A random–covariate approach for distal outcome prediction with latent class analysis

Roberto Di Mari
Department of Economics and Business, University of Catania, Italy

Zsuzsa Bakk
Statistics and Methodology Unit, Institute of Psychology, Leiden University, The Netherlands

Antonio Punzo
Department of Economics and Business, University of Catania, Italy

Abstract

While latent class (LC) models with distal outcomes are becoming popular in literature as a consequence of the increasing use of stepwise estimators, these models still suffer from severe shortcomings. Namely, using the currently available stepwise estimators the direct effects between the distal outcome and the indicators of the LC membership cannot be easily modeled. At the same time using the traditional Full Information Maximum Likelihood (FIML) approach the LC solution can become dominated by the distal outcome, especially when model misspecifications occur, and the relationship between the distal outcome and LC is strong. In this paper, we consider a more general formulation, typical in cluster-weighted models, which embeds both the latent class regression and the distal outcome models. This allows us to test simultaneously both whether the distribution of the distal outcome differs across classes, and whether there are significant direct effects of the distal outcome on the indicators, by including most of the information about the distal outcome - latent variable relationship. We propose a two-step estimator for these models that makes it possible to separate the estimation of the measurement and structural model, that is much desired for distal outcome models, while keeping the possibility of modeling direct effects open. We show the advantages of the proposed modeling approach through a simulation study and an empirical application on assets ownership of Italian households.

Keywords: latent class analysis, continuous distal outcomes, direct effects, cluster-weighted models, random covariates, two–step approach, household wealth, assets ownership

Introduction

Latent class (LC) analysis (McCutcheon, 1985) is widely used in the social and behavioral sciences to locate latent subgroups of observations in the sample based on a set of J observed response variables $\boldsymbol{Y}$. Examples of applications include identification of types of mobile internet usage in travel planning and execution (Okazaki et al., 2015), types of political involvement (Hagenaars & Halman, 1989), classes of treatment engagement in adolescents with psychiatric problems (Roedelof et al., 2013), a typology of infant temperament (Loken, 2004), modeling phases in the development of transitive reasoning (Bouwmeester & Sijtsma, 2007), or classes of self disclosure (Maij-de Meij et al., 2005).

In many empirical studies, interest lies in investigating the relationship between the LC model and its antecedents and consequences for building more complex theoretical frameworks. Latent class models with covariates (Dayton & Macready, 1988) are a well-known extension of the baseline model, in which external variables $\boldsymbol{Z}$ are included in the latent class modeling framework as predictors of the class membership X (Collins & Lanza, 2010). Stegmann & Grimm (2017), for instance, discuss the inclusion of covariates in more complicated LC models. Using LC membership as a predictor of - possibly continuous - distal outcomes $\boldsymbol{Z}$ as depicted in Figure 1, is also becoming a popular modeling approach (Bakk et al., 2013; Lanza et al., 2013). For instance, Roberts & Ward (2011) predict distal pain outcomes based on class memberships defined by patterns of barriers to pain management and Mulder et al. (2012) compared average measures of recidivism in clusters of juvenile offenders. In this paper, without loss of generality, but for simplicity, we will consider the case of a univariate external variable, notationally indicated as Z.

The widespread use of distal outcome models is relatively new, due to the challenges of building these types of models. For a long time there have been two approaches for estimating distal outcome models, both with severe limitations. On the one hand the most common LCA estimation approach - FIML, or one-step approach - is not recommended for distal outcome models, mainly because of its theoretical limitation: the outcome that is predicted by the LC variable contributes to the definition of the LC variable, creating an unintended circularity. While when all model assumptions are met the one-step approach is equivalent to stepwise approaches, because of the theoretical limitations researchers tend to avoid one-step modeling for distal outcome scenarios. On the other hand, the alternative three-step approach in which in the first step the LC variable is estimated based on its indicators only, in step two the respondents are assigned to posterior classes, and the posterior classification is related in step three to a distal outcome using an ANOVA model has its own limitations too. A main disadvantage of this approach is that in step two a classification error is introduced that leads to bias in the step three estimates. Recently, correction methods have been introduced for the three-step approach, known as the bias-adjusted three-step approaches (Bolck et al., 2004; Vermunt, 2010) that make this approach the obvious choice in applied modeling. An overview of existing three-step approaches and recommendations about their usage is available in Nylund-Gibson et al. (2019) - in particular, see their Table 1. Nevertheless using the three-step approach building complex models with multiple distal outcomes with possible residual associations and multiple covariates is still very challenging in practice.

The major challenge in building (complex) distal outcome models is that the distal outcome and the J response variables $\boldsymbol{Y}$'s are assumed to be conditionally independent given the latent

Address:
Room 20,
Corso Italia, 55, 95129 Catania
Phone: +390957537646
Email: roberto.dimari@unict.it

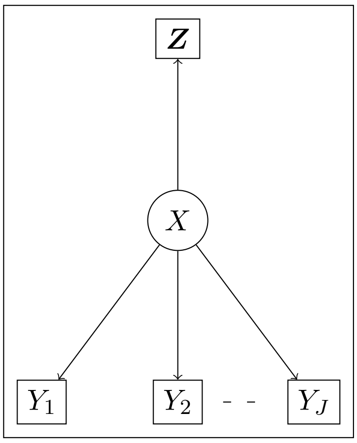


Figure 1. Latent class model with distal outcomes $\boldsymbol{Z}$.

variable X (Bakk et al., 2013; Lanza et al., 2013). A direct effect of Z on $\boldsymbol{Y}$ is therefore not allowed for, neither its presence tested. In latent variable modeling, it is well known that FIML estimation is subject to severe bias when direct effects (DEs) are present and not accounted for in LC and latent trait models (Asparouhov & Muthén, 2014), regression mixture models (Kim et al., 2016; Nylund-Gibson & Masyn, 2016), and latent Markov models (Di Mari & Bakk, 2017). Furthermore Zhu et al. (2017) showed how ignoring DEs for distal outcome models can lead to wrong class enumeration. Using the three-step approach it is even not possible to model or easily test for such direct effects (Asparouhov & Muthén, 2014). If DEs are hypothesized they need to be included in the step one model, and if the model is re-estimated with another covariate having DE, the step one model needs to be re-estimated, thus defeating the purpose of stepwise modeling. The more complex a model is, the more likely the presence of violations of the conditional independence assumption, and thus an increase in model misspecification. FIML is known to be very sensitive to model misspecifications, in most extreme cases the full LC variable can change its meaning if a misspecified distal outcome is added to the model, that drives the ML solution away from the initial LC model (Petras & Masyn, 2010). Using the three-step approach these re-estimation of the LC variable is not possible, however the misspecification becomes masked, and leads to bias (Bakk & Vermunt, 2016).

An alternative two step approach was also recently proposed (Bakk & Kuha, 2017), that after estimating the LC model conditions directly on the step one estimates to estimate the distal outcome model of interest. Using this approach, in the second step direct effects between the external variables and the LC indicators can be estimated. Janssen et al. (2019) have shown this for the case of a categorical covariates. Alternatively direct effects can be handled using a FIML approach proposed by Masyn (2017). This later approach estimates multiple FIML models choosing the one that fits all the necessary DEs. While promising, the approach is cumbersome, and the effects of multiple testing on power are not known. Furthermore both of these approaches for modeling direct effects/ differential item functioning were proposed only for models with covariates, not with distal outcomes.

Given the restrictiveness and the widespread use of the conditional independence assumption and the possible severity of its violation, we propose a more general model that can account for direct effects between the distal outcome, LC membership, and the indicators of the LC model. The proposed approach is able to handle circular models based on the literature on cluster-weighted modeling (also known as mixture with random covariates approach). In regression mixtures, a "circular" relation among $\boldsymbol{Y}$-X-Z is commonly considered in the random covariate (cluster-weighted modeling) approach (Ingrassia et al., 2012, 2014, 2015, Ingrassia & Punzo, 2016 and Punzo & Ingrassia, 2016). That is, a more general model is specified, where next to modeling the class specific distribution of Z (distal outcome situation), also the direct effect of Z on $\boldsymbol{Y}$ is modeled Using this approach all possible direct effects are modeled, similarly to the most general model allowing for DE proposed by Masyn (2019). But instead of estimating multiple nested models, we simply allow all DEs to be present. Using this overparametrization we avoid misspecification in modeling DEs, and insure that the parameters of interest are estimated unbiased. With standard inference, the statistical significance of each effect can then be tested and interpreted when necessary. Using the proposed approach a very flexible framework for latent class modeling with distal outcomes can be established. In structural equation models similar approaches are known as non-recursive models, models that include a feedback loop between two variables, instead of having a linear directionality (Finch & French, 2015; Bollen, 1996). These models are usually estimated using a form of a stepwise estimator, namely the two-stage least square approach (2SLS), using Instrumental Variables to account for the feedback loop (Bollen, 1996). A strength of this 2SLS estimator is that it is consistent even when misspecifications are present, that would lead to bias in the FIML approach (Bollen et al., 2007) Based on this literature we also propose a two-step estimator of our model that is robust against misspecifications.

Some possible interesting substantive applications of such a circular model can be an LC model of social status at career start defined by education, parents education etc. Social status can be associated with future income. We can imagine that at different levels of income the effect of indicators of social status differs. Another example - presented in our real data section - shows how an LC model of household asset ownership can differ at levels of predicted wealth (the distal outcome).

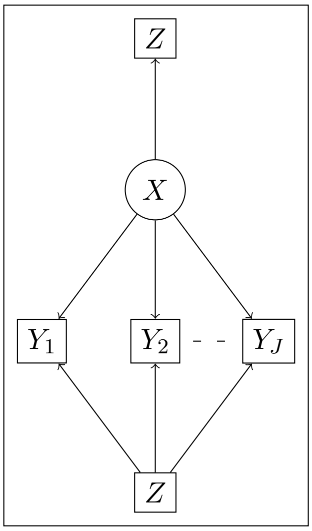


Figure 2. Latent class model with a random covariate.

The proposed approach is optimal for handling direct effects of the distal outcome Z on the LC indicators $\boldsymbol{Y}$. Furthermore it is well known that with a continuous Z and binary $\boldsymbol{Y}$, the likelihood often gets dominated by the distribution for Z (Petras & Masyn, 2010), and even relatively minor distributional specification errors for Z can have strong effects on class enumeration and prediction. That is, the classes may wander away from the optimal profiles for ensuring local independence of $\boldsymbol{Y}$ in the service of better approximating the distribution of Z. We highlight how the proposed random covariate approach is robust towards such misspecifications as well.

In principle, both two- and three-step approaches can be implemented to fit a random-covariate latent class model. However, the three-step method would wrongly identify latent groups in its first step from the conditional distribution of $\boldsymbol{Y}$ given Z - a LC regression model - rather than from jointly $\boldsymbol{Y}$, X and Z. Furthermore using the three–step approach all the direct effects would have to be included in step one, and in case a new Z variable is added to the model that has direct effects, the full step one model would need to be re-estimated, defeating the purpose of stepwise modeling. The misspecification can be as severe as to impact class enumeration (Kim et al., 2016; Zhu et al., 2017; Nylund-Gibson & Masyn, 2016). We propose a two–step approach to fit the random-covariate latent class model. In the first step, a LC model is fitted to the data; in the second step, direct effects and the part on the distal outcome are added to the model, and the parameters of the marginal distribution of X and the intercept terms of the logistic regression of $\boldsymbol{Y}$ given X and Z are kept fixed at their first step value. This has the advantage of 1) keeping model complexity under control, even in the case of many items / latent classes, and 2) allowing class enumeration and identification not to depend on the distributional assumption for Z. We will show, based on simulation and real-data results, that this approach constitutes a more robust, practical and user-friendly alternative to the one–step approach.

The paper proceeds as follows. We introduce the general LCA framework - which includes a description of the one-step distal outcome model and the proposed random-covariate latent class model in Section “The different modeling approaches in details”. Then (in Section “Parameter estimation of the random-covariate latent class model”) we give details for both one-step and two-step approaches for parameter estimation for the proposed random-covariate latent class model, and we discuss how to perform correct inference on the model parameters of the second step model. In Section “Simulation study” we describe the design of our simulation study along with the results. In Section “A latent class model of households’ assets ownership to predict wealth”, we analyze data from the Household Finance and Consumption Survey, and conclude with some final remarks in Section “Conclusion”.

The different modeling approaches in details

The latent class model

Let $\boldsymbol{Y} = (Y_1, \dots, Y_J)'$ be the vector of the full response pattern and $\boldsymbol{y}$ its realization. Let us denote as X the categorical latent variable, with latent classes $s = 1, \dots, S$. A simple latent class model (Goodman, 1974; McCutcheon, 1987; Hagenaars, 1990) for $P(\boldsymbol{Y})$ is defined as

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{s=1}^{S} P(X = s) P(\boldsymbol{Y} = \boldsymbol{y} \,| X = s), \tag{1}$$

where $P(X = s)$ is the unconditional probability of belonging to class s - structural component - and $P(\boldsymbol{Y} = \boldsymbol{y} \,| X = s)$ is the probability of observing response configuration $\boldsymbol{y}$ given the class membership s - measurement component.

Classical Latent Class Analysis (LCA) assumes local independence of the response variables given the class membership. This implies that $P(\boldsymbol{Y} = \boldsymbol{y}\,|X = s)$ can be written as

$$P(\boldsymbol{Y} = \boldsymbol{y}\,|X = s) = \prod_{j=1}^{J} P(Y_j = y_j|X = s). \tag{2}$$

For estimating the model in Equation (1), we assume each Y_j in (2) to be conditionally Bernoulli distributed, with success probability π_{js}, and parametrize the conditional response probabilities through the following log-odds

$$\log\left(\frac{\pi_{js}}{1-\pi_{js}}\right) = \beta_{0,js}, \tag{3}$$

for $1 \leq s \leq S$. The latent class unconditional probabilities can as well be parametrized using logistic regressions. We opt for the following parametrization

$$\log\left[\frac{P(X = s)}{P(X = 1)}\right] = \rho_s, \tag{4}$$

for $1 < s \leq S$, where we take the first category as reference, and we set to zero the related parameter. Under these parametrizations, the total number of free parameters to be estimated in model (1) is $JS + S - 1$.

The model of Equation (1) can be used to assign observations to latent classes based on the posterior membership probabilities

$$P(\mathrm{X} = s|\,\boldsymbol{Y} = \boldsymbol{y}) = \frac{P(X = s)P(\boldsymbol{Y} = \boldsymbol{y}\,|X = s)}{P(\boldsymbol{Y} = \boldsymbol{y})}, \tag{5}$$

according to some assignment rules. The most common assignment rules are modal or proportional assignment; for details see Vermunt (2010)

The latent class with distal outcome model

The latent class variable can be taken to be a predictor of the external variable Z (distal outcome), yielding the following latent class with distal outcome model

$$P(\boldsymbol{Y} = \boldsymbol{y}, Z = z) = \sum_{s=1}^{S} P(X = s)P(Z = z|X = s)P(\boldsymbol{Y} = \boldsymbol{y}\,|X = s), \tag{6}$$

where the structural and measurement components are as above, and $P(Z = z|X = s)$ is the external variable component and models the class specific distribution of Z. Under the local independence assumption of the items given the latent class variable, the response conditional probabilities are as in Equation (2). That is, $\boldsymbol{Y}$ is assumed to be independent of Z given X, which is a standard, and rather very strong, assumption of LCA.

The model of Equation (6) can be used to cluster observations, according to modal or proportional assignment rules, based on the following posterior membership probabilities

$$P(\mathrm{X} = s|\,\boldsymbol{Y} = \boldsymbol{y}, Z = z) = \frac{P(X = s)P(Z = z|X = s)P(\boldsymbol{Y} = \boldsymbol{y}\,|X = s)}{P(\boldsymbol{Y} = \boldsymbol{y}, Z = z)}. \tag{7}$$

The external variable Z is assumed, conditional to the latent class, to be Gaussian with mean μ_s and variance σ_s^2, for $1 \leq s \leq S$, whereby an equivalent specification as in Equation (4) can be used to parametrize the latent class unconditional probabilities. This yields $JS+2S+S-1$ free parameters to be estimated.

While Equation 6 describes the complete distal outcome model that can be estimated using FIML, this is hardly ever done in practice for reasons already described. In practice the three- step (Vermunt, 2010; Asparouhov & Muthén, 2014) or two-step (Bakk & Kuha, 2017) approaches are used, that are already extensively presented in previous literature, and we will not discuss in detail. Currently none of these stepwise approaches is able to easily incorporate DEs between indicators and distal outcomes. Existing literature focuses on modeling DEs for models with covariates only (Asparouhov & Muthén, 2014; Janssen et al., 2019; Masyn, 2017).

The random–covariate latent class model

A more general form of association between $\boldsymbol{Y}$, X and Z, involves modeling the following joint probability

$$P(Z=z, X=s, \boldsymbol{Y}=\boldsymbol{y}) = P(Z=z, X=s)P(\boldsymbol{Y}=\boldsymbol{y}\,|Z=z, X=s), \tag{8}$$

where the common assumption in LCA of $\boldsymbol{Y}$ and Z being conditionally independent given the latent process is relaxed. From Equation (8), several submodels can be specified (covariate model, distal outcome model, etc). If substantive theoretical arguments postulate the latent variable to be a predictor of the external variable Z, the random–covariate latent class model specifies the probability of observing a response pattern $\boldsymbol{y}$ as

$$P(\boldsymbol{Y}=\boldsymbol{y}, Z=z) = \sum_{s=1}^{S} P(X=s)P(Z=z|X=s)P(\boldsymbol{Y}=\boldsymbol{y}\,|Z=z, X=s), \tag{9}$$

which contains a structural part as in the previous models; a measurement component connecting the latent class to the observed responses with a direct effect of Z, as in the latent class regression model; an external variable component, which models the latent class specific distribution of Z. That is, the random–covariate latent class model incorporates features of 1) the simple latent class model (in the structural component), 2) the latent class regression model (in the measurement component), and 3) the distal outcome model (in the external variable component). For the measurement component, as for the classical LC model in Equation (1), we assume each Y_j to be conditionally Bernoulli distributed, with success probability π_{js}, and parametrize the conditional response probabilities through the following log-odds

$$\log\left(\frac{\pi_{js}}{1-\pi_{js}}\right) = \beta_{0,js} + Z\beta_{js}. \tag{10}$$

The parametrization for the structural component is as in Equation (4), under the assumption of Z being conditionally Gaussian with mean μ_s and variance σ_s^2, for $1 \leq s \leq S$. Equally, the posterior membership probabilities are computed as

$$P(\mathrm{X}=s|\,\boldsymbol{Y}=\boldsymbol{y}, Z=z) = \frac{P(X=s)P(Z=z|X=s)P(\boldsymbol{Y}=\boldsymbol{y}\,|Z=z, X=s)}{P(\boldsymbol{Y}=\boldsymbol{y}, Z=z)}, \tag{11}$$

and sample units can be assigned to classes according to, for instance, modal or proportional assignment rules.

One key feature is that sub-models of the random–covariate latent class model can be backed up by imposing suitable restrictions on the model parameters. In particular, by setting the β_{js}'s of Equation (10) to zero, the random–covariate latent class model reduces to a standard LC with distal outcome model (6). Although not formally nested, also the simple LC can be thought of as a sub-model in which Z is completely excluded from the model. Furthermore similarities can be seen also with the latent class regression model, that allows for DEs between the external variable and the indicators - with the main difference being that while DEs are modeled, the external variable is not a covariate of the LC model but rather a distal outcome.

	Dir. Eff.	Z modeled	#par
Latent Class	×	×	$JS + S - 1$
Latent Class with distal outcome	×	✓	$JS + 2S + (S - 1)$
Random–covariate Latent Class	✓	✓	$2JS + 2S + (S - 1)$

Table 1

Summary of different modeling assumptions and number of free parameters to be estimated.

Table 1 summarizes how Z enters each of the three models, and the total number of free parameters to be estimated. Intuitively, this shows that the first two models can be seen as special cases of the third model, which therefore models the relationship between the 3 sets of variables in the most exhaustive manner.

Parameter estimation of the random–covariate latent class model

The (FIML) one–step approach

Let $\boldsymbol{\theta} = \{\boldsymbol{\rho}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2\}$ and let $\{(\boldsymbol{Y}_i, Z_i)\}_n = \{(\boldsymbol{Y}_1, Z_1), \ldots, (\boldsymbol{Y}_n, Z_n)\}$ be a sample of n independent observations. The model parameters $\boldsymbol{\theta}$ can be estimated by maximizing the following sample log-likelihood function

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \{P(\boldsymbol{Y}_i, Z_i)\} = \sum_{i=1}^{n} \log \left\{ \sum_{s=1}^{S} P(X = s) P(Z_i | X = s) P(\boldsymbol{Y}_i | Z_i, X = s) \right\}. \quad (12)$$

The maximization of (12) is typically done either by means of the EM algorithm or of quasi-Newton methods, or a combination of both. However, due to the model complexity and the assumptions on the Z distribution, the simultaneous (one–step) approach might suffer a number of problems. First, especially in the case of many indicators and possibly also latent classes, standard procedures like the EM algorithm might become too unstable and even fail to converge. Second, jointly modeling the external variable and the indicators has the consequence of predicting latent classes not only based on the conditional independence assumption, but also on the distributional assumption on the continuous external variable's distribution. In case this assumption is violated, the whole LC solution might be affected, possibly requiring additional latent classes than optimal to accommodate for a misspecified distribution (Bauer & Curran, 2003). Third, even if the distribution of the external variable is not misspecified, the external variable's contribution might overly influence the log–likelihood - therefore the clustering solution. Motivated by these considerations, in the next Section we introduce a feasible two–step approach.

The two–step approach

The key feature of the two–step approach is that selection and estimation of the measurement component is done separately from estimation of the structural component. Once the body of the measurement component is obtained, the latent class variable can be taken as fixed even if, together with the external variable component, also the external variable's direct effects on the indicators are added to the model.

Let us consider decomposing the parameter set $\boldsymbol{\theta}$ into $\boldsymbol{\theta}_1 = \{\boldsymbol{\rho}, \boldsymbol{\beta}_0\}$, i.e. the parameters of the simple LC model of Equation (1), and $\boldsymbol{\theta}_2 = \{\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2\}$, i.e. the direct effects of Z on $\boldsymbol{Y}$, and the location and scale parameters of the external variable component. We propose to estimate $\boldsymbol{\theta}$ in the following two steps.

Step 1 In the first step, a simple LC model as defined in Equation (1) is estimated by maximizing the the following log–likelihood

$$\ell(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \log\left\{P(\boldsymbol{Y}_i)\right\} = \sum_{i=1}^{n} \log\left\{\sum_{s=1}^{S} P(X=s)P(\boldsymbol{Y}_i\,|X=s)\right\}, \tag{13}$$

by means of standard algorithms (like the EM algorithm).

Step 2 We let $\widehat{\boldsymbol{\theta}}_1$ denote the ML estimator of $\boldsymbol{\theta}_1$ obtained from the first step. The second step consists of maximizing the following log-likelihood function

$$\ell(\boldsymbol{\theta}_2\,|\,\boldsymbol{\theta}_1 = \widehat{\boldsymbol{\theta}}_1) = \sum_{i=1}^{n} \log\left\{P(\boldsymbol{Y}_i, Z_i)\right\}, \tag{14}$$

where ℓ is defined as in (12).

Note that the log-likelihood of Equation (14) is maximized only with respect to $\boldsymbol{\theta}_2$, whereby $\boldsymbol{\theta}_1$ are held fixed at their first step values $\widehat{\boldsymbol{\theta}}_1$. Although this class of estimators is consistent (Gong & Samaniego, 1981), standard theory would suggest that this is less efficient than the full information (one–step) ML estimator. However, recent simulation results (Bakk & Kuha, 2017) have shown that the loss in efficiency is significant only when the classes are almost completely overlapping and, in all other conditions, the two–step estimator is almost as efficient as the one–step estimator. In addition, standard errors based on the inverse of the Fisher information (negative of the Hessian matrix) computed at the second step would not take the uncertainty in the first step into account. Still, robust standard errors based on sandwich formulas are available for reliable inference on the model parameters (Bakk & Kuha, 2017; Gong & Samaniego, 1981).

Let the Fisher information matrix of the joint (one-step) model for $\boldsymbol{\theta}$ be denoted by

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) = \begin{bmatrix} \boldsymbol{\mathcal{I}}_{11} & \\ \boldsymbol{\mathcal{I}}'_{12} & \boldsymbol{\mathcal{I}}_{22} \end{bmatrix},$$

where $\boldsymbol{\theta}^*$ denotes the true value of $\boldsymbol{\theta}$ and the partitioning corresponds to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The asymptotic variance matrix of the one-step estimator $\hat{\boldsymbol{\theta}}$ is thus $\boldsymbol{V}_{ML} = \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}^*)$, which is estimated by $\widehat{\boldsymbol{V}}_{ML} = \boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}})$. Let $\boldsymbol{\Sigma}_{11}$ denote the asymptotic variance matrix of the step one estimator $\tilde{\boldsymbol{\theta}}_1$ of the two-step method, obtained similarly from the Fisher information matrix of the model described in Equation (13). The asymptotic variance matrix of the two-step estimator $\tilde{\boldsymbol{\theta}}_2$ is then

$$\boldsymbol{V} = \boldsymbol{\mathcal{I}}_{22}^{-1} + \boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{12}\,\boldsymbol{\Sigma}_{11}\,\boldsymbol{\mathcal{I}}_{12}'\boldsymbol{\mathcal{I}}_{22}^{-1} \equiv \boldsymbol{V}_2 + \boldsymbol{V}_1. \tag{15}$$

Here $\boldsymbol{V}_2$ describes the variability in $\tilde{\boldsymbol{\theta}}_2$ given the step one parameters $\boldsymbol{\theta}_1$, and $\boldsymbol{V}_1$ the additional variability arising from the fact that $\boldsymbol{\theta}_1$ are not known but rather estimated by $\widehat{\boldsymbol{\theta}}_1$ with their own sampling variability. The variance matrix $\boldsymbol{V}$ is estimated by subtituting $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)$ for $\boldsymbol{\theta}^*$.

The two–step approach can be implemented by all software used for one–step modeling, as typically those allow for estimation with some fixed parameters specified in input. It should be noted that while LC literature emphasizes the issue of separation, it is well–known in econometrics theory (Amemiya, 1985) that independence of the parameters of the first and second step models is by no means a requirement for consistency of two–stage estimators.

Simulation study

Design

The objective of this simulation study is to evaluate the proposed random–covariate approach for latent class analysis in the presence of a distal outcome, both in its one–step and two–step versions, under crossed sample size and class separation conditions, and its robustness against misspecifications on the measurement and structural components.

We generate the data from a two–class LC model with $J = 6$ dichotomous indicators and one external variable Z. Class separation, strongly related to classification error, and sample size are known to affect both one–step and stepwise modeling. We choose to manipulate separation only through the strength of the relationship between items and latent-class conditional response probabilities ($\beta_{0,js}$), although other options are available - like the number of items, the number of item categories, and the class sizes. The intercepts in the two latent classes, in the model with no direct effect of Z on the indicators, were set to

$$\boldsymbol{\beta}_{0,1} = (0.847, 0.847, 0.847, -0.847, -0.847, -0.847)' \text{ and } \boldsymbol{\beta}_{0,2} = -\boldsymbol{\beta}_{0,1},$$

such that the entropy-based R^2 is about 0.7 (moderate class separation), and to

$$\boldsymbol{\beta}_{0,1} = (2.197, 2.197, 2.197, -2.197, -2.197, -2.197)' \text{ and } \boldsymbol{\beta}_{0,2} = -\boldsymbol{\beta}_{0,1},$$

such that the entropy-based R^2 is about 0.9 (large class separation). Although smaller entropy values can be considered, stepwise approaches are known to work well with entropy-based R^2 above 0.6 (see, for instance, Vermunt, 2010). The sample sizes used were 500 (moderate), and 2000 (large).

We consider misspecifications both in the measurement model, involving direct effects of Z on the items, and on the structural model, related to the distribution of Z, to evaluate the robustness of the proposed estimator. In the first part of the simulation study, we will assess all approaches with three levels of misspecifications in the measurement component: 1) no DEs, 2) DEs on half of the indicators, 3) DEs on all the indicators, under moderate and large sample size and class separation. In case one or more DE's are present, the related coefficient is of 0.2 - which can be considered a moderate magnitude in logistic scale. Z is generated from class–specific normal distributions with means -1 and 1, respectively in the first and second latent class, with common unit variance. Together with evaluating the performance of the proposed RC approach, in this first part of the simulation study we compare the one–step and two–step random covariate approaches (1–step RC and 2–step RC respectively) with the standard distal outcome one–step and two–step LCA (“1-step dist" and “2–step dist" respectively), where direct effects are not modeled. Each method will be assessed in terms of parameter estimates, percentage bias, relative efficiency (averaged standard error over averaged standard deviation) and coverage rates. For all 12 crossed conditions obtained by combining the number of DE’s scenarios, the class separation levels and sample sizes, we considered 500 replications and fitted all four approaches.

In the second part of the simulation study, we will stress test all approaches by adding also a misspecification in the structural component. The overall setting is the same as in the small sample size ($n = 500$), moderate separation and direct effects on three indicators scenario. We consider three among the possible departures from within–class normality that may occur in practice (see also Zhu et al., 2017, for a similar choice): skewness, excess kurtosis and bimodality (see Table 2). The skewness scenario is obtained by taking the skew-normal as the within–class distribution of Z, with location $\mu \in \mathbb{R}$, scale $\omega > 0$ and shape $\alpha \in \mathbb{R}$ (see, e.g., Azzalini & Capitanio, 2014). In detail, $\mu = -1$ and $\alpha = 0.1$ in class 1, while $\mu = 1$ and $\alpha = -0.1$ in class 2, with $\omega = 1$ in both classes. The excess kurtosis scenario is instead obtained with the t distribution with location $\mu \in \mathbb{R}$, scale $\omega > 0$, and ν degrees of freedom. In order to have bimodal within–class distribution we follow the same setup as in Bakk & Kuha (2017): we consider a mixture of the following two normal distributions $N(-0.5, 0.5)$ and $N(0.75, 1.3375)$, with weights 0.6 and 0.4 respectively; this mixture is such that its variance is 1 but has positive skewness, with an index of skewness of 1.16 - approximately that of a χ^2_6. In this second simulations study, we will also include the three-step BCH estimator (Bolck et al., 2004; Vermunt, 2010) in our comparison: the reason for this is that BCH is known in literature to be robust against distributional assumptions on Z (Janssen et al., 2019). For all 3 non-normality conditions, we considered 500 replications and fitted all five approaches.

Within-class scenarios	Within-class distribution	Parameters	
		First class	Second class
Skewness	Skew-normal	$\mu = -1$, $\alpha = 0.1$, $\omega = 1$	$\mu = 1$, $\alpha = -0.1$, $\omega = 1$
Excess kurtosis	t	$\mu = -1$, $\omega = 1$, $\nu = 3$	$\mu = 1$, $\omega = 1$, $\nu = 3$
Bimodality	Normal mixture	class prop = 0.6, $\mu = -0.5$, $\sigma^2 = 0.5$	class prop = 0.4, $\mu = 0.75$, $\sigma^2 = 1.3375$

Table 2
Non-normal within-class scenarios for the Z distribution.

Results

First part - normal within-class Z. Table 3 summarizes the results for the 12 simulation conditions where we have no misspecification on the distribution of Z.

Overall, in terms of bias, the RC 1-step approach performs well except for the small sample size and no direct effect condition. In fact, this is the only case where the 1-step Dist approach has the lowest bias. We observe that both 2-step approaches perform relatively well compared to their 1-step counterparts. Interestingly, in the no direct effect and small sample size conditions, the 2-step RC approach performs better than 1-step RC, indicating perhaps that the unconditional first step of the 2-step RC brings in higher flexibility to this kind of misspecification.

In terms of coverage and SE estimation, although the RC 1- and 2-step approaches tend to underestimate the true variability in some conditions, the 1-step approach has always approximately 95% coverage whereby we observe some slight undercoverage in the 2-step RC estimator. This might indicate that the SE correction is not able to account for both first step variability and variability due to model saturation. Not surprisingly, both distal outcome approaches tend to have poor coverage in the 3 and 6 direct effect conditions.

Class	Sample	#DE's	1-step RC			1-step Dist			2-step RC			2-step Dist		
Separation	Size		%Bias	CVG	SE/SD	%Bias	CVG	SE/SD	%Bias	CVG	SE/SD	%Bias	CVG	SE/SD
Moderate	500	0	0.006	0.938	0.904	0.003	0.943	0.955	0.006	0.944	0.931	0.006	0.944	0.928
	2000	0	0.001	0.940	0.985	0.001	0.941	1.000	0.001	0.930	0.951	0.001	0.942	0.947
	500	3	0.001	0.935	0.975	0.001	0.943	1.018	0.000	0.936	0.896	0.001	0.920	0.820
	2000	3	0.003	0.946	0.955	0.003	0.931	1.026	0.003	0.946	0.979	0.003	0.950	0.865
	500	6	0.002	0.927	0.935	0.004	0.866	0.878	0.005	0.926	0.788	0.007	0.900	0.583
	2000	6	0.001	0.946	0.992	0.002	0.798	0.910	0.002	0.920	0.877	0.003	0.922	0.541
Large	500	0	0.008	0.933	0.930	0.005	0.955	1.007	0.005	0.936	0.940	0.005	0.948	0.934
	2000	0	0.000	0.948	0.970	0.000	0.942	0.973	0.002	0.934	0.919	0.002	0.942	0.912
	500	3	0.006	0.923	0.906	0.001	0.947	0.952	0.001	0.926	0.846	0.002	0.918	0.772
	2000	3	0.001	0.941	1.057	0.001	0.937	1.009	0.001	0.922	0.917	0.001	0.932	0.816
	500	6	0.003	0.922	0.918	0.003	0.858	0.829	0.002	0.906	0.755	0.001	0.898	0.572
	2000	6	0.002	0.947	0.989	0.001	0.794	0.896	0.000	0.928	0.883	0.000	0.916	0.545

Table 3

Simulation results *of estimation of the means of Z (averaged across components) for the one-step LC with Random Covariates ("1-step RC"), the one-step LC with distal outcome ("1-step dist"), the two-step LC with Random Covariates ("2-step RC") and the two-step LC with distal outcome ("2-step dist") for all simulation condition class separation × sample size × number of DE's for normal within class distribution of Z.*

Second part - non-normal within-class Z.

Skewness. In Figure 3 we display the boxplot for bias reported at each simulation replication for the first-class mean of Z, under the scenario small sample size ($n = 500$), moderate separation, direct effects on three indicators scenario and non-normal (skewed) within class distribution of Z.

All estimators have approximately between 5 and 10 % bias in the estimated mean: one possible explanation could be that unmodelled skewness translates into a shift in the estimated centers. As we speculated, we observe that the 2–step RC approach is the least sensitive to misspecification of the distribution of Z, even with respect to the 3-step BCH approach. Among the one-step approaches, the 1–step RC has the lowest bias, perhaps because modeling the DEs it deals only with a misspecified structural component - whereby the 1–step Dist has also unmodelled DEs.

Excess kurtosis. Figure 4 displays the boxplot for bias reported at each simulation replication for the first-class mean of Z, under the scenario small sample size ($n = 500$), moderate separation, direct effects on three indicators scenario and non-normal (leptokurtic) within class distribution of Z.

Even in this case, the 2–step RC approach seems to be the least sensitive to misspecification of the distribution of Z. The performances of the one-step approaches are qualitatively the same. By contrast, we find that the 3-step BCH approach performs relatively worse compared to the other approaches: this is not surprising, since all three-step approaches in LCA are known to perform poorly when there are direct effects of the external variable on the indicators (Bakk & Vermunt, 2016; Vermunt, 2010).

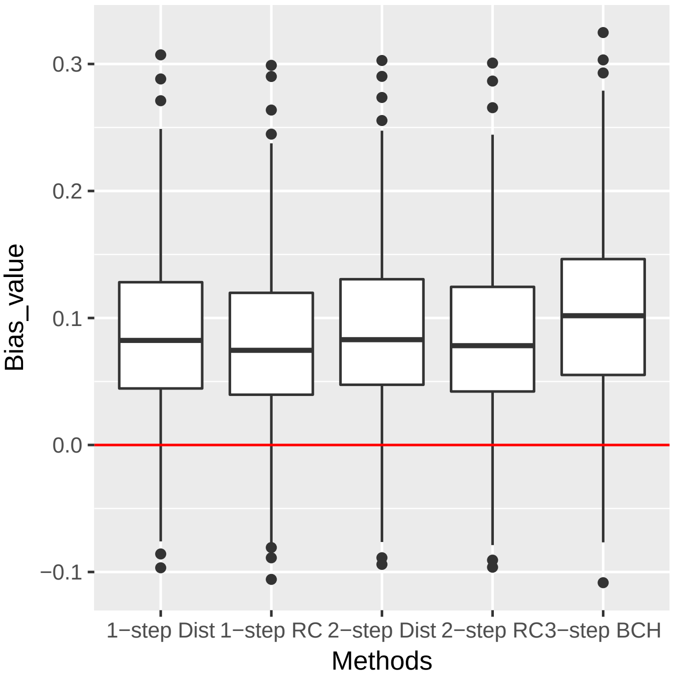


Figure 3. Boxplots of bias in the first estimated mean of Z for the non-normal (skewness) scenario, for the one-step LC with Random Covariates (“1-step RC"), the one-step LC with distal outcome (“1-step dist"), the two-step LC with Random Covariates (“2-step RC"), the two-step LC with distal outcome (“2-step dist") and the three-step BCH approach (“3-step BCH") for $n = 500$, moderate class separation and three direct effect scenario. 500 Monte Carlo replicates.

Bimodality. Figure 5 displays the boxplot for bias reported at each simulation replication for the first-class mean of Z, under the scenario small sample size ($n = 500$), moderate separation, direct effects on three indicators scenario and non-normal (bimodal) within class distribution of Z.

Both distal-outcome approaches do well under the bimodality condition, with the 2-step RC following up close and the 1-step RC and the 3-step BCH approaches doing slightly worse. Results regarding the 3-step BCH approach are not surprising, since all three-step approaches in LCA are known to perform poorly when there are direct effects of the external variable on the indicators (Bakk & Vermunt, 2016; Vermunt, 2010).

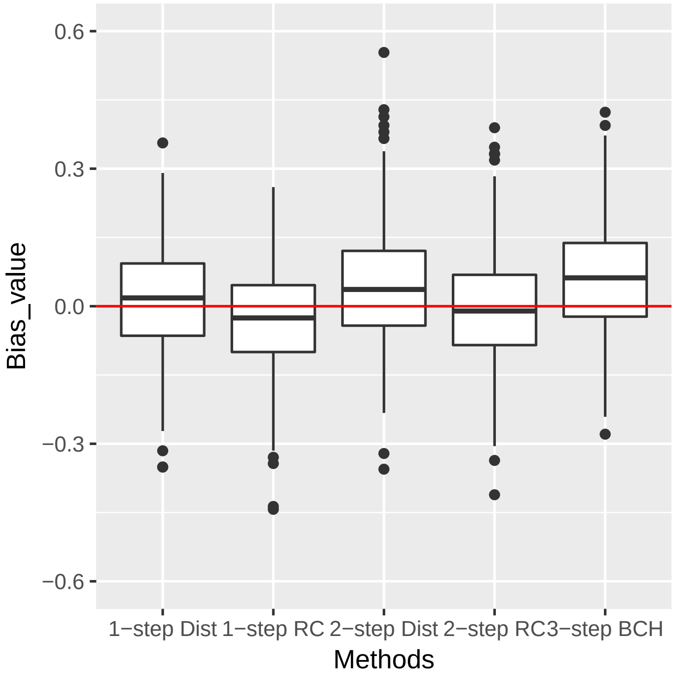


Figure 4. Boxplots of bias in the first estimated mean of Z for the non-normal (excess kurtosis) scenario, for the one-step LC with Random Covariates (“1-step RC"), the one-step LC with distal outcome (“1-step dist"), the two-step LC with Random Covariates (“2-step RC"), the two-step LC with distal outcome (“2-step dist"), and the three-step BCH approach (“3-step BCH") for $n = 500$, moderate class separation and three direct effect scenario. 500 Monte Carlo replicates.

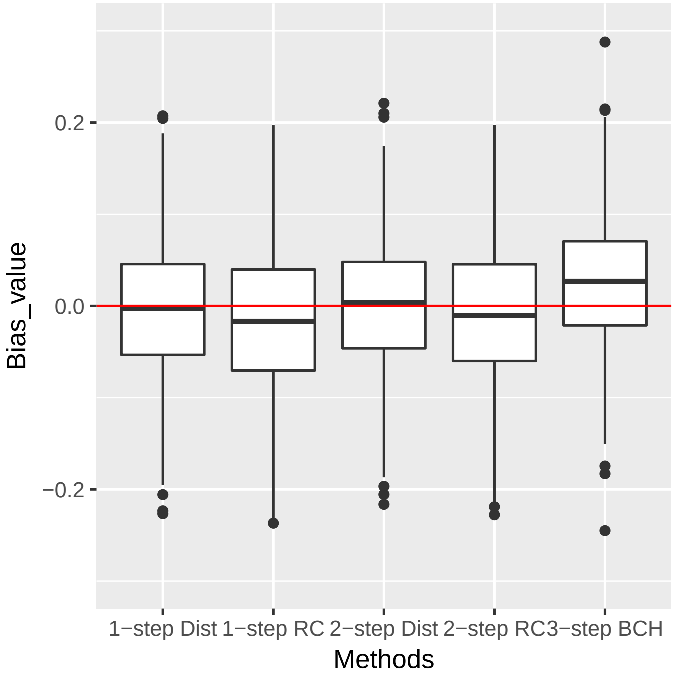


Figure 5. Boxplots of bias in the first estimated mean of Z for the non-normal (bimodality) scenario, for the one-step LC with Random Covariates (“1-step RC"), the one-step LC with distal outcome (“1-step dist"), the two-step LC with Random Covariates (“2-step RC"), the two-step LC with distal outcome (“2-step dist"), and the three-step BCH approach (“3-step BCH") for $n = 500$, moderate class separation and three direct effect scenario. 500 Monte Carlo replicates.

A latent class model of households' assets ownership to predict wealth

Household wealth cannot be directly observed. Nonetheless, measuring it is a crucial issue for any policy maker. Notably, survey measures of income - or expenditure, if available - are affected by substantial measurement error and systematic reporting bias (Ferguson et al., 2003; Moore & Welniak, 2000). In addition, if wealth as a measure of permanent income (Friedman, 1957) is of interest, current income, even if measured without error, is likely to be a poor approximation. In more recent surveys (like the Household Finance and Consumption Survey, from the European Central Bank), a measure of net wealth - value of total household assets minus the value of total liabilities - is provided. However, relying on each household's subjective evaluation of the current value of each asset they own, such a measure is prone to considerable measurement error as well. Furthermore, having such a complex measure is complicated to understand for a more general audience, to whom a simpler classification/index would appeal.

Latent Trait (LTA) and Latent Class Analysis have been used in order to model wealth - or, inversely, deprivation - from observed assets ownership. Whereby in the LTA framework, wealth is modeled as a continuous trait (Szeles & Fusco, 2013; Vandemoortele, 2014), LCA was used (Moisio, 2004; Pérez-Mayo, 2005) based on the idea that wealth (poverty) can be seen as a multidimensional latent construct. Although determining which ownership indicators to use can be a problem, arguably they all attempt to identify subgroups in the population based on the same multidimensional phenomenon (Moisio, 2004): different dimensions of wealth are measured by different (sets of) indicators. In addition, within an LCA framework, response probabilities can be used to evaluate *ex post* each indicator's validity of measuring the (latent variable) wealth.

We analyze data from the first wave of the Household Finance and Consumption Survey, conducted by the European Central Bank. We focus on a sample of Italian households, for which we have information on real and financial assets, liabilities, different income measures, consumption expenditures, and a measure of total wealth in euro. The latter is defined as total household assets value, excluding public and occupational pension wealth, minus total outstanding household's liabilities (Household Finance and Consumption Network, 2013). The value of each asset is provided by asking to the interviewees how much they think each asset is worth. For instance, related to the item "owning any car" (HB4300, Table 4), the interviewer asks "For the cars that you/your household own, if you sold them now, about how much do you think you could get?" (HFCS Core Variables Catalogue, 2013). In fact, different households might have very different (and possibly wrong) perceptions about the value of the assets they own. This is why we cannot rely only on DN3001 as a measure of households wealth, but rather use a model known for its strengths in correcting for measurement error using multiple indicators.

We selected a set of 10 items related to a financial type of wealth, and we included also 4 items related to a broader type of wealth, for a total of $J = 14$ items. The variable concerning household residence tenure status (HB0300) was recoded to have "entirely owned main residence" or "partially owned main residence" merged into one category. The resulting variable on tenure status (hometen) has 3 categories, and enters all models with dummy coding.

We restrict ourselves to analyze only households having positive wealth. Doing so allows us to set up a model for log-wealth rather than for wealth, as is commonly done in the economic literature studying elasticities (see, for instance, Charles & Hurst, 2003). Imposing this restriction leads us to drop only 188 sample units (about 2% of the total sample).

The one-step LC with Random Covariates ("1-step RC"), the one-step LC with distal outcome (“1-step dist"), the two-step LC with Random Covariates (“2-step RC") and the two-step LC with distal outcome (“2-step dist") models are estimated using Latent GOLD 5.1 (Vermunt & Magidson, 2016) in combination with R[1]. The model comparison is done with the purpose of investigating how cluster membership is able to predict classes of (log) wealth. The random covariate modeling approach does so by relaxing the conditional independence assumption and allowing for possible direct effects of log-wealth on the indicators.

Name	Description	Type
HB0300	Household main residence - Tenure status (1 if entirely owned) (2 if partially owned) (3 if rented) (4 if for free use)	Nominal
hometen	Household main residence - Tenure status (1 if entirely or partially owned) (2 if rented) (3 if for free use)	Nominal
HB2400	Household owns other properties	Dichotomous
HB4300	Household owns any car	Dichotomous
HB4700	Ownership of other valuables	Dichotomous
HC0200	Household has a credit line or overdraft	Dichotomous
HC0300	Household has a credit card	Dichotomous
HC0400	Household has a non collaterized loan	Dichotomous
HD0100	Household has any investment in business	Dichotomous
HD1100	Household owns a sight account	Dichotomous
HD1200	Household owns a savings account	Dichotomous
HD1300	Household has any investment in mutual fund	Dichotomous
HD1400	Household owns bonds	Dichotomous
HD1500	Household owns managed accounts	Dichotomous
DN3001	Net Wealth	Continuous

Table 4
Variables list, with description and type. hometen obtained by recoding HB0300 into 3 categories, where “partially owned" and “entirely owned" are merged into one. DN3001 is obtained as total household assets, excluding public and occupational pension wealth, minus total outstanding household’s liabilities. We focus on a sample of observations with positive net wealth, and work with its logarithmic transformation. Further details on variables definitions are available from the ECB website.

In order to avoid distortions in model selection due to the distributional assumption on log-wealth, we select the number of classes based on a simple LC model with items only (first step model). In Table 6 we display a line chart of BIC values (y axis) vs. number of classes (x axis), from 1 to 9. We observe that the curve reaches a minimum only at $G = 8$[2], but there seem to be an elbow at $G = 2$. This is somewhat consistent with previous literature on LCA for measuring wealth (Moisio, 2004; Pérez-Mayo, 2005).

[1] Sample LG syntax and R code is available for each model from the corresponding author upon request.

[2] This is due to a ceiling effect, as for a number of classes greater than 8 also BIC increases.

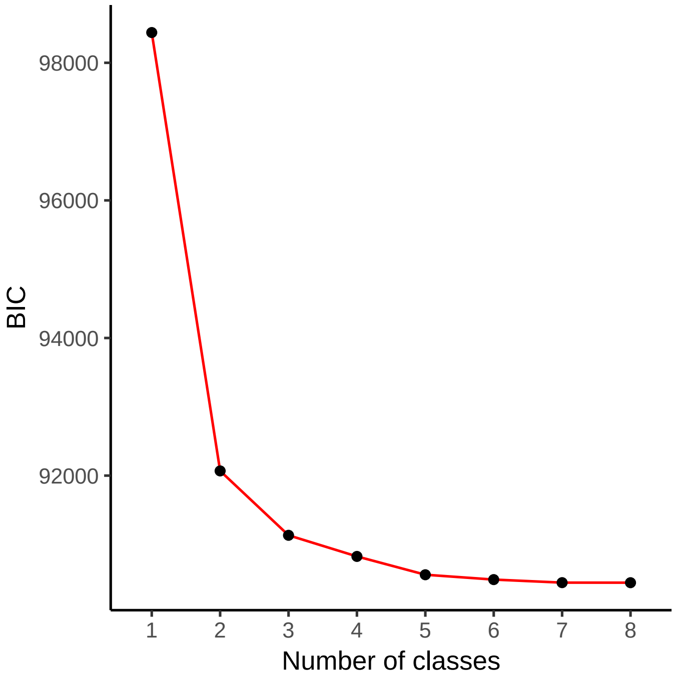


Figure 6. BIC values of the first step model for 1 to 9 latent classes.

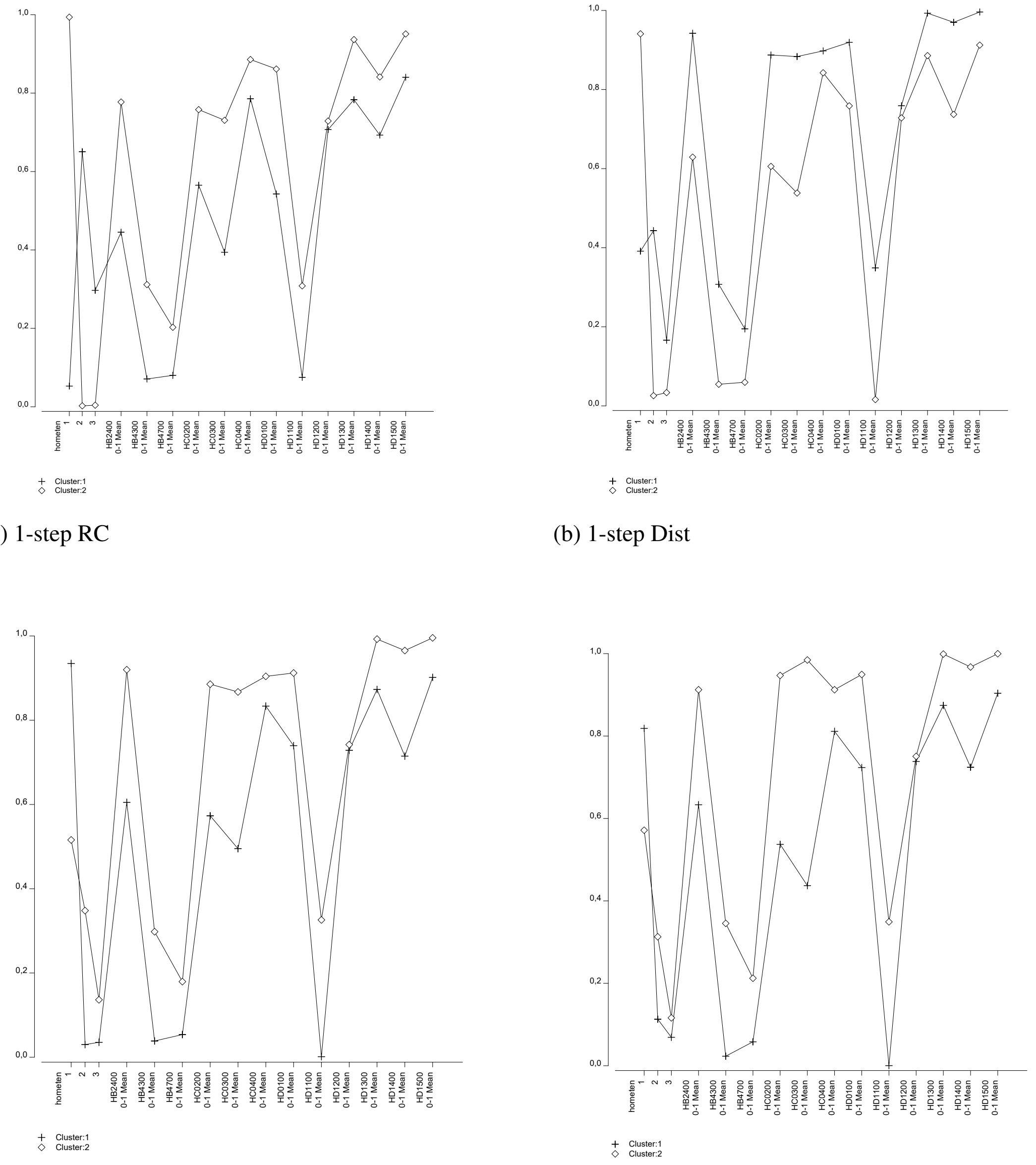
We report (Figure 7) the probability profile plot for 1-step RC (Figure 7a), 1-step Dist (Figure 7b), 2-step RC (Figure 7c) and 2-step Dist.

We observe that the one-step RC approach predicts a wealthier class, with higher ownership probabilities for all items, and a higher probability of owning (partially or entirely) the household main residence - relative to “free use" and “rent" categories. One-step dist, two-step dist and two-step RC tend to agree overall on class ownership probabilities. We observe that the second latent class has higher ownership probabilities than the one-step RC’s second class. Interestingly, they all predict less wealthy HH’s (first latent class) to have on average a higher probability of owning their home. That is, they predict that if a HH has relatively little wealth this is invested to buy the place the HH lives in. We note that also the two–step approaches both predict home ownership probability greater than 0.5 also for the wealthier class.

Figure **??** shows how the class composition of the 1-step and 2-step RC and dist predict classes in log-wealth. The first class seems to capture observations with lower log-wealth, although - in 1-step dist and the two-step approaches - the relatively fat tails in the second class allocate non-zero density also to few of the wealthiest observations. Consistently with findings in the probability profiles, classes in the 1-step RC separate households on (log) wealth more neatly.

To gain further insights on the estimated wealth distribution, we report (Table 5) estimated means and variances of log-wealth for all four approaches.

Inference on the estimated parameters of the log-wealth (p-values of all tests are below 0.01) validates the model assumption of clustered distribution, with heteroscedastic components, of log-wealth in all approaches. The mean in the second class of wealth is larger for all approaches. As we observed above, the first class with lower average log-wealth - but larger variance



(a) 1-step RC

(b) 1-step Dist

(c) 2-step RC

(d) 2-step Dist

Figure 7. Probability profile plot for the one-step LC with Random Covariates (“1-step RC"), the one-step LC with distal outcome (“1-step dist"), the two-step LC with Random Covariates (“2-step RC") and the two-step LC with distal outcome (“2-step dist"). For the variable “hometen", the levels indicate the four category ownership probabilities within each wealth class. Similarly, levels refer to average item ownership probability in wealth classes for the remaining (dichotomous) items.

- in 1-step Dist and 2-step RC and Dist absorbs also some households in the right tale of the distribution.

Finally, Table 6 demonstrates that both 1-step and 2-step RC provide an easy way to test

	Means		Wald(=) p	Variances		Wald(=) p
1-step dist	10.0835***	12.6173***	0.0000	3.0910	0.5573	0.0000
	(0.0451)	(0.0129)		(0.0843)	(0.0137)	
1-step RC	9.4489***	12.4928***	0.0000	2.5960	0.6304	0.0380
	(0.0391)	(0.0112)		(0.0865)	(0.0131)	
2-step dist	10.3851***	12.5471***	0.0000	3.4766	0.7900	0.0000
	(0.0345)	(0.0146)		(0.0874)	(0.0212)	
2-step RC	10.3134***	12.6548***	0.0000	3.2234	0.5545	0.0380
	(0.0356)	(0.0128)		(0.0807)	(0.0151)	

Table 5

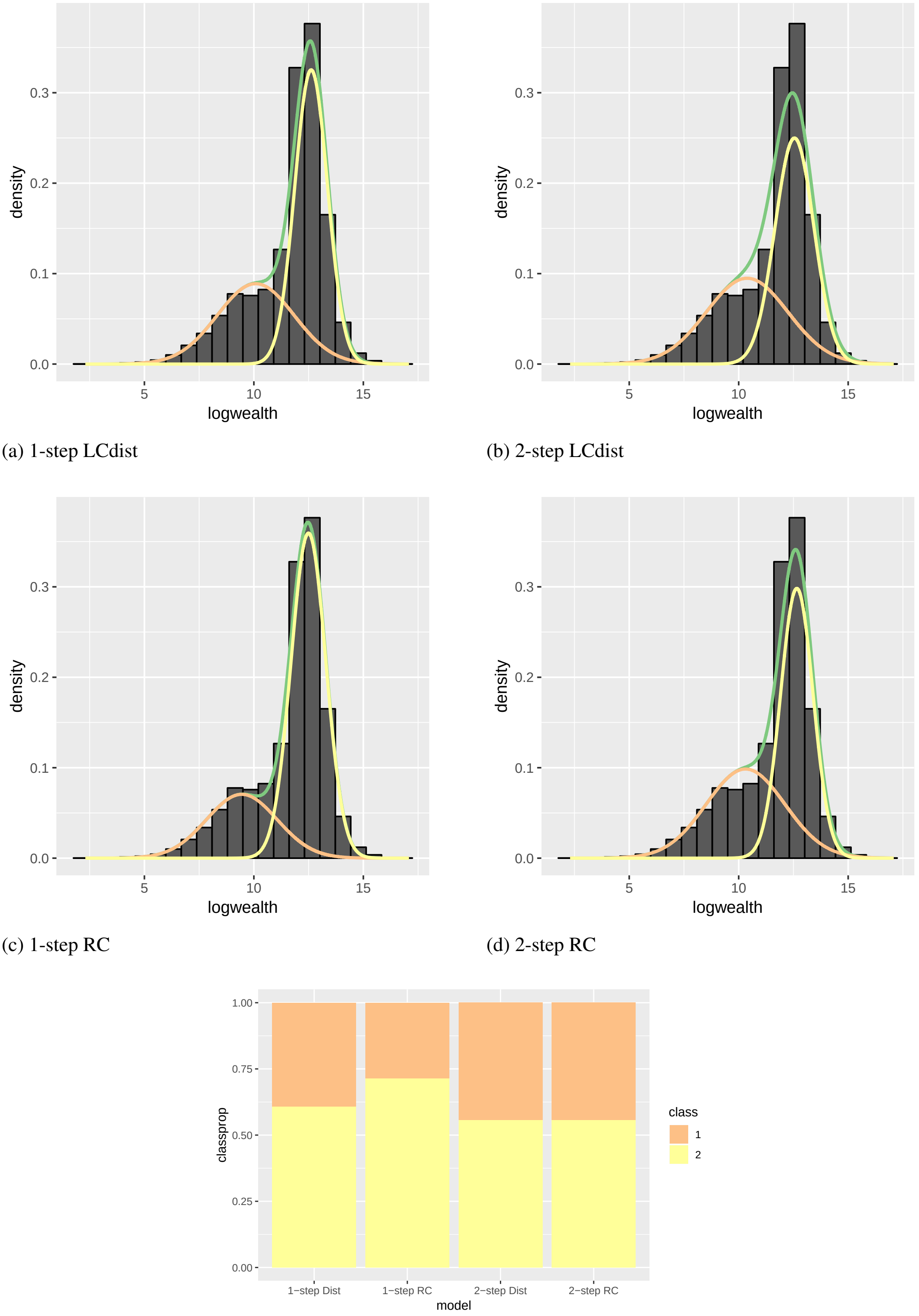
*Estimated means (*** p-value<0.01, ** p-value<0.05, * p-value<0.1) and variances of log-wealth, and p-values from Wald test of equality of component means and variances for the one-step LC with Random Covariates (“1-step RC"), the one-step LC with distal outcome (“1-step dist"), the two-step LC with Random Covariates (“2-step RC") and the two-step LC with distal outcome (“2-step dist"). Standard errors in parentheses.*

whether there are significant direct effects of log-wealth on the response variables. Interestingly, inference on such effects points out that there are variables for which we have no significant direct effects of log-wealth (hometen), as well as indicating that some effects are the same across latent classes (HB2400, HD1300, HD1400 and HD1500). In the logic of an applied researcher, this suggests intermediate and more parsimonious modeling options where effects on some variables can be constrained to be zero, or to be the same across classes. This can easily be discovered with a random covariate modeling approach, especially with the more practical two-step approach that perfectly serves the need of refitting multiple times the model to compare different specifications without having to re-estimate the full model from scratch every time.

	1-step RC				2-step RC			
Log-wealth on	Coefficients		Wald(0) p	Wald(=) p	Coefficients		Wald(0) p	Wald(=) p
	Class 1	Class 2			Class 1	Class 2		
hometen(2)	-1.8014***	1.8844***	0.0000	0.0000	-0.1278***	0.0180***	0.0000	0.0000
(Main res. tenure stat)	(0.0663)	(0.6175)			(0.0088)	(0.0041)		
hometen(3)	-1.5633***	0.9826**	0.0000	0.0000	-0.0690***	0.0226***	0.0000	0.0000
(Main res. tenure stat)	(0.0669)	(0.4042)			(0.0073)	(0.0055)		
HB2400	-0.9990***	-1.6839***	0.0000	0.0000	-0.0103***	0.0083	0.0003	0.0240
(HH owns other properties)	(0.0454)	(0.0788)			(0.0026)	(0.0075)		
HB4300	-0.6239***	-1.3334***	0.0000	0.0000	0.0449***	-0.0187***	0.0000	0.0000
(HH owns any car)	(0.0362)	(0.0729)			(0.0077)	(0.0039)		
HB4700	-0.3953***	-0.7162***	0.0000	0.0001	-0.0068	-0.0180***	0.0000	0.1700
(HH owns other valuables)	(0.0310)	(0.0761)			(0.0060)	(0.0046)		
HC0200	-0.5636***	-1.0027***	0.0000	0.0000	0.0126***	-0.0715***	0.0000	0.0000
(HH has a credit line)	(0.0286)	(0.0688)			(0.0026)	(0.0061)		
HC0300	-0.8510***	-1.3491***	0.0000	0.0000	0.0201***	-0.1927***	0.0000	0.0000
(HH has a credit card)	(0.0388)	(0.0732)			(0.0026)	(0.0061)		
HC0400	-0.2018***	-0.3100***	0.0000	0.3000	0.0131***	-0.0085	0.0003	0.0035
(HH has a non-coll. loan)	(0.0258)	(0.0968)			(0.0034)	(0.0062)		
HD0100	-0.5305***	-1.3817***	0.0000	0.0000	0.0070**	-0.0504***	0.0000	0.0000
(HH has any invest. in business)	(0.0312)	(0.0919)			(0.0029)	(0.0068)		
HD1100	-0.8756***	-1.8493***	0.0000	0.0000	0.1566*	-0.0089**	0.0280	0.0840
(HH owns a sight account)	(0.0411)	(0.0887)			(0.0950)	(0.0038)		
HD1200	-0.1533***	-0.0437	0.0000	0.0710	-0.0042	-0.0040	0.1700	0.9700
(HH owns a savings account)	(0.0228)	(0.0529)			(0.0028)	(0.0041)		
HD1300	-0.8373***	-1.3818***	0.0000	0.0015	-0.0011	-0.1565***	0.0000	0.0000
(HH has invest. in mutual funds)	(0.0549)	(0.1467)			(0.0037)	(0.0232)		
HD1400	-0.7800***	-1.0292***	0.0000	0.0150	-0,0042	-0,0057	0.2600	0.8900
(HH owns bonds)	(0.0456)	(0.0781)			(0.0028)	(0.0112)		
HD1500	-0.8652***	-1.4764***	0.0000	0.0023	-0.0021	-0.2039***	0.0000	0.0000
(HH owns managed accounts)	(0.0607)	(0.1732)			(0.0041)	(0.0316)		

Table 6

*Estimated direct effects of log-wealth on each variable per latent class (*** p-value<0.01, ** p-value<0.05, * p-value<0.1), p-values from Wald test of joint equality of each variable's direct effect to zero - Wald(0) - and from Wald test of equality of effects across latent classes - Wald(=) - for the 1-step and 2-step RC approaches. Short description below variable names. HH stands for Household. Standard errors in parentheses.*



(a) 1-step LCdist

(b) 2-step LCdist

(c) 1-step RC

(d) 2-step RC

(e) Class proportions

Figure 8. Density plots of log-wealth as predicted by the four competing approaches.

Conclusion

The focus of this paper has been to motivate the use of the random covariate modeling approach for distal outcome models with possible misspecifications of the distribution of Z, and direct effect between the indicators and Z. To do so, we have introduced a feasible two-step estimator. A simulation study has been used to illustrate this idea and investigate the final sample properties of both one-step and two-step estimators, and also the one-step and two-step distal outcome models have been included for comparison. Interestingly, we found that 1) the random covariate approach yields the lowest bias, and 2) the 2-step RC approach is the least sensitive to misspecification of the distribution of the external variable. The actual advantage of random covariate modeling was also showed through an application on Italian household asset ownership data.

In the applied researcher perspective, the proposed approach has several advantages. First, it allows to deal with direct effects if they are of substantive interest, as well as if those represent a source of noise to be handled. That is, if direct effects are present, our approach, contrary to the distal outcome model, yields unbiased estimates of the distal outcome cluster specific means and variances. Second, it guarantees a safe and flexible option, in which one starts from the most general model. Then, proceeding backwards, the user can test the model assumptions of the distal outcome and the latent class regression models. We propose to applied researchers to use our approach in any situation where the distal outcome is supposed to have DE on the indicators and also the class specific distribution of the continuous distal outcome is not normal. In situations where such complex interdependencies are not present we recommend using one of the stepwise bias-adjusted approaches that cannot model DEs using for example the overview and recommendation proposed by Nylund-Gibson et al. (2019).

The approach we suggest has some limitations as well. First, it can be unstable if some of the observed response patterns are lacking, or are simply too small in number to estimate the model parameters (see for instance, for standard sufficient conditions for identifiability of LCA, Bandeen-Roche et al., 1997). Especially in an exploratory stage of the analysis, in such cases simpler models can be more attractive. Second, depending on the goal of the analysis, the random covariate modeling approach can generate a final output which might be harder to interpret than that of simpler models. In other words, the final model interpretability relies on the final goal of the analysis, which must be clear in mind in this as well as in any other modeling approach. Third, we found that the SE correction does not work well for saturated models like the random covariate model. How to handle such cases can be an interesting subject for future research.

While in the current paper we proposed the random covariate approach in the context of distal outcome models, the extension to models with covariates, or more complex models involving both covariates and distal outcomes is straightforward from equation 12. Currently in LCA research it is common practice to use either co-variate or distal outcome model, but hardly ever the combination of the two because of the technical limitations of classical approaches when potentially many external variables are present. In such cases, classical approaches might fail because of the sparseness of the analyzed frequency table and the potentially large number of parameters (Goetghebeur et al., 2000; Huang & Bandeen-Roche, 2004; Clark & Muthén, 2009). This practice is very different from the general structural equation modeling mindset, where continuous latent variables are included in complex models both as antecedents of some distal outcomes possibly predicting the outcome in conjunction with other (latent and/or observed) predictors, and having predictors of their own. These type of extensions can be especially practical using the proposed two-step estimator: while keeping the measurement model fixed the structural model can be extended gradually, without the need of re-estimating the full model, or the risk that the measurement model changes due to misspecifications introduced in the estimation of the structural model. The extension to this more complex models is not trivial and warrants further research, to better understand for example how robust the approach is in the presence of multiple possibly correlated distal outcomes, direct effects etc. It will be important to see how misspecifications in one part of a complex model affect the full model, and what model selection approaches work the best. Further research on these more complex extensions is needed.

References

Amemiya, T. (1985). *Advanced econometrics*. Harvard University Press.

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

Azzalini, A., & Capitanio, A. (2014). *The skew-normal and related families*. Cambridge University Press. Retrieved from `https://books.google.it/books?id=-tkaAgAAQBAJ`

Bakk, Z., & Kuha, J. (2017). Two-step estimation of models between latent classes and external variables. *Psychometrika*, 1–22.

Bakk, Z., Tekle, F. T., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, 272-311.

Bakk, Z., & Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 20–31.

Bandeen-Roche, K., Miglioretti, D., & Zegger, P., S.L. Rathouz. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*, 1375-86.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implication for overextraction of latent trajectory classes. *Psychological methods*, *8*(3), 338-363.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*, 3-27.

Bollen, K. A. (1996). An alternative 2sls estimator for latent variable models. *Psychometrika*, *61*, 109-121.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: Two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods and Research*, *36*, 48-86.

Bouwmeester, S., & Sijtsma, K. (2007). Latent class modeling of phases in the development of transitive reasoning. *Multivariate Behavioral Research*, *42*(3), 457–480.

Charles, K. K., & Hurst, E. (2003). The correlation of wealth across generations. *Journal of Political Economy*, *111*(6), 1155–1182.

Clark, S. L., & Muthén, B. (2009). Relating latent class analysis results to variables not included in the analysis. *Retrieved June 16, 2012 (http://statmodel2.com/download/relatinglca.pdf)*.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). Wiley.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent class models. *Journal of the American Statistical Association*, *83*, 173-178.

Di Mari, R., & Bakk, Z. (2017). Mostly harmless direct effects: a comparison of different latent markov modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal*.

Ferguson, B. D., Tandon, A., Gakidou, E., & Murray, C. J. (2003). Estimating permanent income using indicator variables. *Health systems performance assessment: debates, methods and empiricism. Geneva: World Health Organization*, 747–760.

Finch, W. H., & French, B. F. (2015). Modeling of nonrecursive structural equation models with categorical indicators. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 416-428.

Friedman, M. (1957). *A theory of the consumption function*. Princeton: Princeton University Press.

Goetghebeur, E., Liinev, J., Boelaert, M., & Van der Stuyft, P. (2000). Diagnostic test analyses in search of their gold standard: latent class analyses with random effects. *Statistical Methods in Medical Research*, *9*(3), 231–248.

Gong, G., & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, 861–869.

Goodman, L. A. (1974). 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*, 79-259.

Hagenaars, J. A. (1990). *Categorical longitudinal data- loglinear analysis of panel, trend and cohort data*. Newbury Park, CA:Sage.

Hagenaars, J. A., & Halman, L. C. (1989). Searching for idealtypes: the potentialities of latent class analysis. *European Sociological Review*, *5*(1), 81–96.

Household Finance and Consumption Network. (2013). *The eurosystem household finance and consumption survey-methodological report* (Tech. Rep.). ECB Statistics Paper.

Huang, G.-H., & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, *69*(1), 5–32.

Ingrassia, S., Minotti, S., & Vittadini, G. (2012). Local statistical modeling via a cluster–weighted approach with elliptical distributions. *Journal of Classification*, *29*, 363–401.

Ingrassia, S., Minotti, S. C., & Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics & Data Analysis*, *71*, 159–182.

Ingrassia, S., & Punzo, A. (2016). Decision boundaries for mixtures of regressions. *Journal of the Korean Statistical Society*, *45*(2), 295–306.

Ingrassia, S., Punzo, A., Vittadini, G., & Minotti, S. C. (2015). The generalized linear mixed cluster-weighted model. *Journal of Classification*, *32*(1), 85–113.

Janssen, J. H., van Laar, S., de Rooij, M. J., Kuha, J., & Bakk, Z. (2019). The detection and modeling of direct effects in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 280-290.

Kim, M., Vermunt, J. K., Bakk, Z., Jaki, T., & Van Horn, M. L. (2016). Modeling predictors of latent classes in regression mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 601-614.

Lanza, T. S., Tan, X., & Bray, C. B. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling*, *20*(1), 1-26.

Loken, E. (2004). Using latent class analysis to model temperament types. *Multivariate Behavioral Research*, *39*(4), 625–652.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2005). Latent-trait latent-class analysis of self-disclosure in the work environment. *Multivariate Behavioral Research*, *40*(4), 435–459.

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*, 180-197.

McCutcheon, A. L. (1985). A latent class analysis of tolerance for nonconformity in the american public. *Public Opinion Quarterly*, *49*(4), 474–488.

McCutcheon, A. L. (1987). *Latent class analysis.* Newbury Park, CA:Sage.

Moisio, P. (2004). A latent class application to the multidimensional measurement of poverty. *Quality & Quantity*, *38*(6), 703–717.

Moore, J. C., & Welniak, E. J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics*, *16*(4), 331.

Mulder, E., Vermunt, J., Brand, E., Bullens, R., & Van Merle, H. (2012). Recidivism in subgroups of serious juvenile offenders: Different profiles, different risks? *Criminal Behaviour and Mental Health*, *22*, 122–135.

Nylund-Gibson, K., Grimm, R. P., & Masyn, K. E. (2019). Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models. *Structural Equation Modeling: A Multidisciplinary Journal.*

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 782-797.

Okazaki, S., Campo, S., Andreu, L., & Romero, J. (2015). A latent class analysis of spanish travelers' mobile internet usage in travel planning and execution. *Cornell Hospitality Quarterly*, *56*(2), 191–201.

Pérez-Mayo, J. (2005). Identifying deprivation profiles in spain: a new approach. *Applied Economics*, *37*(8), 943–955.

Petras, H., & Masyn, K. (2010). General growth mixture analysis with antecedents and consequences of change. In A. Piquero & D. Weisburd (Eds.), (p. 69-100). Springer, New York.

Punzo, A., & Ingrassia, S. (2016). Clustering bivariate mixed-type data via the cluster-weighted model. *Computational Statistics*, *31*(3), 989–1013.

Roberts, T. J., & Ward, S. E. (2011). Using latent transition analysis in nursing research to explore change over time. *Nursing research*, *60*(1), 73–79.

Roedelof, A. J. M., Bongers, I. L., & van Nieuwenhuizen, C. (2013). Treatment engagement in adolescents with severe psychiatric problems: a latent class analysis. *European Child and Adolescent Psychiatry*, *22*(8), 491–500.

Stegmann, G., & Grimm, K. J. (2017). A new perspective on the effects of covariates in mixture models. *Structural Equation Modeling: A Multidisciplinary Journal.*

Szeles, M. R., & Fusco, A. (2013). Item response theory and the measurement of deprivation: evidence from luxembourg data. *Quality & Quantity*, 1–16.

Vandemoortele, M. (2014). Measuring household wealth with latent trait modelling: an application to malawian dhs data. *Social Indicators Research*, *118*(2), 877–891.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450-469.

Vermunt, J. K., & Magidson, J. (2016). Technical guide for latent gold 5.1: Basic, advanced, and syntax. *Belmont, MA: Statistical Innovations Inc.*.

Zhu, Y., Steele, F., & Moustaki, I. (2017). A general 3-step maximum likelihood approach to estimate the effects of multiple latent categorical variables on a distal outcome. *Structural Equation Modeling: A Multidisciplinary Journal.*