LcaBootstrap SAS Macro Users' Guide (Version 4.0)

John J. Dziak
Stephanie T. Lanza
Penn State

© 2016 The Pennsylvania State University

Please send questions and comments to *MChelpdesk@psu.edu*.

The development of these SAS Macros was supported by the National Institute on Drug Abuse Grant P50-DA10075 to The Center for Prevention and Treatment Methodology. Shu Xu made significant contributions to a previous version of this software.

The suggested citation for this users' guide is
Dziak, J. J., & Lanza, S. T. (2016). *LcaBootstrap SAS macro users' guide* (version 4.0). University Park: The Methodology Center, Penn State. Available from http://methodology.psu.edu.

Contents

The Methodology Center is pleased to release the %**LcaBootstrap** macro, which can assist users in choosing the number of classes for latent class analysis (LCA) models. It works in conjunction with the SAS®[1] software package (version 9.1 or higher) and the PROC LCA procedure (version 1.2.5 or higher). This macro can perform the bootstrap likelihood ratio test to compare the fit of a latent class analysis (LCA) model with k classes ($k \geq 1$) to one with $k + 1$ classes.

Version 4 of the macro represents a small bug fix to version 1.1 following some additional testing, and is being released in conjunction with a function for the Stata software system that provides similar functionality.

1 About the %LcaBootstrap Macro

The parametric bootstrap likelihood ratio test for LCA is described in McLachlan and Peel (2000); Nylund, Asparouhov, and Muthén (2007); and Collins, Fidler, Wugalter, and Long (1993). It is used to choose the number of classes in an LCA. More specifically, it tests the null hypothesis that a given k-class LCA model is adequate to describe the population a particular sample came from, versus the alternative hypothesis that a $(k + 1)$-class LCA model is required. To do this, the null and alternative models are first fit to a given empirical dataset, and then the difference between them (in terms of a likelihood ratio test statistic) is recorded. Many random samples are then generated from a population in which the k-class null hypothesis is true, and then analyzed under both the k-class and the $(k + 1)$-class models, to get an estimate of what the likelihood ratio test statistic distribution would be if the null hypothesis were true. If the observed likelihood ratio test statistic is bigger than most (say, 95%) of the simulated likelihood ratio test statistics, then H_0 is rejected.

The bootstrap p-value as used here is $(s + 1)/(B + 1)$ where B is the total number of bootstrap samples generated and s is the number of bootstrap datasets having a likelihood ratio test statistic larger than that of the observed sample.

One limitation of the bootstrap test is that it can take hours to perform, since many LCA datasets are being simulated and analyzed. Also, the current version of this macro is only for classic LCA models without covariates or special sampling features. Specifically, it is assumed that the models being compared do not have polytomous items as indicators (those with more than 2 possible responses) and do not have COVARIATES, GROUPs, WEIGHTs, CLUSTERs, or special parameter restrictions (RESTRICT option), even though these features are allowed in PROC LCA (See PROC LCA User's Guide; Lanza, Dziak, Huang, Xu & Collins, 2011).

Note: The current version of the %LcaBootstrap macro only handles LCA models based on dichotomous items (e.g., yes/no). Items with more than 2 possible responses are not currently supported in this macro.

[1] SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

2 Using the Macro

2.1 Preparation

A SAS macro is a special block of SAS commands. The block is first defined, and then called when it is needed. The procedure for using an LCA macro is very straightforward.
Several steps need to be completed before running the macro:

1. If you haven't already done so, download and save the macro at the designated path (e.g., *S:\myfolder*).
2. Direct SAS to read the macro code from the path, using a SAS %INCLUDE statement such as
   ```
   %INCLUDE "S:\myfolder\LcaBootstrap.sas";
   ```
3. Fit the null (smaller) and alternative (larger) LCA models in PROC LCA and save their OUTEST and OUTPARAM data files (see the PROC LCA & PROC LTA Users' Guide). The null model must have exactly one fewer class than the alternative model. The %LcaBootstrap macro needs to read certain results from PROC LCA, in the form of SAS datasets, from both the null and alternative models. The needed datasets can be automatically created using OUTPARAM= and OUTEST= in PROC LCA, which are described further in the PROC LCA Users' Guide. For example, in the first line of your SAS syntax for PROC LCA, instead of only specifying
   ```
   PROC LCA DATA=mydata;
   ```
   the user can specify
   ```
   PROC LCA DATA=mydata OUTEST=est1 OUTPARAM=param1;
   ```
   The datasets est1 and param1 for a specified model will be created and the relevant parts of the LCA results will be recorded there.

2.2 Syntax and Input

Call the macro using a percent sign, its name, and user-defined arguments in parentheses. The "arguments" or "parameters" of the macro (i.e., the information in parentheses provided to the macro) are shown below.

```
%LcaBootstrap(   null outest = filename1,
                 alt outest = filename2,
                 null outparam = filename3,
                 alt outparam = filename4,
                 n = number,
                 num bootstrap = number,
                 num starts for null = number,
                 num starts for alt = number,
                 cores = number);
```

Arugment	Required	Description
null_outest	Y	Name of the OUTEST= datasets from PROC LCA run on the null (*k*-class) model
alt_outest	Y	Name of the OUTEST= datasets from PROC LCA run on the alternative ((*k*+1)-class) model
null_outparam	Y	Name of the OUTPARAM= dataset from PROC LCA run on the null model
alt_outparam	Y	Name of the OUTPARAM= dataset from PROC LCA run on the alternative model
n	Y	Original sample size used for the LCA, counting only those cases included in the analysis
num_bootstrap	N	Number of bootstrap replications, which should be at least 99. A value of 999 is preferable but calculations may take several days. **(default = 99)**
num_starts_for_null	N	Number of starting values to fit under the null hypothesis in each bootstrap replication to help find the global maximum likelihood under this hypothesis. Must be at least 2, but 20 or more is recommended. **(default = 20)**
num_starts_for_alt	N	Number of starting values to fit under the alternative hypothesis in each bootstrap replication to help find the global maximum likelihood under this hypothesis. It must be less than *num_starts_for_null* as the alternative model is more complex. It must be at least 2, but 20 or more is recommended **(default = 20)**
cores	N	Number of processor cores to use if more than one is available. **(default = 1)**

2.3 Output

If the calculations for the test are successful, the bootstrap *p*-value is displayed on the screen. Also, three temporary SAS datasets are created:

- **CalculatedLrtForBootstrap** contains the likelihood ratio test statistic (specifically, negative two times the difference in alternative and null log-likelihoods) for the original dataset provided.
- **LcaBootstraps** contains this test statistic for each of the generated bootstrap datasets. This list of numbers is the basis for estimating the null hypothesis distribution of the test statistic to obtain a *p*-value. All of the entries in the *loglikRatioTest* column should be nonnegative. If any are negative, it suggests that the alternative model is very poorly identified and *num_starts_for_alt* should be increased.
- **BootstrapResult** contains the *p*-value for the test.

  Note: The BootstrapResult dataset also contains the variables startvals_agree_H0 and startvals_agree_H1. These provide some information on how well-identified the bootstrap datasets were: they are the average over the bootstrap datasets of the proportion of starting values agreeing with the best fit value. Although there are no generally agreed-upon standards for interpreting these proportions, they may be of some help in model selection.

3 Appendix: Example

The following example is based on simulated data, which are loosely based on the adolescent delinquent behaviors example shown on pages 11-12 of Collins and Lanza (2010). The analysis presented here is based on N = 2000 adolescents' responses on five questionnaire items: *Rowdy*, *Vandal*, *Shoplift*, *Steal*, and *Fight*. First a 2-class and then a 3-class model are fit to the data. The bootstrap likelihood ratio test is used to test whether the 2-class model is adequate relative to the 3-class alternative.

```
%INCLUDE "C:\Work\lca\LcaBootstrap.sas";
DATA Delinquency;
INPUT  Rowdy Vandal Shoplift Steal Fight count;
DATALINES;
       1   1   1   1   1   50
       1   1   1   1   2   66
       1   1   1   2   1   12
       1   1   1   2   2   24
       1   1   2   1   1   15
       1   1   2   1   2   9
       1   1   2   2   1   41
       1   1   2   2   2   71
       1   2   1   1   1   37
       1   2   1   1   2   107
       1   2   1   2   1   13
       1   2   1   2   2   32
       1   2   2   1   1   6
       1   2   2   1   2   17
       1   2   2   2   1   107
       1   2   2   2   2   379
       2   1   1   1   1   2
       2   1   1   1   2   27
       2   1   1   2   1   1
       2   1   1   2   2   8
       2   1   2   1   2   1
       2   1   2   2   1   4
       2   1   2   2   2   25
       2   2   1   1   1   13
       2   2   1   1   2   67
       2   2   1   2   1   6
       2   2   1   2   2   45
       2   2   2   1   1   2
       2   2   2   1   2   4
       2   2   2   2   1   52
       2   2   2   2   2   749
;
RUN;
PROC LCA DATA=delinquency OUTEST=est2 OUTPARAM=par2;
    NCLASS 2;
    ITEMS Rowdy Vandal Shoplift Steal Fight ;
    FREQ Count;
    CATEGORIES 2 2 2 2 2;
    SEED 1000;
    NSTARTS 10;
RUN;
PROC LCA DATA=delinquency OUTEST=est3 OUTPARAM=par3;
    NCLASS 3;
    ITEMS Rowdy Vandal Shoplift Steal Fight ;
    FREQ Count;
    CATEGORIES 2 2 2 2 2;
    SEED 1000;
    NSTARTS 10;
```

```
RUN;
%LcaBootstrap(null outest = est2,
              alt outest = est3,
              null_outparam = par2,
              alt outparam = par3,
              n = 2000,
              num bootstrap = 99,
              num_starts_for_null = 20,
              num starts for_alt = 20,
              cores = 1);
```

Example results:

The *p*-value is .01, suggesting that the 2-class model is too restrictive.

```
                    ANSWER_FROM_BOOTSTRAP

                                The bootstrap p-value was        0.01
```

Note that because the bootstrap procedure is partially random, your *p*-value may not be exactly the same, even with the same data. To reduce the random variability of the *p*-value, use a higher value for `num_bootstrap`, such as 499 or 999.

Model comparisons:

Sequential analyses have been conducted for selecting the number of latent classes of adolescent delinquent behaviors based on bootstrap likelihood ratio tests. The results are summarized in the table below. Because no difference was detected between a 4-class model and a 5-class model in terms of likelihood ratio test statistics, we select a four-class model of adolescent delinquent behaviors.

Null model	**Vs.**	**Alternative model**	**p-value**
1-class		2-class	.01
2-class		3-class	.01
3-class		4-class	.01
4-class		**5-class**	**.45**

4 References

Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. L. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research, 28,* 375-389.

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

Lanza, S. T., Dziak, J. J., Huang, L., Xu, S., & Collins, L. M. (2011). *Proc LCA & Proc LTA users' guide* (version 1.2.6). University Park: The Methodology Center, Penn State. Available from http://methodology.psu.edu.

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York, NY: Wiley.

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, 14,* 535-569.