An R Package for Multilevel Latent Class Analysis with Covariates

Johan Lyrvall [1] , Roberto Di Mari[1] , Zsuzsa Bakk[2] , Jennifer Oser[3] and Jouni Kuha[4]

[1] Department of Business and Economics, University of Catania, (e-mail: johan.lyrvall@phd.unict.it)

[2] Department of Methodology and Statistics, Leiden University

[3] Department of Politics and Government, Ben-Gurion University

[4] Department of Methodology, London School of Economics

Abstract: In this article we introduce *multilevLCA* - an R package for efficient estimation of single-level and multilevel latent class models with covariates.

Keywords: Multilevel latent class analysis, R package, two-step estimation.

1 Introduction

Latent class (LC) analysis is to create a discrete classification of units based on a set of observed variables, which are taken as observed indicators of an unknown nominal variable with some number of latent classes. Multilevel LCA has been developed to account for hierarchical data structures, i.e., when lower-level units are nested within higher-level ones (e.g., survey respondents nested within countries, pupils within schools). The multilevel LC model can be extended to allow for external covariates as predictors of class membership.

The general recommendation for fitting single-level and multilevel LC models with covariates is to use stepwise estimators. In particular, the two-step (Di Mari *et al.*, 2023) and two-stage approaches (Bakk *et al.*, 2022) for multilevel LCA, and the two-step approach for single-level LCA (Bakk & Kuha, 2018) have some attractive properties with respect to model construction, and estimation efficiency and algorithmic stability.

In the current paper we introduce the R package `multilevLCA` - the first to implement two-step estimation, in a functional and user-friendly way, for single-level and multilevel latent class analysis with covariates.

2 Modelling framework

Let Y_{ijh} denote the observed response of low-level unit (individual) i in high-level unit (group) $j = 1,\dots,J$ on the categorical indicator variable $h = 1,\dots,H$. The full response vector for the same unit is denoted $\mathbf{Y}_{ij} = (Y_{ij1},\dots,Y_{ijH})$. For simplicity of exposition, we focus below on dichotomous indicators, with a conditional Bernoulli distribution, $P(Y_{ih} = y_{ih}|X_i = t) = \phi_{h|t}^{y_{ih}}(1-\phi_{h|t})^{1-y_{ih}}$.

Let W_j be a group-level latent class variable, with possible value $m = 1,\dots,M$, and probabilities $P(W_j = m) = \omega_m > 0$. Given a realization of W_j, let X_{ij} be a individual-level latent class variable, with possible values $t = 1,\dots,T$, and conditional probabilities $P(X_i = t|W_j = m) = \pi_{t|m} > 0$.

We assume that individual response probabilities are conditionally independent from each other given low-level class membership (the classical *local independence* assumption). We further assume that individual response probabilities depend on high-level class membership only through X_{ij} (a common assumption in multilevel LCA; Vermunt, 2003; Lukociene *et al.*, 2010). Then, an unconditional multilevel LC model for $\mathbf{Y}_{ij}$ can be specified as follows:

$$P(\mathbf{Y}_{ij}) = \sum_{m=1}^{M} P(W_j = m) \sum_{t=1}^{T} P(X_{ij} = t|W_j = m) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t). \quad (1)$$

High-level and low-level covariates can be included in order to predict class membership. Let $\mathbf{Z}_{ij} = (1, \mathbf{Z}'_{1j}, \mathbf{Z}'_{2ij})'$ be a vector K covariates, with the sub-vector $\mathbf{Z}'_{1j}$ being defined at the high level, and $\mathbf{Z}'_{2ij}$ being defined at the low level. Let $\mathbf{Z}^*_{1j} = (1, \mathbf{Z}'_{1j})'$. For high-level and low-level latent class membership, respectively, we consider the multinomial logistic models

$$P(W_j = m|\mathbf{Z}^*_{1j}) = \frac{\exp(\alpha'_m \mathbf{Z}^*_{1j})}{1 + \sum_{l=2}^{M} \exp(\alpha'_l \mathbf{Z}^*_{1j})}, \quad (2)$$

$$P(X_{it} = t|W_j = m, \mathbf{Z}_{ij}) = \frac{\exp(\gamma'_{tm} \mathbf{Z}_{ij})}{1 + \sum_{s=2}^{T} \exp(\gamma'_{sm} \mathbf{Z}_{ij})}, \quad (3)$$

In Equation (2), α_m are regression coefficients for $m = 2,\dots,M$, and $m = 1,\dots,M$. In Equation (3), γ_{tm} is a vector of regression coefficients for each $t = 2,\dots,T$. When only the intercept is included in Equation (2), or (3), the corresponding vector of regression coefficients is equal to the log-odds of the class proportions (i.e., $\log(\omega_m/\omega_1)$, or $\log(\pi_{t|m}/\pi_{1|m})$).

In addition, we assume that the observed indicators Y_{ijh} are conditionally independent from the covariates given low-level class membership. Thus, the multilevel LC model for $P(\mathbf{Y}_{ij}|\mathbf{Z}_{ij})$ can be written as:

$$\begin{aligned} &P(\mathbf{Y}_{ij}|\mathbf{Z}_{ij}) = \\ &\sum_{m=1}^{M} P(W_j = m|\mathbf{Z}^*_{1j}) \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t). \end{aligned} \tag{4}$$

The class profiles are defined by the measurement parameters $\phi_{h|t}$, $\pi_{t|m}$, and ω_m. The other parameters of interest are the structural parameters α_m, and γ_{tm}. It is straightforward to reduce the multilevel LC structural model in Equation (4) to the multilevel measurement model, the single-level structural model, or the single-level measurement model.

3 Estimating the multilevel LC model in multilevLCA

The default estimator of Equation (4), in the R package `multilevLCA`, is the two-step approach (Di Mari *et al.*, 2023). We add that future versions of the package will relax the assumptions of Equation (4) to allow for local dependencies. Other options are the two-stage (Bakk *et al.*, 2022) and the simultaneous approaches. A basic function call requires the following arguments:

- `data` The input data (matrix or data frame)
- `Y` The names of the item columns
- `iT` The number of low-level latent classes
- `id_high` The name of the high-level id column
- `iM` The number of high-level latent classes
- `Z` The names of the low-level covariates columns
- `Zh` The names of the high-level covariates columns

Estimation is performed via the function `multiLCA`,

```
out = multiLCA(data,Y,iT,id_high,iM,Z,Zh)
```

The list `out` contains a lot of information about class profiles, structural parameters, and estimation details. A summary of this information can be printed by executing `out` in the prompt. To create a plot of the response probabilities, the user types `plot(out)` in the prompt.

In practice, the number of low-level and high-level classes is unknown to the researchers. Selecting these values is a distinct, yet fundamental task. The `multilevLCA` package includes two state-of-the art model selection strategies, namely sequential model selection (Lukociene et al. 2010) and simultaeneous model selection. Both approaches implement the BIC selection criterion on low and high level, reporting also the AIC and ICL BIC.

To implement the former, `iT` and (or) `iM` is replaced by a range of values. The latter is implemented in the same way, but with the extra argument `sequential` set to `FALSE`. For example, to perform simultaneous model selection over 1-4 low-level classes, and 3-4 high-level classes, we execute the following call:

```
out = multiLCA(data,Y,iT=1:4,id_high,iM=3:4,
sequential=FALSE)
```

The list `out` contains the model estimation results as if the selected specification had been estimated directly. Note that specifying `Z` and `Zh` is redundant; in `multilevLCA`, model selection is always performed without covariates.

The tools for model selection, and visualization are available for any LC model, i.e., the multilevel structural model, multilevel measurement model, single-level structural model, and single-level measurement model.

References

BAKK, Z., & KUHA, J. 2018. Two-step estimation of models between latent classes and external variables. *Psychometrika.*, **83**, 871–892.

BAKK, Z., DI MARI, R., OSER, J., & KUHA, J. 2022. Two-stage multilevel latent class analysis with covariates in the presence of direct effects. *Structural Equation Modeling: A Multidisciplinary Journal.*, **29**(2), 267–277.

DI MARI, R., BAKK, Z., OSER, J., & KUHA, J. 2023. A two-step estimator for multilevel latent class analysis with covariates. *arXiv preprint arXiv:2303.06091.*

LUKOCIENE, O., VARRIALE, R., & VERMUNT, J. K. 2010. The simultaneous decision (s) about the number of lower-and higher-level classes in multilevel latent class analysis. *Sociological Methodology.*, **40**(1), 247–283.

VERMUNT, J. K. 2003. Multilevel latent class models. *Sociological Methodology.*, **33**(1), 213–239.