Session 1

Introduction to Latent Class Cluster Models

Session Outline:
A. Basic ideas of latent class analysis
B. The general probability model for categorical variables
C. Determining the number of classes/clusters
D. Fit measures, model specification and selection strategies
E. Classifying cases into latent class segments
F. Interpreting Latent GOLD output (*including an application to Latent Class Trees*)
G. Example from survey analysis
H. Including covariates in LC models
I. Boundary, identification and local solution issues; Bayes constants
J. Extension to continuous variables and other scale types
K. Including direct effects to relax the assumption of local independence
L. Example with Diabetes data (obtaining scoring equations)
A. Basic ideas of latent class analysis

The basic idea behind traditional latent class (LC) models is that responses to variables come from K distinct mutually exclusive and exhaustive populations called latent classes. Respondents in a given latent class are homogeneous with respect to model parameters that characterize their responses. (The specific model parameters associated with traditional LC models will be formalized in topic B). Since the goal of identifying homogeneous groups of cases is the same as in traditional cluster analysis, we refer to the traditional LC model as the LC Cluster Model. Specifically, the LC Cluster model includes a K-category latent variable, each category representing a latent class (cluster, segment).

There is a close connection between the maximum likelihood (ML) algorithm used in estimating the LC Cluster model and the K-means algorithm used in cluster analysis, the latter being the most widely used technique for performing cluster analysis currently. While the K-means algorithm uses Euclidean distance to group cases that are close to each other based on their values on continuous (or at least, quantitative) variables, the LC approach utilizes probabilities to measure distance, and thus is not limited to quantitative variables. The LC approach may be viewed as a way of formalizing the K-means approach in terms of a statistical model, and extending it in many directions.

Cluster Analysis - 2 Approaches PowerPoint presentation. This is a non-technical presentation:

“Session 1 Cluster Analysis.ppt”

Assigned Reading:

“Session 1 Reading.pdf”

Latent Class Models Article:

A. Latent class models for clustering (pages 2-9)

Reference:

This article presents a more technical comparison.
One important advantage of LC modeling over the K-means algorithm for clustering is that with LC a scoring formula can be used to classify new cases regardless whether the variables are continuous, categorical or both. For the simulated data example described in the Magidson and Vermunt (2002) article, as shown in the Appendix (page 8 of Session 1 Assigned Reading), the formula for computing the log-odds (called ‘logit’) of being in Class 1 vs. Class 2 is exactly linear in the 2 variables Y1 and Y2.

Specifically,

\[
\text{Logit}(y_1, y_2) = 15.455 - 3.280 \times y_1 + 0.629 \times y_2
\]

Thus, for new cases where values on Y1 and Y2 are available, the probability of being in class 1 can be computed as:

\[
\text{Prob(Class 1|y_1, y_2)} = \frac{\exp[\text{Logit}(y_1, y_2)]}{1 + \exp[\text{Logit}(y_1, y_2)]}
\]

Since \( \text{Logit}(y_1, y_2) = 0 \) is equivalent to \( \text{Prob(Class 1|y_1, y_2)} = .5 \)

and \( \text{Logit}(y_1, y_2) > 0 \) is equivalent to \( \text{Prob(Class 1|y_1, y_2)} > .5 \),

cases are classified into class 1 or 2 based on the following:

if \( \text{Logit}(y_1, y_2) > 0 \), case is classified into class 1, otherwise case is classified into class 2.

Version 5.1 of Latent GOLD contains an option to display the scoring equations for classifying new cases. We will see how to obtain scoring equations in Exercise E3.
B. The general probability model for categorical variables

The formal LC model may be expressed in terms of probability parameters or log-linear parameters. We will primarily use the probability formulation here.

**Assigned Reading:**

“Session 1 Reading.pdf”

**Sage Article:**

- B1: Introduction, section 1: (pages 11-15)
- B2: Example, section 2.1: (pages 16-20)
- B3: Bivariate Residuals, section 3.1: (pages 21-22)

**LG Tutorial #1 link:** [Latent GOLD Tutorial 1](#)

- B1: Model Setup (pages 1-6)
- B2: Model Estimation (pages 7-8)
- B3: Parameters, Profile and ProbMeans Output: (pages 9-15)

**Exercise B.**

1. For the 3-class model, how many probability parameters are there? How many unconditional probabilities (associated with the size of each latent class)? How many conditional probabilities?

2. Regarding the K unconditional probabilities, since they sum to 1, only K-1 are distinct - the last one can always be computed from the others. In total, for the 3-class model, how many distinct parameters are there? Does this agree with the number reported under the Npar column as shown in Figure 7-9 of LG Tutorial 1?
C. Determining the number of classes/clusters

There are various criteria that can be used to assist in determining the number of classes. That is for choosing between one candidate model that hypothesizes say 3 latent classes over say a 4-class model. No single criteria is generally agreed upon as best. The standard practice of obtaining the model p-value associated with the $L^2$ fit statistic from a chi-squared table lookup is quite limited since the p-value is not valid if based on sparse data, such as when at least one variable is continuous. That is, with sparse data the $L^2$ statistic does not follow a chi-squared distribution. Since in practice data is often sparse, alternative approaches are needed to determine how well a model fits the data.

The assigned reading material (given below) discusses the use of Information Criteria, such as the BIC statistic, which favors more parsimonious models (e.g., models specifying fewer classes) by penalizing the Log-likelihood (LL) statistic obtained for each model according to the number of parameters that are estimated for that model.

An alternative approach for comparing candidate models is Cross-validation (CV), most often applied using the K-fold cross-validation technique. The way that K-fold CV works is as follows:

1) Randomly assign cases into K groups (each group is called a ‘fold’). Generally, K = 10.
2) Estimate each candidate model K times, the kth estimation being performed after omitting the cases from 1 of the folds (i.e., from the kth fold, k = 1, 2,…, K).
3) Apply the estimates from the kth model to the cases in the omitted fold. For example, cases in the omitted fold k are classified based on the parameters estimated using only the other cases (i.e., parameter estimates from model k).
4) Accumulate the results applied to the omitted folds. For example, the log-likelihood statistic obtained by cross-validation, called the ‘Validation Log-likelihood’, is obtained by summing over the K Log-likelihood components evaluated for cases in each of the K folds.
5) Choose the model that yields the highest Validation LL statistic.

The BIC and other information criteria are provided as standard Latent GOLD output. In addition, K-fold cross-validation is implemented in the syntax module in Latent GOLD Version 5.

In addition, a common validation approach that can be applied when subgroups can be defined by a variable on the analysis file is available in the 5.0 GUI version of Latent GOLD using the Select option. In this approach, a model is estimated on a pre-defined subgroup of cases (or replications in the case of data with repeated observations per case) and the remaining cases (or replications) are treated as hold-out records for purposes of validation. For further details on this, see Section 2.5 of the Latent GOLD 5.0 Upgrade Manual: http://www.statisticalinnovations.com/technicalsupport/LG5manual.pdf
Assigned Reading:

"Session 1 Reading.pdf"

Sage Article:

C: Sparse Data, section 2.1: (pages 23-24)

LG Tutorial 1:

C1: Bootstrap, (pages 8-9)
C2: Conditional Bootstrap, (pages 19-21)

Exercise C.

1. What criteria do you use to determine the number of classes?
D. Fit measures, model specification and selection strategies

Since the LC model attempts to account for all of the associations among the variables, another useful measure of “local fit” is the bivariate residual, which evaluates the extent to which the model adequately explains the association between pairs of variables.

Assigned Reading:

“Session 1 Reading.pdf”

Sage Article:

D. Bivariate Residuals (page 21)

LG Tutorial 1:

D: Bivariate Residuals, (pages 17-18)

Exercise D.

1. Reproduce the table shown in Table 5 of the SAGE article (page 21) for the 1-class model H0 by computing the Pearson chi-square test for independence using the raw data. Hint: Be sure to divide by the appropriate number of degrees of freedom. Do some of the 4 variables appear to be independent? Which ones?
E. Classifying cases into latent class segments

Given the model, a case can be assigned to the most likely latent class based on the response pattern observed for that case.

Assigned Reading:

“Session 1 Reading.pdf”

Sage Article:
E: Classification, section 2.3, (pages 25-26)

LG Tutorial 1:
E: Classification, (pages 14-17)

Exercise E.

1. If you have access to SPSS, use the ClassPred Tab to obtain the standard classification output to the file ‘data3.sav’ as indicated in Figure 7-21 of LG Tutorial 1 (page 16). If you do not have access to SPSS, use the Edit Copy command to copy the standard classification output (shown in Figure 7-20 on page 16) to the Clipboard and paste it into Excel. Then sort by ‘Modal’.

2. Compute the frequency distribution for the modal assignment class and confirm that it is the same as shown in the Total row in Figure 7-10 (LG Tutorial 1, page 9) - 805(1), 178(2), and 219(3). Why does this distribution not match the cluster sizes as reported in the Profile Output in Figure 7-15 of LG Tutorial 1 (page 12).

2. Obtain scoring equations for classifying new cases using the file ‘data3.sav’ or the copy of this file ‘data3_copy.sav’ which works with the demo version of Latent GOLD. These equations can be obtained as an output option in LG 5.1 (1-Step Scoring), or by using the Step-3 module in Latent GOLD. Hint: See "Step 3 Tutorial 2".
F. Interpreting Latent GOLD output

Parameters and Profile Output

The primary results from a latent class model estimated in Latent GOLD® are provided in the Parameters and Profile output listings. The parameters displayed in the Parameters Output are log-linear parameters, and are estimated directly by the program. These parameters are then transformed into the more easily interpreted probability model parameters, which are displayed in the Profile Output. The log-linear parameters displayed in the Parameters Output are used primarily for significance testing as explained in Sage Section 2.2.

See Magidson and Vermunt, 2001 ("SOME.pdf") for technical details of the relationship between the log-linear and probability parameters (section 2.1 provides the log-linear form of the model, while page 255 provides the corresponding probability parameter form of the latent class model):

"SOME.pdf"

Beginning with version 5.1 of Latent GOLD a special sub-category within the Parameters Output was included in the program that uses Wald tests to compare each pair of latent classes. This allows, for example, testing which Clusters are significantly different in terms of the indicators. For further details of this Paired Comparisons Output and associated Wald tests see section 8.2 of the Technical Guide.

Output from Latent Class Tree Models

Up to now, we have focused on the traditional (standard) approach to latent class modeling. Latent GOLD 6.0 (forthcoming later in 2018) will include the ability to estimate latent class tree (LCT) models which utilize a hierarchical paradigm for performing latent class analysis somewhat similar to hierarchical clustering. Participants in this online course will have the opportunity to experiment with LCT models and view new output for these models that will appear in Latent GOLD 6.0.

In the introductory tutorial for LCT (see Exercise F2), you will develop a 3-class model in the usual way, and then for comparison you will develop a 3-class LCT model. The LCT model is developed automatically by Latent GOLD by first splitting the sample into two subgroups (two parent classes which are equivalent to the latent classes from a standard 2-class LC model), and then further splits parent class #2 into two child classes in order to obtain an acceptable overall model fit. The three resulting segments are represented as terminal nodes in the tree diagram below, where segment numbers ‘1’, ‘2’ and ‘3’ appear below the terminal nodes associated with the three segments.
Latent Class Tree consisting of 3 Segments

The entries contained in the tree nodes are explained in the tutorial, where you will verify that they differ from the three segments obtained by estimating a standard 3-class model.

In this course you will have the opportunity to experiment with LCT models and view new output for LCT models that will appear in Latent GOLD 6.0. In particular:

Parameters and Profile output listings are extended to include LCT output:

- Output from the first (parent) level of the Tree is displayed in the usual way in the standard Parameters and Profile Output listings.

- Output from child classes spawned by splitting of each parent class is provided in a separate output section.

Exercise F3 introduces this new output along with a new graphical tree display using a simple example.
Assigned Reading:

“Session 1 Reading.pdf”

Sage Article:

F1: Significance Tests, section 2.2: (pages 26-27)

F2: Graphical Displays, section 2.4: (pages 28-31)

"Building LC Trees"

Latent GOLD Tutorial 1:

F1: Parameters Output, (pages 9-10)

F2: Pairwise Comparisons (page 10)

F3: ProbMeans Output, (page 14)

Latent GOLD Tutorial on Estimating a LC Tree Model [LCT Tutorial](#)
Exercise F1: Standard Latent Class Model.

1. The Tri-plot makes clear that the 3-class model may be considered to be a 2-dimensional model. How many dimensions are associated with a 4-class model? Is the tri-plot meaningful in the case of more than 3 classes?

Exercise F2: Latent Class Tree Model.

Use the new “tree” keyword in the experimental version of Latent GOLD to develop a LCT model (see the Latent Class Tree Tutorial).

1. Use the Latent GOLD GUI to reproduce manually the tree output generated automatically (using the “tree” command in the syntax) for the split of parent class #2 into the child classes which we label as “troubled” and “depressed.”

Hint: As indicated at the end of the LCT tutorial, you will use the Advanced tab of Latent GOLD to include the variable clu#2 as a sampling weight. In order to obtain the correct sample sizes, you will need to change the default ‘Rescale’ option so that this weight is used without being rescaled.
Exercise F3: LCT Analysis of Social Capital Data

1. Use the demo dataset socialcapital.sav to reproduce the LC tree analysis performed in “Building Latent Class Trees, With an Application to Social Capital” paper (van den Bergh et al., 2017).

**Background:** Owen and Videras (2009) used data gathered from 14,527 respondents of the 1975, 1978, 1980, 1983, 1984, 1986, 1987 through 1991, 1993, and 1994 samples of the General Social Survey to construct a typology of social capital that accounts for the different incentives that networks provide (van den Bergh et al., 2017). The latent model for this data set can be viewed as an exploratory latent class tree analysis.¹

Table 2. Social capital variables and description used in the exercise.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair</td>
<td>= 1 if &quot;people are fair&quot; (= 0 if &quot;people try to take advantage&quot;)</td>
</tr>
<tr>
<td>Trust</td>
<td>= 1 if people can be trusted</td>
</tr>
<tr>
<td>Church</td>
<td>= 1 if membership in church organization</td>
</tr>
<tr>
<td>Service</td>
<td>= 1 if membership in service group</td>
</tr>
<tr>
<td>Veteran</td>
<td>= 1 if membership in veteran group</td>
</tr>
<tr>
<td>Union</td>
<td>= 1 if membership in labor union</td>
</tr>
<tr>
<td>Political</td>
<td>= 1 if membership in political club</td>
</tr>
<tr>
<td>Youth</td>
<td>= 1 if membership in youth group</td>
</tr>
<tr>
<td>School</td>
<td>= 1 if membership in school service</td>
</tr>
<tr>
<td>Farm</td>
<td>= 1 if membership in farm organization</td>
</tr>
<tr>
<td>Fraternal</td>
<td>= 1 if membership in fraternal group</td>
</tr>
<tr>
<td>Sport</td>
<td>= 1 if membership in sports club</td>
</tr>
<tr>
<td>Hobby</td>
<td>= 1 if membership in hobby club</td>
</tr>
<tr>
<td>Greek</td>
<td>= 1 if membership in school fraternity</td>
</tr>
<tr>
<td>Nationality</td>
<td>= 1 if membership in nationality group</td>
</tr>
<tr>
<td>Literary</td>
<td>= 1 if membership in literary or art group</td>
</tr>
<tr>
<td>Professional</td>
<td>= 1 if membership in professional society</td>
</tr>
<tr>
<td>Other</td>
<td>= 1 if membership in any other group</td>
</tr>
</tbody>
</table>

Introduction to Latent Class Modeling using Latent GOLD

SESSION 1

1. Using the Latent GOLD GUI and syntax, specify a latent class tree model with 2-, 3-, and 4-classes. Which model makes the most substantive sense? Provide a rationale for your decision. Hypothesized models for social capital with 2- and 3-class parent classes are represented in Figures A and B.

2. Are there any branches in the tree you would consider modifying for substantive reasons? Which are they?

Figure A. Layout of a LCT with root of 2 classes of social capital (adapted from van den Bergh et al., 2017).

Figure B. Layout of a LCT with root of 3 classes of social capital (adapted from van den Bergh et al., submitted manuscript).
G. Example from survey analysis

Exercise G.

1. From a substantive perspective, how might you interpret the results as displayed in the Tri-plot (see LG Tutorial 1, Figure 7-18, page 14 of Session 1 Assigned Reading 1)? In particular, the categories of UNDERSTANDING appear to trace out a horizontal dimension in the tri-plot, while the categories of the other variables seem to trace out more of the vertical dimension.

Optional Reading:

H. Including covariates in LC models

Often, it is desired to profile the latent class segments in terms of demographics or other exogenous variables (called covariates, and denoted Z₁, Z₂, …) to help better understand them and to see how they might differ from each other. In addition, it may be desired to predict segment membership for new cases not included among the sample used to estimate the model. Since information on the indicators may not be available for new cases, predictions for new cases may be based on covariate information alone.

Covariates may be included in a LC model in an active or inactive manner. Specifying covariates as inactive yields output tables that show the relationship between the latent classes and the covariates, but does not alter model parameters; inclusion of inactive covariates yields the same model parameter estimates as obtained when no covariates are specified at all. Specifying covariates as active causes additional log-linear parameters to be included in the LC model (gammas), and estimated simultaneously with the other parameters (betas) and hence affect (somewhat) these model parameters. Like the other model parameters (betas), statistical tests are available for the gammas.

While Latent GOLD allows various kinds of model restrictions to be placed on the betas no restrictions may be placed on the gammas. For further details, see section 3.7 of

Latent GOLD Technical Guide,

Inclusion of active covariates in a model enhances the relationship between the covariates and the classes beyond the relationship that exists when the covariates are treated as inactive. Some researchers prefer to specify covariates as active, others as inactive.

Assigned Reading:

“Session 1 Reading.pdf”

Sage Article:

H1: Multi-group Models, Section 3.3 (pages 70-71)

H2: Covariates, Section 3.4 (page 75)

Latent GOLD Technical Guide

H: Sections 3 through 3.3 (pages 90-95)

CAMBRIDGE:

H: Introduction & Covariates (page 109)
Exercise H.

1. When do you think covariates should be treated as active and when inactive?

I. Boundary, identification and local solution issues; Bayes constants

Certain problems may occur during model estimation. These problems are:

1. Boundary solutions may be encountered
2. One or more model parameters may not be identified
3. Local solutions may be encountered

1) Boundary solutions may be encountered

Occasionally, maximizing the likelihood function yields a boundary solution; that is, a solution in which certain multinomial probabilities, Poisson rates or error variances in normal models turn out to be zero, or may converge to zero. Such problems are prevented in Latent GOLD by the imposition of prior distributions on the parameters through the use of Bayes constants as a default technical parameter setting.

By default, the technical parameters for Bayes constants in Latent GOLD are set to alpha = 1, which causes alpha (= 1) artificial observations to be added to the data for the purpose of parameter estimation. Following the model estimation, the artificial observations are ‘subtracted’ from the data, so that the number of cases reported in the summary output does not reflect the additional alpha observations. The artificial observation(s) are generated from a conservative null model. Strictly speaking, when a non-zero Bayes constant is used, rather than maximum likelihood estimation, the estimation procedure utilized is referred to as ‘posterior mode’ estimation and the modified log likelihood function that is maximized is called the ‘log-posterior’. For further information on this and the prior distributions used by Latent GOLD constants see section 6.3 of the Technical Guide (pages 29-31 of Session 2 Assigned Reading), and the section on Estimation in CAMBRIDGE (page 45 of Session 2 Assigned Reading).

In exercise #B1 (below) we will change the Latent GOLD setting for Bayes constants from the default of 1 to 0, to illustrate a situation where the resulting maximum likelihood solution is a boundary solution, and interpret these results.
2) Identification issues

If insufficient information is available to obtain unique maximum likelihood estimates for one or more model parameters, such parameters are said to be ‘not identified’. (For a more formal definition of parameter identification, see section 6.8 of the Technical Guide) The Latent GOLD program provides estimation warning messages when it encounters boundary or unidentified parameters. In exercise #B2 (below), we will examine some of these situations. One issue to be aware of is that the additional information provided by a non-zero Bayes constant may in some cases cause unidentified parameters to become identified.

3) Local solutions

Because the likelihood function is not guaranteed to be concave, the estimation algorithm for maximizing it may sometimes yield a solution that provides a maximum only within a local range of parameter values rather than globally over all possible combinations of parameter values. Depending upon the particular starting value used for the parameter estimates in the estimation algorithm, the resulting solution may be local. The best way to prevent ending up with a local solution is to use multiple sets of randomly generated starting values. By default, the Latent GOLD program uses 10 sets of random start values. See section 6.6 of the Technical Guide (pages 35-36 of Session 2 Assigned Reading) for further information on this topic.

Assigned Reading:

“Session 1 Reading.pdf”

Latent GOLD Technical Guide:

I1: Sections 6.3 - 6.6 (pages 96-103)
I2: Section 6.8 (page 104)

CAMBRIDGE:

I: Estimation (page 112)
Exercise I.

1. Re-estimate the 3-class model estimated earlier in Tutorial #1. Now, change the technical parameter setting for the ‘Bayes Constants’ for both Categorical Variables and as well as for Latent Variables’ from ‘1’ to ‘0’ and estimate new model. (You will find the ‘Bayes Constants’ settings labeled in the upper right portion of the Technical Tab.)

Notice that the L-squared statistic is now slightly better than the original model ($L^2 = 21.8920$ vs. the original value of $22.0872$). An Estimation Warning message is produced along with an Iteration Output file. At the bottom of the Iteration Output is a message saying that 2 boundary solutions were encountered. Go to the Profile output. Where are the 2 boundary solutions that were encountered? How do these estimates compare to the corresponding estimates obtained in the original model?

Next, change the Bayes constant from ‘0’ to ‘2’ and estimate the model. Compare the profile output in these two models. Notice that as the Bayes constant is increased, the extreme parameter estimates become less extreme. The greater the value of the Bayes constant, the greater the weight that is placed on a conservative null model (‘prior distribution’) which specifies that all variables are mutually independent. Use of the default Bayes constant of 1 provides a fairly small weight for this ‘prior’ distribution.

2. Return to the 3-class model estimated in Exercise B1 above when the Bayes constant was set to 0. Open the Variables Tab, remove the variable COOPERATE from the model and estimate it. Again, you will get an Estimation Warning Message. Estimate the model once again. Do you get the same L-squared value? If not, re-estimate it again until you get the same L-squared value at least twice. Examine the ‘Parameters’ Output for models having the same L-square value. Notice that some of the parameter estimates are different! This is an indication that these parameter estimates are not identified.

For this exercise, you may obtain an L-squared value of .0705 (which is actually an unidentified ‘local’ solution), or .0220 which is the unidentified ‘global solution’ (At least I believe that it is the global solution. When local solutions exist, it is never 100% certain that the solution you obtain is a global solution.) You may also encounter other L-squared values associated with other local solutions when
estimating this model. See the section on ‘Local solutions’ above.)

Now, repeat the exercise after restoring the Bayes Constant to its default value of ‘1’. Notice that for models having the same L-squared value, the parameters estimates no longer are different. That is because the information provided by non-zero Bayes constant is sufficient to uniquely identify the model.

How many degrees of freedom are associated with this model? When Bayes Constants = 0, negative degrees of freedom indicate that the model is not identified.

3. If you estimate and re-estimate a model several times and always get the same L-squared value and always the same parameter estimates, and the output is very interpretable from a substantive perspective, can you be comfortable with your interpretation? What if you notice that the degrees of freedom are negative?
J. Extension to continuous variables and other scale types

The field of finite mixture (FM) modeling developed as mixtures from K latent populations of normally distributed variables. Hence, the extension to LC models with continuous observed variables (utilizing the normal distribution) formalizes the connection between LC modeling and finite mixture (FM) modeling. In traditional LC modeling with nominal indicators, the multinominal distribution is used. For continuous variables, the joint distribution used is the multivariate normal. For count variables, other distributions are used (Poisson, binomial count).

To make clear that the T indicators / response/ dependent variables may be quantitative, some of the reading materials change the notation for the indicators to $Y_1, Y_2, \ldots, Y_T$.

An important difference from traditional LC modeling with categorical indicators, is that with some other scale types, fewer indicators are required to achieve identification. In traditional LC models, 3 dichotomous indicators are required for a 2-class model to be identified. In contrast, only a single indicator is required for any K-class model to be identified when the indicator is continuous or a Poisson count variable. As an example of this, see SAGE, section 4.1 (pages 9-13 of Session 2 Assigned Reading).

Assigned Reading:

“Session 1 Reading.pdf”

Sage Article:

J: Section 4.1 (pages 76-80)

Latent GOLD Technical Guide

J1: Finite Mixture models for Continuous Response Variables, Section 3.5 (pages 105-106)

J2: LC Cluster models for mixed mode data, Section 3.6 (page 107)

CAMBRIDGE

J: Continuous Indicator Variables (pages 113-116)
Exercise J.

1. See example from diabetes data below (Section L)

In the case of a nominal indicator with ordered categories, the ordinal scale type may be used to take into account the ordered nature of the categories. By default, the categories are assumed to be equally spaced, which is accomplished within the Latent GOLD framework using equidistant category scores. Thus, for indicator t, the score used for category m, would be $y_m = m$. See page 93 of the Session 1 Assigned Reading (excerpts from the LG Technical Guide) for further discussion.
K. Including direct effects to relax the assumption of local independence

In traditional LC models, the latent classes account for all associations between the response variables/indicators. That is, for cases in the same latent class, the indicators are statistically independent. In some cases, this ‘local independence’ property is not desired. In such cases, one or more direct effect parameters can be included in a model to account for certain bivariate associations outside the LC portion of the model. For example, suppose that 6 items are used to identify a dichotomous scale (say, depressed vs non-depressed), and that very similar wording is used for 2 of the 6 indicators creating additional (extraneous) association between these 2 items. This additional association can be accounted for outside the LC portion of the model using a direct effect, so that the latent classes would be based only on the common association shared by all 6 indicators.

Another example of direct effects is given in Exercise #L1 (below), where it is desired that the latent classes represent different types – persons with Chemical diabetes, those with Overt diabetes, and cases know to have neither of these types of diabetes (‘Normal’ individuals). In this example, we will see that a correlation between 2 of the indicators is not relevant to distinguishing between these 3 latent classes. As such, a direct effect is included in the final model to account for this association.

Assigned Reading:

“Session 1 Reading.pdf”

Latent GOLD Technical Guide

K: Local Dependencies, Section 3.4 (page 108)

CAMBRIDGE:

K: Diabetes Example (pages 117-119)
L. Example with Diabetes data

Exercise L.

1. Read the diabetes example in SAGE section 4.3 (starting on page 81 of the Session 1 Assigned Reading).

Download the associated data files diabetes.dat and diabetes.lgf for these data.

After estimating a model, double click on that model and click the Residuals Tab. Here you will see the bivariate residuals associated with each pair of indicators, sorted from high to low. A checkmark preceding an indicator pair indicates that a direct effect parameter for that pair has been included in the model. In the Model tab, you will see which (if any) of the effects are specified as class independent. Which model do you think is best? What is your criteria? Add the true diagnosis – the variable TRUE -- as an inactive covariate in each of these models. Examine the Profile and ProbMeans output to see which model most closely relates the latent classes to the desired true states.

2. Re-estimate model type 5, requesting the posterior membership probabilities (Classification - Posterior) be output to a file. Then open the newly created outfile and use the new Step 3 option to obtain the scoring formula that can be used to score new cases as a function of the 3 indicators. Hint: Since model type 5 does not assume the variances and covariances to be equal within each of the 3 latent classes, a quadratic function must be specified in order to obtain an $R^2 = 1$ (i.e., to perfectly reproduce the posterior membership probabilities). Which quadratic terms entered into the model have non-zero coefficients? What is the formula for the posterior membership probabilities as a function of the 3 indicators? For assistance, see: ‘Step 3 Tutorial 3.

3. Using only the 2 variables GLUCOSE and INSULIN, how well are you able to distinguish persons with overt diabetes from the others using a 2-class model? How can you tell that only 3 cases are misclassified?

4. Optional: If you have access to SPSS, use the K-Means procedure (using the Analyze/Classify menu), specifying 2 clusters and requesting that cluster membership probabilities be used. Confirm that 7 cases are misclassified. Now repeat the analysis after standardizing the variable to Z scores (using the Analyze/Description Statistics/Descriptive menu), check "Save standardized values". Are there more or less misclassifications? Show that the latent class model is unchanged when Z scores are used.
Assigned Reading:

“Session 1 Reading.pdf”

SAGE Article:

L: Section 4.3 (pages 81-85)