Latent Class Analysis: Assumptions, Equations, and Estimation

Jeroen K. Vermunt
Department of Methodology and Statistics, Tilburg University
www.jeroenvermunt.nl

Introduction

- I will discuss model assumptions and equations
- I will explain maximum likelihood estimation

- I will use a data set with three dichotomous observed variables from GSS 1987 (antireli.dat)
    - Y_1="allow anti-religionists to speak"(1 = allowed, 2 = not allowed)
    - Y_2="allow anti-religionists to teach" (1 = allowed, 2 = not allowed)
    - Y_3="remove anti-religious books from the library" (1 = do not remove, 2 = remove)

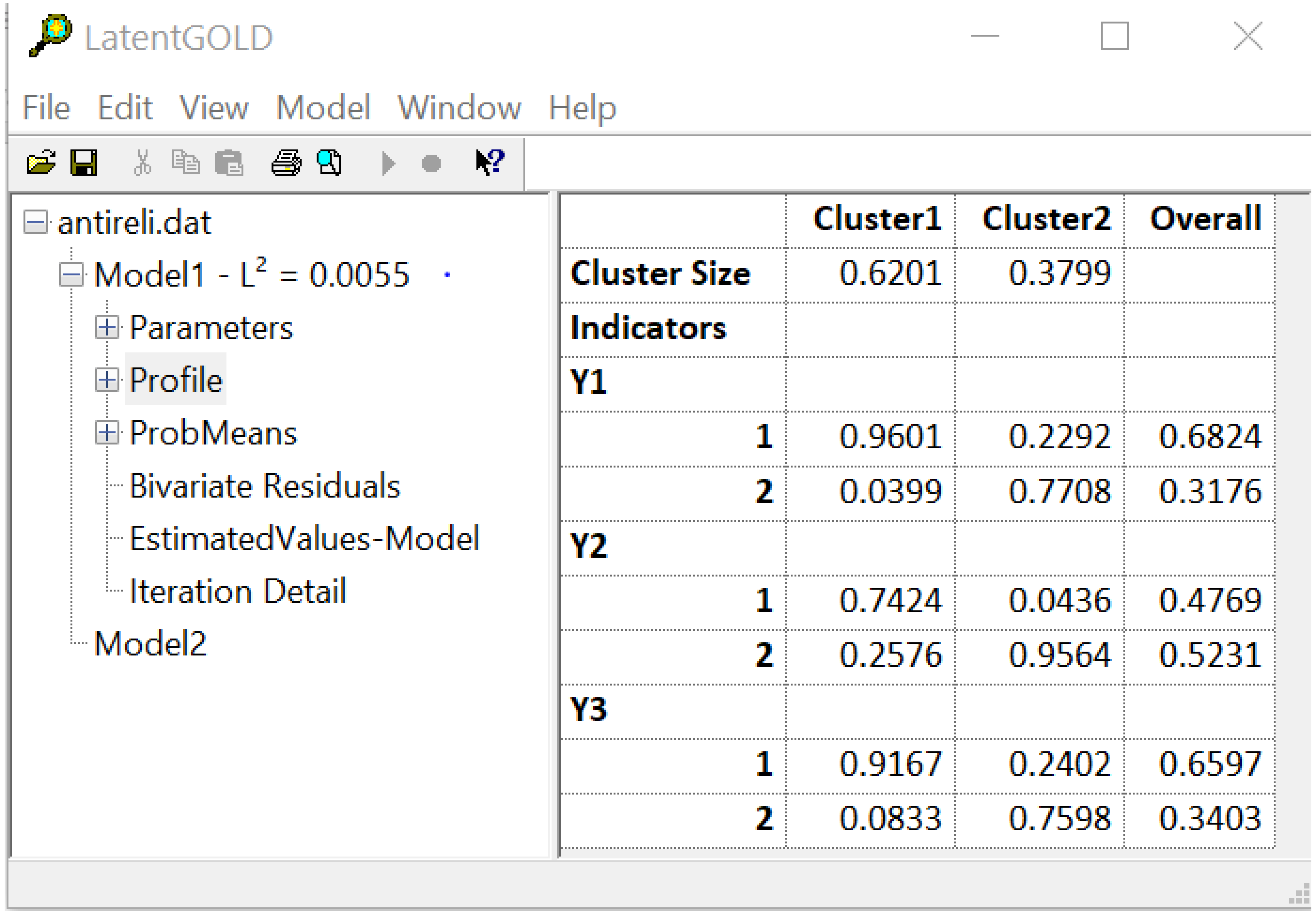
- This is the smallest possible application of LC analysis

Data set in the form of a multidimensional frequency table

Y_1	Y_2	Y_3	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

Latent GOLD Profile output for 2-class model

(caseweight=n; indicators=nominal)



	Cluster1	Cluster2	Overall
Cluster Size	0.6201	0.3799	
Indicators			
Y1			
1	0.9601	0.2292	0.6824
2	0.0399	0.7708	0.3176
Y2			
1	0.7424	0.0436	0.4769
2	0.2576	0.9564	0.5231
Y3			
1	0.9167	0.2402	0.6597
2	0.0833	0.7598	0.3403

Cluster1 = liberals
Cluster2 = conservatives

Questions of interest …

- How does the statistical model look like that we are estimating?
- How is the estimation of this model performed?
- Or … opening the black box!

- Some notation:
  - The discrete latent variable is called X
  - The observed variables are called y_1, y_2, y_3. etc.

Assumptions 2-class model for y_1, y_2, and y_3

- We define a model for $P(y_1, y_2, y_3)$, the joint probability of a particular response pattern
- Two key model assumptions:

  1. The joint probability/distribution $P(y_1, y_2, y_3)$ is a mixture of 2 class-specific distributions (some persons with this pattern are from class 1 and others from class 2)

  2. Within class X=1 and X=2, responses are independent (local independence)

  (knowing your response on y_1 doesn't tell me anything about y_2 if I know your class membership)

Equations of 2-class model for y_1, y_2, and y_3

1. Joint probability is a <u>mixture</u> of 2 class-specific distributions

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

2. Within classes responses are independent (<u>local independence</u>)

$$P(y_1,y_2,y_3|X=1) = P(y_1|X=1)\,\mathrm{P}(y_2|X=1)\,\mathrm{P}(y_3|X=1)$$

$$P(y_1,y_2,y_3|X=2) = P(y_1|X=2)\,\mathrm{P}(y_2|X=2)\,\mathrm{P}(y_3|X=2)$$

Using the numbers from the Profile output

$P(y_1\!=\!1, y_2\!=\!1, y_3\!=\!1) = 0.620 * 0.653 + 0.380 * 0.002 = \underline{0.406}$

$P(y_1\!=\!1, y_2\!=\!1, y_3\!=\!1 \mid X\!=\!1) = 0.960 * 0.742 * 0.917 = 0.653$

$P(y_1\!=\!1, y_2\!=\!1, y_3\!=\!1 \mid X\!=\!2) = 0.229 * 0.044 * 0.240 = 0.002$

$P(y_1\!=\!1, y_2\!=\!2, y_3\!=\!1) = 0.620 * 0.227 + 0.380 * 0.053 = \underline{0.161}$

$P(y_1\!=\!1, y_2\!=\!2, y_3\!=\!1 \mid X\!=\!1) = 0.960 * 0.258 * 0.917 = 0.227$

$P(y_1\!=\!1, y_2\!=\!2, y_3\!=\!1 \mid X\!=\!2) = 0.229 * 0.956 * 0.240 = 0.053$

The Excel file antirel.xls shows the computations for all 8 patterns

The general case: a *C*-class LC model for *J* indicators

1. Mixture of *C* classes:

$$P(y_1,\ldots,y_J)=\sum_{c=1}^{C}P(X=c)P(y_1,\ldots,y_J\mid X=c)$$

2. Local independence of *J* indicators:

$$P(y_1,\ldots,y_J\mid X=c)=\prod_{j=1}^{J}P(y_j\mid X=c)$$

1. and 2. combined:

$$P(y_1,\ldots,y_J)=\sum_{c=1}^{C}P(X=c)\prod_{j=1}^{J}P(y_j\mid X=c)$$

Maximum likelihood (ML) estimation

- Finding the parameter values which maximize the likelihood, the probability of observing the data you have
- Likelihood = product across observations of the probability of having the observed response pattern
- Log-likelihood = sum across observations of the logarithm (ln) of the probability of having the observed response pattern

$$LL = \sum_{i=1}^{N} \ln P(\mathbf{y}_i) = \sum_{all\ pattern\ p} n_p \ln P\left(\mathbf{y}_p\right)$$

ML solution for antirel.dat

- LL = -2795.38 (see Latent GOLD and Excel sheet)
- You can verify that other values for the model probabilities give a lower (more negative) LL value

- How do we find the ML solution? We need an algorithm for this.
- In LC analysis we use the Expectation-Maximization (EM) algorithm and the Newton-Raphson algorithm