Comparing Frequentist and Bayesian latent class analyses approaches

Waesberghe, Evianne van

Comparing Frequentist and Bayesian Latent Class Analysis Approaches

Evianne van Waesberghe

Master's Thesis Psychology,

Methodology and Statistics Unit, Institute of Psychology

Faculty of Social and Behavioral Sciences, Leiden University

Date: December 2022

Supervisor: Zsuzsa Bakk

Abstract

Psychological latent class analysis research is based on the Frequentist approach stating that conclusions are only based on the data that is used in that particular dataset. In political research the Bayesian approach is fairly common, which uses the information from previous research and the current dataset to base their conclusions on. Here, the Frequentist and Bayesian approaches are compared using a latent class model with a distal outcome in a simulation and application study, based on their bias on strength of the predictors and strength of the outcome variable. Multiple software programs are available for doing LCA, such as Mplus and LatentGold, which are often used in research. Various R-packages also offer the possibility of doing LCA, but with a limited number of approaches available. The aim of this study is to use the poLCA- and BayesLCA-packages to compare the Frequentist and Bayesian approaches to evaluate the estimations of the strength of the predictors, the strength of the distal outcome and the class sizes. Previous research has shown that the distal outcome in the naïve three step-approach can influence the class sizes, but it is not yet known if this problem also occurs in the Bayesian approaches. The simulation study results show that the Variational Bayes and three-step approach from the poLCA- and BayesLCA-packages give very similar results on strengths of the predictors, strength of the distal outcome and class size. The Gibbs Sampling method shows a better performance in estimating the strength of the distal outcome for smaller sample sizes, but performs worse in estimating the strength of the predictors. In the application study, the three step-approaches from the poLCA- and BayesLCA-packages again perform almost identical, but the Variational Bayes shows (very) different estimations on class sizes and multiple indicator variables. The Gibbs Sampling method is relatively close to the three-step approach estimations. All four methods show the same strength for the distal outcome variable.

Inhoudsopgave

1. Introduction

Latent class (LC) analysis is a method widely used in the social and behavioural sciences to group individuals based on their responses on a set of observed variables (Goodman, 1974). Often in LC analysis applications the interest lies not only in obtaining a clustering, but also in determining whether the classes differ with respect to one or more, possibly continuous, distal outcome variables. Examples for this can be found in Mulder et al. (2012) where juvenile offender profiles were related to more than 80 outcomes and research from Quirk, Nylund-Gibson and Furlong (2013) in which latent classes of school readiness were linked to later academic outcomes of children. An example in the political science field can be found in Bonikowksi & DiMaggio (2016), where different groups of American popular nationalism were linked to attitudes towards ethnic minorities, immigration and national sovereignty.

Many LCA-methods have been created, each with their pros and cons in classifying people into groups. Psychology and Political science research both use LCA methods to group people based on their responses, but they have different schools of thought on this. Psychological research is commonly based on the frequentist approach, which means that conclusions are based only on the data that the researcher is currently using. This approach begins research with the assumption that the null hypothesis is true (before even collecting the data) and then deciding if the collected data is unlikely based on this null hypothesis. If not likely, the null hypothesis will be rejected since it does not explain the data adequately. Hard conclusions can only be made when research has been done repeatedly on different datasets. The frequentist approach does not use prior probabilities since it makes the analysis subjective and less accurate in their view (Fornacon-Wood et al., 2022).

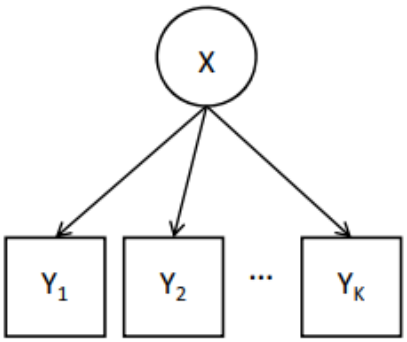
In Bayesian statistics the probabilities are based on the data that is currently used and information from previous research and/ or historical data, also known as the priors. These probabilities are continuously updated as new information is added to the already available prior distribution, creating the posterior distribution. Choosing your prior is therefore an important and delicate task, since a poorly chosen prior can steer your results in the wrong direction or it can be used to create a falsely positive result. Another option is using uninformative priors, which might be preferable when there is little prior knowledge and/ or when you do not want to rely on subjective beliefs. Uninformative priors are thus used to make the Bayesian inferences as objective as possible. However, even uninformative priors still contain some information that might lead the data to a certain direction, so the term uninformative prior should not be taken literally. A good uninformative prior has a negligible contribution with respect to that provided by the data. There are multiple types of uninformative priors, but the most commonly used one is the uniform prior, which gives each value of θ the same

prior probability. Here you can also decide between a closed interval or an infinite number of possible θ-values (Tobago, 2021). In my simulation study I will use uninformative priors, since these are commonly used in Bayesian statistics and provide the most honest way to compare the outcomes to that of the frequentist approach, since the frequentist approach does not use priors in their analysis.

1.1. Defining the LCA-model

When defining a model, we first create the measurement model, also known as the basic latent class model. A measurement model gives the class-specific probability of a pattern of responses to the indicators. Here we look for class-specific response patterns and their probabilities that could indicate the existence of different groups within the data before adding covariates and/or distal outcome variables to the model. LCA assumes that respondents belong to one of the *T* classes of an underlying latent variable X which affects the responses (Goodman, 1974; McCutcheon, 1987).

Figure 1. **Measurement model**



Note: extracted from Bakk & Kuha (2020).

The number of classes is selected by comparing different goodness of fit models with different numbers of classes using model selection. Typically, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are used for this, choosing the model with the lowest AIC and/or BIC values as the best fitting model. The number of classes in the data is held constant after deciding what the optimal number of classes is in the data. If not held constant, the number of classes might change when adding covariates or distal outcomes to your model in the next steps (Masyn, 2017). We can then add covariates and/or distal outcomes to create the structural model. The structural model gives the unconditional probability of belonging to latent class *t.* There are many different combinations of covariates and distal outcomes that are possible in structural models, but most often only either covariates or distal outcomes are included. In this research, a latent class model with one

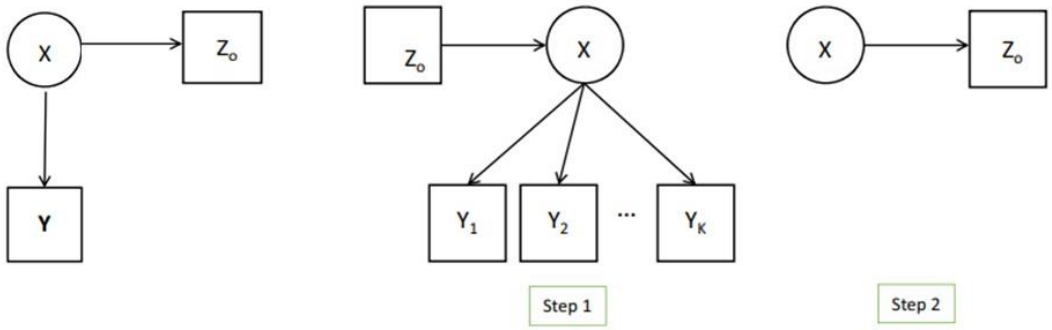
distal outcome will be used. Another statistic that is also looked at when checking the AIC and BIC, but not used for model selection (Masyn, 2013; Muthén, 2008), is the entropy-value. The entropy measures how accurately one could assign individuals to classes, based on the uncertainty in the assignments. There is more certainty about the class membership when the probability of belonging to a class is near 0 or 1, which implies that there will be few errors of assignment. When the probabilities move away from 0 and 1, this means that there is more uncertainty about class membership. The entropy summarises this uncertainty of class membership based across all individuals, showing how accurately individuals can be assigned in the model (Curran & Bauer, 2021). Muthén (2008) indicates that an entropy-value of .8 can be considered as good, but that it is difficult to specify which value indicates bad entropy. However, a low entropy value does not necessarily mean that the number of classes is bad or that the model does assign all individuals poorly. A lower entropy might be caused by certain groups showing less than ideal probabilities for belonging to certain classes (e.g., 40% probability group 1, 60% probability group 2), while other groups have probabilities near 0 or 1 for belonging to a certain group. Next to calculating the entropy-value, it is therefore useful to also investigate the classification table to see if the whole model shows class assignment issues or if this is only the case for certain groups (Masyn, 2013; Muthén, 2008). Another possible reason for having a low entropy-value, might be due to poor selection of indicator variables, with variables not differentiating well enough between groups. If classes show similar values on an indicator variable, it might be wise to remove it or replace it with a new indicator variable based on literature (Muthén, 2008).

1.2. Frequentist approach

In the literature, many frequentist methods and overviews can be found for latent class analysis (Nylund-Gibson et al., 2019; Bakk & Kuha, 2020; Vermunt, 2010). One of the original approaches is the one-step approach by Bandeen-Roche, Miglioretti, Zeger & Rathouz (1997), which fits the whole model at once, thus estimating both its measurement and structural model at the same time. This estimation of the full model and its standard errors are obtained using the standard maximum likelihood estimation. For models with a distal outcome, this results in an analysis where the distal outcome variable serves as an indicator variable, since the relationship between the latent class variable, the indicator variables and the class-specific distal outcome distributions are estimated simultaneously. No separation is thus made between these types of variables in the analysis and because of this, the class-specific distribution of the outcome cannot be interpreted as being a direct effect of being in a certain class (Nylund-Gibson et al., 2019). Another disadvantage of the one-step approach is that the inclusion of the distal outcome in the model can alter the meaning of the classes, since there is no separation of measurement and structural model estimation (Bakk & Vermunt, 2016). Alternative

methods were created to prevent the problems that are present in the one step approach. An example of this, is the LTB approach created by Lanza, Tan & Bray (2013), which is a two-step approach. In the first step, a latent class model is estimated in which the distal outcome is included as a covariate instead of a response variable. In the second step, the class specific means for the distal outcome are calculated, based on the estimates from step one. The LTB approach is able to perform well when the distal outcome is normally distributed with different means and equal variances across classes, resulting in unbiased estimations of the class-specific means. However, the LTB approach performs poorly when the relationship between *X* and the distal outcome are not linear-logistic, resulting in biased estimation of the class-specific means (Asparouhov & Muthén, 2014; Bakk & Vermunt, 2016). This approach also does not offer the possibility to use multiple distal outcomes at the same time. It is possible to work around this by repeating the analysis for every separate distal outcome, but this does not guarantee that the latent class solution will be the same for every analysis (Bakk & Vermunt, 2016)

Figure 2. One-step approach and LTB approach



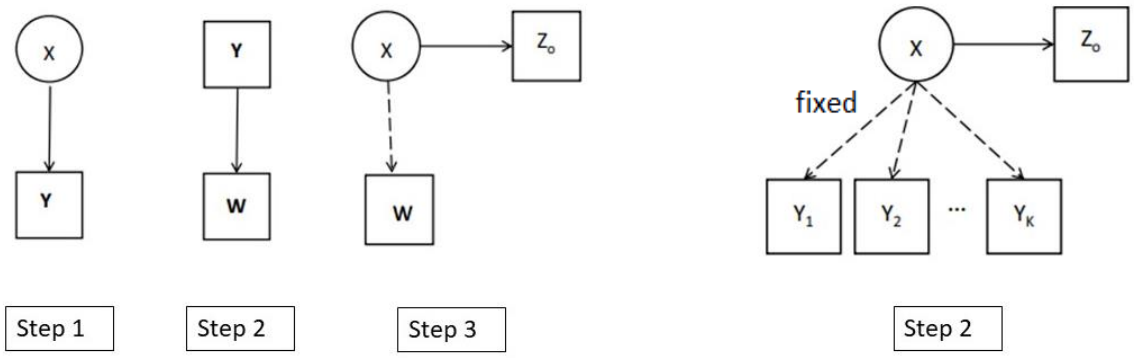
Note: Left: one-step approach. Right: LTB approach.

For the frequentist approach the BCH-method and the two-step approach are the most robust methods that can be currently used. The BCH-method is a three-step method for doing latent class analysis, first created by Bolck, Croon & Hagenaars (2004). Later on, this method has been modified to accommodate continuous covariates (Vermunt, 2010) and distal outcomes (Bakk et al., 2013), still using a three-step approach. Step one is building a basic latent class (*X*) model based on the categorical response variables (*Y*). The second step is assigning individuals to their predicted latent classes (*W*) and step three is estimating the association between *X* and *Z* using the assigned class membership, taking into account that these contain classification errors (Bakk & Vermunt, 2016). These classification errors are computed for each individual and the inverse logits of these individual error rates are used as weights in the third step. Advantages of this approach are that it is quite resistant to shifts in latent class membership between step one and three and that it can be used when the distal outcome

variances are either equal or unequal across latent classes. However, this method is sensitive to low entropy-values and small sample sizes, which can cause the weights to take on negative values if the distal outcome variances are unequal (Nylund-Gibson, 2019).

The two-step approach created by Bakk & Kuha (2017) needs only two steps for latent class analysis. The first step is exactly the same as for the three-step approaches such as the BCH-method. In the second and final step, we then maximize the joint likelihood, but with the parameters of the measurement model and of exogenous latent variables fixed at their estimated values from the first step, so that only the parameters of the rest of the structural model are estimated in the second step (Bakk & Kuha, 2018). By doing this, it avoids the problem of the misclassification error that the three-step approaches try to solve with their third step. However, these two methods are not (yet) available in software program R and only available in commercial software programs.

Figure 3. Bias-adjusted three-step approach & Two-step approach with distal outcome

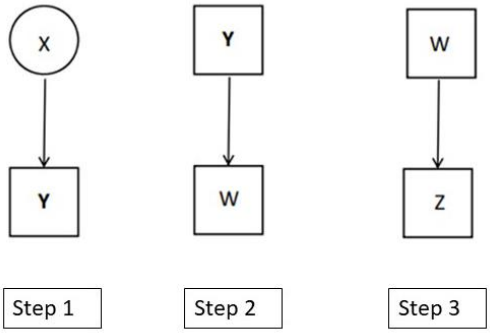


Note: extracted from Bakk & Kuha (2020).

Currently, only the naïve three-step method is available in R in the poLCA-package (Linzer & Lewis, 2011) and BayesLCA-package (White & Murphy, 2014). This three-step method is a predecessor of the BCH-method and does not account for the classification error that is introduced in step two which can lead to biased estimations in step three. This bias comes from the fact that the assigned class *W* is not necessarily equal to the true *X*, thus creating a misclassification error which can bias the estimations in step 3 (Bakk & Kuha, 2020).

The poLCA- and BayesLCA-packages do not offer the possibility to do a latent class distal outcome model. To create a latent class model with a distal outcome variable *Z*, the posteriors (probabilities for the classes) will be saved and then be used in an ANOVA treating them as observed scores to work around this problem.

Figure 4. Naive three-step approach



Note: extracted from Bakk & Kuha (2020).

1.3. Bayesian approach

Finding and comparing Bayesian methods proved to be more of a challenge, due to the small number of articles stating which specific LCA method they used and a lack of overview articles for the Bayesian LCA approaches. Some articles can be found using Bayesian approaches in the political science field (Alvarez, Katz, Levin & Núñez, 2021; Alvarez, Levin & Núñez, 2017; Katz & Levin, 2018; Oser & Jenkins, 2022), but they often focus on basic LCA or with a covariate and not LCA with a distal outcome. For the Bayesian approach two Markov Chain Monte Carlo (MCMC) based approaches have been selected, named Gibbs Sampling and Variational Bayes. Gibbs Sampling uses iterative sampling from each conditional distribution from which samples of the joint posterior distribution are indirectly obtained. Gibbs sampling relies on the Markov assumption, which means that samples drawn with iteration k +1 depend only on parameter values during the previous iteration k. Samples are repeatedly drawn until it is decided that a reasonable representation of the joint posterior distribution has been obtained (White & Murphy, 2014). Variational Bayes can be seen as a combination of both maximizing the joint posterior and iterative sampling from the conditional distribution. It can be used to obtain parameter estimates which maximize a fully factorized posterior approximation to the joint posterior (White & Murphy, 2014). It has gained popularity due to its relatively low computational cost and good empirical approximation. Both of these methods can be used in R using the BayesLCA-package (White & Murphy, 2014), but it does not offer a possibility of using a distal outcome modal. To solve this, the posteriors will be saved and then be used in an ANOVA treating them as observed scores.

1.4. Research question

Quite some research has been done on the differences between the frequentist methods (Bakk & Kuha, 2019; Bakk & Vermunt, 2016; Nyland-Gibson, Grimm & Masyn, 2019), but little research has been done on the Bayesian methods and their differences. Furthermore, no research has been done yet, comparing methods from these two different approaches. For my research I will compare Frequentist and Bayesian methods that are available in software program R while using a latent class distal outcome model. In the Frequentist approaches we know that stepwise approaches are needed in distal outcome models with dependent continuous variables to prevent the distal outcome of influencing the latent class solution. In this research I want to compare the Bayesian and Frequentist approaches that are available in R to identify if this influencing problem can also occur in the Bayesian approaches.

For this I will compare three methods from two different packages. For the frequentist approach I will use the naïve three-step approach from the poLCA- and BayesLCA-packages, combined with ANOVA. For the Bayesian methods I will use the partly Bayesian methods Gibbs-sampling and Variational Bayes from the BayesLCA-package combined with ANOVA. Comparisons will be made based on the prediction precision of each of these methods. This cannot be done when doing latent class analysis on real data, but in this simulation study we know all the relationships and their strengths, which makes it possible to see how well each method can predict the latent classes. Finally, all methods will be used on a real dataset and a comparison will be done based on their estimation of the class sizes, strength of *Y* and strength of *Z*. Since this is real data, there will be no true values available to compare them to, so here I will focus on comparing the methods purely to each other.

1.5. Theoretical framework

1.5.1 Basic LC model

Let Y_{ik} denote the response of individual *i* on one of *K* categorical response variables, where $1 \leq k \leq K$ and $1 \leq i \leq N$. A particular latent class is denoted by *t*, and the model can then be formulated as follows:

$$P(Y_i) = \sum_{t=1}^{T} P(X = t)\, P(Y_i|X = t) \quad (1)$$

Where $P(X = t)$ represents the unconditional probability of belonging to class t and $P(Y_i|X = t)$ represents the class specific distributions of the responses Y_i. The class-specific distributions can be

simplified by assuming that the *K* response variables are independent within classes, also known as the local independence assumption (Bakk & Vermunt, 2016). This gives us the following equation:

$$P(Y_i) = \sum_{t=1}^{T} P(X = t) \prod_{k=1}^{K} P(Y_{ik}|X = t) \quad (2)$$

1.5.2. One-step approach

In the one-step approach the measurement and structural model are calculated together. The basic latent class model as described in equation 2 can be extended to include a continuous distal outcome (Z_i), creating the one-step approach as seen in Figure 2:

$$P(Y_i, Z_i) = \sum_{t=1}^{T} P(X = t) \prod_{k=1}^{K} P(Y_{ik}|X = t) f(Z_i|X = t) \quad (3)$$

where $f(Z_i|X = t)$ denotes the class-specific distribution of Z_i. For continuous distal outcome variables, this usually is defined to be a normal distribution with mean μ and variance σ^2. In the one-step approach, the distal outcome is equal to predictors *Y*, since the measurement and structural model are measured simultaneously to estimate the full latent class model. Due to this predictor role for the distal outcome in the analysis, the estimation of the full model is sensitive to the distribution of this variable. If the distal outcome is not normally distributed, this can lead to misclassification problems and thus a distortion of the full model (Bauer & Curran, 2003).

1.5.3. Naive three-step approach

In the naive three-step approach, the optimal number of classes is first decided on using equation 2, so without yet adding the distal outcome in the model (Bakk & Vermunt, 2016). The assignment of respondents to these classes in step 2 is based on:

$$p(X = t|Y = y) = \frac{p(X=t)p(Y=y|X=t)}{p(Y=y)} \quad (4)$$

Which gives us the posterior probabilities that a respondent belongs to each of the classes given the respondent's observed response *y*, derived from the model in step 1. Commonly, a respondent is assigned to the class for which he or she has the highest posterior probability. This assigned class will be denoted as *W*. In step 3 this assigned class *W* replaces *X* when estimating the structural model. In the naïve three-step approach there is no adjustment for the use of *W*, giving us:

$$p(W = s|Y = y) = \frac{p(W=s)p(Y=y|W=s)}{p(Y=y)} \quad (5)$$

As discussed earlier, without adjustment for the difference between true class membership *X* and assigned class membership *W* a misclassification error is introduced into the model, since these two variables are not necessarily equal to each other. This misclassification error can lead to bias in the estimation of step 3 (Bakk & Kuha, 2020). The equations so far only focussed on obtaining the optimal

number of classes in our data, but do not yet take in account a distal outcome *Z* in our model. A distal outcome can be added after assigning class membership *W* and can be described as step 3 of the three-step approach. In this third step the association between *X* and the distal outcome *Z* is estimated, using the assigned class memberships *W*. The poLCA- and BayesLCA-packages do not offer a possibility of adding this distal outcome to the model, so the posteriors (probabilities for being assigned to a certain class) obtained in step two are needed to complete the three-step approach. Normally, these posteriors are used to assign class membership *W*, but we can also use them as input for an ANOVA to test the relationship between these class memberships and the distal outcome.

1.5.4. EM

In the BayesLCA-package the EM-algorithm will be used to perform the naïve three-step approach. The expected complete-data log-posterior is defined as

$$Q(\theta, \tau|\theta^k, \tau^k) := \mathrm{E}[\log p\,(\theta, \tau|X, L)|X, \theta^k, \tau^k] \tag{6}$$

Here, θ (item-probability) and τ (class-probability) are iteratively maximised, where $\theta^k$and τ^k denote the values of θ and τ at iteration *k*. At the *k*th stage of the algorithm the parameter estimates are updated in two steps: E-step and M-step. These steps are repeated until the algorithm is deemed to converge. The algorithm for these two steps can be written as:

E- step:
$$L_{ig}^{(k+1)} = \frac{\tau_g^{(k)}\, p(X_i|\theta_g^{(k)})}{\Sigma_{h=1}^{G} \tau_h^{(k)}\, p(X_i|\theta_h^{(k)})} \tag{7}$$

M-step:
$$\theta_{gm}^{(k+1)} = \frac{\Sigma_{i=1}^{N} X_{im} L_{ig}^{(k+1)} + \alpha_{gm} - 1}{\Sigma_{i=1}^{N} L_{ig}^{(k+1)} + \alpha_{gm} + \beta_{gm} - 2} \tag{8}$$

$$\tau_g^{(k+1)} = \frac{\Sigma_{i=1}^{N} L_{ig}^{(k+1)} + \delta_g - 1}{N + \Sigma_{h=1}^{G} \delta_h - G} \tag{9}$$

The L_i represents the true class membership of X_i. Since the L_i is not known, the posterior probability for the class membership of observation *i* is given by

$$p(L_i|X_i, \tau, \theta) = \prod_{g=1}^{G} \left[\frac{\tau_g p(X_i|\theta_g)}{\Sigma_{h=1}^{G} \tau_h p(X_i|\theta_h)}\right]^{L_{ig}} \tag{10}$$

More detailed information on the working of the EM-algorithm can be retrieved from White & Murphy (2014).

1.5.5. Gibbs Sampling

Sampling is done by iteratively sampling from each conditional distribution in turn. This method uses the Markov assumption that samples drawn at iteration *k* + 1 depend only on the parameters that are

drawn in the previous iteration *k*. By drawing many samples and using this assumption, a good representation of the joint posterior distribution can be reached (White & Murphy, 2014).

1.5.6. Variational Bayes

Variational Bayes is an approximate combination of both the EM and Gibbs Sampling techniques, since it can be used to obtain parameter estimates which maximize a fully factorized posterior approximation to the joint posterior. In this method we introduce the variational distribution with several variational parameters, which will then go through multiple updates to attempt to find the true posterior, using an arbitrary distribution Q(θ). Optimising Q(θ) to find the optimal posterior, is achieved through iterative optimisation, as also used in the EM algorithm (Marwade, 2021).

2. Simulation study

2.1. Methodology

A simulation study will be done where the naive three-step approach, Gibbs-sampling and Variational Bayes with distal outcome models from different R-packages will be compared to each other. The model will have two classes and weights set to be unequal at .4 and .6. There will be four indicator variables with different strengths to the two classes. In the first condition, class one will have a strength of .8 and class two a strength of .2 related to each of the four indicator variables. In the second condition the strengths are .9 and .1 respectively. So, in both conditions class one will score high on all indicator variables and class two will score low on all four indicator variables (Bakk, Di Mari, Oser & Kuha, 2022; Bakk & Vermunt, 2016). The distal outcome will have a class specific normal distribution and I will manipulate the strength of the latent class model with sample sizes of 250, 500 and 1000 (Bakk, et al., 2022; Bakk & Vermunt, 2016). and strengths of association with the distal outcome. For this I will use the strong and weak set up for the distal outcome- latent class relationship of a regression coefficient equal to .2 and -.2 in the weak condition and .8 and -.8 in the strong condition (Bakk & Vermunt, 2016). Furthermore, I will compare the different approaches and R-packages when used on real data. The R-package BayesLCA will be used for the Bayesian analyses Gibbs sampling and Variational Bayes together with

Sample size	Strength Y	Strength Z
250	.8	.8
500	.8	.8
1000	.8	.8
250	.8	.2
500	.8	.2
1000	.8	.2
250	.9	.8
500	.9	.8
1000	.9	.8
250	.9	.2
500	.9	.2
1000	.9	.2

Table 1. *Simulation conditions*

ANOVA. For the Frequentist approach I will use the poLCA- and BayesLCA-packages (EM), also used in combination with ANOVA.

2.2. Results simulation study

2.2.1. Label Switching

Before presenting the results, it should be mentioned that label switching was prevented as much as possible using the options offered in the corresponding R-packages. However, even with these measures, there were still a few rows that show signs of label switching (see Table 2). For example, in one of the poLCA-columns measuring the strength of *Y*, there were five rows that showed a strength of approximately .06 to .25 instead of the correct .8 or .9 strengths in these conditions. When label switching occurs for a certain row, it tends to occur for all four LCA-methods as well. Considering that there were a total of 1200 rows in this dataset, five rows showing different results will not influence the results that much. For this reason, it was decided to keep all rows in the dataset, so that all conditions would have the same number of cases that the results would be based on.

Row	poLCA	EM	Gibbs	Variational Bayes
41	.280	.158	.168	.161
203	.249	.167	.169	.167
424	.277	.168	-	-
501	.210	.187	.190	.187
952	.109	.080	.088	.081

Table 2. *Overview of the label switching rows in the dataset, with corresponding Y-values.*

2.2.2. Comparing LCA-methods

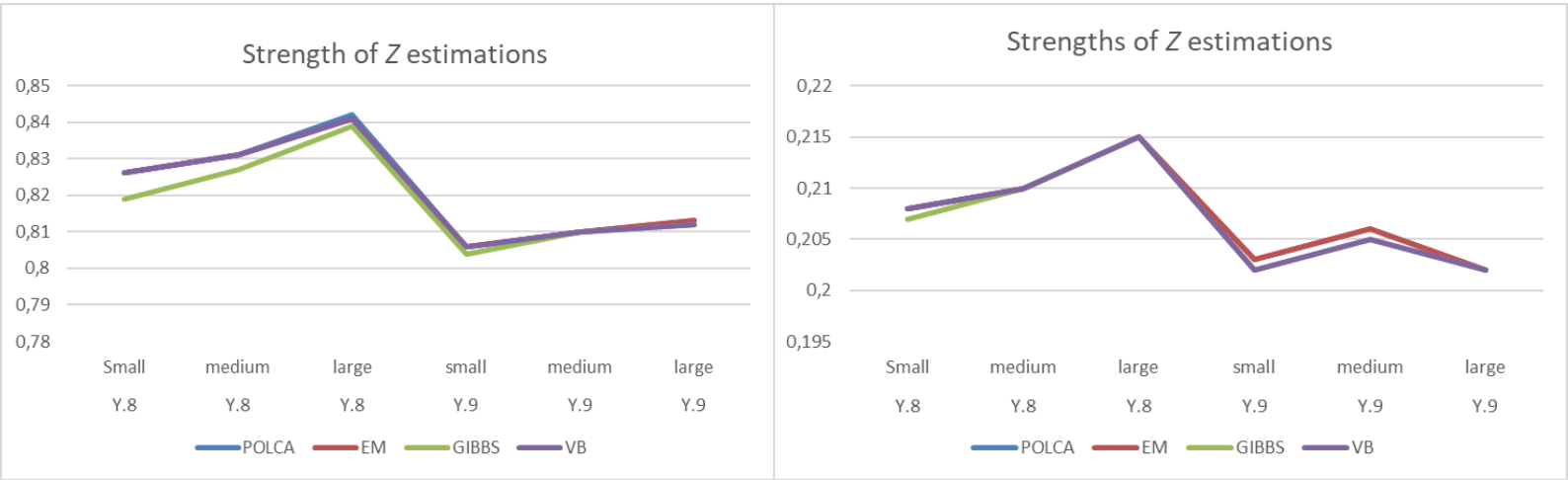
In the simulation study, the four different methods are compared based on multiple variables such as group sizes, the strength of *Y* and strength of *Z*. When looking at the group sizes over the different conditions, we can see that the distribution of the cases strongly resembles the true sample distribution of 40% in group 1, and 60% in group 2. The only condition where this clearly deviates is for a sample size of 250 and strengths of *Y* and *Z* being .8, for all four methods. Here, group one in poLCA, EM and Variational Bayes has a size of 39.1% and group two is 60.9%, for Gibbs Sampling group one is 39.5% and group two is 60.5%.

The strength of the relationship between *Y* and *X* is very similar for poLCA, EM and Variational Bayes, with estimations of the strength only deviating .002 from each other or being exactly the same in most conditions. These three methods all estimate the strength of the relationship quite well, in

most conditions measuring the exact strength as set in the simulation or only deviating .002 from the real strength of *Y* when *Y*= .9. The estimations are still quite well when *Y*= .8, with deviations up to .005 in most conditions. The estimations of the relationship between *Y* and *X* with Gibbs Sampling generally perform worse compared to the other methods. Especially when the sample size is relatively small (*N*=250), Gibbs sampling struggles to get near the true strength of *Y*, with deviations that are bigger than .010 from the true strength. Gibbs Sampling tends to improve quite well when the sample size increases, but its estimates still perform a bit worse when compared to poLCA, EM and Variational Bayes. The methods all tend to underestimate the strength of Y when the sample size is small, but this effect is more visible when the strength of Y is set to be .8 than for strength .9 in the simulation.

The estimations of *Z* by poLCA, EM and Variational Bayes are almost identical copies of each other, with only a couple conditions showing a very small difference between them. The estimations of Gibbs Sampling are structurally lower than the other methods when the sample size is small, but gets very close or even equal to the other methods when the sample size is *N*=1000. Table 3 and Figure 5 also shows that the strength of *Z* is overestimated in every condition, but that this overestimation is larger when the strength of *Y* is .8 than in conditions where the strength of *Y*=.9. The lower limit of the 95% confidence intervals of *Z* is often larger than the true value of *Z,* which indicates that only a very small number of simulations get near the true value.

Figure 5. Estimations of the strength of *Z* by poLCA, EM, Gibbs Sampling and Variational Bayes



Note: Left: estimations when the true value of *Z* is .8. Right: estimations when the true value of Z is .2.

To test whether the four methods are indeed different or similar to each other, an ANOVA was done using the residuals of the *Z* estimations. This test showed that there is indeed a difference between the methods, with F(3)= 3.598 and p=.013, which indicates that there is at least one method that is different from the other methods. The Tukey post hoc test was used to test which methods might be significantly different from each other and to control for the multiple comparisons.

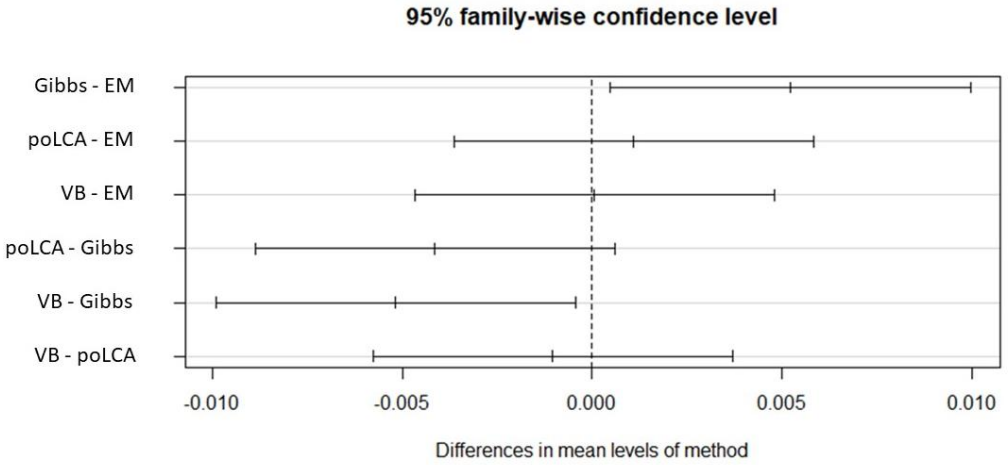
Study conditions			poLCA								EM _ BayesLCA								
Sample size	Strength Y	Strength Z	Class 1	Class 2	Strength Y	Strength Z	Lower CI 95%	Upper CI 95%	Std. Error	Variance	RMSE	Class 1	Class 2	Strength Y	Strength Z	Lower CI 95%	Upper CI 95%	Std. Error	Variance	RMSE
250	.8	.8	.391	.609	.792	.826	.812	.839	.038	.0014	.194	.391	.609	.795	.826	.812	.839	.038	.0014	.194
500	.8	.8	.398	.602	.801	.831	.819	.843	.027	.0007	.163	.398	.602	.804	.831	.819	.843	.027	.0007	.163
1000	.8	.8	.402	.598	.796	.842	.833	.850	.019	.0004	.137	.402	.598	.796	.841	.833	.850	.019	.0004	.137
250	.8	.2	.398	.602	.799	.208	.202	.214	.033	.0011	.181	.398	.602	.803	.208	.201	.214	.033	.0011	.181
500	.8	.2	.399	.601	.789	.210	.204	.216	.024	.0006	.154	.399	.601	.788	.210	.204	.215	.024	.0006	.154
1000	.8	.2	.399	.601	.798	.215	.211	.219	.017	.0003	.129	.399	.601	.803	.215	.211	.219	.017	.0003	.129
250	.9	.8	.403	.597	.900	.806	.797	.815	.036	.0013	.189	.403	.597	.898	.806	.797	.815	.036	.0013	.189
500	.9	.8	.400	.600	.901	.810	.803	.817	.025	.0006	.159	.400	.600	.900	.810	.803	.817	.025	.0006	.159
1000	.9	.8	.400	.600	.901	.813	.808	.818	.018	.0003	.134	.400	.600	.900	.813	.808	.818	.018	.0003	.134
250	.9	.2	.398	.602	.892	.203	.195	.210	.033	.0011	.181	.398	.602	.892	.203	.195	.210	.033	.0011	.181
500	.9	.2	.399	.601	.899	.206	.201	.210	.023	.0005	.152	.399	.601	.901	.206	.201	.210	.023	.0005	.152
1000	.9	.2	.400	.600	.901	.202	.198	.205	.016	.0003	.128	.400	.600	.900	.202	.198	.205	.016	.0003	.128

Study conditions			Gibbs Sampling								Variational Bayes								
Sample size	Strength Y	Strength Z	Class 1	Class 2	Strength Y	Strength Z	Lower CI 95%	Upper CI 95%	Std. Error	Variance	RMSE	Class 1	Class 2	Strength Y	Strength Z	Lower CI 95%	Upper CI 95%	Std. Error	Variance	RMSE
250	.8	.8	.395	.605	.785	.819	.804	.834	.038	.0014	.195	.391	.609	.794	.826	.812	.839	.038	.0014	.194
500	.8	.8	.400	.600	.798	.827	.814	.840	.027	.0007	.164	.398	.602	.803	.830	.818	.843	.027	.0007	.163
1000	.8	.8	.403	.597	.794	.839	.829	.848	.019	.0004	.138	.402	.598	.796	.841	.832	.850	.019	.0004	.137
250	.8	.2	.402	.598	.793	.207	.201	.213	.033	.0011	.181	.399	.601	.802	.208	.202	.214	.033	.0011	.181
500	.8	.2	.401	.599	.789	.210	.204	.216	.024	.0006	.154	.399	.601	.793	.210	.204	.215	.024	.0006	.154
1000	.8	.2	.400	.600	.801	.215	.211	.219	.017	.0003	.129	.399	.601	.803	.215	.211	.219	.017	.0003	.129
250	.9	.8	.404	.596	.889	.804	.794	.813	.036	.0013	.189	.403	.597	.898	.806	.797	.815	.036	.0013	.189
500	.9	.8	.401	.599	.895	.810	.803	.816	.025	.0006	.159	.400	.600	.900	.810	.803	.817	.025	.0006	.159
1000	.9	.8	.400	.600	.897	.812	.807	.817	.018	.0003	.134	.400	.600	.900	.812	.807	.817	.018	.0003	.134
250	.9	.2	.399	.601	.883	.202	.195	.210	.033	.0011	.181	.398	.602	.891	.202	.195	.210	.033	.0011	.181
500	.9	.2	.401	.599	.897	.205	.200	.210	.023	.0005	.152	.400	.600	.901	.205	.200	.210	.023	.0005	.152
1000	.9	.2	.402	.598	.898	.202	.198	.205	.016	.0003	.128	.400	.600	.900	.202	.198	.205	.016	.0003	.128

Table 3. *Summary of simulation output. Class sizes, strength of Y, strength of Z, 95% CI of Z and standard error of Z are shown per method and per simulation condition.*

This analysis shows that the poLCA, EM and Variational Bayes methods are indeed not significantly different from each other with p-values greater than .900. However, Gibbs Sampling is significantly different from EM (p=.024) and Variational Bayes (p=.026), but not from poLCA (p=.112). These results are also visualized in Figure 6 with the 95% confidence intervals.

Figure 6



Note: 95% Confidence Interval for ANOVAs between the four methods with Tukey post hoc analysis.

3. Application study

3.1. Methodology

Next to a simulation study, I will also compare the four methods to each other based on a real-life dataset. For this I will use data that was obtained in a survey of students' math course in secondary school in Portugal (Banerjee, n.d.). This dataset contains 34 variables, of which thirteen were binary variables that could be used in an LCA. The dataset also has three outcome variables, which are the math grades in the first, second and final period, with a range from zero to twenty. Variables were selected based on Evans & Field (2020) and intuition (e.g., extra help in a subject can improve your grades). LCA will be performed based on the students' home address (urban or rural), parent's cohabitation status (living together or apart), extra educational support (yes or no), extra paid classes within course subject (yes or no) and if they have internet access at home (yes or no). Weekly study time was also added as a binary variable, where a study time up to five hours is considered low and five hours or more as high. Finally, the final math grade was selected as a distal outcome for the LCA.

3.2. Results

In this part of the study, the different packages and methods are tested on more realistic data. First, the number of classes was determined based on the six predictor variables using the poLCA-package. Six different models were tested, each with a different number of classes. The model selected based

on the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) is the two-class model (BIC= 2363 and AIC= 2311). The entropy for this model is *entropy-based* $R^2 = .59$, which indicates that the model cannot distinguish very well between the classes, as the ideal entropy would be around .8 or higher (Muthén, 2008). The classification table (Table 4) shows that the model has difficulty correctly predicting who will be in class 2, with one-third of the individuals being predicted to be in class 1. The predictions for class 1 are much better, with all individuals correctly assigned to class 1 and zero assigned to class 2. The analysis will be continued using this model, since the entropy is not part of the model selection and a model can still statistically perform well with a less than ideal entropy (Muthén, 2008).

		Predicted class	
		1	2
Observed class	1	**300.5**	0.0
	2	28.5	**66.0**

Table 4. *Classification table, with correctly predicted classes in bold*.

First, the classes and their characteristics will be explained, to get an idea of what the classes look like (see Table 6), since the methods are very alike in their characterisation of these classes. After this short description, the focus will lay on the differences between the methods. Class one in this model can be characterized by multiple variables, such as their high scores on extra paid classes in mathematics, high scores on internet availability and almost all students living in an urban area. The students in class two score relatively low on having extra paid classes in mathematics and score low on their internet availability as well. This second class also consists of a close to equal mix of students from rural and urban areas. The two classes do not differ a lot based on the amount of time that they study per week, with most students studying less than 5 hours a week. They also do not differ a lot based on the parental cohabitation status, with many parents living together and only a small number living apart.

	Class 1	Class 2	Strength Z	Std. Error
poLCA	.239	.761	-1.208	.616
EM	.238	.762	-1.208	.616
Gibbs	.255	.745	-1.208	.616
Variational Bayes	.179	.821	-1.208	.616

Table 5. *Summary of output non-simulated dataset, with class sizes, strength of Z and standard error of Z shown per LCA-method.*

Class	Method	Address	Pstatus	Studytime	Schoolsup	Paid	Internet
1	poLCA	.839	.911	.257	.126	.514	1.000
1	EM	.839	.911	.257	.126	.513	.999
1	Gibbs	.844	.910	.260	.127	.529	.959
1	VB	.826	.908	.251	.127	.501	.976
2	poLCA	.581	.850	.155	.138	.282	.301
2	EM	.580	.850	.156	.138	.281	.301
2	Gibbs	.561	.840	.167	.147	.261	.391
2	VB	.553	.843	.150	.139	.262	.178

Table 6. *Summary of output non-simulated dataset, with the strengths of Y for each variable per class and LCA-method.*

When comparing the methods, the three-step approach from poLCA and EM give very similar results on all fronts of this analysis: class sizes, strength of *Z* and the strengths of *Y* are all exactly the same or only differ by .001 from each other. The Gibbs Sampling method is quite similar to poLCA and EM on class sizes and on most of the estimating variables. However, Gibbs Sampling does deviate quite a bit from poLCA and EM on the predictors 'paid tutoring' and 'having internet'. The Variational Bayes method stands out from the other methods, with results such as the class sizes being very different when compared to the results from poLCA, EM and Gibbs Sampling. Table 5 shows that Variational Bayes gives estimations that are either close to the poLCA and EM estimations or to the Gibbs sampling estimation, with exception of the 'having internet' variable where it has an estimation for class 2 that is completely different from the other methods.

4. Discussion

For the simulated part of this research, it can be concluded that the three-step approaches from the poLCA- and BayesLCA-package (EM), are very similar to each other based on sample size, estimation of *Y* and estimation of *Z*, but not exactly the same. Regarding their estimation ability, it can be found that the estimation of *Y* is very close to the true value of *Y* when the sample size is 500 or 1000, but that the methods tend to underestimate the value of *Y* when the sample size is small (*N*=250). The estimation of *Z* on the other hand tends to be overestimated by quite a bit in certain conditions, especially when the sample size is large. An interesting find in this simulation is that the Variational Bayes method is almost identical in its estimations for class size, strength of *Y* and strength of *Z* when compared to two different three-step approaches (poLCA-package and BayesLCA-package). While the poLCA and EM methods are Frequentist, the Variational Bayes method is Bayesian, but they still are very similar in their estimations. The Gibbs Sampling method however, produces very different estimations for some conditions for *Y* and *Z* when compared to the other three methods. The Gibbs Sampling method underestimates the values of *Y* quite a bit more for the small samples when compared to the other three methods, but the estimations improve very well when the sample size increases. The estimations of *Z* are all overestimated, but they are closer to the real value of *Z* when

compared to the other methods, when the sample sizes are small (*N*=250). The estimations for *Z* are very close between the different methods (but still overestimated) when the sample size increases. An extra analysis, comparing the residuals of *Z* for the different methods confirms that the Gibbs Sampling method is significantly different from EM and Variational Bayes, but not from the three-step approach used in poLCA.

For the application latent class analysis, it was found that the three-step approaches from poLCA and BayesLCA (EM) are again very similar to each other. However, in contrast to the simulated data, the Variational Bayes shows different results for the estimation of class sizes and of the strength of *Y* for some of the predicting variables. The Gibbs Sampling method is relatively similar to poLCA and EM in class size estimation and for most of the predicting variables, but not as similar as the three-step approaches are to each other. The estimations of *Z* and its standard error are exactly the same for all four LCA methods.

Label switching was prevented as much as possible using the options offered in the corresponding R-packages. However, label switching can still occur using the poLCA- and BayesLCA-packages even when prevention measures have been taken, but the effects of this are minimal. A limitation of this simulation study is that there is no R-package available for creating a complete latent class with distal outcome model. Because of this, the data had to be created in two separate steps, which is not ideal. A dataset was created for latent classes with 0-1-data, after which poLCA was used to decide which combinations on the *Y* variables belong to each group. Based on this grouping, the distal outcome was added with the rnorm-function. So, the results of the BayesLCA methods are in a way influenced by the poLCA grouping method.

5. References

Alvarez, R. M., Katz, G., Levin, I. & Núñez, L. (2021). Conventional and unconventional participation in Latin America: a hierarchical latent class approach. *Political Science Research and Methods, 9*, 878, 888. DOI: 10.1017/psrm.2020.35

Alvarez, R. M., Levin, I., & Núñez, L. (2017). The four faces of political participation in Argentina: using latent class analysis to study political behavior. *The Journal of Politics, 79*, 1386-1402.

Asparouhov, T., & Muthén, B. (2014). *Auxiliary variables in mixture modeling: Using the BCH method in Mplus to estimate a distal outcome model and an arbitrary secondary model* (Mplus Web Note: No. 21). Retrieved from: http://www.statmodel.com/download/asparouhov_muthen_2014.pdf

Banerjee, P. (n.d.). *Student Alcohol Consumption*. Retrieved November 17, 2022 from https://www.kaggle.com/datasets/prashant111/student-alcohol-consumption.

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. DOI:10.1080/10705511.2021.1980882

Bakk, Z. & Kuha, J. (2017). Two-step estimation of models between latent classes and external variables. *Psychometrika*. DOI: 10.1007/s11336-017-9592-7

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. DOI: 10.1080/10705511.2014.955104

Bakk, Z. & Kuha, J. (2020). Relating latent class membership to external variables: An overview. *British Journal of Mathematical and Statistical Psychology, 74*, 340-362. DOI: 10.1111/bmsp.12227

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association, 92,* 1375-1386.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implication for overextraction of latent trajectory classes. *Psychological Methods, 8*,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. doi:10.1093/pan/mph001

Bonikowski, B., Feinstein, Y., & Bock, S. (2021). The Partisan Sorting of "America": How Nationalist Cleavages Shaped the 2016 U.S. Presidential Election. *American Journal of Sociology, 127*(2), 492-561. DOI: 10.1086/717103

Curran, P., & Bauer, D. (2013). What's the best way to determine the number of latent classes in a finite mixture analysis? Available at centerstat.org/class-ensumeration/. Accessed on 13-12-2022.

Elliot. M.R., Zhou, Z., Mukherjee, B., Kanaya, A., & Needham, B.L. (2020). Methods to account for uncertainty in latent class assignments when using latent classes as predictors in regression models, with application to acculturation strategy measures. *Epidemiology, 31*(2), 194-204. DOI: 10.1097/EDE.0000000000001139

Evans, D., & Field, A.P. (2020). Predictors of mathematical attainment trajectories across the primary-to-secondary education transition: parental factors and the home environment. *Royal society open science, 7*(7), 200422–200422. https://doi.org/10.1098/rsos.200422

Fornacon-Wood, I., Mistry, H., Johnson-Hart, C., Faivre-Finn, C., O'Connor, J.P.B. & Price, G.J. (2022). Understanding the Differences Between Bayesian and Frequentist Statistics. *Statistics for the people, vol. 112*(5), 1076-1082.

Katz, G., & Levin, I. (2018). Varieties of political support in emerging democracies: A cross-national analysis. *Social Science research, 70*, 55-70. DOI: 10.1016/j.ssresearch.2017.11.002

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

Linzer, D. A. & Lewis, J., B. (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. Journal of Statistical Software, 42(10), 1-29. URL https://www.jstatsoft.org/v42/i10/

Marwade, A. (2021). Variational Bayes: the intuition behind Variational Auto-Encoders (VAEs). Retrieved 27 November 2022, from towardsdatascience.com/variational-bayes-4abdd9eb5c12

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Model Ing: A Multidisciplinary Journal, 24,* 180-197. DOI: 10.1080/10705511.2016.1254049

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. DOI: 10.1002/cbm.1819

Muthén, B. O. (2008) What is a good value of entropy?. Available at http://www.statmodel.com/discussion/messages/13/2562.html?1487458497. Accessed December 12, 2022.

Nylund-Gibson, K., Grimm, R. P. & Masyn, K. (2019). Prediction from Latent Classes: A Demonstration of Different Approaches to include Distal Outcomes in Mixture Models. *Structural Equation Modeling: A Multidisciplinary Journal, 26:6,* 967-985. DOI: 10.1080/10705511.2019.1590146

Oser, J. (2022). Protest as One Political Act in Individuals' Participation Repertoires: Latent Class Analysis and Political Participant Types. The American Behavioral Scientist (Beverly Hills), 66(4), 510–532. DOI: 10.1177/00027642211021633

Quirk, M., Nylund-Gibson, K., & Furlong, M. (2013). Exploring patterns of Latino/a children's school readiness at kindergarten entry and their relationship with Grade 2 achievement. *Early Childhood Research Quarterly, 28,* 437-449. DOI: 10.1016/j.ecresq.2012.11.002

Taboga, Marco (2021). "Uninformative prior", Lectures on probability theory and mathematical statistics. Kindle Direct Publishing. Online appendix. https://www.statlect.com/fundamentals-of-statistics/uninformative-prior.

White, A. & Murphy, T. B. (2014). BayesLCA: An R Package for Bayesian Latent Class Analysis. *Journal of Statistical Software, 61*(13), 1–28. DOI: 10.18637/jss.v061.i13

Appendix A. R-code Simulation study

```
#Loading packages
library("BayesLCA"); library("poLCA"); library("readr"); library("Rmisc")

#Generate data conditions
nrep <-100  #number of simulations
sampsize <- c(250, 500, 1000)
weights <- c(.4, .6)
strength_Z <- c(.8,.2)
strength_y <- c(.8,.9)
conds <- expand.grid(sampsize, strength_Z, strength_y)
names(conds) <- c("sample size", "strength_Z", "strength_y")

###########1. Generate Data-function
MAKE_DATA <- function(sampsize, strength_Z, strength_y){
  #generating 0|1-data based on item-probability and sample weights
  strength_Y <- c(strength_y, strength_y, strength_y, strength_y)
  x <- as.data.frame(rlca(sampsize, rbind(strength_Y, 1-strength_Y), weights))
  x$sum <- rowSums(x[,1:4], na.rm = TRUE)
  #changing y-variables to 1-2 data, since poLCA does not work on 0-1-data
  x[, 1:4] <- ifelse(x[,1:4] == 0, 1, 2)
  f <- cbind(V1, V2, V3, V4) ~ 1 #formula
  model_poLCA <- poLCA(f, x, nclass= 2, probs.start = 1)
  posteriors <- data.frame(model_poLCA$posterior, predclass = model_poLCA$predclass)
  #dataset + posteriors
  x_complete <- cbind(x, posteriors)
  x_complete$predclass <- as.factor(x_complete$predclass)
  Group1 <- x_complete[x_complete$predclass == 1,]
  error <- rnorm(nrow(Group1))
  Group1$z <- 10 + strength_Z * Group1$sum + error
  Group2 <- x_complete[x_complete$predclass == 2,]
  error2 <- rnorm(nrow(Group2))
  Group2$z <- 10 -strength_Z * Group2$sum + error2

  x_complete <- rbind(Group1, Group2)
  #extra code to change it back to 0-1code for Bayes-analyses.
  x_complete[, 1:4] <- ifelse(x_complete[,1:4] == 1, 0, 1)
  return(x_complete)
  return(model_poLCA)
}
########### 2. Analyses  ########################
############ POLCA #################
POLCA <- function(x_complete){
  x_complete[, 1:4] <- as.data.frame(ifelse(x_complete[,1:4] == 0, 1, 2)) #extra code for poLCA, since it
     does not work with 0-1-data, I change it to 1-2 data
  #running regression using GLM with posterior class membership as multinomial predictor
  f <- cbind(V1, V2, V3, V4) ~ 1 #formula
```

```
  model_poLCA <- poLCA(f, x_complete[,1:4], nclass= 2, probs.start = 1)

  probs.start.new <- poLCA.reorder(model_poLCA$probs.start,order(model_poLCA$P,decreasing = TRUE))

  model_poLCA <- poLCA(f, x_complete, nclass= 2, probs.start= probs.start.new)

  model_poLCA_dis <- glm(z ~ predclass , data = x_complete)

  output_p1 <- summary(model_poLCA_dis)$coefficients[2,1:2]

  output_p2 <- c(min(model_poLCA$P),max(model_poLCA$P))

  output_p3<-               c(model_poLCA$probs$V1[2,1],model_poLCA$probs$V2[2,1],model_poLCA$probs$V3[2,1],
     model_poLCA$probs$V4[2,1],
     model_poLCA$probs$V1[1,1],model_poLCA$probs$V2[1,1],model_poLCA$probs$V3[1,1],
     model_poLCA$probs$V4[1,1])

  output1 <- c(output_p1,output_p2, output_p3)

}

############ BayesLCA ~ EM #################

EM <- function(x_complete){

  fit_em <- blca.em(x_complete[,1:4], 2, restarts = 20)

  # setting probabilities for unique datapoints as seperate dataframe

  Z1 <- as.data.frame(fit_em$Z)

  #if the probability for being in group1 is >.5, person is set in group 1, else in group 2

  Z1$class <- as.factor(ifelse(Z1$`Group 1`>.5, 1, 2))

  #to set the unique datapoints as a seperate column in the dataset so that I can link them to the other
     other dataset

  Z1$code <- rownames(Z1)

  #creating a variable consisting of the answers of the participant (same set up as the unique datapoints
     from the row above)

  x_complete$code <- apply(x_complete[,1:4], 1, paste, collapse= "")

  #linking the datapoints to their classes

  x_complete2 <- merge(x_complete, Z1[, c("code", "class")], by= "code")

  model_EM_dis<- glm(z ~ 1+class , data = x_complete2)

  output_EM1 <- summary(model_EM_dis)$coefficients[2,1:2]

  output_EM2 <- c(min(fit_em$classprob),max(fit_em$classprob))

  output_EM3 <- c(fit_em$itemprob[1,], fit_em$itemprob[2,])

  output2 <-  c(output_EM1,output_EM2, output_EM3)

}

############ BayesLCA ~ GIBBS #################

Gibbs <- function(x_complete){

  fit_gibbs <- blca.gibbs(x_complete[,1:4], 2, relabel = TRUE)

  Z_gibbs <- as.data.frame(fit_gibbs$Z)

  #if the probability for being in group1 is >.5, person is set in group 1, else in group 2

  Z_gibbs$class <- as.factor(ifelse(Z_gibbs$`Group 1`> 0.5, 1, 2))

  #to set the unique datapoints as a seperate column in the dataset so that I can link them to the other
     other dataset

  Z_gibbs$code <- rownames(Z_gibbs)

  #creating a variable consisting of the answers of the participant (same set up as the unique datapoints
     from the row above)

  x_complete$code <- apply(x_complete[,1:4], 1, paste, collapse= "")

  #linking the datapoints to their classes

  x_complete2 <- merge(x_complete, Z_gibbs[, c("code", "class")], by= "code")

  model_gibbs_dis<- glm(z ~ 1+class , data = x_complete2)

  summary(model_gibbs_dis)

  output_G1 <-summary(model_gibbs_dis)$coefficients[2,1:2]
```

```
  output_G2 <- c(min(fit_gibbs$classprob),max(fit_gibbs$classprob))

  output_G3 <- c(fit_gibbs$itemprob[1,], fit_gibbs$itemprob[2,])

  output3 <- c(output_G1,output_G2, output_G3)

}

############ BayesLCA ~ Variational Bayes #################

VB <- function(x_complete){

  fit_vb <- blca.vb(x_complete[,1:4], 2, restarts = 20)

  Z_vb <- as.data.frame(fit_vb$Z)

  #if the probability for being in group1 is >.5, person is set in group 1, else in group 2 Z_vb$class <-
     as.factor(ifelse(Z_vb$`Group 1`> 0.5, 1, 2))

  #to set the unique datapoints as a seperate column in the dataset so that I can link them to the other
     other dataset

  Z_vb$code <- rownames(Z_vb)

  #creating a variable consisting of the answers of the participant (same set up as the unique datapoints
     from the row above)

  x_complete$code <- apply(x_complete[,1:4], 1, paste, collapse= "")

  #linking the datapoints to their classes

  x_complete2 <- merge(x_complete, Z_vb[, c("code", "class")], by= "code")

  model_vb_dis<- glm(z ~ 1+class , data = x_complete2)

  output_vb1 <- summary(model_vb_dis)$coefficients[2,1:2]

  output_vb2 <- c(min(fit_vb$classprob),max(fit_vb$classprob))

  output_vb3 <- c(fit_vb$itemprob[1,], fit_vb$itemprob[2,])

  output4 <- c(output_vb1,output_vb2, output_vb3)

}

############ 3. GENERATE DATA MATRIX PER CELL AND WRITE TO FILE ############

cellData <- NULL

################ Complete run #################

for (r in 1:12){

  for (k in 1:nrep){

    set.seed(1000*r+k)

    output0 <- c(r, k, conds[r,1], conds[r,2], conds[r,3])

    x_complete <- MAKE_DATA(conds[r,1], conds[r,2], conds[r,3])

    output1 <- POLCA(x_complete)

    output2 <- EM(x_complete)

    output3 <- Gibbs(x_complete)

    output4 <- VB(x_complete)

    output  <- c(output0, output1, output2, output3,output4)

    cellData <- rbind(cellData, output)

  }

  rownames(cellData) <- NULL

  colnames(cellData) <- c("r", "k", "sampsize", "strength_z", "strength_y","estimatePOLCA", "std.errPOLCA",
     "Group1POLCA","Group2POLCA","y1_1",  "y1_2",  "y1_3",  "y1_4",  "y2_1",  "y2_2",  "y2_3",  "y2_4",
     "estimateEM", "std.errEM", "Group1EM","Group2EM","y1_1", "y1_2", "y1_3", "y1_4", "y2_1", "y2_2", "y2_3",
     "y2_4","estimateGibbs",  "std.errGibbs",  "Group1Gibbs", "Group2Gibbs","y1_1", "y1_2", "y1_3", "y1_4",
     "y2_1", "y2_2", "y2_3", "y2_4","estimateVB", "std.errVB", "Group1VB","Group2VB","y1_1", "y1_2", "y1_3",
     "y1_4", "y2_1", "y2_2", "y2_3", "y2_4")

  outfile <- paste("MP2022output",r,".csv",sep="")

  write.csv(cellData, file = outfile)

}
```