

Solutions to Exercise #2

Latent class analysis

We are going to define a latent class variable with 2 classes based on responses to the **rspspsgv ractrolg rpsppi1 rcpttpol rptcp1t retapap1** variables.

TASKS:

1. Using as a basis the previous “BASIC.inp” file, specify a 2-class latent variable model using the **rspspsgv ractrolg rpsppi1 rcpttpol rptcp1t retapap1** variables. Ensure that the variable **cou** (country) is used as a clustering variable and clustering is accounted for in the analyses.

ANSWER:

See input file “Exer2_1.inp”.

The indicators variables **rspspsgv** etc. are listed after USEVARIABLES=,

They are defined as ordered categorical by listing them after CATEGORICAL=.

The VARIABLE: command contains the name and number of latent classes estimated following the option CLASSES=.

The clustering variable is indicated after CLUSTER=. To ensure the analyses take into account this clustering, the ANALYSIS: command includes “COMPLEX” , thus ensuring clustering is controlled for:

```
Data:
  File is ess_ex1.dat ;
Variable:
  Names are
    essround idno polintr rspspsgv ractrolg rpsppi1 rcpttpol rptcp1t
    retapap1 cou ess7id nutslen nuts2en nuts3en;
  Missing are all (-999) ;
  Usevariables =
    rspspsgv ractrolg rpsppi1 rcpttpol rptcp1t retapap1;

    CATEGORICAL=  rspspsgv ractrolg rpsppi1 rcpttpol rptcp1t retapap1 ;

    Classes=class(2);

CLUSTER=cou;

Analysis:
  Type = MIXTURE COMPLEX;

MODEL:

OUTPUT: Tech1 Tech10 Tech11  Svalues;
```

2. Check the log-likelihood values of the final stage optimisation: does it appear as a trustworthy solution?

ANSWER: The solution would appear acceptable since loglikelihood has been replicated at least at least twice in two final stages: -47174.029. See output file “Exer2_1.out” and related messages:

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:

-47174.029	76974	16
-47174.029	107446	12
-47174.029	93468	3
-47174.029	533738	11

THE BEST LOGLIKELIHOOD VALUE HAS BEEN REPLICATED. RERUN WITH AT LEAST TWICE THE RANDOM STARTS TO CHECK THAT THE BEST LOGLIKELIHOOD IS STILL OBTAINED AND REPLICATED.

THE MODEL ESTIMATION TERMINATED NORMALLY

Note the warning provided by Mplus: It is good practice to follow this advice and ensure the same loglikelihood is replicated when the number of starts in the EM algorithm are increased.

3. What are the proportions of individuals in the two classes based on the estimated model?

ANSWER: The proportions of the two classes were 60.75% and 39.25% respectively.

FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES
BASED ON THE ESTIMATED MODEL

Latent Classes		
1	5419.26654	0.60747
2	3501.73346	0.39253

4. How can you interpret the two classes?

ANSWER: We should look at the conditional item response probabilities. Overall, individual in class 1 are more likely to report the lowest score to most of the items. They might be described as “sceptical” about the possibility to participate and influence politics:

RESULTS IN PROBABILITY SCALE

Latent Class 1

RSPPSGV

Category 1	0.761	←
Category 2	0.199	
Category 3	0.040	

RACTROLG

Category 1	0.717	←
Category 2	0.190	
Category 3	0.093	

RSPPIPL

Category 1	0.840	←
Category 2	0.147	
Category 3	0.014	

RCPTPOL

Category 1	0.615	←
Category 2	0.252	
Category 3	0.133	

RPTCPPLT

Category 1	0.796	←
Category 2	0.183	
Category 3	0.021	

RETAPAPL

Category 1	0.750	←
Category 2	0.206	
Category 3	0.044	

Overall, individuals in class 2 seem more positive about the chances to influence politics, although not overly optimistic since the highest probabilities of responses were the middle responses, rather than the highest ones:

Latent Class 2

RSPPSGV		
Category 1		0.163
Category 2		0.491
Category 3		0.346
RACTROLG		
Category 1		0.213
Category 2		0.404
Category 3		0.382
RSPPIPL		
Category 1		0.113
Category 2		0.572
Category 3		0.315
RCPTPOL		
Category 1		0.112
Category 2		0.461
Category 3		0.427
RPTCPPLT		
Category 1		0.228
Category 2		0.563
Category 3		0.209
RETAPAPL		
Category 1		0.177
Category 2		0.505
Category 3		0.318

5. What does the Vuong-Lo-Mendell-Rubin Likelihood Ratio test indicate?

The TECHNICAL 11 OUTPUT reports the results of a test where the fit of a 2-class solution is compared to a 1-class solution. A significant p-value ($p < .05$) indicates we should reject the null hypothesis that the two models (2-classe and 1-class) provide the same fit, so the 2-class model does seem to provide a different (better) fit. We would thus accept a 2-classe model over a 1-class model.

TECHNICAL 11 OUTPUT

Random Starts Specifications for the k-1 Class Analysis Model

Number of initial stage random starts	20
Number of final stage optimizations	4

VUONG-LO-MENDELL-RUBIN LIKELIHOOD RATIO TEST FOR 1 (H0) VERSUS 2 CLASSES

H0 Loglikelihood Value	-52416.892
2 Times the Loglikelihood Difference	10485.727
Difference in the Number of Parameters	13
Mean	-487698.081
Standard Deviation	976724.706
P-Value	0.0000

LO-MENDELL-RUBIN ADJUSTED LRT TEST

Value	10397.796
P-Value	0.0000


6. Now, run a model with 5 classes

ANSWER: See input file "Exer2_2.inp" and the number of classes indicated after CLASSES= :

```
Data:
File is "D:\Olivers Laptop\WiP\Zeyun\exer1.dat" ;
Variable:
Names are
  rlagey ragender rldraw rlslfmem rlimrc rldlrc rlser7 raeduc
  ID householdID communityID;
Missing are all (-999) ;

Usevar= rldraw rlslfmem rlimrc rlser7 ;
CATEGORICAL = rldraw rlslfmem rlimrc rlser7 ;

Classes=class(5);
```



7. Check the loglikelihood values of the solution: does it appear as a trustworthy solution?

ANSWER: The solution may be considered trustworthy solution because the loglikelihood values were replicated at least twice in the last stages. However, it is worth checking it by increasing the number starts (as the Mplus output suggests), particularly when we consider that there was some variation in the loglikelihood values observed. A rule-of-thumb that some suggest using states that the loglikelihood must have been replicated at least 20 times (e.g.: Sinha, P., Calfee, C.S., & Delucchi, K.L. (2021). Practitioner's Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls. *Critical Care Medicine*, 49(1), e63-e79. doi:

10.1097/CCM.0000000000004710).

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:

-44357.807	107446	12
-44357.807	399671	13
-44357.807	195873	6
-44486.723	533738	11

THE BEST LOGLIKELIHOOD VALUE HAS BEEN REPLICATED. RERUN WITH AT LEAST TWICE THE RANDOM STARTS TO CHECK THAT THE BEST LOGLIKELIHOOD IS STILL OBTAINED AND REPLICATED.

THE MODEL ESTIMATION TERMINATED NORMALLY

8. Indicate the number of starts and final stage interactions to be 1000 and 100 respectively, and check the loglikelihood values of the solution

ANSWER: Check output file "exer2_3.out". The solution appears trustworthy as the loglikelihood is replicated (over 20 times) and it is the same obtained in the previous run with fewer starts (see solution to Question 7):

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:

-44357.807	284716	713
-44357.807	79212	517
-44357.807	396614	918
-44357.807	531546	702
-44357.807	942848	852
-44357.807	499150	216
-44357.807	897782	545
-44357.807	618760	489
-44357.807	881886	608
-44357.807	922042	492
-44357.807	278661	674
-44357.807	915107	54
-44357.807	804616	868
-44357.807	461866	722
-44357.807	496344	808
-44357.807	662718	460
-44357.807	823392	479
-44357.807	748692	204
-44357.807	809240	543
-44357.807	798821	423
-44357.807	373815	618
-44357.807	937588	293
-44357.807	968846	970
-44357.807	608460	244
-44357.807	641794	591
-44357.807	737601	891
-44357.807	354624	448
-44357.807	920593	611
-44357.807	626762	704

9. Compare the information criteria, and other parameters of the 2- and 5-class solutions: which appears to provide a better fit?

ANSWER: the 2-class solution provides AIC= 94398.058, BIC= 94575.462, and sample-sized adjusted (aBIC)= 94496.016. The 5-class solution provides AIC= 88843.614, BIC= 89297.768, and aBIC= 89094.387. Overall, all the 3 information criteria are lower for the 5-class solution: the latter provides a preferable solution.

10. Inspect the graph of estimated probabilities for conditional item responses of item category 3 in the 5-class solution

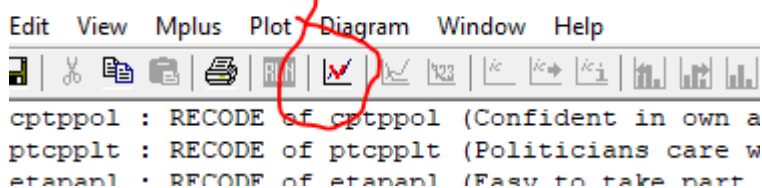
ANSWER:

To include graphs, include these lines in the INPUT files:

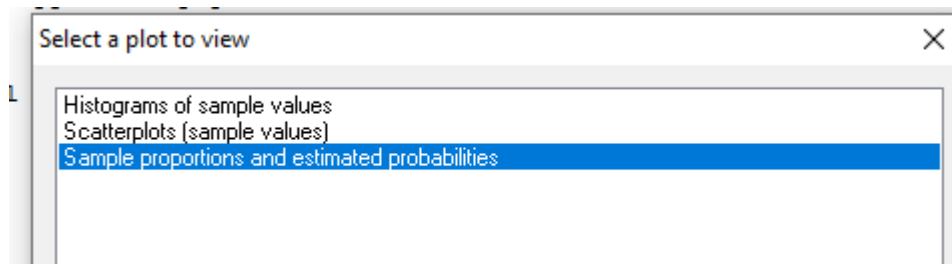
```
PLOT:Type is PLOT3;
series is rpsppsgv (1) ractrolg (2) rpsppipl (3) rcptppol (4) rptcpplt (5) retapapl (6);
```

To open options for visualising graphs, click on the graph icon in the output file “exer2_1.out”:

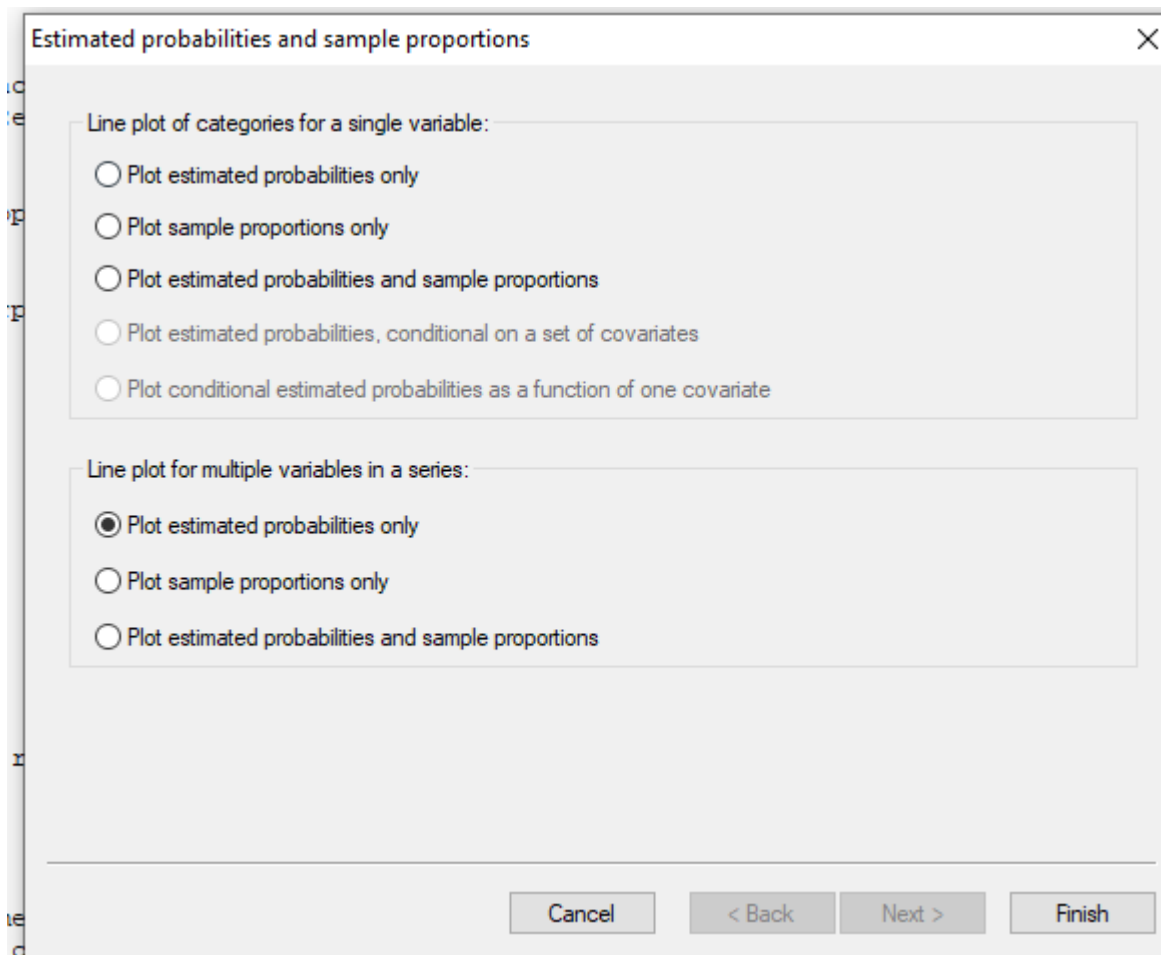
- [exer2_1.out]



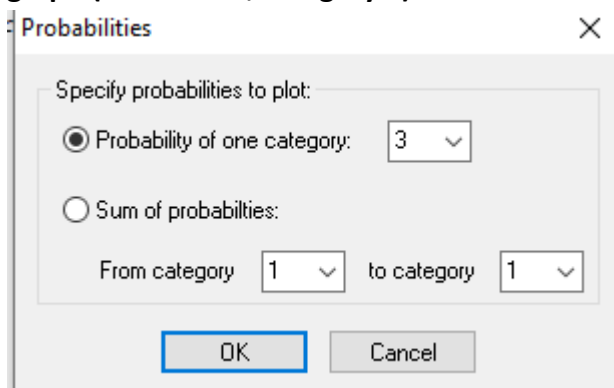
This opens a dialog window: Select “Sample proportions and estimated probabilities”:



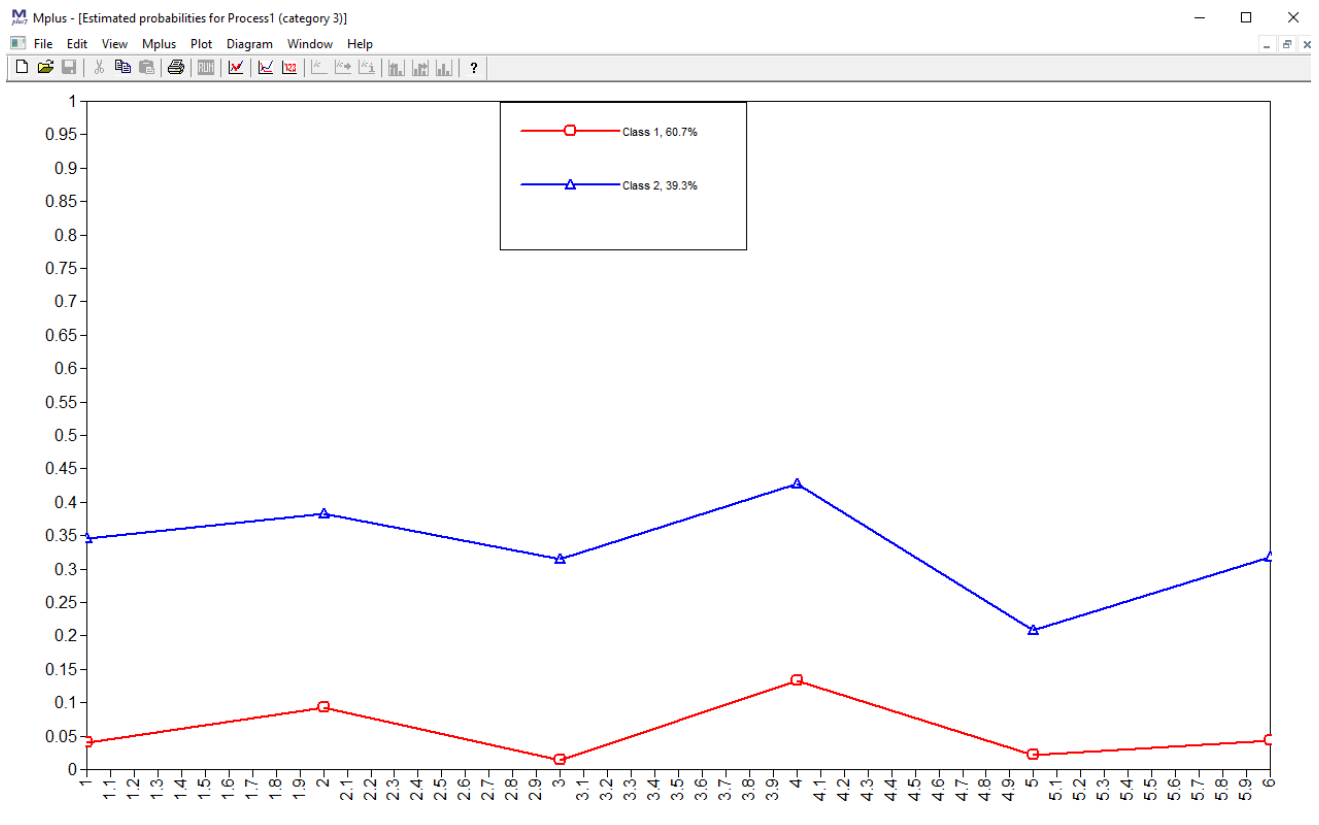
This opens another dialog window. Select “Line plot for multiple variables in a series:” “Plot estimated probabilities only”:



This opens a further dialog window. Select the answer category you want to represent in the graph (in this case, category 3):



This will open this graph, which you can save for further use:



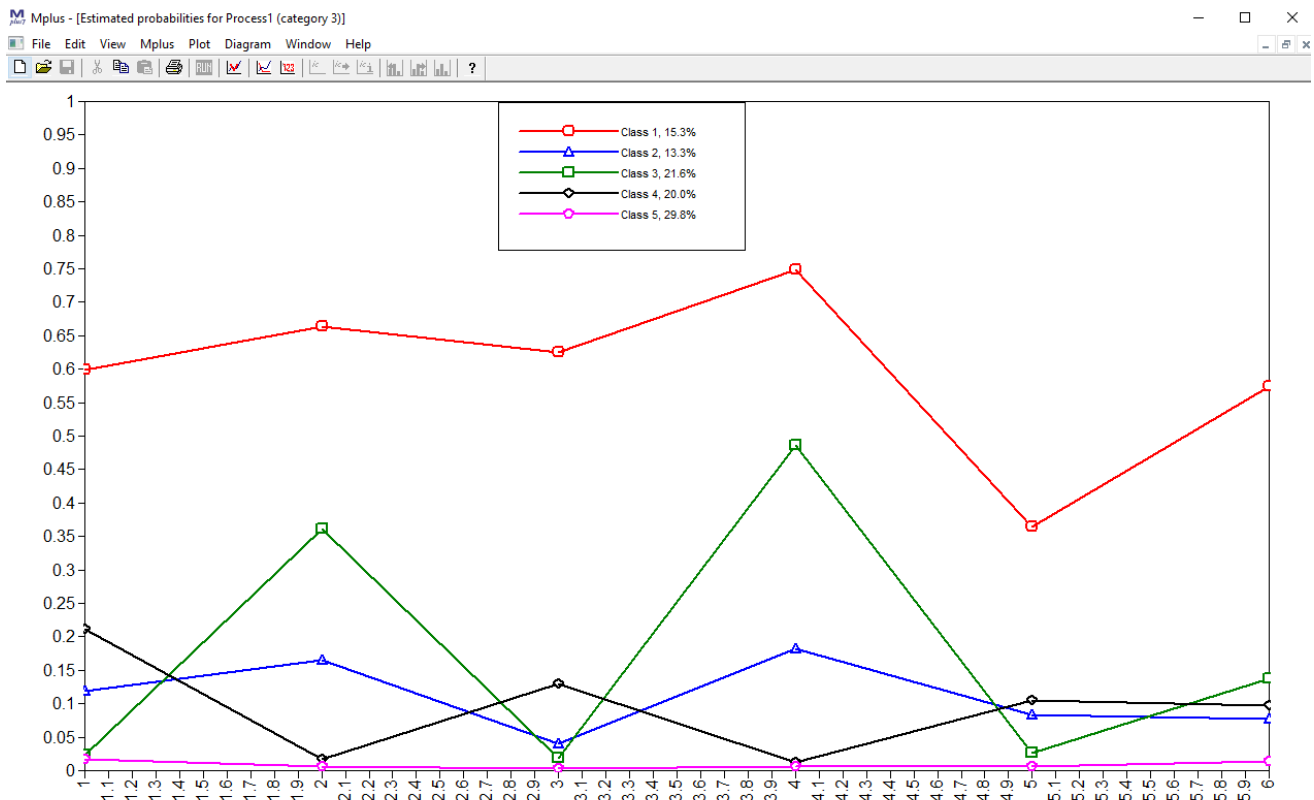
The graph represents the probability of answering category 3 of each item by classes. Items are in x-axis, following the sequence we indicated here:

```
PLOT:Type is PLOT3;
series is rpsppsgv (1) ractrolg (2) rpsppi1 (3) rcptppol (4) rptcpplt (5) retapapl (6);
```

Which means in x=1 we have the conditional probability of answering more positively to item rpsppsgv: “Political system allows people to have a say in what government does”, in x=2 we have the conditional probability of answering more positively to time ractrolg: “Able to take active role in political group”, and so on. As we had seen, Class 1 includes people that are more sceptical about the possibility to influence politics, while Class 2 includes people that are more positively, although not overtly so. Note that the graph also reports the prevalence of the two classes: we can see that most people (approximately 61%) are estimated to be in the “sceptical” class

We can repeat the same operations by opening output file “exer2_3.out”.

We should obtain a similar graph:



This indicates a more complex scenario. People in class 5 seem to be the more sceptical (they have virtually no chance of answering more positively to all 6 items), while people in class 1 seem to be most positive. People in class 3 provide an interesting profile where they seem likely to answer more positively to items *ractrolg*: “Able to take active role in political group” and *rcptppol*: “Confident in own ability to participate in politics” (respectively in $x=2$ and $x=4$), but seem otherwise quite sceptical. These may be people that despite participating actively in politics, seem to be quite sceptical about the possibilities to influence politics. We might call them “Disenchanted” with politics.

11. Estimate models with the number of classes increasing from 1 to 8, and compare the models based on fit statistics, information criteria, entropy, and the Vuong-Lo-Mendell-Rubin Likelihood Ratio test. Which model would you select?

ANSWER:

You have to create different INPUT files with increasing number of classes. Once this is done, you will have to inspect each of these OUTPUT files to extract all the statistics you need. All these files are provided in folder “*exer2_4_classes*”. If you use R, the MplusAutomation suite can greatly help and avoid these tedious tasks, see Hallquist, M. N. & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. Structural Equation Modeling, 25, 621-638. doi: 10.1080/10705511.2017.1402334.

Considering the output in folder “*exer2_4_classes*”, I have obtained these results:

N classes	Log-Likelihood	AIC	BIC	aBIC	Parameters	Estimator	T11_LMR_Value	T11_LMR_PValue	Entropy
2	-47174.03	94398.06	94575.46	94496.02	25	MLR	10397.8	0	0.78
3	-45905.23	91886.47	92156.12	92035.36	38	MLR	2516.31	0.21	0.78
4	-44750.62	89603.24	89965.14	89803.07	51	MLR	2289.87	0.09	0.78
5	-44357.81	88843.61	89297.77	89094.39	64	MLR	779.03	0.33	0.75
6	-44069.98	88293.96	88840.37	88595.67	77	MLR	570.83	0.35	0.74
7	-43931.08	88042.16	88680.82	88394.81	90	MLR	275.47	0.36	0.73
8	-43821.68	87849.36	88580.26	88252.95	103	MLR	216.82	0.44	0.72

Overall, no single model seems to be far more satisfactory than others. The Information Criteria seem to favour the model with more classes. The Vuong-Lo-Mendell-Rubin Likelihood Ratio test (T11 LMR Value and p Value in the Table), seems to favour the model with two classes, or, to some extent, that with 4 classes. If we consider entropy (i.e. the model ability to separate individuals across the estimated classes), the models with 2 to 4 classes seem equivalent, while entropy decreases from estimating 5 or more classes.

Overall, I would consider the model with 4 classes, but this example emphasises the importance of considering, testing and investigate more than one model before accepting one.

For more discussion, see:

- Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, 4(4), 440-461. <https://doi.org/10.1037/tps0000176>
- Perra, O. (2020). *Latent Transition Analysis*. In Atkinson, P., Delamont, S., Cernat, A., Sakshaug, J.W., & Williams, R.A. (Eds.): Sage Research Methods Foundations. London: Sage. <https://doi.org/10.4135/9781526421036878157>
- Ryoo, J. H., Wang, C., Swearer, S. M., Hull, M., & Shi, D. (2018). Longitudinal model building using latent transition analysis: an example using school bullying data. *Frontiers in psychology*, 9, 675. <https://doi.org/10.3389/fpsyg.2018.00675> .

The 4-class solution appears to indicate differences along the dimension of Scepticism-Optimism. In this figure I report the conditional probabilities of the three categories of each item for individuals in classes 2, 3 and 4. These conditional probabilities are reported as stacked percentages for each item within each class:

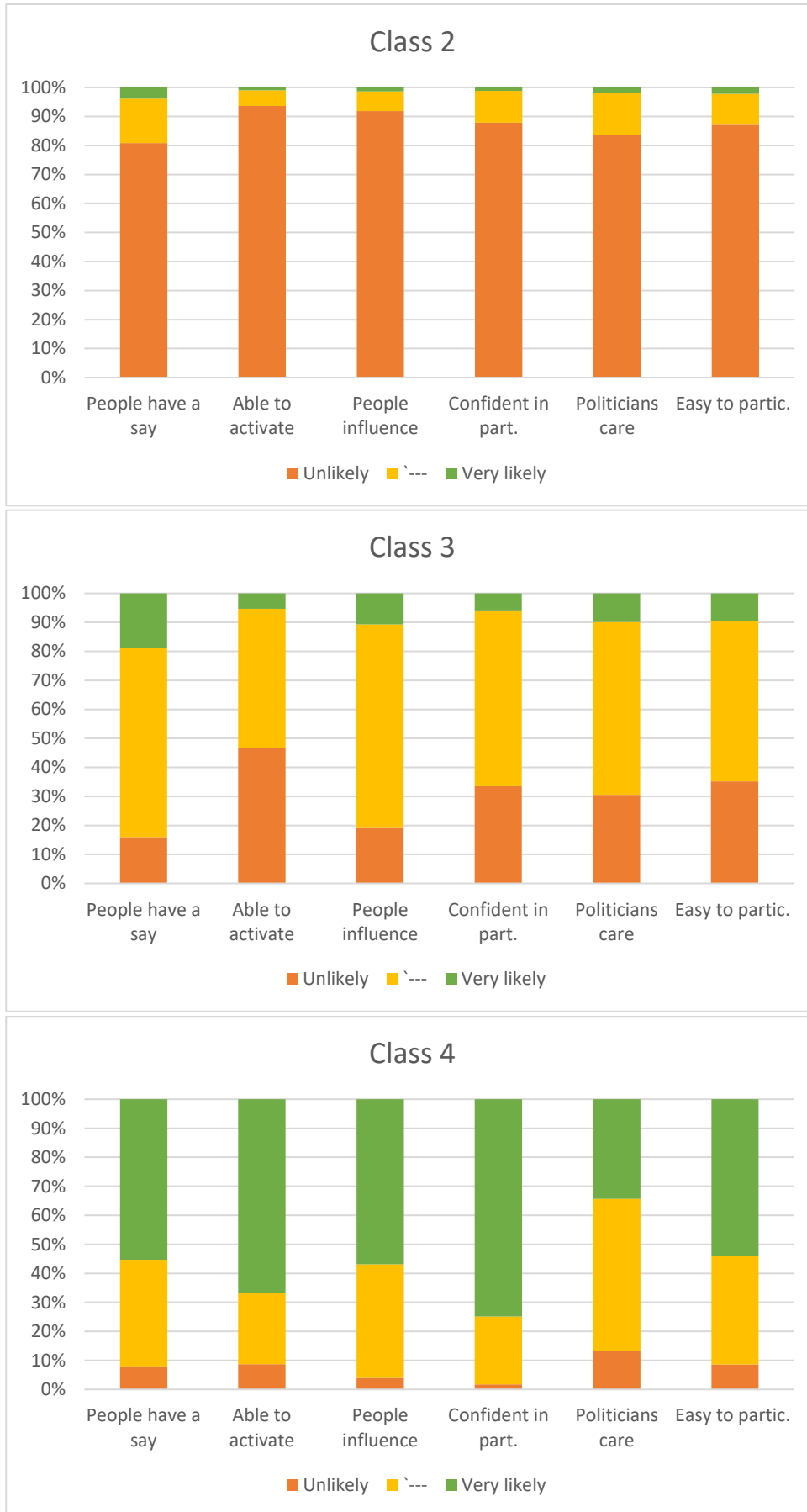


Figure 1: Conditional probabilities of item responses by items and by classes. Conditional probabilities are reported as stacked probabilities within each item.

The graph suggests that we might call individuals according to their position in the Scepticism-Optimism dimension. Individuals in Class 2 are the more sceptical about the political system and their ability to participate and influence politics. Individuals in Class 4 are the most optimistic, while individuals in Class 3 stand in between the two (we could call them “Neutrals”).

Individuals in Latent Class 1 have a different profile, one that does not easily sit in the in the Scepticism-Optimism dimension. Let’s consider their conditional responses:

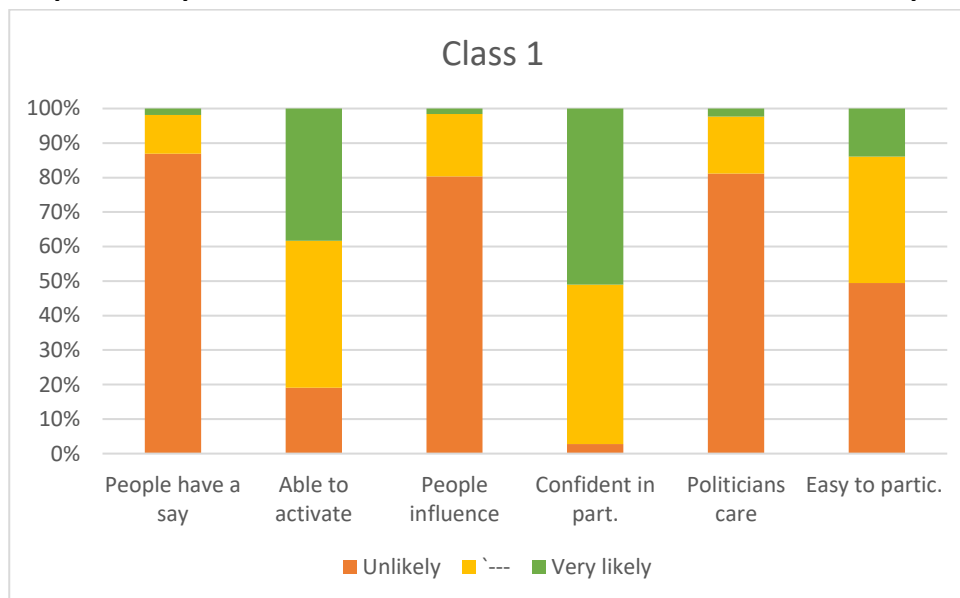


Figure 2: Conditional probabilities of item responses by items for individuals in latent class 1. Conditional probabilities are reported as stacked probabilities within each item.

As noted when we looked at the 5-class solution, these are individuals that seem more optimistic about their ability to be active and take part in politics. However, they seem to be otherwise very sceptical about the political system and the ability to influence it. I would refer to these as “Disenchanted”.

Therefore, the example provides a good example of how latent class analysis allows to identify “qualitative” differences between individuals: in this case, as well a classification of individuals along the dimension of Scepticism-Optimism, we have also identified individuals that differ in important ways from the other individuals.

12. A parallel indicators model is one where the all the categories of response to an indicator have the same probabilities within a class. For example, if the latent classes were “Depression” vs. “No Depression”, a parallel model would assume that people in the “Depression” class will have the same probability of reporting the symptoms of depression, e.g. they will have the same probability of reporting “Low Mood”, “Lack of Pleasure”, “Sleep Problems”, etc. Try to specify a similar model for the latent class more likely to report lower scores across all indicators in the 4-class solution (the “Sceptics”). Compare this constrained model with the unconstrained one using a likelihood-ratio test. Should we accept or reject the constrained “parallel indicators” model?

ANSWER:

You will have to create an INPUT file where the thresholds of at least latent class 2 are specified and constrained. For this purpose, the best way to proceed is to extract the starting values of the model without constraints, the model with 4 unconstrained classes.

The output we are looking for is this:

MODEL COMMAND WITH FINAL ESTIMATES USED AS STARTING VALUES

%OVERALL%

```
[ class#1*0.19175 ];
[ class#2*0.77960 ];
[ class#3*0.46612 ];
```

%CLASS#1%

```
[ rpsppsgv$1*1.89570 ];|
[ rpsppsgv$2*3.95066 ];
[ ractrolg$1*-1.44428 ];
[ ractrolg$2*0.47613 ];
[ rpsppipl$1*1.40571 ];
[ rpsppipl$2*4.14669 ];
[ rcptppol$1*-3.55713 ];
[ rcptppol$2*-0.04216 ];
[ rptcpplt$1*1.46241 ];
[ rptcpplt$2*3.74280 ];
[ retapapl$1*-0.02414 ];
[ retapapl$2*1.82023 ];
```

%CLASS#2%

```
[ rpsppsgv$1*1.44018 ];
[ rpsppsgv$2*3.21359 ];
[ ractrolg$1*2.69610 ];
[ ractrolg$2*4.63803 ];
[ rpsppipl$1*2.42504 ];
[ rpsppipl$2*4.26520 ];
[ rcptppol$1*1.98349 ];
[ rcptppol$2*4.41838 ];
[ rptcpplt$1*1.63715 ];
```

This output shows all the starting values for the parameters of the overall latent class model (%OVERALL%) and for the parameters within each of the estimated latent classes (%CLASS#1% %CLASS#2%...etc.).

We can copy this output and paste it into the INPUT file we are creating after the MODEL: command. This will also ensure that you will be more likely to obtain a solution where the latent classes are in the same order of the previous output:

MODEL:**%OVERALL%**

```
[ class#1*0.19175 ];
[ class#2*0.77960 ];
[ class#3*0.46612 ];
```

%CLASS#1%

```
[ rpsppsgv$1*1.89570 ];
[ rpsppsgv$2*3.95066 ];
[ ractrolg$1*-1.44428 ];
[ ractrolg$2*0.47613 ];
[ rpsppipl$1*1.40571 ];
[ rpsppipl$2*4.14669 ];
[ rcptppol$1*-3.55713 ];
[ rcptppol$2*-0.04216 ];
[ rptcpplt$1*1.46241 ];
[ rptcpplt$2*3.74280 ];
[ retapapl$1*-0.02414 ];
[ retapapl$2*1.82023 ];
```

%CLASS#2%

To ensure that the indicators in latent class 2 (the “Sceptics” class) are parallel, we will need to impose equality constraints to the thresholds of the indicators:

```
[ retapapl$1*-0.02414 ],
```

%CLASS#2%

```
[ rpsppsgv$1*1.44018 ] (1);
[ rpsppsgv$2*3.21359 ] (2);
[ ractrolg$1*2.69610 ] (1);
[ ractrolg$2*4.63803 ] (2);
[ rpsppipl$1*2.42504 ] (1);
[ rpsppipl$2*4.26520 ] (2);
[ rcptppol$1*1.98349 ] (1);
[ rcptppol$2*4.41838 ] (2);
[ rptcpplt$1*1.63715 ] (1);
[ rptcpplt$2*3.99546 ] (2);
[ retapapl$1*1.90766 ] (1);
[ retapapl$2*3.78049 ] (2);
```

By using the numbers between parentheses, we are asking the software to keep the first threshold of the indicator **rpsppsgv** the same as the first threshold of variable **ractrolg**, as well as the same as the first threshold of variable **rpsppipl** and so on.

In other words, within latent class 2, the model will estimate only 2 thresholds, one (1) for the first thresholds of all indicators, and a second (2) one for the second thresholds of the indicators. See file “Exer2_5.inp” for the full input.

In the output “Exer2_5.out” you can check that the equality constraints have been applied in the final solution:

```
Latent Class 2

RSPSPSGV
  Category 1      0.878      0.017      52.404      0.000
  Category 2      0.102      0.015       6.659      0.000
  Category 3      0.020      0.002       8.099      0.000
RACTROLG
  Category 1      0.878      0.017      52.404      0.000
  Category 2      0.102      0.015       6.659      0.000
  Category 3      0.020      0.002       8.099      0.000
RSPSPIPL
  Category 1      0.878      0.017      52.404      0.000
  Category 2      0.102      0.015       6.659      0.000
  Category 3      0.020      0.002       8.099      0.000
RCPTPPOL
  Category 1      0.878      0.017      52.404      0.000
  Category 2      0.102      0.015       6.659      0.000
  Category 3      0.020      0.002       8.099      0.000
RPTCPPLT
  Category 1      0.878      0.017      52.404      0.000
  Category 2      0.102      0.015       6.659      0.000
  Category 3      0.020      0.002       8.099      0.000
RETAPAPL
  Category 1      0.878      0.017      52.404      0.000
  Category 2      0.102      0.015       6.659      0.000
  Category 3      0.020      0.002       8.099      0.000
```

We can see that now the indicators are all “parallel” in class 2: the conditional probability of providing the lowest level of answer to the items is estimated to be 0.88 for individuals in latent class 2 (“Sceptics”), and is the same for all the items.

I report key statistics of the two models here:

	Unconstrained 4 classes	4 Classes with parallel indic.
Log-Likelihood	-44750.618	-44868.714
Scaling correction factor	13.1872	15.5197
Free parameters	51	41
AIC	89603.236	89819.428
BIC	89965.140	90110.371
aBIC	89803.071	89980.080
Pearson χ^2 (df)	1325.508 (673)	1394.165 (683)
Likelihood χ^2 (df)	1084.808 (673)	1141.519 (683)

When using the MLR estimator, as in this case, the Likelihood Ratio test should be calculated considering the “Scaling correction factor”. The formulas for the Likelihood Ratio (LR) test are:

$$\text{LR test} = -2 \frac{L_0 - L_1}{Cd}; \quad Cd = \frac{(p_0 * c_0) - (p_1 * c_1)}{(p_0 - p_1)}; \quad \text{Degrees of freedom} = (p_1 - p_0);$$

Where:

L_0 = Log-Likelihood of the null model (that with equality constraints);

L_1 = Log-Likelihood of the model without the added constraints;

c_0 = Scaling correction factor of the null model (that with equality constraints);

c_1 = Scaling correction factor of the model without the added constraints;

p_0 = Free parameters in the null model (that with equality constraints);

p_1 = Free parameters in the model without the added constraints;

Using the formulae above, we can calculate that the LR test of the two models is 65.18 with $df=10$, which corresponds to $p < .001$. Therefore, we will reject the null hypotheses that the two models have no significant differences in fit, and can accept that the model without the equality constraints provides a significantly better fit compared to the model with parallel indicators. We would thus discard the model with parallel indicators.