Behaviormetrika (2020) 47:291–306
https://doi.org/10.1007/s41237-019-00084-6

ORIGINAL PAPER

# Categorical latent variable modeling utilizing fuzzy clustering generalized structured component analysis as an alternative to latent class analysis

**Ji Hoon Ryoo**[1] · **Seohee Park**[2] · **Seongeun Kim**[3]

Received: 20 November 2018 / Accepted: 29 April 2019 / Published online: 8 May 2019


## Abstract

Latent class analysis is becoming popular in many areas of education, psychology, social and behavioral sciences, public health, and medicine. However, it often suffers from identification issues due to the large number of parameters involved when using maximum likelihood (ML) estimation. Increasing the sample size, reducing sparseness, and strengthening the relationship between the observed variables and the latent variables all improve the information and thus reduce the identification issues, but the identification issue still affects the validity of parameter estimates in ML estimation and the definition of identification is not sufficient to guarantee the existence of an ML solution. In this paper, generalized structured component analysis (GSCA), which is a component-based approach that utilizes optimal scaling and fuzzy clustering, is applied to avoid these identification issues and develop more stable solutions for the heterogeneity of a population based on a set of categorical responses. Testing our proposed new approach, component-based (CB) latent class analysis (LCA), on real world substance use data from Add Health produced not only the same features as those yielded by conventional ML LCA but also stable estimation without identification issues. Comparing the results obtained from ML LCA using Mplus and `poLCA` in *R*, with those from our proposed CB LCA using GSCA in *R* revealed a similar number of latent classes and posterior probabilities and only minor discrepancies in individual latent class classifications when the posterior probabilities of membership are not distinct.

**Keywords** Fuzzy clustering · Generalized structured component analysis · Latent class analysis · Optimal scaling

Communicated by Heungsun Hwang

✉ Ji Hoon Ryoo
jryoo@usc.edu; jryoo@chla.usc.edu

Extended author information available on the last page of the article

## 1 Introduction

Analyses of categorical outcomes to examine underlying constructs are widespread and are becoming almost commonplace in education, marketing, medicine, nursing, psychology, and public health. Even in the case of ordinal variables, statistical models dealing with categorical analyses such as item response theory (IRT) are now more prevalent than those with a continuous variable for both cross-sectional and longitudinal data analyses (Jeon and Rabe-Hesketh 2012; Pastor and Beretvas 2006; Wilson et al. 2012). A major reason for this popularity is the technical development of generalized linear models (Nelder and Wedderburn 1972; Skrondal and Rabe-Hesketh 2004) and of IRT (Lord 1952; Yang and Zheng 20180).

Alongside the growing popularity of analysis for categorical variables, many researchers are also analyzing categorical variables with assumed heterogeneous subgroups within a population. For example, Collins and Lanza (2010) provide a methodological framework for modeling categorical constructs, Masyn (2013) described a conceptual difference between a variable-centered approach and a person-centered approach with heterogeneous subgroups, and Ryoo et al. (2018) suggest a unified framework for fitting a latent transition analysis. However, in component-based analyses such as partial least-squares structural equation modeling (PLS–SEM; Wold 1975) and generalized structured component analysis (GSCA; Hwang and Takane 2004), component-based latent class/transition analysis is not well-structured. One barrier is the methodological challenge involved in defining heterogeneous subgroups in component-based analyses because they do not impose distributional assumptions on the variables. Consequently, the probability of being in a specific group is difficult to measure for individual parameters.

Because of this barrier, previous research in this area has focused on analyzing either continuous variables for heterogeneous subgroups of a population or categorical variables for a homogenous population in component-based analyses. For the former, Hwang et al. (2007) proposed a fuzzy clusterwise GSCA to provide a unified framework that takes into account cluster-level heterogeneity, which is applied for continuous outcomes with assumed heterogeneous subgroups within a population. For the latter, a form of GSCA based on categorical variables for use with a homogenous population was developed by Hwang and Takane (2010), which they called optimal scaling GSCA. Although fuzzy clustering GSCA for continuous variables and GSCA for categorical variables have been developed separately, latent class analysis for component-based analyses has yet not been fully investigated in GSCA. In this study, we, therefore, explored the use of fuzzy clustering for categorical outcome variables in more depth to develop a unified framework to cope with heterogeneity based on fuzzy clustering.

## 2 Literature review

This research focused on the use of component-based latent class analysis (CB LCA) to deal with observed categorical variables and heterogeneous population within the GSCA framework, which has not previously been discussed in the literature. The

first step was thus to consider an alternative approach, a maximum likelihood (ML) based latent class analysis (ML LCA).

ML LCA has been popular ever since it was first proposed by Lazarsfeld and Henry (1968), who derived a mathematical treatment of categorical survey items as an example of LCA to demonstrate the potential of LCA. Its application was initially limited due to the lack of a reliable estimation method, but this was resolved by Goodman (1974a, 1974b, 1979), who developed a simple implementable method for obtaining ML estimates and the EM algorithm is now used in most LCA software packages. Today, ML LCA is a valuable tool in many areas of research and it has also been applied in studies of heterogeneous populations to identify latent classes based on the changes in individual growth in longitudinal data (e.g., Muthén and Shedden 1999; Nagin 2005).

To fully articulate the scope of our study and the importance of the comparison between component-based approach, CB LCA, and factor-based approach, ML LCA, our literature review includes a consideration of the available methods for modeling both continuous and categorical variables in potential heterogeneous subgroups in a population across factor-based and component-based structural equation approaches. In factor-based SEM (FB-SEM), mixture models have been widely discussed in conjunction with maximum likelihood estimation (Collins and Lanza 2010; Lubke and Muthén 2005; Muthén and Asparouhov 2006). However, the distributional assumption required imposes a heavy burden if too many parameters are involved, frequently leading to serious estimation problems (Collins and Lanza 2010). This assumption also prevents researchers from applying mixture modeling approaches for large numbers of heterogeneous groups (Dziak et al. 2014; Gudicha et al. 2016).

Alternative methods that utilize component-based SEMs, PLSPM and GSCA, have also been developed for mixture modeling to resolve the issues associated with both estimation problems and the limited number of heterogeneous groups that can be handled by latent class analyses. In partial least squares path modeling (PLSPM), Hahn et al. (2002) proposed an approach to identify the cluster-level heterogeneity that is now available as the software package SmartPLS 3 (Ringle et al. 2015). However, this only utilizes a structural model that requires both the multivariate normality assumption for latent variables and the invariance assumption for the measure model. This is too restrictive and difficult to justify in many situations; for a more detailed discussion of the issues involved, see Hwang and Takane (2014) and Hair et al. (2017). Esposito Vinzi et al. (2008) suggested another approach based on a response-based procedure for detecting unit segments in PLSPM (REBUS–PLS) that addresses disadvantages such as the distributional assumption and the structural model limitation in Hahn et al.'s approach. However, REBUS–PLS has its own limitations; for example, it can only be used when all indicators are reflective (Hwang and Takane 2014). Becker et al. (2013) recognized these issues and proposed a prediction-oriented segmentation in their PLS–SEM (PLS–POS) that is also available in SmartPLS 3. Population heterogeneity in PLS is discussed in more detail in Hair et al. (2017).

In GSCA, Hwang et al. (2007) proposed a way to integrate cluster analysis utilizing fuzzy clustering that allows a non-zero probability of membership over heterogeneous populations. This is attractive because the estimation of membership probability is based on a global estimation and the function of weights,

commonly referred to as the "*fuzzifier*" (Bezdek 1974). In this framework, Hwang et al. (2007) demonstrated the efficiency of fuzzy clustering GSCA using a small-scale Monte Carlo study and illustrated its utility for latent curve models. However, it is somewhat limited when it comes to discussing observed categorical responses for group-level respondent heterogeneity.

## 3 Research question

Although the efficiency and appropriateness of studying heterogeneous subpopulations utilizing GSCA has been discussed (Hwang et al. 2007; Hwang and Takane 2014, Ch. 4), as yet there has been no consideration of how the observed categorical outcome variables can be dealt with in the GSCA framework. We are, therefore, proposing a new unified framework for fitting latent class analysis in categorical data within GSCA that utilizes both optimal scaling (Hwang and Takane 2010) and fuzzy clustering (Hwang et al. 2007). After describing these methodologies, we will illustrate the component-based latent class analysis using substance usage data from Add Health (Add Health; Harris 2009). We used the *R* program (*R* Core Team 2017) to run fuzzy clustering GSCA for categorical observed responses.

## 4 Method

### 4.1 Latent class analysis using fuzzy clustering in GSCA

#### 4.1.1 GSCA model specification

As stated earlier, GSCA is a component-based approach to SEM. It consists of three sub-models: measurement, structural, and weighted relation models. The first two of these models are the same as the general structural equation system, also known as the LISREL model (Jöreskog 1973, 1977, 1978), while the weighted relation model defines a latent variable as a weighted composite (or component) of indicators in a way that is unique in component-based analyses. Utilizing the same notations as Hwang and Takane (2014), we can write these sub-models in matrix form as follows: Measurement model, $z = C^T\gamma + \varepsilon$ Structural model, $\gamma = B^T\gamma + \zeta$, and Weighted relation model, $\gamma = W^T z$ where $z$ is a $J$ by 1 vector of indicators, $\gamma$ is a $P$ by 1 vector of latent variables, $C$ is a $P$ by $J$ matrix of loadings, $B$ is a $P$ by $P$ matrix of path coefficients, $W$ is a $J$ by $P$ matrix of component weights, $\in$ is a $J$ by 1 vector of the residuals of indicators, and $\zeta$ is a $P$ by 1 vector of the residuals of latent variables. The superscript $T$ signifies a transpose matrix.

#### 4.1.2 Estimation

GSCA estimates model parameters, including weights ($W$), path coefficients ($B$), and loadings ($C$), by minimizing the sum of the squares of the residuals, $e_i$, i.e., consistently minimizing a single least square (LS) criterion defined by

$$\boldsymbol{\Phi} = \sum_{i=1}^{N} e_i^T e_i = \sum_{i=1}^{N} \left(V^T z_i - A^T W^T z_i\right)^T \left(V^T z_i - A^T W^T z_i\right),$$

where $A = \begin{bmatrix} C^T \\ B^T \end{bmatrix}$, $V = \begin{bmatrix} I \\ W^T \end{bmatrix}$, and $N$ is the sample size. To maintain consistent scaling for the indicators and latent variables, both the indicators and the latent variables must be standardized. Standard errors for the parameter estimates are computed using the bootstrap method (Efron 1979, 1982). More details of the computation and algorithms involved are available in Hwang and Takane (2014, Ch. 2) and the corresponding R code is also available upon request from the authors.

The alternating LS algorithm in the GSCA estimation (Hwang and Takane 2014) provides a unique and global solution that is independent of identification issues. Unlike GSCA, identification issues are a common problem in ML estimation, especially when the models are underidentified or unidentified. Models that are not well-identified cannot generate a unique ML solution due to their multimodality (Collins and Lanza 2010). One of the benefits of utilizing GSCA is that there is no identification issue.

#### 4.1.3 Optimal scaling for categorical variables

Hwang and Takane (2010) extended GSCA to include categorical indicators, making it possible to apply GSCA to qualitative data such as nominal and categorical data. The authors dubbed their new method nonlinear GSCA (NL-GSCA). In the same paper, they resolved the linearity issue afflicting LS methods by applying the same optimal scaling method used by other researchers and reported in the IRT literature (see, for example, McDonald 1999). In NL-GSCA, each indicator, $z_j$, where $j = 1, \ldots, J$, is transformed by $s_j = \omega(z_j)$, where $\omega$ depends on the measurement characteristics of the variable, $z_j$. This requires an additional step in the estimation of GSCA but, just as in conventional GSCA, NL-GSCA applies the LS estimation minimizing technique. The LS estimation for NL-GSCA is implemented by minimizing the following criterion:

$$\phi = SS(SV - SWA)$$

with respect to $W, A$, and $S$, where $S = [s_j]$, $V = \begin{bmatrix} I \\ W^T \end{bmatrix}$, and $A = \begin{bmatrix} C^T \\ B^T \end{bmatrix}$, subject to the restrictions that $\text{diag}\left((SW)^T SW\right) = I$, $s_j^T s_j = 1$, and $s_j = \omega(z_j)$. This criterion is minimized by alternating two phases. The first of these phases is identical to the alternating LS estimation procedure used for updating model parameters ($W$ and $A$) in the conventional GSCA for quantitative data. The second is an optimal scaling phase in which qualitative data are transformed to quantitative data $S$ in such a way that they agree maximally with their model predictions while at the same time preserving the measurement characteristics of the data. In practice, variables in the original data matrix $Z$ may consist of a mix of different measurement characteristics; for example, some variables are nominal, others are ordinal, and yet others are numerically defined, as explained in Young (1981). This flexibility in measurement

characteristics allows us to apply NL-GSCA to various IRT models in factor based-IRT including, for example, logistic models and graded response models, among others.

#### 4.1.4 Fuzzy clustering algorithm

Hwang and Takane (2014) described how to utilize fuzzy clustering in GSCA for continuous data. 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 $u_{ki}$:

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

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 parameters in GSCA in class, $k$), subject to the probabilistic condition, $\sum_{k=1}^{K} u_{ki} = 1$. The exponent, $m$, is referred to as the fuzzifier; $m$ becomes fuzzier as it gets larger but less so, or harder, as it approaches 1. In practice, $m = 2$ is most commonly used (Bezdek 1981) and was thus also used in this study.

#### 4.1.5 Fuzzy clustering NL-GSCA for LCA

For binary or ordinal data, we can combine optimal scaling with the fuzzy clustering algorithm to conduct a latent class analysis. Here, binary and ordinal data are dealt within GSCA via optimal scaling by performing the transformation, $s_j = \omega\left(z_j\right)$., which is the first step in estimating the parameters. The fuzzy clustering step can then be applied to update the membership probabilities, $u_{ki}$, based on their distances from the centroids. Alternating this process with the termination criteria provides the final membership based on the posterior probability, with the membership being assigned to the group whose posterior probability is the highest.

In terms of clustering, due to the nature of the difference between CLCA and ML LCA there is no one-to-one correspondence between the parameters in CB LCA and ML LCA when the parameters are interpreted. The fuzzy clustering in LCA identifies the distances from the centroids, whereas ML in LCA relies on the likelihood function. Thus, instead of competing with each other, it is reasonable to suggest that researchers select one of the approaches, either CB LCA or ML LCA, following Hwang et al. (2017), who discussed the conceptual difference between two approaches, namely the component-based SEM and the factor-based SEM, respectively, for such cases.

### 4.2 Model evaluation

For the clustering, FIT and AFIT indices were used to evaluate the fitted GSCA models (see Hwang and Takane, 2014, for more details). Corresponding to the terms

*R* squared and adjusted *R* squared in regression analysis, FIT and AFIT are interpreted as the variance explained by the fitted model. Thus, for both FIT and AFIT, the larger the value, the more of the variance is explained.

To decide the number of clusters, we also used the fuzziness performance index (FPI) and the normalized classification entropy (NCE), as recommended by Roubens (1982) and defined as follows:

$$\mathrm{FPI} = 1 - \frac{\left(K \times \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ki}^2 - 1\right)}{K - 1}$$

and

$$\mathrm{NCE} = \frac{-\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ki} \log u_{ki}}{\log K},$$

where the number of clusters is chosen as an elbow point from the trajectory of those values over the number of clusters. Here, smaller values of FPI and NCE indicate a more distinct separation of the clusters from each other.

### 4.3 Add health data

In this study, we used data from the National Longitudinal Study of Adolescent to Adult Health (Add Health; Harris 2009), which investigates how health factors in childhood affect adult outcomes. Since its inception in 1994, there have been four additional waves of data collection with the fifth and most recent being collected between 2016 and 2018. The first wave was conducted during the 1994–1995 academic year, when respondents were 7th through 12th graders. Here, we focus on wave IV, specifically the section on Tobacco, Alcohol, and Drugs. The 15,701 participants who participated in Wave IV were aged from 24 to 32, and the number of Wave IV participants was markedly lower than the 20,745 participants in Wave I. This study utilized publically available data on 5144 of the participants in the de-identified dataset, hence IRB approval was not required.

In Wave IV, the section of interest examined substance use and abuse, specifically the use of cigarettes, tobacco, marijuana and other substances. For example, the first question asked "*Have you ever smoked an entire cigarette?*" with the answer choices "*no*", "*yes*", "*don't know*", and "*refused*", coded 0, 1, 6, and 8 respectively; the survey responses indicated that 65.0% of the respondents had smoked an entire cigarette (Harris and Udry 2018). In this study, we used five items, namely Smoking, Alcohol, Drug that is not prescribed, Marijuana, and Cocaine, dichotomizing the responses as either "no" or "yes" and treating the other options as missing. The raw frequencies are summarized in Table 1. Although the overall frequencies appear to be relatively high for Smoking, Alcohol, and Marijuana, this does not indicate whether or not these characteristics can be applied to each individual as population parameters. To determine whether there is any heterogeneity in the preferences for

the five substance uses, we therefore fitted both component-based and maximum likelihood based latent class analysis into the Add Health data.

## 5 Results

### 5.1 Fuzzy clustering NL-GSCA for LCA

After applying listwise deletion, 5059 of the 5114 samples were used to enumerate the latent classes and parameter estimates. Based on the responses to the use of the five substances, namely smoking, alcohol, drugs, marijuana, and cocaine, we considered from two-solution to five-solution models. The results from the models were stable and unique except for five-solution model that had an issue of local maxima. Over replication with the different initial values, we had two sets of estimates and presented the one of two estimates in Table 2. The model fit indexes in Table 2 show stable values $>0.998$ in both FIT and AFIT. As the number of classes increased, various distributions of membership probabilities were observed, making it difficult to select the best fit model among these four LCA models based on FIT and AFIT. Looking at the cluster validity measures, the 3-solution model performed the best for both FPI and NCE in terms of deciding the number of clusters (FPI $=0.770$ and NCE $=0.802$) by applying the elbow rule. These results are summarized in Table 2.

In the 3-solution model, the first class (LC1) contains 2781 respondents (55.0%), the second class (LC2) 1187 (23.5%), and the third class (LC3) 1091 (21.6%). LC1 is thus referred to as the *smoking and drinking* class (with both smoking (41%) and alcohol (65%) being relatively higher than the other substance uses, as shown in Fig. 1). LC2 is the *heavy substance user* class (with 91, 100, 68, 91, and 72% of the respondents having smoked cigarettes, drunk alcohol, used non-prescription drugs, smoked marijuana, and experienced cocaine, respectively, as shown in Fig. 1). LC3 is the *binge drinking and heavy smoking* class (with respondents reporting high levels of smoking (99%), alcohol (99%), and marijuana (98%) but very low levels of drug (2%) and cocaine (1%) use, as shown in Fig. 1). As the data presented in Fig. 1 indicate, participants in two groups, LC1 and LC3, are rarely exposed to drugs and cocaine, but those in LC2 tend to be utilizing all five substances.

### 5.2 LCA based on maximum likelihood estimation

The data for the same 5059 participants were then used to enumerate the latent classes and estimate the parameters using Mplus (Muthén and Muthén 1998–2017)

**Table 1** Raw frequency from ADD Health data

| Item | Frequency | Percentage (%) |
|---|---|---|
| Smoking | 3324 | 65.00 |
| Alcohol | 4087 | 79.90 |
| Drugs | 878 | 17.20 |
| Marijuana | 2778 | 54.30 |
| Cocaine | 972 | 19.00 |

**Table 2** Latent class memberships with model evaluation in fuzzy clustering GSCA

| | Membership | | Cluster validity measures | | Model fit indexes | |
|---|---|---|---|---|---|---|
| | Cluster size | Percentage | FPI | NCE | FIT | AFIT |
| 2-Solution | | | | | | |
| Class1 | 2342 | 46.29 | 0.917 | 0.939 | 1.000 | 1.000 |
| Class2 | 2717 | 53.71 | | | | |
| 3-solution | | | | | | |
| Class1 | 2781 | 54.97 | 0.770 | 0.802 | 0.999 | 0.999 |
| Class2 | 1187 | 23.46 | | | | |
| Class3 | 1091 | 21.57 | | | | |
| 4-solution | | | | | | |
| Class1 | 1817 | 35.92 | 0.754 | 0.776 | 0.999 | 0.999 |
| Class2 | 1149 | 22.71 | | | | |
| Class3 | 874 | 17.28 | | | | |
| Class4 | 1219 | 24.10 | | | | |
| 5-solution | | | | | | |
| Class1 | 1275 | 25.20 | 0.724 | 0.728 | 0.998 | 0.998 |
| Class2 | 1096 | 21.66 | | | | |
| Class3 | 682 | 13.48 | | | | |
| Class4 | 874 | 17.28 | | | | |
| Class5 | 1132 | 22.38 | | | | |

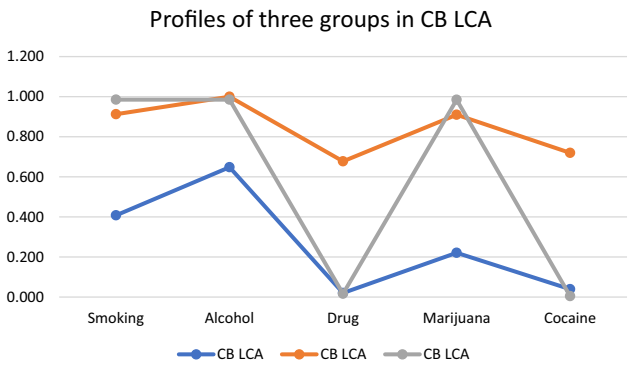


**Fig. 1** Profiles of three latent classes from fuzzy clustering GSCA

utilizing full information maximum likelihood estimation and `poLCA` (Linzer and Lewis 2013) in *R* and the results compared with those obtained using fuzzy clustering GSCA. The two sets of results were almost identical except for the numbering of the latent classes. Note that both Mplus and `poLCA` use identical estimation methods. The 2-solution LCA model was run without any warning message being shown (BIC = 25,141.75; Entropy = 0.71; LC1 = 45.6% and LC2 = 54.4%). However, from the 3-solution LCA, the results were not deemed trustworthy due to issues related to

the estimation, including problems with the optimization, local maxima, and singularity and/or non-positive definiteness, which are listed in Table 3. Nevertheless, we have listed the results from Mplus for the purpose of comparison.

The 3-solution LCA was as follows: BIC = 24,785.54, Entropy = 0.67, LC1 = 46.5%, LC2 = 23.2%, and LC3 = 30.3%, with one warning message related to optimization. In the 4-solution LCA model, three warning messages were generated, for optimization, local maxima and singularity (BIC = 24,825.78; Entropy = 0.73; LC1 = 21.3%, LC2 = 30.3%, LC3 = 46.5%, and LC4 = 1.9%). The 5-solution LCA model showed similar warning messages to those thrown up in the 4-solution LCA model, for optimization, local maxima and singularity (BIC = 24,871.51; Entropy = 0.644; LC1 = 24.2%, LC2 = 29.5%, LC3 = 22.9%, LC4 = 2.0%, and LC5 = 21.6%). Although the 3-solution model did show an optimization issue, we opted to use the 3-solution LCA for this discussion for the purposes of comparison. The 3-solution model also produced the smallest BIC among the models considered.

### 5.3 Comparison of classifications between the two different approaches

Both the fuzzy clustered NL-GSCA for LCA and the maximum-likelihood LCA produced fairly comparable latent classes (71.3%, 3606 out of a total sample of 5059, were classified as being members of the same latent classes; these participants are shown as the diagonal entries in Table 4). The largest discrepancy occurred for the *smoking and drinking* class (LC1) in the fuzzy clustering NL-GSCA for LCA, where 1174 of the total 2781 were instead in the LC3 group (*binge drinking and heavy smoking* class) in ML LCA. Looking at the data presented in Table 5, the average posterior membership probabilities of the inconsistent membership assignments are generally lower. That is, the average posterior membership probabilities of LC1 in CB LCA are 0.695 for LC1 in ML LCA (identical categorization), 0.43 for LC2 in ML LCA, and 0.523 for LC3 in ML LCA. This makes sense because the other membership probabilities for LC 1 in both CB LCA and ML LCA would be relatively high for those whose memberships have changed to LC2 in ML LCA and to LC3 in ML LCA. This suggests that the discrepancy is likely to occur when the posterior probability is not distinct. Another explanation of the differences between two methods is that a different assignment of LC1 in ML LCA yields a probability of saying "yes" to marijuana use of 0.000 (Fig. 2), while LC1 in CB LCA has a probability of 0.221 (Fig. 1). That is, participants placed in LC1 by CB LCA are less likely to be assigned to LC1 by ML LCA and instead much more easily assigned to either LC2 or LC3, where the probabilities of saying "yes" to marijuana use are 0.987 and 0.686, respectively (Fig. 2).

In addition to the discrepancy described above, we also observed a number of small discrepancies: 125 (2.5%) who were assigned to LC1 in CB LCA and LC2 in ML LCA; 39 (0.8%) who were assigned to LC2 in CB LCA and LC1 in ML LCA; 97 (1.9%) who were assigned to LC2 in CB LCA and LC3 in ML LCA; 13 (0.3%) who were assigned to LC3 in CB LCA and LC1 in ML LCA; and 5 (0.1%) who were assigned to LC3 in CB LCA and LC2 in ML LCA (Table 4). It is interesting to note, however, that the differences in the average membership probabilities

**Table 3** Estimation issues in maximum likelihood based latent class analysis that are listed in Mplus

| | Optimization | Local maxima | Singularity and/or non-positive definite |
|---|---|---|---|
| 3-solution | In the optimization, one or more logit thresholds approached and were set at the extreme values. (extreme values are -15.000 and 15.000.)<br>The following thresholds were set at these values:<br>Threshold 1 of class indicator marijuana for class 2 at iteration 69<br>Threshold 1 of class indicator cocaine for class 2 at iteration 69 | | |
| 4-solution | In the optimization, one or more logit thresholds approached and were set at the extreme values. (extreme values are -15.000 and 15.000.)<br>The following thresholds were set at these values:<br>Threshold 1 of class indicator marijuana for class 2 at iteration 105<br>Threshold 1 of class indicator cocaine for class 2 at iteration 105 | Warning: the best loglikelihood value was not replicated. the solution may not be trustworthy due to local maxima. increase the number of random starts. | One or more parameters were fixed to avoid singularity of the information matrix. the singularity is most likely because the model is not identified, or because of empty cells in the joint distribution of the categorical variables in the model.<br>The following parameters were fixed:<br>Parameter 17, %c#4%: [alcohol$1] |
| 5-solution | In the optimization, one or more logit thresholds approached and were set at the extreme values. (extreme values are -15.000 and 15.000.)<br>The following thresholds were set at these values:<br>threshold 1 of class indicator cocaine for class 2 at iteration 158<br>Threshold 1 of class indicator alcohol for class 3 at iteration 158<br>Threshold 1 of class indicator smoking for class 4 at iteration 158<br>Threshold 1 of class indicator alcohol for class 4 at iteration 158 | The best loglikelihood value has been replicated. rerun with at least twice the random starts to check that the best loglikelihood is still obtained and replicated. | The standard errors of the model parameter estimates may not be trustworthy for some parameters due to a non-positive definite first-order derivative product matrix. this may be due to the starting values but may also be an indication of model nonidentification. (the condition number is 0.104d-13.)<br>Problem involving the following parameter:<br>parameter 4, %c#1%: [marijuana$1] |

**Table 4** Crosstab of memberships obtained from ML-based LCA and fuzzy clustering GSCA

| | Fuzzy clustering GSCA | | | Total |
|---|---|---|---|---|
| | Smoking and drinking class (LC1) | Heavy substance user class (LC2) | Binge drinking and heavy smoking class (LC3) | |
| ML LCA | | | | |
| LC1 | 1482 | 39 | 13 | 1534 |
| LC2 | 125 | 1051 | 5 | 1181 |
| LC3 | 1174 | 97 | 1073 | 2344 |
| Total | 2781 | 1187 | 1091 | 5059 |

**Table 5** Average posterior probabilities of memberships where diagonal entries represent identical assignment and off-diagonal entries represent different assignment across CB LCA and ML LCA, where the values in parentheses are the standard deviations of the probabilities

| | Fuzzy clustering GSCA | | | |
|---|---|---|---|---|
| | Smoking and drinking class (LC1) | Heavy substance user class (LC2) | Binge drinking and heavy smoking class (LC3) | Total |
| ML LCA | | | | |
| LC1 | CB: 0.695 (0.047) | CB: 0.378 (0.000) | CB: 0.385 (0.000) | 1534 |
| | ML: 0.854 (0.081) | ML: 0.601 (0.000) | ML: 0.917 (0.000) | |
| LC2 | CB: 0.430 (0.042) | CB: 0.730 (0.094) | CB: 0.406 (0.043) | 1181 |
| | ML: 0.837 (0.148) | ML: 0.898 (0.143) | ML: 0.747 (0.196) | |
| LC3 | CB: 0.523 (0.043) | CB: 0.472 (0.078) | CB: 0.619 (0.000) | 2344 |
| | ML: 0.781 (0.156) | ML: 0.780 (0.046) | ML: 0.899 (0.000) | |
| Total | 2781 | 1187 | 1091 | 5059 |

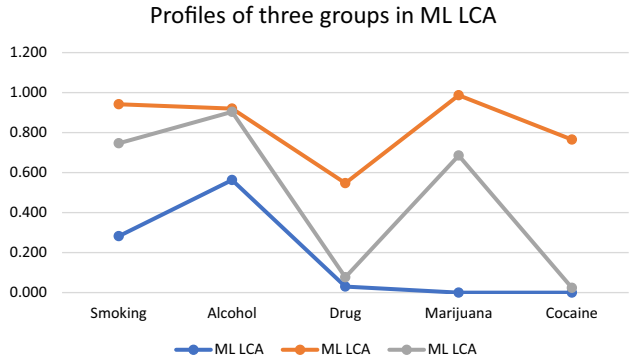


**Fig. 2** Profiles of three latent classes from maximum likelihood based latent class analysis

for ML LCA are not distinct across other membership classes. For example, ML LCA assigned 46.3% into the *binge drinking and heavy smoking* class, but CB LCA instead distributed this class into 50.1% (1174 out of 2344) in LC1 in and 4.1% (97

out of 2344) in LC2 (Table 4). The average membership probabilities for LC3 in ML LCA in Table 5 are 0.781, 0.780, and 0.899 for LC1, LC2, and LC3 in CB LCA, respectively, which does not follow the same trend as that shown for LC3 in CB LCA.

## 6 Discussion

In this study, we introduced a new approach for conducting latent class analyses using fuzzy clustering NL-GSCA. It extends the fuzzy clustering GSCA (Hwang and Takane 2014, Ch. 4) to categorical variables by utilizing optimal scaling (Hwang and Takane 2010). Using five types of substance usage, namely Smoking, Alcohol, Drugs, Marijuana, and Cocaine, for which data were gathered as part of the Add Health survey (Harris 2009), we fitted component-based latent class analyses, finding three latent classes that provided the best fitting model from the 2-solution to 5-solution options tested. This result is consistent with the results obtained from a maximum likelihood based latent class analysis using Mplus or `poLCA` in *R*, with 71.3% of the total sample assigned to the same groups. However, the results are not identical and nor should they be. For example, the posterior probabilities are not the same (compare Figs. 1 and 2) due to the two different estimation methods.

The different estimation methods used, least square estimation and maximum likelihood estimation, mean that this discrepancy was expected but it is interesting that the discrepancy arises mainly due to the relatively low posterior probabilities in CB LCA. The standard deviations for the average membership probabilities of CB LCA, which averaged 0.039, are relatively small compared to those for ML LCA, which averaged 0.086 (Table 5). However, this study was not designed to examine the efficiencies of these two approaches and it is also important to note that the two conceptual LCA frameworks are not fundamentally different from one another. Rather, the difference between them can be considered simply as selecting whichever of the two approaches, CB LCA and ML LCA, incurs the lowest computational cost and provides the best conceptual fit for a particular research study, as Hwang et al. (2017) pointed out. In some cases, the different approaches may provide classifications of the underlying heterogeneous subpopulations that complement each other. For example, a small sample case may benefit from CB LCA. Thus, instead of further arguing which one is better, we plan to explore the benefits or advantages of each approach in a future study.

As always, this study suffers from two main limitations. Missing data are common in social and behavioral data. Unfortunately, GSCA is somewhat limited in its ability to handle missing data, which it normally deals with by applying either listwise deletion or multiple imputation. In this study, we used listwise deletion for about 1% of missing data. The other important limitation is that as yet there are far fewer model evaluation tools for GSCA compared with those available for the more widely used FB-SEM. There have been some attempts to address this lack: for example, Nylund et al. (2007) studied the performance of model evaluation tools including information criteria and likelihood based tests in mixture modeling and suggested BIC and the bootstrap likelihood ratio test based on the results of their

simulation study. More research is needed that focuses specifically on examining the performance of the existing model evaluation tools available in GSCA such as confirmatory tetrad analysis (Ryoo and Hwang 2017) and we hope to take this further by developing new model evaluation tools in the future (Ryoo et al. 2015).

## 7 Conclusion

Despite the popularity of maximum likelihood (ML) based latent class analysis (LCA), such estimations are in practice limited to very small numbers of latent classes of up to three or four. ML based LCA requires large sample sizes, typically 1000 or more, which is another barrier for applied researchers in education, for example. The new LCA approach utilizing fuzzy clustering NL-GSCA proposed in this paper expands the applicability and capability of LCA, which is a major breakthrough in both the GSCA literature and for ML based SEM.

### Compliance with ethical standards

**Conflict of interest** "On behalf of all authors, the corresponding author states that there is no conflict of interest." Categorical Latent Variable Modeling Utilizing Fuzzy Clustering Generalized Structured Component Analysis as an Alternative to Latent Class Analysis.

## References

Becker JM, Rai A, Ringle CM, Völckner F (2013) Discovering unobserved heterogeneity in structural equation models to avert validity threats. MIS Q 37(3):665–694

Bezdek JC (1974) Numerical taxonomy with fuzzy sets. J Math Biol 1:57–71

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York

Collins L, Lanza S (2010) Latent class and latent transition analysis: with applications in the social, behavioral, and health sciences. Wiley, New York

Dziak JJ, Lanza ST, Tan X (2014) Effect size, statistical power, and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. Struct Equ Model 21(4):534–552

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26

Efron B (1982) The jackknife, the bootstrap and other resampling plans. SIAM, Philadelphia

Esposito Vinzi V, Trinchera L, Squillacciotti S, Tenenhaus M (2008) REBUS–PLS: a response-based procedure for detecting unit segments in PLS path modeling. Appl Stoch Models Bus Industry 24:439–458

Goodman LA (1974a) The analysis of systems of qualitative variables when some of the variables are unobservable. Part I—a modified latent structure approach. Am J Sociol 79:1179–1259

Goodman LA (1974b) Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61:215–231

Goodman LA (1979) On the estimation of parameters in latent structure analysis. Psychometrika 44:123–128

Gudicha DW, Schmittmann VD, Vermunt JK (2016) Power computation for likelihood ratio tests for the transition parameters in latent Markov models. Struct Equ Model 23:234–245

Hahn C, Johnson DM, Herrmann A, Huber F (2002) Capturing customer heterogeneity using a finite mixture PLS approach. Schmalenbach Bus Rev 54:243–269

Hair JF, Hult GTM, Ringle CM, Sarstedt M (2017) A primer on partial least squares structural equation modeling (PLS–SEM), 2nd edn. Sage, Thousand Oaks

Harris KM (2009) The national longitudinal study of adolescent to adult health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002; Wave IV, 2007–2009 (Machine-readable data file and documentation). Chapel Hill: Carolina Population Center, University of North Carolina at Chapel Hill. Retrieved from https://doi.org/10.3886/ICPSR21600.v21

Harris KM, Udry JR (2018) National longitudinal study of adolescent to adult health (Add Health), 1994–2008 [Public Use]. Ann Arbor, MI: Carolina Population Center, University of North Carolina-Chapel Hill [distributor], Inter-university Consortium for Political and Social Research [distributor], 2018-08-06. https://doi.org/10.3886/ICPSR21600.v21

Hwang H, Takane Y (2004) Generalized structured component analysis. Psychometrika 69:81–99

Hwang H, Takane Y (2010) Nonlinear generalized structured component analysis. Behaviormetrika 34:95–109

Hwang H, Takane Y (2014) Generalized structured component analysis: a component-based approach to structural equation modeling. CRC Press, Boca Raton

Hwang H, DeSarbo SW, Takane Y (2007) Fuzzy clusterwise generalized structured component analysis. Psychometrika 72:181–198

Hwang H, Takane Y, Jung K (2017) Generalized structured component analysis with uniqueness terms for accommodating measurement error. Front Psychol 8:2137

Jeon M, Rabe-Hesketh S (2012) Profile-likelihood approach for estimating generalized linear mixed models with factor structures. J Educ Behav Stat 37:518–542

Jöreskog KG (1973) A general method for estimating a linear structural equation system. In: Goldberger AS, Duncan OD (eds) Structural equation models in the social sciences. Seminar Press, New York

Jöreskog KG (1977) Structural equation models in the social sciences. In: Krishnaiah PR (ed) Applications of statistics. North-Holland, Amsterdam

Jöreskog KG (1978) Structural analysis of covariance and correlation matrices. Psychometrika 43:443–477

Lazarsfeld PF, Henry NW (1968) Latent structure analysis. Houghton, Mifflin, New York

Linzer DA, Lewis J (2013) "poLCA: polytomous variable latent class analysis." R package version 1.4. http://dlinzer.github.com/poLCA

Lord FM (1952) A theory of test scores. Psychometric Monograph, No, p 7

Lubke GH, Muthén B (2005) Investigating population heterogeneity with factor mixture models. Psychol Methods 10(1):21–39

Masyn KE (2013) Latent class analysis and finite mixture modeling. In: Little TD (ed) Oxford library of psychology. The oxford handbook of quantitative methods: statistical analysis. Oxford University Press, New York, pp 551–611

McDonald RP (1999) Test theory: a unified treatment. Lawrence Erlbaum Associates, Mahwah

Muthén B, Asparouhov T (2006) Item response mixture modeling: application to tobacco dependence criteria. Addict Behav 31:1050–1066

Muthén LK, Muthén BO (1998–2017) Mplus User's Guide. Eighth Ed. Los Angeles, CA: Muthén & Muthén

Muthén BO, Shedden K (1999) Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics 55:463–469

Nagin D (2005) Group-based modeling of development. Harvard University Press, Cambridge

Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc Ser A 135:370–384

Nylund KL, Asparouhov T, Muthén BO (2007) Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. Struct Equ Model 14(4):535–569

Pastor DA, Beretvas SN (2006) Longitudinal Rasch modeling in context of psychotherapy outcomes assessment. Appl Psychol Meas 30(2):100–120

Ringle C, Wende S, Becker J-M (2015) *SmartPLS 3*. Bönningstedt: SmartPLS. http://www.smartpls.com. Accessed 30 Nov 2018

Roubens M (1982) Fuzzy clustering algorithms and their cluster validity. Eur J Oper Res 10:294–301

Ryoo JH, Hwang H (2017) Model evaluation in the generalized structured component analysis using the confirmatory tetrad analysis. Front Psychol Quant Psychol Meas 8:916

Ryoo JH, Chatterjee S, Shi D (2015) New variable selection criteria in model selection. In: Paper presented at the annual meeting of the modern modeling methods conference, Storrs, CT

Ryoo JH, Wang C, Swearer S, Hull M, Shi D (2018) Longitudinal model building using latent transition analysis: an example of using school bullying data. Front Psychol Quant Psychol Meas 9:675

Skrondal A, Rabe-Hesketh S (2004) Generalized latent variable modeling: multilevel, longitudinal, and structural equation models. Chapman & Hall/CRC, Boca Raton

R Core Team (2017). R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. https://www.R-project.org/. Accessed 30 Nov 2018

Wilson M, Zheng X, McGuire L (2012) Formulating latent growth using an explanatory item response model approach. J Appl Meas 13(1):1–22

Wold H (1975) PLS path models with latent variables: the NIPALS approach. In: Blalock HM, Aganbegian A, Borodkin FM, Boudon R, Cappecchi V (eds) Quantitative sociology: international perspectives on mathematical and statistical modeling. Academic Press, New York, pp 307–357

Yang JS, Zheng X (2018) Item response data analysis using Stata item response theory package. J Educ Behav Stat 43(1):116–129

Young FW (1981) Quantitative analysis of qualitative data. Psychometrika 46:347–388



## Affiliations

**Ji Hoon Ryoo[1] · Seohee Park[2] · Seongeun Kim[3]**

[1] 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, 300B Smith Research Tower, 4650 Sunset Blvd., #160, Los Angeles, CA 90027, USA

[2] The University of Iowa, Iowa City, IA, USA

[3] The University of North Carolina at Greensboro, Greensboro, NC, USA