ARTICLE TEMPLATE

Multilevel latent class analysis: state-of-the-art methodologies and their implementation in the R package multilevLCA

Johan Lyrvall[ab], Roberto Di Mari[a], Zsuzsa Bakk[b], Jennifer Oser[c] and Jouni Kuha[d]

[a]University of Catania, Department of Economics and Business; [b]Leiden University, Department of Methodology and Statistics; [c]Ben-Gurion University of the Negev, Department of Politics and Government; [d]London School of Economics, Department of Statistics

ARTICLE HISTORY

Compiled February 24, 2025

ABSTRACT

Latent class (LC) analysis is a model-based clustering approach for categorical data, with a wide range of applications in the social sciences and beyond. When the data have a hierarchical structure, the multilevel LC model can be used to account for higher-level dependencies between the units by means of a further categorical LC variable at the group level. The research interest of LC analysis typically lies in the relationship between the LCs and external covariates, or predictors. To estimate LC models with covariates, researchers can use the one-step approach, or the generally recommended stepwise estimators, which separate the estimation of the clustering model from the subsequent estimation of the regression model. The package multilevLCA has the most comprehensive set of model specifications and estimation approaches for this family of models in the open-source domain, estimating single- and multilevel LC models, with and without covariates, using the one-step and stepwise approaches.

KEYWORDS

latent class analysis; multilevel; categorical data; stepwise estimation; R

CONTACT Johan Lyrvall. Email: johan.lyrvall@phd.unict.it

1. Introduction

Latent class (LC) analysis (Goodman, 1974a; Lazarsfeld & Henry, 1968; McCutcheon, 1979) is used to classify units into discrete types based on a set of observed categorical variables. The clustering is modeled as an underlying discrete variable with some number of categories or *latent classes*. LC analysis has been applied in diverse research domains in the social sciences and beyond. For example, in political research, Oser (2022) identified repertoires of political participation; in educational research, Hickendorff, van Putten, Verhelst, and Heiser (2010) identified patterns of mental strategies for division problems among elementary school students; in substance use research, Bray, Watson, Salisbury-Afshar, Taylor, and McGuire (2023) identified types of opioid users among patients in the emergency department.

A basic assumption of standard LC analysis is that the units of analysis are independent of each other. This conditional independence assumption is often violated when the data have a multilevel, or hierarchical structure, for example when we observe voters within countries, students within schools, or patients within hospitals. In hierarchical data, units within groups are likely to be systematically more similar than units across groups.

To account for the higher-level dependencies in the hierarchical data, the baseline LC model can be extended by modeling a second categorical LC variable at the higher (group) level. In such a *multilevel LC model*, the distribution of the lower-level classes is allowed to vary between the higher-level classes. This random effect is effectively nonparametric (Aitkin, 1999; Finch & French, 2014; Laird, 1978; Vermunt, 2003), thus avoiding strict distributional assumptions. For instance, in their multilevel LC analysis of financial product ownership across European countries, Bijmolt, Paas, and Vermunt (2004) identified 14 individual-level consumer segments and found that the prevalence of these segments varied between 7 country-level clusters. For example, the consumer segment that was the largest in the cluster of countries in North-Central Europe was rather small in the cluster of countries in North-Western Europe.

In LC analysis, identifying the clustering structure of the data is usually only the first step of the empirical investigation. The research interest usually lies in the relationship between the classes and some covariates, or predictors. In the multilevel LC model, covariates can be included both on the lower level and on the higher level. For instance, in their multilevel LC analysis of adolescent smoking behavior across communities, Henry and Muthén (2010) first identified three individual-level clusters - heavy smokers, moderate smokers, and nonsmokers, and two community-level clusters - low-use communities and high-use communities. Subsequently, they analyzed the regression relationship between smoking behavior and lower-level covariates such as school performance and academic aspirations, and the regression relationship between community type and higher-level covariates such as the proportion of youth living in poverty.

Historically, multilevel LC models were estimated using the traditional *one-step approach*, which involves fitting the full model simultaneously (Lazarsfeld & Henry, 1968; Vermunt, 2003). While the one-step approach has attractive statistical propoperties - when the LC model is correctly specified, it is efficient and asymptotically unbiased - it also comes with serious defects (see e.g. the discussion in Bakk & Kuha, 2018). Whenever covariates are added or removed, the whole model needs to be refitted and the effective definitions of the latent classes can change. This complicates model interpretation and model selection. Furthermore, the one-step approach does not fit with the logic of most applied researchers, who tend to view the regression model as a distinct component that should be estimated only after the clustering model has been built. Therefore, the general recommendation is to use *stepwise estimation approaches* (Asparouhov & Muthén, 2014). These were traditionally only available in single-level LC analysis, but recent methodological advancements have shown how they can be extended to multilevel LC models (Bakk, Di Mari, Oser, & Kuha, 2022; Di Mari, Bakk, Oser, & Kuha, 2023b; Lyrvall, Bakk, Oser, & Di Mari, 2024).

Stepwise approaches avoid the defects of the one-step approach by separating the estimation of the measurement model from the subsequent estimation of the structural model. Among the available stepwise approaches, the *two-step approach* is known to be the most efficient, least biased, most direct, and most flexible option (Bakk & Kuha, 2018; Di Mari et al., 2023b). The *two-stage approach* (Bakk et al., 2022) is slightly less direct but otherwise largely shares the same properties as the two-step approach.

Compared to the one-step approach, the two-step and two-stage approaches come with enhanced algorithmic stability and improved speed of convergence (Di Mari et al., 2023b). Regardless of which estimation approach is applied, the number of classes on the higher level and the lower level is taken as given. Because the complexity of the underlying clustering structure in the data tends to be unknown a priori, identification of the optimal number of classes is typically the first step of applied LC analysis.

In light of these recent methodological contributions, the first aim of this article is to provide a compilation of state-of-the-art methods for multilevel LC analysis with covariates. We describe benchmark model specifications and estimation approaches. In addition, we detail initialization issues and model selection alternatives. Targeting both beginning LC analysts and more advanced LC analysts, we hope to strike a satisfying balance between user-friendly ground-up exposition and technical detail.

A lack of general and easily available software solutions has limited the dissemination of these estimation and model selection approaches in the applied multilevel LC analysis literature. The recently published `R` package `multilevLCA` (Di Mari & Lyrvall, 2024) was developed to fill this gap. The package is available from the Comprehensive `R` Archive Network at `http://cran.r-project.org/package=multilevLCA`. The second aim of this article is to propose the `multilevLCA` package to the open-source statistical software literature. While the functionalities discussed in this paper can be implemented in specialized software like Latent GOLD (Vermunt & Magidson, 2021) and Mplus (Muthén & Muthén, 2017), these software options are commercial and offer fewer automatic implementations of stepwise and sequential routines. In this paper we focus on open-source software. We present the capabilities and syntax of `multilevLCA`. The presentation is organized in the article alongside the corresponding LC analysis methodological exposition, to closely connect software implementation with theory. The software contribution has been written in such a way that we hope that this article can serve as a stand-alone reference for application of `multilevLCA`.

The `multilevLCA` package is both the first freeware-software to implement stepwise estimation of multilevel LC models with covariates and the first to estimate multilevel LC models with both dichotomous and polytomous indicators. `multilevLCA` has the most comprehensive set of model specifications and estimation approaches;

estimating single- and multilevel LC models, with and without covariates, using the one-step, two-stage, and two-step approaches. The semi-automatic implementation of model selection in the package is more straightforward and efficient compared to when each model of interest needs to be fitted separately, which is the case when using other freeware-software for LC analysis.

The only existing freeware-software for multilevel LC analysis with covariates is the `R` package `glca` (Kim, Jeon, Chang, & Chung, 2022), but it is limited to the one-step approach, with no implementation of stepwise approaches. Moreover, it does not have the capacity to model polytomous indicators, which are typically used in applied research. As such, the scope of the use of `glca` is somewhat limited compared to `multilevLCA`. The comprehensive functionalities of `multilevLCA` also extend the freeware-software state-of-the-art in single-level LC analysis with covariates. Existing packages for it include the `R` packages `poLCA` (Linzer & Lewis, 2011) and `MultiLCIRT` (Bartolucci, Bacci, & Gnaldi, 2014), but they estimate only single-level models using the one-step approach. The more complete alternative for single-level LC modeling is the `Python` package `StepMix` (Morin et al., 2023), with `R` interface `stepmixr` (Lacourse et al., 2024), which also implements stepwise estimation. However, unlike `multilevLCA`, `StepMix` does not compute maximum-likelihood standard errors of the regression parameters for the covariates, which is the statistical benchmark, instead applying the bootstrap method.

This article offers a comprehensive review of the key aspects of multilevel LC analysis with covariates, and a hands-on guide to the implementation of these techniques using the `multilevLCA` package. In the next section, we present the multilevel LC model and the `multilevLCA` syntax. Then, we describe possible estimation strategies for the model and their implementation in `multilevLCA`, including strategies for class selection and initialization and a benchmark simulation study of performance and estimation times. Next, we illustrate key features of `multilevLCA` by means of an empirical example, and conclude with a summary.

2. Model specifications

2.1. *Theoretical framework*

Let Y_{ih} denote the response of unit $i = 1, \ldots, N$ on the categorical item $h = 1, \ldots, H$, with possible values $Y_{ih} = 1, \ldots, R_h$, and let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iH})'$ denote the full response vector for the same unit. The elements of the vector are treated as observed indicators of the categorical latent variable X_i, with possible values $\{1, \ldots, T\}$. The single-level latent class (LC) model defines the unconditional probability of observing a particular response pattern $\mathbf{Y}_i$ as a mixture of T class-specific probabilities, that is,

$$P(\mathbf{Y}_i) = \sum_{t=1}^{T} P(X_i = t) P(\mathbf{Y}_i | X_i = t). \tag{1}$$

Here, the mixture weight $P(X_i = t)$ describes the unconditional probability that unit i belongs to class t, while the mixture component $P(\mathbf{Y}_i | X_i = t)$ describes the conditional probability of a particular response pattern $\mathbf{Y}_i$ given class t. The responses of the different indicators are assumed to be conditionally independent given class membership (the *local independence assumption*), leading to

$$P(\mathbf{Y}_i) = \sum_{t=1}^{T} P(X_i = t) \prod_{h=1}^{H} P(\mathbf{Y}_{ih} | X_i = t) = \sum_{t=1}^{T} P(X_i = t) \prod_{h=1}^{H} \prod_{r=1}^{R_h} \phi_{rh|t}^{I(Y_{ih}=r)}, \tag{2}$$

where the quantity $\phi_{rh|t}$ is the probability of giving response r on item h given class t, and $I(Y_{ih} = r)$ is equal to 1 if unit i gives response r on item h, and 0 otherwise. For ease of notation, we will use $P(\mathbf{Y}_{ih} | X_i = t)$ to denote $\prod_{r=1}^{R_h} \phi_{rh|t}^{I(Y_{ih}=r)}$ in what follows.

Figure 1 graphically illustrates the single-level LC model defined in (2). The arrows describe a causal relationship from the LC variable X_i to the indicators Y_{ih}. There are no arrows between the indicators, reflecting the local independence assumption.

In a multilevel LC model we take the lower-level units $i = 1, \ldots, n_j$ (e.g. individual respondents) to be nested within higher-level units $j = 1, \ldots, J$ (groups, e.g. countries). Let W_j be a higher-level categorical latent variable with possible categories $m = 1, \ldots, M$, and probabilities $P(W_j = m) = \omega_m > 0$, and let X_{ij} now be a lower-level categorical latent variable that is defined conditional on the values of W_j, with possible values $t = 1, \ldots, T$ and conditional probabilities $P(X_{ij} = t|W_j = m) = \pi_{t|m} > 0$. We collect all ω_m and $\pi_{t|m}$ respectively in the M-vector ω, and the $M \times T$ matrix $\mathbf{\Pi}$. The multilevel (random-effect) LC model for $\mathbf{Y}_{ij}$ can be specified as

$$P(\mathbf{Y}_i) = \sum_{m=1}^{M} P(W_j = m) \sum_{t=1}^{T} P(X_{ij} = t|W_j = m) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t), \tag{3}$$

where we assume that the conditional response probabilities of items Y_{ijh} depend on higher-level class membership only through X_{ij}. The model specified in (3) is similar to the multilevel item response model (Gnaldi, Bacci, & Bartolucci, 2016), but with categorical latent variables on both levels.

While the assumption of conditional independence between Y_{ijh} and W_j given X_{ij} is not necessary for model identification, it is a standard assumption in multilevel LC analysis for enhancing model interpretation (Lukočienė, Varriale, & Vermunt, 2010; Vermunt, 2003). The higher-level LC variable is typically included when it cannot be assumed that the distribution of the lower-level LCs be invariant across higher-level units j (this point is exemplified in a substantive analysis in Section 5).

In Figure 2, we graphically illustrate the multilevel LC model defined in (3). The absence of arrows from the higher-level LC variable W_j to the indicators Y_{ih} reflect their conditional independence given the lower-level LC variable X_{ij}.

Higher-level and lower-level covariates can be included to predict class membership. Let $\mathbf{Z}_{ij} = (1, \mathbf{Z}'_{1j}, \mathbf{Z}'_{2ij})'$ be a vector of K covariates, which can be defined on the higher level ($\mathbf{Z}'_{1j}$) and the lower level ($\mathbf{Z}'_{2ij}$). On the higher level, we consider the following multinomial logistic model

$$P(W_j = m|\mathbf{Z}_j^H) = \frac{\exp(\alpha'_m \mathbf{Z}_j^H)}{1 + \sum_{l=2}^{M} \exp(\alpha'_l \mathbf{Z}_j^H)}, \tag{4}$$

where $\mathbf{Z}_j^H = (1, \mathbf{Z}'_{1j})'$, and α_m are regression coefficients for $m = 2, \ldots, M$. When only the intercept term is included, then α_m is equal to the log-odds $\log(\omega_m/\omega_1)$.

On the lower level, class membership probabilities can be parameterized in the following analogous way,

$$P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}) = \frac{\exp(\gamma'_{tm} \mathbf{Z}_{ij})}{1 + \sum_{s=2}^{T} \exp(\gamma'_{sm} \mathbf{Z}_{ij})}, \tag{5}$$

where γ_{tm} is a vector of regression coefficients for each $t = 2, \ldots, T$, and $m = 1, \ldots, M$. When only the intercept term is included, so that $\mathbf{Z}_{ij} = 1$, then γ_{tm} is equal to the log-odds $\log(\pi_{t|m}/\pi_{1|m})$. As can be seen, this parametrization allows the effects of $\mathbf{Z}_{ij}$ on X_{ij} to vary across different m. The methodological exposition throughout this article holds also for the equivalent constrained parametrization in which the slopes are held fixed across different m and only the intercepts are allowed to vary (Di Mari et al., 2023b; Vermunt, 2005). For generality of exposition, we focus on the unconstrained parametrization without fixed slopes in (5).

We further assume that the indicators Y_{ijh} are conditionally independent from the covariates given lower-level class membership. With these assumptions, the multilevel LC model for $P(\mathbf{Y}_{ij}|\mathbf{Z}_{ij})$ can be written as

$$P(\mathbf{Y}_{ij}|\mathbf{Z}_{ij}) = \sum_{m=1}^{M} P(W_j = m|\mathbf{Z}_j^H) \left[\sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t) \right]. \quad (6)$$

The conditional response probabilities $P(Y_{ijh}|X_{ij} = t)$ define the LC *measurement model*, while the conditional class membership probabilities $P(W_j = m|\mathbf{Z}_j^H)$ and $P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij})$ define the LC *structural models*.

The multilevel LC model with covariates defined in (6) is graphically illustrated by means of a path diagram in Figure 3. The assumption of conditional independence between the indicators and the covariates given lower-level class membership is reflected in the absence of arrows from $\mathbf{Z}_{ij}$ and $\mathbf{Z}_j^H$ to the Y_{ijh}.

As noted above, multilevel LCA is typically applied when the distribution of the lower-level LCs X_{ij} cannot be assumed to be invariant across higher-level units j. The strategy of capturing this invariance by means of a higher-level clustering structure is known as the random-effect approach. This is the approach on which we focus. For completeness, we now briefly describe the alternative fixed-effect approach. In this approach the distribution of X_{ij} is allowed to vary across each of the J higher-level units. This is achieved by treating higher-level unit membership as a (categorical) covariate in a single-level LC model. Let $\mathrm{I}_i^H = (\mathrm{I}_i(1), \ldots, \mathrm{I}_i(J))'$ be a collection of vectors $\mathrm{I}_i(j)$ which are equal to unity if i belongs to j and zero otherwise. A fixed-effect multilevel LC model with covariates can be specified as

$$P(\mathbf{Y}_i|\mathbf{Z}_{ij}) = \sum_{t=1}^{T} P(X_i = t|\mathrm{I}_i^H, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t), \tag{7}$$

where, like in the random-effect specification, $P(X_i = t|\mathrm{I}_i^H, \mathbf{Z}_{ij})$ can be parameterized by means of multinomial logistic equations.

2.2. Implementation in multilevLCA

The syntax used in the `R` package `multilevLCA` is aligned with the notation used in (6). The package's multilevel modeling focuses on standard specifications with conditional independence between the items Y_{ijh} and the higher-level LC variable W_j are given the lower-level LC variable X_{ij}. LC models are specified using the function `multiLCA()`, based on some combination of statements about the variables to be included in the model. This is structured by means of the following arguments:

- `data`: Matrix or data frame containing the observed data
- `Y`: Names of `data` columns with indicators
- `iT`: Number of lower-level classes
- `id_high`: Name of `data` column with higher-level id
- `iM`: Number of higher-level classes
- `Z`: Names of `data` columns with covariates in the model for the lower-level classes
- `Zh`: Names of `data` columns with covariates in the model for the higher-level classes

The multilevel LC model with covariates on the higher level and the lower level includes all the variables corresponding to these statements - the indicators $\mathbf{Y}$, specified by `Y`; the lower-level LC variable $X = 1, \ldots, T$, specified by `iT`; the higher-level LC variable $W = 1, \ldots, M$, specified by `id_high` and `iM`; the covariates in the model for the lower-level classes $\mathbf{Z}$, specified by `Z`; and the covariates in the model for the higher-level classes $\mathbf{Z}^H$, specified by `Zh`. The syntax for specifying this model is[1]

[1] `multilevLCA` also estimates multilevel LC models in which the slopes for the lower-level structural model are held fixed across the higher-level classes. This constraint is managed by means of the argument `fixedslopes` in the `multiLCA()` function. The specification `fixedslopes = TRUE` fixes the slopes in the lower-level structural model. The default specification `fixedslopes = FALSE` estimates models without these constraints, which is the focus of this article.

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

Single-level LC models with covariates and multilevel fixed-effect LC models can be estimated by omitting to specify `id_high`, `iM`, and `Zh` (which default to `NULL`). More specifically, multilevel fixed-effect LC models can be estimated by specifying `Z` as the column name which in random-effect modeling is specified for `id_high`. We illustrate the `multiLCA()` syntax in greater detail by means of real-data examples in Section 5.

The next section describes the currently existing approaches for estimating (6).

3. Methodology

3.1. Theoretical framework

Let $\mathbf{Y}_j = (\mathbf{Y}_{1j}, \ldots, \mathbf{Y}_{n_j j})'$ denote the full set of item responses for all lower-level units belonging to higher-level unit j[2]. Let $\theta = (\theta_1', \theta_2')'$ denote the full set of model parameters in (6), where θ_1' contains the measurement parameters $\phi_{rh|t}$, and θ_2' contains the structural parameters α_m and γ_{tm}.

Figure 4 graphically illustrates the measurement parameters θ_1' by red arrows, and the structural parameters θ_2' by blue arrows.

Maximum-likelihood estimates $\widetilde{\theta}$ can be obtained by maximizing the observed-data log-likelihood function

$$\ell(\theta) = \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m|\mathbf{Z}_j^H) \prod_{i=1}^{n_j} \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t) \right]. \tag{8}$$

This is the classical *one-step* approach (Lazarsfeld & Henry, 1968; Vermunt, 2003). It is efficient and asymptotically unbiased when the LC model is correctly

[2] By default, `multilevLCA` discards any rows with missing values on the items, or incomplete item-response patterns, before estimation. An alternative strategy involves including incomplete item-response patterns by means of full-information maximum-likelihood (FIML) estimation, only discarding any rows with missing values on all the items. The choice between these strategies is managed by means of the argument `incomplete` in the function `multiLCA()`. The default specification `incomplete = FALSE` implements row-wise deletion of incomplete item-response patterns. The alternative specification `incomplete = TRUE` implements the FIML strategy, including incomplete item-response patterns (except fully missing item-response patterns). Regardless of strategy for handling missing values, if covariates are included in the model, rows with missing values in the covariates are removed only in the estimation of the structural part of the LC model, i.e. (see below) step 2 in the two-step estimator, stage 2 in the two-stage estimator, or the single step in the one-step estimator.

specified. However, simultaneous estimation of the measurement model and structural models has serious disadvantages when the correct specification is not known a priori (see e.g. the discussion in Bakk & Kuha, 2018). Whenever the structural model is changed - for example adding or removing covariates - the measurement model will be affected, which distorts the class definitions. In practice, this problem can occur to an extent that makes comparisons of estimated models meaningless. As such, the one-step approach complicates model interpretation and model selection. Moreover, simultaneous estimation of complex models involves demanding computations, which renders the one-step approach the more time consuming modeling option for multilevel LC analysis with covariates (Di Mari et al., 2023b).

Stepwise methods overcome the drawbacks of the one-step approach by separating the estimation of the measurement model and structural model. The first stepwise method that was proposed in multilevel LC modeling with covariates is the *two-stage* approach (Bakk et al., 2022; Di Mari, Bakk, Oser, & Kuha, 2023a). Its first stage involves estimating the measurement parameters. This is further broken down into three sub-steps. In the first sub-step, the single-level LC model without covariates is estimated, ignoring the hierarchical structure of the data, by maximizing the log-likelihood function

$$\ell_{\text{stage1.1}}(\theta_1) = \sum_{i=1}^{N} \log \left[\sum_{t=1}^{T} P(X_{ij} = t) \prod_{h=1}^{H} P(Y_{ijh} | X_{ij} = t) \right], \tag{9}$$

where $N = \sum_{j=1}^{H} n_j$, to obtain measurement estimates $\widetilde{\theta}_1$. In the second sub-step, the multilevel LC model without covariates is estimated, keeping the measurement parameters θ_1 fixed at their values from sub-step 1, by maximizing the log-likelihood function

$$\begin{aligned} &\ell_{\text{stage1.2}}(\theta_2 | \theta_1 = \widetilde{\theta}_1) = \\ &\sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m) \prod_{i=1}^{n_j} \sum_{t=1}^{T} P(X_{ij} = t | W_j = m) \prod_{h=1}^{H} P(Y_{ijh} | X_{ij} = t, \theta_1 = \widetilde{\theta}_1) \right], \end{aligned} \tag{10}$$

where the structural parameters θ_2 now contain only the intercept terms, to obtain structural estimates $\widetilde{\theta}_2$. In the third sub-step, to stabilize the measurement estimates, the multilevel LC model is estimated again, this time keeping the structural parameters θ_2 fixed at their values from sub-step 2, by maximizing the log-likelihood function

$$\ell_{\text{stage1.3}}(\theta_1|\theta_2 = \widetilde{\theta}_2) = \\ \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m|\theta_2 = \widetilde{\theta}_2) \prod_{i=1}^{n_j} \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \theta_2 = \widetilde{\theta}_2) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t) \right]. \tag{11}$$

Stage 2 of the two-stage approach involves adding the covariates to the multilevel LC model, and estimating the intercept and slope terms θ_2, keeping the measurement parameters fixed at their stage-1 values, by maximizing the log-likelihood function

$$\ell_{\text{stage2}}(\theta_2|\theta_1 = \widetilde{\theta}_1) = \\ \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m|\mathbf{Z}_j^H) \prod_{i=1}^{n_j} \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t, \theta_1 = \widetilde{\theta}_1) \right]. \tag{12}$$

The two-stage approach simplifies model interpretation and improves computation time compared to the one-step method, while demonstrating very similar properties when the model assumptions hold (Bakk et al., 2022). However, a difficulty of this approach is estimating asymptotic standard errors of the structural parameters. In the second stage, conditioning on the measurement parameters as if they were known, rather than estimated with sampling error, yields underestimation of the standard errors. Conditioning on this first-stage variability is complicated due to the multiple sub-steps of the first stage.

To address this difficulty, the more straightforward *two-step* approach (Di Mari et al., 2023b) was developed. It simplifies the estimation of the measurement model by means of a single first step. This involves maximizing the log-likelihood function

$$\ell_{\text{step1}}(\theta_1) = \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m) \prod_{i=1}^{n_j} \sum_{t=1}^{T} P(X_{ij} = t|W_j = m) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t) \right], \tag{13}$$

to obtain measurement estimates $\widetilde{\theta}_1$. The second step involves estimating the structural parameters, keeping the measurement parameters fixed at their step-1 values, by maximizing the log-likelihood function for the second step as

$$\ell_{\text{step2}}(\theta_2|\theta_1 = \widetilde{\theta}_1) = \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m|\mathbf{Z}_j^H) \prod_{i=1}^{n_j} \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t, \theta_1 = \widetilde{\theta}_1) \right]. \tag{14}$$

The two-step approach retains the attractive properties of the two-stage method, with the additional benefits of easy-to-derive asymptotic standard errors, and even greater computational efficiency (Di Mari et al., 2023b).

The estimation approaches that were presented in this section take the number of classes on the higher level, M, and the lower level, T, as given. Selecting these values is a distinct but equally fundamental task. In Section 4, two model selection approaches are described.

3.2. Implementation in multilevLCA

Because of its attractive properties, the two-step approach is the default estimator in the R package multilevLCA. Users can also choose to estimate LC models using the one-step and two-stage approaches. This makes multilevLCA the first R package, and the first freeware software in any programming language, to implement stepwise estimation of multilevel LC models with covariates.

Estimation approaches are managed using the argument fixedpars in the function multiLCA(). One-step, two-stage, and two-step estimation of the multilevel LC model with covariates on the higher level and the lower level are implemented by means of the syntax

```
# One-step estimation:
multiLCA(data, Y, iT, id_high, iM, Z, Zh, fixedpars = 0)

# Two-stage estimation:
multiLCA(data, Y, iT, id_high, iM, Z, Zh, fixedpars = 2)
```

```
# Two-step estimation (the default):
multiLCA(data, Y, iT, id_high, iM, Z, Zh, fixedpars = 1)

# Equivalent two-step estimation:
multiLCA(data, Y, iT, id_high, iM, Z, Zh)
```

The estimators are labeled by the total number of fixed parameters; in one-step estimation, no parameters are kept fixed (`fixedpars = 0`); in two-stage estimation, the fixed parameters are obtained from two consecutive sub-steps (`fixedpars = 2`); in two-step estimation, the fixed parameters are obtained from a single step (`fixedpars = 1`).

Regardless of which estimator is used, estimation is performed using the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). When covariates are included, the M step of the EM algorithm uses a Newton-Raphson (NR) algorithm. For computational efficiency, the EM and NR algorithms are implemented by integration of `C++` code (Eddelbuettel & François, 2011; Eddelbuettel & Sanderson, 2014).

In stepwise estimation, the starting values for the EM algorithm are particularly important because subsequent steps are conditional on estimates from previous steps. `multilevLCA` implements an initialization strategy based on Di Mari et al. (2023b).

For the measurement model, the initialization strategy involves the following hierarchical procedure:

(1) Fit a single-level LC model with T classes to the pooled data $(\mathbf{Y}_{11}, \dots, \mathbf{Y}_{n_J J})$, ignoring the multilevel structure. To initialize the class proportions $P(X_i = t)$, perform a k-modes clustering on the dummy-coded data, with $k = T$. Use the relative sizes of the resulting clusters for the initialization. From the single-level class solution, retain the estimates for the conditional response probabilities $P(Y_{ijh}|X_{ij} = t)$, and the modal posterior class assignments[3] $\widetilde{X}_{ij}$. The estimates

[3] The modal posterior class assignment is the class for which the posterior class membership probability $P(X_{ij} = t|\mathbf{Y}_{ij})$, which describes the probability of belonging to class t given the observed response pattern $\mathbf{Y}_{ij}$, is the greatest. Using the Bayes rule (Goodman, 1974a, 1974b; Hagenaars, 1992; MacLahlan & Peel, 2000),

for $P(Y_{ijh}|X_{ij} = t)$ are passed to the EM-algorithm as starting values.

For computational speed and stability, the class proportions $P(X_i = t)$ can be initialized by the following alternative strategy. First, perform a principal component analysis on the dummy-coded data. Retain the first principal components that together explain at least 85% of the total variance, or retain the first half of all principal components, if this is a greater number. Second, perform a k-means clustering on the reduced data, with $k = T$. Use the relative sizes of the resulting clusters for the initialization.

(2) Compute the relative sizes of $\widetilde{X}_{ij}$ within each higher-level unit j. On the resulting $J \times T$ table, perform a k-means clustering, with $k = M$. Let $\widetilde{W}_j$ be the resulting clusters. The relative sizes of $\widetilde{W}_j$ are passed to the EM-algorithm as starting values for the higher-level class proportions $P(W_j = m)$.

In the function `multiLCA()`, the choice between the k-modes strategy and the k-means on principal components strategy is managed using the logical argument `kmea`. The default argument is `kmea = TRUE`, which indicates the k-means on principal components strategy. The user also has the option to specify custom starting values. This can be done by specifying, in the `multiLCA()` call, the argument `startval` (which defaults to `NULL`) as the name of the `data` column containing starting values for the lower-level class membership of each lower-level unit. The three initialization strategies are implemented by means of the syntax

```
# k-means on principal components initialization:
multiLCA(data, Y, iT, id_high, iM, Z, Zh)

# k-modes initialization:
multiLCA(data, Y, iT, id_high, iM, Z, Zh, kmea = FALSE)

# user-specified starting values:
multiLCA(data, Y, iT, id_high, iM, Z, Zh, startval)
```

this quantity can be computed as

$$P(X_{ij} = t|\mathbf{Y}_{ij}) = \frac{P(X_{ij} = t)P(\mathbf{Y}_{ij}|X_{ij} = t)}{P(\mathbf{Y}_{ij})} \tag{15}$$

For the structural model, the initialization strategy is used to handle label switching on the higher level. Keeping the conditional response probabilities fixed cannot prevent that higher-level class labels can be switched, as there are still $M!$ equivalent permutations of them. This is handled by initializing the intercept in α_m and the intercept in γ_{tm} at the measurement model estimates for $\log(\omega_m/\omega_1)$ and $\log(\pi_{t|m}/\pi_{1|m})$, respectively, while initializing the slope parameters in α_m and the slope parameters in γ_{tm} at zero.

4. Model selection

4.1. Theoretical framework

The general recommendation in LC analysis with covariates is to perform model selection on the model without covariates, defined in (3), and then estimate the full model given this value (Masyn, 2017). In multilevel LC analysis, different approaches can be used to identify the locally optimal number of higher-level classes, M, and lower-level classes, T, among a set of specifications. Using the straightforward *simultaneous* approach, all crossed combinations of the values of interest for M and T are estimated.

Using the generally recommended *sequential* approach (Lukočienė et al., 2010), the optimal values for M and T are selected in a stepwise procedure. First, single-level LC models, defined in (2), are estimated to select the optimal number of lower-level classes, T^*. Second, multilevel LC models are estimated, keeping the number of lower-level classes fixed at the step-1 value T^*, to select the optimal number of higher-level classes, M^*. Third, multilevel LC models are estimated again, this time keeping the number of higher-level classes fixed at the step-2 value M^*, to re-select the number of lower-level classes.

The optimal model can be selected based on standard information criteria, such as the Bayesian information criterion (BIC) or the Akaike information criterion (AIC). BIC can be evaluated on the higher level and the lower level separately (e.g. Lukočienė et al., 2010). Another information criterion is the BIC-type approximation of the integrated complete likelihood (ICL-BIC; e.g. Morgan, 2015), which can be defined on the higher level and the lower level separately, wherein a penalty for class separation is added to the BIC.

4.2. Implementation in multilevLCA

The R package multilevLCA implements semi-automatic[4] model selection, for model specifications without covariates, using the simultaneous and sequential approaches. This is done using the same syntax as for standard model estimation, in the function multiLCA(), with the number of classes on the higher level and the lower level specified as a range of consecutive integers, and model selection approaches managed using the argument sequential. The argument sequential = TRUE indicates sequential model selection, and the argument sequential = FALSE indicates simultaneous model selection. The sequential approach is the default model selection approach.

Consider, for example, the multilevel LC model with an unknown number of lower-level classes, which is taken to be within the range 1-5, and an unknown number of higher-level classes, taken to be within the range 1-3. The syntax for implementing simultaneous and sequential model selection is

```
# Sequential model selection:
multiLCA(data, Y, iT = 1:5, id_high, iM = 1:3)

# Simultaneous model selection:
multiLCA(data, Y, iT = 1:5, id_high, iM = 1:3, sequential = FALSE)
```

Regardless of which model selection approach is implemented, the function call returns the optimal model, and information criteria for all the estimated models. The information criteria include higher- and lower-level BIC, AIC, and higher- and lower-level ICL-BIC. The optimal model is selected based on BIC; with simultaneous model selection, the lower-level BIC, and with sequential model selection, the lower-level BIC for step 1, the higher-level BIC for step 2, and again the lower-level BIC for step 3. This is illustrated by means of a real-data example in Section 5.

[4] Semi-automatic in the sense that the package implements model selection only over the range of specifications which is specified by the user.

4.3. Performance and estimation time of model selection

To examine the performance and estimation time for the semi-automatic implementation of the simultaneous and sequential model selection approaches in the `multilevLCA` package, we conduct a simulation study. The population model has twelve binary items Y_{ijh}. For all the lower-level classes, the probability of the most likely response is set to 0.8. We vary the number T of lower-level classes X_{ij} from three to five, and the number M of higher-level classes W_j from two to three. The sample sizes on the lower level and the higher level are 500 and 30, respectively.

In all the simulation conditions, the first lower-level class $X_{ij} = 1$ has high probabilities (0.8) of endorsement for all the items and the last lower-level class $X_{ij} = T$ low probabilities (0.2) of endorsement for all the items.

When the number of lower-level classes is $T = 3$, the second class has high probabilities for the first six items $Y_{ij1}, \ldots, Y_{ij6}$ and low probabilities for the last six items $Y_{ij7}, \ldots, Y_{ij12}$. The lower-level class proportions within the first and second higher-level classes are:

- $P(Xij = 1|Wj = 1) = P(Xij = 3|Wj = 2) = 0.19$
- $P(Xij = 2|Wj = 1) = P(Xij = 2|Wj = 2) = 0.31$
- $P(Xij = 3|Wj = 1) = P(Xij = 1|Wj = 2) = 0.51$
- $P(Xij = t|Wj = 3) = 1/T = 0.33$ for all t, when a third higher-level class is modeled

When the number of lower-level classes is $T = 4$, the second and third classes have high probabilities only for the first and last six items, respectively. The lower-level class proportions within the first and second higher-level classes are:

- $P(Xij = 1|Wj = 1) = P(Xij = 4|Wj = 2) = 0.10$
- $P(Xij = 2|Wj = 1) = P(Xij = 3|Wj = 2) = 0.17$
- $P(Xij = 3|Wj = 1) = P(Xij = 2|Wj = 2) = 0.28$
- $P(Xij = 4|Wj = 1) = P(Xij = 1|Wj = 2) = 0.46$
- $P(Xij = t|Wj = 3) = 1/T = 0.25$ for all t, when a third higher-level class is modeled

When the number of lower-level classes is $T = 5$, the second, third, and fourth classes have high probabilities only on the first, mid, and last four items, respectively. In this context, the lower-level class proportions within the first and second higher-level classes are:

- $P(Xij = 1|Wj = 1) = P(Xij = 5|Wj = 2) = 0.06$
- $P(Xij = 2|Wj = 1) = P(Xij = 4|Wj = 2) = 0.10$
- $P(Xij = 3|Wj = 1) = P(Xij = 3|Wj = 2) = 0.16$
- $P(Xij = 4|Wj = 1) = P(Xij = 2|Wj = 2) = 0.26$
- $P(Xij = 5|Wj = 1) = P(Xij = 1|Wj = 2) = 0.43$
- $P(Xij = t|Wj = 3) = 1/T = 0.20$ for all t, when a third higher-level class is modeled

In all the simulation conditions, model selection is performed over a range of values for T and M. The smallest value for these ranges is one, while we vary the highest values by means of the excess above the true number of classes, considering excesses equal to one or three. For example, with a lower-level excess of three for $T = 3$ and a higher-level excess of one for $T = 2$, we perform model selection over 1-6 lower-level classes and 1-3 higher-level classes.

Table 1 summarizes the resulting 24 fully crossed simulation conditions. For each of them, we generate 50 random samples.

The sequential model selection approach correctly identified the true number of lower-level and higher-level classes for all the simulation conditions and random samples. The simultaneous approach performed equally well for the lower level, while, for the higher level, it yielded a 50/50 success rate across the random samples for 16 of the 24 simulation conditions. For the other simulation conditions, it yielded a success rate of 47-49/50 across the random samples[5].

Figure 5 reports the average estimation time for the sequential and simultaneous model selection approaches across the 24 simulation conditions and 50 replications. As expected, the time cost for both approaches tends to be greater when the range of values for the number of classes is larger on the lower level or the higher level.

[5] The success rate was 47/50 for simulation condition 21; 48/50 for simulation condition 5; 49/50 for simulation conditions 1, 4, 7, 8, 16 and 22.

It can clearly be seen that the sequential approach is consistently faster than the simultaneous approach. The time cost for the sequential approach is less sensitive to the range of values for T or M, so that the time cost difference increases when these ranges increase.

5. Empirical example: citizenship norms

To illustrate the functionalities of the R package multilevLCA, we analyze data from the International Civic and Citizenship Education Study 2016 (Schulz et al., 2018) of the International Association for the Evaluation of Educational Achievement (IEA), which have been used to advance political research on citizenship norms (Hooghe & Oser, 2015; Hooghe, Oser, & Marien, 2016; Oser & Hooghe, 2013; Oser, Hooghe, Bakk, & Di Mari, 2023). For details on data cleaning and recoding, see Oser, Di Mari, and Bakk (2023). These data are contained in multilevLCA as the data frame dataIEA. We can load the package and the data by executing

```
library(multilevLCA)
data("dataIEA")
```

We interpret the substantive results in relation to the LC analysis of the same data by Oser, Hooghe, et al. (2023). Prior to their investigation, the political literature on citizenship norms had been focusing on societal-level analyses. The LC analysis informs the literature by taking a person-centered approach and investigating how individuals in different sub-groups of the population adhere to distinct citizenship norms.

As part of a comprehensive evaluation of education systems, the IEA conducted surveys in school classes of 14–year olds to investigate civic education. The use of responses from adolescents to analyze citizenship norms is justified by political research showing that stabilization of individual political attitudes and behaviors occurs rather early in the life cycle (Prior, 2010; Van Deth, Abendschön, & Vollmar, 2011). The survey lists a variety of activities for respondents to rate in terms of importance in order to be considered a good adult citizen. These can be categorized as self-expressive, engaged normative ideals: promoting human rights (*rights*), participating in local activities (*local*), supporting activities to protect the environment (*envir*), participating in peaceful protest (*protest*), and engaging in political conversations (*discuss*); and traditional, duty-based normative ideals: obeying the law (*obey*), working hard (*work*), voting (*vote*), learning about the country's history (*history*), showing respect for government representatives (*respect*), following political news (*news*), and joining a political party (*party*). The answer options "very important" and "quite important" are here coded as 1, while the answer options "not very important" and "not important at all" are coded as 0.

Similar to Oser, Hooghe, et al. (2023), in our LC analysis, we treat the items as observed indicators $\mathbf{Y}_{ij}$ of an underlying structure of citizenship norms X_{ij}, where i denotes a particular student, and j denotes the country in which the school is located. The data contain 90,221 students from 22 countries.

To illustrate the observed response patterns, we print the first three rows below (the observed responses to the questionnaire items are located in columns 5-16).

```
head(dataIEA[,5:16], 3)
obey rights local work envir vote history respect news protest discuss party
   1      1     1    1     1    1       1       1    1       1       0     0
   1      1     1    1     1    1       1       1    1       1       0     0
   1      1     1    1     1    1       1       1    1       0       0     0
```

We begin the illustrative analysis with the five-class single-level LC model without covariates, which was defined in (2), replicating the analysis of Oser, Hooghe, et al. (2023), by executing

```
set.seed(2023)
multiLCA(data = dataIEA, Y    = colnames(dataIEA)[5:16], iT   = 5)
```

```
CLASS PROPORTIONS:

P(C1) 0.3956
P(C2) 0.3509
P(C3) 0.1111
P(C4) 0.1147
```

```
P(C5) 0.0277

RESPONSE PROBABILITIES:

                 C1     C2     C3     C4     C5
P(obey|C)    0.9801 0.9742 0.6335 0.9594 0.3408
P(rights|C)  0.9802 0.9601 0.7386 0.2999 0.0485
P(local|C)   0.9678 0.9079 0.7267 0.3517 0.0527
P(work|C)    0.9364 0.8894 0.5991 0.8532 0.3150
P(envir|C)   0.9800 0.9767 0.7135 0.4771 0.1241
P(vote|C)    0.9727 0.7893 0.6644 0.7476 0.1605
P(history|C) 0.9399 0.8361 0.5992 0.7031 0.1744
P(respect|C) 0.9384 0.8569 0.5357 0.8351 0.1465
P(news|C)    0.9621 0.7171 0.5150 0.7015 0.0783
P(protest|C) 0.8713 0.5701 0.6315 0.1672 0.0516
P(discuss|C) 0.8400 0.1782 0.3945 0.1797 0.0122
P(party|C)   0.6071 0.1439 0.3071 0.1519 0.0177

---------------------------

MODEL AND CLASSIFICATION STATISTICS:

ClassErr      0.1966
EntR-sqr      0.6181
```

At the bottom of the partial `multiLCA()` output above, we can see class separation statistics for the class solution, namely, the average proportion of classification error (`ClassErr`; see Vermunt & Magidson, 2021), and the entropy-based R^2 (`EntR-sqr`; see Magidson, 1981). To interpret these statistics, consider the task of predicting class membership based on the model parameters (using the modal assignment rule). Based on the average proportion of classification error, we can expect 20% of the respondents to be assigned to the wrong class. Based on the entropy-based R^2, we can expect a 62% improvement of the class prediction when using the response probabilities and class proportions, compared to the prediction using only the class proportions.

The results show that estimated 11.1% and 11.5% of the respondents belong to class 3 and class 4, respectively. Class 3 is corresponding to the “Engaged” class and class 4 to the “Duty” class in Oser, Hooghe, et al. (2023). The youth belonging to class 3 have consistently high conditional probabilities to score 1 (i.e., indicate high importance) on the self-expressive and engaged notions of good citizenship, and consider the traditional and duty-based items to be less important. Class 4 places high importance on the traditional items, except for joining a political party, while placing relatively low importance on the self-expressive items. From a theoretical perspective, the capacity of LCA to identify these two distinctive citizenship norms allows us to address longstanding questions in the literature regarding the socio-demographic characteristics of people who adhere to these different norms.

We can automatically plot the estimated response probabilities by executing

```
plot(out)
```

The resulting plot is shown in Figure 6.

To investigate whether the proportion of classification error differs between the classes, we request extensive `multiLCA()` output using the specification `extout = TRUE`. The quantities of interest are contained in the element `mClassErrProb`, which we display below, rounded to two decimal points. The rows of the matrix correspond to true class membership, while columns correspond to predicted class membership. As shown, the expected proportion of correct classification for class 3 (Engaged) and class 4 (Duty) are 73% and 76%, respectively. The youth belonging to class 3 have 9% probability of being assigned to class 4, and those belonging to class 4 a 10% probability of being assigned to class 3.

```
out = multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 5, extout = TRUE)
round(out$mClassErrProb, 2)
        C1_pred C2_pred C3_pred C4_pred C5_pred
C1_true    0.87    0.11    0.02    0.01    0.00
C2_true    0.13    0.76    0.07    0.04    0.00
C3_true    0.04    0.13    0.73    0.09    0.01
C4_true    0.01    0.11    0.10    0.76    0.02
```

```
C5_true    0.00    0.00    0.04    0.06    0.90
```

The element mU_modal, which is returned when extout = TRUE, contains the modal class assignment of the units. As shown below, respondents scoring 0 on all the items are estimated to belong to class 5.

```
head(out$mU_modal, 1)
obey rights local work envir vote history respect news protest discuss party
   0      0     0    0     0    0       0       0    0       0       0     0
C1 C2 C3 C4 C5
 0  0  0  0  1
```

Next, we extend the analysis of Oser, Hooghe, et al. (2023) by accounting for the hierarchical structure of the data using the multilevel LC model. The higher-level unit is the country of the respondent (the dataIEA column COUNTRY). The rationale of this multilevel modeling is that we do not assume the distribution of citizenship norms to be invariant across countries. We could reasonably accept that this distribution would vary across different clusters of countries. We perform model selection on the higher level and, to illustrate how multilevel LC analysis is typically carried out, the lower level. For simplicity of illustration, we consider a small range of values; 1-2 classes on the higher level and 4-5 classes on the lower level (in a more substantive LC analysis of these data, we should reasonably consider larger ranges, such as 1-4 on the higher level and 1-6 on the lower level). In applied LC analysis, the one-class specification is often included in model selection to test for the presence of a clustering structure in the data. We perform model selection using the sequential approach by executing

```
out = multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 4:5,
               id_high = "COUNTRY", iM = 1:2)
$step1
     BIClow    BIChigh   AIC       ICL_BIClow ICL_BIChigh
iT=4 877289.33 876869.28 876813.64 -          -
iT=5 872987.19 872460.07 872390.24 -          -

$step2
```

```
         BIClow    BIChigh   AIC       ICL_BIClow ICL_BIChigh
iT*,iM=1 872987.19 872460.07 872390.24 -          -
iT*,iM=2 869122.92 868554.62 868479.34 952146.46  868554.62

$step3
         BIClow    BIChigh   AIC       ICL_BIClow ICL_BIChigh
iT=4,iM* 873450.73 872997.73 872937.72 942352.88  872997.73
iT=5,iM* 869122.92 868554.62 868479.34 952146.47  868554.62

$optimal

iT= 5
iM= 2
```

The `multiLCA()` output above shows that the model with two higher-level classes and five lower-level classes was selected as the local optimum across the considered specifications. The value $T = 5$ was selected based on the lower-level BIC in the first step, $M = 2$ selected based on the higher-level BIC in the second step, and $T = 5$ re-selected based on the lower-level BIC in the third step.

The function call for model selection returns the results for the optimal model. This is equivalent to directly estimating the model of interest, if it were "known" to be the locally optimal specification, that is, by executing

```
out = multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 5,
               id_high = "COUNTRY", iM = 2)
```

For brevity, we do not print the output for this model. The equivalent fixed-effect model can be estimated by executing

```
out = multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 5,
               Z = "COUNTRY", fixedpars = 0)
```

Again, for brevity, we do not print the resulting output.

Next, we add covariates on both levels, specifying the model defined in (6). On the higher level, we consider as covariate the country's gross domestic product (GDP) per capita in constant terms with log transformation (log_gdp_constant). These data are obtained from the International Monetary Fund, and included in dataIEA. On the lower level, we consider as covariates the respondent's gender (female; 1 if the respondent is a girl, 0 if the respondent is a boy) and immigration status of the family (immigrantfam; 1 if the respondent comes from a family of immigrants, 0 otherwise). We estimate this model by executing

```
multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 5,
         id_high = "COUNTRY", iM = 2,
         Z = c("female","immigrantfam"), Zh = "log_gdp_constant")

GROUP PROPORTIONS (SAMPLE MEAN):

P(G1) 0.5909
P(G2) 0.4091

CLASS PROPORTIONS (SAMPLE MEAN):

            G1     G2
P(C1|G) 0.2904 0.5494
P(C2|G) 0.4135 0.2729
P(C3|G) 0.1193 0.0884
P(C4|G) 0.1467 0.0667
P(C5|G) 0.0300 0.0226

---------------------------

LOGISTIC MODEL FOR HIGHER-LEVEL CLASS MEMBERSHIP:


MODEL FOR G2 (BASE G1)

                             Alpha   S.E.  Z-score   p-value
alpha(Intercept|G2)         9.2286 0.1748  52.7958 0.0000***
alpha(log_gdp_constant|G2) -0.9376 0.0171 -54.6772 0.0000***
```

```
---------------------------


LOGISTIC MODEL FOR LOWER-LEVEL CLASS MEMBERSHIP:



MODEL FOR C4 (BASE C1) GIVEN G1

                             Gamma   S.E.  Z-score   p-value
gamma(Intercept|C4,G1)     -0.7142 0.1052  -6.7866 0.0000***
gamma(female|C4,G1)         0.0789 0.0377   2.0914 0.0365**
gamma(immigrantfam|C4,G1) -0.2994 0.0697  -4.2963 0.0000***

MODEL FOR C4 (BASE C1) GIVEN G2

                             Gamma   S.E.  Z-score   p-value
gamma(Intercept|C4,G2)     -2.0883 0.1337 -15.6217 0.0000***
gamma(female|C4,G2)        -0.0653 0.0499  -1.3093 0.1904
gamma(immigrantfam|C4,G2)  0.4719 0.0867   5.4434 0.0000***


 *** p < 0.01, ** p < 0.05, * p < 0.1
```

As shown in the partial `multiLCA()` output above, the results suggest that 59% of the countries belong to higher-level class 1, while 41% belong to higher-level class 2. The countries belonging to higher-level class 1 emphasize the citizenship norms of lower-level class 2, while the countries belonging to higher-level class 1 emphasize the citizenship norms of lower-level class 1. As such, we can label higher-level class 1 “Mainstream-emphasizing countries”, and higher-level class 2 “Maximalist-emphasizing countries”. The prevalence of the citizenship norms of lower-level class 4 (Duty) within higher-level class 1 (Mainstream-emphasizing countries) is about twice as high compared to the prevalence within higher-level class 2 (Maximalist-emphasizing countries).

Below the class separation statistics and information criteria, we can see the estimated logistic regression model for higher-level class membership. The negative and highly statistically significant estimate for the effect of GDP per capita - `alpha(log_gdp_constant|G2)` - suggests that wealthier countries have smaller probabilities of belonging to higher-level class 2 relative to higher-level class 1 than less wealthy countries.

Furthermore, we can see the logistic regression parameter estimates for lower-level class membership, conditional on higher-level class membership. For brevity, we comment only on the logistic regression coefficient for gender in the model for membership lower-level class 4 (Duty) relative to lower-level class 1, given higher-level class 1. This coefficient is labeled `gamma(female|C4,G1)` in the above output. The positive sign and statistical significance (at the 5%-level) suggest that, in the countries belonging to higher-level class 1, girls have larger probabilities than boys of belonging to lower-level class 4 relative to lower-level class 1, when controlling for immigration background.

To investigate the posterior class membership probabilities, we specify `extout = TRUE`. We focus on the posterior higher-level class membership probabilities for the countries, which is contained in the element `mPW`, rounding the values to two decimal points (`R` does not display decimal points when the values are very close to 0 or 1). In the printed partial output below, we can see that higher-level class 1 includes, for example, the Nordic countries: Denmark (`DNK`), Finland (`FIN`), Norway (`NOR`) and Sweden (`SWE`). Higher-level class 2 includes, for example, the Asian areas: Hong Kong (`HKG`), South Korea (`KOR`) and Taiwan (`TWN`).

```
out = multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 5,
               id_high = "COUNTRY", iM = 2,
               Z = c("female","immigrantfam"), Zh = "log_gdp_constant",
               extout = TRUE)
round(out$mPW, 2)
    log_gdp_constant G1 G2
DNK            10.70  1  0
FIN            10.56  1  0
HKG            10.89  0  1
KOR            10.44  0  1
```

```
NOR           11.08  1  0
SWE           10.73  1  0
TWN           10.69  0  1
```

6. Concluding remarks

We presented the state of the art of multilevel latent class modeling with covariates. The focus was on estimation approaches, model selection, and freeware-software. We presented the theoretical modeling framework, the most advantageous estimation approaches, and recommendations for model selection, including a benchmark simulation study of performance and estimation times for model selection. We gave a tutorial of the user-friendly syntax of the `R` package `multilevLCA` that executes this estimation, visualizes the results, and implements semi-automatic model selection.

The aim of the article was to disseminate the use of advanced multilevel latent class modeling among applied researchers from a variety of academic disciplines. Multilevel latent class analysis has a wide range of applications in fields such as the educational, political, economic, health and behavioral disciplines. There is considerable appeal in this methodology, which allows great flexibility in the parametrization of individual differences in a (possibly multidimensional) phenomenon of interest.

Disclosure statement

Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Funding details

This work was supported by Grant 101077659 from the European Research Council.

Acknowledgement(s)

The authors would like to thank Yuqi Liu for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the author's institutions or the European Research Council is not intended and should not be inferred.

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*(1), 117–128.

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 329–341.

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.

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

Bartolucci, F., Bacci, S., & Gnaldi, M. (2014). Multilcirt: An R package for multidimensional latent class item response models. *Computational Statistics & Data Analysis*, *71*, 971–985.

Bijmolt, T. H. A., Paas, L. J., & Vermunt, J. K. (2004). Country and consumer segmentation: Multi-level latent class analysis of financial product ownership. *International Journal of Research in Marketing*, *21*(4), 323–340.

Bray, B. C., Watson, D. P., Salisbury-Afshar, E., Taylor, L., & McGuire, A. (2023). Patterns of opioid use behaviors among patients seen in the emergency department: Latent class analysis of baseline data from the point pragmatic trial. *Journal of Substance Use and Addiction Treatment*, *146*, 208979.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Di Mari, R., & Lyrvall, J. (2024). multilevLCA: Estimates and plots single-level and multilevel latent class models [Computer software manual]. (R package)

Di Mari, R., Bakk, Z., Oser, J., & Kuha, J. (2023a). Multilevel latent class analysis with covariates: Analysis of cross-national citizenship norms with a two-stage approach. *arXiv preprint arXiv:2307.10720*.

Di Mari, R., Bakk, Z., Oser, J., & Kuha, J. (2023b). A two-step estimator for multilevel latent class analysis with covariates. *Psychometrika*, *88*, 1144-1170.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*, 1–18.

Eddelbuettel, D., & Sanderson, C. (2014). Rcpparmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data analysis*, *71*, 1054–1063.

Finch, W. H., & French, B. F. (2014). Multilevel latent class analysis: Parametric and non-parametric models. *The Journal of Experimental Education*, *82*(3), 307–333.

Gnaldi, M., Bacci, S., & Bartolucci, F. (2016). A multilevel finite mixture item response model to cluster examinees and schools. *Advances in Data Analysis and Classification*, *10*, 53–70.

Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I-A Modified latent structure approach. *American Journal of Sociology*, *79*(5), 1179–1259.

Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231.

Hagenaars, J. A. (1992). *Exemplifying longitudinal log-linear analysis with latent variables.* European Science Foundation, Scientific Network on Household Panel Studies.

Henry, K. L., & Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(2), 193–215.

Hickendorff, M., van Putten, C. M., Verhelst, N. D., & Heiser, W. J. (2010). Individual differences in strategy use on division problems: Mental versus written computation. *Journal of Educational Psychology*, *102*(2), 438.

Hooghe, M., & Oser, J. (2015). The rise of engaged citizenship: The evolution of citizenship norms among adolescents in 21 countries between 1999 and 2009. *International Journal of Comparative Sociology*, *56*(1), 29–52.

Hooghe, M., Oser, J., & Marien, S. (2016). A comparative analysis of 'good citizenship': A latent class analysis of adolescents' citizenship norms in 38 countries. *International Political Science Review*, *37*(1), 115–129.

Kim, Y., Jeon, S., Chang, C., & Chung, H. (2022). glca: An R package for multiple-group latent class analysis. *Applied Psychological Measurement*, *46*(5), 439–441.

Lacourse, É., de la Sablonnière, Giguère, C.-É., Morin, S., Legault, R., & Laliberté, F. (2024). stepmixr: Interface to Python package StepMix [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=stepmixr` (R package version 0.1.2)

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*(364), 805–811.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Houghton Mifflin.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*, 1–29.

Lukočienė, 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.

Lyrvall, J., Bakk, Z., Oser, J., & Di Mari, R. (2024). Bias-adjusted three-step multilevel latent class modeling with covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(4), 592–603.

MacLahlan, G., & Peel, D. (2000). *Finite mixture models.* John Wiley & Sons.

Magidson, J. (1981). Qualitative variance, entropy, and correlation ratios for nominal dependent variables. *Social Science Research*, *10*(2), 177–194.

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 180–197.

McCutcheon, A. L. (1979). *Analysis of qualitative data: New developments.* Academic Press.

Morgan, G. B. (2015). Mixed mode latent class analysis: An examination of fit index performance for classification. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 76–86.

Morin, S., Legault, R., Laliberté, F., Bakk, Z., Giguère, C.-É., de la Sablonnière, R., & Lacourse, É. (2023). StepMix: A Python package for pseudo-likelihood estimation of generalized mixture models with external variables. *arXiv preprint arXiv:2304.03853*.

Muthén, B., & Muthén, L. (2017). Mplus. In *Handbook of item response theory* (pp. 507–518). Chapman and Hall/CRC.

Oser, J. (2022). Protest as one political act in individuals' participation repertoires: Latent class analysis and political participant types. *American Behavioral Scientist*, *66*(4), 510–

532.

Oser, J., Di Mari, R., & Bakk, Z. (2023). Data preparation for citizenship norm analysis, International Association for the Evaluation of Educational Achievement (IEA) 1999-2009-2016. *Open Science Framework*, *10*. Retrieved from `https://doi.org/10.17605/OSF.IO/AKS42`

Oser, J., & Hooghe, M. (2013). The evolution of citizenship norms among Scandinavian adolescents, 1999–2009. *Scandinavian Political Studies*, *36*(4), 320–346.

Oser, J., Hooghe, M., Bakk, Z., & Di Mari, R. (2023). Changing citizenship norms among adolescents, 1999-2009-2016: A two-step latent class approach with measurement equivalence testing. *Quality & Quantity*, *57*, 4915–4933.

Prior, M. (2010). You've either got it or you don't? the stability of political interest over the life cycle. *The Journal of Politics*, *72*(3), 747–766.

Schulz, W., Ainley, J., Fraillon, J., Losito, B., Agrusti, G., & Friedman, T. (2018). *Becoming citizens in a changing world: IEA International Civic and Citizenship Education Study 2016 international report.* Springer Nature.

Van Deth, J., Abendschön, S., & Vollmar, M. (2011). Children and politics: An empirical reassessment of early political socialization. *Political Psychology*, *32*(1), 147–174.

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

Vermunt, J. K. (2005). Mixed-effects logistic regression models for indirectly observed discrete outcome variables. *Multivariate Behavioral Research*, *40*(3), 281–301.

Vermunt, J. K., & Magidson, J. (2021). LG-syntax user's guide: Manual for Latent GOLD syntax module version 6.0. *Arlington, MA: Statistical Innovations*.

Sim. cond.	T	M	T-exc.	M-exc.
1	3	2	1	1
2	4	2	1	1
3	5	2	1	1
4	3	3	1	1
5	4	3	1	1
6	5	3	1	1
7	3	2	3	1
8	4	2	3	1
9	5	2	3	1
10	3	3	3	1
11	4	3	3	1
12	5	3	3	1
13	3	2	1	3
14	4	2	1	3
15	5	2	1	3
16	3	3	1	3
17	4	3	1	3
18	5	3	1	3
19	3	2	3	3
20	4	2	3	3
21	5	2	3	3
22	3	3	3	3
23	4	3	3	3
24	5	3	3	3

Table 1. Fully crossed simulation conditions based on the true and excess number of lower-level classes T, and the true and excess number of higher-level classes M.

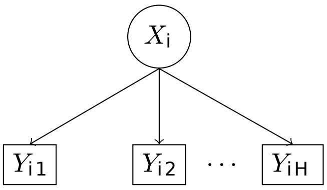

Figure 1. The single-level latent class model, with categorical indicators Y and a categorical latent class variable X.

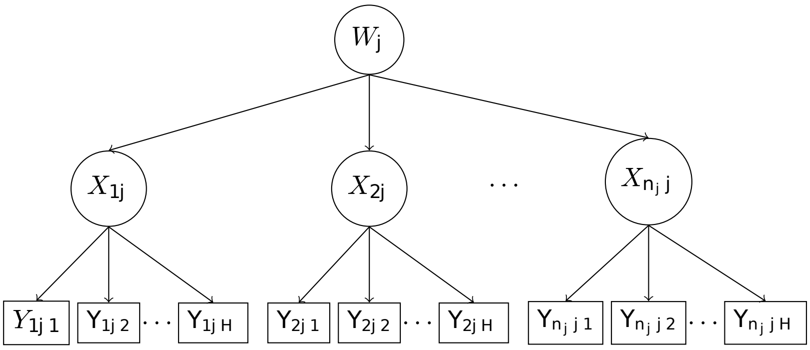

Figure 2. The multilevel latent class model, with categorical indicators Y, a categorical lower-level latent class variable X, and a categorical higher-level latent class variable W.

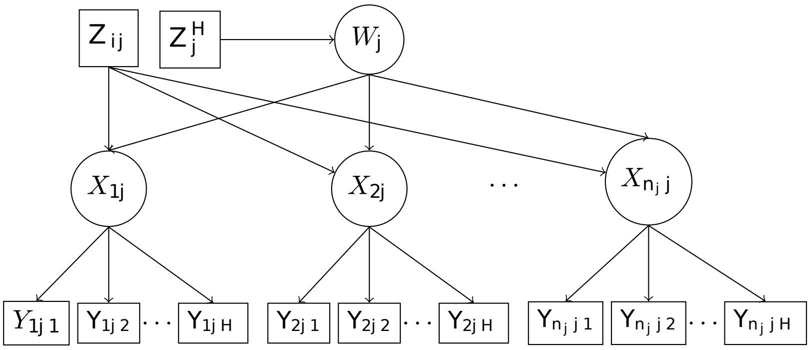

Figure 3. The multilevel latent class model with covariates.

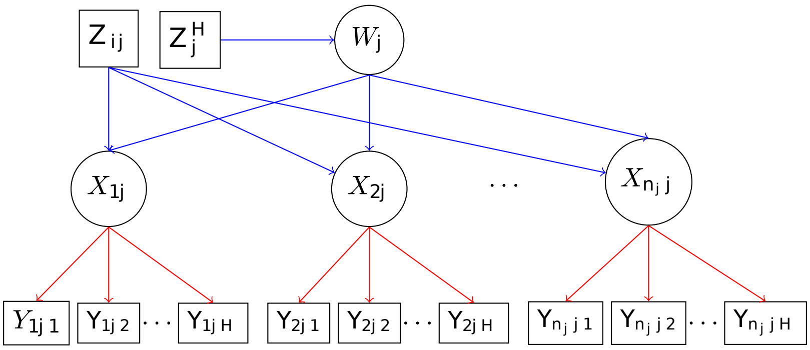

Figure 4. The measurement model (red) and structural models (blue).

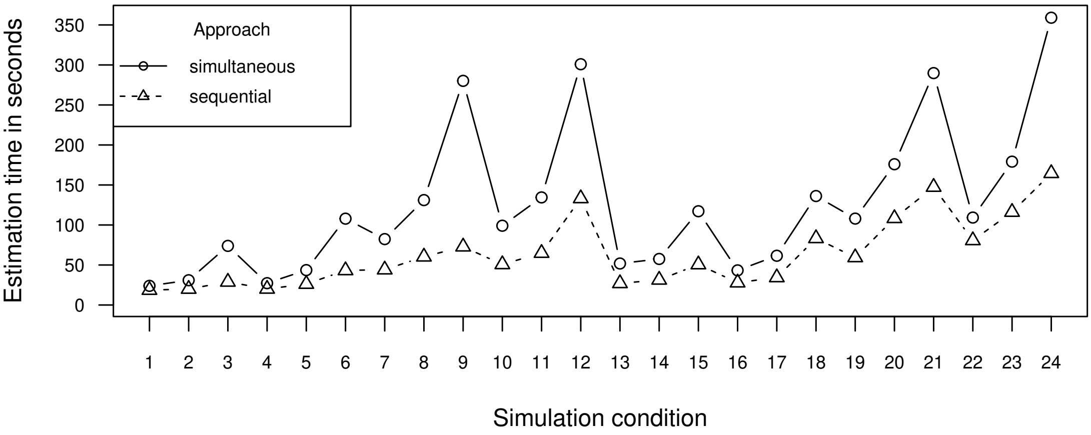

Figure 5. Estimation time for the sequential model selection approach and the simultaneous model selection approach, averaged across the 24 simulation conditions and the 50 replications.

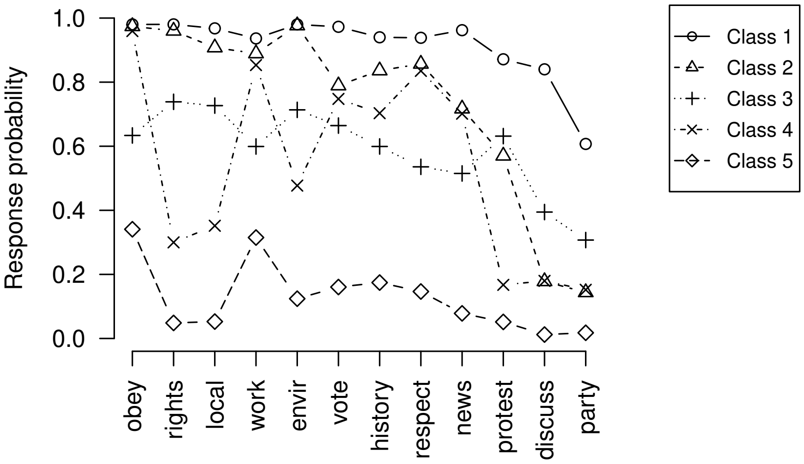

Figure 6. Plot generated using the function `multiLCA()`.