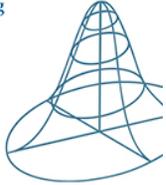


Stepwise Latent Class analysis in the presence of missing values on the class indicators

Authors	Alagöz, Emre C.; Vermunt, Jeroen K.
Published in	Structural Equation Modeling
DOI	10.1080/10705511.2022.2030743
Publication Date	2022
Document Version	publishersversion
Link	https://research.tilburguniversity.edu/en/publications/75b45b3e-59d2-43c4-b5bb-e66e62ee12c6
Citation	Alagöz, E C & Vermunt, J K 2022, 'Stepwise Latent Class analysis in the presence of missing values on the class indicators', Structural Equation Modeling, vol. 29, no. 5, pp. 784-790. https://doi.org/10.1080/10705511.2022.2030743
Download Date	2025-10-13 22:36:54
Rights	<p>General rights</p> <p>Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.</p> <ul style="list-style-type: none"> - Users may download and print one copy of any publication from the public portal for the purpose of private study or research. - You may not further distribute the material or use it for any profit-making activity or commercial gain - You may freely distribute the URL identifying the publication in the public portal" <p>Take down policy</p> <p>If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.</p>



Stepwise Latent Class Analysis in the Presence of Missing Values on the Class Indicators

Ö. Emre C. Alagöz & Jeroen K. Vermunt

To cite this article: Ö. Emre C. Alagöz & Jeroen K. Vermunt (2022) Stepwise Latent Class Analysis in the Presence of Missing Values on the Class Indicators, Structural Equation Modeling: A Multidisciplinary Journal, 29:5, 784-790, DOI: [10.1080/10705511.2022.2030743](https://doi.org/10.1080/10705511.2022.2030743)

To link to this article: <https://doi.org/10.1080/10705511.2022.2030743>



© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 25 Mar 2022.



Submit your article to this journal [↗](#)



Article views: 1426



View related articles [↗](#)



View Crossmark data [↗](#)

Stepwise Latent Class Analysis in the Presence of Missing Values on the Class Indicators

Ö. Emre C. Alagöz^a  and Jeroen K. Vermunt^b 

^aUniversity of Mannheim; ^bTilburg University

ABSTRACT

While latent class (LC) modeling using bias-adjusted stepwise approaches has become widely popular, little is known on how these methods are affected by missing values. Using synthetic data sets, we illustrate under which conditions missing values introduce biases in the estimates of the relationship between class membership and auxiliary variables. We apply three-step LC analysis with both modal and proportional class assignments, as well as the recently proposed two-step LC analysis method. Our results show that stepwise LC analysis yields unbiased parameter values as long as the MAR assumption holds in the step-one model. When this assumption does not hold because covariates are omitted from the step-one model, each of the stepwise approaches yields some bias, but bias is much larger with modal class assignments. The amount of bias is affected by the amount of deviation from MAR, the proportion of missing values, and the separation between the classes.

KEYWORDS

Missing data; mixture modeling; three-step modeling; auxiliary variables

1. Introduction

In social and behavioural sciences and related fields, latent class (LC) analysis (Goodman, 1974; Lazarsfeld & Henry, 1968) has become a popular tool for classifying respondents into a small number of subgroups based on their response patterns on a set of observed indicators. More extended LC models allow the inclusion of auxiliary variables (e.g., covariates, distal outcomes) to examine the cause of the class formation or the effect of these classes on other constructs. While researchers are often confronted with missing values on the class indicators, maximum likelihood estimation of LC models with missing data is straightforward as long as the missingness can be assumed to be missing at random (MAR; Dong & Peng, 2013; Little & Schenker, 1995). This approach is implemented in most of the current software for LC analysis, such as Mplus (Muthén & Muthén, 2015), Latent GOLD (Vermunt & Magidson, 2013, 2021a), poLCA in R (Linzer & Lewis, 2011), and PROC LCA in SAS (Lanza, Dziak, Huang, Wagner, & Collins, 2015). Note that in an LC analysis, MAR implies the missingness is independent of the actual value of the indicators with missing values conditional on the auxiliary variables and the indicators without missing values. If the assumption of MAR is violated, the missing data mechanism is called not missing at random (NMAR), in which case maximum likelihood estimation under MAR yields biased estimates (Allison, 2001; Little & Rubin, 1989).

During the past years, the practice of stepwise latent class (LC) modeling using bias-adjusted stepwise approaches has become widely popular (Asparouhov & Muthén, 2014; Bakk & Kuha, 2018; Vermunt 2010). However, little is known on how these new approaches are affected by missing values on the class indicators. In the first step of the stepwise approaches, an LC model is estimated without the inclusion of the auxiliary variables. In the estimation of this step-one model, missing values can be handled in the usual way as long as the MAR assumption holds. The second step in a three-step analysis involves obtaining classifications based on the observed responses and the estimated parameters from step one. In the third step of a three-step LC analysis, we estimate the relationship between class membership and auxiliary variables using the predicted class memberships from step two while correcting for classification errors (Bolck, Croon, & Hagenaars, 2004; Vermunt, 2010). However, the current three-step approaches ignore the fact that the classification errors are larger for cases with missing values since they apply a single classification error correction matrix to all observations. The first question of interest is, therefore, whether we can simply ignore the differences in classification errors resulting from missing data, or whether we should account in some way for the missing value problem also in the third step of a three-step LC analysis.

The second potential problem arises when the missingness depends on the covariates or distal outcomes of interest. That is, when auxiliary variables affect the probability of having missing values on the class indicators, this

CONTACT Jeroen K. Vermunt  j.k.vermunt@uvt.nl  Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, PO box 90153, 5000 LE Tilburg, The Netherlands.

This article was originally published with errors, which have now been corrected in the online version. Please see Correction (<http://dx.doi.org/10.1080/10705511.2022.2123641>)

© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

corresponds to a MAR mechanism in a one-step LC analysis, but yields a not MAR (NMAR) mechanism in the first step of a stepwise LC analysis because the auxiliary variables are excluded from this analysis. The second question of interest, therefore, is how strongly parameter estimates of the stepwise LC approaches are affected by this type of violation of the MAR assumption.

To summarize, in this paper, we address the following two questions:

1. Is it correct to ignore the fact that classification errors are larger for observations with missing values, or should we address this in some way in the third step of a three-step LC analysis?
2. Are estimates of the relationship between class membership and auxiliary variables obtained with stepwise LC approaches strongly affected by possible violations of the MAR assumption in the step-one model?

Note that the second question is relevant for both three-step and two-step LC analysis, while the first question is relevant only for three-step approaches.

The next section describes the design of our study, which is based on the analysis of synthetic data sets corresponding to LC models with covariates and with missing values on the class indicators. These data sets vary in the missing data mechanism, the proportion of missing values, and the separation between classes. Next, we present the results of the analyses of these synthetic datasets, where we focus on the amount of bias in the covariate effects when using stepwise LCA methods. The paper ends with a conclusion and discussion section.

2. Method

Figure 1 depicts the four LC population models we are going to focus on. These models consist of three covariates (Z_1 – Z_3) affecting class membership (X), and six dichotomous indicators (Y_1 – Y_6). The three covariates have five equidistant values ranging from -2 to 2 . Indicators Y_3 and Y_6 may contain missing values for some persons, which is indicated using the missing value indicators I_3 and I_6 . Figure 1a represents the missing completely at random (MCAR) mechanism since I_3 and I_6 are independent of the other variables in the model. Figure 1b corresponds with a MAR mechanism in which I_3 depends on Y_1 and Y_2 and I_6 on Y_4 and Y_5 . Figure 1c assumes that I_3 depends on Z_1 and I_6 on Z_2 , which is in agreement with a MAR mechanism when Z_1 and Z_2 are included in the model, but which becomes an NMAR mechanism when estimating the model without the covariates included. Figure 1d represents a specific type of NMAR mechanism in which missingness on Y_3 and Y_6 (thus I_3 and I_6) depends on the latent variable X . We will refer to these four missing data mechanisms as MCAR, MAR-Y, MAR-Z, and NMAR-X.

For the LC model part of these population models, we used the same specifications as in Vermunt (2010). The model was a three-class model with equal class proportions,

where the class-specific response probabilities for the six dichotomous indicators were chosen to create low, moderate, and high class separation conditions (corresponding with entropy R-squared values of .36, .65, and .90, respectively, without missing data). For the moderate separation condition, in Class 1, the success probabilities were .80 for all indicators, in Class 2, .80 for the first three and .20 for the last three indicators, and in Class 3, .20 for all indicators. These probabilities were replaced with .70 (.30) for the low and with .90 (.10) for the high separation conditions. The Z_1 , Z_2 , and Z_3 slope parameters in the logistic model for covariate effects of the classes are set to 2, 0, and 0 for Class 2 and to 2, -1 , and 0 for Class 3. By setting the Class 2 and 3 intercepts to .867 and .709, we obtained equal class proportions.

Depending on the missing data mechanism, the likelihood of having a missing value on Y_3 and/or Y_6 depended on values chosen for $P(I_3)$ and $P(I_6)$, $P(I_3|Y_1, Y_2)$ and $P(I_6|Y_4, Y_5)$, $P(I_3|Z_1)$ and $P(I_6|Z_2)$, or $P(I_3|X)$ and $P(I_6|X)$. These probabilities were modelled using logistic equations with main effects. With the value for the intercept, we varied the overall proportion of missing values, yielding conditions with a small, a medium, and a large proportion of missing data on Y_3 (16%, 26%, 36%, respectively) and Y_6 (24%, 34%, 44%, respectively). In the MAR-Y condition, we set the slope parameters for the effects of Y_1 and Y_2 on I_3 and the effects of Y_4 and Y_5 on I_6 to .5, 1, and 1.5, yielding conditions with weak, medium, and strong effects of indicators on missingness. Similarly, in the MAR-Z condition, we set the slopes for the effect of Z_1 effect on I_3 and of Z_2 on I_6 to .5, 1, and 1.5 to manipulate the effect of covariates on missingness. In the NMAR-X condition, we manipulated the contrast between Class 3 and the other two classes (Class 3 having higher missing value probabilities), with slope parameters equal to .5, 1, and 1.5 for weak, medium, and strong NMAR effects.

By varying the overall proportion of missing values and the MAR and NMAR effect sizes, we created 3 MCAR, 9 MAR-Y, 9 MAR-Z, and 9 NMAR-X conditions (thus 30 missing data conditions). Each of the missing data conditions was combined with the 3 different class separation conditions, yielding a total of 90 conditions.

Since we are only interested in bias and not in sampling variability, instead of randomly generating a large number of replication data sets, for each condition in our study design, we created a single synthetic data set that is exactly in agreement with the population concerned. These are data sets containing all possible response patterns (including those with missing values) with frequency weights equal or proportional to the population probability for the response pattern concerned. We created these data sets using the Latent GOLD “writeexemplarydata” output option and used R to transform the Y_3 value to missing when I_3 equals 1 and the Y_6 value to missing when I_6 equals 1. These “exemplary” data sets were analyzed with Latent GOLD using three-step LC analysis with modal class assignments and ML bias adjustment, three-step LC analysis with proportional class assignments and ML bias adjustment, and two-step LC analysis. Appendix A illustrates how the

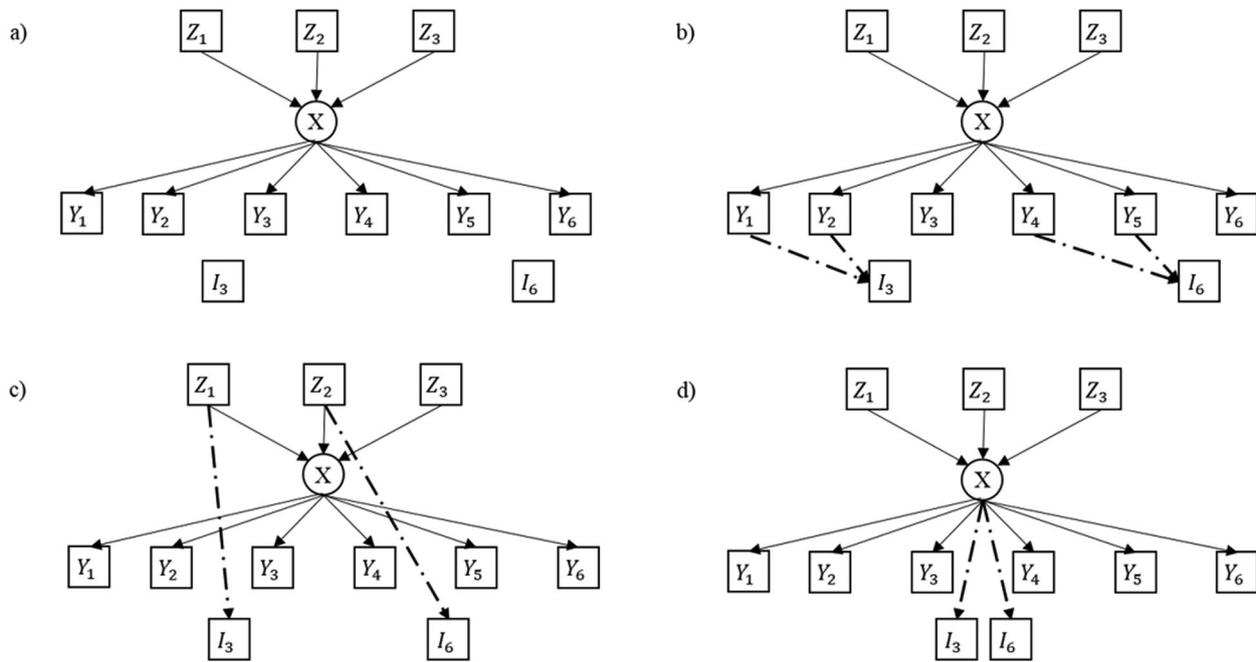


Figure 1. LC model with covariates and MCAR (a), MAR depending on observed indicators (b; MAR-Y), MAR depending on covariates (c; MAR-Z), and NMAR depending on latent classes (d, NMAR-X) missing data mechanisms.

Table 1. Probability of modal class assignment (W) given the true class membership (X), both overall and per missing data pattern, for the MCAR condition with moderate class separation and medium proportion of missing values.

		Assigned class W			
		True class X	1	2	3
		1	0.90	0.08	0.02
		2	0.18	0.74	0.08
		3	0.06	0.14	0.80
Missing data	Pattern	True class X	Assigned class W		
I_3	I_6		1	2	3
0	0	1	0.89	0.09	0.01
		2	0.10	0.80	0.09
		3	0.05	0.09	0.85
0	1	1	0.92	0.04	0.04
		2	0.33	0.57	0.10
		3	0.05	0.07	0.88
1	0	1	0.88	0.10	0.02
		2	0.10	0.86	0.04
		3	0.04	0.32	0.63
1	1	1	0.90	0.09	0.01
		2	0.30	0.67	0.04
		3	0.10	0.28	0.61

The probability of a correct assignment is printed in bold face.

synthetic data sets were created and how the different steps of the analyses were performed. For comparison with the stepwise approaches, the data sets were also analyzed using one-step LC models which include the covariates directly.

We expect that the stepwise approaches will yield biased estimates of the covariate effects on the classes under the MAR-Z condition. The NMAR-X condition was added for comparison purposes only and can be expected to yield biased estimates with both one-step and stepwise estimation.

For the three-step LC analysis, we hypothesized that even MCAR or MAR-Y may be problematic because the amount of classification errors depends on the missing data pattern.

To illustrate this point, in Table 1, we take the MCAR case with moderate class separation and medium proportion of missing values as an example, and present the overall probabilities of modal class assignment W conditional of the true class membership X , as well as the values for the four the patterns with $(I_3 = 0, I_6 = 0)$, $(I_3 = 1, I_6 = 0)$, $(I_3 = 0, I_6 = 1)$, and $(I_3 = 1, I_6 = 1)$. As can be seen, the classification probabilities $P(W|X)$ are affected by the presence of missing values. The class 2 predictions are more uncertain when $I_3 = 1$, and the class 3 predictions when $I_6 = 1$.

Note that besides the missing data mechanism, we manipulated the class separation, the effects of the Y s, Z s, and X on the missingness, and the proportion of missing values in order to see whether these factors affect the amount of bias of a three-step LC analysis. More specifically, we expect to encounter larger biases with lower class separation since classification errors are larger in those situations, with larger effects of the Z s and X on missingness because of the resulting larger deviation from MAR in the step-one model, and with larger proportions of missing values because of larger overall impact of missingness.

3. Results

This section presents the results obtained with the 90 investigated conditions. We summarize the overall bias as the mean absolute bias (MAB) across the six covariate effects on the classes. We also look at the bias in one selected parameter β_{12} , representing the effect of Z_1 for $X = 2$.

3.1. MCAR and MAR-Y Conditions

The average absolute bias was exactly 0 for all stepwise approaches and the one-step approach under all MCAR and

Table 4. Mean absolute bias (MAB) across the six covariate effects when missingness depends on the latent classes (NMAR-X condition).

Method	Missingness proportion	Low separation			Moderate separation			High separation		
		Weak effect	Medium effect	Strong effect	Weak effect	Medium effect	Strong effect	Weak effect	Medium effect	Strong effect
Three-step modal	Small	0.02	0.03	0.04	0.02	0.03	0.03	0.01	0.02	0.02
Three-step modal	Medium	0.04	0.07	0.10	0.03	0.05	0.07	0.02	0.03	0.04
Three-step modal	Large	0.06	0.12	0.20	0.04	0.08	0.11	0.02	0.05	0.06
Three-step proportional	Small	0.02	0.03	0.04	0.01	0.01	0.02	0.00	0.01	0.01
Three-step proportional	Medium	0.02	0.04	0.05	0.01	0.02	0.03	0.01	0.01	0.01
Three-step proportional	Large	0.03	0.05	0.06	0.02	0.03	0.04	0.01	0.02	0.02
Two-step	Small	0.01	0.01	0.02	0.01	0.01	0.01	0.00	0.01	0.01
Two-step	Medium	0.01	0.03	0.04	0.01	0.02	0.03	0.00	0.01	0.01
Two-step	Large	0.02	0.04	0.06	0.01	0.03	0.04	0.01	0.01	0.02
One-step	Small	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
One-step	Medium	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.01
One-step	Large	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01

Table 5. Bias in the β_{12} parameter with a true value 2.00 when missingness depends on the latent classes (NMAR-X condition).

Method	Missingness proportion	Low separation			Moderate separation			High separation		
		Weak effect	Medium effect	Strong effect	Weak effect	Medium effect	Strong effect	Weak effect	Medium effect	Strong effect
Three-step modal	Small	-0.04	-0.07	-0.09	-0.04	-0.06	-0.07	-0.03	-0.05	-0.06
Three-step modal	Medium	-0.08	-0.17	-0.23	-0.06	-0.12	-0.15	-0.04	-0.08	-0.10
Three-step modal	Large	-0.13	-0.28	-0.49	-0.09	-0.18	-0.26	-0.06	-0.11	-0.15
Three-step proportional	Small	0.04	0.06	0.07	0.01	0.02	0.02	0.00	0.00	0.00
Three-step proportional	Medium	0.05	0.07	0.07	0.02	0.02	0.02	0.00	0.00	0.00
Three-step proportional	Large	0.06	0.07	0.04	0.02	0.02	0.01	0.00	-0.01	-0.02
Two-step	Small	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00
Two-step	Medium	0.00	-0.01	-0.02	0.00	-0.01	-0.01	0.00	-0.01	-0.01
Two-step	Large	0.00	-0.03	-0.06	-0.01	-0.02	-0.03	0.00	-0.01	-0.01
One-step	Small	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
One-step	Medium	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00
One-step	Large	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.00	-0.01

4. Conclusions and Discussion

In this study, we examined the performance of stepwise LC methods with regard to the recovery of covariate effects in the presence of missing data on the class indicators. We examined four mechanisms for the missing data, namely MCAR, MAR-Y, MAR-Z, and NMAR-X, and manipulated three factors within each mechanism, namely the proportion of missing data, the effect of indicators/covariates/latent classes on missingness, and the class separation.

Contrary to what we expected, estimates obtained using stepwise LC methods are not biased with missing values on the class indicators when the MAR assumption holds in the step-one model estimation stage. This assumption holds when the missingness is MCAR or when it depends only on the indicators that are observed (our MAR-Y condition). This result answers the first research question we formulated in the introduction; that is, no modifications are needed when applying three-step LC analysis with missing values as long as the MAR assumption holds.

As expected, when missingness depends on covariates (our MAR-Z condition), the stepwise approaches may yield biased parameter estimates, where bias increases with a larger proportion of missing values, a stronger effect of covariates on missingness, and a lower separation between classes. Our most important and rather unexpected finding is that amount of bias varies strongly across the various stepwise approaches. More specifically, three-step LC analysis with modal class assignments is much more strongly

affected by the resulting NMAR missing data than the other two stepwise LC approaches.

An explanation for the rather small biases encountered with the two-step approach is that the step-one measurement model parameters were not strongly affected by violating MAR assumption, even in the least favorable conditions. For instance, in the most difficult condition (low class separation, large effects of covariates on missingness, and high proportion of missingness), the largest bias in the class-specific response probabilities was -0.02 (.28 instead of .30). These almost correct step-one response probabilities are treated as fixed measurement model parameters in the two-step approach, which explains the low bias. Proportional class assignment performs slightly worse than the two-step approach because it also uses the estimated class proportions from the step-one model to obtain the posteriors that serve as weights in the step-three analysis. The largest bias in the class proportions was -0.017 (0.316 instead of 0.333) in the least favorable condition, which is again rather small and therefore explains why step-three with proportional assignment perform well. However, these seemingly small biases in the step-one parameters may have a much larger impact when using modal class assignments in which one transforms the highest posterior into a weight of 1 and the other ones to 0. An assignment to class 2 may suddenly change in an assignment to class 3. Most probably, this increases the number of classification errors quite a bit in the less favorable conditions, which cannot be compensated by the applied correction for classification errors which itself is based on the biased step-one parameters.

As expected, in the NMAR-X conditions, we always find a certain amount of bias, where again the three-step proportional and two-step approaches are less affected than the three-step modal approach. The step-one parameters showed a larger bias than in the MAR-Z condition, which explains why the two-step and three-step proportional approaches perform slightly worse in the NMAR-X condition. The one-step LC analysis approach performed very well in the NMAR-X condition, which can be explained by the fact that inclusion of covariates in the model improves the class separation substantially (in the least favorable condition, entropy R-squared increased from .30 to .57) and, moreover, causes one gets closer to MAR when the covariates are strongly related to the classes (they serve as a kind of proxy for the latent classes).

Based on our results, the practical recommendation for researchers who wish to use a stepwise LC analysis is to be cautious when there is missing data on the class indicators. If there is some evidence that missingness is related to auxiliary variables of interest (e.g., if males and females have clearly different missingness probabilities and one is interested in gender differences in class membership), it can be recommended to use either a three-step approach with proportional assignment, which is the default option in the Latent GOLD software, or a two-step approach, which is also available in Latent GOLD (Vermunt & Magidson, 2021a).

An alternative way to deal with the MAR-Z situation could be to make use of multiple imputation; that is, to impute the missing values on the indicators using a good imputation model (containing the auxiliary variables) prior to performing the stepwise LC analysis (Allison, 2000; Schafer, 1997; Vermunt, Van Ginkel, Van der Ark, & Sijtsma, 2008). Another option could be to include the auxiliary variables that are related to the missingness in the step-one model, which yields a procedure similar to the one proposed by Vermunt and Magidson (2021b) for dealing with stepwise LC analysis in the presence of measurement non-invariance.

As any study based on constructed data sets, also our study has certain limitations. First of all, because we analyzed data sets that are exactly in agreement with the assumed populations, we did not study the effect of sampling fluctuation on estimates of parameters and their standard errors. Another limitation is that we postulated rather simplified missing data mechanisms, whereas in practice, the actual missing data mechanism may be much more complex, such as missingness being affected simultaneously by auxiliary variables, observed indicators, missing indicators, and latent classes. Moreover, we created missing values only on two of the six indicators, but in empirical applications, a larger portion of the indicators may contain missing values.

For practical reasons, we restricted ourselves to studying the bias in the covariate effects in LC models for dichotomous responses. However, we expect that our results also apply to the estimation of the association between class membership and distal outcomes, in which case one may prefer using the BCH instead of the ML estimation approach (Asparouhov & Muthén, 2021; Bakk & Vermunt, 2016, Nylund-Gibson, Grimm, & Masyn, 2019). Moreover, it can be expected that our results generalize to LC models with continuous

indicators, also referred to as latent profile models (Lazarsfeld & Henry, 1968; Oberski, 2016), which may also contain missing values. Finally, our results are also relevant for mixture growth or latent trajectory models (Muthén, 2004, Van de Schoot, Sijbrandij, Winter, Depaoli, & Vermunt, 2017), in which it is very common to have different numbers of measurements per individual, something that can also be seen as a missing data problem. In all these situations, it can be expected that missing data is not an issue when the MAR assumption holds in the step-one model. But when missingness depends on auxiliary variables, also in these situations it can be recommended not to use a three-step LC analysis with modal class assignments, but instead, a three-step LC analysis with proportional class assignments or a two-step LC analysis.

ORCID

Ö. Emre C. Alagöz  <http://orcid.org/0000-0003-3305-6564>
 Jeroen K. Vermunt  <http://orcid.org/0000-0001-9053-9330>

References

- Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods & Research*, 28, 301–309.
- Allison, P. D. (2001). *Missing data (quantitative applications in the social sciences)* (1st ed.). SAGE Publications, Inc.
- Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling*, 21, 329–341.
- Asparouhov, T., & Muthén, B. (2021). Auxiliary variables in mixture modeling: Using the BCH method in Mplus to estimate a distal outcome model and an arbitrary secondary model. *Mplus Web Notes*, 21
- Bakk, Z., & Kuha, J. (2018). Two-step estimation of models between latent classes and external variables. *Psychometrika*, 83, 871–892.
- Bakk, Z., & Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling*, 23, 20–31.
- 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.
- Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2, 222
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215–231.
- Lanza, S. T., Dziak, J. J., Huang, L., Wagner, A., & Collins, L. M. (2015). *Proc LCA & Proc LTA users' guide (Version 1.3. 2)*. The Methodology Center, Penn State.
- Lazarsfeld, P., & Henry, N. (1968). *Latent structure analysis*. Houghton Mifflin.
- Linzer, D. A., & Lewis, J. B. (2011). polCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42, 1–29.
- Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, 18, 292–326.
- Little, R. J., & Schenker, N. (1995). Missing data. In *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39–75). Springer.
- Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345–368). Sage
- Muthén, L. K., & Muthén, B. O. (2015). *Mplus (Version 7.4)*. Muthén & Muthén.
- 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*, 26, 967–985.

- Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. In *Modern statistical methods for HCI* (pp. 275–287). Springer.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall.
- Van de Schoot, R., Sijbrandij, M., Winter, S. D., Depaoli, S., & Vermunt, J. K. (2017). The GROLTS-checklist: Guidelines for reporting on latent trajectory studies. *Structural Equation Modeling, 24*, 451–467.
- Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*, 450–469.
- Vermunt, J. K., & Magidson, J. (2013). *Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax*. Statistical Innovations Inc.
- Vermunt, J. K., & Magidson, J. (2021a). *Upgrade manual for Latent GOLD 6.0*. Statistical Innovations Inc.
- Vermunt, J. K., & Magidson, J. (2021b). How to perform three-step latent class analysis in the presence of measurement non-invariance or differential item functioning. *Structural Equation Modeling, 28*, 356–364.
- Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology, 38*, 369–397.

Appendix A. Latent GOLD 6.0 Syntax

Latent GOLD 6.0 (Vermunt & Magidson, 2021a) was used to create synthetic data sets which are exactly in agreement with the assumed populations. This is example syntax for the MAR-Z model with medium effects of z_1 and z_2 on i_3 and i_6 , medium proportion of missing values, and moderate class separation:

```
options
  algorithm emiterations = 0, niterations = 0;
  output parameters=first writeexemplarydata='data.txt';
variables
  caseweight freq1000;
  dependent y1 2, y2 2, y3 2, y4 2, y5 2, y6 2, i3
  2, i6 2;
  independent z1, z2, z3;
  latent Class nominal 3;
equations
  Class <- 1 + z1 + z2 + z3;
  y1 - y6 <- 1 | Class;
  i3 <- -1 + z1;
  i6 <- -1 + z2;
  {0.867 0.709 2 2 0 -1 0 0
  1.386294361 1.386294361 -1.386294361
  1.386294361 1.386294361 -1.386294361
  1.386294361 1.386294361 -1.386294361
  1.386294361 -1.386294361 -1.386294361
  1.386294361 -1.386294361 -1.386294361
  1.386294361 -1.386294361 -1.386294361
  -1.49 1
  -.95 1}
```

The input data file contains one record for each of the 125 covariate patterns, with the eight dependent variables (six items and two missing data indicators) set to 0 and a frequency weight yielding an arbitrary total sample size (here it equals to 8 for each pattern, yielding a total of sample size of 1000). Specific in this syntax is that we set the number of EM and Newton-Raphson iterations to 0 (to fix the parameter values to their starting values), that we use the output option “writeexemplarydata” (to obtain the data file that we need), and that we specify “starting values” for all model parameters at the end of the equations section (to define the population values). As can be seen, the variables section specifies the caseweight and the variables which are part of the model. For the dependent variables, we have to specify their number of categories, and for the latent variable we have to provide a name, define its scale types, and specify its number of categories. The population parameters are specified between “{,” where the first row defines the logit parameters of the model for the classes (the intercepts, and the effects of the three covariates), the next six rows define the

class-specific response logits for the six items, and the last two rows contain the logit parameters of the models for the two missing value indicators. Note that the output option “parameters=first” indicates that dummy coding is used for the logit parameters with the first category as the reference category. The output data file “data.txt” will contain all possible response patterns (thus $125 \times 6^2 \times 2^2$ rows) a with frequency weight equal to the population proportion derived for the specified parameter values times the total sample size (here 1,000).

Subsequently, in data file “data.txt,” the value of y_3 (y_6) should be replaced by a missing value if i_3 (i_6) equals 1, which can, for example, be done using R. Then, using the resulting data set, a step-one analysis can be performed, while writing the classification information to an output data file. For the two-step approach, the log of the class-specific response densities should be saved in the output data file. That is,

```
options
  missing includeall;
  output parameters=first standarderrors
  profile;
  outfile 'classification.txt' classification
  logdensity
  keep z1 z2 z3;
variables
  caseweight frequency;
  dependent y1, y2, y3, y4, y5, y6;
  latent Class nominal 3;
equations
  Class <- 1;
  y1 - y6 <- 1 | Class;
```

Important is the missing values option “includeall,” which is used to indicate that records with missing values should be kept in the analysis. The “outfile” option is used to write the posterior class memberships and the log of class-specific response densities to an output data file, which in addition should contain the three covariates (indicated with the “keep” option). The “variables” and “equations” sections are similar to those showed above, though quite a bit simpler since the covariates and the missing value indicators are not part of the step-one model. Moreover, there is no need to specify number of categories of the items since these can be derived from the input data file. Note that the “caseweight” contains the frequency counts which are in agreement with the specified population model. Though not really needed, starting values may be provided to make sure the classes come out in the “right” order.

The step-three model can be estimated using the data file “classification.txt.” With modal class assignments, the model syntax the looks as follows:

```
options
  step3 modal ml;
  output parameters=first standarderrors = robust
  estimatedvalues;
variables
  caseweight frequency;
  independent z1, z2, z3;
  latent Class nominal posterior=(Class#1
  Class#2 Class#3);
equations
  Class <- 1 + z1 + z2 + z3;
```

The option “step3” indicates the type of step-three analysis one wishes to perform. Note that Class#1, Class#2, and Class#3 are names of the variables in the data file “classification.txt,” which contain the posterior class membership probabilities. The “caseweight” is the same frequency count as was used in the step-one model. The equations section contains the logistic regression equations for the latent classes.

With proportional class assignments, we use “step3 proportional ml” instead of “step3 modal ml.” For a the step-2 analysis, the line “step3 modal ml;” can be removed, and posterior = (Class#1 Class#2 Class#3)” is replaced by “logdensity = (logdensity1 logdensity2 logdensity3).” The variables logdensity1, logdensity2, and logdensity3 contain the log of the classification response densities from the step-one model.