Latent Class Analysis

This is just a very brief introduction to the general concepts of latent class analysis. The topic deserves much more space than this, but this handout will give you a general idea of the purpose of the analysis and some suggestions for further readings.

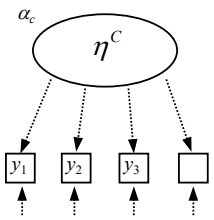
Latent Variables

Latent class analysis (Lazarsfeld & Henry, 1968; Goodman, 1974) is a kind of measurement model which estimates an unobserved construct, or *latent variable*, defined by a set of observed variables. The idea is much like a traditional factor analysis model in which a set of observed variables define an underlying continuous construct. A traditional factor analysis, for example, might involve a set of questions about various political attitudes to try to define an underlying construct of political conservativism. The advantage of such an approach is that the shared variance among the questions can be extracted to create a more reliable measure of political conservativism, removing measurement error and variance that is unique to any of the observed variables. That latent variable can then be used in regression model to improve the estimates of the associates with other variables by correcting for the attenuation that occurs with measurement error. In contrast to the factor analysis model, the latent class model groups individuals in order to identify types of voters, perhaps obtaining patterns that reflected groups such as moderate liberal, radical progressive, apolitical, and libertarian. The classes are assumed to represent nominal categories of voters and do not represent a continuum. As with factor models, the observed variables, or *indicators*, can be continuous or binary measures (ordinal indicators are also possible). The term *latent profile analysis* is used for the special case in which indicators are continuous, but latent class analysis is used more generally to refer to models whether binary or continuous indicators are involved.

Latent Class Analysis

The latent class measurement model (i.e., there are no predictors of the latent class and the latent class does not predict anything) seeks to find some set of mutually exclusive and exhaustive categories that group cases based on a set of observed variables. Either exploratory or confirmatory approaches to latent class models are possible, analogous to the distinction between exploratory and confirmatory factor analysis. For exploratory models, the number of latent classes is not specified, usually because no clear hypothesis exists about the number of latent classes. In the confirmatory form, which I will mostly focus on here, the number of classes is specified and the software provides an estimate of the fit of the data to the hypothesized number of classes. Though the true number of classes is still unknown in the confirmatory model, a number ultimately must be specified in the application. The meaning of each latent class must be inferred from the data or theory, and it is up to the researcher to name and interpret them. Contrast the example of unknown political class described above from an example of a known political class, such as when the respondent can be assigned as a Democrat, Independent, or Republican based on self-declaration or registration documents, for instance. Latent class membership of any one individual is estimated in a probabilistic fashion.

The figure below follows confirmatory factor analysis and structural equation modeling conventions to depict a latent class model. The ellipse is used to represent the latent variable, called η (the Greek eta) with a superscript C for some number of classes, and the square boxes represent the measured variables, y_j.



Each arrow represents a type of regression, where the observed variable y_j is predicted by the latent variable. The v (the Greek nu) is the intercept for that regression. If the indicators were binary, we would use τ, the Greek letter tau, for a threshold, as with generalized linear modeling thresholds.[1] Latent class models do not have traditional factor loadings (the regression slope), and this is why the arrows are represented with dotted lines here. Instead, the pattern of measurement intercepts (for continuous indicators) or response probabilities (for binary indicators) across classes indicate the strength of the relationship between factor and item. The epsilons, ε_j, represent measurement residuals that include unaccounted for (unique) variance including measurement error.

The model provides $C-1$ latent class factor mean estimates, α_c.[2] These values are like the intercepts from a logistic regression. Because there are no predictors of this latent class variable, we have an intercept only model and no regression coefficient. If there are only two classes, then then there will be one intercept value. Just like in logistic regression, the estimates of α_c are in logit form and we would need to use the logistic transformation to obtain an estimated probability of membership in a given class, exactly as we would with a logistic model.

$$\hat{\pi}_c = P\left(\eta_i^C = c\right) = \frac{e^{\alpha_c}}{1+e^{\alpha_c}}$$

The *class membership probability* is the estimate of the proportion of the sample that belongs to a certain class. In other words, for a two class model, $\hat{\pi}_1$ would tell us the proportion of cases we expect to be members of the first class, where $c=1$. If there are three classes, there will be two intercept values. The program may choose the last category as the referent class by default, but the referent group is the same as the use of $Y=0$ as the referent group in binary or multinomial logistic regression. When there are more than two classes, we extend the transformation, using the multinomial logistic transformation.

$$\hat{\pi}_c = \frac{1}{1+\sum_{c=1}^{C} e^{\alpha_c}}$$

The probabilities for all of the classes must sum to 1.0.

$$\sum_{c=1}^{C} \hat{\pi}_c = 1$$

Mixture Models

A focus on multiple indicators in a more traditional latent class analysis has served to provide an introduction to some of the underlying principles of categorical latent variables. A more general framework, suggested by Muthén and colleagues (Clark & Muthén, 2009; Lubke & Muthén, 2005; Muthén, 2001; 2002; Muthén & Muthén, 2000), combines categorical and continuous latent variables in the same model. These *structural equation mixture models* open a variety of possible avenues for investigating hypotheses involving unknown groups in two important ways—(1) associations among categorical latent variables, continuous latent variables, continuous observed variables, or categorical observed variables can be examined, and (2) categorical latent variables can be used in a flexible manner that allows classification of observed or latent variables.

[1] When the indicators are binary the thresholds, τ_j, can be converted using the logistic cdf transformation to obtain the estimated response probabilities, which are closely related to response probabilities in item response theory (IRT), except that the latent variable is categorical in the latent class model but usually assumed to be continuous in the standard IRT model.

[2] I apologize that my notation differs from the Collins and Lanza course reading. They use L for a particular latent class and p for the probability. I'm trying to be consistent with more writings on the subject and our main text.

Estimation, Model Identification, and Fit

The most common estimator for latent class models is maximum likelihood using an expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). In the EM steps of the ML process, conditional expectations and the posterior class membership probabilities are computed in the expectation step and parameter estimates are updated. The fit is then maximized through iterations in the maximization step. This process alternates between the two steps until an optimization criterion is reached. Estimation can be sensitive to start values and it is wise to retest any model with different start values to be certain that convergence was reached at a global solution not a local solution (Hipp & Bauer, 2006), a testing process that may be automated within the software program. Many packages now employ random starts, and the user can specify the number of sets of random start values the computers uses.[3] A log-likelihood value obtained upon convergence is used to compute fit indices.

For identification, the number of classes must be less than the number of indicators unless there are additional constraints (e.g., equal variances for the indicators across classes). Thus, with three indicators, only two classes can be specified without special restrictions. For latent class models that have binary indicators and do not include covariates or latent variables, a likelihood ratio chi-squared (G^2) and Pearson chi-squared are computed to assess fit. Degrees of freedom for these tests can be calculated by $df = \left[2^J - 1\right] - q$, where q is equal to the number of latent class means estimated ($C - 1$) plus the number of class-specific intercepts. For example, with binary 4 indicators and 2 classes, there are 6 degrees of freedom. With three indicators and two classes, the model is just identified. Though the model may be theoretically identified, it may not be empirically identified and restrictions may be necessary for convergence. Models with a large number of binary indicators, fewer cases, large number of classes, low membership proportions in one or more classes, or sparse data (i.e., low frequency in the contingency table) may be more susceptible to convergence issues (Lubke & Muthén, 2005; 2007).

The likelihood ratio chi-squared may be problematic for sparse data, where the Pearson chi-squared is sometimes substituted. The likelihood ratio chi-squared is not valid for comparing different number of classes (Lanza, Bray, & Collins, 2013), however. No chi-squared model fit is available for latent class models with continuous indicators. Instead assessment of fit must rely on likelihood-based fit indices, such as the Akaike Information Criteria and the Bayesian Information Criteria, which are commonly used for evaluation of fit relative to comparison models. The sample size adjusted BIC (aBIC; Sclove, 1987) seems to perform better than other information criteria (e.g., Nylund, Asparouhov, & Muthén, 2007).

$$\text{aBIC} = \text{-}2LL + q \cdot \ln\left(\frac{N+2}{24}\right)$$

The quantity -2LL is the -2 log likelihood value for the H_0 model, q is the number of free parameters, and *N* is the sample size. The number of free parameters q is $2JC - (J-1)(C-1)$.

Another concept related to model fit is *entropy*. Entropy is a measure of the overall accuracy of classification or class separation. Although there are several possible measures of accuracy, the entropy index, E, which represents a kind of average of the natural log of all class membership probabilities, is the most frequently employed (Ramaswamy, DeSarbo, Reibstein, & Robinson, 1993):

$$E = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C}\left(-\hat{\pi}_{ic}\ln\hat{\pi}_{ic}\right)}{N\ln C}$$

C is the number of classes and $\hat{\pi}_{ic}$ is the predicted class membership probability for an individual. Values of E can range between 0 and 1. Higher values indicate greater separation and therefore better fit in one

[3] I use a minimum of 20 and increase the number with more complex models, often rerunning models with more random starts to confirm that the the results do not change

sense.[4] Although some authors give cutoff suggestions for acceptable entropy values (I often hear .8), entropy is probably best used for comparing different models (Celeux & Soromenho, 1996).

Determining the Number of Classes

Part of the process of latent class analysis involves deciding on the correct number of classes, sometimes called *class enumeration*. Although the researcher may have *a priori* hypotheses about the number of classes, comparisons are usually made among models with different numbers of classes to provide evidence that the number of classes is correct. The difference in the log likelihoods for two models with a different number of classes is not distributed as a chi-squared, so an exact test to compare models does not exist. The BIC or adjusted BIC is commonly used for this purpose (lower values indicating better fit) and performs fairly well (Tofighi & Enders, 2006), but a number of simulation studies suggest that more precise methods may be preferable. These methods are designed to compare two models that differ by only one latent class. Among the several proposed alternatives are a bootstrapped likelihood ratio test (Nylund et al., 2007), the Lo-Mendell-Rubin adjusted likelihood ratio test, and the Vuong-Lo-Mendell-Rubin likelihood ratio test (Lo, Mendell, Rubin, 2007; Vuong, 1989). Nylund and colleagues provide evidence that the parametric bootstrap standard error and p-value adjustment to the Vuong-Lo-Mendell-Rubin (VLMR) comparison is performers preferably.

An additional complication is that for every latent class model there is an equivalent continuous factor model that can account for the same covariance matrix, where the continuous factor model has one fewer factors than classes in the latent class model (e.g., Bauer & Curran, 2004; Bartholomew, 1987). The equivalence of these models has sparked considerable discussion among statisticians that is likely to be continued, but, for the applied researcher, there is currently no simple way to distinguish among the two types of models on an empirical basis. There is no choice for the researcher but to decide which type of model is most appropriate based on theoretical considerations in the context of the questions most of interest.

Software

Most software programs use a maximum likelihood estimation via the expectation maximization (EM) algorithm, but a Bayesian process is also possible. There are a variety of software programs that estimate latent class models, including the `poLCA` and `lcca` packages in R, PROC LCA, which is a free macro for SAS (Lanza, Collins, Lemmon, & Schafer, 2007) and Latent Gold (Vermunt & Magidson, 2005), and structural equation modeling packages, such as Mplus (Muthén & Muthén, 1998–2012) and Mx (Boker et al., 2012) integrate latent class variables within the larger structural equation modeling framework (the so-called mixture modeling approach). Mplus uses the maximum likelihood-EM approach with a robust standard error adjustment as the default.

Examples

The example below uses a set of items about sleep difficulty taken from the Australian sleep study.

R

I use the `poLCA` package in R (Linzer & Lewis, 2011). It requires numeric variables coded 1 and 2 (or they must be positive integers). The `nrep` keyword is for the number of start values.

```
>
> library(haven)
> d = read_sav("c:/jason/spsswin/cdaclass/sleep.sav")
>
> library(summarytools)
>
> library(poLCA)
> lcamod = poLCA(cbind(trubslep, trubstay, wakenite, liteslp,
+                      refreshd, medhelp,problem, stopb, restlss, drvsleep,
+                      drvresul) ~ 1,maxiter=50000, nclass=2, nrep=25, data=d)
Model 1: llik = -1178.771 ... best llik = -1178.771
Model 2: llik = -1178.771 ... best llik = -1178.771
Model 3: llik = -1181.842 ... best llik = -1178.771
Model 4: llik = -1178.771 ... best llik = -1178.771
Model 5: llik = -1178.771 ... best llik = -1178.771
Model 6: llik = -1181.842 ... best llik = -1178.771
Model 7: llik = -1178.771 ... best llik = -1178.771
```

[4] The usage of entropy is seemingly the opposite of the use in physics, where it is a tendency toward disorganization.

```
Model 8: llik = -1181.842 ... best llik = -1178.771
Model 9: llik = -1178.771 ... best llik = -1178.771
Model 10: llik = -1178.771 ... best llik = -1178.771
Model 11: llik = -1178.771 ... best llik = -1178.771
Model 12: llik = -1178.771 ... best llik = -1178.771
Model 13: llik = -1178.771 ... best llik = -1178.771
Model 14: llik = -1178.771 ... best llik = -1178.771
Model 15: llik = -1181.842 ... best llik = -1178.771
Model 16: llik = -1181.842 ... best llik = -1178.771
Model 17: llik = -1178.771 ... best llik = -1178.771
Model 18: llik = -1178.771 ... best llik = -1178.771
Model 19: llik = -1178.771 ... best llik = -1178.771
Model 20: llik = -1178.771 ... best llik = -1178.771
Model 21: llik = -1181.842 ... best llik = -1178.771
Model 22: llik = -1181.842 ... best llik = -1178.771
Model 23: llik = -1181.842 ... best llik = -1178.771
Model 24: llik = -1178.771 ... best llik = -1178.771
Model 25: llik = -1178.771 ... best llik = -1178.771
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)

$trubslep
           Pr(1)  Pr(2)
class 1:  0.4043 0.5957
class 2:  0.7951 0.2049

$trubstay
           Pr(1)  Pr(2)
class 1:  0.2275 0.7725
class 2:  0.9185 0.0815

$wakenite
           Pr(1)  Pr(2)
class 1:  0.0733 0.9267
class 2:  0.3180 0.6820

$liteslp
           Pr(1)  Pr(2)
class 1:  0.3812 0.6188
class 2:  0.7828 0.2172

$refreshd
           Pr(1)  Pr(2)
class 1:  0.2156 0.7844
class 2:  0.5444 0.4556

$medhelp
           Pr(1)  Pr(2)
class 1:  0.8874 0.1126
class 2:  1.0000 0.0000

$problem
          Pr(1) Pr(2)
class 1:  0.223 0.777
class 2:  0.914 0.086

$stopb
           Pr(1)  Pr(2)
class 1:  0.9269 0.0731
class 2:  0.9027 0.0973

$restlss
           Pr(1)  Pr(2)
class 1:  0.4778 0.5222
class 2:  0.9195 0.0805

$drvsleep
           Pr(1)  Pr(2)
class 1:  0.7966 0.2034
class 2:  0.9307 0.0693

$drvresul
           Pr(1)  Pr(2)
class 1:  0.8716 0.1284
class 2:  0.9246 0.0754

Estimated class population shares
 0.5152 0.4848

Predicted class memberships (by modal posterior prob.)
 0.5134 0.4866

=========================================================
Fit for 2 latent classes:
=========================================================
number of observations: 224
number of estimated parameters: 23
residual degrees of freedom: 201
maximum log-likelihood: -1178.771

AIC(2): 2403.543
BIC(2): 2482.011
G^2(2): 403.8194 (Likelihood ratio/deviance statistic)
X^2(2): 2460.182 (Chi-square goodness of fit)
```

SAS

SAS does not have a built in LCA procedure, but Lanza and colleagues (2015) have developed a macro that can be installed into the SAS program folder (see the SAS LCA macro installation instructions: https://www.methodology.psu.edu/files/2019/03/Installing_PROC_LCA_M4-23myjrn.pdf).[5]

(proc syntax was omitted in original version)

```
proc lca data=one;
nclass 2;
items trubslep trubstay wakenite liteslp refreshd medhelp
problem stopb restlss drvsleep drvresul;
categories 2 2 2 2 2 2 2 2 2 2;
seed 51921;
nstarts 40;
rho prior=1;
run;
```

```
               Data Summary, Model Information, and Fit Statistics (EM Algorithm)

Number of subjects in dataset:         224
Number of subjects in analysis:        224

Number of measurement items:            11
Response categories per item:            2 2 2 2 2 2 2 2 2 2 2
Number of groups in the data:            1
Number of latent classes:                2

NOTE: A data-derived prior was applied to the rho parameters to help
      avoid parameter estimates on boundary values of zero and one.

Rho starting values were randomly generated (seed = 51921).

No parameter restrictions were specified (freely estimated).

Seed selected for best fitted model:    1231039985
Percentage of seeds associated with best fitted model:  100.00%

The model converged in 56 iterations.

Maximum number of iterations: 5000
Convergence method: maximum absolute deviation (MAD)
Convergence criterion:  0.000001000

============================================
Fit statistics:
============================================

Log-likelihood:     -1178.81
G-squared:            403.90
AIC:                  449.90
BIC:                  528.37
CAIC:                 551.37
Adjusted BIC:         455.48
Entropy:                0.76
Degrees of freedom:     2024
                        Parameter Estimates

Class membership probabilities: Gamma estimates (standard errors)
Class:                      1          2
                       0.5126     0.4874
                      (0.0493)   (0.0493)
```

[5] PROC LCA & PROC LTA (Version 1.3.2) [Software]. (2015). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

```
Item response probabilities: Rho estimates (standard errors)
  Response category  1:
Class:                    1         2
  trubslep    :       0.4043    0.7929
                     (0.0499)  (0.0461)
  trubstay    :       0.2275    0.9148
                     (0.0521)  (0.0420)
  wakenite    :       0.0734    0.3167
                     (0.0283)  (0.0482)
  liteslp     :       0.3812    0.7806
                     (0.0499)  (0.0483)
  refreshd    :       0.2151    0.5431
                     (0.0468)  (0.0521)
  medhelp     :       0.8871    0.9997
                     (0.0305)  (0.0019)
  problem     :       0.2228    0.9105
                     (0.0544)  (0.0409)
  stopb       :       0.9266    0.9032
                     (0.0252)  (0.0293)
  restlss     :       0.4779    0.9171
                     (0.0511)  (0.0374)
  drvsleep    :       0.7963    0.9303
                     (0.0397)  (0.0265)
  drvresul    :       0.8716    0.9244
                     (0.0335)  (0.0283)

  Response category  2:
Class:                    1         2
  trubslep    :       0.5957    0.2071
                     (0.0499)  (0.0461)
  trubstay    :       0.7725    0.0852
                     (0.0521)  (0.0420)
  wakenite    :       0.9266    0.6833
                     (0.0283)  (0.0482)
  liteslp     :       0.6188    0.2194
                     (0.0499)  (0.0483)
  refreshd    :       0.7849    0.4569
                     (0.0468)  (0.0521)
  medhelp     :       0.1129    0.0003
                     (0.0305)  (0.0019)
  problem     :       0.7772    0.0895
                     (0.0544)  (0.0409)
  stopb       :       0.0734    0.0968
                     (0.0252)  (0.0293)
  restlss     :       0.5221    0.0829
                     (0.0511)  (0.0374)
  drvsleep    :       0.2037    0.0697
                     (0.0397)  (0.0265)
  drvresul    :       0.1284    0.0756
                     (0.0335)  (0.0283)
```

References and Further Reading

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. London: Griffin.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychological methods*, *9*, 3-29.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Estabrook, R., Bates, T., Mehta, P., Oertzen, T. v., Gore, R., Hunter, D., Hackett, D., Karch, J. & Brandmaier, A. (2012) *OpenMx version 1.3*. Retrieved from . http://openmx.psyc.virginia.edu.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*, 195-212.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311-359). New. York: Plenum.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part IA modified latent structure approach. *American Journal of Sociology, 79*, 1179–1259.

Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological methods*, *11*, 36-53.

Lanza, S. T., Bray, B. C., & Collins, L. M. (2013). An introduction to latent class and latent transition analysis. In J. A. Schinka, W. F. Velicer, & I. B. Weiner (Eds.), *Handbook of psychology* (2nd ed., Vol. 2, pp. 691-716). Hoboken, NJ: Wiley.

Lanza, S. T., Dziak, J. J., Huang, L., Wagner, A., & Collins, L. M. (2015). *PROC LCA & PROC LTA users' guide (Version 1.3.2)*. University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling*, *14*(4), 671-694.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software, 42*(10), 1-29.

Ramaswamy, V., DeSarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science, 12*, 103-124.

Lo, Y., Mendell, N., & Rubin, D. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767- 778.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods*, *10*, 21.

Lubke, G., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling: A Multidisciplinary Journal,14*, 26-47.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus User's Guide, Seventh Edition*. Los Angeles: Muthén & Muthén.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal,14*, 535-569.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333–343.

Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in a growth mixture model. In G. R. Hancock (Ed.), Mixture models in latent variable research (pp. 317–341). Greenwich, CT: Information Age.

Vermunt, J. K., & Magidson, J. (2005). Latent Gold 4.0. *User's Guide*.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57*, 307-333.