What is Latent Class Analysis

Tarani Chandola
methods@manchester

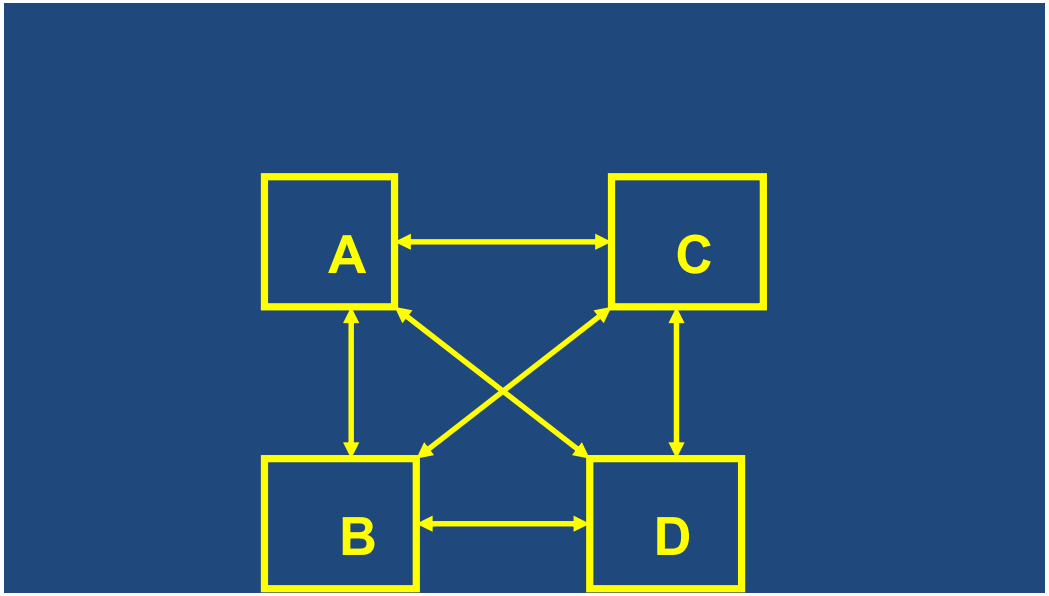
Many names- similar methods

- (Finite) Mixture Modeling
- Latent Class Analysis
- Latent Profile Analysis

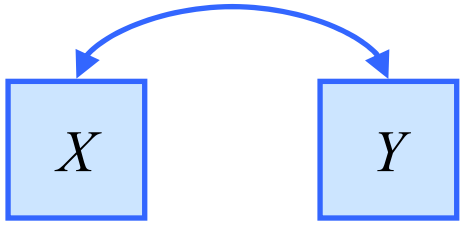
Latent class analysis (LCA)

- LCA is a similar to factor analysis, but for categorical responses.

- Like factor analysis, LCA addresses the complex pattern of association that appears among observations….

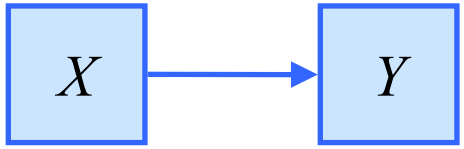


Factor Analysis

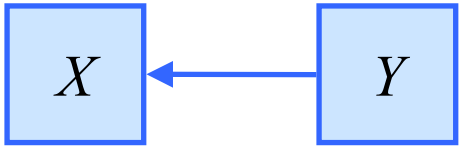


We observe a correlation between two variables. Why?

1. X causes Y ?

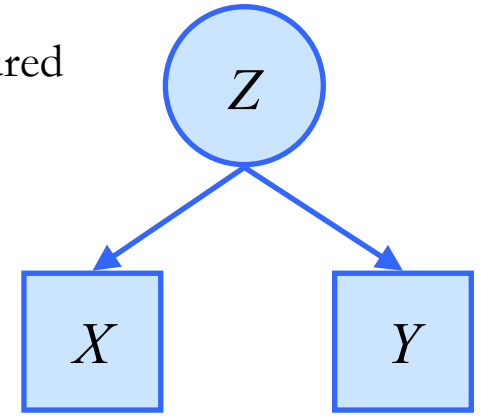


2. Y causes X ?



3. Reciprocal causation ($X \leftrightarrow Y$) ?

4. A third, unmeasured cause ?



Unmeasured Causes: Factor Models

Variables may be related due to the action of unobserved influences.

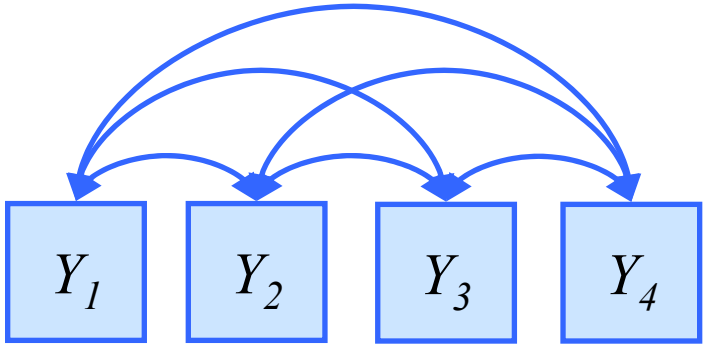
Sometimes these are confounding variables, but many constructs of interest are not directly observed (or even observable)

Unobserved Construct	Observed Measures
Social Capital	Bowling club membership
	Local newspaper reading
Ethnic prejudice	Housing segregation
	Ethnic intermarriage

Factor Models

Correlations may not be due to causal relations among the observed variables at all, but due to these unmeasured, latent influences - factors

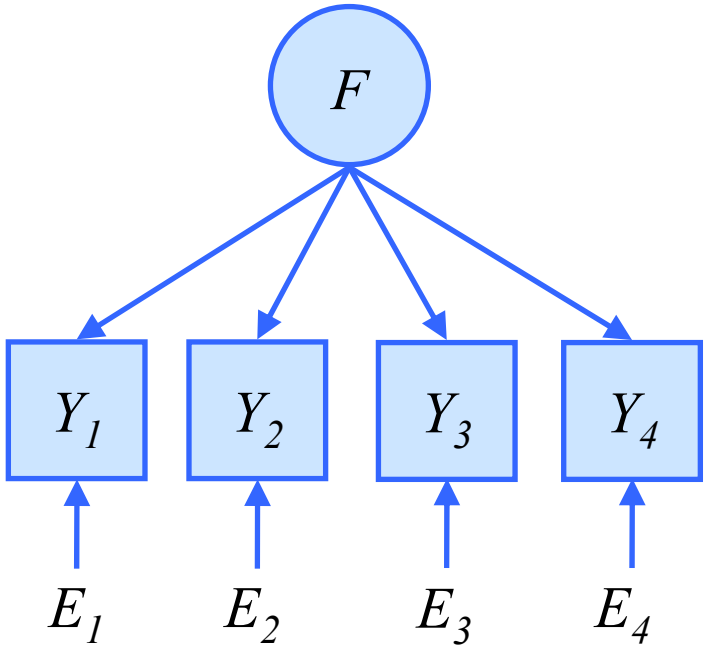
	Y_1	Y_2	Y_3	Y_4
Y_1	1.0			
Y_2	0.6	1.0		
Y_3	0.7	0.6	1.0	
Y_4	0.5	0.6	0.8	1.0



Factor Models

The observed correlations may be due to each observed measure sharing an unobserved component (F)

	Y_1	Y_2	Y_3	Y_4
Y_1	1.0			
Y_2	0.6	1.0		
Y_3	0.7	0.6	1.0	
Y_4	0.5	0.6	0.8	1.0



Factor Models

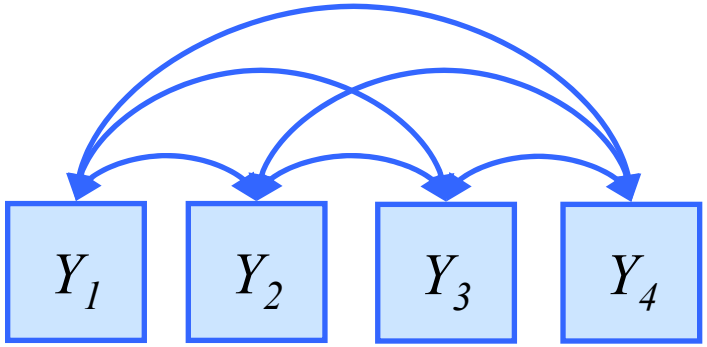
Example: Four questionnaire items that have highly correlated answers

Y1 “I Often feel blue”

Y2 “I dislike myself”

Y3 “I have a low opinion of myself”

Y4 “My life lacks direction”



Factor Models

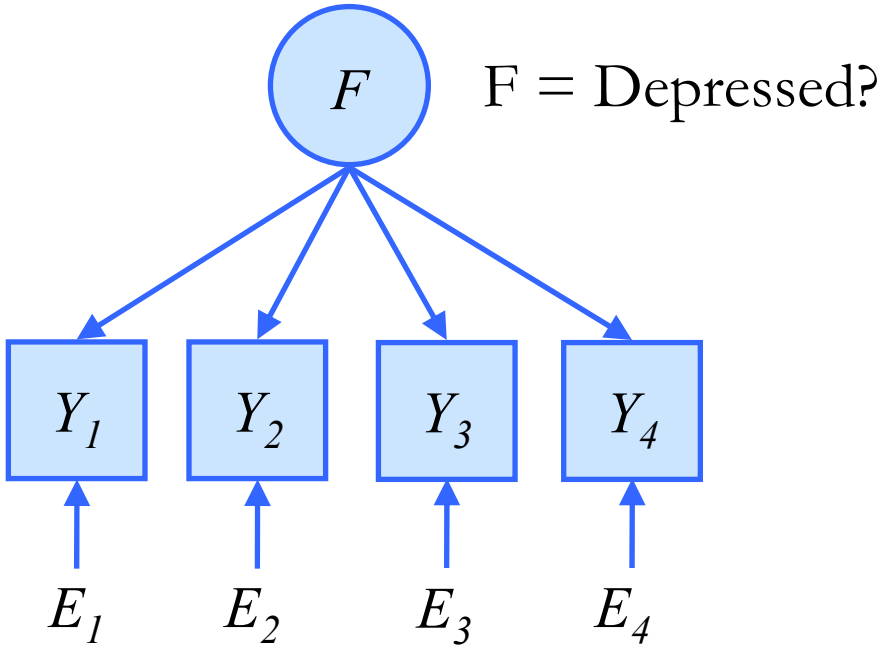
The items may be correlated due to the influence of the respondent's mood state, which we can't observed directly

Y1 "I Often feel blue"

Y2 "I dislike myself"

Y3 "I have a low opinion of myself"

Y4 "My life lacks direction"



Factor Models

1	F	e1	e2	e3	e4
2	1.2	-0.4	0.2	-1.5	-1.4
3	3.3	0.8	-0.2	-0.1	0.9
4	2.2	0.8	-1.8	0.0	1.5
5	1.3	0.6	-1.9	0.3	1.0
6	1.5	-0.9	0.1	1.6	1.0
7	1.6	-1.5	1.0	0.5	-0.4
8	2.2	1.5	1.2	-0.7	0.7
9	2.1	-0.6	0.7	0.1	0.2
10	0.7	0.3	0.2	-0.4	1.5
11	1.9	0.5	-1.3	0.2	-0.1

Hypothesised Factor Model

y1	y2	y3	y4
0.8	1.4	-0.3	-0.2
4.1	3.1	3.2	4.2
3.0	0.4	2.2	3.7
1.9	-0.6	1.6	2.3
0.6	1.6	3.1	2.5
0.1	2.6	2.1	1.2
3.7	3.4	1.5	2.9
1.5	2.8	2.2	2.3
1.0	0.9	0.3	2.2
2.4	0.6	2.1	1.8

Observed data

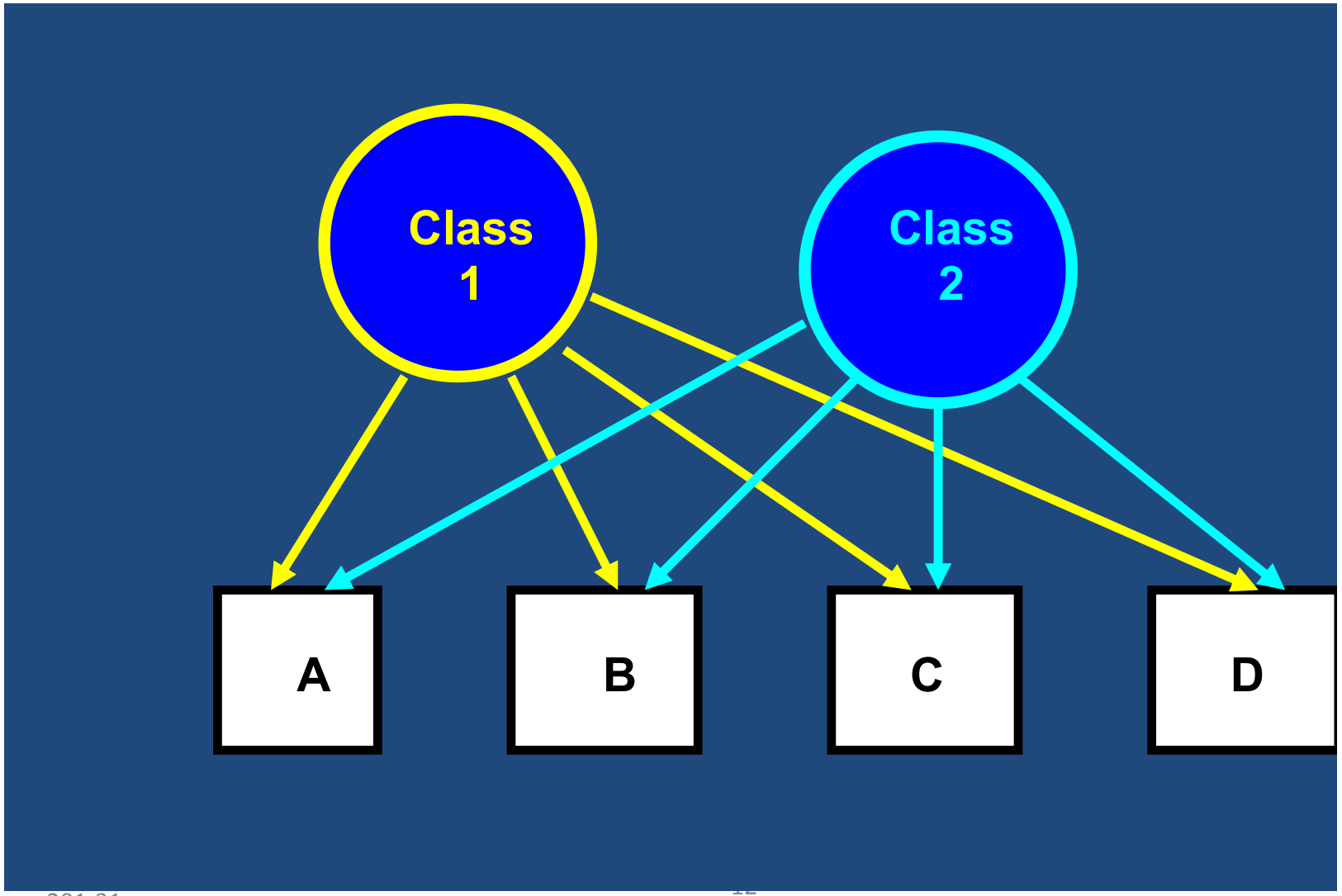
Model Fit

- Standard measure of 'observed' vs. 'expected' fit?
  - Pearson χ^2 (Chi-Square) test
  - Sum of the squared differences between observed (O) and expected (E) (co)variances divided by the expected

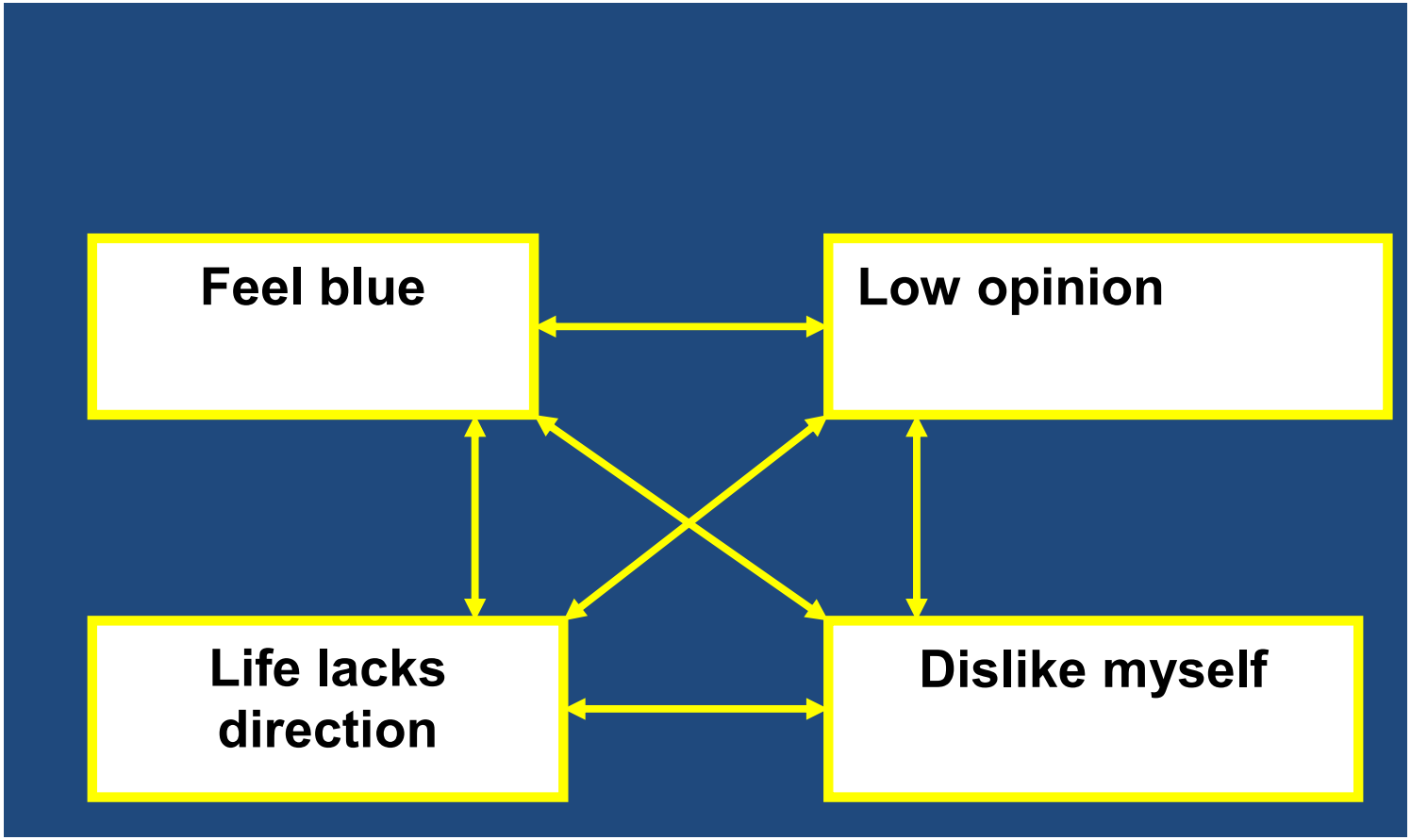
$$\chi^2 = \Sigma[(O-E)^2/E]$$

  - The larger the χ^2 the greater the model **mis**fit
  - Can test if $\chi^2 = 0$ using the model df

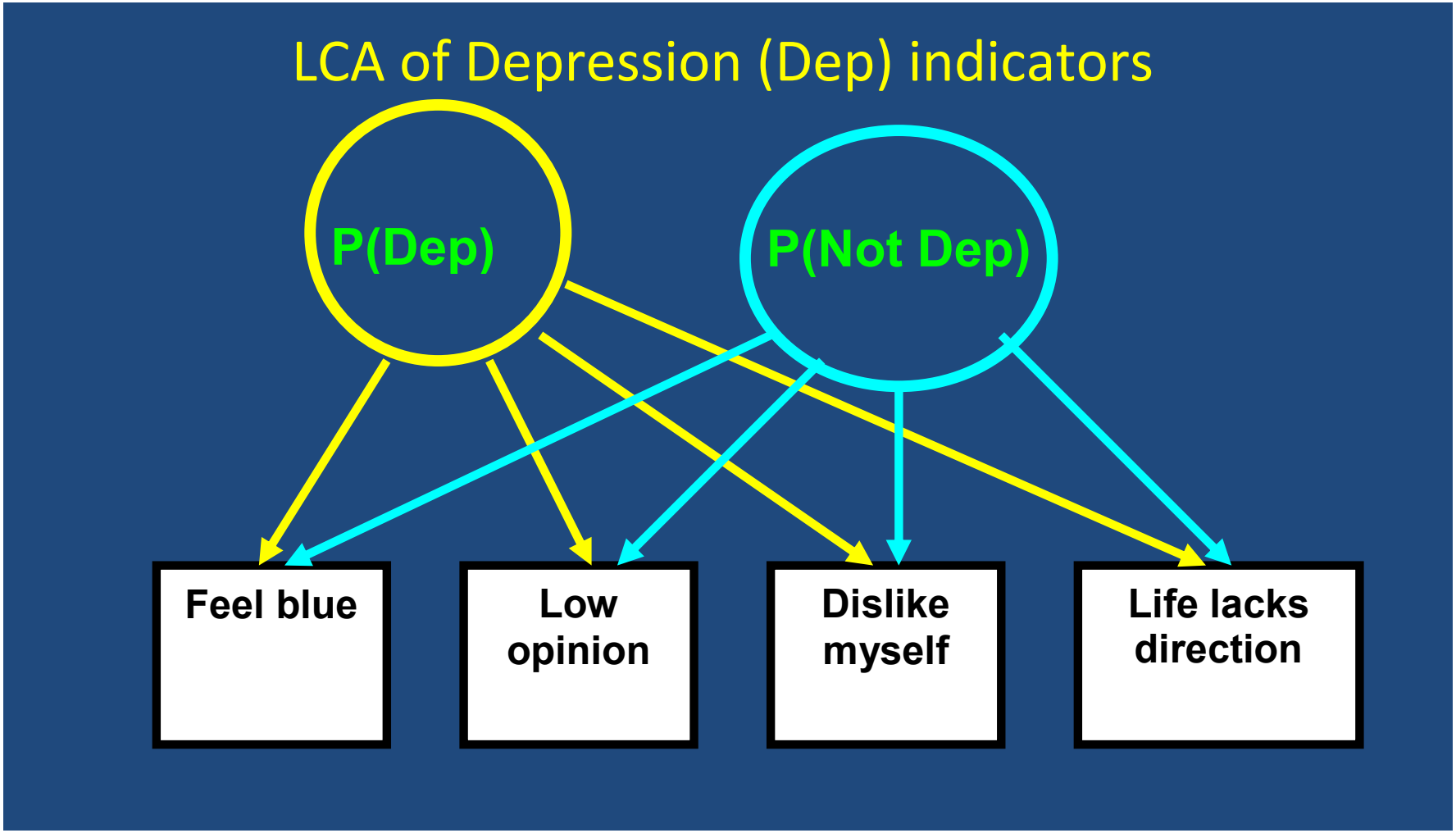
In LCA, the underlying unobserved variables are not continuous (dimensions) but classes/categories/discrete



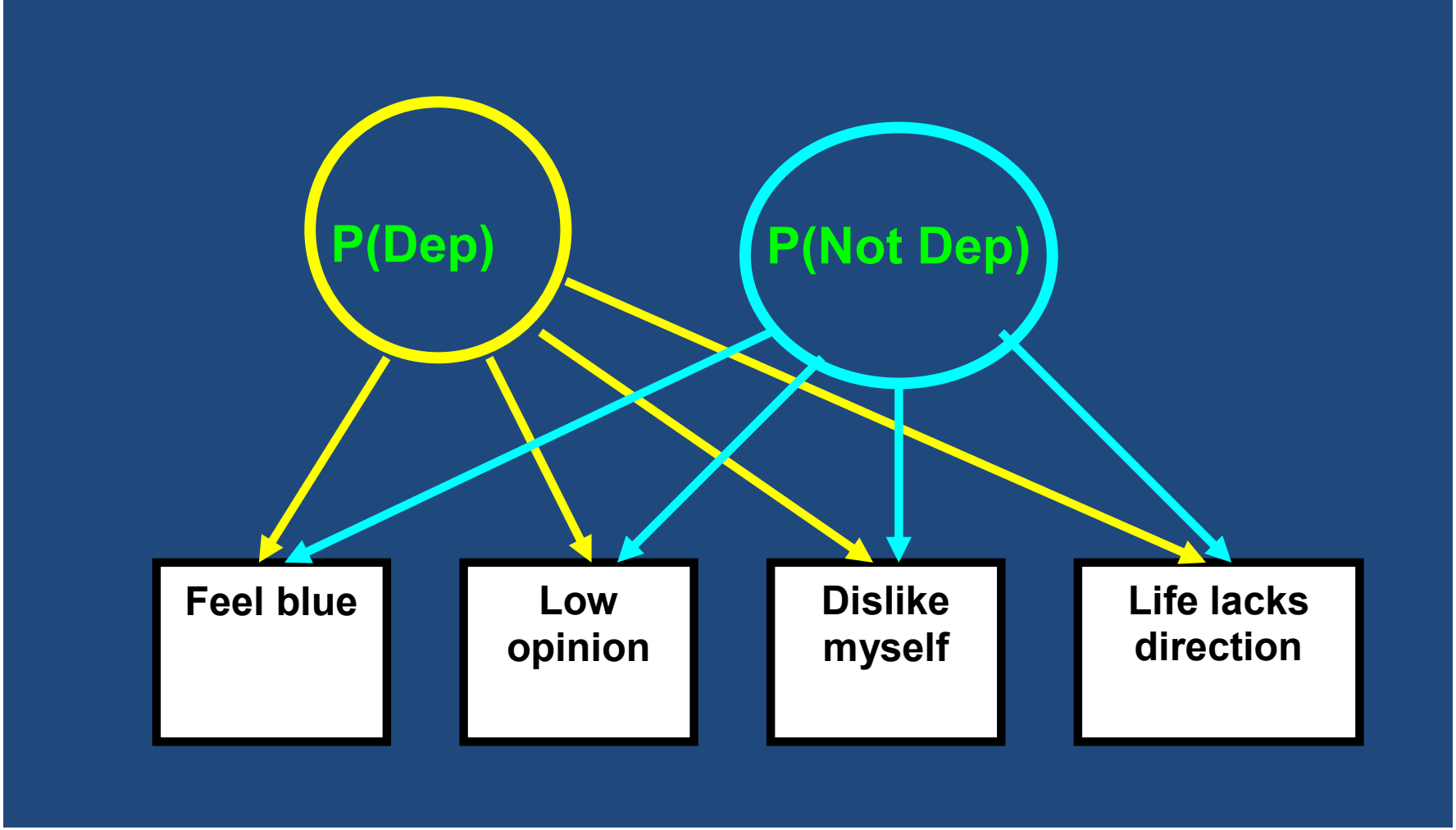
What if you do not know how to classify people into (depressed vs not depressed) groups? What if there is no gold standard to assess a pattern of “yes/no”signs and symptoms?



Rindskopf, R., & Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in Medicine, 5*, 21-27.



LCA predicts latent class membership such that the observed variables are independent.



LCA estimates

- Latent class prevalences
- Conditional probabilities: probabilities of specific response, given class membership

LCA works on unconditional contingency table (no information on latent class membership)

Feel blue	Low opinion	Dislike myself	Life lacks direction	n_{ijkl}
0	0	0	0	15
0	0	0	1	14
0	0	1	0	11
0	0	1	1	8
0	1	0	0	23
.	.	.	.	.
1	1	1	1	9

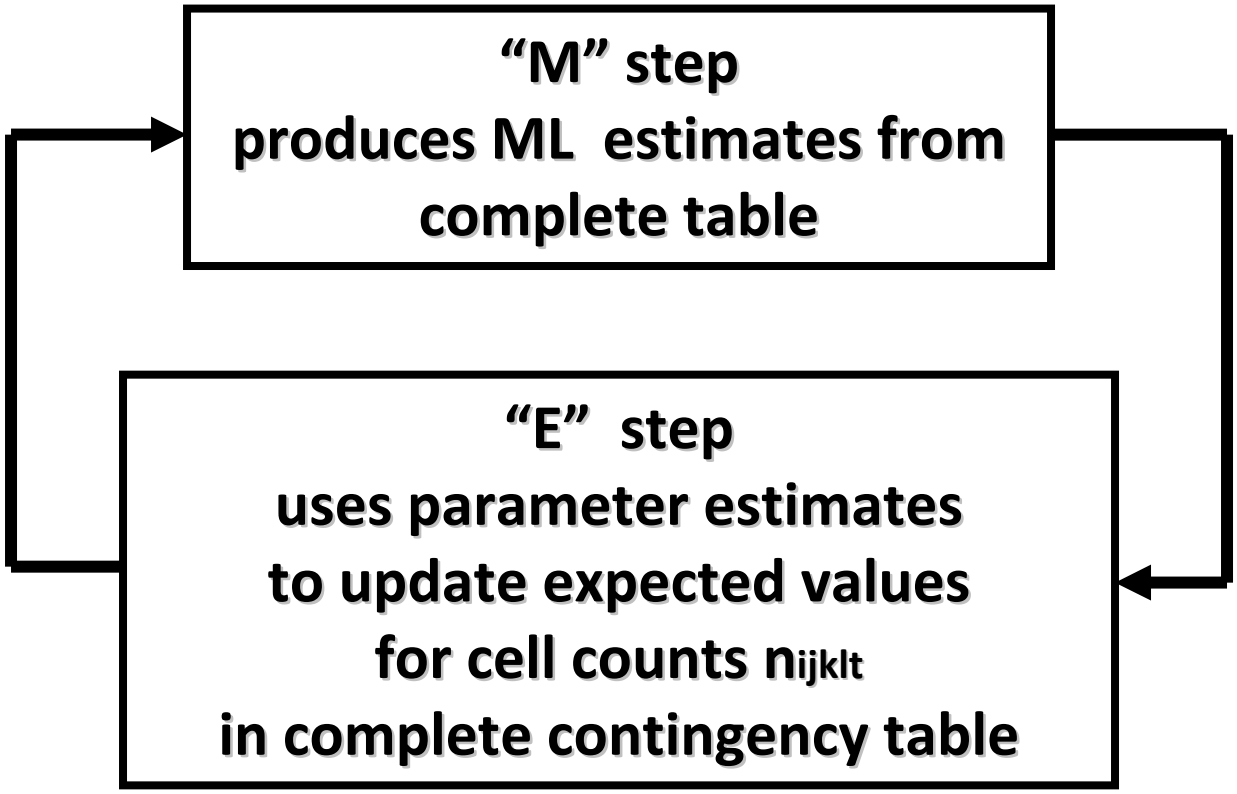
LCA's goal is to produce
a complete (conditional) table
that assigns counts for each latent class:

Feel blue	Dislike myself	Low opinion	Life lacks direction	Latent Class X=t	n_{ijklt}
0	0	0	0	1	9
0	0	0	1	2	6
0	0	1	0	1	3
0	0	1	1	2	11
.	.	.	.	.	.
1	1	1	1	2	9

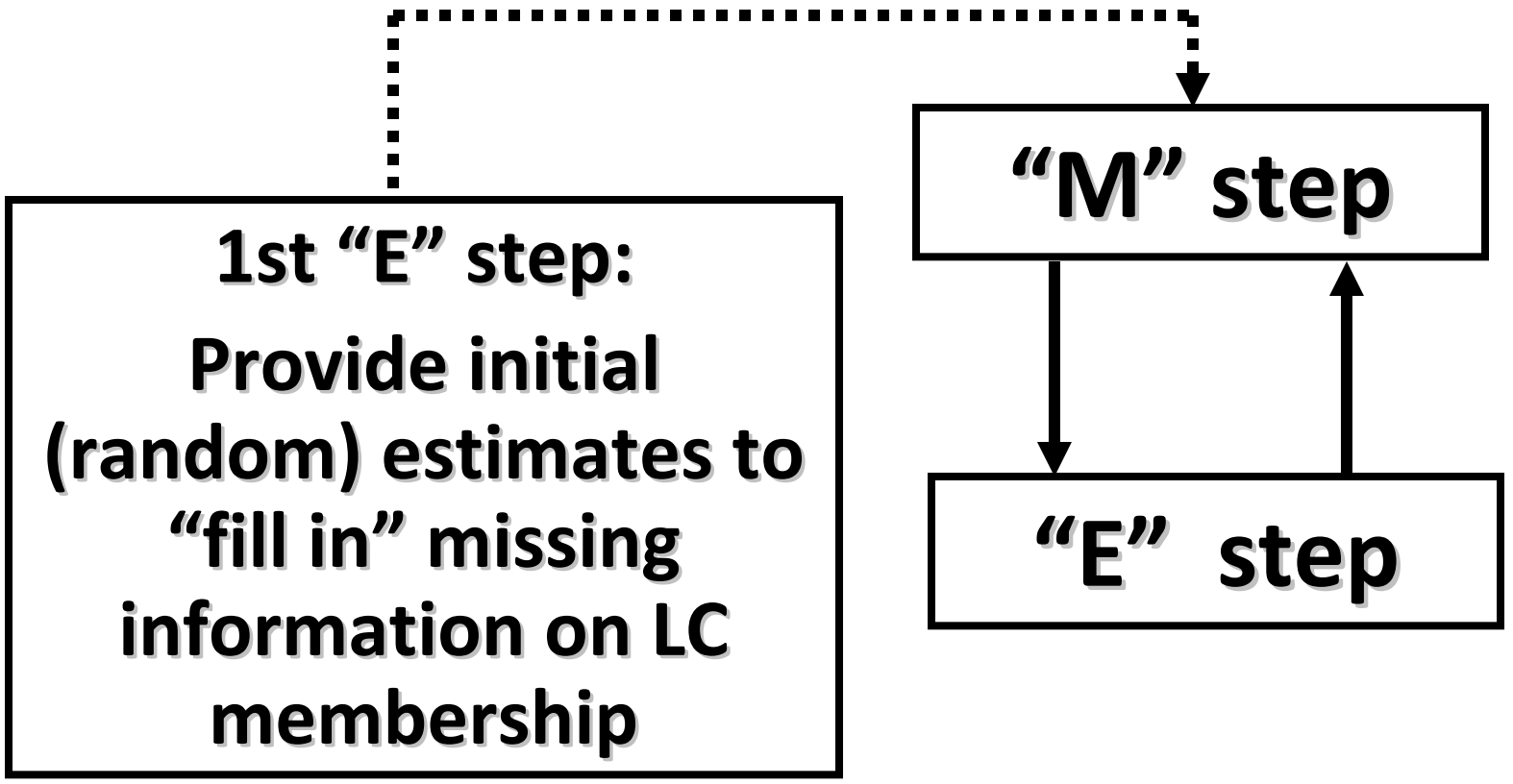
Estimating LC parameters

- Maximum likelihood approach
- Because LC membership is unobserved, the likelihood function, and the likelihood surface, are complex.

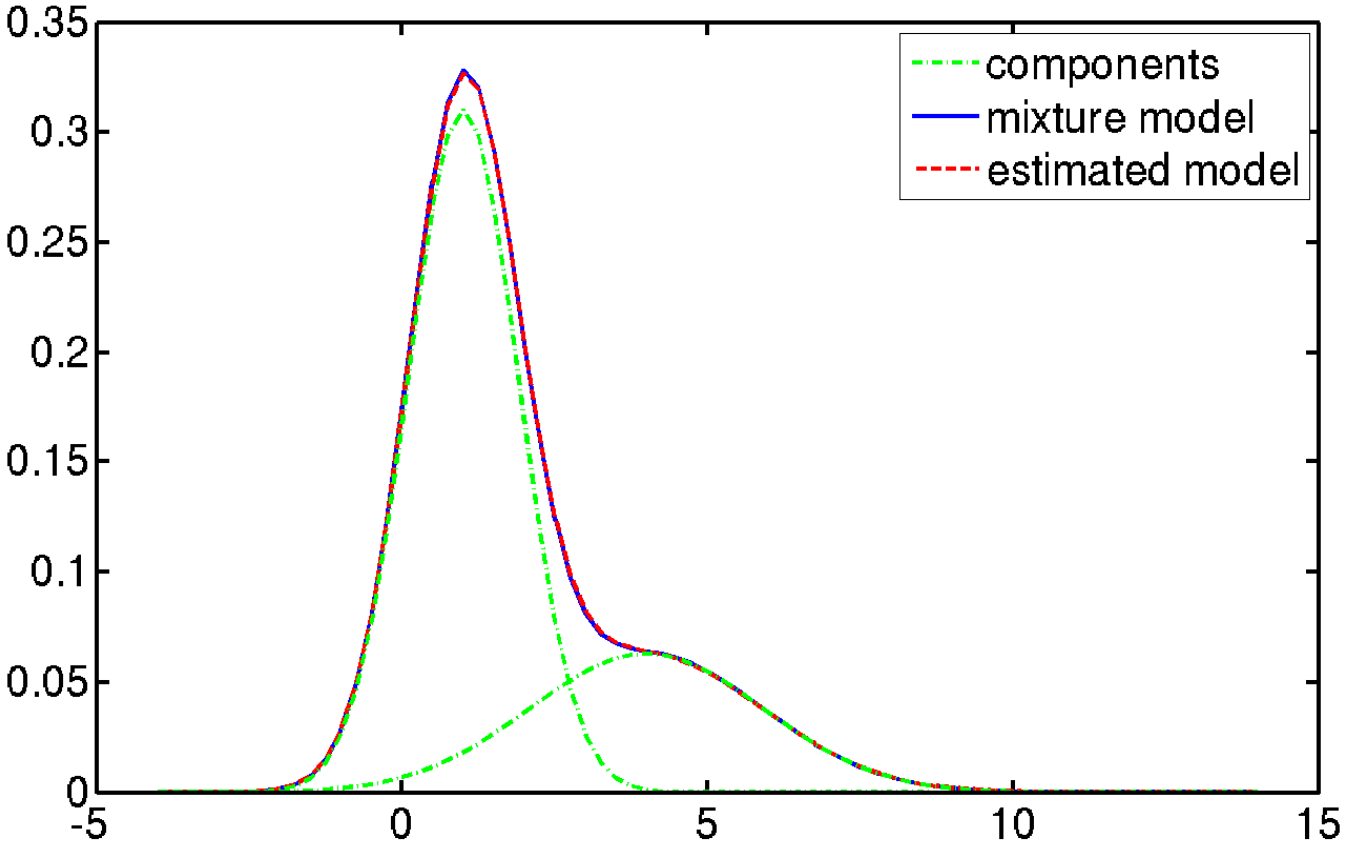
EM algorithm calculates L when some data (X) are unobserved

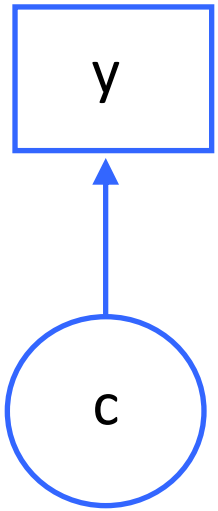


EM algorithm requires initial estimates

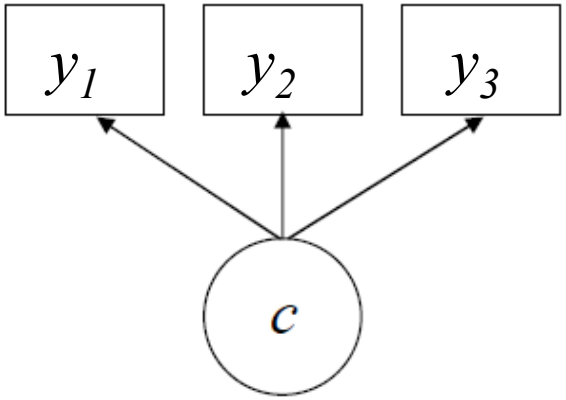


Mixture modeling





Latent Profile Analysis Model



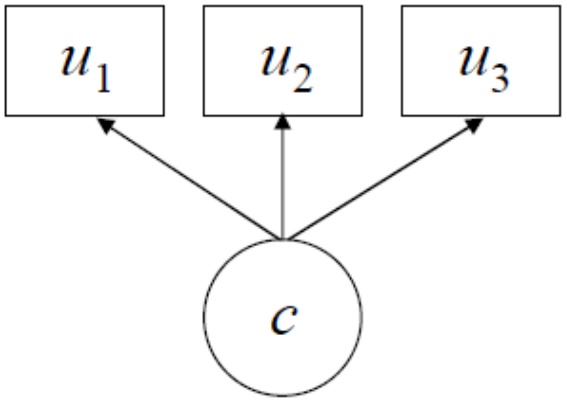
Continuous indicators

y: $y_1, y_2, \ldots, y_r$

Categorical latent variable

c: $c = k$; $k = 1, 2, \ldots, K$.

Latent Class Analysis Model



Dichotomous (0/1) indicators

u: $u_1, u_2, \ldots, u_r$

Categorical latent variable

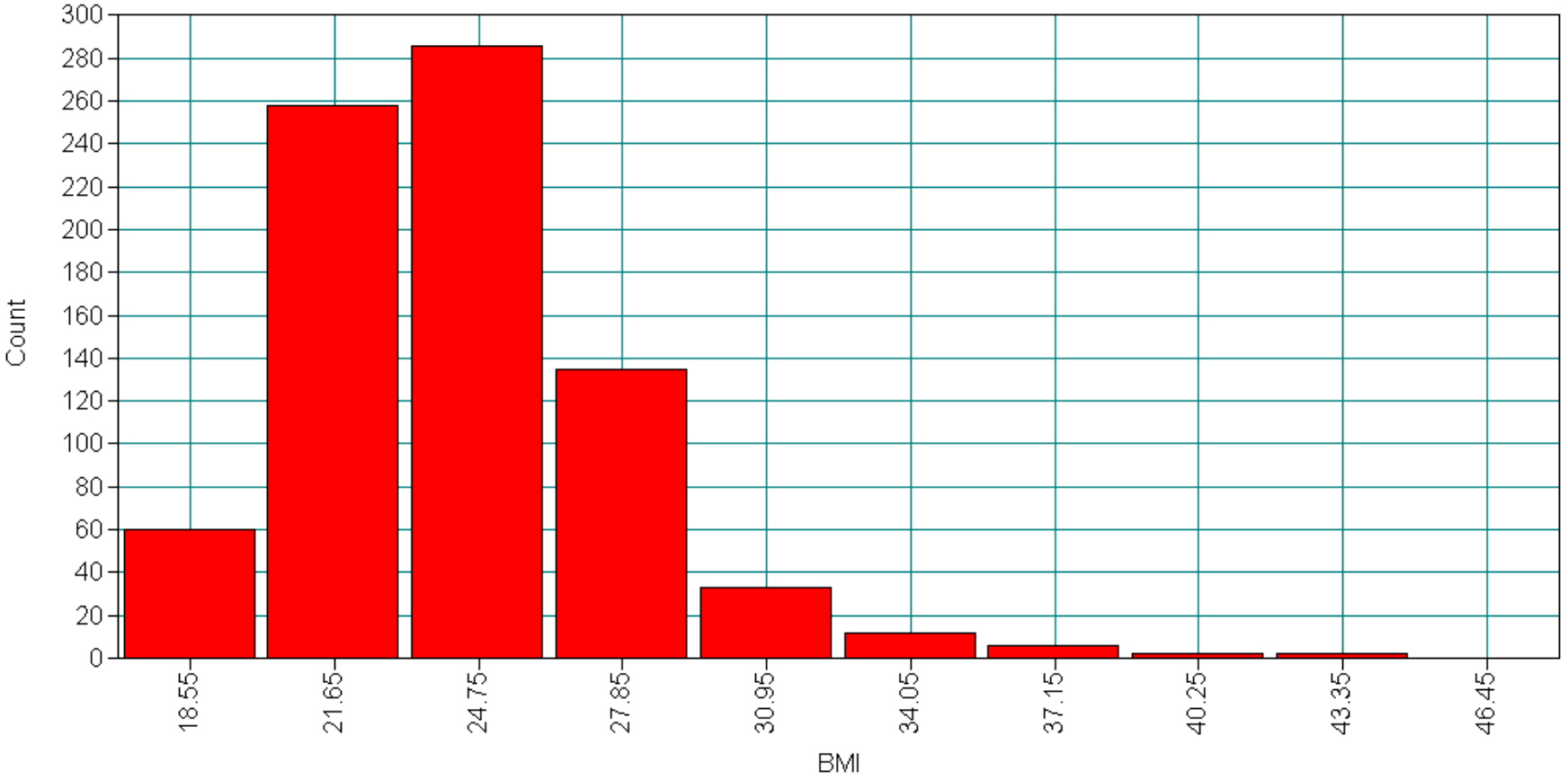
c: $c = k$; $k = 1, 2, \ldots, K$.

Model Results

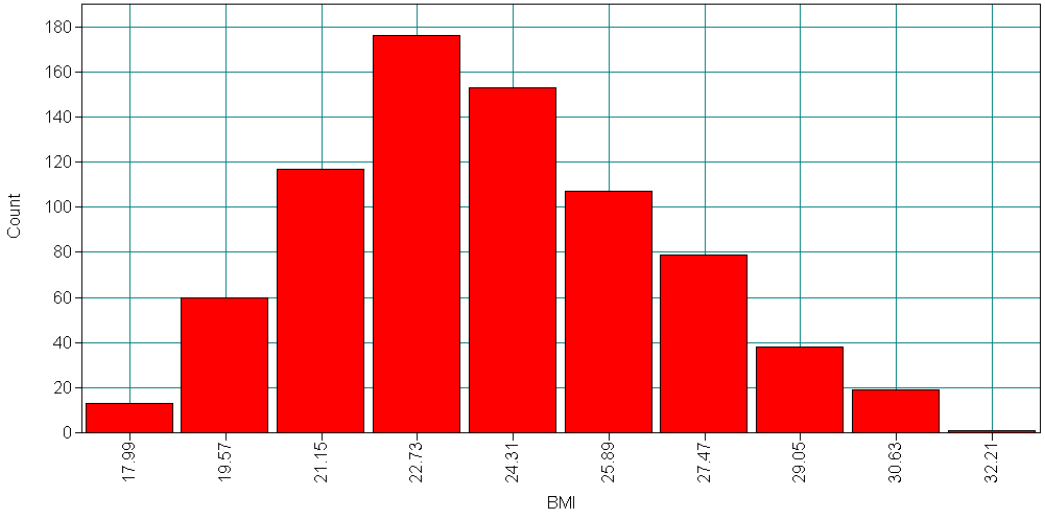
	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Latent Class 1				
Means				
BMI	25.166	0.139	181.262	0.000
Variances				
BMI	15.305	1.279	11.970	0.000

Mean BMI of 25.2 (and variance of 15.3) in the whole population

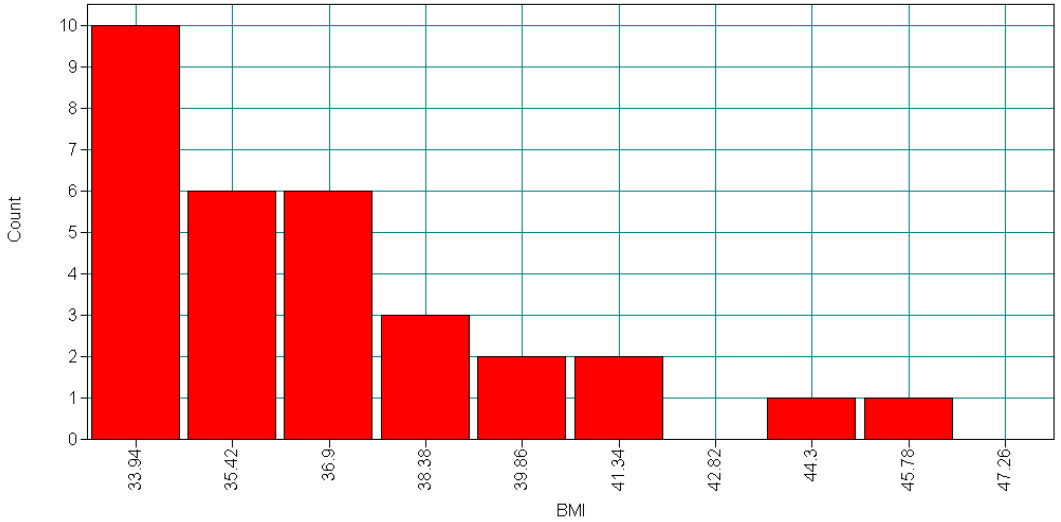
Histogram of BMI (1 class solution)



Class 1



Class 2

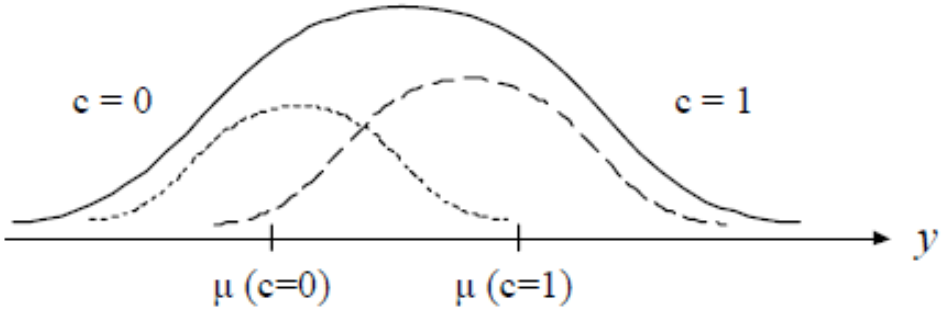


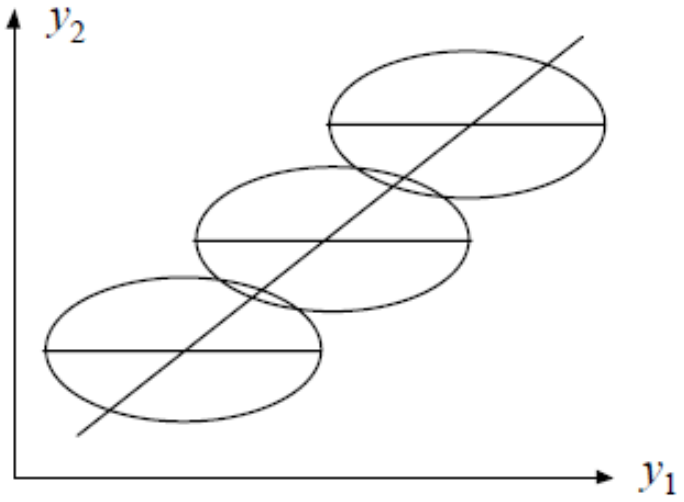
Mixture Model of BMI with 3 classes

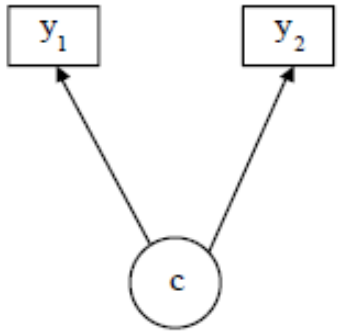
	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Latent Class 1				
Means				
BMI	32.100	16.503	1.945	0.052
Variances				
BMI	8.319	4.653	1.788	0.074
Latent Class 2				
Means				
BMI	40.685	18.724	2.173	0.030
Variances				
BMI	8.319	4.653	1.788	0.074
Latent Class 3				
Means				
BMI	24.414	1.342	18.190	0.000
Variances				
BMI	8.319	4.653	1.788	0.074
Categorical Latent Variables				
Means				
C#1	-2.561	2.973	-0.861	0.389
C#2	-4.277	5.878	-0.728	0.467

! Those in Class 2 have much higher mean BMI than those in classes 1 and 3

Latent Profile/Class analysis with 2 and 3 latent classes







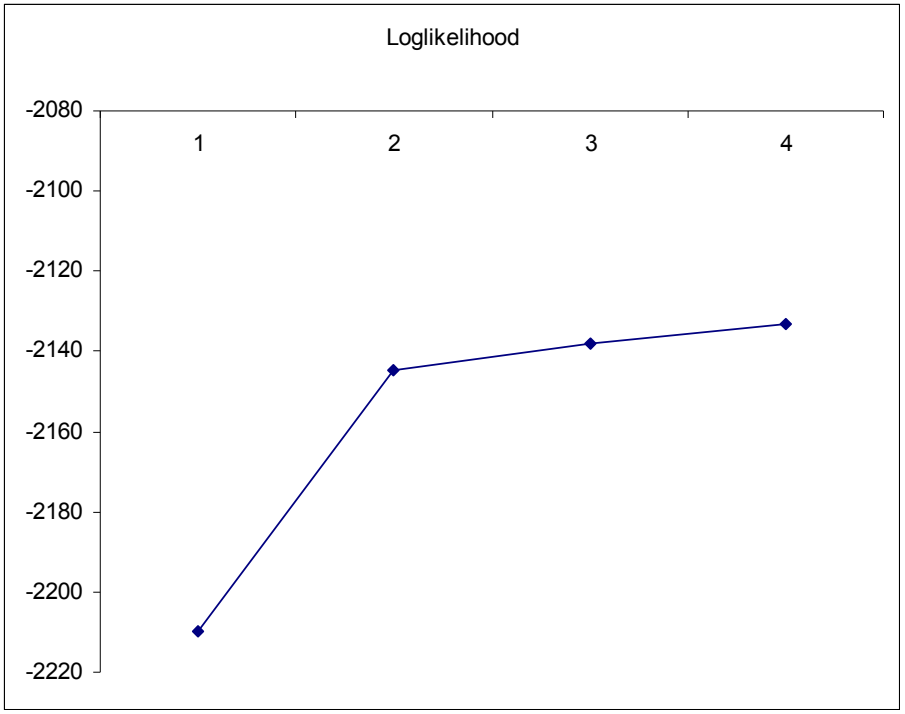
3 classes

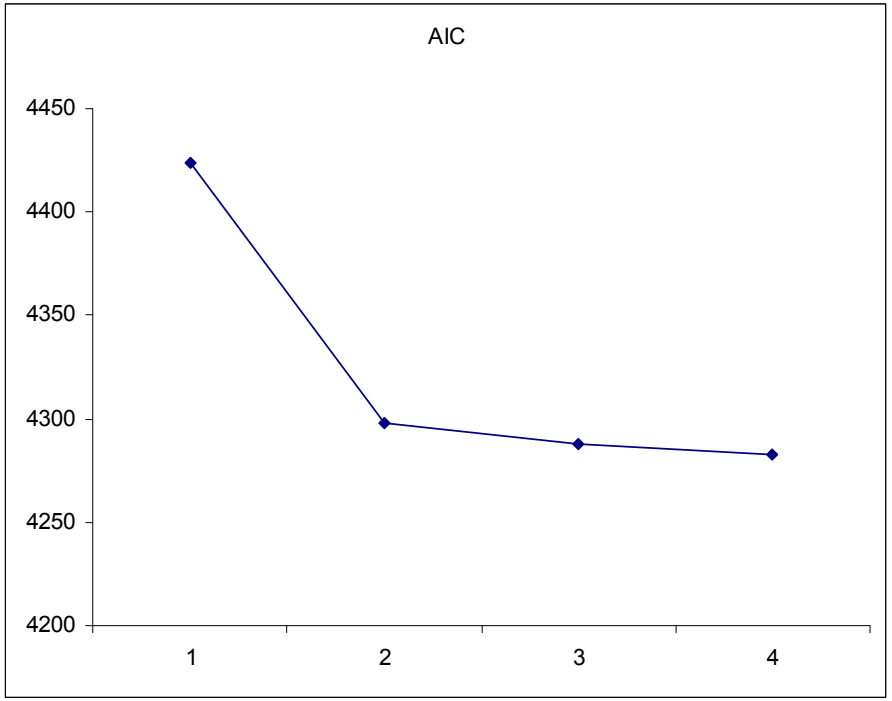
Deciding on number of latent classes

- Start with the simplest (a one class) solution, and add more classes stepwise.
- Examine the model evaluation statistics: Chi-square difference tests are not appropriate for likelihood ratio test comparisons of models with higher numbers of classes. Models that ***maximize* the log likelihood** are generally better fitting, although this comes at the expense of fitting more parameters to the model. Look for ***low*** values on the **Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC) and sample size adjusted BIC statistics**. In addition, **Tech 11:** modification to the likelihood ratios test that adjusts the conventional likelihood ratio test for K vs K-1 classes for violation of regularity conditions (***p>0.05*** indicates K-1 classes are sufficient).
- Examine **entropy measure** (***higher*** values indicate better fit).
- Usefulness of the latent classes in practice. This can be determined by examining the trajectory shapes for similarity, the number of individuals in each class, and whether the classes are associated with observed characteristics in an expected manner.

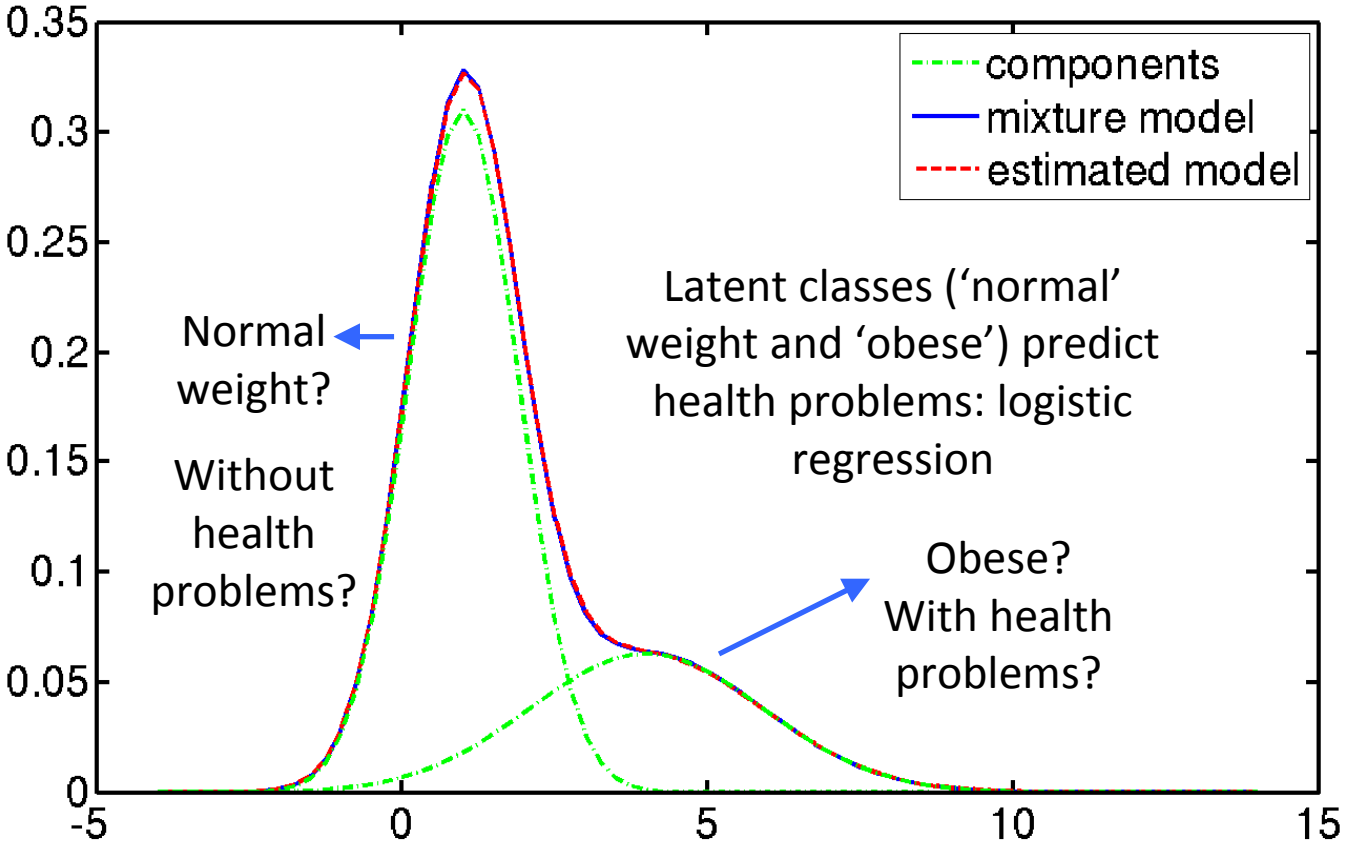
Deciding on number of classes- BMI example

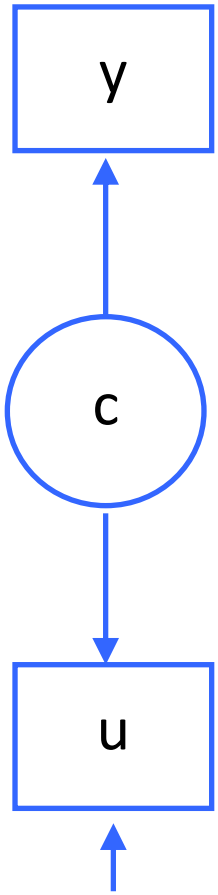
No. of classes	Loglikelihood	# par.	AIC	BIC	Entropy	LRT p-value for k-1
1	-2209.712	2	4423.424	4432.778	NA	NA
2	-2144.898	4	4297.797	4316.505	0.952	0.0000
3	-2137.826	6	4287.652	4315.714	0.901	0.8237
4	-2133.359	8	4282.718	4320.135	0.745	0.0326



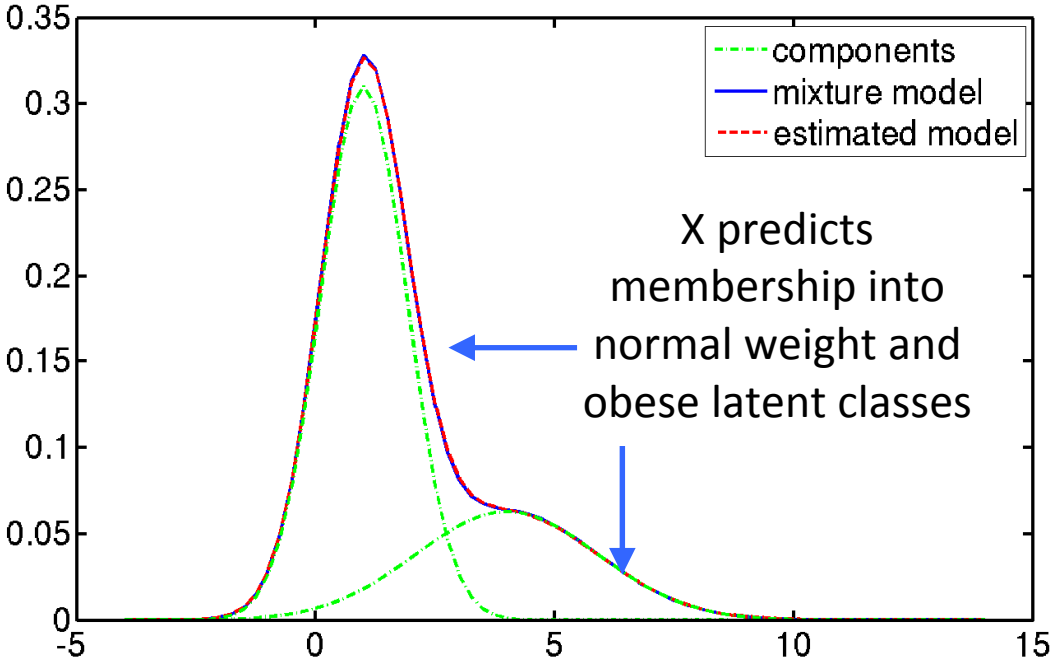


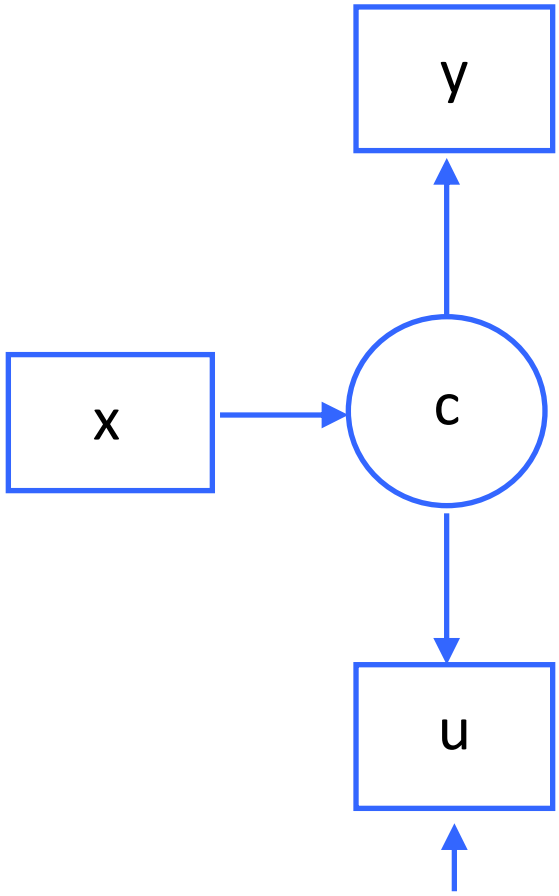
Mixture modeling with categorical dependent variables





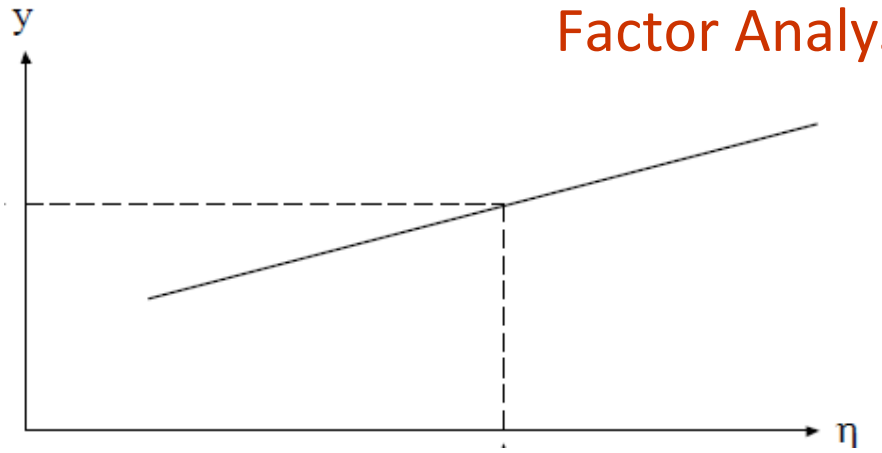
Mixture model with covariates and categorical dependent variables

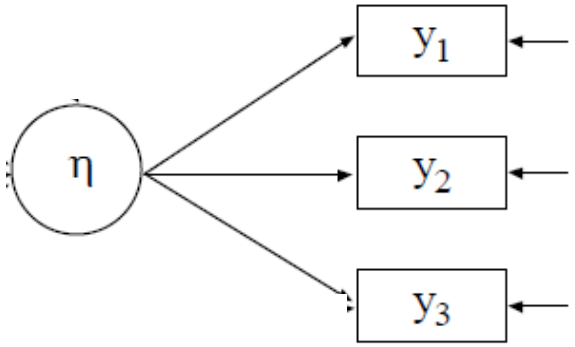




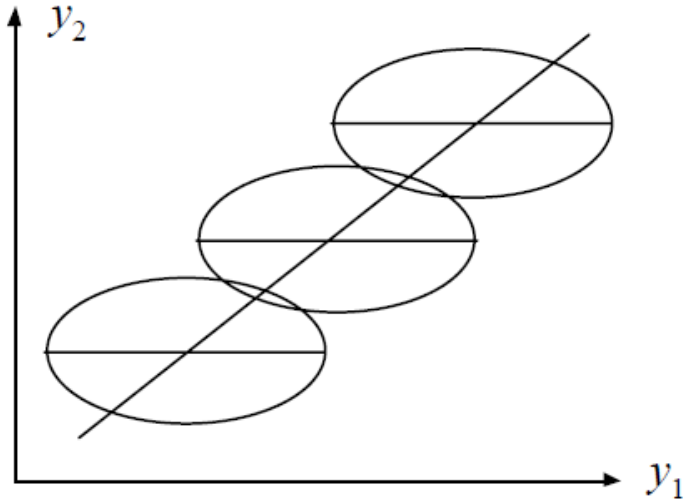
Are you a joiner or a splitter?

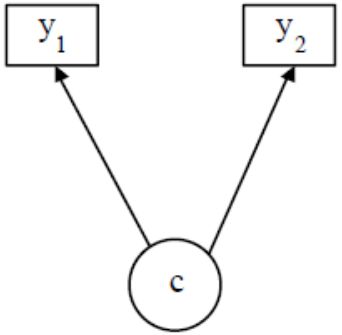
Factor Analysis





vs. Latent Profile/Class Analysis





3 classes

Resources

Introduction to LCA:

http://www.john-uebersax.com/stat/faq.htm

http://www.ccsr.ac.uk/methods/festival/programme/wiwp/francis.pdf

McCutcheon AC. Latent class analysis. Beverly Hills: Sage Publications, 1987

Software:
http://www.john-uebersax.com/stat/soft.htm

Short courses:
Latent Trait and Latent Class Analysis for Multiple Groups Using Mplus
http://www.ccsr.ac.uk/courses/congnitiveInterviewing/LatT.html

Introduction to Structural Equation Modelling using Mplus
http://www.ccsr.ac.uk/courses/semintro/