Using Mplus TECH11 and TECH14 to test the number of latent classes

Tihomir Asparouhov & Bengt Muthén

Mplus Web Notes: No. 14
May 22, 2012

Abstract

This note describes how to determine the number of classes in latent class analysis using Mplus. The Lo-Mendell-Rubin test and the bootstrapped likelihood test using TECH11 and TECH14 are considered. The note describes the K-1STARTS and LRTSTARTS options that are used with TECH11 and TECH14. It is pointed out that K-1STARTS makes it unnecessary to arrange the classes so that the last class is the largest. It is shown how to use LRTSTARTS to avoid warnings about loglikelihoods not being replicated in the bootstrap draws. The advantage of using OPTSEED is discussed. It is demonstrated how to do the analysis stepwise and how to avoid local optima. Following these steps can also reduce computational time and avoid ineffective analysis attempts. It is shown that the common mistake of using STARTS together with TECH11 and TECH14 leads to significantly longer computational time than using the proposed steps.

1 Introduction

Mplus can test the number of classes in a mixture analysis using the Lo-Mendell-Rubin (LMR; TECH11) test and the bootstrapped likelihood ratio test (BLRT; TECH14). A simulation study discussing these approaches is given in Nylund et al. (2007). Using an example, this web note gives a simple description of how to efficiently use the TECH11 and TECH14 options in practice in a stepwise fashion, avoiding practical problems such as local optima. This avoids the time-consuming approach of requesting TECH11 and TECH14 in all the analysis steps, where the steps are slow due to requesting many random starts. An example shows that the proposed stepwise approach can be more than three times faster.

2 Brief background for TECH11 and TECH14

As discussed in Nylund et al. (2007), the usual likelihood ratio chi-square test (2 times the loglikelihood difference) should not be used to test a k-1 versus k class model because 2 times the loglikelihood difference is not chi-square distributed in this situation. Mplus uses two alternatives, TECH11 and TECH14.

The TECH11 LMR test (Lo et al., 2001; see also Vuong, 1989) uses the correct distribution of 2 times the loglikelihood difference to perform this test. This is obtained by the k class run also doing a $k-1$ class analysis and using the derivatives from both models to compute the p-value. A low p-value rejects the $k-1$ class model in favor of the k class model.

The TECH14 BLRT test (McLachlan & Peel, 2000) is obtained by bootstrapping using the following procedure:

- (a) In the k class run, estimate also the $k-1$ class model to compute 2 times the loglikelihood difference

- (b) Use the $k-1$ class model estimates of step (a) to generate data and analyze by a $k-1$ class model (correct number of classes) and a k class model (one class too many) and calculate 2 times the loglikelihood difference. Repeat this step several times to give the true (bootstrap) distribution of 2 times the loglikelihood difference

- (c) Estimate the p-value for testing the $k-1$ class model against the k class model by comparing the step (a) value of 2 times the loglikelihood difference with the bootstrap distribution obtained in step (b).

The algorithm for the step (c) p-value computation is discussed in the appendix of Nylund et al. (2007). Rejection of the $k-1$ class model is obtained with a low p-value as usual. Note that step (a) uses the real data while step (b) uses generated data. Steps (a) - (c) are done by Mplus in the k class run when TECH14 is requested. The Mplus ANALYSIS options related to TECH11 and TECH14 are K-1STARTS and LRTSTARTS. These options are discussed next.

The K-1STARTS option can be used in conjunction with both TECH11 and TECH14 and refers to real-data analysis of the $k-1$ class model. The default setting in Version 7 is

K-1STARTS = 20 4;

(Version 6.12 has the setting 10 2). The two numbers of K-1STARTS work the same as for the STARTS option so that the first number refers to number of initial iterations and the second number refers to the number of final iterations. For TECH11, K-1STARTS is used for the $k-1$ class real-data analysis in the k

class run. For TECH14, K-1STARTS is used for the $k-1$ class real-data analysis in the k class run in step (a) above. The technical approach to the $k-1$ class analysis carried out in the k class run is to delete the first class. This means that the first class should not have any parameter restrictions. The absence of parameter restrictions is the typical case with latent class analysis. Because the K-1STARTS option uses random perturbations of the starting values for the $k-1$ class model, there is no need to arrange the classes in the k class run to have the largest class last (this was recommended before the K-1STARTS option was introduced).

The LRTSTARTS option is used only in conjunction with TECH14 and refers to the generated-data analyses of step (b). The default setting in Version 7 is

LRTSTARTS = 0 0 40 8;

(Version 6.12 has the setting 0 0 20 5). The first two numbers refer to the estimation of the $k-1$ class model and the last two numbers refer to the estimation of the k class model. Each pair of numbers works as for the STARTS option. In step (b), data are generated according to a $k-1$ class model using the real-data estimates. Analyzing these data using a $k-1$ class model uses the correct number of classes and this is why the first two numbers of LRTSTARTS are 0 0. Estimating the same data by the k class model, that is using one too many classes, typically requires more starts, motivating the default of 40 8. These last two numbers sometimes need to be increased to obtain replicated solutions. The following example shows how the different steps play out in practice.

3 SVALUES as an alternative to K-1STARTS

When the number of classes is fairly large, the K-1STARTS command may fail to find the best solution for the K-1 class. This is because in the current implementation, the K-1 class model estimation always uses as starting values the estimates for classes 2,3,..., K. Direct control to the starting values for the K-1 class estimation is currently unavailable. In Mplus 8.5, the SVALUES option has been generalized so that it can automatically reorder the classes for the K-class run. This can be very useful for the purposes of TECH11 and TECH14. The idea behind this approach is as follows. In the K-class run, using SVALUES, we can reorder the classes so that the smallest class appears first. When the K-1 class is estimated as a part of the TECH11 or TECH14 computation, that first class which is now also the smallest class will be removed, thereby yielding the best chance to quickly find the optimal K-1 solution (without any random starting values based on K-1STARTS). The assumption that underlines the idea is that the best K-1 class solution is similar to the best K class solution without its smallest class.

The procedure can be outlined as follows. Both the K and K-1 class models are estimated (without TECH11 and TECH14) with a large number of STARTS to ensure that the most optimal solutions are found. Then using the best OPTSEED of the K class solution and the SVALUES option we can produce the best K class solution where the first class is the smallest. For example, if class 3 is the smallest class, using OPTSEED and OUTPUT:SVALUES(3 2 1 4 ... K) will place the smallest class as the first class. To that run, we can add TECH11 and TECH14. If the best K-1 solution is still not found, then adding large K-1STARTS can be attempted. If that also fails to produce the best solution, yet

another avenue to pursue is facilitating the STSCALE option on top of that. To do so, the OPTSEED plus SVALUES reordering will be used to produce starting values for the K class solution with the smallest class as class 1. These starting values are conveniently printed as a model setup in the SVALUE output. Using that model setup in combination with STARTS=0 (without OPTSEED), varying values of STSCALE and K-1STARTS in the K class run offers another possibility for getting to the best K-1 solution.

4 An example: Analyzing the number of latent classes of antisocial behavior

Data on 17 antisocial behavior items are obtained from the National Longitudinal Survey of Youth (NLSY). A sample of $n = 7,326$ subjects ages 16 to 23 is considered here. The items concern the frequency of various behaviors during the past year. For the present purpose, these items are dichotomized and scored 0/1 with 0 representing never in the last year. The items are: damaged property, fighting, shoplifting, stole less than $50, stole more than $50, use of force, seriously threaten, intent to injure, use marijuana, use other drugs, sold marijuana, sold hard drugs, "con" someone, take auto, broken into building, held stolen goods, gambling operation. Consider a latent class analysis of the 17 antisocial behavior items using 4 and 5 latent classes, representing the $k-1$ and k class models of interest. Following are the three recommended analysis steps. The analyses are carried out in Mplus Version 6.12 and it is expected that different defaults are used in later versions.

4.1 Step I: Replicating the best loglikelihood for $k-1$ and k classes in the real data

As a first step, the best loglikelihood values for the 4- and 5-class models should be found. In this step, TECH11 and TECH14 are not requested.

The 4-class solution replicates the best loglikelihood several times even when using the Version 7 default number of random starting value pertubations, using STARTS = 20 4 (Version 6.12 has 10 2). The best loglikelihood is −41007.498. To verify that a better loglikelihood cannot be obtained, a second run increases the number of random starting value perturbations to STARTS = 100 20 and finds the same best replicated loglikelihood.

With 5 classes, the default starts setting of 20 4 is not sufficient to obtain replication of the best loglikelihood as evidenced when comparing to a run with STARTS = 200 40. This run, however, does not replicate the best loglikelihood. Making a further increase to STARTS = 600 120 replicates the best loglikelihood value also found with STARTS = 200 20. Table 1 shows the input for the 5-class run with STARTS = 600 120. Table 2 shows that the best loglikelihood is −40808.314 and that it is replicated four times. Using STARTS = 600 160 further supports that this is the best loglikelihood value, replicating it 12 times.

Table 1: Step I input for 5-class LCA for 17 ASB items

TITLE:	
DATA:	FILE = asb.dat;
	FORMAT = 34X 54F2.0;
VARIABLE:	NAMES = property fight shoplift lt50 gt50 force threat injure pot drug soldpot solddrug con auto bldg goods gambling dsm1-dsm22 sex black hisp single divorce dropout college onset f1 f2 f3 age94 cohort dep abuse;
	USEVARIABLES = property-gambling;
	CATEGORICAL = property-gambling;
	CLASSES = c(5);
ANALYSIS:	TYPE = MIXTURE;
	STARTS = 600 120;
	PROCESSORS = 4(STARTS);
OUTPUT:	

Table 2: Step I loglikelihood values for 15 random starts for 5-class LCA for 17 ASB items

RANDOM STARTS RESULTS RANKED FROM THE BEST TO THE WORST LOGLIKELIHOOD VALUES

Final stage loglikelihood values at local maxima, seeds, and initial stage start numbers:

-40808.314	991329	182
-40808.314	195353	225
-40808.314	329127	185
-40808.314	862607	468
-40808.406	66276	217
-40808.406	566739	575
-40808.406	248742	556
-40808.406	783165	170
-40808.406	863691	481
-40809.657	723035	538
-40815.960	486622	522
-40815.960	373505	88
-40815.960	928287	197
-40815.960	741484	441
-40815.960	404510	442

4.2 Step II: Using OPTSEED and adding TECH11

As a second step, a 5-class analysis is done requesting TECH11 in the OUTPUT command. TECH11 gives faster computations than TECH14 and should therefore be investigated first, making sure that the TECH11 output for the $k-1$ class model, here 4 classes, shows the best loglikelihood value found in Step I. To do this run efficiently, the OPTSEED value 991329 from the run shown in Table 2 is used in a 5-class run that uses STARTS=0. This means that the run uses the starting values that gave the best loglikelihood so that no random perturbation of the starting values is needed. This saves considerable computational time. The input for this run is shown in Table 3.

Table 4 shows that the TECH11 test of the 4-class model against the 5-class model has the loglikelihood value -41007.498 for the H_0 hypothesis, which is the 4-class model. This is the correct value also obtained in the search of Step I above. Given the agreement, this means that in this example the K-1STARTS option does not need to be used to search for the best loglikelihood. The 2 times the loglikelihood difference is 398.368, which uses the 4- and 5-class loglikelihood values also obtained in Step I. The mean and standard deviation of the derived distribution for 2 times the loglikelihood difference is 38.441 and 41.386, respectively. Although the distribution is not normal, this indicates that a value of 398.368 is unlikely. The p-value is zero, rejecting the 4-class model when comparing to the 5-class model.

Table 3: Step II input for 5-class LCA for 17 ASB items using OPTSEED and TECH11

```
TITLE:
DATA:       FILE = asb.dat;
            FORMAT = 34X 54F2.0;
VARIABLE:   NAMES = property fight shoplift lt50 gt50 force threat
            injure pot drug soldpot solddrug con auto bldg goods
            gambling dsm1-dsm22 sex black hisp single divorce dropout
            college onset f1 f2 f3 age94 cohort dep abuse;
            USEVARIABLES = property-gambling;
            CATEGORICAL = property-gambling;
            CLASSES = c(5);
ANALYSIS:   TYPE = MIXTURE;
            STARTS = 0;
            OPTSEED = 991329;
            PROCESSORS = 4(STARTS);
OUTPUT:     TECH11;
```

Table 4: Step II output for for 5-class LCA for 17 ASB items using OPTSEED and TECH11

TECHNICAL 11 OUTPUT

Random starts specifications for the k-1 class analysis model	
Number of initial stage random starts	10
Number of final stage optimizations	2
Vuong-Lo-Mendell-Rubin likelihood ratio test for 4 (H0) versus 5 classes	
H0 loglikelihood value	-41007.498
2 times the loglikelihood difference	398.368
Difference in the number of parameters	18
Mean	38.441
Standard deviation	41.386
P-value	0.0000
Lo-Mendell-Rubin adjusted LRT test	
Value	395.896
P-value	0.0000

4.3 Step III: Using OPTSEED and adding TECH14

As a third step, a 5-class analysis is done requesting TECH14 in the OUTPUT command. The same OPTSEED value is used as in Step II. The input is shown in Table 5. Table 6 shows that the TECH14 test receives a warning that the best loglikelihood has not been replicated during the bootstrap runs. This implies that the default values of 20 5 (Version 6.12; Version 7 uses 40 8) for the last two numbers of the LRTSTARTS option did not use a sufficient number of random starts perturbations but have to be increased. The analysis therefore has to be modified.

Table 7 shows that increasing the perturbations in LRTSTARTS as LRTSTARTS = 0 0 100 20; avoids the TECH14 warning as seen in Table 8. The p-value is zero, rejecting the 4-class model when comparing to the 5-class model.

In this application, TECH11 and TECH14 do not help in deciding on the number of latent classes because their p-values are zero for any reasonable number of latent classes. This outcome is not surprising given that the 17 items of the antisocial behavior measurement instrument were not designed to capture latent classes. Other models can be tried, such as a factor analysis model or a factor mixture model. On the other hand, the 5-class latent class model may be a good enough approximate model and does give a reasonable interpretation. The estimated model can be used to modify the measurement instrument, deleting and adding items to better represent latent classes.

Table 5: Step III input for 5-class LCA for 17 ASB items using OPTSEED and TECH14

TITLE:	
DATA:	FILE = asb.dat;
	FORMAT = 34X 54F2.0;
VARIABLE:	NAMES = property fight shoplift lt50 gt50 force threat injure pot drug soldpot solddrug con auto bldg goods gambling dsm1-dsm22 sex black hisp single divorce dropout college onset f1 f2 f3 age94 cohort dep abuse;
	USEVARIABLES = property-gambling;
	CATEGORICAL = property-gambling;
	CLASSES = c(5);
ANALYSIS:	TYPE = MIXTURE;
	STARTS = 0;
	OPTSEED = 991329;
	PROCESSORS = 4(STARTS);
OUTPUT:	**TECH14**;

Table 6: Output for 5-class LCA for 17 ASB items using OPTSEED and TECH14

TECHNICAL 14 OUTPUT	
Random starts specifications for the k-1 class analysis model	
Number of initial stage random starts	10
Number of final stage optimizations	2
Random starts specifications for the k-1 class analysis model for generated data	
Number of initial stage random starts	0
Number of final stage optimizations for the initial stage random starts	0
Random starts specifications for the k class model for generated data	
Number of initial stage random starts	20
Number of final stage optimizations	5
Number of bootstrap draws requested	Varies
Parametric bootstrapped likelihood ratio test for 4 (H0) versus 5 classes	
H0 loglikelihood value	-41007.498
2 times the loglikelihood difference	398.368
Difference in the number of parameters	18
Approximate p-value	0.0000
Successful bootstrap draws	5
WARNING: THE BEST LOGLIKELIHOOD VALUE WAS NOT REPLICATED IN 5 OUT OF 5 BOOTSTRAP DRAWS. THE P-VALUE MAY NOT BE TRUSTWORTHY DUE TO LOCAL MAXIMA. INCREASE THE NUMBER OF RANDOM STARTS USING THE LRTSTARTS OPTION.	

Table 7: Step III input for 5-class LCA for 17 ASB items using OPTSEED, TECH14 and adding LRTSTARTS 0 0 100 20

TITLE:	
DATA:	FILE = asb.dat;
	FORMAT = 34X 54F2.0;
VARIABLE:	NAMES = property fight shoplift lt50 gt50 force threat injure pot drug soldpot solddrug con auto bldg goods gambling dsm1-dsm22 sex black hisp single divorce dropout college onset f1 f2 f3 age94 cohort dep abuse;
	USEVARIABLES = property-gambling;
	CATEGORICAL = property-gambling;
	CLASSES = c(5);
ANALYSIS:	TYPE = MIXTURE;
	STARTS = 0;
	OPTSEED = 991329;
	PROCESSORS = 4(STARTS);
	LRTSTARTS = 0 0 100 20;
OUTPUT:	TECH14;

Table 8: Step III output for 5-class LCA for 17 ASB items using OPTSEED, TECH14 and adding LRTSTARTS 0 0 100 20

TECHNICAL 14 OUTPUT	
Random starts specifications for the k-1 class analysis model	
Number of initial stage random starts	10
Number of final stage optimizations	2
Random starts specifications for the k-1 class model for generated data	
Number of initial stage random starts	0
Number of final stage optimizations for the initial stage random starts	0
Random starts specifications for the k class model for generated data	
Number of initial stage random starts	100
Number of final stage optimizations	20
Number of bootstrap draws requested	Varies
Parametric bootstrapped likelihood ratio test for 4 (H0) versus 5 classes	
H0 loglikelihood value	-41007.498
2 times the loglikelihood difference	398.368
Difference in the number of parameters	18
Approximate p-value	0.0000
Successful bootstrap draws	5

5 Time savings

Consider the computational time spent on the 5-class runs using the proposed three steps. In Step I, the first 5-class run (STARTS = 200 40) takes 18 seconds and increasing to STARTS = 600 120 in the second run takes 1 minute, 12 seconds. The Step II addition of TECH11 takes only 3 seconds due to using OPTSEED. The Step III addition of TECH14 takes 11 seconds and increasing to LRTSTARTS = 0 0 100 20 takes 35 seconds. The total is 139 seconds.

Compare the above time consumption to that of a commonly used alternative approach. This approach makes the mistake of using STARTS, TECH11, and TECH14 in all runs. First, a 5-class run with STARTS = 200 40 takes 1 minute, 7 seconds. Second, a 5-class run with STARTS = 600 120 takes 2 minutes, 59 seconds. Third, a 5-class run with STARTS = 600 120 and LRTSTARTS = 100 20 takes 3 minutes, 27 seconds. The total is 453 seconds. This is over three times slower than the proposed approach. For larger and more complex models, the time difference can be even greater.

References

Lo, Y., Mendell, N., & Rubin, D. (2001). Testing the number of components in a normal mixture. Biometrika, 88, 767 778.

McLachlan, G., & Peel, D. (2000). Finite mixture models. New York: Wiley.

Muthén, L. & Muthén, B. (1998-2012). Mplus User's Guide. Version 7. Los Angeles, CA: Muthén & Muthén.

Nylund, K.L., Asparouhov. T. & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural equation modeling, 14, 535-569.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57, 307 - 333.