Latent Class Analysis

Diana Mindrila
University of West Georgia, USA

Abstract

A variety of phenomena examined in education can be described using categorical latent models, which help identify groups of individuals who have a series of characteristics in common. Usually, data are too complex to identify and describe groups through inspection alone; therefore, multivariate grouping techniques such as latent class analysis (LCA) should be employed. LCA refers to a set of classification procedures used to identify subgroups of individuals who have several characteristics in common [1]. The following article provides a general description of LCA. It describes the latent class model and explains the steps involved in latent class modeling. Finally, the article includes an empirical application of LCA with binary indicators using the Mplus statistical software.

1. Introduction

A variety of phenomena examined in education can be described using categorical latent variables, which help identify groups of individuals who share a set of characteristics. Usually the array of observed data is too complex to identify and describe groups through inspection alone; therefore, researchers use person-oriented classification procedures such latent class analysis (LCA) to accomplish this goal. Although LCA has been discussed in the methodological literature, most authors focused either on computational procedures [2], on conceptual explanations [1,3], or on describing empirical applications [4]. The present article provides an overview of all aspects of LCA, by discussing both the theoretical foundations and the practical applications of this method. For the applied researchers, the article includes a step-by-step description of LCA with a clear example of how to apply this procedure using the M*plus* statistical software and how to interpret the results.

2. Milestones in the Development of LCA

Lazarfeld and Henry [5] brought the first major contribution to the development of LCA. Although they were not the first suggesting the possibility of estimating categorical latent variables, they were the first authors to provide a detailed, comprehensive, mathematical and conceptual description of LCA. Nevertheless, the absence of a reliable and general method for calculating parameter estimates posed difficulties in the implementation of this method. The use of LCA became more widespread after 1974, when Goodman [6] developed a more general procedure for computing maximum likelihood parameter estimates. Later, Dempster, Laird, and Rubin [7] developed the “expectation-maximization algorithm”, which placed LCA in the log-linear model framework and increased its generalizability [8]. The expectation-maximization algorithm allowed researchers to assess model fit [9] and to estimate more complex models such as LCA with covariates [10] and longitudinal “individual growth trajectories” [11]. From this point on, researchers developed several procedures for estimating changes in latent class membership over time [12].

3. A General Description of LCA

The LCA method is a multivariate classification procedure that allows researchers to categorize individuals into homogeneous groups [2]. It is sometimes referred to as “mixture modeling based clustering”, “mixture-likelihood approach to clustering”, or “finite mixture modeling” [13]. In fact, “finite mixture modeling” is a more general term for latent variable modeling where latent variables are categorical. The latent categories represent a set of sub-populations of individuals, and individuals’ memberships to these sub-populations are inferred based on patterns of variations in the data [13].

LCA postulates the existence of a categorical variable free of measurement error [1]. The latent construct that underlies the data is not measured directly, it is estimated based on the variance shared by a set of observed variables. Researchers take a similar approach when conducting factor analysis, which also infers underlying latent variables based on the relationships among observed variables. Nevertheless, factor analysis groups variables, and is, therefore, a variable-oriented classification procedure. In contrast, LCA groups individuals and is considered a person-oriented classification procedure [14]. Further, in latent class models the latent variable is categorical rather than continuous [2], which means that groups do not necessarily differ quantitatively but have distinct combinations of characteristics and represent a phenomenon that is inherently categorical rather than continuous [1].

LCA is similar to cluster analysis (CA), which is also a person-oriented classification procedure. An important distinction between CA and LCA is that cluster analysis does not estimate the error of measurement and is, therefore, conducted “at the observed level” [3]. In contrast, LCA assumes that a latent categorical variable underlies the data; it estimates the error of measurement of the latent categorical variable, which is taken into account in the calculation of parameter estimates and the estimation of class membership probabilities. Additionally, LCA allows the computation of fit indices, which show how well the latent class model fits the data [3].

LCA relies on the assumption that homogeneous sub-populations exist within the data. These subgroups have distinct probability distributions and are mutually exclusive [15]; therefore, the percentages of individuals assigned to each latent class add up to 100%. LCA also relies on the assumption of local independence. Specifically, LCA assumes that the variance shared by the observed indicators is accounted for only by the latent categorical variable [16]. Further, LCA assumes that the number of latent classes specified by the latent model is correct [3].

4. The LCA Model

A mixture model includes a measurement model and a structural model. LCA is the measurement model, which consists of a set of observed variables, also referred to as observed indicators, regressed on a latent categorical variable [2]. Figure 1 illustrates relationships between a set of *r* observed indicators *i* and an underlying categorical variable *C* [2].

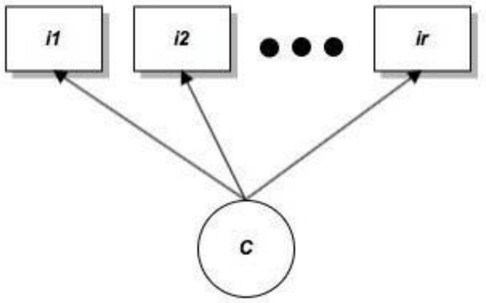

Figure 1. Diagram of a general latent class model

Observed variables can be continuous, counts, ordered categorical, binary, or unordered categorical variables [2]. When estimating a latent variable *C* with *k* latent classes (*C*=*k*; *k*=1, 2,…*K*), the “marginal item probability” for item i_j=1 can be expressed as:

$$P(i\,j\ = 1) = \sum_{k=1}^{K} \quad P(C = k)P(i\,j\ = 1|C = k)$$

Assuming that the assumption of local independence is met, the joint probability for all observed variables can be expressed as:

$$P(i\,1\,, i\,2\,, \ldots\ldots, i\,r\,) = \textstyle\sum\_(k = 1)\text{^}K\ P(C = k)P(C = k)P(C = k) \ldots P(i\,r\ \Big| C = k)$$

When the observed indicators are continuous, parameters are represented by the means and variances of the observed variables and the LCA model can be expressed as:

$$f(y\,p\,) = \sum_{k=1}^{K} \quad P(C = k)f(y\,p\,|C = k)$$

where y_p represents the set of responses provided by a person *p* on all the observed indicators [1].

The statistical model described above is the basis of all types of LCA: exploratory LCA, confirmatory LCA, latent profile analysis (LPA), latent transition analysis (LTA), multilevel LCA, etc. [1]. Although most often used as an exploratory procedure, LCA can also be used as a confirmatory procedure by constraining model parameters based on the researchers’ hypotheses [17]. LPA is a special case of LCA where observed indicators are continuous rather than ordered categorical or binary [2]. LPA is based on the assumption of multivariate normality, meaning that the multivariate distribution of the continuous observed variables is normal within each group [3]. The LTA approach is a variation of LCA for longitudinal data, which allows researchers to examine changes in latent class memberships across time [1]. Multilevel LCA is employed with data from complex sampling designs and allows the estimation of a multilevel latent class model to account for the nested structure of the data. M*plus* permits the estimation of exploratory and confirmatory LCA models, LTA models, as well as multilevel LCA models. Further, M*plus* allows the estimation of relationships among latent categorical variables and second order factors, covariates, or observed dependent variables also known as distal outcomes [2]. The current article will further describe the exploratory LCA model.

4.1 Model estimation

The computation procedures used for estimating model parameters are based on the type of variables used as observed indicators (Table 1). The default estimation method for mixture modeling in M*plus* is robust maximum likelihood (MLR), which uses “log-likelihood functions derived from the probability density function underlying the latent class model” [2]. Other estimation procedures such as maximum likelihood (ML), or Bayesian estimation (BAYES) can be specified in M*plus* using the ESTIMATOR option of the ANALYSIS command. Although MLR corrects standard errors and test statistics, other estimators such as BAYES may provide more accurate results with smaller sample sizes, ordinal data, and non-normal continuous variables [2].

Table 1. Computation procedures by variable type [3]

Variable Type	Computation Procedure
Continuous	Linear regression equations
Censored	Censored-inflated normal regression
Count	Poisson or zero-inflated Poison regression equations
Ordered categorical	Logistic regression
Binary	Logistic regression
Nominal	Multinomial logistic regression

Individual cases are assigned to latent classes through an iterative procedure. Researchers have the option of specifying starting values or to use automatic, random starts. The specification of LCA starting values is very similar to specifying seed values for the *k*-means clustering algorithm. Estimation is repeated until the same solution is obtained from multiple sets of starting values, at which point parameters are considered most likely to be representative of a latent class [1]. If the model does not converge even with many sets of starting values, the latent variable that underlies the data may not have the number of classes specified by the latent class model [2].

Estimated model parameters include item means and variances by latent class. Results also include, for each case, the probability of membership to each class. These probabilities add up to one across latent classes. Latent class membership is determined through modal assignment, by placing each person in the latent class for which the probability of membership is the highest [3].

4.2. Model selection

With exploratory LCA the researcher does not know a priori the number of classes of the latent categorical variable; therefore, several models with different numbers of classes are estimated and compared. The process of determining the number of latent classes is often referred to as "class enumeration" [3]. The optimal number of latent classes is determined by examining several criteria, such as the interpretability of model parameters in relation to theory and prior research, classification accuracy, and the extent to which LCA models fit the data.

Like factor analysis, LCA requires researchers to examine the interpretability of latent class solutions [3]. Item means and variances along with the size and demographic composition of latent classes are examined to make sure that each latent class describes a distinct and meaningful pattern and makes sense based on existing information on the topic. Each latent class is labeled based on its distinct set of characteristics while making sure that the definitions of these subgroups have substantive meaning in relation to theory and prior research [2].

Measures of classification precision are another criterion for selecting the optimal latent class model. For each individual, LCA calculates the probability of membership to each one of the classes specified in the LCA model. These probabilities are also referred to as "posterior probabilities" [16]. Individuals are then assigned to the latent class for which the probability of membership is the highest. When probabilities of membership are close to one for one class and close to zero for all other latent classes, the model has a high level of classification precision.

Membership probabilities for the entire sample are summarized in a *K* x *K* table, where *K* is the number of latent classes specified in the LCA model. The M*plus* LCA output includes a *K* x *K* table of "classification probabilities for the most likely latent class membership (column) by latent class (row)", and a *K* x *K* table of "average latent class probabilities for most likely latent class membership (row) by latent class (column)" [2]. The diagonal elements of these tables represent the average probabilities of membership to the assigned class, or the proportions of correctly classified cases. They are considered indices of classification certainty and should be close to one [4]. In contrast, the off-diagonal elements of the *K* x *K* table represent the average probabilities of membership to the other latent classes specified in the model. These probabilities represent the proportions of misclassified cases and should be close to zero [3].

Entropy is an omnibus index of classification certainty, which relies on the class posterior probabilities reported in the *K* x *K* table. The entropy coefficient indicates the degree to which the entire LCA model accurately predicts individual class memberships, or the extent to which latent classes are distinct from each other [3]. Entropy values can range from zero to one, where values closer to one indicate superior classification precision [3].

The examination of goodness of fit indices, which show the degree to which a hypothesized model fits the data, is another step in selecting the optimal LCA model. The goodness of fit indices estimated with the M*plus* software are a) the Bayesian Information Criteria (BIC), b) the sample-size adjusted BIC, and c) the Akaike Information Criteria (AIC). These coefficients do not have specific cut-offs. When comparing models with a different number of latent classes or different specifications, lower AIC and BIC values indicate better model fit to the data [3]. When the number of parameters to be estimated is increasing, AIC and BIC values tend to also increase. The goal is to obtain a parsimonious model that identifies all the distinct latent classes that underlie the data while maintaining an acceptable model fit [16, 3].

The M*plus* LCA output also includes the results of the Lo-Mendell-Rubin (LMR) likelihood ratio test. The LMR test helps determine whether the inclusion of an additional latent class while maintaining the same model specifications significantly improves the model fit. For a model with *K* latent classes, LMR tests the hypothesis that *K*-1 classes are in fact underlying the data. A significant test statistics indicates that the model with *K* classes is superior to the model with *K*-1 latent classes. Additional classes are specified in the LCA model until the LMR test yields a non-significant test statistic [18].

5. LCA Application Using M*plus*

The following example illustrates the estimation of a latent class model with binary observed indicators using the M*plus* 7.4 software. The goal of this research project was to differentiate latent classes of bullying and cyberbullying victimization based on the prevalence of different forms face-to-face (traditional) bullying victimization and cyberbullying victimization in U.S. adolescents.

5.1 Data Sources

Participants in the study (N=4,939) were U.S. adolescents (ages 12-18) who responded to the 2013 School Crime Supplement of the National Crime Victimization Survey. This survey is administered every two years by the U.S. National Center for Education Statistics and the U.S. Bureau of Justice Statistics using a stratified, multistage cluster sampling design. In 2013, the SCS response rate on all items exceed 85%; therefore, imputation of missing values was not necessary. The SCS sample weights, which are a combination of household weights and person-level weights [19], were used to account for the nested structure of the data.

5.2 Model Specification and Estimation

Fourteen binary survey items measuring bullying (7 items) and cyberbullying (7 items) were specified as observed indicators of a categorical latent variable *C* (see Figure 2). Seven of these variables (bul1-bul7) measured face-to-face (traditional) bullying victimization, while the other seven variables measured cyberbullying victimization (cyb1-cyb7).

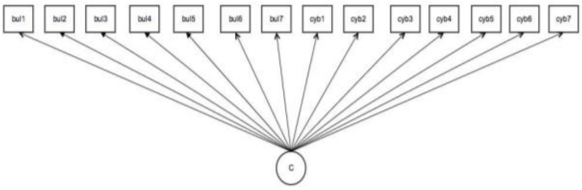

Figure 2. Latent class measurement model

The optimal number of latent classes was determined by estimating and comparing results from models with two, three, four, and five latent classes. The estimation procedure was MLR with automatic random starts. Models were compared based on the interpretability of latent classes, measures of classification precision, and goodness of fit indices.

5.3 Results

Although only the two-class model had a significant test statistic on the LMR test, the four-class model had the highest entropy (0.92) and the lowest BIC value (see Table 2). Further, the classes included in the four-class model were more clearly defined. The four groups differed in the extent to which individuals were victims of traditional bullying and/or cyberbullying (see Figure 3). The largest of the four latent classes (*N*=4,274, 86.5%) was labeled “Non-victims” (NV), because individuals in this group experienced little or no bullying victimization. The second largest latent class (*N*=486, 9.8%) was labeled “Traditional victims” (TV); the majority of individuals in this group experienced mostly traditional forms of bullying such as “being made fun of, called names, or insulted” (75%) or “being the subject of rumors” (65%) [20]. The third latent class included 107 (2.2%) individuals and was labeled “Cyber-victims” (CV); a large proportion of this latent class was the subject of rumors (67%) and also experienced high levels of other forms of cyberbullying such as being insulted “through text messaging” (49%), “through instant messaging or chat” (44%), or “through hurtful Internet posts” (44%) [35]. The fourth latent class was the smallest in size (N=72, 1.5%) but experienced increased levels of both traditional bullying and cyberbullying and was, therefore, labeled “Traditional Victims and Cyber-victims” (TVCV). Most individuals in this group were the subject of rumors (100%), have been “made fun of, called names, or insulted” (96%), were “excluded from activities on purpose” (67%), were “pushed, shoved, tripped, or spit on” (57%), and “threatened with harm” (57%) [20]. Individuals in this latent class also experienced high levels of cyberbullying by being “threatened or insulted through text messaging” (75%), “through hurtful posts” on the Internet (53%), or “through instant messaging” (48%) [20]. Classification probabilities and average latent class probabilities for this model ranged between 70.1% and 99.6% (see Table 3 and Table 4).

Table 2. Goodness of fit indices

Index	Two-class model	Three-class Model	Four-class model	Five-class model
AIC	16698.567	16335.625	16193.244	**16128.342**
BIC	16887.210	16621.842	**16577.034**	16609.706
Sample-Size Adjusted BIC	16795.058	16482.025	16389.553	**16374.560**
Lo-Mendell-Rubin adjusted LRT test				
Statistic	**4515.458**	389.886	171.041	94.164
p-value	**0.0000**	0.3822	0.4337	0.9134
Entropy	0.916	0.909	**0.920**	0.836

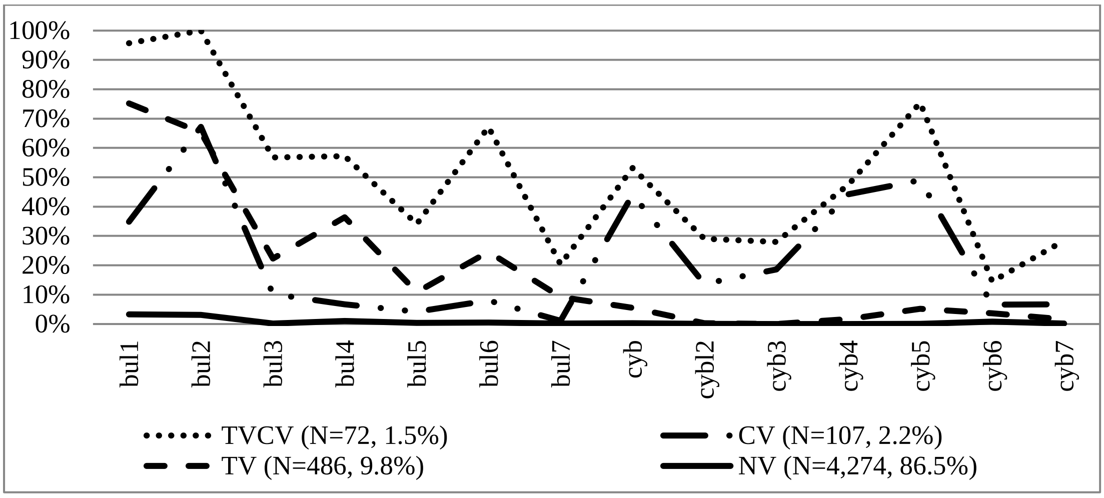

Figure 3. Latent class results in probability scale

Table 3. Classification probabilities

	TVCV	CV	TV	NV
TVCV	**0.818**	0.042	0.140	0.000
CV	0.043	**0.701**	0.157	0.099
TV	0.006	0.016	**0.810**	0.168
NV	0.000	0.000	0.004	**0.996**

Table 4. Average latent class probabilities

	TVCV	CV	TV	NV
TVCV	**0.878**	0.079	0.043	0.000
CV	0.031	**0.872**	0.080	0.017
TV	0.022	0.043	**0.904**	0.031
NV	0.000	0.003	0.021	**0.976**

For each latent class, M*plus* calculated the probability estimates of the observed indicators along with the corresponding *t* statistics (estimate/*SE*) and two-tailed *p* values. For each observed indicator, M*plus* also provided odds ratios to compare item probabilities across latent classes. Statistically significant odds ratio values showed that adolescents assigned to the TVCV latent class were significantly more likely than the adolescents assigned to the CV latent class to be victimized through “hurtful posts” on the Internet [20]. Similarly, the TVCV latent class recorded significantly higher probabilities of traditional victimization and victimization via hurtful Interned posts than individuals in the TV and NV latent classes. Compared to NV, adolescents included in the TV latent class were significantly more likely to be the victims of traditional forms of victimization as well as hurtful Internet posts. Individuals in the CV latent class were significantly more likely to be the victims of hurtful Internet posts and rumors, but significantly less likely to be “made fun of, called names, or insulted” than individuals in the TV latent class [20]. The CV latent class was significantly more likely than the NV latent class to be the target of hurtful Internet posts and rumors, or to be “made fun of, called names, or insulted” [20].

6. Conclusion

Latent class modeling aims to identify unobservable subgroups of individuals who have a set of characteristics in common [1]. Although similar in purpose to cluster analytic algorithms such as *k* means, LCA is conducted at the latent level rather than the observed level. Unlike factor analysis, LCA allows the estimation of categorical latent variables and groups individuals rather than variables.

The M*plus* software allows researchers to specify complex mixture models using several types of observed variables and several estimation methods. M*plus* also calculates measures of classification precision and a series of goodness of fit indices, which permit comparisons across latent class models with different specifications. With a relatively simple code, researchers can run this complex latent modeling procedure and obtain a wide range of results.

7. References

[1] L. M. Collins, and S. T. Lanza (2010). *Latent class and latent transition analysis for the social, behavioral, and health sciences*. New York: Wiley.

[2] L. K. Muthén, and B. O. Muthén, (2010). *Mplus User's Guide: Statistical Analysis with Latent Variables: User's Guide*. Muthén and Muthén.

[3] DiStefano, C. (2012). “Cluster analysis and latent class clustering techniques”, *Handbook of developmental research methods,* 645-666.

[4] K. L. Nylund, T. Asparouhov, and Muthén, B. O. (2007). “Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study”, *Structural equation modeling, 14,* 535-569.

[5] P. F. Lazarsfeld, and N. W. Henry (1968). “Latent structure”, *Psychology, a study of science, 3*.

[6] L. A. Goodman, (1974b). “The analysis of systems of qualitative variables when some of the variables are unobservable. Part IA modified latent structure approach”, *American Journal of Sociology, 79,* 1179-1259.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, (1977). “Maximum likelihood from incomplete data via the EM algorithm”, *Journal of the royal statistical society. Series B (methodological),* 1-38.

[8] J. A. Hagenaars, (1998). “Categorical causal modeling: Latent class analysis and directed log-linear models with latent variables”, *Sociological Methods and Research, 26,* 436-486.

[9] D. Rindskopf (1984). “Linear equality restrictions in regression and loglinear models”, *Psychological Bulletin, 96,* 597.

[10] McReady, G. B. (1988). “Concomitant-Variable Latent Class Models”, *Journal of the american statistical association, 83,* 173-178.

[11] D. S. Nagin, and D. Nagin, (2005). *Group-based modeling of development*. Harvard University Press.

[12] L. M. Collins, and S. E. Wugalter, (1992). “Latent class models for stage-sequential dynamic latent variables”, *Multivariate Behavioral Research, 27,* 131-157.

[13] J. D. Banfield, and A. E. Raftery, (1993). “Model-based Gaussian and non-Gaussian clustering”, *Biometrics,* 803-821.

[14] L. R. Bergman, D. Magnusson, and B. M. El Khouri, (2003). “*Studying individual development in an interindividual context: A person-oriented approach”*. Psychology Press.

[15] L. R. Bergman, and D. Magnusson, (1997). “A person-oriented approach in research on developmental psychopathology”, *Development and psychopathology, 9,* 291-319.

[16] J.K. Vermunt, and J.Magidson, (2002). “Latent class cluster analysis”, In: J.A. Hagenaars and A.L. McCutcheon (Eds.), *Applied latent class analysis*. Cambridge University Press, Cambridge, 89–106.

[17] W. H. Finch, and K. C. Bronk, (2011). “Conducting confirmatory latent class analysis using M plus”, *Structural Equation Modeling, 18,* 132-151.

[18] Y. Lo, N. R. Mendell, and D. B. Rubin, (2001). “Testing the number of components in a normal mixture”, *Biometrika, 88,* 767-778.

[19] S. Burns, X. Wang, and A. Henning, (2011). NCES Handbook of Survey Methods. NCES 2011-609. *National Center for Education Statistics*.

[20] S. Robers, J. Kemp, A. Rathbun, and R. E. Morgan, (2014). Indicators of School Crime and Safety: 2013. NCES 2014-042/NCJ 243299. *National Center for Education Statistics*.