 5 \\ 1 & 67 & 5 \\ 2 & 77 & 3 \end{bmatrix} \approx \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{21} & c_{22} & c_{23} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

K-Modes: The procedure

Minimize $\left| \mathrm{D}_{IxJ} - \mathrm{M}_{IxR}\mathrm{C}_{RxJ} \right|^{p, p \to 0}$

Find **C**
Given **M**

$$\begin{bmatrix} 1 & 67 & 0 \\ 1 & 67 & 0 \\ 2 & 77 & 5 \\ 1 & 67 & 5 \\ 2 & 77 & 3 \end{bmatrix} \approx \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{21} & c_{22} & c_{23} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

To estimate c_{21}

$$\text{Minimize} \,|\, 2 - \mathrm{c}_{21} \,|^{0.0001} + |\, 1 - \mathrm{c}_{21} \,|^{0.0001} + |\, 2 - \mathrm{c}_{21} \,|^{0.0001}$$

To get c_{21} = mode (2,1,2) = 2. Similarly solve for all c's

K-Modes: The procedure

Minimize $\left| D_{IxJ} - M_{IxR} C_{RxJ} \right|^{p, p \to 0}$

Find **M**
Given **C**

$$\begin{bmatrix} 1 & 67 & 0 \\ 1 & 67 & 0 \\ 2 & 77 & 5 \\ 1 & 67 & 5 \\ 2 & 77 & 3 \end{bmatrix} \approx \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \\ m_{31} & m_{32} \\ m_{41} & m_{42} \\ m_{51} & m_{52} \end{bmatrix} \begin{bmatrix} 1 & 67 & 0 \\ 2 & 77 & 5 \end{bmatrix}$$

First Row

$$\begin{bmatrix} 1 & 67 & 0 \end{bmatrix} \approx \begin{bmatrix} m_{11} & m_{12} \end{bmatrix} \begin{bmatrix} 1 & 67 & 0 \\ 2 & 77 & 5 \end{bmatrix}$$

K-Modes: The procedure

Minimize $\left| D_{IxJ} - M_{IxR} C_{RxJ} \right|^{p, p \to 0}$

First Row $\begin{bmatrix} 1 & 67 & 0 \end{bmatrix} \approx \begin{bmatrix} m_{11} & m_{12} \end{bmatrix} \begin{bmatrix} 1 & 67 & 0 \\ 2 & 77 & 5 \end{bmatrix}$

$$|1 - 1m_{11} - 2m_{12}|^{0.0001} + |67 - 67m_{11} - 77m_{12}|^{0.0001} + |0 - 0m_{11} - 5m_{12}|^{0.0001}$$

- Try both (m_{11} = 0 and m_{12} = 1) and (m_{11}= 1 and m_{12} = 0).
- Whichever yields minimum, that is the membership assignment

K-Modes: The procedure

Minimize $\left| D_{IxJ} - M_{IxR} C_{RxJ} \right|^{p,p\to 0}$

Find **C**
Given **M**

$$\begin{bmatrix} M & Young & High \\ M & Young & High \\ F & Old & Low \\ F & Young & Low \\ M & Old & High \end{bmatrix} \approx \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

$$\begin{bmatrix} M & Young & High \\ M & Young & High \\ F & Old & Low \\ F & Young & Low \\ M & Old & High \end{bmatrix} \approx \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{21} & c_{22} & c_{23} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

K-Modes: The procedure

Minimize $\left| D_{IxJ} - M_{IxR} C_{RxJ} \right|^{p,\,p\to 0}$

Find **C**
Given **M**

$$\begin{bmatrix} M & Young & High \\ M & Young & High \\ F & Old & Low \\ F & Young & Low \\ M & Old & High \end{bmatrix} \approx \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{21} & c_{22} & c_{23} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

To estimate c_{21}

$$\text{Minimize} \, | F - c_{21} |^{0.0001} + | F - c_{21} |^{0.0001} + M - c_{21} |^{0.0001}$$

To get c_{21} = mode (F,F,M) = F.

K-Modes: The procedure

Minimize $\left| D_{IxJ} - M_{IxR} C_{RxJ} \right|^{p, p \to 0}$

Find **M**
Given **C**

$$\begin{bmatrix} M & Young & High \\ M & Young & High \\ F & Old & Low \\ F & Young & Low \\ M & Old & High \end{bmatrix} \approx \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \\ m_{31} & m_{32} \\ m_{41} & m_{42} \\ m_{51} & m_{52} \end{bmatrix} \begin{bmatrix} M & Young & High \\ F & Old & Low \end{bmatrix}$$

First Row

$$\begin{bmatrix} M & Young & High \end{bmatrix} \approx \begin{bmatrix} m_{11} & m_{12} \end{bmatrix} \begin{bmatrix} M & Young & High \\ F & Old & Low \end{bmatrix}$$

K-Modes: The procedure

Minimize $\left| \mathrm{D}_{IxJ} - \mathrm{M}_{IxR}\mathrm{C}_{RxJ} \right|^{p,p\to 0}$

First Row $[M \quad Young \quad High] \approx [m_{11} \quad m_{12}] \begin{bmatrix} M & Young & High \\ F & Old & Low \end{bmatrix}$

- Try both ($m_{11} = 0$ and $m_{12} = 1$) and ($m_{11} = 1$ and $m_{12} = 0$).
- Whichever yields minimum, that is the membership assignment

ANY QUESTIONS?

K-Modes clustering: Election Data set

Package poLCA in R implements Latent Class Analysis Election Data.

- Education
  - (1) 8 grades or less
  - (2) 9-11 grades, no further schooling
  - (3) High school diploma or equivalency
  - (4) More than 12 years of schooling, no higher degree
  - (5) Junior or community college level degree
  - (6) BA level degrees, no advanced degree
  - (7) Advanced degree
- Gender
  - (1) Male
  - (2) Female
- Party
  - (1) Strong Democrat
  - (2) Weak Democrat
  - (3) Independent-Democrat
  - (4) Independent-Independent
  - (5) Independent-Republican
  - (6) Weak Republican
  - (7) Strong Republican

K-Modes clustering: Election Data set

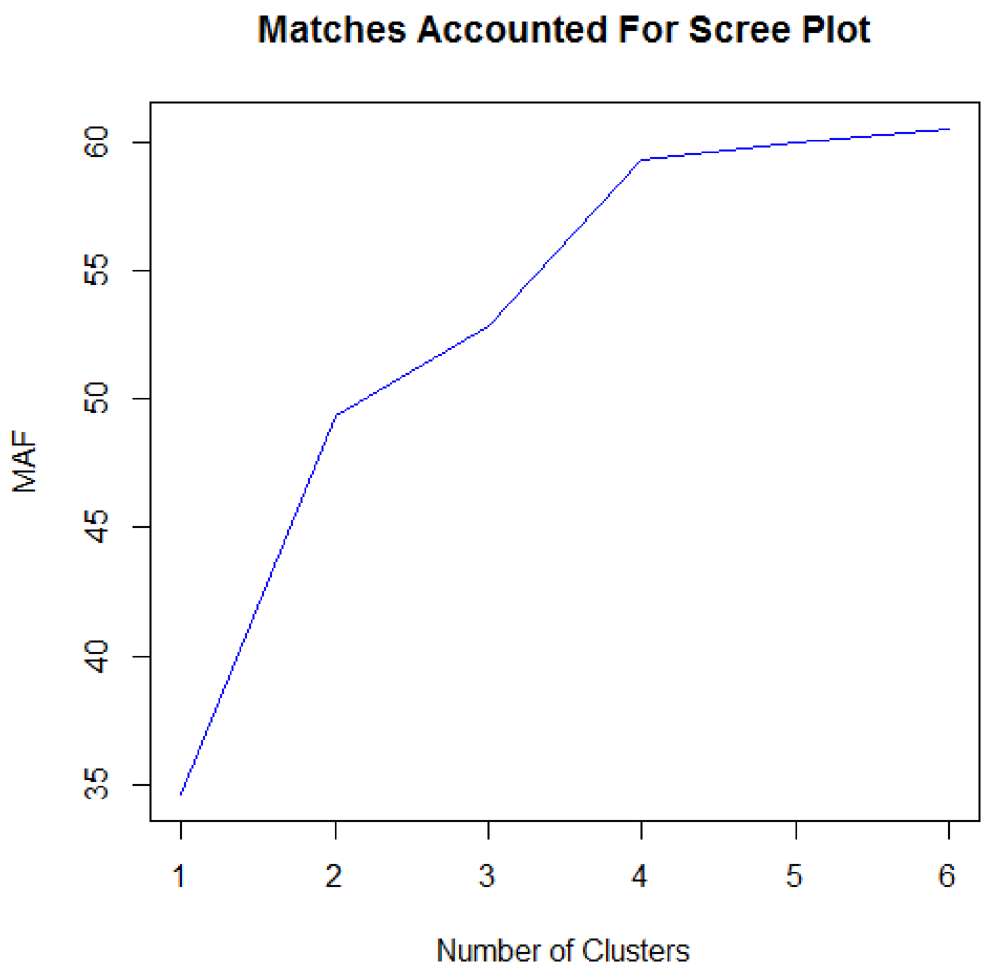
#	EDUC	GENDER	PARTY
1	5	1	5
2	4	2	3
3	3	2	1
4	4	1	3
5	5	2	7
6	2	1	1
7	4	1	6
8	7	2	1
9	6	2	1
10	3	2	1
11	3	2	6
12	3	1	2
13	3	1	3
14	3	2	5
15	4	2	4
16	4	1	7
17	2	2	6
18	4	2	2
19	3	2	1
20	6	2	7

K-modes clustering: Election Data set

Number of Clusters	Matches Accounted For
1	34.64
2	**49.32**
3	52.83
4	59.30
5	60.04
6	60.49

MAF = Total Data values matched/Number of Data elements

K-Modes clustering: Election Data set Scree Plot 3 Variables: Gender, Education, and Party



K-Modes clustering: Election Data set
Cluster Modes

Cluster	Male	Female	Cluster Size
1	100%	0	786
2	0	100%	999

BUT NOW LET US USE THE MORE INTERESTING CASE: USE ALL CATEGORIAL VARIABLES

K-Modes clustering: Election Data

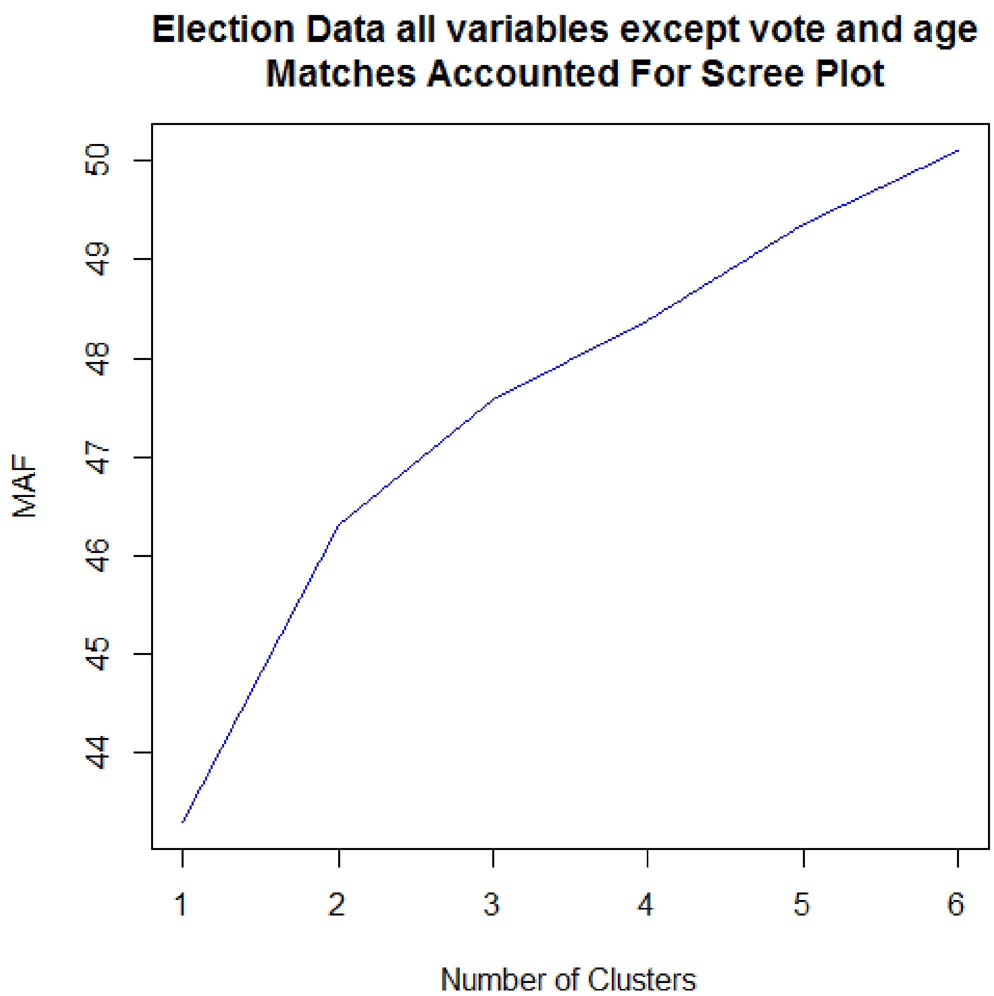
```
x=kmodes(data=election[,-c(13,14)],nclust=6, nloops=30,seed=123121)
MAF2=c(43.31,46.32,47.58,48.38,49.36,50.10)
plot(1:6,MAF2,main="Matches Accounted For Scree Plot", xlab = "Number of Clusters",
ylab="MAF",col=4,type="l")
```

K-modes clustering: Election Data set Using all variables except Vote and Age

Number of Clusters	Matches Accounted For
1	43.31
2	46.32
3	47.58
4	48.38
5	49.36
6	**50.10**

MAF = Total Data values matched/Number of Data elements

K-Modes clustering: Election Data set (All Variables) Scree Plot for selecting number of clusters



K-Modes clustering: Election Data set
Cluster Modes

Cluster	1	2	3	4	5	6
Cluster Size	34.60%	17.90%	6.50%	12.20%	8.30%	14.60%
MORALG	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well
CARESG	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well
KNOWG	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well
LEADG	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well
DISHONG	4 Not well at all	3 Not too well	4 Not well at all	3 Not too well	3 Not too well	3 Not too well
INTELG	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well
MORALB	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well
CARESB	3 Not too well	2 Quite well	2 Quite well	2 Quite well	3 Not too well	2 Quite well
KNOWB	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well
LEADB	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well
DISHONB	3 Not too well	3 Not too well	3 Not too well	4 Not well at all	3 Not too well	3 Not too well
INTELB	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well	2 Quite well
EDUC	HS	HS	HS	HS	HS	HS
GENDER	Female	Female	Female	Female	Female	Male
PARTY	SD	WD	SD	SD	SD	IR

K-Means clustering
Watch-outs and Properties

1. K-Modes shares all the strengths and weaknesses of K-means and K-medians.
2. Suffers from severe local optima problems. Never know if we got a global solution or not. (But this is a shared problem with almost every known clustering methodology)
3. Intractable with Big Data.

ANY QUESTIONS ON K-MODES?

Latent Class Analysis

1. Developed in 1950 by Lazarsfeld

2. Widely used currently in many domains

   - Business and marketing
   - Medicine
   - Healthcare
   - Services
   - ….

Lazarsfeld's Latent Class: The Basics

Cluster	Read A	Did Not Read A	Total
Read B	260	140	400
Did Not Read B	240	360	600
Total	500	500	1000

```
chisq.test(matrix(c(260,240,140,360),2,2),correct=FALSE)

        Pearson's Chi-squared test

data:  matrix(c(260, 240, 140, 360), 2, 2)
X-squared = 60, df = 1, p-value = 9.486e-15
```

Latent Class: The Basics

Cluster	Read A	Did Not Read A	Total
Read B	260	140	400
Did Not Read B	240	360	600
Total	500	500	1000

- Strong Dependence between readers of A and readers of B'
  - 52% of A's readers also read B
  - Only 28% of Non-readers of A read B

Latent Class: The Basics

Cluster	Read A	Did Not Read A	Total
Read B	260	140	400
Did Not Read B	240	360	600
Total	500	500	1000

- How many groups do we see in the Data?
- We all can see 4. Right?
- Can they explain dependence?
- Why or why not?

Latent Class: The Basics

Cluster	Read A	Did Not Read A	Total
Read B	260	140	400
Did Not Read B	240	360	600
Total	500	500	1000

- Lazarsfeld (1950, 1968) postulated
  - two latent or hidden groups – within this data – **related to a missing variable**
  - Each group with its own 2x2 table
  - Each table where A and B are independent

Latent Class: The Basics

Cluster	Read A	Did Not Read A	Total
Read B	260	140	400
Did Not Read B	240	360	600
Total	500	500	1000

	Read A	Did Not Read A	Total
Read B	240	60	300
Did Not Read B	160	40	200
Total	400	100	500

	Read A	Did Not Read A	Total
Read B	20	80	100
Did Not Read B	80	320	400
Total	100	400	500

Latent Class: The Basics

	Read A	Did Not Read A	Total
Read B	260	140	400
Did Not Read B	240	360	600
Total	500	500	1000

High Educ	Read A	Did Not Read A	Total
Read B	240	60	300
Did Not Read B	160	40	200
Total	400	100	500

Low Educ	Read A	Did Not Read A	Total
Read B	20	80	100
Did Not Read B	80	320	400
Total	100	400	500

Latent Class: The Basics

High Educ	Read A	Did Not Read	Total
Read B	240	60	**300**
Did Not Read B	160	40	**200**
Total	**400**	**100**	**500**

Low Educ	Read A	Did Not Read A	Total
Read B	20	80	**100**
Did Not Read B	80	320	**400**
Total	**100**	**400**	**500**

High Educ	Read A	Did Not Read	Total
Read B			**300**
Did Not Read B			
Total	**400**		**500**

Low Educ	Read A	Did Not Read A	Total
Read B			**100**
Did Not Read B			
Total	**100**		

Latent Class: The Basics

High Educ	Read A	Did Not Read	Total
Read B	240	60	300
Did Not Read B	160	40	200
Total	400	100	500

Low Educ	Read A	Did Not Read A	Total
Read B	20	80	100
Did Not Read B	80	320	400
Total	100	400	500

High Educ	Read A	Did Not Read	Total
Read B			300
Did Not Read B			
Total	400		500

Low Educ	Read A	Did Not Read A	Total
Read B			100
Did Not Read B			
Total	100		

- LCA is about finding the 5 numbers

Latent Class: The Basics

High Educ	Read A	Did Not Read	Total
Read B	240	60	300
Did Not Read B	160	40	200
Total	400	100	500

Low Educ	Read A	Did Not Read A	Total
Read B	20	80	100
Did Not Read B	80	320	400
Total	100	400	500

High Educ	Read A	Did Not Read	Total
Read B			300
Did Not Read B			
Total	400		500

Low Educ	Read A	Did Not Read A	Total
Read B			100
Did Not Read B			
Total	100		

- LCA is about finding the 5 numbers (parameters)

Latent Class: The problem

Cluster	Read A	Did Not Read A	Total
Read B	260	140	400
Did Not Read B	240	360	600
Total	500	500	1000

Given the categorical data above,

(A) how do we find K latent or hidden groups (missing data) wherein the multi-level tables are statistically independent?

AND

(B) Put individual people into the K groups

Latent Class: The Basics

High Educ	Read A	Did Not Read	Total
Read B	240	60	300
Did Not Read B	160	40	200
Total	400	100	500

Low Educ	Read A	Did Not Read A	Total
Read B	20	80	100
Did Not Read B	80	320	400
Total	100	400	500

High Educ	Read A	Did Not Read	Total
Read B			.6
Did Not Read B			.4
Total	.8	.2	.5

Low Educ	Read A	Did Not Read A	Total
Read B			.2
Did Not Read B			.8
Total	.2	.8	.5

- LCA is about finding the 5 numbers (parameters)

Latent Class: Estimation via EM Algorithm

Latent Classes are estimated by the Classic, Nobel-winning caliber work: EM-Algorithm of Dempster, Laird, and Rubin (1977) – The Expectation-Maximization (EM) Algorithm.

- Based on Maximum Likelihood
- Applicable to Missing Data problems
- Treats latent classes as missing data
- Is a two step algorithm:
  - The Expectation Step
  - The Maximization Step

Latent Class Analysis: EM Algorithm

THE E-STEP
(Assigns observations to classes)

- **Assume that all parameters are known from M step or chosen at random. Then apply the following two steps:**
  - **Using the parameters and Bayes rule, compute the posterior probability of <u>each</u> observation being in <u>each</u> class.**
  - **Take the expectation (mean) of all the observation’s posterior probabilities for a class. This gives the size (in proportions) of each class.**

Latent Class Analysis: EM Algorithm

THE M-STEP
(Finds the classes)

- **Assume that all posterior probabilities for all observations are known (and hence, latent class sizes or proportions are known) from E step. Then apply the following step**
  - **Maximize the Likelihood function (or the Log-likelihood) function, to estimate parameters (marginal probabilities of all variables for all classes). Often - using non-linear optimization techniques such as**
    - **Newton search methods (Using Gradients of the likelihood functions) such as Newton Raphson**
    - **Conjugate Gradients**
    - **Steepest Descent (or Ascent)**
    - **Nelder Mead approaches, etc.**

Latent Class: Estimation via EM Algorithm

Caveats on EM-Algorithm based estimation

- Prone to local optima problems
- Multiple random starts are recommended
- The eternal search for a global optimum continues even with this very general, creative, and widely used algorithm

LCA Analysis: Election Data set

Package poLCA in R implements Latent Class Analysis Election Data.

- Education
  - (1) 8 grades or less
  - (2) 9-11 grades, no further schooling
  - (3) High school diploma or equivalency
  - (4) More than 12 years of schooling, no higher degree
  - (5) Junior or community college level degree
  - (6) BA level degrees, no advanced degree
  - (7) Advanced degree
- Gender
  - (1) Male
  - (2) Female
- Party
  - (1) Strong Democrat
  - (2) Weak Democrat
  - (3) Independent-Democrat
  - (4) Independent-Independent
  - (5) Independent-Republican
  - (6) Weak Republican
  - (7) Strong Republican

LCA Analysis: Election Data set

```
#set.seed
library(poLCA)
f1=cbind(EDUC,GENDER,PARTY)~1
data(election)
names(election)
results.2=poLCA(f1,election,nclass=2,nrep=10,tol=.001,verbose=FALSE, graphs=TRUE)
results.3=poLCA(f1,election,nclass=2,nrep=10,tol=.001,verbose=FALSE, graphs=TRUE)
results.4=poLCA(f1,election,nclass=2,nrep=10,tol=.001,verbose=FALSE, graphs=TRUE)
results.5=poLCA(f1,election,nclass=2,nrep=10,tol=.001,verbose=FALSE, graphs=TRUE)
attributes(results)
results$npar
Table(results$predclass)
results$posterior
```

Latent Class Analysis: Election Data set

Number of Latent Classes	AIC	BIC
2	15311.77	**15459.47**
3	**15311.62**	15535.96
4	15313.66	15614.52

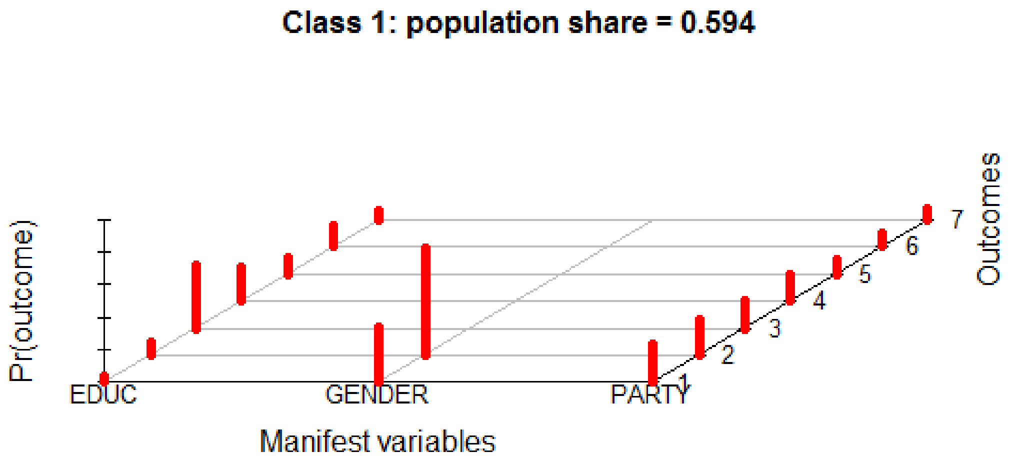
AIC: Akaike Information Criterion

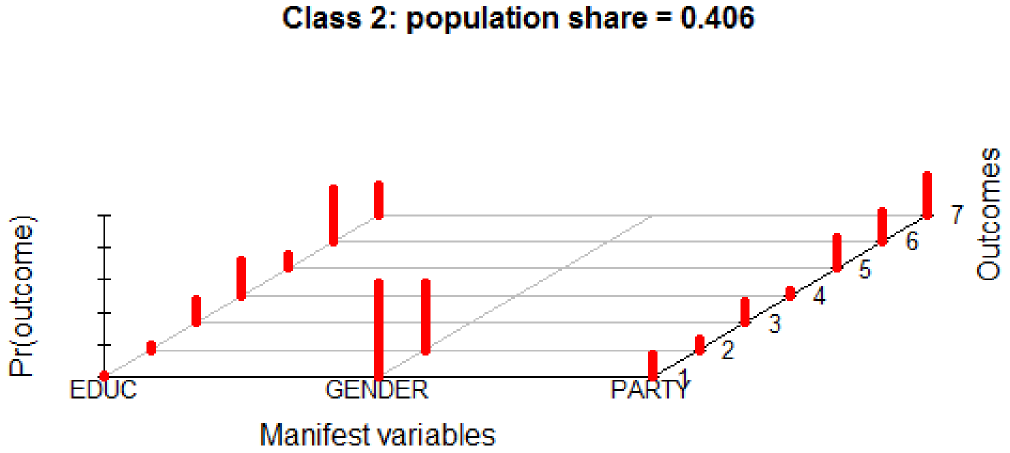
-2LL + 2P = Deviance+2*#Parameters

BIC: Bayesian Information Criterion

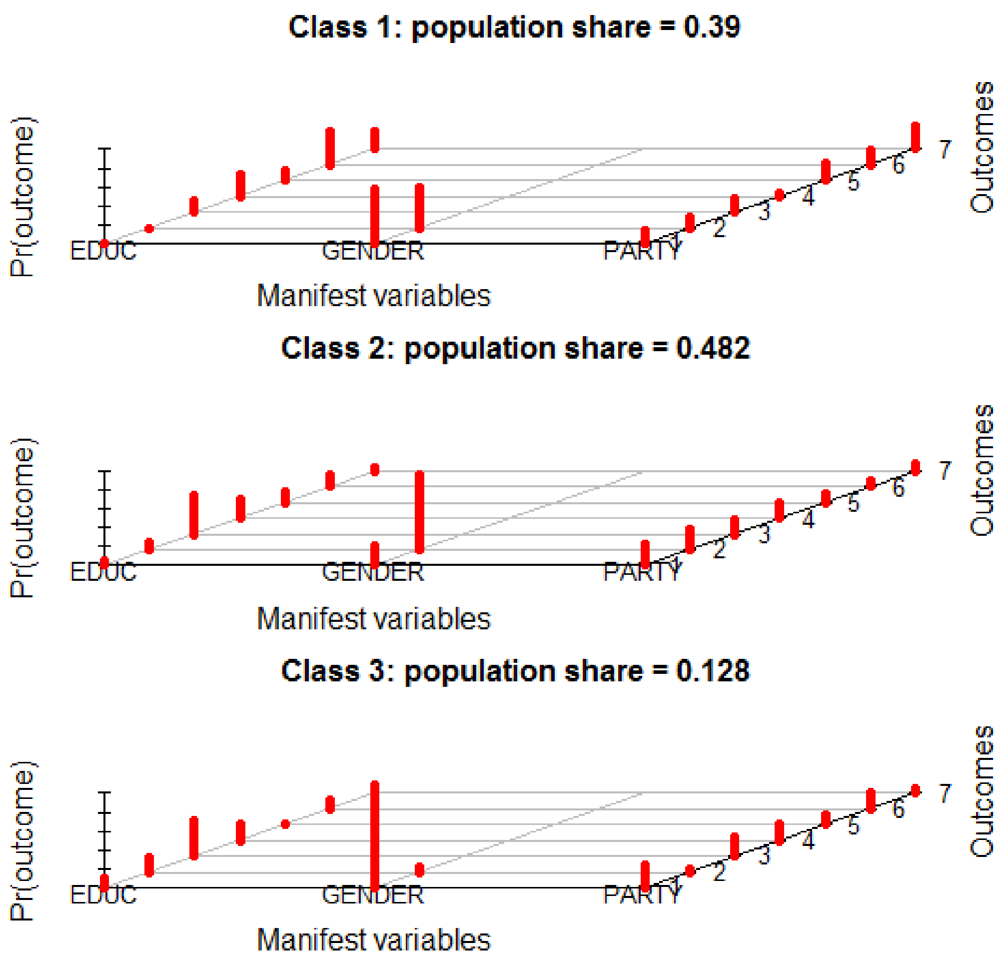
-2LL + PLog(n) = Deviance + log(samplesize)*#parameters

Latent Class Analysis: Election Data set

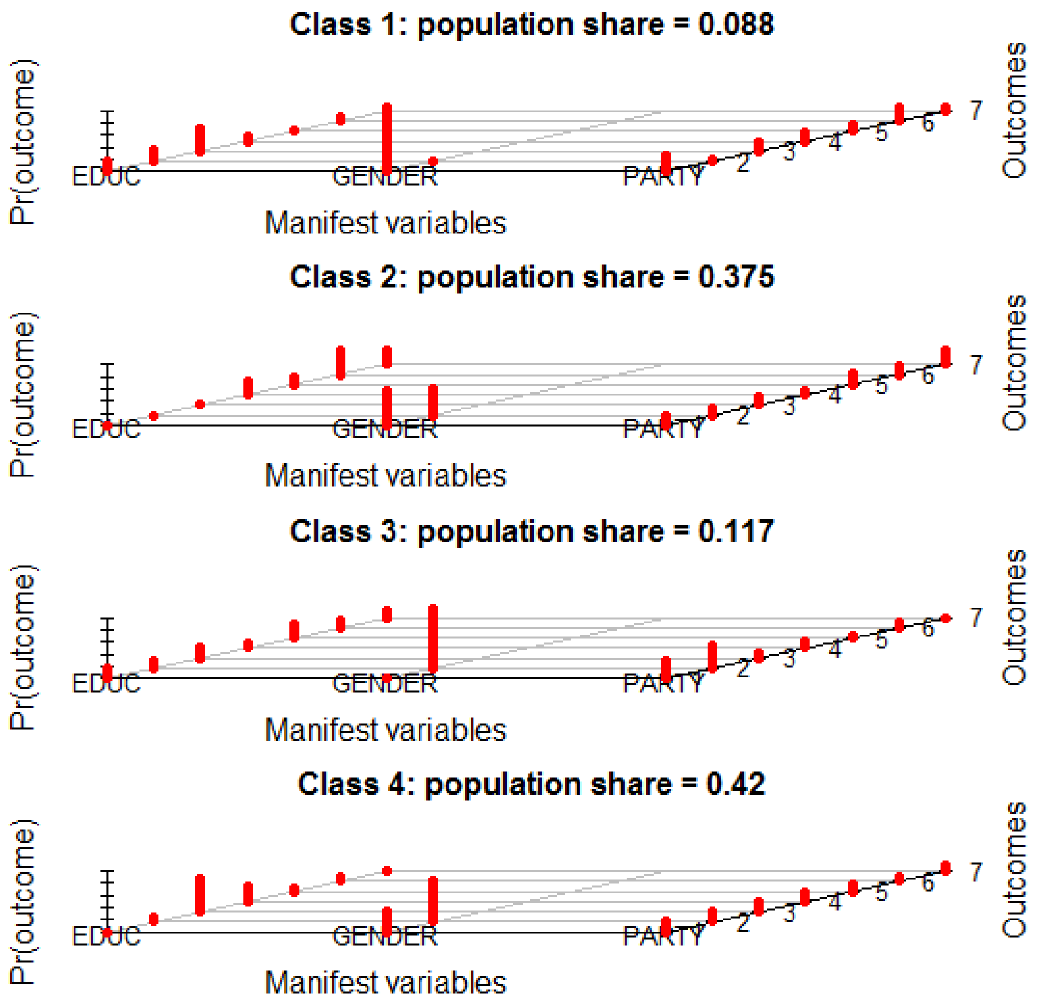




Latent Class Analysis: Election Data set



Latent Class Analysis: Election Data set



Basic Stats Primer for Latent Class

1. Assume two events A and B – with probability of occurrence p(A) and (B)

2. Joint probability of A and B = p(A and B) = p(AB)

3. If B has occurred (B given), then the joint probability of A and B is $p(AB) = p(A|B)p(B)$

4. If A has occurred (A given) then joint probability of A and B is $p(AB) = p(B|A)p(A)$

5. If A and B independent then $p(AB) = p(A)p(B)$

Basic Stats for Latent Class
Bayes Rule

If A and B are independent random variables then

$$p(AB) = p(A)p(B)$$

If A and B are NOT independent random variables then

$$p(AB) = p(A|B)p(B) = p(B|A)p(A)$$

From $P(A|B)(B) = p(B|A)p(A)$ we can see that

$$P(B|A) = [p(A|B)p(B)]/p(A)$$

Basic Stats for Latent Class
Bayes Rule

$$p(B) = p(B \text{ and } A) + p(B \text{ and } \bar{A})$$

$$p(B) = p(BA) + p(B\bar{A})$$

Basic Stats for Latent Class
Bayes Rule

$$p(B \mid A) = \frac{p(AB)}{p(A)} = \frac{p(A \mid B)p(B)}{p(A)}$$

$$p(B \mid A) = \frac{p(A \mid B)p(B)}{p(AB) + p(A\bar{B})}$$

Latent Class Analysis: EM Algorithm

- Variables j, j = 1,…,J
- Levels of variable j is k, k = 1, …, K_j
- Observations be denoted by i, i = 1,….I
- f be probability of data for observations i given class c
- d_{ijk} is data for observation i, variable j, level k (1 or 0)
- π_{cjk} is marginal probability for class c, variable j, level k
- Put random values in all marginal probability parameters π_{cjk}

$$f(\text{obs} = i \mid \text{class} = c) = \prod_{j=1}^{J} \prod_{k=1}^{k_j} (\pi_{cjk})^{d_{ijk}}$$

Latent Class Analysis: EM Algorithm

- Variables j, j = 1,…,J
- Levels of variable j is k, k = 1, …, K_j
- Observations be denoted by i, i = 1,….I
- f be probability of data for observations i given class c
- d_{ijk} is data for observation i, variable j, level k
- π_{cjk} is marginal probability for class c, variable j, level k
- S_c is size of class c (proportion). Put random values in all S_c
- C is number of latent classes

$$p(\text{obs} = i \mid C \text{ classes}) = \sum_{c=1}^{C} s_c f(\text{obs} = i \mid \text{class} = C)$$

$$p(\text{obs} = i \mid C \text{ classes}) = \sum_{c=1}^{C} s_c \prod_{j=1}^{J} \prod_{k=1}^{k_j} (\pi_{cjk})^{d_{ijk}}$$

Latent Class Analysis: EM Algorithm
E-Step: Equations for determining class membership via Posterior Probability computations

$$p(\text{class}=c \mid \text{obs}.\,i) = \frac{p(\text{class}=c \;\&\; \text{obs}=i)}{p(\text{obs}=i)}$$

$$p(\text{class}=c \mid \text{obs}.\,i) = \frac{p(\text{class}=c \;\&\; \text{obs}=i)}{p(\text{obs}=i \;\&\; c=1)+\ldots+p(\text{obs}=i \;\&\; \text{class}=C)}$$

$$p(\text{class}=c \mid \text{obs}.\,i) = \frac{p(\text{obs}=i \mid \text{class}=c)p(\text{class}=c)}{p(\text{obs}=i \mid c=1)p(\text{class}=1)+\ldots+p(\text{obs}=i \mid \text{class}=C)p(\text{class}=C)}$$

$$p(\text{class}=c \mid \text{obs}.\,i) = \frac{f(i \mid \text{class}=c)s_c}{\sum_{q=1}^{C} f(i \mid \text{class}=q)s_q}$$

Latent Class Analysis: EM Algorithm
E-Step: Equation for determining class sizes

$$s_c = \frac{1}{I}\sum_{i=1}^{I} \mathrm{p}(\mathrm{class} = \mathrm{c} \mid \mathrm{obs.\,i})$$

Latent Class Analysis: EM Algorithm
M-Step: Equations for maximizing Log likelihood to determine parameters (marginal probabilities for each class)

$$p(\text{obs}=i\,|\,C\text{ classes}) = \sum_{c=1}^{C} s_c \prod_{j=1}^{J}\prod_{k=1}^{k_j} (\pi_{cjk})^{d_{ijk}}$$

$$L = \prod_{i=1}^{I} p(\text{obs}=i\,|\,C\text{ classes})$$

$$L = \prod_{i=1}^{I}\left(\sum_{c=1}^{C} s_c \prod_{j=1}^{J}\prod_{k=1}^{k_j} (\pi_{cjk})^{d_{ijk}}\right)$$

$$\log L = \sum_{i=1}^{I}\log\left(\sum_{c=1}^{C} s_c \prod_{j=1}^{J}\prod_{k=1}^{k_j} (\pi_{cjk})^{d_{ijk}}\right)$$

Latent Class Analysis: EM Algorithm
M-Step: Equations for maximizing Log likelihood to determine parameters (marginal probabilities for each class)

$$\text{Maximize } \log L = \sum_{i=1}^{I} \log\left(\sum_{c=1}^{C} s_c \prod_{j=1}^{J} \prod_{k=1}^{k_j} (\pi_{cjk})^{d_{ijk}} \right)$$

Solution of parameters is given by following expression

$$\pi_{cjk} = \frac{\sum_{i=1}^{I} d_{ijk}\,\mathrm{prob}(\mathrm{class} = c \mid \mathrm{obs} = i)}{\sum_{i=1}^{I} p(\mathrm{class} = c \mid \mathrm{obs} = i)}$$