Session 3

Introduction to Some Advanced Topics

Session Outline:

A. Incorporating/ accommodating cases with known class membership

B. Latent Markov modeling for longitudinal data analysis

C. Multilevel models

D. Continuous factors (CFactor) and individual-level parameters

E. Controlling for the ‘level’ effect: specifying a random intercept model as an alternative to ‘centering’

F. Controlling for the ‘scale’ effect: using scale classes and scale CFactors

G. Using previously estimated models to score new cases

A. Incorporating/ accommodating cases with known class membership

Sometimes, you might have a priori information – perhaps, from an external source -- on the class membership of some individuals. For example, in a four-class situation, one may know that case 5 belongs to latent class 2 and case 11 to latent class 3. Similarly, you may have a priori information on which class cases do not belong. For example, again in a four-class situation, you may know that case 19 does not belong to latent class 2 and that case 41 does not belong to latent classes 3 or 4. In Latent GOLD, there is an option -- called "Known Class" -- for indicating to which latent classes cases do not belong. For technical details as to how the likelihood function is modified to incorporate this option see section 2.5 of the Latent GOLD Technical Guide (see pages 17-18 of Session 3 Assigned Reading).

Common applications include:

1. Multiple group models (we will explore an example of this in an exercise below)
2. using new data to refine old segmentation models while maintaining the segment classifications of the original sample
3. archetypal analysis – define class membership a priori based on extreme response patterns that reflect theoretical “archetypes”

4. partial classification -- high cost (or other factors) may preclude all but a small sample of cases from being classified with certainty. These cases can be assigned to their respective classes with 100% certainty, and the remaining would be classified by the LC model in the usual way

5. certain cases may be known to be “type 1 OR type 2” (e.g., ‘clinically depressed’ or ‘troubled’). By excluding such cases from being in say class 3 = ‘healthy’, such cases can be pre-assigned to be in class 1 or 2, while additional cases may be freely classified into any class

6. post-hoc refinement of class assignment where modal assignment for certain cases is judged to be implausible based on the desired interpretation of the classes.

Multiple group example: Open the file gss82.sav and specify a full (complete heterogeneity) 2-group LC cluster model that is equivalent to specifying separate 3-class models for the White and Black samples. To accomplish this, follow the following steps:

1. Include the 4 indicators and specify them as Nominal.

2. Specify 6 latent classes.

3. In the ClassPred Tab specify the variable RACE as the known class indicator and select SCAN.

4. Assign checkmarks so that Whites are assigned to classes 1,2 and 3 and Blacks to classes 4, 5, and 6.

5. Estimate the model.

Optional Reading:

“Session 3 Reading.pdf”

Latent GOLD Technical Guide

- Section 2.5 (pages 17-18)

Exercise A.

This exercise uses results from the multiple group example above.

- From the Profile output, interpret and name the 3 classes for the White sample (classes 1, 2 and 3). Repeat for the Black sample (classes 4, 5 and 6).

- Did you assign the same name to classes 1 and 4? To classes 2 and 5? 3 and 6?

- For the White sample, what percentage of the cases are in class 1? class 2? and class 3? (Hint: since these 3 probabilities must sum to 1, divide the corresponding probabilities shown in the Profile output for classes 1, 2, and 3 by the sum of these 3 probabilities. Repeat for the Black sample, using the sum of probabilities associated with classes 4, 5 and 6.

- Now estimate a 3-class model using RACE as an active covariate instead of a known class indicator. Under this model,
  a) what percentage of the White sample are in class 1?, 2? 3?
  b) what percentage of the Black sample are in class 1?, 2? 3?
  Hint: To obtain the class probabilities for each sample, look at the covariate portion of the ProbMeans output.

- Note that the 3-class model is a special case of the 6-class model where the conditional probability parameters are restricted to be identical for each sample. To test whether these restrictions are consistent with the data, we will use a L-square difference test. Compute the difference between the Lsquare fit statistics for 6 class and the restricted 3 class models. Use the chi-prob calculator (in the View menu) to compute the p-value associated with this Lsquare difference statistic. How many degrees of freedom should you use for this test?
  (Hint: why is the number of parameters estimated under the 6-class model not equal to exactly twice that of the 3-class model?)

Note: Should you need assistance setting up the 6-class model, see the section in the Latent GOLD User's Manual - ClassPred Tab: Restricting Cases Known (Not) to Belong to a Certain Class or Classes, (pages 26-28 of Session 3 Assigned Reading).

B. Using Mixture Latent Markov Models for Analyzing Change with Longitudinal Data

Mixture latent Markov (MLM) models are latent class models containing both time-constant and time-varying discrete latent variables. They can be extremely useful in analyzing data arising from longitudinal surveys, clinical trials and related designs. Such models often fit data better than latent growth models since the autocorrelation structure in data often satisfies the Markov assumption made by the MLM model.

MLM models are introduced and illustrated in three tutorials with real world data examples using the new GUI implemented in Latent GOLD 5.0 which can easily accommodate even hundreds of time points.

Common variations of the MLM can be tested easily such as restricting the transition structure (change) to be time homogeneous, and restricting the mover-stayer structure so that 1 latent class (stayer class) does not change. We will see that ‘longitudinal bivariate residuals’, new in version 5.0 of Latent GOLD, can assist in selecting the most appropriate variation.

The three examples are:

Tutorial 1: Simple latent Markov model for brand switching
Tutorial 2: Analysis of Satisfaction Data
Tutorial 3: Mover-Stayer latent Markov model for studying change in drug usage

Assigned Reading:

“markov.ppt”

PowerPoint: "Using Mixture Latent Markov Models for Analyzing Change in Longitudinal Data"

View power-point slides

- B1: (pages 1-11)

“LGtutorial.Markov1.pdf”

“Markov Tutorial #1: Latent GOLD Longitudinal Analysis of Brand Loyalty”

Exercise B1.

- On page 9 of Latent Markov Tutorial 1 we see that a simple latent Markov model with only 5 parameters fits the loyalty data very well. Of the 5 parameters, how many correspond to ‘b0’, how many correspond to ‘b’ and how many to ‘a’- parameters (see Fig. 13)?

Assigned Reading:

“markov.ppt”

PowerPoint: "Using Mixture Latent Markov Models for Analyzing Change in Longitudinal Data"

View power-point slides

- B1: (slides 12-18)

“LGtutorial.Markov2.pdf”

“Markov Tutorial #2: Latent GOLD Longitudinal Analysis of Life Satisfaction”

Exercise B2.

- The overall trend in Satisfaction of the 5 years measured is nonlinear (a decline followed by an increase). Which model best explains this trend?

Optional (advanced application)

“markov.ppt”

PowerPoint: "Using Mixture Latent Markov Models for Analyzing Change in Longitudinal Data"

View power-point slides

- B3: (slides 19-32)

“LGtutorial.Markov3.pdf”

“Markov Tutorial #3: Latent GOLD Longitudinal Analysis of Sparse Data”

For a theoretical introduction to MLM models and comparisons to latent class and latent growth models see**:**

Vermunt, Tran and Magidson (2008) Latent Class Models in Longitudinal Research (2008). S. Menard (Ed), *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, 373-385. Academic Press: Burlington, MA.

http://www.statisticalinnovations.com/technicalsupport/MarkovUSEmployment.pdf

C. Multilevel models

The multilevel modeling option can be used to extend LC models to account for additional correlation in nested data, such as employees nested within departments, pupils nested within schools, clients nested within stores, patients nested within hospitals, citizens nested within regions, and repeated measurements nested within individuals. Simultaneously with the identification of latent classes at the individual level (level 1), 2 or more latent group level classes (GClasses) may also be identified. The basic idea of a multilevel LC analysis is that one or more parameters of the model of interest is allowed to vary across groups via the GClasses.

The variant of the multilevel LC model that we will focus on involves including group-level random effects in the model for the latent classes, which is a way to take into account that groups differ with respect to the distribution of their members across latent classes. Not only the intercept, but also the covariate effects may have a random part. Such models are especially useful when there are many groups.

To introduce this type of model extension, we will utilize the first example in Vermunt (2003) where there are 886 employees nested within 88 teams (groups).

Assigned Reading:

“Session 3 Reading.pdf”

Multilevel Latent Class Models, by Jeroen Vermunt, 2003

- C1: (pages 3-8)
- C2: (pages 9-10)
- C3: (pages 11-15)

Exercise C1.

- Download mierlo_socmeth.txt and multilevel.lgf

- Open multilevel.lgf. Notice that the variable TEAM is specified as the group ID in the advanced tab, and the number of GClasses is specified as 1, 2 or 3 in the GClasses box. After estimating the models, compare the output for the standard 2-class model (model 2) containing only 1 GClass with the 2-class model which includes a 3-class structure on the groups (e.g., 3 GClasses).

- Compare the Gprofile output in model 2 with that in model 5. How do you interpret these parameters? Under model 5, what percentage of the 88 teams are in GClass 1? Are these teams more or less likely to be in latent class 1?

  Note: For further technical details on the implementation of these models in Latent GOLD, see section 10 in the Latent GOLD Technical Guide (pages 19-24 of Session 3 Assigned Reading).

D. Continuous factors (CFactors) and individual- level parameters

As mentioned in Section A of Session 2, in traditional or ‘fixed effects’ regression, a common set of regression coefficient estimates apply to all cases. In contrast, random effects regression models allow for heterogeneity to exist in the coefficients, so that each case may have its own set of regression coefficients.

A standard random effects regression model, frequently estimated using Bayesian methods – referred to as Hierarchical Bayes (HB) -- assumes that the underlying heterogeneity is continuous which yields separate individual-level coefficients for each case. LC regression assumes that the underlying heterogeneity is discrete and provides separate coefficients for each latent class. See Section E for one important application of CFactors.

E. Controlling for the ‘level’ effect: specifying a random intercept model as an alternative to ‘centering’

Frequently, treating the heterogeneity as continuous in the intercept works better than treating it as discrete. Such traditional random intercept models (and other random effects models which allow for continuous heterogeneity) can be estimated in conjunction with LC models by adding up to 3 continuous factors (CFactors) in a model using the Advanced Module in Latent GOLD.

Assigned Reading:

"cfactors.ppt"

PowerPoint: "Continuous Factors"

Exercise E1.

Exercise E1:

In this exercise, we will explore the use of continuous factors (CFactors) to control for the 'level effect' in ratings data. A latent class regression model is estimated where the dependent variable is preference ratings of 15 crackers, and a nominal 15-category predictor is used for the different crackers. Different classes are identified that exhibit different preferences, controlling for their overall rating level across all crackers. These data are based on a study by the Kellogg company, as described in "Applications of latent class models to food product development: a case study" by Popper et. al.

- Download the files crackers.sav and crackers.lgf
- Estimate the 2 models. Which model is preferred according to the BIC?
- For both models, use the ClassPred Tab to request that the classification output be written to a file. Using the output file, compute the mean rating for each product, separately for each latent class segment. Verify that the segments obtained from model 2 show much clearer differences in product preferences than those obtained from model 1.
- Examine the Parameters output for both models – for which are the product effects more statistically significant?
- For model 2, the CFactor scores are also output to the file (named 'CFactor1'). What is the correlation between these scores and the average rating across all products? The latter has been computed as the variable 'avg' on the original data file, and may be included in the output file if 'avg' is specified as an inactive covariate prior to the model being estimated. How do you interpret this result?

F. Controlling for the 'scale' effect: using scale classes and scale CFactors

Assigned Reading:

"Scale Effects.pdf"

- Download crackers4.sav
- Download crackers4.scale.lgs

"mlcat.pdf"

Exercise F1.

- Estimate the models in crackers4.scale.lgs and examine the Parameters output for Model 5 ("2cl 2scl simp random intercept w/ corr"). What are the scale factors for both scale classes?

G. Using previously estimated models to score new cases and/or additional replications

After estimating a model on a sample of cases, it may be desired to use the model parameter estimates to score new cases and/or additional replications for the same cases. Depending upon the type and complexity of the LC model estimated, the formulae for computing the posterior membership probabilities and predictions for new records may be quite complex. To illustrate some of these issues associated with scoring new cases we will revisit the conjoint data example from Session 2, but use only 4 of the 8 replications for each case to estimate the parameters.

Assigned Reading:

"Session 3 Reading.pdf"

- Download LGtutorial3A.pdf
- Download CONJOINT2.sav or CONJOINT2.txt (for users without SPSS)
- Download CONJOINT2.lgf or CONJOINT2T.lgf (for users without SPSS)

Exercise G1.

- If you use only the 4 replications instead of 8, does the BIC still select the 3-class model as best?
- What if you were to use the AIC, AIC3 or CAIC criteria in place of BIC?

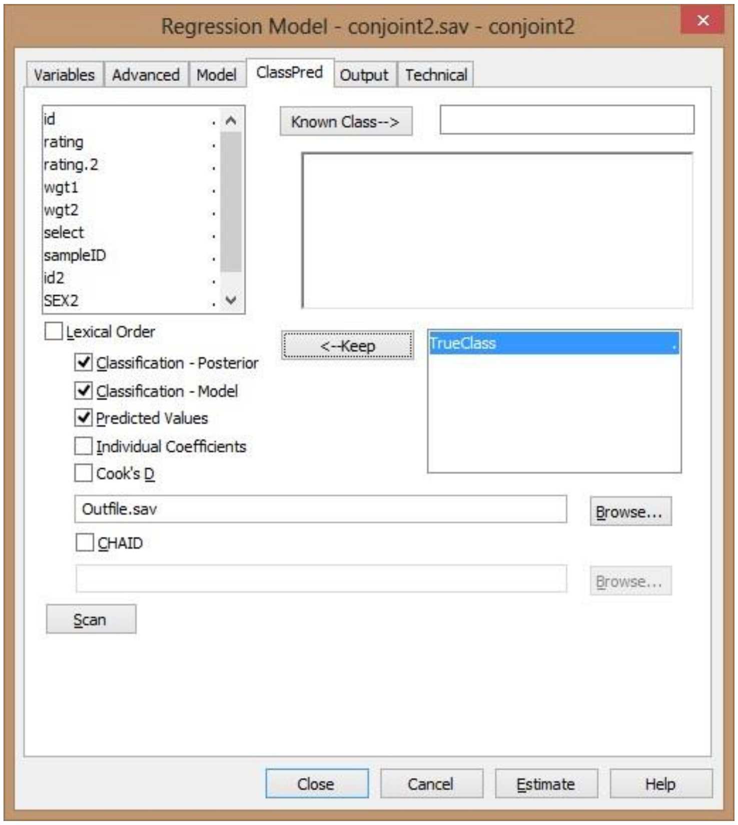
Exercise G2.

- In Tutorial #3A use RATING instead of RATING.1 as the dependent variable and utilize the variable WGT1 as the case weight (when the variable ‘ID1’ is used as the ID variable) or as the replication weight (when the variable ‘ID’ is used). Do you get the same results as obtained from following the tutorial instructions?

Exercise G3.

- When you estimate a 3-class regression model based on only 4 replications per case, is the expected misclassification rate lower than when you use all 8 replications? How about the true misclassification rate?

  Hint: To get the true class membership variable in your output file, when obtaining output to a file in Tutorial #4, specify the variable ‘TrueClass’ in the Keep box. TrueClass contains the true class membership for all cases.



Regression Model - conjoint2.sav - conjoint2

id
rating
rating.2
wgt1
wgt2
select
sampleID
id2
SEX2

TrueClass

□ Lexical Order
☑ Classification - Posterior
☑ Classification - Model
☑ Predicted Values
□ Individual Coefficients
□ Cook's D

Outfile.sav

□ CHAID