Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

An R package for fitting fuzzy clustering analysis using generalized structured component analysis: gscaLCA

Ji Hoon Ryoo
University of Southern California
Children's Hospital Los Angeles

Seohee Park
The University of Iowa

Seongeun Kim
The University of North Carolina at Greensboro

Heungsun Hwang
McGill University

Abstract

gescLCA is a package implemented in the R statistical computing environment, which can be used for conducting latent class analysis. Applying both fuzzy clustering algorithm and generalized structured component analysis in gscaLCA, reseachers are allowed to compute prevalence and item response probabilities as posterior probabilities. As a hybrid model between cluster analysis for classifications and mixture-modeling approach, gscaLCA encompasses many advantages such as efficiency in computing parameter estimates with bootstrap method, robustness to the deviations from multivariate normal distribution, etc. Main function, gscaLCA, works for both binary and ordered categorical variables. Visualization of profiles of latent classes based on the posterior probabilities is also available in gscaLCA.

Keywords: Fuzzy clustering, Generalized structured component analysis, Latent class analysis, Optimal scaling, gscaLCA.

1. Introduction

1.1. Motivation

Latent class analysis (LCA; Lazarsfeld 1950; McCutcheon 1987) has been widely used to identify homogeneous subpopulation from observed categorical variables under the assumption of heterogeneous population in research areas including social, behavioral, and health sciences. Theoretically, its usefulness seems to be well fitted to predictive modeling in a data mining. However, its popularity of LCA is somewhat disconnected from predictive modeling due to its common estimation method, maximum likelihood estimation based on the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) under the assumption of a multivariate normality for variance-covariance matrix. More precisely, the multivariate normality assumption prevents researcher from utilizing concept of big data in latent class analysis, which often produces estimation issues such as non-positive definite in computing a Hessian matrix (Gill and King 2004).

As shown in the comparison between mixture-modeling approach and K-means cluster analysis (Steinley and Brusco 2011), the mixture-modeling approach does not always perform better than K-means cluster analysis. Under a certain condition, Steinley and Brusco (2011) showed an equivalence in terms of statistical modeling. Therefore, conceptually equivalent two clustering methods cannot be said to lead another, which is consistent when latent class, K-means, and K-median methods in Brusco, Shireman, and Steinley (2017). Although Lubke and Muthen (2005) note that "Model-based methods have the advantage that more rigorous methods can be applied for the comparison of alternative models", it would not be applicable when we consider big data and/or data mining due to the model complexity and strict assumption of a multivariate normality. On the other hand, K-means cluster analysis would rather be better fitted due to the simplest estimation algorithm using least square estimation. (Steinley and Brusco 2011) also noted that "it is important to realized that increased complexity and flexibility do not necessarily imply that a better solution will be found if the goal of the analysis is to uncover the unknown cluster membership".

In spite of the advantage, K-means cluster analysis is not a perfect alternative because it often suffers from a poor local optimum and it is also a limitation that its algorithm is based on the convex clustering. The formal is not only an issue for K-means cluster analysis, but mixure-modeling clustering also suffers. However, the latter would be disadvantage compared with mixture-modeling clustering. Such disadvantage can be minimized by applying fuzzy clustering analysis (Authors, 2019; Hwang, Desarbo, and Takane 2007). Furthermore, authors (2019) show how fuzzy clustering analysis can be mingled with generalized structured component analysis (GSCA; Hwang and Takane 2004) within the LCA framework.

1.2. Existing methods and tools

The method utilizing both the fuzzy clustering analysis and GSCA for LCA is new and thus, there is no comparable R package. Although gscaLCA applies fuzzy clustering algorithm, it is designated to fit latent class analysis. In this section, we introduce three well-known packages for LCA as mixture-modeling approach: Mplus, poLCA in R, and SAS procedure LCA as competitors. There are many other software packages available but exhaustive search of packages are out of our scope. It is easy to find papers comparing software packages in LCA

and cluster analysis, for example, Haughton, Legrand, and Woolford (2009) for Latent Gold, poLCA, and MCLUST and Flynt and Dean (2016) for stats, mclust, poLCA, and clustMD.

Mplus

Mplus is the most common software package in structural equation modeling and provides a variety of tools in terms of modelings including LCA. Within a mixture modeling framework, both latent class analysis and latent profile analysis for categorial and continuous observed variables, respectively, are available using the option of TYPE=MIXTURE in Analysis part of Mplus, and also provides various estimation methods in maximum likelihood (Muthen and Muthen 2017). On the other hand, the versatility in modeling does not provide the best stability and robustness of fitting model among statistical software packages from the deviation from normality assumption or any other little violation of assumptions. Thus, it is required fors researcher to investigate which options are best fit to their studies. Here, listed are a couple of LCA examples using Mplus: Van Horn, Jaki, Masyn, Ramey, Smith, and Antaramian (2009) including a syntax, and O'Neill, McLarnon, Xiu, and Law (2016).

poLCA

Among several R packages for latent class analysis, poLCA is one of the most common R package for LCA. By using expectation-maximization and Newton-Raphson algorithm, poLCA finds maximum likelihood estimates of the LCA model parameters (Linzer and Lewis 2011). Latent class regression predicting latent class memberships by one or more covariates. Here are two examples of using poLCA: Schreiber (2017) with a syntax, Miranda, dos Santos Amorim, Bastos, Souza, de Faria, do Carmo Castro Franceschini, and Priore (2019) in public health, van Rijnsoever and Castaldi (2011) in information science, and Xia, Evans, Spilsbury, Ciesielski, Arrowsmith, and Wright (2010) in tourism management.

Proc LCA

Proc LCA was developed for SAS for Windows, along with Proc LTA for latent transition analysis that is a longitudinal version of LCA. Lanza, Dziak, Huang, Wagner, and Collins (2015) listed key features including multi-groups LCA, option to impose measurement invariance across groups, LCA with covariates, binary and multinomial logistic regression options for predicting latent class membership, and the ability to take into account sampling weights and clusters. Collins (2010) describe the whole process of fitting LCA including the key features, although those key features are also available in other tools. Listed are a couple of examples using Proc LCA: Reynolds and Fisher (2019) and Ryoo, Wang, Swearer, and Park (2017).

1.3. Our contribution

As noted, there is no dominating method to run latent class analysis between mixture-modeling approach and cluster analysis. Rather, the choice of statistical model would be related to researcher's discretion (Hwang, Takane, and Jung 2017; Widaman 2007). Our goal and contribution is to provide an analytic tool to researchers who want to run LCA using a traditional, "heuristic" cluster analysis with fuzzy clustering algorithm as a hybrid method. In addition, the utilization of GSCA in gscaLCA allows researchers to analyze data within the full range of structural equation modeling methods. Sections for describing gscaLCA consist

of three sections: Fuzzy clustering GSCA in Section 2, description of main function, gscaLCA, in Section 3, and demonstration of fitting gscaLCA with two empirical datasets in Section 4.

2. Fuzzy clustering GSCA

Fuzziness would be well understood as a soft clustering that each object belongs to every cluster at a certain degree while K-means is a hard clustering that every object belongs only one cluster. As one of centroid-based clustering approaches, Fuzzy clustering overcomes one disadvantage of K-means algorithm that K-means does not work well for non-convex data. In addition to the clustering point of view, the harmonization with GSCA in estimating process provides tools such as statistical modeling and model evaluation that are fruitful compared with K-means. The feature of model evaluation will be discussed at the later this section, which provides cluster validity measures in fuzzy clustering.

2.1. Fuzzy clustering algorithm focusing on fuzzy c means

Based on the description of fuzzy clustering (Bezdek 1981) and terminologies used by Mahata, Sarkar, Das, and Das (2017), we describe fuzzy c means (FCM) algorithm as follows: FCM minimizes the following distance measured by the sum of squares:

$$J_m = \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ki}^m \|x_i - c_k\|^2 \tag{1}$$

where $m \in (1, \infty]$ is a classification index, u_{ki} indicates the membership probability of ith object in class k, and c_k indicates the centroid for class $k \in \{1, \cdots, K\}$. The m is also known as a fuzzifier and indicates the probability of being a class, i.e., $m = 1$ indicates the membership probability is either 0 or 1 such as K-means, which is excluded in this FCM algorithm. On the other than hand, $m = \infty$ indicates equal probability in being any of classes. Both u_{ki} and c_k are defined by

$$u_{ki} = \frac{1}{\sum_{l=1}^{K} (\frac{\|x_i - c_k\|}{\|x_i - c_l\|})^{\frac{2}{m-1}}}$$
$$c_k = \frac{\sum_{i=1}^{N} u_{ki}^m \cdot x_i}{\sum_{i=1}^{N} u_{ki}^m}$$

With a termination criterion such that $\max_{ki}\{|u_{ki}^{(L+1)} - u_{ki}^{(L)}|\} < \epsilon$ for a small value ϵ at $L+1$ repeats, the FCM algorithm is done as follows:

1. (Step 1) Initialize $U^{(0)}$ for $U = [u_{ki}]$,
2. (Step 2) compute c_k and u_{ki} by minimizing J_m, which also update U, and
3. (Step 3) If $\|U^{(L+1)} - U^{(L)}\| \leq \epsilon$ then "STOP". Otherwise, the algorithm repeats from Step 2.

2.2. Fuzzy clustering GSCA

Lin, Chen, and Wu (2004) applied fuzzy clustering for latent class model and authors (2019) applied fuzzy clustering to GSCA for LCA. Fuzzy clustering GSCA for LCA follows two steps:

(1) based on the responses, we classify the clustering where fuzzy clustering works and (2) with u_{ki}^m estimated in Step (1), we estimate the parameters of item response probabilities in the latent class analysis. Hwang and Takane (2014) described how to utilize fuzzy clustering in GSCA for continuous data, which was employed to fit fuzzy clustering GSCA for LCA with the optimal scaling introduced by Young (1981). Briefly, to estimate the memberships for the heterogeneous subgroups, we can estimate the membership parameters, u_{ki}, by minimizing the residual sums of their squares weighted by u_{ki}:

$$\phi = \sum_{k=1}^{K} \sum_{i=1}^{N} u_{ki}^m \cdot SS(V_k' z_i - A_k' W_k' z_i), \tag{2}$$

with respect to u_{ki}, W_k (the matrix of weights in GSCA in class, k), and A_k (the matrix of both measurement and structural in GSCA in class, k), subject to the probabilistic condition, $\sum_{k=1}^{K} u_{ki} = 1$. The algorithm called as "alternating least squares" can be found at in detail. The fuzzifier, m, is often set up at 2 in practice (Bezdek 1981), which is also used as a default in gscaLCA. The item response probabilities are then estimated based on their membership within each class.

2.3. Model evaluation in Fuzzy clustering GSCA

In addition to R squared type of model evaluation tools, FIT and AFIT, from GSCA (Hwang and Takane 2014), gscaLCA uses the fuzziness performance index (FPI) and the normalized classification entropy (NCE) recommended by Roubens (1982), defined as follows:

$$FPI = 1 - \frac{(K \times \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ki}^2 - 1}{K - 1} \tag{3}$$

$$NCE = \frac{-\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ki} \log u_{ki}}{\log K}. \tag{4}$$

With the same criteria as "Smaller is better" between 0 and 1, both FPI and NCE help researchers decide the number of clusters in the process of fitting LCA (Hwang and Takane 2014).

3. gscaLCA

The gscaLCA package enables to run LCA based on fuzzy clsutering GSCA by estimating the parameters of latent class prevalence and item response probability in LCA with a single line comment. Main function, gscaLCA, will be described below with a visualization of results as a key feature.

3.1. Data input and sample datasets

Data are the main input to the gscaLCA function, and it should be a format of data frame containing manifest variables. As a package to fit LCA, gscaLCA requires that the manifest variables are categorical variables. It, however, does not requires whether the variables are integer or character. When the variable is continuous, the function still run by recognizing the type of variable as categorical. Thus, a caution is necessitated. When data set has missing

value, it requires to be coded as NA. In this current version, the algorithm of gscaLCA used the only option of listwise delection for the missing data.

The gscaLCA package also provides two pre-installed sample datasets that would be useful for exploring different types variables: categorical variables (binary and more than two categories).

AddHealth data. This AddHealth data consist of 5,144 of the participants with a randomly generated ID variable and five item variables such as Smoking, Alcohol, Other Types of Illegal Drug, Marijuana, and Cocaine. The responses of the five are dichotomous as either "Yes" or "No" and are treated the other missing codes as systematic missing. These data can be obtained from the National Longitudinal Study of Adolescent to Adult Health (Add Health; Harris, Halpern, Whitsel, Hussey, Tabor, Entzel, and Udry 2009) where the study have mainly focused on the investigation of how health factors in childhood affect adult outcomes. In terms of data collection, there have been four additional waves since 1994. In the gscaLCA package, the data of specific section of substance use at the wave IV is pre-installed.

TALIS data. For the ordered categorical data in gscaLCA package, we used the data of 2,560 US teachers to question about teachers' learning environment as well as working conditions. The data are a part of survey which was conducted by the Organization for Economic Cooperation and Development (OECD) and named the Teaching and Learning International Survey (TALIS) 2018 (OECD 2019). The thirty-four countries participated: 24 of these were from OECD, and 10 from partner jurisdictions. In the gscaLCA pacakge, we utilize publicly available TALIS 2018 U.S. Data focusing on five items: two items are on motivation, two items are on pedagogy, and the last item is on satisfaction. The five items were coded as the ordinal responses from 1 (least) to 3 (most).

3.2. gscaLCA commend line and options

The gscaLCA commend requires six options including the dataset. The datasets were explained in the previous session, and the remaining options are specified below:

varnames: A vector of character elements. The names of variables (column names of data set) to be employed for run gscaLCA.

ID.var: A character element. The name of ID variable. If an ID variable is not specified, gscaLCA will find an ID variable in a given data. The ID of observation are automatically created when data set does not have any ID variable. The default is NULL.

num.cluster: A integer element. The number of cluster to be analyzed. When num.cluster is smaller than 2, gscaLCA terminates with an error message. The default is 2.

num.factor: Either EACH or ALLin1. EACH indicates that observed variables are not correlated to each other. ALLin1 indicates that observed variables are correlated and thus, they have shared variance. However, it is not strong enough to run a factor analysis or an item response theory model. The default is EACH.

Boot.num: The number of bootstrap. The standard deviations of parameters are obtained from the bootstrap option while the alternating least squares algorithm runs in GSCA. The default is 20.

To estimate LCA based on fuzzy clustering GSCA, the default gscaLCA commend is:

```
R> gscaLCA(dat, varnames, ID.var = NULL, num.cluster = 2,
+      num.factor = "EACH", Boot.num = 20)
```

3.3. gscaLCA output

The gscaLCA function returns an object involving the following elements:

N: The number of observations used after listwise delection when missing values exist.

C: The number of cluster.

Boot.num.im: The number of bootstrap implemented. Not all iterations of bootstrap is appropriate to estimate the standard error.

model.fit: The model fit indices. FIT, AFIT, FPI and NCE are provided with the standard error and 95% credible interval lower and upper bounds.

LCprevalence: The latent class prevalence. The percent of class, the number of observation for each class, standard error, and 95% credible interval lower and upper bounds are provided.

RespRrob: The item response probabilities for each variable are reported as elements of a list. Each element consists of a table containing the probabilities with respect to the possible categories of each variable. The standard error and 95% credible interval lower and upper bounds are also reported.

membership: A data frame of the posterior probabilities for each examinee with the predicted class membership.

plot: Graphs for each categories. For example, with two categories, each graph is stored as p1 and p2 in the list of plot. When the number of options across variables are different, the graphs are not provided.

4. Example

To demonstrate the usage of the **gscaLCA** package, we display two examples: the one is AddHealth data with binary variables and the other is TALIS data with ordered categorical variable. It should be noted that the results do not reflect study findings but are for demonstration purposes.

4.1. An Example with binary data

The AddHealth data consist of six variables. The first variable is observations' ID (named as AID), and other 5 variables are usages of Smoking, Alcohol, Drug, Marijuana, and Cocaine. These five variables are dichotomous, and coded as "Yes" or "No". The current paper replicates the process of the study from the authors' previous study (2019). However, the results are not identical because one variable from the original data set is different and the structure model of GSCA is specified differently. The package gscaLCA of the current version enables to assign the structure model based on whether observed variables have each latent variable (num.factor = "EACH") or a common latent variable (num.factor = "ALLin1").

When the model indices were evaluated across the number of clsuters, the three-class model with the num.factor = "EACH" option was the most reasonable for the AddHealth data. Thus, the current paper demonstrates when the number of clusters (classes) is three and each variable has their own latent variable. The following commands enable to run gscaLCA. The gscaLCA command shows the degree of the process completion as percentage while R is running. Once the estimation completes, the sample size for the analysis, model fit indices, estimated latent class prevalence, and item response probabilities are printed out.

```
R> data("AddHealth")
R> head(AddHealth)

       AID Smoking Alcohol Drug Marijuana Cocaine
1 57101310     Yes     Yes   No        No      No
2 57103869     Yes      No   No       Yes      No
3 57109625     Yes     Yes  Yes       Yes     Yes
4 57111071     Yes     Yes  Yes       Yes     Yes
5 57113943      No     Yes   No       Yes      No
6 57117542     Yes     Yes  Yes       Yes     Yes

R> AH.3 = gscaLCA(AddHealth, varnames = names(AddHealth)[2:6],
+      num.factor = "EACH", num.cluster = 3, Boot.num = 100)

============================================================
LCA by using Fuzzing Clusterwise GSCA
============================================================
Fit for 3 latent classes:
 number of used observations: 5066
 number of deleted observation: 48
 number of bootstrap for SE: 46 / 100

MODEL FIT -----------------------------------------------
 FIT      :  0.9993
 AFIT     :  0.9993
 FPI      :  0.4605
 NCE      :  0.5000

Estimated Latent Class Prevalnces (%) -------------------
```

```
 32.02% 47.73% 20.25%

Conditional item response probability --------------------
 $Smoking
           Class Category Estimate
1 Latent Class 1      Yes   0.7472
2 Latent Class 1       No   0.2528
3 Latent Class 2      Yes   0.4566
4 Latent Class 2       No   0.5434
5 Latent Class 3      Yes   0.9591
6 Latent Class 3       No   0.0409

$Alcohol
           Class Category Estimate
1 Latent Class 1      Yes   0.9544
2 Latent Class 1       No   0.0456
3 Latent Class 2      Yes   0.6187
4 Latent Class 2       No   0.3813
5 Latent Class 3      Yes   0.9990
6 Latent Class 3       No   0.0010

$Drug
           Class Category Estimate
1 Latent Class 1      Yes   0.6301
2 Latent Class 1       No   0.3699
3 Latent Class 2      Yes   0.0132
4 Latent Class 2       No   0.9868
5 Latent Class 3      Yes   0.0409
6 Latent Class 3       No   0.9591

$Marijuana
           Class Category Estimate
1 Latent Class 1      Yes   0.9673
2 Latent Class 1       No   0.0327
3 Latent Class 2      Yes   0.0749
4 Latent Class 2       No   0.9251
5 Latent Class 3      Yes   0.9971
6 Latent Class 3       No   0.0029

$Cocaine
           Class Category Estimate
1 Latent Class 1      Yes   0.5530
2 Latent Class 1       No   0.4470
3 Latent Class 2      Yes   0.0132
4 Latent Class 2       No   0.9868
5 Latent Class 3      Yes   0.0409
6 Latent Class 3       No   0.9591
```

The results report that 5,065 observations were used for the anlaysis after listwise deletion. They also show that the model fit of AddHealth data is acceptable. FIT and AFITT were 0.9993 and 0.9993, and they are close to 1. The indice to evaluate the classification were relatively low (FPI = 0.4605 and NCE = 0.5000). The estimated latent class prevalences is 32.02%, 47.73%, and 20.25%. The conditional item response probabiliies are also presented for each category per variable. When the standard error and 95% credible interval of the model fit, the prevalence and conditional reseponse probabilites are required, we can print out the objects through the following commands.

```
R> AH.3$model.fit
```

```
      Estimate           SE 95CI.lower 95CI.upper
FIT  0.9992789 0.0001210061  0.9992857  0.9997049
AFIT 0.9992763 0.0001214375  0.9992832  0.9997038
FPI  0.4604592 0.0385881067  0.3597711  0.4603451
NCE  0.5000078 0.0382382441  0.3958695  0.4995067
```

```
R> AH.3$LCprevalence
```

```
                Percent Count        SE 95.CI.lower 95.CI.upper
Latent Class 1 32.01737  1622  8.339451    24.78583    47.46743
Latent Class 2 47.72996  2418 12.738361    22.04895    54.54402
Latent Class 3 20.25266  1026  4.940015    19.51737    30.99092
```

```
R> AH.3$RespProb
```

```
[1] "Smoking"
```

```
           Class Category   Estimate         SE 95.CI.lower
1 Latent Class 1      Yes 0.74722565 0.05737902 0.738930511
2 Latent Class 1       No 0.25277435 0.05737902 0.052186717
3 Latent Class 2      Yes 0.45657568 0.08452562 0.236603390
4 Latent Class 2       No 0.54342432 0.08452562 0.531225338
5 Latent Class 3      Yes 0.95906433 0.17307900 0.633039765
6 Latent Class 3       No 0.04093567 0.17307900 0.002971357
  95.CI.upper
1   0.9478133
2   0.2610695
3   0.4687747
4   0.7633966
5   0.9970286
6   0.3669602
```

(The results for Alcohol, Drug, Marijuana, and Cocaine, were omitted here.)

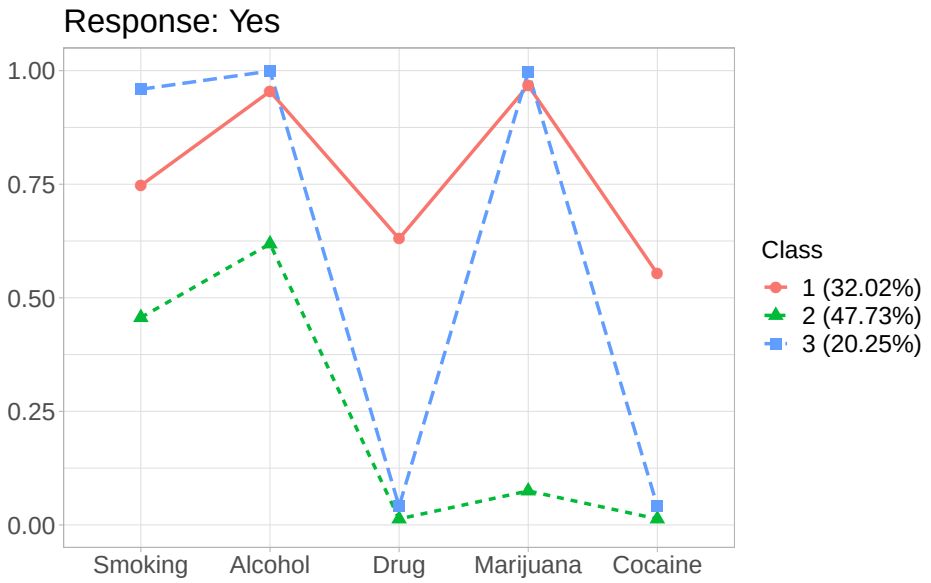

Figure 1: Profiles of three latent classes from fuzzy clustering GSCA in AddHealth data

These response probabilities are used to define latent classes. In order to grasp the patterns of the probabilities, a visual representation of profiles based on the probabilities would be more helpful than numeric quantities in the output above. Once the gscaLCA function is implemented, the graph is automatically generated. When a category of response is binary, the graph shows the probabilities' patterns of one category. In AddHealth data, the responses are binary (Yes or No), thus the graph is involved when the response is "Yes".

NA

```
R> AH.3$plot[[1]]
R> AH.3$plot[[2]]
```

Lastly, the membership probabilities of observations can be obtained from the membership of the saved objects, AH.3.

```
R> AH.3$membership
```

```
           Clus.1     Clus.2     Clus.3          label
57101310 0.3168141 0.47797925 0.20520666 Latent Class 2
57103869 0.1601041 0.68353399 0.15636194 Latent Class 2
57109625 0.9584893 0.02046429 0.02104637 Latent Class 1
57111071 0.9584893 0.02046429 0.02104637 Latent Class 1
57113943 0.4409957 0.42702648 0.13197784 Latent Class 1
57117542 0.9584893 0.02046429 0.02104637 Latent Class 1
```

(The first six observations are listed here.)

4.2. An example with ordinal data

The TALIS data include 2,560 responses with six variables including teacher ID. The five variables are teachers' responses to questions about motivation for becoming a teacher, pedagogy in class, and satisfaction of teachers' learning environment and working conditions in their schools. The first two variables are about motivation, the third and fourth ones are about pedagogy, and the last one is about satisfaction. Teachers' responses are originally coded as four ordered categorical data. Due to too small frequencies at the lowest levels at the five variables, we modified them into three ordered categories by merging the two lowest levels: (Not/low important, moderate important, and high important) in motivation, (not at all/to some extent, quite a bit, and a lot) in pedagogy, and (strongly disagree/disagree, agree, and strongly agree) in satisfaction. Other missing codes were treated as a missing code, (NA). The specific explaination about the categories is presented in the manual, which is accessible via a command of ?TALIS.

Similar as in the AddHealth data, we demonstrate the three-class model with a single factor through the model comparison between the number of classes. The following command was implemented in the gscaLCA.

```
R> data("TALIS")
R> head(TALIS)

  IDTEACH Mtv_1 Mtv_2 Pdgg_1 Pdgg_2 Stsf
1  300101     2     1      2      2    3
2  300102     2     2      2      2    2
3  300103     2     2      2      1    2
4  300104     2     3      2      3    3
5  300105     3     3      1      2    2
6  300106     2     1      1      2    2

R> T.3 = gscaLCA(TALIS, varnames = names(TALIS)[2:6], num.cluster = 3,
+      num.factor = "ALLin1", Boot.num = 100)

============================================================
LCA by using Fuzzing Clusterwise GSCA
============================================================
Fit for 3 latent classes:
 number of used observations: 2365
 number of deleted observation: 195
 number of bootstrap for SE: 91 / 100

MODEL FIT -----------------------------------------------
 FIT      :  0.5046
 AFIT     :  0.5033
 FPI      :  0.8623
 NCE      :  0.8742

Estimated Latent Class Prevalnces (%) -------------------
```

```
 39.20% 34.59% 26.22%

Conditional item response probability --------------------
 $Mtv_1
           Class Category Estimate
1 Latent Class 1        1   0.1607
2 Latent Class 1        2   0.4153
3 Latent Class 1        3   0.4239
4 Latent Class 2        1   0.2482
5 Latent Class 2        2   0.2910
6 Latent Class 2        3   0.4609
7 Latent Class 3        1   0.0419
8 Latent Class 3        2   0.6581
9 Latent Class 3        3   0.3000

$Mtv_2
           Class Category Estimate
1 Latent Class 1        1   0.2546
2 Latent Class 1        2   0.3161
3 Latent Class 1        3   0.4293
4 Latent Class 2        1   0.2958
5 Latent Class 2        2   0.2604
6 Latent Class 2        3   0.4438
7 Latent Class 3        1   0.1758
8 Latent Class 3        2   0.4806
9 Latent Class 3        3   0.3435

$Pdgg_1
           Class Category Estimate
1 Latent Class 1        1   0.2136
2 Latent Class 1        2   0.4132
3 Latent Class 1        3   0.3732
4 Latent Class 2        1   0.1577
5 Latent Class 2        2   0.5391
6 Latent Class 2        3   0.3032
7 Latent Class 3        1   0.3726
8 Latent Class 3        2   0.2758
9 Latent Class 3        3   0.3516

$Pdgg_2
           Class Category Estimate
1 Latent Class 1        1   0.1370
2 Latent Class 1        2   0.4078
3 Latent Class 1        3   0.4552
4 Latent Class 2        1   0.0978
5 Latent Class 2        2   0.4963
6 Latent Class 2        3   0.4059
```

```
7 Latent Class 3        1   0.2484
8 Latent Class 3        2   0.3532
9 Latent Class 3        3   0.3984

$Stsf
           Class Category Estimate
1 Latent Class 1        1   0.0583
2 Latent Class 1        2   0.4865
3 Latent Class 1        3   0.4552
4 Latent Class 2        1   0.0648
5 Latent Class 2        2   0.4902
6 Latent Class 2        3   0.4450
7 Latent Class 3        1   0.1258
8 Latent Class 3        2   0.4984
9 Latent Class 3        3   0.3758
```

The results report that 2,365 observations were used for the anlaysis, excluding 195 incomplete responses. FIT and AFITT were 0.5046 and 0.5033, respectively. With a single factor, FIT and AFIT are typically lower than the larger number of factors. The indices to evaluate the classification were relatively large (FPI = 0.8623 and NCE = 0.8742), but they are better than when the option of num.factor is used with "EACH". The estimated latent class prevalences is 39.20%, 34.59%, and 26.22%. The conditional item response probabiliies for each category per variable are also presented in a table.

Figure 2 present the conditional item response probabilities. Based on the patterns of responses in each class, we can define "Motivated teachers", "Pedagogy focused teachers", and "Balanced teachers". Similar as in the anlaysis of AddHealth data, the model fit indices, estimated latent class prevalence, and item response probabilities with 95% credible interval are accessible.

5. Conclusion

The R package gscaLCA provides a unified framework of fitting an LCA model utilizing fuzzy clustering algorithm and generalized structured component analysis. Not only dichotomized observed variables but also ordered categorical observed variables can be used in the function of gscaLCA. In addition, visual representation of results profiles are a key feature in gscaLCA that helps researchers identify characteristics of classes. It should also be noted that the capacities of GSCA (Hwang and Takane 2014) within gscaLCA will extend the application of gscaLCA in a variety of SEM modeling.

gscaLCA is still undergoing active development, which includes implementation of LCA with covariates, multiple-group LCA including testing measurement invariance, multilevel LCA, and LTA for longitudinal data. Among other things, the driving force of developing gscaLCA as a hybrid method between mixture-modeling approach and cluster analysis is its applicability to data science for big data.

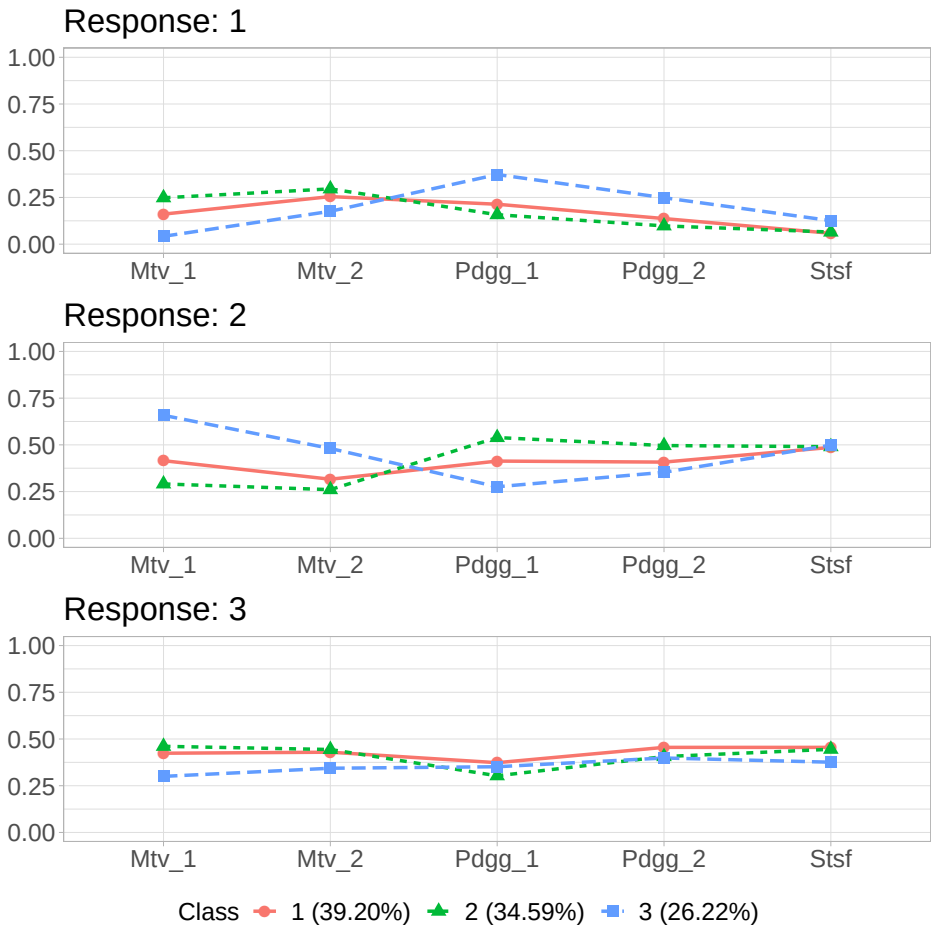

Figure 2: Profiles of three latent classes from fuzzy clustering GSCA by using TALIS data

References

Bezdek JC (1981). *Pattern recognition with fuzzy objective function algorithms*. Advanced applications in pattern recognition. Plenum Press, New York. ISBN 978-0-306-40671-3.

Brusco MJ, Shireman E, Steinley D (2017). "A comparison of latent class, K-means, and K-median methods for clustering dichotomous data." *Psychological methods*, **22**(3), 563–580. ISSN 1082-989X. doi:10.1037/met0000095.

Collins LM (2010). *Latent class and latent transition analysis: with applications in the social behavioral, and health sciences*. Wiley series in probability and statistics. Wiley, Hoboken, N.J. ISBN 978-0-470-22839-5.

Dempster A, Laird N, Rubin D (1977). "Maximum likelihood from incomplete data via the "EM" algorithm." *Journal of the Royal Statistical Society, Series B, Methodological*, **39**, 1. ISSN 0035-9246.

Flynt A, Dean N (2016). "A Survey of Popular R Packages for Cluster Analysis." *Journal of Educational and Behavioral Statistics*, **41**(2), 205–225. ISSN 1076-9986, 1935-1054.

doi:10.3102/1076998616631743. URL http://journals.sagepub.com/doi/10.3102/1076998616631743.

Gill J, King G (2004). "What to Do When Your Hessian is Not Invertible: Alternatives to Model Respecification in Nonlinear Estimation." *Sociological Methods & Research*, **33**(1), 54–87. ISSN 0049-1241, 1552-8294. doi:10.1177/0049124103262681. URL http://journals.sagepub.com/doi/10.1177/0049124103262681.

Harris KM, Halpern CT, Whitsel E, Hussey J, Tabor J, Entzel P, Udry JR (2009). "The National Longitudinal Study of Adolescent to Adult Health." *Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill*.

Haughton D, Legrand P, Woolford S (2009). "Review of Three Latent Class Cluster Analysis Packages: Latent Gold, poLCA, and MCLUST." *The American Statistician*, **63**(1), 81–91. ISSN 0003-1305, 1537-2731. doi:10.1198/tast.2009.0016. URL http://www.tandfonline.com/doi/abs/10.1198/tast.2009.0016.

Hwang H, Desarbo WS, Takane Y (2007). "Fuzzy Clusterwise Generalized Structured Component Analysis." *Psychometrika*, **72**(2), 181. ISSN 1860-0980. doi:10.1007/s11336-005-1314-x. URL https://doi.org/10.1007/s11336-005-1314-x.

Hwang H, Takane Y (2004). "Generalized structured component analysis." *Psychometrika*, **69**(1), 81–99. ISSN 1860-0980. doi:10.1007/BF02295841. URL https://doi.org/10.1007/BF02295841.

Hwang H, Takane Y (2014). *Generalized structured component analysis: a component-based approach to structural equation modeling*. CRC Press, Boca Raton. ISBN 978-1-4665-9294-0.

Hwang H, Takane Y, Jung K (2017). "Generalized Structured Component Analysis with Uniqueness Terms for Accommodating Measurement Error.(Report)(Brief article)." *Frontiers in Psychology*, **8**, 2137. ISSN 1664-1078. doi:10.3389/fpsyg.2017.02137.

Lanza ST, Dziak JJ, Huang L, Wagner A, Collins LM (2015). "PROC LCA & PROC LTA Users' Guide Version 1.3.2."

Lazarsfeld PF (1950). "The logical and mathematical foundation of latent structure analysis." *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction*, pp. 362–412.

Lin CT, Chen CB, Wu WH (2004). "Fuzzy clustering algorithm for latent class model." *Statistics and Computing*, **14**(4), 299–310. ISSN 1573-1375. doi:10.1023/B:STCO.0000039479.56180.d5. URL https://doi.org/10.1023/B:STCO.0000039479.56180.d5.

Linzer DA, Lewis JB (2011). "poLCA: An R Package for Polytomous Variable Latent Class Analysis." *Journal of Statistical Software*, **42**(10). ISSN 1548-7660. doi:10.18637/jss.v042.i10. URL http://www.jstatsoft.org/v42/i10/.

Lubke G, Muthen B (2005). "Investigating Population Heterogeneity With Factor Mixture Models." *Psychological Methods [PsycARTICLES]*, **10**(1), 21–39. ISSN 1082-989X. doi:10.1037/1082-989X.10.1.21.

Mahata K, Sarkar A, Das R, Das S (2017). "Chapter 8 - Fuzzy evaluated quantum cellular automata approach for watershed image analysis." In S Bhattacharyya, U Maulik, P Dutta (eds.), *Quantum Inspired Computational Intelligence*, pp. 259–284. Morgan Kaufmann, Boston. ISBN 978-0-12-804409-4. doi:10.1016/B978-0-12-804409-4.00008-5. URL http://www.sciencedirect.com/science/article/pii/B9780128044094000085.

McCutcheon AL (1987). *Latent class analysis*. Sage, Newbury Park, Calif. :. ISBN 978-0-8039-2752-0.

Miranda VPN, dos Santos Amorim PR, Bastos RR, Souza VGB, de Faria ER, do Carmo Castro Franceschini S, Priore SE (2019). "Evaluation of lifestyle of female adolescents through latent class analysis approach." *BMC Public Health*, **19**(1), 184. ISSN 1471-2458. doi:10.1186/s12889-019-6488-8. URL https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-6488-8.

Muthen B, Muthen L (2017). *Mplus User's Guide*. Eighth edition edition. Muthen & Muthen, Los Angeles, CA. URL https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf.

OECD (2019). *TALIS 2018 Results (Volume I)*. URL https://www.oecd-ilibrary.org/content/publication/1d0bc92a-en.

O'Neill TA, McLarnon MJW, Xiu L, Law SJ (2016). "Core self-evaluations, perceptions of group potency, and job performance: The moderating role of individualism and collectivism cultural profiles." *Journal of Occupational and Organizational Psychology*, **89**(3), 447–473. ISSN 09631798. doi:10.1111/joop.12135. URL http://doi.wiley.com/10.1111/joop.12135.

Reynolds GL, Fisher DG (2019). "A latent class analysis of alcohol and drug use immediately before or during sex among women." *The American Journal of Drug and Alcohol Abuse*, **45**(2), 179–188. ISSN 0095-2990, 1097-9891. doi:10.1080/00952990.2018.1528266. URL https://www.tandfonline.com/doi/full/10.1080/00952990.2018.1528266.

Roubens M (1982). "Fuzzy clustering algorithms and their cluster validity." *European Journal of Operational Research*, **10**(3), 294–301. ISSN 0377-2217. doi:10.1016/0377-2217(82)90228-4. URL http://www.sciencedirect.com/science/article/pii/0377221782902284.

Ryoo JH, Wang C, Swearer SM, Park S (2017). "Investigation of Transitions in Bullying/Victimization Statuses of Gifted and General Education Students." *Exceptional Children*, **83**(4), 396–411. ISSN 0014-4029, 2163-5560. doi:10.1177/0014402917698500. URL http://journals.sagepub.com/doi/10.1177/0014402917698500.

Schreiber JB (2017). "Latent Class Analysis: An example for reporting results." *Research in Social and Administrative Pharmacy*, **13**(6), 1196–1201. ISSN 15517411. doi:10.1016/j.sapharm.2016.11.011. URL https://linkinghub.elsevier.com/retrieve/pii/S1551741116305782.

Steinley D, Brusco MJ (2011). "Evaluating mixture modeling for clustering: Recommendations and cautions." *Psychological Methods*, **16**(1), 63–79. ISSN 1939-1463, 1082-989X. doi:10.1037/a0022673. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/a0022673.

Van Horn ML, Jaki T, Masyn K, Ramey SL, Smith JA, Antaramian S (2009). "Assessing differential effects: applying regression mixture models to identify variations in the influence of family resources on academic achievement." *Developmental psychology*, **45**(5), 1298–1313. ISSN 0012-1649. doi:10.1037/a0016427.

van Rijnsoever FJ, Castaldi C (2011). "Extending consumer categorization based on innovativeness: Intentions and technology clusters in consumer electronics." *Journal of the American Society for Information Science and Technology*, **62**(8), 1604–1613. ISSN 15322882. doi:10.1002/asi.21567. URL http://doi.wiley.com/10.1002/asi.21567.

Widaman KF (2007). "Common factors versus components: Principals and principles, errors and misconceptions." In *Factor analysis at 100: Historical developments and future directions*, pp. 177–203. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US. ISBN 978-0-8058-6212-6 978-0-8058-5347-6 978-1-4106-1580-0.

Xia JC, Evans FH, Spilsbury K, Ciesielski V, Arrowsmith C, Wright G (2010). "Market segments based on the dominant movement patterns of tourists." *Tourism Management*, **31**(4), 464–469. ISSN 02615177. doi:10.1016/j.tourman.2009.04.013. URL https://linkinghub.elsevier.com/retrieve/pii/S0261517709000983.

Young FW (1981). "Quantitative analysis of qualitative data." *Psychometrika*, **46**(4), 357–388. ISSN 1860-0980. doi:10.1007/BF02293796. URL https://doi.org/10.1007/BF02293796.

Affiliation:

Ji Hoon Ryoo, Ph.D.
Department of Pediatrics and Preventive Medicine
Keck School of Medicine
University of Southern California
Biostatistics Core
The Saban Research Institute
Children's Hospital Los Angeles
E-mail: jryoo@usc.edu, jryoo@chla.usc.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd