Getting Started with Latent Class Analysis (LCA)

Yi-Fan Chen

Design and Analysis Core
Center for Clinical and Translational Science
University of Illinois at Chicago

January 13, 2015

What is Latent Class Analysis?

LCA

- In general, to find subgroup of cases from multivariate categorical data
- In statistics,
  - to stratify cases, aggregated as the cross-classification table of observed variables, by an unobserved variable with unordered categories
  - to explore subgroups which follow different parameters of a postulated statistical model
- In applications, for discovering case subtypes, reducing data dimensions, and predicting future cases in marketing, medicine, and behavior science, etc.

How is LCA different from others ?

Comparison with other similar methods

- Cases vs. Variables (Factor analysis)
- Model-based vs. Data-driven method (K-means)
- Categorical vs. Continuous predictors (Discrete latent class)
- Without vs. With an outcome (Tree analysis)

What is a latent class?

Latent Class: a underlying class which satisfies a **conditional independence assumption**

- Within each latent class, variables are independent
- If the effect of latent class membership is removed, all that remains is randomness
- The effect of latent class membership eliminates all confounding between observed variables

How does LCA work?

- Procedure: it estimates parameters of a simple parametric model using observed data.
- Parameters
  1. The prevalence of each latent class
  2. Conditional response probabilities for each combination of latent class and response level

How does LCA work? (cont.)

- Model: the probability of obtaining response pattern is a weighted average of the C class-specific probabilities

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{r=1}^{C} P(R = r) P(\mathbf{Y} = \mathbf{y} | R = r)$$

$$\text{Assumption: } P(\mathbf{Y} = \mathbf{y} | R = r) = \prod_{p=1}^{P} P(Y_p = y_p | R = r)$$

$R = 1, \ldots, C$: latent variable with C classes
$Y_p = 1, \ldots, D_p$: one of P predictors/manifest variables with D_p levels

How does LCA work? (cont.)

- Estimation: maximum likelihood estimator (MLE)

$$ln(L) = \sum_{i=1}^{I} n_i \times ln\{P(\mathbf{Y} = \mathbf{y}_i)\}$$

$I = \prod_{p=1}^{P} D_p$: the number of possible answer patterns

n_i: the observed frequency in i^{th} pattern

- Algorithms
  - Expectation-Maximization (EM)
  - Newton-Raphson (NR)
- Standard error estimates
  - The second derivatives of model parameters
  - The parametric bootstrap method

How does LCA work? (cont.)

- Estimation problems
  - **Local maxima:** a local solution is obtained
    - Try different parameter initial values
  - **Identifiability problem:** more than one solutions exist when having more unknowns than equations
    - Check the rank of the matrix of the second derivatives of model parameters
    - Try different initial values to see if different solutions exist
    - Simplify the model
    - Impose constraints
  - **Boundary solutions:** probability 0 causes numerical problems
    - Impose constraints
    - Include other kinds of prior information on the parameters

How does LCA work? (cont.)

Cases classification

1. A fuzzy/probabilistical classification using the Bayes' theorem to calculate a posterior probability of a case's membership in each class

$$P(R = r|\mathbf{Y} = \mathbf{y}) = \frac{P(R=r)P(\mathbf{Y}=\mathbf{y}|R=r)}{P(\mathbf{Y}=\mathbf{y})}$$

2. Either a modal assignment to a latent class with the highest posterior probability

Model evaluation

Goodness of fit

- Comparing the observed cross classification frequencies to the expected frequencies predicted by using a likelihood ratio Chi-squared statistic (G-squared)

$$G^2 = \sum_{i=1}^{I} 2 \times f(i) \times ln\{f(i)/e(i)\}$$
$$= \sum_{i=1}^{I} 2 \times n_i \times ln\left\{\frac{n_i}{N \times P(\mathbf{Y}=\mathbf{y}_i)}\right\} \sim \chi^2_{df}$$

$$df = \prod_{p=1}^{P} D_p - C \times \{1 + \sum_{p=1}^{P}(D_p - 1)\}$$

N: total number of cases
$f(i)$: the observed frequency of response patterns
$e(i)$: the expected frequency of response patterns

- Sparse table problem: use parametric bootstrapping or parsimony indices

Model evaluation (cont.)

Goodness of fit (cont.)

- Using the difference of G-squared statistics for comparing two nested models
- Information statistics for comparing non-nested models: AIC, BIC

Classification error

$$\text{Proportion of classification error} = E = \sum_{i=1}^{I} \frac{n_i}{N}\{1 - max\{P(R = r|\mathbf{Y} = \mathbf{y}_i)\}\}$$

$$\Rightarrow \text{Reduction of errors measure} = \lambda = 1 - \frac{E}{max\{P(R=r)\}}$$

Determination of the number of latent classes

Methods

- Try different plausible number of latent classes and assess the fit of each other to the data
- Use information indices, such as AIC, BIC with a scree-type test which shows a leveling-off point in a plot of model fit vs. number of latent classes
- Conduct computation-intensive approaches, such as bootstrapping and Monte Carlo

Extensions of LCA

- LC model as a log-linear model by Haberman (1979)

$$ln\{P(R=r,\mathbf{Y}=\mathbf{y})\} = \beta + \beta_r^R + \sum_{p=1}^{P} \beta_{y_p}^{Y_p} + \sum_{p=1}^{P} \beta_{r,y_p}^{R,Y_p}$$

$$P(Y_p = y_p | R = r) = \frac{exp(\beta_{y_p}^{Y_p} + \beta_{r,y_p}^{R,Y_p})}{\sum_{j=1}^{D_p} exp(\beta_j^{Y_p} + \beta_{r,j}^{R,Y_p})}$$

- Inclusion of covariates, Z, that describe the latent variable

$$P(R = r | \mathbf{Z} = \mathbf{z}) = \frac{exp(\alpha_r^R + \sum_{k=1}^{K} \alpha_r^{R,Z_k} \cdot z_k)}{\sum_{l=1}^{C} exp(\alpha_l^R + \sum_{k=1}^{K} \alpha_l^{R,Z_k} \cdot z_k)}$$

- Inclusion of ordering of categories: impose ordinal constraints via association model structures on β_{r,y_p}^{R,Y_p}, such as $\beta_{r,y_p}^{R,Y_p} = \beta_{r,y_p}^{R,Y_p} \cdot y_p$

Extensions of LCA (cont.)

- When the local independence fails,
  - Increase the number of latent classes
  - Include direct effects between certain variables to relax the assumption
- LC model with continuous variables: latent profile model, mixture-model clustering, model-based clustering, latent discriminant analysis, LC clustering

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{r=1}^{C} P(R = r) f(\mathbf{Y} = \mathbf{y} | R = r)$$

Software

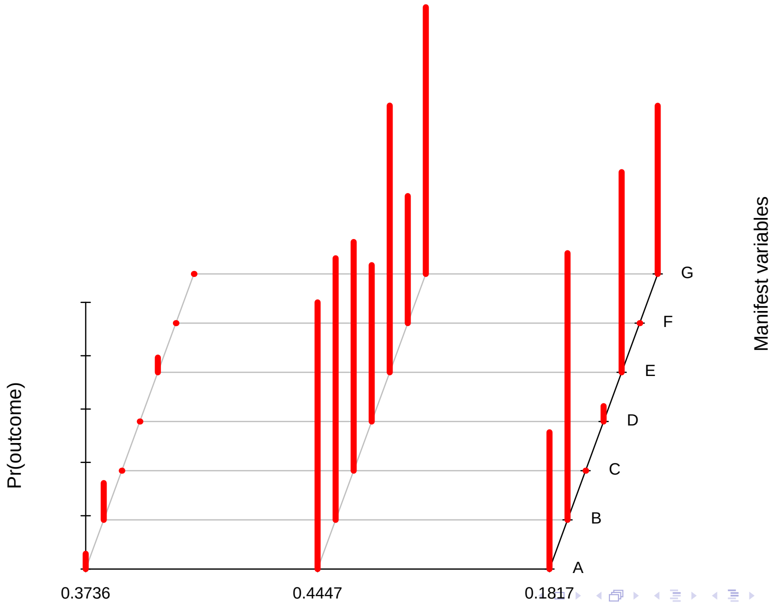
- CDAS/MLLSA
- CLIMMIX
- DILTRAN
- DISTAN
- GLIMMIX 2.0
- Latent GOLD
- LCABIN
- LCAG
- LEM
- LLCA
- Miracle 32
- MLLSA
- Mplus
- Multimix
- NEWTON and LAT
- PANMARK
- PRASCH
- PROC LCA/PROC LTA
- R: LCA, LCMM, poLCA, MCLUST
- WinLTA
- WINMIRA

A Simple Example by using poLCA in R

- poLCA: by Linzer and Lewis, 2014
  - Estimation: EM algorithm and Newton-Raphson
  - Standard error estimation: emprical observed information matrix
  - Data format: the manifest variables must be coded as integer values starting at 1 for the first category
- carcinoma data: from Agresti, 2002
  - Data: 7 binary variables whaich are the ratings by 7 pathologists of 118 slides on the presence or absence of carcinoma
  - Goal: to investigate the interobserver agreement by examing how subjects might be divided into groups depending upon the consistency of their diagnoses

```
> #-- load package
> #install.packages('poLCA')
> library(poLCA)

Loading required package:  scatterplot3d
Loading required package:  MASS

> #-- load built-in data
> data("carcinoma")
> tail(head(carcinoma,66))

   A B C D E F G
61 2 2 1 1 2 1 2
62 2 2 1 1 2 1 2
63 2 2 1 1 2 1 2
64 2 2 1 1 2 1 2
65 2 2 1 1 2 1 2
66 2 2 1 1 2 1 2

> dim(carcinoma)

[1] 118   7
```

```
> #-- LCA
> f <- cbind(A, B, C, D, E, F, G) ~ 1
> lc3 <- poLCA(formula=f, data=carcinoma, nclass=3, graphs=TRUE,
+              na.rm=TRUE, nrep=10, maxiter=1000, tol=1e-10,
+              probs.start=NULL, verbose=TRUE, calc.se=TRUE)

Model 1: llik = -293.705 ... best llik = -293.705
Model 2: llik = -293.705 ... best llik = -293.705
Model 3: llik = -293.705 ... best llik = -293.705
Model 4: llik = -293.705 ... best llik = -293.705
Model 5: llik = -293.705 ... best llik = -293.705
Model 6: llik = -293.705 ... best llik = -293.705
Model 7: llik = -293.705 ... best llik = -293.705
Model 8: llik = -293.705 ... best llik = -293.705
Model 9: llik = -293.705 ... best llik = -293.705
Model 10: llik = -293.705 ... best llik = -293.705
```



Pr(outcome)
Manifest variables
A
B
C
D
E
F
G
0.3736
0.4447
0.1817

```
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)

$A
           Pr(1)  Pr(2)
class 1:  0.9427 0.0573
class 2:  0.0000 1.0000
class 3:  0.4872 0.5128

$B
           Pr(1)  Pr(2)
class 1:  0.8621 0.1379
class 2:  0.0191 0.9809
class 3:  0.0000 1.0000

$C
           Pr(1)  Pr(2)
class 1:  1.0000 0.0000
class 2:  0.1425 0.8575
class 3:  1.0000 0.0000

$D
           Pr(1)  Pr(2)
class 1:  1.0000 0.0000
class 2:  0.4138 0.5862
class 3:  0.9424 0.0576
```

```
$E
           Pr(1)  Pr(2)
class 1:  0.9449 0.0551
class 2:  0.0000 1.0000
class 3:  0.2494 0.7506

$F
           Pr(1)  Pr(2)
class 1:  1.0000 0.0000
class 2:  0.5236 0.4764
class 3:  1.0000 0.0000

$G
           Pr(1)  Pr(2)
class 1:  1.0000 0.0000
class 2:  0.0000 1.0000
class 3:  0.3693 0.6307

Estimated class population shares
 0.3736 0.4447 0.1817

Predicted class memberships (by modal posterior prob.)
 0.3729 0.4322 0.1949

=========================================================
```

```
Fit for 3 latent classes:
=========================================================
number of observations: 118
number of estimated parameters: 23
residual degrees of freedom: 95
maximum log-likelihood: -293.705

AIC(3): 633.41
BIC(3): 697.1357
G^2(3): 15.26171 (Likelihood ratio/deviance statistic)
X^2(3): 20.50335 (Chi-square goodness of fit)
```

```
> #-- Goodness of fit
> capture.output(lc2 <- poLCA(f, carcinoma, nclass = 2), file='NUL')
> capture.output(lc4 <- poLCA(f, carcinoma, nclass = 4), file='NUL')
> lc2$bic

[1] 706.0739

> lc3$bic

[1] 697.1357

> lc4$bic

[1] 726.4629
```

```
> #-- Classification
> round(tail(head(lc3$posterior,66)),2)

      [,1] [,2] [,3]
[61,]    0 0.24 0.76
[62,]    0 0.24 0.76
[63,]    0 0.24 0.76
[64,]    0 0.24 0.76
[65,]    0 0.24 0.76
[66,]    0 0.24 0.76

> tail(head(lc3$predclass,66))

[1] 3 3 3 3 3 3
```

Reference

- Dr. John Uebersax at California Polytechnic State University http://www.john-uebersax.com/stat/faq.htm
- Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. The sage encyclopedia of social sciences research methods, 549-553.
- Magidson, J., & Vermunt, J. K. (2006). Latent Class Models.
- Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. Journal of Statistical Software, 42(10), 1-29.

Thank you!

yfchen2@uic.edu