L

Latent Class Analysis

Jinbo He and Xitao Fan
Chinese University of Hong Kong (Shenzhen), Shenzhen, China

Synonyms

LCA

Definition

Latent class analysis (LCA) is a latent variable modeling technique that used for identifying subgroups of individuals with unobserved but distinct patterns of responses to a set of observed categorical indicators (Lanza et al. 2007).

Introduction

Introduced by Lazarsfeld (1950) and further developed and extended by many methodologists later (e.g., Goodman, Haberman, Hagenaars, and Vermunt), LCA has been increasingly utilized in various research fields (e.g., psychology, education, management, and health sciences) as a useful technique for grouping individuals.

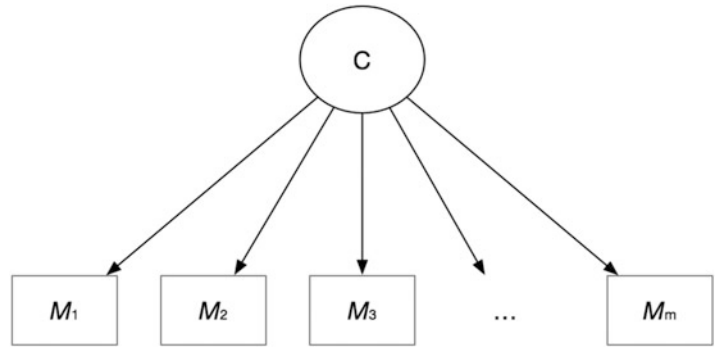
LCA is distinct from traditional clustering approaches (e.g., *k*-means cluster analysis), as LCA offers researchers a model-based (or probability-based) clustering. Because of this, an obvious advantage of LCA over traditional clustering approaches is that the choice of the cluster criterion relies on rigorous statistical tests; as a result, LCA is considered more objective than other approaches.

There are two sets of parameters involved in an LCA model: class membership probabilities (i.e., class proportion parameters, referred as γ_c) and item-response probabilities conditional on class membership (i.e., item parameters, referred as $\rho_{i \mid c}$; Collins and Lanza 2010). Figure 1 is a generic representation of an LCA model, which has c latent classes ($c = 1, 2, \ldots, C$) to be estimated based on a data set with m categorical indicators ($m = 1, 2, \ldots, M$). Suppose indicator m has r_m response categories, $r_m = 1, 2, \ldots, R_m$. If $r_m = 2$, it indicates that the m indicator uses a binary response pattern. Let Y be the vector of response patters and y be a specific response patterns from individuals to the m categorical indicators, $y = (r_1, r_2, \ldots, r_m)$. In a situation where we have four yes/or indicators, one of response patters could be y = (yes, no, no, yes). Then, the probability of observing a particular response pattern can be represented as:

$$P(Y = y) = \sum_{c=1}^{C} \gamma_c \prod_{m=1}^{M} \prod_{r_m=1}^{R_m} \rho_{m,r_m|c}^{I(y_m = r_m)}$$

where $I(y_m = r_m)$ refers to an indicator function that equals 1 if the response to variable $m = r_m$ and

V. Zeigler-Hill, T. K. Shackelford (eds.), *Encyclopedia of Personality and Individual Differences*,
https://doi.org/10.1007/978-3-319-28099-8_2313-1



Latent Class Analysis, Fig. 1 Path diagram of a latent class model

0 otherwise; γ_c is the probability of membership in latent class c, and $\rho_{m,r_m|c}^{I(y_m=r_m)}$ is the probability of response r_m to indicator m in latent class c. For more details on the LCA mathematical definitions, readers are referred to Collins and Lanza (2010).

Currently, there are competing statistical software available for conducting LCA, such as Mplus (Muthén and Muthén 1988–2017), Latent GOLD (Vermunt and Magidson 2005), PROC LCA (Lanza et al. 2007), etc. As each software for LCA may have its strengths and weaknesses, and may differ on dimensions such as cost, usability, data characteristics, and performance (Haughton et al. 2009), researchers need to decide which one to use in their own situations and refer to the manual of a specific software for how to conduct LCA.

Considerations When Conducting LCA

Local Independence

Local independence is a fundamental assumption of LCA (Collins and Lanza 2010), which means that the observed indicators for conducting LCA should be independent of each other such that the probability of a response to each indicator is conditioned on a given latent class membership only. Statistically, the local independence manifests as no highly correlated indicators within the identified latent classes. In LCA, the bivariate residual (BVR) as a diagnostic statistic for local independence can be used to assess the extent to which the association between two observed indicators is reproduced by an LCA model (Magidson and Vermunt 2001).

Sample Size

To date, there is no formal criterion of the minimum sample size for conducting LCA, as the required sample size depends on many factors (e.g., model complexity) and varies in terms of the specific research questions. However, LCA is generally considered as a large-sample size technique, because small sample sizes may lead to convergence issues and failure in identifying small but meaningful subgroups. Nevertheless, most simulation studies have suggested that it might be relatively safe to have around 500 individuals in a sample for LCA studies in applied research (Finch and Bronk 2011).

Model Convergence

Due to the difficulty for modeling a mixture of many different kinds of subdistributions within a sample distribution, convergence issues (e.g., local maxima) are common in the procedure of model estimation in LCA with the method of maximum likelihood (or some variant). To increase the probability of obtaining the best convergence, it is recommended for researchers to set a large number of random starting values when conducting LCA (e.g., 100 or 500 random starts; Jung and Wickrama 2008). Furthermore, most software for LCA provide warning messages if convergence issues happen (e.g., '*the log-*

likelihood is not replicated'). Thus, researchers may determine whether to increase the starting values to a larger number (e.g., 1000 random starts) when they receive such messages.

Missing Data

The occurrence of missing data is very common, and it could have significant influence on the conclusions drawn from the data. Although LCA does not require complete data to run, a large proportion of missing data, and/or data that are not missing at random, are likely to have negative impact on the accuracy of parameter estimation and the model fit indicators. Thus, missing data should be handled properly in LCA. For such purpose, modern methods such as full-information maximum likelihood (FIML) and multiple imputation (MI) can be used (Collins and Lanza 2010).

Model Selection

The most challenging part of LCA might be deciding on the optimal number of classes. In practice, a number of models are evaluated when conducting LCA, starting from a model with one class and adding one more class until the optimal solution is identified. As LCA is based on the framework of Structural Equation Modeling (SEM), model fit indicators are used to indicate the extent to which a particular LCA model fits the empirical data used.

The commonly used model fit indicators for LCA include the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the sample size-adjusted BIC (SABIC), the Vuong-Lo-Mendell-Rubin adjusted likelihood ratio test (VLMR-LRT), the bootstrap likelihood ratio test (BLRT), and Entropy. Specifically, models with lower AIC, BIC, and SABIC have a better balance between model fitting and model parsimony. Statistical *p*-values for the LMR-LRT and BLRT may be used to determine if the model with *k* classes has statistically significant improvement in model fit over the model with one less class (i.e., *k*-1 classes). The Entropy indicator is used to evaluate classification quality (or classification precision), which ranges from 0 to 1 (closer to 1 indicates better classification quality). Furthermore, small classes (e.g., less than 5% of the sample) are generally considered as spurious (unless strong justification can be provided); thus, the relative sizes of the emerged classes need to be considered when choosing the optimal model.

In addition, although the use of model fit indicators for model selection is straightforward, this can be very tricky in practice, because model fit indicators often do not suggest a single solution. Thus, it has been recommended to jointly consider model fit indicators, previous findings, theoretical meaningfulness, conceptual interpretability, as well as classification diagnostics, to determine the optimal model (Masyn 2013).

Inclusion of Covariates and Distal Outcomes

After the identification of latent classes, researchers are often interested in examining whether there are certain antecedents and/or consequences of latent class membership. In LCA, this can be achieved by including covariates and/or distal outcomes. Specifically, the inclusion of a covariate (e.g., gender) is used to explore whether class prevalence is equivalent across the levels of the covariate (e.g., male vs. female), which is usually conducted with multinomial logistic regressions. The inclusion of a distal outcome is used to explore whether individuals in different latent classes have statistically significant differences in the outcome variables (e.g., life satisfaction). It has been well documented that the inclusion of covariates and/or distal outcomes may influence the formation of the latent classes when a one-step approach (i.e., the joint estimation of the latent class variable and its relations to covariates and distal outcomes) is used (Asparouhov and Muthén 2014). Thus, it is recommended to use a three-step approach (i.e., the estimation of latent class variable is prior to the estimation of the relations to covariates and distal outcomes) when including covariates and distal outcomes in LCA models (Asparouhov and Muthén 2014).

Extensions of LCA

There are some advanced extensions of LCA that can be used in certain situations, and/or for exploring additional research questions beyond identifying latent classes. Latent Transition Analysis (LTA), as a longitudinal extension of LCA, is used to model the longitudinal changes among the subgroups over time. Multilevel LCA (MLCA) is preferred when the empirical data have multilevel structures. Confirmatory LCA (CLCA) is used when researchers are interested in testing hypotheses about response patterns in the observed variables.

Conclusion

In this chapter, we introduced LCA and described the major considerations when conducting LCA in research studies. Overall, as an analytic approach to identifying unobserved subgroups with categorical indicators, LCA is a powerful tool for researchers to develop better understanding about individual differences.

Cross-References

- ▶ Latent Profile Analysis
- ▶ Latent Variable Model
- ▶ Structural Equation Modeling

References

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using M plus. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 329–341. https://doi.org/10.1080/10705511.2014.915181.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis. With applications in the social, behavioral, and health sciences*. Hoboken: Wiley.

Finch, W. H., & Bronk, K. C. (2011). Conducting confirmatory latent class analysis using M plus. *Structural Equation Modeling, 18*(1), 132–151. https://doi.org/10.1080/10705511.2011.532732.

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. https://doi.org/10.1198/tast.2009.0016.

Jung, T., & Wickrama, K. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass, 2*(1), 302–317. https://doi.org/10.1111/j.1751-9004.2007.00054.x.

Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling, 14*(4), 671–694. https://doi.org/10.1080/10705510701575602.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Star, J. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Studies in social psychology in world war II Vol. IV: Measurement and prediction* (pp. 362–412). Princeton: Princeton University Press.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology, 31*(1), 223–264. https://doi.org/10.1111/0081-1750.00096.

Masyn, K. E. (2013. Latent class analysis and finite mixture modeling). In T. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (pp. 375–393). Oxford, UK: Oxford University Press.

Muthén, L. K., & Muthén, B. O. (1988–2017). *Mplus: Statistical analysis with latent variables: User's guide* (8th ed.). Los Angeles: Muthén & Muthén.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for latent GOLD 4.0: Basic and advanced*. Belmont: Statistical Innovations.