Appendices for Auxiliary Variables in Mixture Modeling: 3-Step Approaches Using Mplus

Tihomir Asparouhov and Bengt Muthén

1 Appendix A: Step 3 of the 3-step method done manually versus all steps done automatically using R3STEP

Step 3 of the 3-step method done manually:

VARIABLE: NAMES = u1-u5 x p1-p3 n;
USEVARIABLES = x n;
CLASSES = c(3);
NOMINAL = n;
DATA: FILE = man3step2.dat;
ANALYSIS: TYPE = MIXTURE; STARTS = 0;
MODEL: %OVERALL%
c ON x;
%c#1%
[n#1@1.901];
[n#2@-0.990];
%c#2%
[n#1@-0.486];
[n#2@1.936];
%c#3%
[n#1@-2.100];
[n#2@-2.147];

3-step method done automatically using R3STEP:

```
VARIABLE: NAMES = u1-u5 x;
          CATEGORICAL = u1-u5;
          CLASSES = c(3);
          AUXILIARY = x(R3STEP);
DATA:     FILE = 3step.dat;
ANALYSIS: TYPE = MIXTURE;
MODEL:    !no model is needed, LCA is default
```

2 Appendix B: Input file for conducting a simulation study with a latent class auxiliary predictor

```
Montecarlo:
Names are u1-u5 x;
Generate = u1-u5(1);
Categorical = u1-u5;
Genclasses = c(2);
Classes = c1(2);
Nobservations = 500;
Nreplications = 500;
Auxiliary = x(R3STEP);

Analysis: Type = Mixture;

Model Population:
%Overall%
[x@0];
x@1;
[c#1*0.3];
c#1 on x*0.5;
%c#1%
[u1$1-u5$1*-1.25];
```

```
%c#2%
[u1$1-u5$1*1.25];

Model:
%Overall%
[c1#1*0.3];
c1#1 on x@0; ! This command is needed so that the LCA model
             ! is estimated with no influence from the predictor
             ! variable on the class formation
%c1#1%
[u1$1-u5$1*-1.25];
%c1#2%
[u1$1-u5$1*1.25];
```

3 Appendix C: Input file for generating data for manual 3-step estimation

```
Montecarlo:
Names are u1-u10 y x;
Generate = u1-u10(1);
Categorical = u1-u10;
Genclasses = c(3);
Classes = c(3);
Nobservations = 1000;
Nrep = 1;
save=man3step.dat;

Analysis: Type = Mixture;

Model Population:

%Overall%
[x@0]; x@1;
y*1;
y on x*0;

%c#1%
[u1$1-u10$1*-1];
[y*0];
```

```
y on x*0.5;

%c#2%
[u1$1-u10$1*1];
[y*1];
y on x*-0.5;

%c#3%
[u1$1-u5$1*1];
[u6$1-u10$1*-1];
[y*-1];
y on x*0;
```

Note that in this input file we do not need a model statement because we only use this input file to generate data.

4 Appendix D: Input file for step 1 in the manual 3-step estimation

```
variable:
Names are u1-u10 y x;
Categorical = u1-u10;
Classes = c(3);
usevar are u1-u10;
auxiliary=y x;

data: file=man3step.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%

%c#1%
[u1$1-u10$1*-1];

%c#2%
[u1$1-u10$1*1];

%c#3%
```

[u1$1-u5$1*1];

[u6$1-u10$1*-1];

SAVEDATA: FILE= man3step2.dat; SAVE=CPROB;

Note here that the MODEL statement is not needed. We have included that, however, so that the order of the classes remains the same as in the data generation. This is done just to make easy comparison between the true and the estimated parameters. In a real-data application if the measurement part is an LCA model, the MODEL section of this input can be removed. Note also that we specified the number of random starting values to be 0 in the ANALYSIS command with the option STARTS. This is again done to avoid class order switching between the data generation procedure and the estimation procedure. This option should not be used in a real-data application setting. Finally we need to clarify the use of the AUXILIARY option in the VARIABLE command. This use of the AUXILIARY option is completely different from the ones discussed in the previous sections. In this situation we do not specify a type for the auxiliary variables such as (R3STEP) or (DU3STEP). This means that the auxiliary variables are not used in the estimation. They are only included in the SAVEDATA file which will be used in the following steps. The SAVEDATA command is also used in this input file with the option SAVE=CPROB. This option produces 2 types of outputs. It produces the posterior class probabilities for each observation, which we don't actually need, as well as the most likely class variable N that we will use as a latent class indicator in the final stage estimation.

5 Appendix E: Input file for step 3 in the manual 3-step estimation

```
variable:
Names are u1-u10 y x p1-p3 n;
usevar are y x n;
classes = c(3);
nominal=n;

data: file=man3step2.dat;

Analysis: Type = Mixture; starts=0;

Model:

%overall%
Y on X;

%C#1%
[N#1@1.901];
[N#2@-0.990];
Y on X; Y;

%C#2%
[N#1@-0.486];
```

[N#2@1.936];

Y on X; Y;

%C#3%

[N#1@-2.100];

[N#2@-2.147];

Y on X; Y;

Note that in this step we use the data file obtained from the SAVEDATA command in Step 1. The most likely class variable is specified as a nominal variable and all the parameters $[N\#i]$ of the conditional distribution $[N|C]$ are fixed to the log ratios computed in Step 2. The parameters $[N\#1]$ and $[N\#2]$ in class 1 are fixed to the log ratios obtained from row 1 in the measurement error table: 1.901 and -0.990. The parameters $[N\#1]$ and $[N\#2]$ in class 2 are fixed to the log ratios obtained from row 2 in the measurement error table etc. In this third step we also specify the auxiliary model. In our example this is just a simple linear regression model with class-varying residual variances.

6 Appendix F: Input file for LTA data generation

Appendices F-I describe how to generate LTA data and carry out the different analysis steps. The input file in Appendix F is used to generate data according to the true LTA model. The input file in Appendix G is used to estimate the LCA measurement model for the first class variable C_1 and to obtain the most likely class variable N_1 which will be used in step 3 as a C_1 indicator. The measurement error for N_1 is computed using the log ratios as in Section **??**. The input file in Appendix H is used to estimate the LCA measurement model for the second class variable C_2 and to obtain the most likely class variable N_2 which will be used in step 3 as a C_2 indicator. The measurement error for N_2 is computed using the log ratios as in Section **??**. In real-data applications neither Appendices F or G need a model statement. We provide model statements here simply to order the classes according to the way we generated the data. The final third step is to estimate an LTA model where the variable N_1 is used as a class indicator variable for the first latent variable with prefixed error rates and the variable N_2 is used as a class indicator variable for the second latent class variable with prefixed error rates. This input file is included in Appendix I.

```
    Montecarlo:
Names are u11-u15 u21-u25;
Generate = u11-u15(1) u21-u25(1);
Categorical = u11-u15 u21-u25;
Genclasses = c1(2) c2(2);
```

```
Classes = c1(2) c2(2);
Nobservations = 2000;
Nrep = 1;
save=conc3step.dat;

Analysis: Type = Mixture;

Model Population:

%Overall%
[c1#1*0.3];
[c2#1*0.3];
c2#1 on c1#1*0.5;

MODEL population-c1:

%c1#1%
[u11$1-u15$1*-1];

%c1#2%
[u11$1-u15$1*1];

MODEL population-c2:

%c2#1%
```

[u21$1-u25$1*-1];

%c2#2%

[u21$1-u25$1*1];

7 Appendix G: Input file for 3-step LTA analysis, estimating LCA for C_1

```
variable:
Names are u11-u15 u21-u25;
usevar are u11-u15;
Categorical = all;
Classes = c1(2);
auxiliary=u21-u25;

data: file=conc3step.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%
[c1#1*0.3];

%c1#1%
[u11$1-u15$1*-1];

%c1#2%
[u11$1-u15$1*1];
```

savedata: file=c1.dat; save=cprob;

8 Appendix H: Input file for 3-step LTA analysis, estimating LCA for C_2

variable:

Names are u11-u15 u21-u25 p1 p2 n1;

usevar are u21-u25;

Categorical = all;

Classes = c2(2);

auxiliary=u11-u15 n1;

data: file=c1.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%

[c2#1*0.3];

%c2#1%

[u21$1-u25$1*-1];

%c2#2%

[u21$1-u25$1*1];

savedata: file=c2.dat; save=cprob;

9 Appendix I: Input file for 3-step LTA analysis, estimating the final auxiliary model

```
variable:
Names are u21-u25 u11-u15 n1 p1 p2 n2;
usevar are n1 n2;
nominal n1 n2;
Classes = c1(2) c2(2);

data: file=c2.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%
[c1#1*0.3];
[c2#1*0.3];
c2#1 on c1#1*0.5;

MODEL c1:

%c1#1%
[n1#1@1.864];
```

%c1#2%

[n1#1@-2.138];

MODEL c2:

%c2#1%

[n2#1@1.841];

%c2#2%

[n2#1@-1.842];

10 Appendix J: Input file for LTA data generation with measurement invariance and a covariate

```
Montecarlo:
Names are u11-u15 u21-u25 x;
Generate = u11-u15(1) u21-u25(1);
Categorical = u11-u15 u21-u25;
Genclasses = c1(2) c2(2);
Classes = c1(2) c2(2);
Nobservations = 2000;
Nrep = 1;
save=conc3step.dat;

Analysis: Type = Mixture;

Model Population:

%Overall%
[c1#1*0.3];
[c2#1*0.3];
c2#1 on c1#1*0.5 x*0.3;
c1#1 on x*-0.5;
x*1;
```

MODEL population-c1:

%c1#1%

[u11$1-u15$1*-1];

%c1#2%

[u11$1-u15$1*1];

MODEL population-c2:

%c2#1%

[u21$1-u25$1*-1];

%c2#2%

[u21$1-u25$1*1];

Model:

%Overall%

[c1#1*0.3];

[c2#1*0.3];

c2#1 on c1#1*0.5 x*0.3;

c1#1 on x*-0.5;

MODEL c1:

%c1#1%

[u11$1-u15$1*-1];

%c1#2%

[u11$1-u15$1*1];

MODEL c2:

%c2#1%

[u21$1-u25$1*-1];

%c2#2%

[u21$1-u25$1*1];

11 Appendix K: Input file for 3-step LTA estimation with measurement invariance: step 1

variable:

Names are u11-u15 u21-u25 x;

Categorical = u11-u15 u21-u25;

Classes = c1(2) c2(2);

auxiliary=x;

data: file=conc3step.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%

[c1#1*0.3];

[c2#1*0.3];

MODEL c1:

%c1#1%

[u11$1-u15$1*-1] (t1-t5);

%c1#2%

[u11$1-u15$1*1] (tt1-tt5);

MODEL c2:

%c2#1%

[u21$1-u25$1*-1] (t1-t5);

%c2#2%

[u21$1-u25$1*1] (tt1-tt5);

output: svalues;

Note that in Appendix K we request the OUTPUT option SVALUES which provides the model input commands for the next two input files. The SVALUES output contains the final results of the model estimation formatted as an input file. At this point in the SVALUES output one has to replace the * symbol with the @ symbol because in the next two inputs we are holding the parameters fixed to the results of the joint LCA estimation from Appendix K.

12 Appendix L: Input file for 3-step LTA estimation with measurement invariance: step 1 for C1

```
variable:
Names are u11-u15 u21-u25 x;
usevar are u11-u15;
Categorical = all;
Classes = c1(2);
auxiliary=u21-u25 x;

data: file=conc3step.dat;

Analysis: Type = Mixture; starts=0;

Model:

%OVERALL%
[ c1#1@0.19434 ];

%C1#1%
[ u11$1@-0.97524 ] (t1);
[ u12$1@-0.98527 ] (t2);
[ u13$1@-0.96129 ] (t3);
```

[u14$1@-0.97072] (t4);

[u15$1@-0.89841] (t5);

%C1#2%

[u11$1@1.02624] (tt1);

[u12$1@1.00941] (tt2);

[u13$1@1.03036] (tt3);

[u14$1@1.05849] (tt4);

[u15$1@1.08370] (tt5);

savedata: file=c1.dat; save=cprob;

13 Appendix M: Input file for 3-step LTA estimation with measurement invariance: step 1 for C2

```
variable:
Names are u11-u15 u21-u25 x p1 p2 n1;
usevar are u21-u25;
Categorical = all;
Classes = c2(2);
auxiliary=u11-u15 x n1;

data: file=c1.dat;

Analysis: Type = Mixture; starts=0;

Model:

%OVERALL%
[ c2#1@0.66961 ];

%C2#1%
[ u21$1@-0.97524 ] (t1);
[ u22$1@-0.98527 ] (t2);
[ u23$1@-0.96129 ] (t3);
```

[u24$1@-0.97072] (t4);

[u25$1@-0.89841] (t5);

%C2#2%

[u21$1@1.02624] (tt1);

[u22$1@1.00941] (tt2);

[u23$1@1.03036] (tt3);

[u24$1@1.05849] (tt4);

[u25$1@1.08370] (tt5);

savedata: file=c2.dat; save=cprob;

14 Appendix N: Input file for 3-step LTA estimation with measurement invariance: step 3

variable:

Names are u21-u25 u11-u15 x n1 p1 p2 n2;

usevar are n1 n2 x;

nominal n1 n2;

Classes = c1(2) c2(2);

data: file=c2.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%

[c1#1*0.3];

[c2#1*0.3];

c2#1 on c1#1*0.5 x*0.3;

c1#1 on x*-0.5;

MODEL c1:

%c1#1%

[n1#1@1.925];

%c1#2%

[n1#1@-2.020];

MODEL c2:

%c2#1%

[n2#1@1.787];

%c2#2%

[n2#1@-2.084];

15 Appendix O: Input file for a 3-step analysis with an auxiliary variable used as a predictor and for a direct effect

```
variable:
Names are u1-u10 x;
usevar are u1-u10 x x2;
Categorical = u1-u10;
Classes = c(2);
Auxiliary = x(R3STEP);

define: x2=x; ! duplication of variable

data: file=dup3st.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%
[c#1*0.3];
u1 on x2*0;

%c#1%
```

```
[u1$1-u10$1*-0.75];
u1 on x2*1;

%c#2%
[u1$1-u10$1*0.75];
u1 on x2*0;
```

To estimate Method 2 in Mplus the covariate X has to be used in the model as well as in the AUXILIARY option. In Mplus Version 7 this is not allowed, although within a Montecarlo simulation it is allowed. To easily estimate Method 2 the covariate should be duplicated using the DEFINE command and the duplicate variable should be used in the model.

16 Appendix P: Input file for conducting a simulation study with a distal outcome

```
Montecarlo:
Names are u1-u5 y;
Generate = u1-u5(1);
Categorical = u1-u5;
Genclasses = c(2);
Classes = c1(2);
Nobservations = 500;
Nreplications = 500;
Auxiliary = y(DU3STEP);

Analysis: Type = Mixture;

Model Population:
%Overall%
[y@0];
y@1;
[c#1*0.3];
%c#1%
[u1$1-u5$1*-1.25];
[y*0];
%c#2%
[u1$1-u5$1*1.25];
```

```
[y*0.7];


Model:
%Overall%
[c1#1*0.3];
[y] (1); y (2); ! This command is needed so that the LCA model
                 ! is estimated with no influence from the distal
                 ! variable on the class formation
%c1#1%
[u1$1-u5$1*-1.25];
%c1#2%
[u1$1-u5$1*1.25];
```