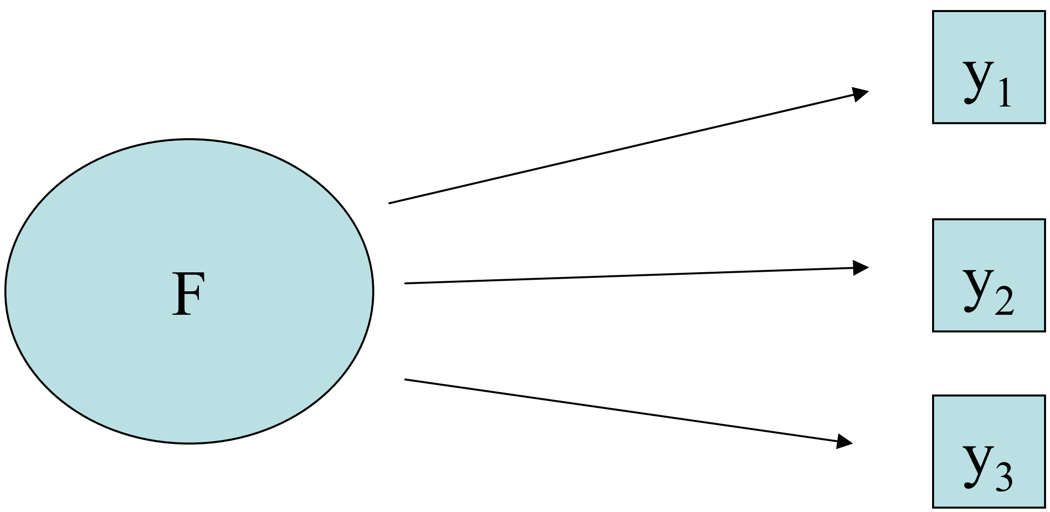
Latent Class Analysis

October 9 and 11, 2006

Statistics for Psychosocial Research

Introduction:
Types of Latent Variable Models

		Observed Variables	
		Continuous	Categorical
Latent Variable	Continuous	Factor analysis	Latent trait Analysis
	Categorical	Latent profile analysis	Latent class analysis



Latent Class Analysis

Definitions:

- method for describing associations in a multidimensional contingency table
- method for describing the patterns in which multiple categorical variables co-occur
- views populations as comprising several subpopulations, and responses as surrogates that imperfectly measure to which subpopulation an individual belongs

Motivating Example: Functional Disability

- Functional disability in elderly women (over age 65)
- Women's Health and Aging Study (WHAS), 1991
- "Screener" data subsample (755 of 3500 women interviewed)
- For each woman in the sample we have answers to questions about difficulty with mobility tasks:
  - Do you have difficulty…..?
  - 0=no, 1=yes: $y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5},)$
- Random community sample ⇨ reflects true prevalence

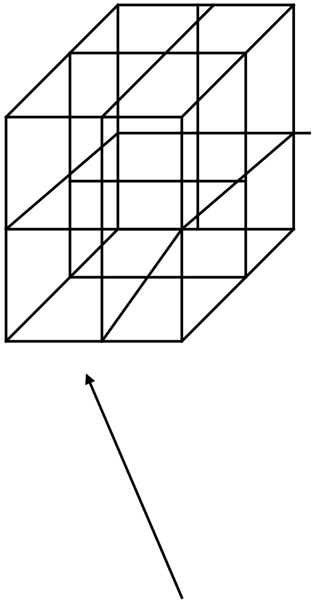
WHAS Example

- Five indicators of mobility disability
  - do you have difficulty **walk**ing 1/2 mile?
  - do you have difficulty walking up 10 **steps**?
  - do you have difficulty getting in and out of a bed or **chair**?
  - do you have difficulty with **h**eavy **h**ouse**w**ork?
  - do you have difficulty **lift**ing 10 lbs.?

	% with difficulty
Walk	0.39
Steps	0.25
Chair	0.16
Hhw	0.42
Lift	0.30

Multidimensional Contingency Table

- WHAS example: 2x2x2x2x2 table
- With five items
  - 00000
  - 10000
  - 01000
  - ….
  - 01111
  - 11111
- $2^5 = 32$ possible patterns



2x2x2 table(?)

Latent Class Model: Main Ideas

- There are M classes of disability (e.g. none, mild, severe). π_m represents the proportion of individuals in the population in class m (m=1,…,M)

- Each person is a member of one of the M classes, but we do not know which. The latent class of individual i is denoted by η_i.

- Symptom prevalences vary by class. The prevalence for difficulty with task k in class m is denoted by p_{km}.

☼ Conditional Independence: Given class membership, the tasks are independent. BIG assumption!

How do we use the latent class model?

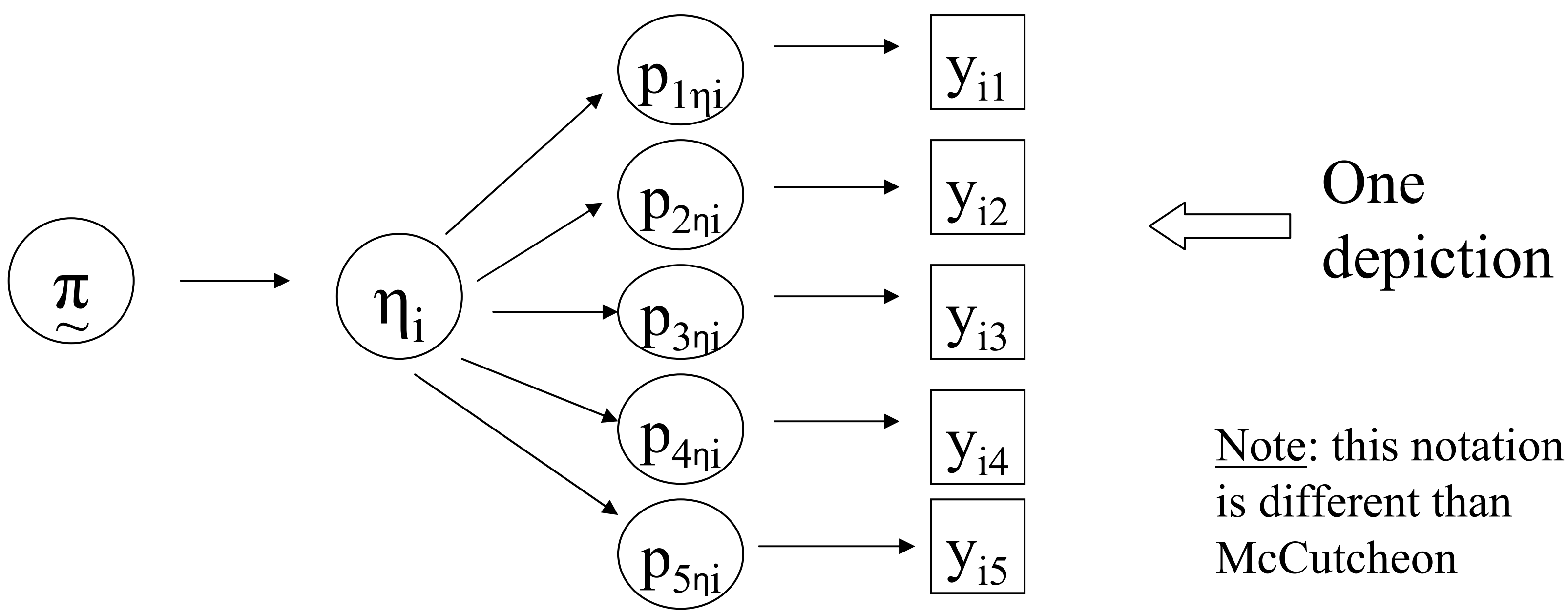
- <u>Classification</u>: Individuals can be clinically diagnosed into disability categories. Identification can help allocate resources more efficiently.

- <u>Description</u>: By describing the prevalence of disability in populations, we can better understand the resources necessary for treatment and early interventions

- <u>Prediction</u>: Prediction of prevalence of disability can aid in allocation of services

Latent Class Model

- M: number of latent classes
- K: number of tasks
- p_{km}: probability of having difficulty with item k given latent class m.
- π_m: probability of being in class m
- η_i: the true latent class of individual i, i = 1,…,N
- m = 1,…,M; k = 1,…,K



Note: this notation is different than McCutcheon

The latent class model is useful when….

- Items measure “diagnoses” rather than underlying scores
- Patterns of responses are thought to contain information above and beyond “aggregation” of responses
- The goal is “clustering” individuals rather than response variables

Choosing Items

- Descriptive Analysis
  - pattern frequencies: look at prevalences and 2x2 tables
  - multi-dimensional tables
  - recall hierarchy from association lecture.
- Want items that have variability
  - not useful to include p(difficulty eating) if prevalence is close to 0
  - not useful to include p(difficulty running 8 minute mile) if prevalence is close to 1.
- Do all items behave in same “direction”?
  - ”do you have difficulty walking up 10 steps?”
  - “is it easy for you to walk up 10 steps?”
- <u>Taken together, items should “define” construct</u>

  (issue of validity!)

Example: 2 Classes of Disability

	Class 1	Class 2
Lift	0.07	0.73
Walk	0.15	0.89
Step	0.04	0.66
Chair	0.02	0.46
Hhw	0.15	0.91
Class Size	0.67	0.33

p_{32}

π_1

2, 3, and 4 class models

	2 Class Model		3 Class Model			4 Class Model			
	Class 1	Class 2	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 4
Lift	0.07	0.73	0.04	0.42	0.85	0.06	0.48	0.29	0.85
Walk	0.15	0.89	0.10	0.58	0.96	0.09	0.58	0.72	0.96
Step	0.04	0.66	0.03	0.23	0.91	0.02	0.22	0.28	0.89
Chair	0.02	0.46	0.02	0.13	0.60	0.02	0.12	0.00	0.62
Hhw	0.15	0.91	0.06	0.78	0.93	0.06	0.78	0.71	0.94
Class size	0.67	0.33	0.56	0.25	0.19	0.53	0.19	0.03	0.20

Recall Binomial Distribution….

If we know the probability of difficulty with items 1 and 2 for a woman AND we can assume that they are independent:

$$P(y_{i1}) = p_1^{y_{i1}}(1 - p_1)^{(1-y_{i1})}$$

$$P(y_{i1}, y_{i2}) = \left(p_1^{y_{i1}}(1 - p_1)^{(1-y_{i1})}\right) \times \left(p_2^{y_{i2}}(1 - p_2)^{(1-y_{i2})}\right)$$

$$= \prod_{k=1}^{2} p_k^{y_{ik}}(1 - p_k)^{(1-y_{ik})}$$

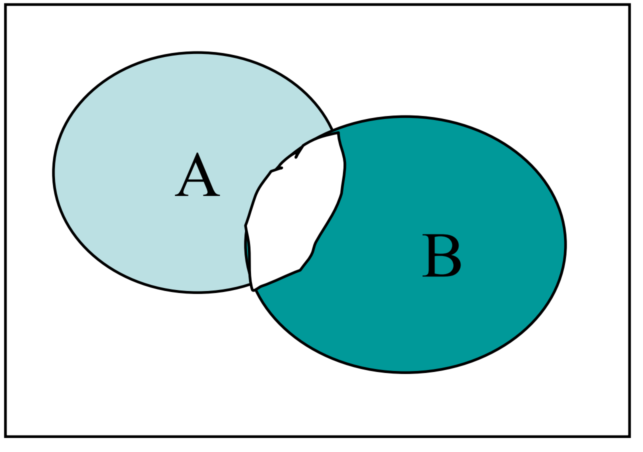
Using class specific-probabilities and K items, the probability of a woman reporting a specific response pattern given she is in class m is defined by:

$$P(\underset{\sim}{y_i} | \eta = m) = \prod_{k=1}^{K} p_{km}^{y_{ik}}(1 - p_{km})^{(1-y_{ik})}$$

Aside: Probability Concepts

$$P(A) = P(A \cap B) + P(A \cap B^C)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Leftrightarrow P(A|B)P(B) = P(A \bigcap B)$$



Deriving the Likelihood Function

We need to account for chance that a woman could be in any of the classes:

$$P(Y_i = y_i) = P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3}, Y_{i4} = y_{i4}, Y_{i5} = y_{i5})$$

$$= P(Y_i = y_i, \eta_i = 1) + P(Y_i = y_i, \eta_i = 2) + \ldots + P(Y_i = y_i, \eta_i = M)$$

$$= P(Y_i = y_i \mid \eta_i = 1) P(\eta_i = 1) + P(Y_i = y_i \mid \eta_i = 2) P(\eta_i = 2) +$$

$$\ldots + P(Y_i = y_i \mid \eta_i = M) P(\eta_i = M)$$

$$= \sum_{m=1}^{M} P(Y_i = y_i \mid \eta_i = m) P(\eta_i = m)$$

$$= \sum_{m=1}^{M} P(\eta_i = m) P(Y_i = y_i \mid \eta_i = m)$$

$$= \sum_{m=1}^{M} \pi_m \prod_{k=1}^{K} p_{km}^{y_{ik}} (1 - p_{km})^{(1 - y_{ik})}$$

Deriving the Likelihood Function

Take the product over all individuals in the dataset to get the likelihood function:

$$L(\pi, p \mid Y) = \prod_{i=1}^{N} \left[\sum_{m=1}^{M} \pi_m \prod_{k=1}^{K} p_{km}^{y_{ik}} (1 - p_{km})^{(1-y_{ik})} \right]$$

Issues r.e. Likelihood Function

- Local/Conditional Independence assumption allows us to take product over items
- Describes observed patterns (2^K possible patterns)
- Intuitively: sum up probability of observed pattern given membership in each of classes and weight by class prevalence.

item probabilities (conditional probabilities)

- prevalence of symptom in class
- degree of "measurement error"
- heterogeneity within a class
- If all p's in class 0 or 1: "no error" (one possible pattern in class
- If all p's in class 0.5: "noise" (every possible pattern has equal probability)

Aside: Probability Concepts

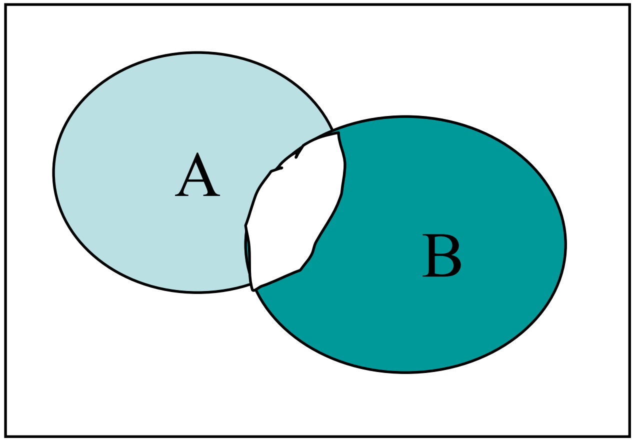
$$P(A) = P(A \cap B) + P(A \cap B^C)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Leftrightarrow P(A|B)P(B) = P(A \cap B)$$

<u>Bayes Rule:</u>

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$= \frac{P(B|A)P(A)}{\sum_i P(B_i)}$$



Clinically Relevant: Posterior Probabilities of Class Membership

$$P(\eta_i = j \mid Y_i = y_i) = \frac{P(\eta_i = j \text{ and } Y_i = y_i)}{P(Y_i = y_i)}$$

$$= \frac{P(Y_i = y_i \mid \eta_i = j)P(\eta_i = j)}{\sum_{m=1}^{M} P(Y_i = y_i \mid \eta_i = m)P(\eta_i = m)}$$

$$= \frac{\prod_{k=1}^{K} p_{kj}^{y_{ik}} (1 - p_{kj})^{(1-y_{ik})} \pi_m}{\sum_{m=1}^{M} \prod_{k=1}^{K} p_{km}^{y_{ik}} (1 - p_{km})^{(1-y_{ik})} \pi_m}$$

Example: What are the posterior probabilities of class membership for a woman who reports difficulty with only heavy housework and walking 1/2 mile?

2 class model

$$P(Y_i = 01001 \mid \eta_i = 1) = (1-0.07)(0.15)(1-0.04)(1-0.02)(0.15) = 0.02$$

$$P(Y_i = 01001 \mid \eta_i = 2) = (1-0.73)(0.89)(1-0.66)(1-0.46)(0.91) = 0.04$$

$$P(\eta_i = 1 \mid Y_i = 01001) = (0.02*0.67) / (0.02*0.67 + 0.04*0.33) = \mathbf{0.50}$$

$$P(\eta_i = 2 \mid Y_i = 01001) = (0.04*0.33) / (0.02*0.67 + 0.04*0.33) = \mathbf{0.50}$$

2, 3, and 4 class models

	2 Class Model		3 Class Model			4 Class Model			
	Class 1	Class 2	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 4
Lift	0.07	0.73	0.04	0.42	0.85	0.06	0.48	0.29	0.85
Walk	0.15	0.89	0.10	0.58	0.96	0.09	0.58	0.72	0.96
Step	0.04	0.66	0.03	0.23	0.91	0.02	0.22	0.28	0.89
Chair	0.02	0.46	0.02	0.13	0.60	0.02	0.12	0.00	0.62
Hhw	0.15	0.91	0.06	0.78	0.93	0.06	0.78	0.71	0.94
Class size	0.67	0.33	0.56	0.25	0.19	0.53	0.19	0.03	0.20

3 class model

$P(Y_i = 01001 | \eta_i = 1) = (1-0.04)(0.10)(1-0.03)(1-0.02)(0.06)$

$= 0.0055$

$P(Y_i = 01001 | \eta_i = 2) = (1-0.42)(0.58)(1-0.23)(1-0.13)(0.78)$

$= 0.176$

$P(Y_i = 01001 | \eta_i = 3) = (1-0.85)(0.96)(1-0.91)(1-0.60)(0.93)$

$= 0.0048$

$P(\eta_i = 1 | Y_i = 01001) = (0.0055*0.56) / (0.0055*0.56 + 0.176*0.25 + 0.0048*0.19)$

= **0.06**

$P(\eta_i = 2 | Y_i = 01001) = (0.176*0.25) / (0.0055*0.56 + 0.176*0.25 + 0.0048*0.19)$

=**0.92**

$P(\eta_i = 3 | Y_i = 01001) = (0.0048*0.19) / (0.0055*0.56 + 0.176*0.25 + 0.0048*0.19)$

= **0.02**

Estimation Approaches

Maximum Likelihood Approach:

Find estimates of p, π, and η that are most consistent with the data that we observe conditional on number of classes, *M*.

Often used: EM algorithm (iterative fitting procedure)

Bayesian Approach:

Quantify beliefs about p, π, and η before and after observing data.

Often used: Gibbs sampler, MCMC algorithm

Bayesian Terminology

- <u>Prior Probability</u>: What we believe about unknown parameters **before** observing data.

- <u>Posterior Probability</u>: What we believe about the parameters **after** observing data.

Bayesian Estimation Approach

- Specify prior probability distribution:

$$P(p, \pi, \eta)$$

- Combine prior with likelihood to obtain posterior distribution:

$$P(p, \pi, \eta | Y) \propto P(p, \pi, \eta) \times L(Y | p, \pi, \eta)$$

- Estimate posterior distribution for each parameter using iterative procedure.

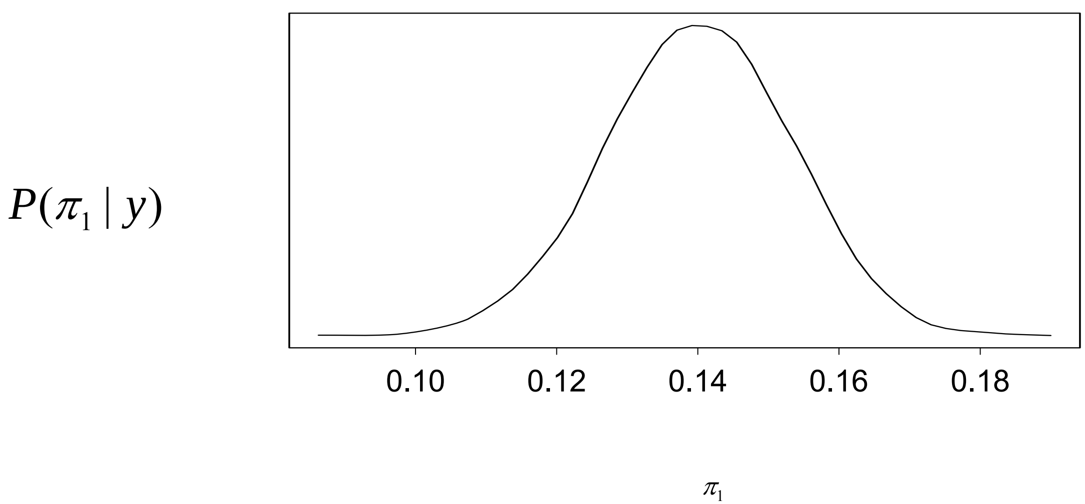
$$P(\pi_1 | Y) = \int P(p, \pi, \eta | Y)$$

Bayesian Estimation Approach

The <u>Gibbs Sampler</u> is an iterative process used to estimate posterior distributions of parameters.

- we sample parameters from conditional distributions

  e.g. $P(\pi_1|Y, p, \eta, \pi_2, \pi_3)$
- At each iteration, we get 'sampled' values of p, π, and η.
- We use the samples from the iterations to estimate posterior distributions by averaging over other parameter values.



Estimation Issues

- ML estimation: Initial values needed
  - some programs provide them
  - some require you to provide them
- Bayesian: Choosing parameter estimates
  - mean, median, or mode of posterior?
- Convergence: both ML and Bayes
  - may be more than one “optimal” solution
  - different starting values can give different solutions
- Time consuming

Caveat with ML estimation

- Once p_{km} reaches an estimate of 0 or 1, it will never move
- Recall iterative fitting procedure.
- If results give 0 or 1, rerun with different <u>starting/initial</u> values
- NEVER give 0 or 1 as starting value for p_{km} unless you want to “force” it (more later).

Software for LC analysis

- Maximum Likelihood Estimators:
  - *Latent variable specific:*
    - ***Mplus***
    - LEM
    - LatentGOLD
    - MLLSA
    - LISREL
  - SAS, Splus, Mathematica, etc
- Bayesian Estimators:
  - *WinBugs*

- http://ourworld.compuserve.com/homepages/jsuebersax/soft.htm

Choosing Estimation Procedure

- Inference should NOT depend on estimation procedure!
- “Philosophical”: Bayesian versus Frequentist.
- Pros:
  - ML: intuitive, canned packages
  - Bayesian: can assess identifiability, more certainty in “convergence to truth”
- Cons:
  - ML: solution can depend on starting values
  - Bayesian: estimation procedure more complicated

Latent Class Output

- Not always do we get p_{km} and π_m
- Very common to see the 'logit' transformation used.
- On logit scale, γ_{km} ranges from $-\infty$ to ∞.
  - If $\gamma_{km} = 4$, $p_{km} = 0.98$
  - If $\gamma_{km} = 2$, $p_{km} = 0.88$
  - If $\gamma_{km} = 1$, $p_{km} = 0.73$
  - If $\gamma_{km} = 0.5$, $p_{km} = 0.62$
  - If $\gamma_{km} = 0$, $p_{km} = 0.5$

$$p_{km} = \frac{e^{\gamma_{km}}}{1 + e^{\gamma_{km}}}$$

$$\gamma_{km} = \log\left(\frac{p_{km}}{1 - p_{km}}\right)$$

What if $\gamma_{km} = -4$?

Latent Class Output

- How do we tell if our estimates ‘got stuck’ in Mplus?
- Technically, for p = 0, then

$$\mathrm{logit}(0) = \log(\tfrac{0}{1}) = \log(0) = -\infty$$

- And for p = 1, then

$$\mathrm{logit}(1) = \log(\tfrac{1}{0}) = \log(\infty) = \infty$$

- But, Mplus doesn’t give you these: If you see logit values of -15 or +15, then you “GOT STUCK”!!!!!!!!!

Identifiability of Models

- Are there several solutions? That is, are there two sets of parameters?
  - **Necessary** condition for "theoretical identifiability":
    - Number parameters is smaller than the number of possible patterns (minus 1)
    - $M-1+M*K < 2^K - 1$
  - **Caveat**: Not **sufficient** condition.
    - 3 class model not identifiable with 4 items.
- "Empirical identifiability": Is there enough evidence in the data to estimate all of the parameters in the model? Is the data set big enough?

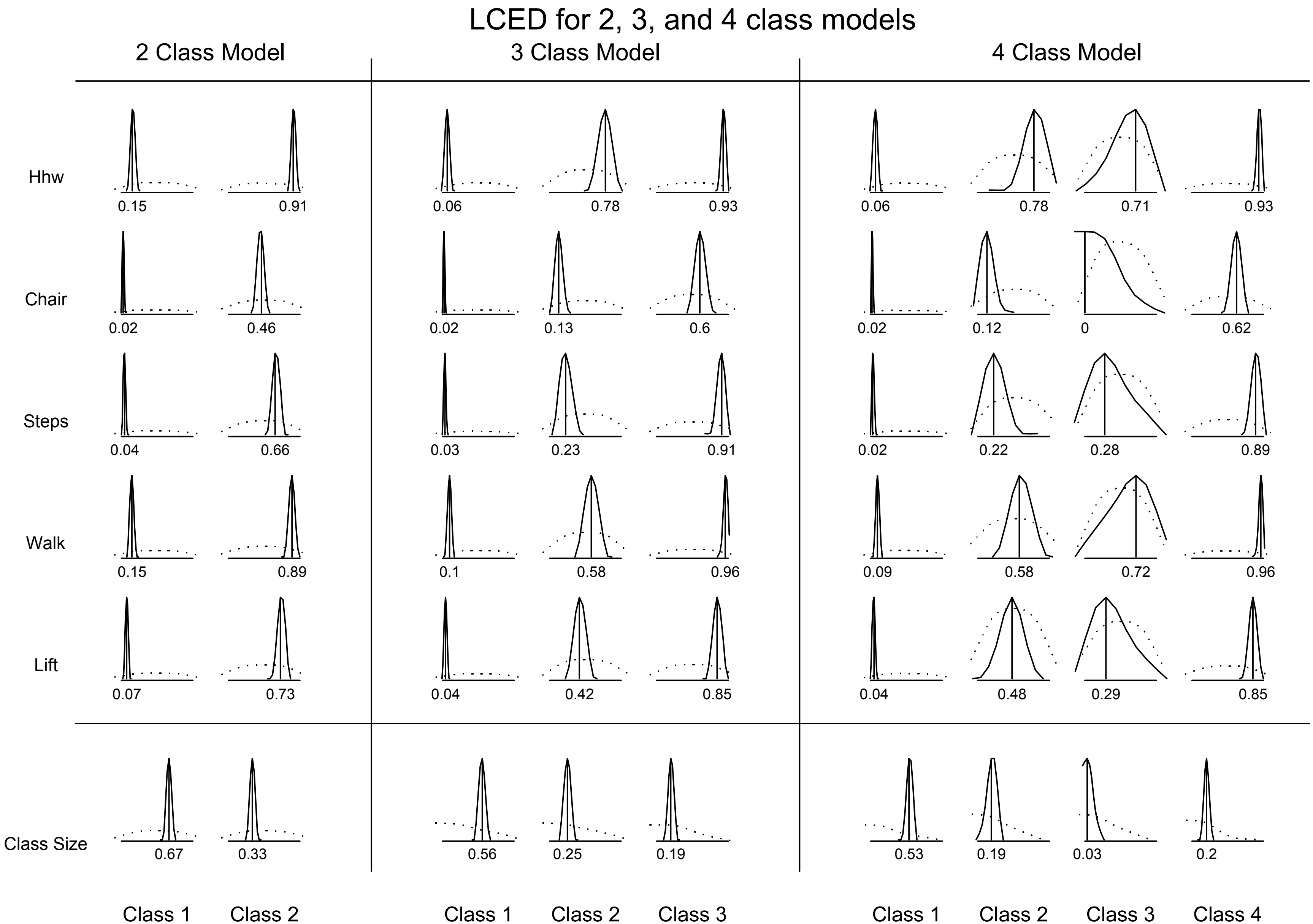
Ways to improve identifiability

- Put constraints on item probabilities, *p's.*
  - e.g. set p_{km} to 0 or 1
  - Or, for example, $p_{km}>0.90$
  - must have good scientific rationale

- More data!

Investigating Identifiability via Bayesian Approach

- Look at how much reliance is placed on prior distribution
- Compare **posterior** distribution of parameter to **prior** distribution of parameter
- If prior and posterior are similar, then we say model is only “weakly identifiable” or “weakly estimable”
- Latent Class Estimability Display (LCED)


LCED for 2, 3, and 4 class models
2 Class Model
3 Class Model
4 Class Model
Hhw
Chair
Steps
Walk
Lift
Class Size
0.15
0.91
0.06
0.78
0.93
0.06
0.78
0.71
0.93
0.02
0.46
0.02
0.13
0.6
0.02
0.12
0
0.62
0.04
0.66
0.03
0.23
0.91
0.02
0.22
0.28
0.89
0.15
0.89
0.1
0.58
0.96
0.09
0.58
0.72
0.96
0.07
0.73
0.04
0.42
0.85
0.04
0.48
0.29
0.85
0.67
0.33
0.56
0.25
0.19
0.53
0.19
0.03
0.2
Class 1
Class 2
Class 1
Class 2
Class 3
Class 1
Class 2
Class 3
Class 4

“Estimability”

- Is there enough data to estimate all of the parameters
- How can we tell if we should “believe” the symptoms probabilities in a class?
- Consider precision of estimates
  - ML: confidence interval
  - Bayesian: posterior interval

Assessing Estimability: 3 class model

	estimate	se	95% confidence interval 2.5%	97.5%	width of 95%CI
pi[1]	0.25	0.036	0.18	0.32	0.14
pi[2]	0.55	0.033	0.48	0.61	0.13
pi[3]	0.20	0.028	0.15	0.26	0.11
p[1,1]	0.42	0.06	0.31	0.54	0.23
p[1,2]	0.04	0.01	0.01	0.07	0.06
p[1,3]	0.84	0.04	0.76	0.93	0.17
p[2,1]	0.59	0.07	0.45	0.73	0.28
p[2,2]	0.09	0.02	0.05	0.14	0.09
p[2,3]	0.96	0.02	0.91	0.99	0.08
p[3,1]	0.22	0.06	0.11	0.35	0.24
p[3,2]	0.03	0.01	0.01	0.06	0.05
p[3,3]	0.89	0.04	0.80	0.97	0.17
p[4,1]	0.13	0.04	0.06	0.21	0.15
p[4,2]	0.02	0.01	0.01	0.04	0.03
p[4,3]	0.61	0.06	0.51	0.73	0.22
p[5,1]	0.76	0.06	0.63	0.88	0.25
p[5,2]	0.06	0.02	0.02	0.11	0.09
p[5,3]	0.93	0.03	0.87	0.98	0.11

	estimate	se	95% confidence interval 2.5%	97.5%	width of 95%CI
pi[1]	0.08	0.07	0.01	0.28	0.27
pi[2]	0.20	0.02	0.16	0.25	0.09
pi[3]	0.18	0.07	0.02	0.28	0.26
pi[4]	0.54	0.03	0.48	0.60	0.12
p[1,1]	0.37	0.22	**0.06**	**0.92**	**0.86**
p[1,2]	0.85	0.04	0.76	0.92	0.16
p[1,3]	0.45	0.13	0.16	0.74	**0.58**
p[1,4]	0.04	0.01	0.01	0.07	0.06
p[2,1]	0.62	0.22	**0.12**	**0.96**	**0.84**
p[2,2]	0.96	0.02	0.91	0.99	0.08
p[2,3]	0.60	0.12	0.32	0.87	0.55
p[2,4]	0.09	0.02	0.05	0.13	0.08
p[3,1]	0.34	0.23	**0.04**	**0.88**	**0.84**
p[3,2]	0.89	0.05	0.79	0.98	0.21
p[3,3]	0.25	0.14	0.08	0.71	**0.63**
p[3,4]	0.03	0.01	0.01	0.05	0.04
p[4,1]	0.23	0.19	**0.03**	**0.78**	**0.75**
p[4,2]	0.61	0.05	0.50	0.72	0.22
p[4,3]	0.15	0.09	0.05	0.39	0.34
p[4,4]	0.02	0.01	0.01	0.04	0.03
p[5,1]	0.63	0.20	**0.14**	**0.94**	**0.80**
p[5,2]	0.94	0.03	0.88	0.98	0.10
p[5,3]	0.76	0.13	0.33	0.94	**0.61**
p[5,4]	0.06	0.02	0.02	0.11	0.09

4 class
model

How do we choose among models?

- Likelihood Ratio?
  - Assumes models are “nested”
  - But, 2 class model is not a subset of 3 class model
  - Parameters have different interpretations
- Other procedures?
  - Information Criterion (AIC, BIC)
  - Graphical diagnostics

Model Fit:
How can we compare model to observed data?

- What is observed? patterns frequencies
- Measure “distance” between predicted and observed patterns
- Statistics:

  -2Log-Likelihood (-2LL) (see next page for note)

  Information Criteria: Pick model with smallest IC

  Akaike: AIC = -2LL + 2*s

  Schwarz: BIC = -2LL + s*log(N)

  (s = number of parameters in model)

  (N = sample size)

Traditional Fit Statistics

	"2LL"	AIC	BIC	s
2 class model	3478.30	3500.30	3551.19	11
3 class model	3403.71	3437.71	3516.36	17
4 class model	3413.35	3524.30	3630.71	23

(Note: LL is not really log-likelihood in this case because ML was not procedure used for estimation. It is the log-likelihood based on the Bayesian parameter estimates.)

Goodness of Fit

- Assumption:
  - 2LL is valid goodness of fit statistic assuming that cell counts are large
  - That is, if the number of individuals reporting each pattern is relatively large (at <u>least</u> 5)
- Generally true in LC analysis?
- Also, 2LL has been shown to be best for sample sizes around 200-300 (for larger, it is conservative).

Another approach: Estimating Pattern Prevalence

$$
\begin{aligned}
P(Y_i = y_i) &= \sum_{m=1}^{M} P(Y_i = y_i \text{ and } \eta_{\mathrm{i}} = m) \\
&= \sum_{m=1}^{M} P(Y_i = y_i \mid \eta_{\mathrm{i}} = m) P(\eta_{\mathrm{i}} = m) \\
&= \sum_{m=1}^{M} \left(\prod_{k=1}^{K} p_{mk}^{y_{ik}} (1 - p_{mk})^{1 - y_{ik}} \right) \pi_m
\end{aligned}
$$

Example: Probability of Reporting "Walk" and "Hhw" in 3 class model

Recall:

$P(Y_i = 01001|\eta_i = 1) = (1\text{-}0.04)(0.10)(1\text{-}0.03)(1\text{-}0.02)(0.06) = 0.0055$

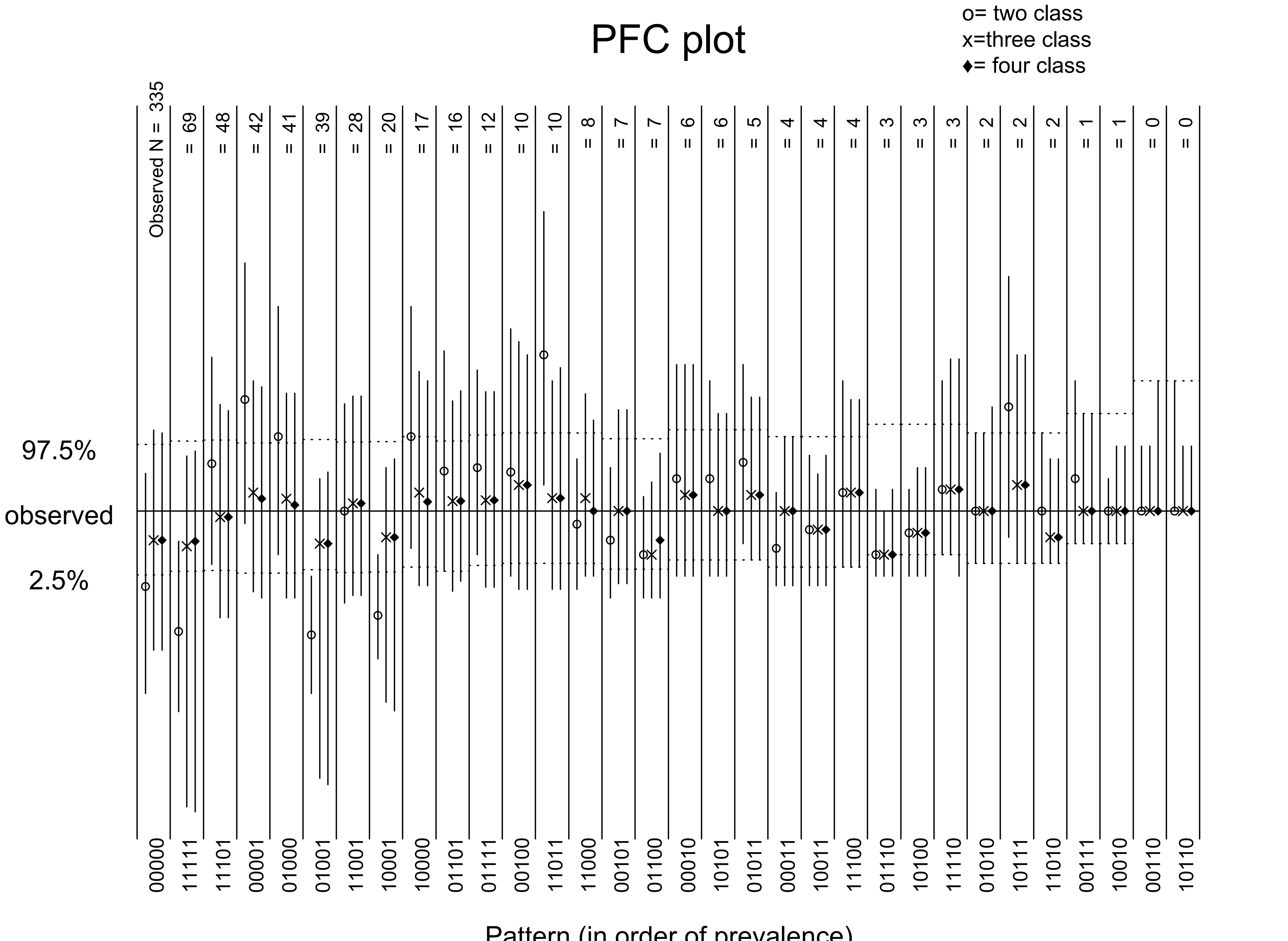
$P(Y_i = 01001|\ \eta_i = 2) = (1\text{-}0.42)(0.58)(1\text{-}0.23)(1\text{-}0.13)(0.78) = 0.176$

$P(Y_i = 01001|\ \eta_i = 3) = (1\text{-}0.85)(0.96)(1\text{-}0.91)(1\text{-}0.60)(0.93) = 0.0048$

and $P(\eta_i = m) = \pi_m = (0.56,0.25,0.19)$

➡ $P(Y_i = 01001) = 0.0055*0.56 + 0.176*0.25 + 0.0048*0.19 = 0.05$

- That is, 5% of the sample is "predicted" to report this pattern based on the <u>three class model</u>. Recall sample size is 755.
- Observed number: **39**
- Expected number: 0.05*755 = **37.75**
- FYI: Two class model
  - Percent expected is 2.7%
  - Expected number: 0.027*755 = **20.38**



PFC plot
o= two class
x=three class
♦= four class
Observed N = 335
= 69
= 48
= 42
= 41
= 39
= 28
= 20
= 17
= 16
= 12
= 10
= 10
= 8
= 7
= 7
= 6
= 6
= 5
= 4
= 4
= 4
= 3
= 3
= 3
= 2
= 2
= 2
= 1
= 1
= 0
= 0
97.5%
observed
2.5%
00000
11111
11101
00001
01000
01001
11001
10001
10000
01101
01111
00100
11011
11000
00101
01100
00010
10101
01011
00011
10011
11100
01110
10100
11110
01010
10111
11010
00111
10010
00110
10110
Pattern (in order of prevalence)

Guttman Scaling

- A type of latent class model
- Different than what we've seen so far
- So far:
  - Identify items
  - Find structure/definitions of classes
- Guttman
  - Define structure
  - See if data are consistent with structure

Guttman Scaling

- "Pure" hierarchy
- 756 individuals who have used any illicit drugs at least 5 times
  - 97% used marijuana
  - 24% used cocaine
  - 13% used heroin
- K items ► K+1 classes
- **Very restrictive** latent class model.

Guttman Model: 4 classes

	Marijuana	Cocaine	Heroin
Class 1:	0	0	0
Class 2:	1	0	0
Class 3:	1	1	0
Class 4:	1	1	1

Measurement Error

- What if you report only cocaine (010) ?
- 8 possible reporting patterns with 3 binary items (2^3)
- “Allowed” vs. “Disallowed” patterns
  - Allowed: 000, 100, 110, 111
  - Disallowed: 010, 001, 101, 011
- Measurement error:
  - An “error” in reporting
  - An ‘aberrant’ user—one who doesn’t follow the normal pattern of use.
- 2 general types of models:
  - **assume that there is a measurement error associated with each drug.**
  - assume that there is a measurement error associated with each class.

What about those with "disallowed" patterns?

- Based on response pattern, they get posterior probabilities of class membership
- (And so do those with allowed patterns)
- But, class definitions are 'predefined'
- However, actual values of 'p_{km}' are not defined: they depend on measurement errors.
- Parameters estimated:
  - Class sizes
  - Amount of measurement error (either by item (drug), or by class).
- <u>Drug use example</u>: 4 class sizes, and three measurement error parameters (one for each drug).
  - p_m = measurement error for marijuana = 0.03
  - p_c = measurement error for cocaine = 0.04
  - p_h = measurement error for heroin = 0.04

Guttman Model: Latent Class Presentation

	"Non-User" Class 1 (000)	*"Light User"* Class 2 (100)	*"Moderate User"* Class 3 (110)	*"Heavy User"* Class 4 (111)
Marijuana	p_m	$1-p_m$	$1-p_m$	$1-p_m$
Cocaine	p_c	p_c	$1-p_c$	$1-p_c$
Heroin	p_h	p_h	p_{3h}	$1-p_h$
π	π_1	π_2	π_3	π_4

Guttman Model: Latent Class Presentation

	"Non-User" Class 1 (000)	*"Light User"* Class 2 (100)	*"Moderate User"* Class 3 (110)	*"Heavy User"* Class 4 (111)
Marijuana	0.03	0.97	0.97	0.97
Cocaine	0.04	0.04	0.96	0.96
Heroin	0.04	0.04	0.04	0.96
π	0.00007	0.78	0.12	0.10

Fitted Model Results versus Observed Data

Marijuana	Cocaine	Heroin	Predicted Prevalence of pattern	Observed Prevalence of Pattern
0	**0**	**0**	**0.02**	**0.01**
1	**0**	**0**	**0.70**	**0.72**
0	1	0	0.004	0.003
0	0	1	0.001	0.008
1	**1**	**0**	**0.14**	**0.14**
1	0	1	0.03	0.03
0	1	1	0.003	0.01
1	**1**	**1**	**0.10**	**0.09**

What are the posterior class probabilities for someone reporting only cocaine?

What are the posterior probabilites of class membership for the pattern 010?

"Non-user":

$$P(\text{true } 000 \mid \text{report } 010) = \frac{P(\text{report } 010 \mid \text{true } 000)\text{P}(\text{true } 000)}{\text{P}(\text{report } 010)}$$

$$= \frac{(0.97)(0.04)(0.96)(0.00007)}{0.0043} < 0.001$$

"Light user":

$$P(\text{true } 100 \mid \text{report } 010) = \frac{(0.03)(0.04)(0.96)(0.78)}{0.0043} = 0.21$$

"Moderate user":

$$P(\text{true } 110 \mid \text{report } 010) = \frac{(0.03)(0.96)(0.96)(0.12)}{0.0043} = 0.77$$

"Heavy user":

$$P(\text{true } 111 \mid \text{report } 010) = \frac{(0.03)(0.96)(0.04)(0.10)}{0.0043} = 0.03$$

What are the posterior class probabilities for someone reporting marijuana and cocaine?

What are the posterior probabilites of class membership for the pattern 110?

"Non-user": $$P(\text{true } 000 \mid \text{report } 110) = \frac{P(\text{report } 110 \mid \text{true } 000)\text{P}(\text{true } 000)}{\text{P}(\text{report } 110)}$$

$$= \frac{(0.03)(0.04)(0.96)(0.00007)}{0.14} < 0.00001$$

"Light user": $$P(\text{true } 100 \mid \text{report } 110) = \frac{(0.97)(0.04)(0.96)(0.78)}{0.14} = 0.21$$

"Moderate user": $$P(\text{true } 110 \mid \text{report } 110) = \frac{(0.97)(0.96)(0.96)(0.12)}{0.14} = 0.77$$

"Heavy user": $$P(\text{true } 111 \mid \text{report } 110) = \frac{(0.97)(0.96)(0.04)(0.10)}{0.14} = 0.03$$