Latent class analysis

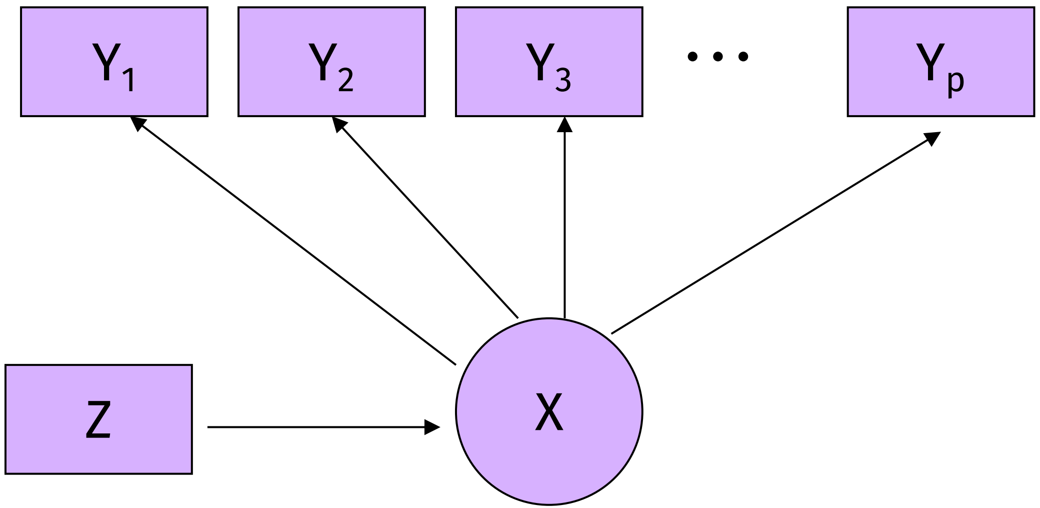
LCA Basics

DL Oberski & L Boeschoten

	Models for means	
	Latent	
	Continuous	Discrete
Observed		
Continuous	Factor analysis	**Latent profile analysis**
Discrete	Item response theory	**Latent class analysis**

Table 1 Names of different kinds of latent variable models.

The Latent Class Model



- Observed (continuous or) **categorical** Items (**Y**)
- **Categorical** Latent Class Variable (X)
- Continuous or Categorical Covariates (Z)

Four main applications of LCM

- Clustering (model based / probabilistic)
- Scaling (discretized IRT/factor analysis)
- Random-effects modelling (mixture regression / nonparametric multilevel)
- Density estimation

Why latent class models?

For substantive analysis:

- Creating typologies of respondents, e.g.:
  - McCutcheon 1989: tolerance,
  - Rudnev 2015: human values
  - Savage et al. 2013: “A new model of Social Class”
  - …
- Nonparametric multilevel model (Vermunt 2013)
- Longitudinal data analysis
  - Growth mixture models
  - Latent transition (“Hidden Markov”) models

Why would survey researchers need latent class models?

For survey methodology:

- As a method to evaluate questionnaires, e.g.
  - Biemer 2011: Latent Class Analysis of Survey Error
  - Oberski 2015: latent class MTMM
- Modeling extreme response style (and other styles), e.g.
  - Morren, Gelissen & Vermunt 2012: extreme response
- Measurement equivalence for comparing groups/countries
  - Kankaraš & Moors 2014: Equivalence of Solidarity Attitudes
- Identifying groups of respondents to target differently
  - Lugtig 2014: groups of people who drop out panel survey
- Flexible imputation method for multivariate categorical data
  - Van der Palm, Van der Ark & Vermunt

A small example

(showing the basic ideas and interpretation)

Small example: data from GSS 1987

Y1: “allow anti-religionists to speak” (1 = allowed, 2 = not allowed),
Y2: “allow anti-religionists to teach” (1 = allowed, 2 = not allowed),
Y3: “remove anti-religious books from the library” (1 = do not remove, 2 = remove).

Y1	Y2	Y3	Observed frequency (n)	Observed proportion (n/N)
1	1	1	696	0.406
1	1	2	68	0.040
1	2	1	275	0.161
1	2	2	130	0.076
2	1	1	34	0.020
2	1	2	19	0.011
2	2	1	125	0.073
2	2	2	366	0.214

N = 1713

Estimating the 2-class model in `R`

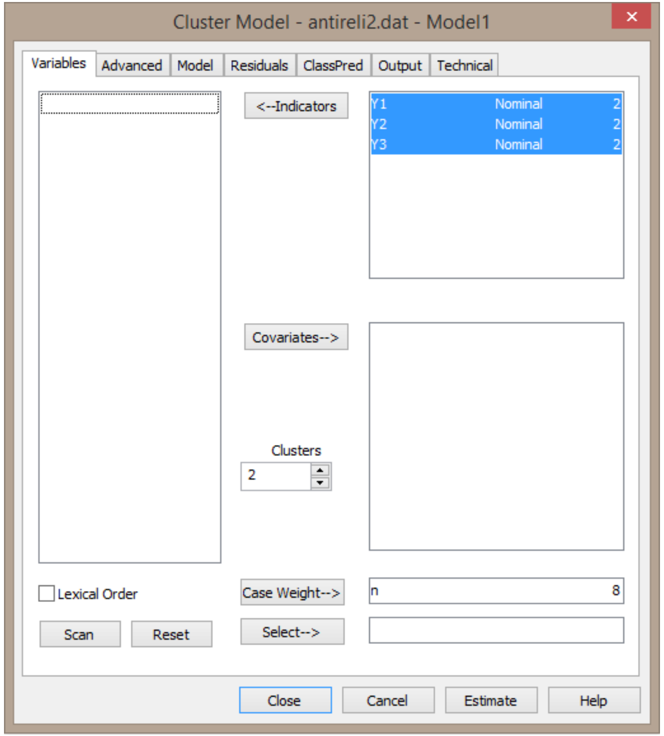
```
antireli  <- read.csv("antireli_data.csv")

library(poLCA)

M2 <- poLCA(cbind(Y1, Y2, Y3)~1, data=antireli, nclass=2)
```

Profile for 2-class model

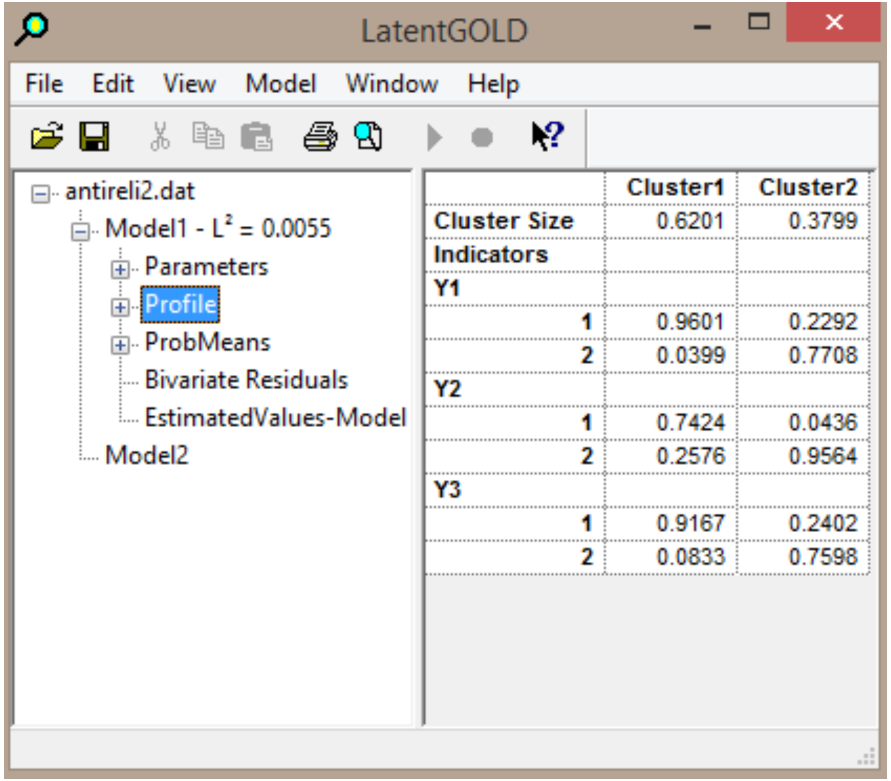
```
$Y1
            Pr(1)  Pr(2)
class 1:  0.9601 0.0399
class 2:  0.2284 0.7716

$Y2
            Pr(1)  Pr(2)
class 1:  0.7424 0.2576
class 2:  0.0429 0.9571

$Y3
            Pr(1)  Pr(2)
class 1:  0.9166 0.0834
class 2:  0.2395 0.7605
```

```
Estimated class population shares
 0.6205 0.3795
```

2-class model in Latent GOLD

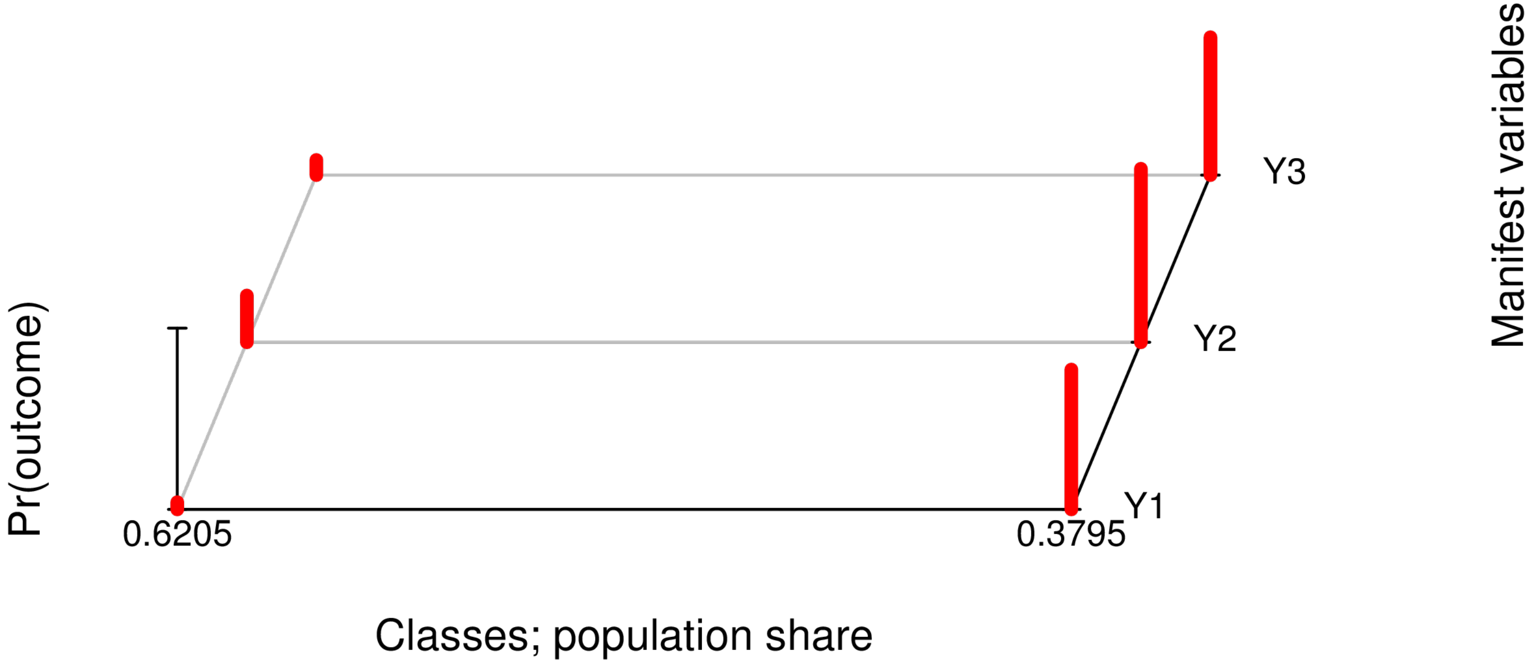


Profile for 2-class model



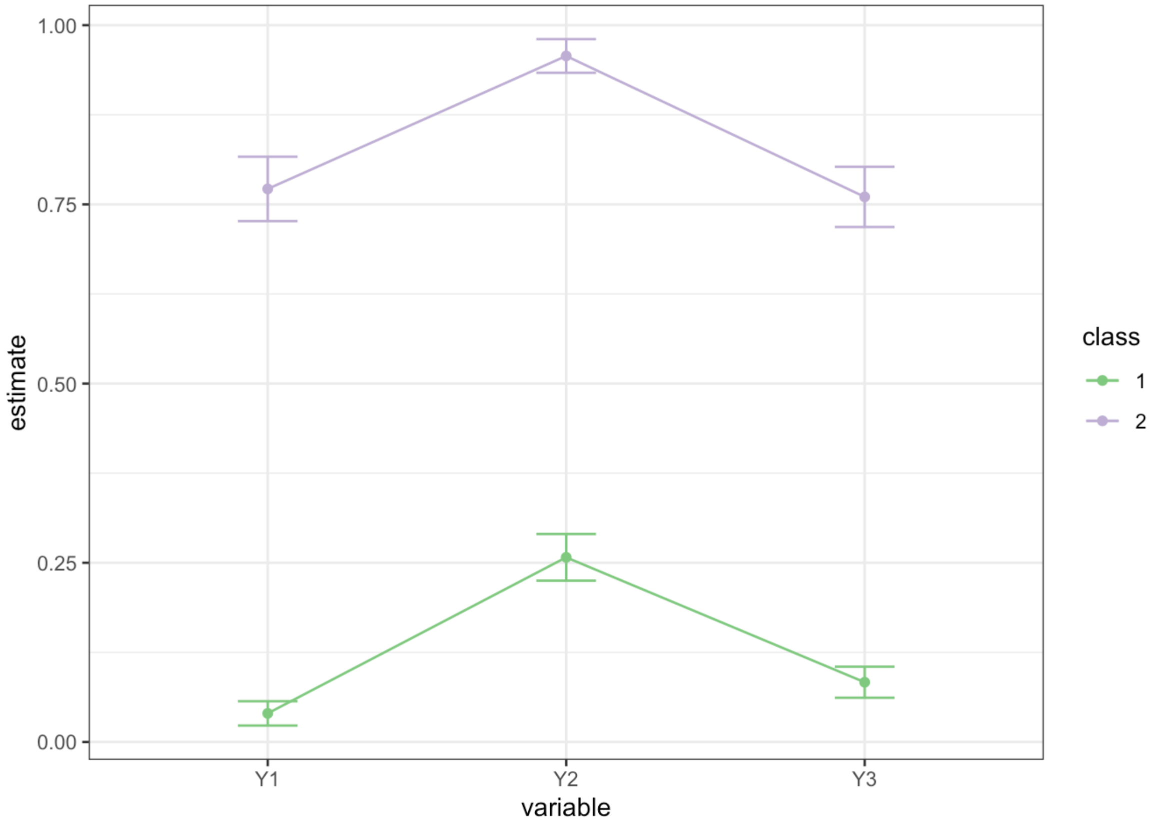
	Cluster1	Cluster2
Cluster Size	0.6201	0.3799
Indicators		
Y1		
1	0.9601	0.2292
2	0.0399	0.7708
Y2		
1	0.7424	0.0436
2	0.2576	0.9564
Y3		
1	0.9167	0.2402
2	0.0833	0.7598

```
> plot(M2)
```



Pr(outcome)
Manifest variables
Y3
Y2
Y1
0.6205
0.3795
Classes; population share

Profile plot for 2-class model



Model equation
for 2-class LC model for 3 indicators

Model for

$$P(y_1, y_2, y_3)$$

the probability of a particular response pattern.

For example, how likely is someone to hold the opinion
“allow speak, allow teach, but remove books from library:

P(Y1=1, Y2=1, Y3=2) = ?

Two key model assumptions

(X is the latent class variable)

1. (MIXTURE ASSUMPTION)

Joint distribution mixture of 2 class-specific distributions:

$$P(y_1, y_2, y_3) = P(X=1)P(y_1, y_2, y_3 \mid X=1) + P(X=2)P(y_1, y_2, y_3 \mid X=2)$$

2. (LOCAL INDEPENDENCE ASSUMPTION)

Within class $X=x$, responses are independent:

$$P(y_1, y_2, y_3 \mid X=1) = P(y_1 \mid X=1)P(y_2 \mid X=1)P(y_3 \mid X=1)$$

$$P(y_1, y_2, y_3 \mid X=2) = P(y_1 \mid X=2)P(y_2 \mid X=2)P(y_3 \mid X=2)$$

Example: model-implied proprtion

	X=1	X=2
P(X)	0.620	0.380
P(Y1=1|X)	0.960	0.229
P(Y2=1|X)	0.742	0.044
P(Y3=1|X)	0.917	0.240

P(Y1=1, Y2=1, Y3=2) =

(Mixture assumption)

P(Y1=1, Y2=1, Y3=2 | X=1) P(X=1) +

P(Y1=1, Y2=1, Y3=2 | X=2) P(X=2)

Example: model-implied proportion

	X=1	X=2
P(X)	0.620	0.380
P(Y1=1|X)	0.960	0.229
P(Y2=1|X)	0.742	0.044
P(Y3=1|X)	0.917	0.240

P(Y1=1, Y2=1, Y3=2) =

(Mixture assumption)

$P(Y1=1, Y2=1, Y3=2 \mid X=1)\ 0.620\ +$

$P(Y1=1, Y2=1, Y3=2 \mid X=2)\ 0.380\ =$

(Local independence assumption)

$P(Y1=1|X=1)\ P(Y2=1|X=1)\ P(Y2=2|X=1)\ 0.620\ +$

$P(Y1=1|X=2)\ P(Y2=1|X=2)\ P(Y2=2|X=2)\ 0.380$

Example: model-implied proprtion

	X=1	X=2
P(X)	0.620	0.380
P(Y1=1|X)	0.960	0.229
P(Y2=1|X)	0.742	0.044
P(Y3=1|X)	0.917	0.240

P(Y1=1, Y2=1, Y3=2) =

(Mixture assumption)

P(Y1=1, Y2=1, Y3=2 | X=1) 0.620 +

P(Y1=1, Y2=1, Y3=2 | X=2) 0.380 =

(Local independence assumption)

(0.960) (0.742) (1-0.917) (0.620) +

(0.229) (0.044) (1-0.240) (0.380) ≈

≈ 0.0396

Small example: data from GSS 1987

Y1: “allow anti-religionists to speak” (1 = allowed, 2 = not allowed),
Y2: “allow anti-religionists to teach” (1 = allowed, 2 = not allowed),
Y3: “remove anti-religious books from the library” (1 = do not remove, 2 = remove).

	Y1	Y2	Y3	Observed frequency (n)	Observed proportion (n/N)
	1	1	1	696	0.406
	1	1	2	68	0.040
	1	2	1	275	0.161
	1	2	2	130	0.076
	2	1	1	34	0.020
	2	1	2	19	0.011
	2	2	1	125	0.073
	2	2	2	366	0.214

N = 1713

Implied is 0.0396, observed is 0.040.

Activity

You can play around with the implied probabilities in the Excel file on the course website!

(thanks to Jeroen Vermunt).