# Introduction to Latent Class Analysis

## Video 1 transcript

https://www.ncrm.ac.uk/resources/online/all/?id=20806

Oliver Perra: So, to summarise what I have discussed so far, I have said the Latent Class Analysis is a person-centred model. It does assume a sample includes a mixture of individuals that show different patterns of behaviour. It's also a measurement model, so we assume that the latent classes are the causes of different behaviours we observe in the sample. And it is a probabilistic model. The associations between offset behaviours and latent classes is observed with arrow and individuals’ latent class affiliation is also uncertain to some degree.

In this final presentation, I will talk more about how to estimate the model, the latent class model, and then I will talk about how we should decide on the number of latent classes we want to estimate. And I will talk about some issues that arise when we are trying to include covariates and distal outcomes in the model.

So, let's start with some information about the estimation of the latent class model.

And estimated latent class models require using observed information to estimate unobservable parameters, for example, latent class frequencies and so on. Commonly the maximum likelihood approaches are used for this purpose. These identify the set of model parameters that maximise the likelihood that the process described by the model produced the data we observe. So, within maximum likelihood, the two step expectation maximisation algorithm is commonly used. The algorithm gives parameters in the models on values drawn randomly to start with and then re-estimate these parameters until any new draws meet a given criterion, which is usually convergence. That is, the difference between the likelihood in previous draws and the likelihood in any new iteration and any new draw is below a pre-set threshold. At that point, the algorithm stops and provides what is considered to be the best solution.

Other algorithms can be used, for example, the Newton-Raphson, but the problem that these algorithms have is that they may converge on solutions that are not necessarily the best solution. And I will try to explain what I mean by referring to a fictional example where here I have drawn a fictional likelihood distribution.

So, assume the likelihood distribution has a similar shape. The X axis here represents different sets of model parameters. We cannot observe this distribution, and we need to identify basically the parameter values that correspond to the highest likelihood value.

And so, the algorithm tries to do that by drawing parameter values at random and then checking the likelihood associated with these parameter values. So, here these points, for example, in the X value correspond to different set of parameter values drawn at random, and the dots represent the likelihood associated with these parameter values.

So, the algorithm continues to draw values of the parameters until the likelihood associated with any new draw of this value does not exceed the highest likelihood we have already identified which here will be this one.

So, when the algorithm finds that any new draws of parameter values does not correspond to a higher value of the likelihood compared to the one we have estimated here, the algorithm will stop.

[0:04:38]

But according to what we have observed here, the highest likelihood value that we find corresponds to this value of set parameters that we find here. But we are failing to detect the highest likelihood indicated by the black arrow which corresponds to another different set of parameter values. So, in this case, in this fictional example, the solution provided by the algorithm is failing to identify the parameter set values that correspond with the highest likelihood.

And we call the solution that we have found in this fictional example as identifying a local maximum and the problem with this is a solution that looks good when we run the algorithm, but this is failing to find the global maximum, the highest value of likelihood that we are aiming to identify.

So, this simplified example illustrates how it may be possible using this algorithm to miss the best solution and so we need to be careful when we run latent class models. And what we can actually do is a series of steps. The most important ones are first of all to check that the different rows in the algorithm return the same likelihood, the same value of the best likelihood several times. For example, more than 20 times. So, if the same value of the likelihood is returned on different occasions, on different rounds of the algorithm, this increases our confidence that the solution we find may not be a local one, but may be the actual best one.

And we can also then, after estimating a model, run the same model a second time, but increasing the number of runs of the expectation maximisation algorithm. And in this way, ensuring, for example, that we have more random draws of parameter set values and therefore reducing that probability that we will miss the highest peak in the distribution of the likelihood.

And there are other possibilities I have listed here. For example, increasing the number of iterations or reducing the threshold to stop the running of the algorithm so that the algorithm continues to look for the best solution and no better solution is being found. So, we can and we should always try to doublecheck how the latent class solution to make sure that the solution is replicated that the likelihood is replicated and the solution is stable.

Now I'm going to talk about how we can decide on the number of latent classes we want to estimate. And provided we have a stable solution, we want to know how many latent classes can reliably explain the different patterns of behaviours we have observed in our participants. For example, do we need only two categories of depression to explain differences in symptoms of depression?

So, there are several statistics that we need to consider when we are deciding on the appropriate number of classes. And the first ones are the fit indices. These report comparisons between the frequencies expected by the model and the observed frequencies. So, if the latent class model produced expected frequencies that are very similar to the ones we have observed, the value of these indices will be low. Conversely, if the value of these indices are high and significantly different from zero, this will indicate that the latent class model we are testing provides a full feed.

However, when we use these fit indices, we need to be cautious because the large contingency tables of categorical data, so the large contingency tables we create with latent class models often, are quite sparse which causes problems in estimating the distribution of these tests because with past data, the distribution of these fit indices may not fall of the non-chi square distribution. So, we might use this fit indices with some caution.

But latent class models with different number of classes appear nested, so one model contains all the parameters of the other one and at least one additional parameter. However, latent class models with different number of classes cannot be compared directly using the likelihood ratio test as in other instances. And that is because some of the necessary prerequisite for using the likelihood ratio tests are not met. However, there are different tests that are used in other approaches, for example, bootstrap approach, can test models with different number of classes. This test basically compared the feed of a model with N classes, for example, four against the feed of a model with one less Class 3 example. If the test has a P value below 0.05, we should reject the new hypothesis that the model with N classes provides the same feed as the model with one less class, which means that we would accept a model with four classes if the P value of this test is less than 0.05. And so, on we can run then other comparisons of a model with four classes against a model with five classes and so on until we find a model whose feed is not significantly different from the model with one less class.

[0:11:18]

Information criteria are also very useful in comparing latent class models. The information criteria we use are based on the likelihood ratio chi square, but penalise more complex models, those with increased number of parameters. So, the rule is to choose models that are associated with lower values of the information criteria and that also means that we tend to prefer models that are more parsimonious, those with less number of parameter has been estimated.

And the most common ones are the ones I have listed here. For example, the Bayesian information criteria or the sample size adjusted Bayesian information criteria.

But when we run Latent Class Analysis, we are often interested in finding latent classes that effectively separate the individuals in a sample into categories. In the same presentation, I introduced the index of entropy which measures the precision of the latent class model classification. So, if researchers want to find models that provide more precise categorisations of individuals, they may select solutions with higher entropy.

And finally, it's also important to consider if the models selected are providing plausible solutions when we want to choose the best or the optimal model. And the best solution may, and the choice of these models should be also based on theory and or substantive knowledge. And this is also because if there is a high degree of heterogeneity in the data observed due to measurement error, this can determine the emergence of latent classes that are allowing in a way for error patterns. So, it's always advisable to rely on substantive knowledge and theory to check if the models we are estimating, the classes we're estimating, are sensible, make sense based on what we know about that phenomena.

Finally, I will talk about issues that emerge when we are trying to include covariates and distal outcomes in latent class models.

Here I represented a latent class estimation on the left, but I also added the distal outcome on the left, age of retirement. I wonder if you can spot what type of problems the inclusion of the distal outcome can create if we estimate the latent classes at the same time?

The problem that emerges when we estimate the latent classes using indicators and we add a distal outcome is that the latent classes we are estimating in this case are taking into account the heterogeneity in the indicators and the heterogeneity in the distal outcome.

So, this means that when we add a distal outcome, the measurement parameters of the latent classes will shift often and often we will have a different solution where the number of latent classes may even change when we add these outcomes.

The same problem happens when we add covariates. The problem is basically that the latent classes are then hunting for heterogeneity in the indicators and the other covariates or these or distal outcomes were included.

We might want to use latent class affiliation as covariates, but in doing so, we will also need to take into account that latent classes are probabilistic, so we need to control for the ascent in latent class membership if we want to avoid bias in the results when we use latent classes as covariates.

[0:15:37]

One solution is to use multiple pseudo-class draws. Basically, this involves assigning participants at random into latent classes in multiple times in multiple datasets, and this assignment of participants into latent classes will be based on the distribution of posterior probabilities, the probabilities of being one class or another. And then the results can be combined across draws using the same methods as multiple imputation.

But another approach that has been recently developed is the so-called three step approach. And this basically involves a step where we first estimate the measurement model of the latent class model and we assign individuals to the most likely class. After this we can inspect the model we have estimated and check the measurement error, the uncertainty in latent class affiliation, and use that to create a latent class model where respondents are assigned to latent classes but with the assignment taking into account the uncertainty we have estimated.

So, at that stage we'll have a latent class model that is fixed based on the information we have collected from Step 1 and 2. And we cannot covariate or distal outcomes without the need to re-estimate the measurement model. So, to estimate a model like this one and assign individuals to the most likely class.

In the second step, we assess the measurement error of the model. That is the degree of uncertainty in allocating people to the latent classes.

And in Step 3, we can use the most likely class variable of each participant that we have estimated in Step 1 as a latent class indicator but with uncertainty rates fixed at the level assessed in Step 2. This basically ensures that the measurement of the latent class variable is not affected by inclusion of covariates and distal outcomes.

And in fact, we can include then covariates and other distal outcomes and test models like this one because the measurement model has been already estimated in Step 1.

So, the three-step approach really separates the steps of assessing the measurement model from the step of adding covariates and distal outcomes and allows them to run more complex models without re-estimated and measurement model that we are using.

This also allows to test other type of models, for example, moderation models, where, for example, we may assume that the latent class is moderate in the association between, say, social demographic variables and age of retirements.

So, to summarise, it's important that we check solutions that we obtain when we run Latent Class Analysis and we ensure that these solutions are stable. and in deciding the number of latent classes we need to consider different statistics.

And finally, when we are considering latent classes as outcomes or predictors or moderators, we need to use advanced methods that control for the uncertainty in latent class affiliation.

So, thank you very much for your attention and please look at the exercises and the other material I've put online and get back to me if you have any questions or queries.

Thanks a lot. Bye now.

National Centre for Research Methods (NCRM)  
Social Sciences  
Murray Building (Bldg 58)  
University of Southampton  
Southampton SO17 1BJ  
United Kingdom

**Web** www.ncrm.ac.uk   
**Email** info@ncrm.ac.uk  
**Tel** +44 23 8059 4539  
**Twitter** @NCRMUK