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

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 and Jeroen K. Vermunt

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) 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.

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...
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 $Y_5$ 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 −1, 0, 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 $Z_1$ 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 "writexemplarydata" 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
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.

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 $b_{12}$, 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

![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.](image)

<table>
<thead>
<tr>
<th>Assigned class $W$</th>
<th>True class $X$</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.90</td>
<td>0.08</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.18</td>
<td>0.74</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.06</td>
<td>0.14</td>
<td>0.80</td>
<td></td>
</tr>
</tbody>
</table>

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.

The probability of a correct assignment is printed in bold face.
MAR-Y conditions, thus irrespective of the proportion of missing values, the strength of the MAR-Y mechanism, and the class separation.

### 3.2. MAR-Z Conditions

The results obtained with the stepwise LC approaches for the MAR-Z mechanism are shown in Tables 2 and 3. As can be seen, the three-step modal approach has the largest absolute bias in all conditions, whereas the three-step proportional and two-step methods show almost zero absolute bias. The absolute bias in the covariate effect estimates increases with a larger proportion of missing data, a stronger MAR-Z effect, and a lower class separation. Furthermore, results in Table 2 show that a high class separation reduces the negative effect of large missing data proportions and strong covariate effects on missingness.

These results are confirmed if we look at the bias encountered for a selected parameter \( \beta_{12} \) (see Table 3). Again, the three-step modal approach yields estimates with a problematic amount of bias in almost all conditions. Especially in the more difficult scenarios (i.e., low class separation, large missing data proportion, and strong covariate effects on missingness), it performs much worse than the other two stepwise LC methods. Again, we can see that the three-step proportional and two-step methods yield estimates with either zero or close to zero bias in the moderate and high separation conditions. The two-step method performs slightly better than the three-step proportional method, mainly in the scenarios with extremely low class separation.

We also estimated step-one LC models, which as expected yielded no bias since the MAR assumption holds when the covariates affecting missingness are included in the model.

### 3.3. NMAR-X Conditions

Tables 4 and 5 present the mean absolute bias across the six covariate effects and the bias in the selected parameter \( \beta_{12} \) for the NMAR-X mechanism. As expected, we see biases with all LC methods. As with in the MAR-Z conditions, the three-step modal approach yields the largest bias in all conditions, whereas the three-step proportional and two-step LC methods produced estimates with relatively small biases. Among the latter two, the two-step LC approach has a slightly smaller bias than the three-step proportional approach. As the missing data proportions and the effects of class membership on missingness increase, the bias increases as well. Similar to what we saw for the MAR-Z mechanism, a high class separation reduces the bias.

The one-step approach performed much better than the stepwise approaches, especially in the less favorable conditions with low class separation, large effects of \( X \) on missingness, and large proportion of missing values.

### Table 2: Mean absolute bias (MAB) across the six covariate effects when missingness depends on covariates (MAR-Z condition).

<table>
<thead>
<tr>
<th>Method</th>
<th>Missingness proportion</th>
<th>Low separation</th>
<th>Moderate separation</th>
<th>High separation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Weak effect</td>
<td>Medium effect</td>
<td>Strong effect</td>
</tr>
<tr>
<td>Three-step modal</td>
<td>Small</td>
<td>0.04</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>0.11</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>0.13</td>
<td>0.22</td>
<td>0.28</td>
</tr>
<tr>
<td>Three-step proportional</td>
<td>Small</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Two-step</td>
<td>Small</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>One-step</td>
<td>Small</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

### Table 3: Bias in the \( \beta_{12} \) parameter with a true value 2.00 when missingness depends on covariates (MAR-Z condition).

<table>
<thead>
<tr>
<th>Method</th>
<th>Missingness proportion</th>
<th>Low separation</th>
<th>Moderate separation</th>
<th>High separation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Weak effect</td>
<td>Medium effect</td>
<td>Strong effect</td>
</tr>
<tr>
<td>Three-step modal</td>
<td>Small</td>
<td>0.09</td>
<td>0.19</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>0.15</td>
<td>0.30</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>0.20</td>
<td>0.37</td>
<td>0.51</td>
</tr>
<tr>
<td>Three-step proportional</td>
<td>Small</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>0.02</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>0.03</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Two-step</td>
<td>Small</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>One-step</td>
<td>Small</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Three-step modal Small
Three-step modal Medium
Three-step modal Large
Three-step proportional Small
Three-step proportional Medium
Three-step proportional Large
Two-step Small
Two-step Medium
Two-step Large
One-step Small
One-step Medium
One-step Large

<table>
<thead>
<tr>
<th>Method</th>
<th>Missingness proportion</th>
<th>Low separation</th>
<th>Moderate separation</th>
<th>High separation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Weak effect</td>
<td>Medium effect</td>
<td>Strong effect</td>
</tr>
<tr>
<td>Three-step modal</td>
<td>Small</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>Three-step modal</td>
<td>Medium</td>
<td>0.04</td>
<td>0.07</td>
<td>0.10</td>
</tr>
<tr>
<td>Three-step modal</td>
<td>Large</td>
<td>0.06</td>
<td>0.12</td>
<td>0.20</td>
</tr>
<tr>
<td>Three-step proportional</td>
<td>Small</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>Three-step proportional</td>
<td>Medium</td>
<td>0.02</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>Three-step proportional</td>
<td>Large</td>
<td>0.03</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>Two-step</td>
<td>Small</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Two-step</td>
<td>Medium</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>Two-step</td>
<td>Large</td>
<td>0.02</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>One-step</td>
<td>Small</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>One-step</td>
<td>Medium</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>One-step</td>
<td>Large</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

|                      | Weak effect | Medium effect | Strong effect |
| Three-step modal     | 0.02        | 0.03          | 0.03          |
| Three-step modal     | 0.03        | 0.05          | 0.07          |
| Three-step modal     | 0.04        | 0.08          | 0.11          |
| Three-step proportional | 0.01       | 0.01          | 0.02          |
| Three-step proportional | 0.01       | 0.02          | 0.03          |
| Three-step proportional | 0.02       | 0.03          | 0.04          |
| Two-step             | 0.01        | 0.01          | 0.02          |
| Two-step             | 0.01        | 0.02          | 0.03          |
| Two-step             | 0.02        | 0.03          | 0.04          |
| One-step             | 0.00        | 0.00          | 0.00          |
| One-step             | 0.00        | 0.00          | 0.00          |
| One-step             | 0.00        | 0.00          | 0.00          |

|                      | Weak effect | Medium effect | Strong effect |
| Three-step modal     | 0.01        | 0.02          | 0.02          |
| Three-step modal     | 0.02        | 0.03          | 0.04          |
| Three-step modal     | 0.03        | 0.05          | 0.07          |
| Three-step modal     | 0.04        | 0.08          | 0.11          |
| Three-step proportional | 0.00       | 0.01          | 0.02          |
| Three-step proportional | 0.00       | 0.02          | 0.03          |
| Three-step proportional | 0.01       | 0.03          | 0.04          |
| Two-step             | 0.01        | 0.01          | 0.02          |
| Two-step             | 0.01        | 0.02          | 0.03          |
| Two-step             | 0.02        | 0.03          | 0.04          |
| One-step             | 0.00        | 0.00          | 0.00          |
| One-step             | 0.00        | 0.00          | 0.00          |
| One-step             | 0.00        | 0.00          | 0.00          |

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 −.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 response probabilities 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.

**References**


**ORCID**

Ö. Emre C. Alagöz [orcid.org/0000-0003-3305-6564](http://orcid.org/0000-0003-3305-6564)

Jeroen K. Vermunt [orcid.org/0000-0001-9053-9330](http://orcid.org/0000-0001-9053-9330)
Chapman and Hall.
Van de Schoot, R., Biesebroek, A., J. K., W. Deurwaerder, M., Winter, S. D., Depaoli, S., & Vermunt,
Vermunt, J. K. (2010). Latent class modeling with covariates: Two
improved three-step approaches. Political Analysis, 18, 450–469.
GOLD 5.0. Basic, advanced, and syntax. Statistical Innovations Inc.
GOLD 6.0. Statistical Innovations Inc.
Vermunt, J. K., & Magidson, J. (2021b). How to perform three-step latent
class modeling with covariates: Two
improved three-step approaches.

### 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 z1 and z2 on i3 and i6, medium proportion of missing
values, and moderate class separation:

```latex
options
  algorithm emiterations = 0, nriterations = 0;
  output parameters = first writeexemplarydata = 'data.txt';
  caseweight freq1000;
  dependent y1, y2, y3, y4, y5, y6;
  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
     -1.49 1
     -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 num-
er of EM and Newton-Raphson iterations to 0 (to fix the parameter
value to their starting values), that we use the output option
`algorithm emiterations` (to obtain the data file that we need), and that
we set `caseweight frequency` (to change the 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 y3 (y6) should be
replaced by a missing value if i3 (i6) 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,

```latex
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 "output" 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 pro-
vided 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:

```latex
options
  step3 modal ml;
  output parameters = first standarderrors
    robust estimatedvalues;
  caseweight frequency;
  independent z1, z2, z3;
  latent Class nominal posterior = (Class1:
    Class2 Class3);

  equations
    Class <- 1 + z1 + z2 + z3;
```

The option "step3" indicates the type of step-three analysis one wishes
to perform. Note that Class1, Class2, and Class3 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 con-
tains 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 = (Class1 Class2 Class3)
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.