Nonlinear modeling – Mixture Modeling

Holger Brandt

June 06, 2017

1 SEMM 1: Direct application in a growth curve mixture model

The first example is a longitudinal data set on CD4 cell counts in HIV patients. CD4 cell counts are one of the main indicators (biomarkers) for the severity of the HIV infection; if they fall below a certain threshold, a patient is indicated with having AIDS. Here, we investigate the progress of the CD4 cell counts over the course of 32 weeks and four measurement occasions (each 8 week). Use an intercept and a linear slope to describe the development. Are there different groups of patients?
Use the `Mplus` demo version for this example (and the `MplusAutomation` package in `R`).

1. Draw a path diagram for the model. Specify which parameters should be class-specific and which ones should be class-invariant. How many parameters do you have for a 2, 3, or 4 group model?

2. How many classes do you need in this data set? Use the AIC/BIC to determine this number by running 3 different models (2 to 4 classes). A syntax file for a two-class model is provided. Extend it to 3 and 4 classes.

3. Investigate the results in detail and check for Heywood cases (negative variances).

4. Illustrate the results by using the factor scores obtained in Mplus for a (valid) two group solution. [code is provided.]

   Save the factor scores by including these two lines of code before the output:

   ```
   SAVE:   save = fscores cprob;
           file = fscores_gcmm2.dat;
   ```

   which saves factor scores and class membership probabilities.

5. In a second step, include the covariates cd4base and age that are available in the data set to predict the class membership.

- Does the number of necessary classes change?
- Which of the covariates are relevant?

6. An alternative to the growth curve mixture modeling (GCMM) approach is the heterogeneous growth curve model. It uses a specific nonlinear function for the residual of the slope factor to account for heteroskedasticity. Run the model `hgcm1.inp` and investigate the results. Compare the factor score estimates for the GCMM with 2 groups and those from the HGCM. Try to describe the differences in these solutions.

2 SEMM 2: An indirect application

The second example uses the PISA data set from Day 3. The data is saved in `pisa_sem_noname.dat` (for `Mplus`) and in `pisa_sem.dat` (for `R`). Again use the `Mplus` demo version for the model estimation and `R` for the rest. Due to the limitations of the demo version, we will only use two indicators for each latent construct.
We will use the mixture model here in an indirect approach, i.e. we do not try to extract subgroups of students but describe the (nonlinear) relationship between the two latent predictor variables (online activities and reading attitude) and the latent outcome (reading skills).

1. Draw a path diagram for the model. Specify which parameters should be class-specific and which ones should be class-invariant. How many parameters do you have for a 2, 3, or 4 group model? Assume that there are nonlinear relationships that can be approximated with linear regression coefficients in each group.

2. How many classes do you need in this data set? Use the AIC/BIC to determine this number by running 3 different models (2 to 4 classes). Start with the model `semm2_pisa_start.inp` file that leads you step by step through the necessary syntax (also have a look at the handout on Mplus syntax). Extend the model to 3 and four classes.

3. Illustrate the results for the two groups using the `plotSEMM` package. [code is provided.]

2.1 NLSEM

This model can also be specified in `nlsem`. We will go through the model code. However, the model does not converge in this package.

3 NSEMM

We now return to the pisa data set. In the script of day 3, we illustrated the distribution of the variables and found at least some nonnormality. Often this kind of nonnormality leads to biased parameter estimates when using the pi approaches or LMS. We now extend the syntax for a single class model to analyze the data set to include several classes.

1. Use the input file `lms_mplus.inp` and run the analysis.
2. Then extend the the model syntax to include class-specific means and (co)variances for the latent predictor variables. Fix all other parameters as class-invariant (check this by investigating the TECH1 part of the output).
3. How many classes do you think should one use to approximate the nonnormality? Which of the predictor variables do you think is more nonnormally distributed?
4. How to the results change across the different solutions?
5. Create standardized coefficients for the two-class solution.
6. Illustrate the model-implied distribution of the latent variables and the multivariate relationship between all latent variables. [use code provided.]