ResearchReport

Daniëlle Remmerswaal

3-12-2021

# 2. Methods

## 2.1. MILC

The MILC method, developed by Boeschoten et al. (2019), is a combination of latent class (LC) analysis and multiple imputation (MI), and is used to identify and correct classification errors. The typical use of latent variable models is for analysing multivariate response data. In contrast, the LC model in MILC is used to estimate the “true” category of units in combined datasets, that consist of multiple unit-linked categorical variables from different sources measuring the same attribute. Here, an overview of the six MILC steps is given and visualised in figure 1. In the next sections, each step will be explained more elaborately.  
1. m bootstrap samples are taken with replacement from a dataset with multiple variables. With m being equal to the number of imputations (see step 4).  
2. We apply a LC model with C classes on each of those m bootstrap samples. With C being equal to the number of categories of the variable of interest.  
3. We calculate the posterior membership probabilities for each of the m LC models  
4. By sampling from the obtained posteriors we create m imputations of the latent class and store them in a dataset with the original variables.  
5. Calculate the estimate of interest for each of the m imputations  
6. Pool the m estimates with Rubin’s pooling rules

### 2.1.1. Bootstrap

To reflect uncertainty caused by classification errors of the data in the m imputations, we create m bootstrap samples from our original dataset. The dataset consists of multiple unit-linked categorical variables from different sources measuring the same attribute. We use a multinomial distribution to sample from, with the frequencies of each response pattern as probabilities. This step results in m bootstrap samples of the same size as the original dataset.

### 2.1.2.LC

On each of the m bootstrap samples we apply an LC model. The basic idea behind latent class analysis (LCA) is to find latent variables based on the correlation structure of observed categorical “indicator” variables. Each of the J categorical variables has Kj possible outcomes for individuals i=1,,N. In our specific case each source measures the same attribute and has the same number of categories, so Kj = K. Each individual has a response on each of the J variables Y\_j (j=1,,J); there is no missing data. The vector of responses Y for one individual is called a response pattern. There are \_{j=1}^{J}K\_j , and in our case K^J, possible response patterns. In general, the number of latent classes is not known and has to be determined by the user of the LC model. When an LC model is used for correction of classification errors and all sources have the same number of categories, K\_j=C for all j=1,,J, and the number of latent classes is also fixed to C. Most LC models make use of modal assignment: an individual is assigned to the class for which he has the highest membership likelihood based on his response pattern on the indicator variables. Without classification errors, all sources would contain the correct category and all individuals would be assigned to the class to which they belong. In MILC, modal assignment is not used. Instead, values are assigned to the latent class by drawing from the posterior membership distribution (see below).

The two key model assumptions for latent class analysis are the local independence assumption and the mixture assumption. By assuming that responses within one class are independent of each other, we assume the correlation between them was solely caused by their class membership. With this assumption we can calculate the likelihood of a response pattern Y (combination of responses on the indicator variables) occurring given the observation is part of class c.

P(|X=c)=\_{j=1}^{J}P(Y\_j|X=c) (1) By assuming that the probability of obtaining a specific response pattern is the weighted sum of all C class-specific conditional response probabilities (mixture assumption) we can calculate the probability density function of a response pattern Y across all classes.

P()=\_{c=1}^{C}P(X=c)P(|X=c) (2)

class C=1

### 2.1.3. Posterior membership probabilities

By combining the mixture and local independence assumption and using Bayes’ rule, we are able to calculate the posterior membership probabilities of each response pattern, and thus for each individual.  
P(X=c|=)= (3)

The posterior membership probabilities of a response pattern for each class sums to one. \_{c=1}^{C}P(X=c|=)=1

### 2.1.4. Multiple Imputation (MI)

There are numerous methods for multiple imputation. Since we are not imputing a small proportion of missing categorical values, but the entire dataset, methods like imputing the mode are inappropriate. One other possible method is logistic regression based on auxiliary variables, but this is not applicable to the type of data we have. We use the posteriors, as calculated in the previous step, to impute the assign the classes to the individuals. The most straightforward method to impute the classes with the posteriors is modal assignment, which is used by the R package poLCA. Here, individuals are assigned to the class with the highest posterior membership probability. We use the posteriors as probabilities and assign individuals to a class by sampling from the posterior membership probabilities. We impute m variables based on the posteriors of each of the m bootstrapped samples. Five imputations are sufficient, as demonstrated by Boeschoten et al. (2019).

### 2.1.5. Calculation of estimates

We are interested in the proportions of the class sizes. … Frequencies of the classes.

### 2.1.6. Pooling of estimates

We pool the estimates of the imputations with Rubin’s pooling rules. (not adjusted, because in our data simulation with polca we sample with replacement)

## 2. Tree-MILC

Tree-MILC is potentially a useful method to distinguish the different (selection and classification) errors in the dataset, and obtain the correct estimates. The idea of implementing a tree-step originates from latent class tree (LCT) analysis, as developed by Bergh et al. (2018). In LCT analysis, the data is split and structured into classes by method of a sequential comparison of 1- and 2-class models. The top-down approach continues until the information criterion (e.g. BIC) no longer chooses 2-class models over 1-class models. The splits are performed based on the posterior class membership probabilities of each new class (child nodes) conditional on the class before the splitting (parent node). The main advantage of this method of latent class analysis is that it provides clear insight in splitting of the classes, the structure of the data, and how models with different number of classes are related to each other. The same steps as for MILC are followed, twice, with the LC model in the second phase being applied on a subset of the data. Subsequently, the two LC models extract a different number of classes, as will be elaborated on later. The process is visualised in figure 2.

Option 1: in step 4 we place the imputations next to the original dataset. In the second phase we continue with the subset of the original dataset based on the imputations (we delete the wrongfully selected units). Since we continue with a subset of the original dataset we have to bootstrap again to incorporate uncertainty due to classification errors.  
Option 2: in step 4 we place the imputations next to the corresponding bootstrap sample. We continue with the a subset of each bootstrap sample based on the corresponding imputations (we delete the wrongfully selected units). Since we are still working with the bootstraps the uncertainty due to classification errors is still reflected and we do not have to bootstrap again.

### 2.2.1. First phase of tree-MILC

We apply a LC model with S classes on each of those m bootstrapped samples. S = 2 when we have units that should be either selected or not. Now we select the cases we are interested in and delete the cases with selection errors from our pooled imputed dataset Step 4: combine with subset of original data? Of gewoon met de bootstrap samples? Dan kunnen we gelijk door. (zie option 1 en 2 hierboven)

### 2.2.1. Second phase of tree-MILC

with the subset of selected cases? M bootstrap samples are taken with replacement from the dataset. (this step is possibly unnecessary) We apply a LC model with C classes on each of those m bootstrapped samples. With C being the categories prone to classification error.

# References

Bergh, M. van den and Vermunt, J. K., Building Latent Class Growth Trees, *Structural Equation Modeling: A Multidisciplinary Journal*, vol. **25**, no. 3, pp. 331–42, accessed September 15, 2021, from <https://www.tandfonline.com/doi/full/10.1080/10705511.2017.1389610>, May 4, 2018. DOI: [10.1080/10705511.2017.1389610](https://doi.org/10.1080/10705511.2017.1389610)

Boeschoten, L., Waal, T. and Vermunt, J. K., Estimating the Number of Serious Road Injuries Per Vehicle Type in the Netherlands by Using Multiple Imputation of Latent Classes, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. **182**, no. 4, pp. 1463–86, accessed September 15, 2021, from <https://onlinelibrary.wiley.com/doi/10.1111/rssa.12471>, October 2019. DOI: [10.1111/rssa.12471](https://doi.org/10.1111/rssa.12471)