Supplemental material for: Brandenburger, M., & Schwichow, M. (2023): Utilizing Latent Class Analysis (LCA) to Analyze Response Patterns in Categorical Data. In: Liu, X., & Boone, W. J. (Eds.), *Advances in Applications of Rasch Measurement in Science Education*. Springer. DOI: 10.1007/978-3-031-28776-3_6

How to Run a Latent Class Analysis with R

In the following, we will describe how to run the LCA with the poLCA (Polytomous Variable Latent Class Analysis) package in R (Linzer & Lewis, 2011; Linzer & Lewis, 2022). Additionally, we provide the used data set and the R script as supplemental material.

Data Preparation

When using your own data, you need to prepare the data set. As discussed, the LCA needs a complete data set without missing values. You need to remove cases with missing responses or use a data imputation method (e.g., hot-deck imputation). However, poLCA can handle missing data by taking advantage of the EM algorithm (Linzer & Lewis, 2011).

Even if there is no restriction for the number of categories within each item, you need to recode your categorical responses in a way they start with "1" and go up to the max number without gaps. In our example, item 1 and 2 had the original categories 1, 2, 5 and 7 (Table 6.2). We recoded them, so the categories in the provided data set are 1, 2, 3 and 4. Therefore, you need to be careful when interpreting your results – a "3" in item 1 is a different category than a "3" in item 3.

Running the Analysis

We run the analysis with poLCA. In line 8, we define the number of classes. Line 9 defines the formula of the poLCA model with the variables from the data file used for the LCA, in our case `ID1`, `ID2`, `IN`, `UN`. In line 18, we start the analysis. If needed, you can change the max number of iterations (`maxiter=1000`) or the convergence criterion (`tol=.001`). If you want to retain cases with missing values, you have to code the missings with `NA` and add `na.rm=FALSE` to the poLCA function.

To verify you have found a global maximum, you have to run the analysis several times and make sure the estimations produce the highest log-likelihood (section 6.4.3). poLCA has automated this procedure. `nrep` defines the number of times to estimate the model using different start values. poLCA will choose the model with the greatest log-likelihood. You should at least use `nrep=10`. With the data in the supplemental material you will find the global maximum of -2257.13, but also several other very similar log-likelihoods, so we recommend a higher `nrep`, depending on your data.

1: Running the LCA with poLCA

```
1  # install.packages("poLCA")
2  library(poLCA) # loading poLCA
3
4  setwd(your working directory)
5
6  file <- "CVSdata.csv" # name of the data file
7
8  nclasses <- 3 # number of classes
9  f1 <- as.formula(cbind(ID1, ID2, IN, UN)~1) # selecting the variables
10
11 ######################################################################
12
13 mydata <- read.csv2(file) # reading the data file
14
15 # Running the analysis
16 # max number of iterations: 1000, convergence criteria: .001
17 # number of times to estimate the model, using different start values: 30
18 LCA <- poLCA(f1, data=mydata, nclass=nclasses, maxiter=1000, tol=.001, nrep=30)
```

poLCA can simulate data based on given model parameters. Therefore, we can perform a simple parametric bootstrap for χ^2 (section 6.4.4.3). If you do not want to perform the bootstrap, skip line 23 to 51.

First, we specify the number of bootstrap samples (`bootstrapN`) in line 23. If you want to use the bootstrap for estimating the concrete empirical distribution of χ^2, you have to choose a large number. If the bootstrap is only used for testing the model fit, 20 to 40 is sufficient (Davier, 1997a). In line 29, we resample a data set based on the model parameters of the LCA and use this resampled data set to run a new LCA in line 31 - 33. We use the same iteration criteria but use a lower `nrep=1` to save computing time. Choose a higher `nrep` for a better accuracy. We save the observed pattern frequency test statistic χ^2 in a list (line 35). This procedure will be repeated `bootstrapN` times (line 27).

As a simple bootstrap approach suggested by Rost (2004), we calculate the %-rank of the original χ^2 test static in the resampled distribution (ordered by size) and calculate based on the %-rank of χ^2 an empirical p-value (line 39 - 40). For an optical check, we create a histogram of the resampled χ^2 distribution (line 49 – 51). The red line marks the χ^2 test statistic of the original LCA, the black dotted line cuts off the 5% highest χ^2 values of the resampled data (Figure 1). If the red line is left of the dotted line, than the original LCA model fits the data because there is no significant difference between the original model and most of the resampled models.

2: Performing the bootstrap (optional)

```
23 bootstrapN <- 40 # max number of bootstrap samples
24
25 bootstrapCHIsq <- LCA$Chisq
26
27 for (i in 1:bootstrapN){
28   # simulating a data set based on the calculated model parameters
29   simdata <- poLCA.simdata(N=nrow(mydata),probs=LCA$probs,P=LCA$P)
30   # running the LCA for the resampled data set
31   simLCA <- poLCA(
32     as.formula(paste0("cbind(",paste0("Y",1:ncol(LCA$y),collapse=","),")~1")),
33     data=simdata$dat, nclass=nclasses, maxiter=1000, tol=.001, nrep=1)
34   # save the resampled CHI^2 value in a list
35   bootstrapCHIsq <- c(bootstrapCHIsq, simLCA$Chisq)
36 }
37
38 # rank the original CHI^2 within the resampled distribution
39 bootstrapCHIsq <- sort(bootstrapCHIsq)
40 empP <- 1-(match(LCA$Chisq, bootstrapCHIsq)/(bootstrapN+1))
41
42
43 # creating a histogram of the resampled CHI^2 distribution
44 # red line with CHI^2 value of the original LCA
```

```
45 # black dotted line cuts off the 5% highest CHI^2 values of the resampled data
46 # if the red line is left of the dotted line, than the original LCA model
47 # fits the data because there is no significant difference between
48 # the original model and most of the resampled models
49 hist(bootstrapCHIsq)
50 abline(v=LCA$Chisq, col="red")
51 abline(v=bootstrapCHIsq[floor((bootstrapN+1)-((bootstrapN+1)*0.05))],lty=3)
```

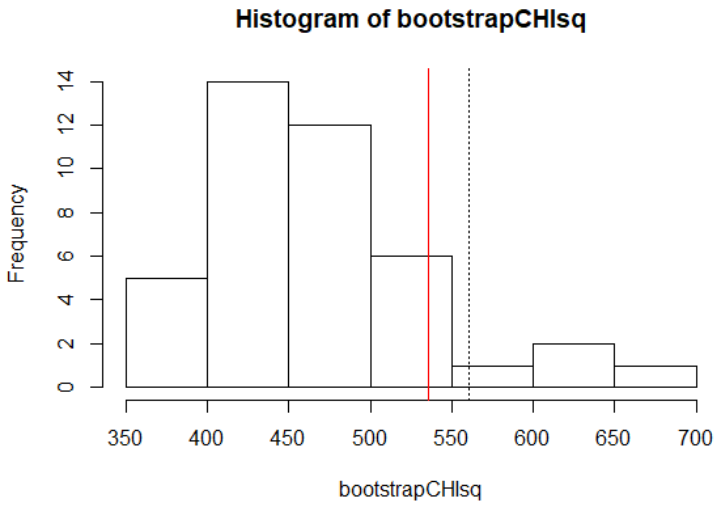

Figure 1: Histogram of the resampled χ^2 test statistics. red line: χ^2 value of the original LCA. black dotted line: cuts off the 5% highest χ^2 values of the resampled data.

Lastly, we combine the class membership and classification probability (section 6.4.1.6, line 57 – 58) into a *.cvs and calculate the hit-rate (section 6.4.1.7) and standard deviations of the hit rates (line 62 – 65). The lines following line 68 print the model summary into a text file. The text file and the *.csv document with the class membership and classification probabilities will be saved into the current working directory of R and are named according to the date and time when the analysis was conducted.

3: Model summary

```
56 # classification probability and assigned class membership
57 classMembership <- data.frame(LCA$posterior, class=LCA$predclass)
58 write.csv2(classMembership,paste0(format(Sys.time(),"%y%m%d-%H%M%S-"),nclasses,"-
       classMembership.csv", collapse=""))
59
60 # hit rate / mean classification probability
61 # we get the hit rate by calculating the mean classification probability per class
62 hitrate <- aggregate(classMembership, list(classMembership$class), FUN=mean)
63
64 # we can also calculate the SD of the hit rate
65 hitrateSD <- aggregate(classMembership, list(classMembership$class), FUN=sd)
66
67 # Printing a model summary into a text file named with model and current time
68 sink(paste0(format(Sys.time(),"%y%m%d-%H%M%S-"),nclasses,"-classSummaryLCA.txt", collapse=""))
69 ...
```

Interpreting the Output

When running the LCA, we will print the model summary into a text file. We are interested if the iteration was stopped within the convergence criterion before the max number of iterations (line 3 is smaller than `maxiter=1000`) in the log-likelihood (line 6), the number of parameters (line 5), and the information criteria (line 11 - 14). Line 15 presents the likelihood ratio statistic G^2, that we cannot use with sparse data (section 6.4.4.2). Line 16 shows the Person χ^2 test statistic (section 6.4.4.2), line 20 the empirical p-value of this statistic in the resampled distribution. The p-value is > 0.050, therefore the estimated model fits the data. Note that for repeated calculations, a lower `bootstrapN` will result in different p-values, since only a small part of the actual distribution is taken into account.

4: Interpreting the model summary – model description, information criteria and result of the bootstrap

```
1  Model summary for the 3 class model
2  ----------------------------------
3  number of iterations:  51
4  number of observations:  496
5  number of estimated parameters:  50
6  maximum log-likelihood:  -2257.13
7
8  ------------------------
9  Goodness of fit
10 ------------------------
11 AIC: 4614
12 BIC: 4825
13 CAIC: 4875
14 aBIC: 4666
15 G^2: 346.68 (Likelihood ratio/deviance statistic)
16 CHI^2: 535.46 (Chi-square goodness of fit)
17 CHI^2 emp. p-value: 0.146
```

The summary also shows the model parameters. The expected class prevalence π_c (Table 6) for each class starting line 22 and the conditional item-response probabilities π_{ixc} (Appendix Table 6.1) starting line 27. Be careful with label switching when running the LCA several times, it is helpful to re-name the classes in descending order of their expected prevalence. Note that because of how the MLE works, there might be a small, but within rounding reasonable, deviations in the estimated model parameters.

Furthermore, the output shows the hit rate and the standard deviation of the hit rate (starting line 36). The classification probability of every student in every class and the resulting assigned class membership (section 6.4.1.6) were exported to a *.csv.

5: Interpreting the output – model parameters and hit rate

```
19  Expected class prevalence for each class
20  ----------------------------------------
21  class 1 class 2 class 3
22    0.534   0.238   0.228
23  ----------------------------------------------------
24  Conditional item-response (column) probabilities,
25  by outcome variable, for each class (row)
26  ----------------------------------------------------
27 $ID1
28            Pr(1) Pr(2) Pr(3) Pr(4)
29 class 1:  0.893 0.092 0.000 0.015
30 class 2:  0.736 0.177 0.056 0.032
31 class 3:  0.535 0.193 0.109 0.162
32 ...
33 ------------------------
34  Hit Rate
35  ------------------------
36  Group.1    X1    X2    X3 class
37        1 0.847 0.078 0.075     1
38        2 0.051 0.815 0.134     2
39        3 0.094 0.082 0.824     3
```

6: Complete R script to run an LCA with poLCA

```
# install.packages("poLCA")
library(poLCA) # loading poLCA

setwd("")

file <- "CVSdata.csv" # name of the data file

nclasses <- 3 # number of classes
f1 <- as.formula(cbind(ID1, ID2, IN, UN)~1) # selecting the variables for the LCA

#####################################################################

mydata <- read.csv2(file) # reading the data file

# Running the analysis
# max number of iterations: 1000, convergence criteria: .001
# number of times to estimate the model, using different start values: 30
LCA <- poLCA(f1, data=mydata, nclass=nclasses, maxiter=1000, tol=.001, nrep=30)

#####################################################################
# start bootstrap

bootstrapN <- 40 # max number of bootstrap samples

bootstrapCHIsq <- LCA$Chisq

for (i in 1:bootstrapN){
  # simulating a data set based on the calculated model parameters
  simdata <- poLCA.simdata(N=nrow(mydata),probs=LCA$probs,P=LCA$P)
  # running the LCA for the resampled data set
  simLCA <- poLCA(
    as.formula(paste0("cbind(",paste0("Y",1:ncol(LCA$y),collapse=","),")~1")),
    data=simdata$dat, nclass=nclasses, maxiter=1000, tol=.001, nrep=1)
  # save the resampled CHI^2 value in a list
  bootstrapCHIsq <- c(bootstrapCHIsq, simLCA$Chisq)
}

# rank the original CHI^2 within the resampled distribution
bootstrapCHIsq <- sort(bootstrapCHIsq)
empP <- 1-(match(LCA$Chisq, bootstrapCHIsq)/(bootstrapN+1))


# creating a histogram of the resampled CHI^2 distribution
# red line with CHI^2 value of the original LCA
# black dotted line that cuts off the 5% highest CHI^2 values of the resampled data
# if the red line is left of the dotted line, than the original LCA model
# fits the data because there is no significant difference between
# the original model and most of the resampled models
hist(bootstrapCHIsq)
abline(v=LCA$Chisq, col="red")
abline(v=bootstrapCHIsq[floor((bootstrapN+1)-((bootstrapN+1)*0.05))],lty=3)

# End bootstrap
#####################################################################

# classification probability and assigned class membership
classMembership <- data.frame(LCA$posterior, class=LCA$predclass)
write.csv2(classMembership,paste0(format(Sys.time(),"%y%m%d-%H%M%S-"),nclasses,"-
    classMembership.csv", collapse=""))

# hit rate / mean classification probability
# we get the hit rate by calculating the mean classification probability per class
hitrate <- aggregate(classMembership, list(classMembership$class), FUN=mean)

# we can also calculate the SD of the hit rate
hitrateSD <- aggregate(classMembership, list(classMembership$class), FUN=sd)

# Printing a model summary into a text file named with model and current time
sink(paste0(format(Sys.time(),"%y%m%d-%H%M%S-"),nclasses,"-classSummaryLCA.txt", collapse=""))
```

```
cat(c("Date of analysis:", format(Sys.time(),"%x %X")))

cat(c(
  "\n \n", paste0(rep("-", times=36), collapse=""), "\n Model summary for the", nclasses, "
      class model \n ", paste0(rep("-", times=36), collapse=""),
  "\n number of iterations: ", LCA$numiter,
  "\n number of observations: ", LCA$Nobs,
  "\n number of estimated parameters: ", LCA$npar,
  "\n maximum log-likelihood: ", round(LCA$llik, digits=2),
  "\n \n", paste0(rep("-", times=23), collapse=""), "\n Goodness of fit \n", paste0(rep("-",
      times=23), collapse=""),
  "\n AIC:", round(LCA$aic, digits=0),
  "\n BIC:", round(LCA$bic, digits=0),
  "\n CAIC:", round(-2*LCA$llik+(log(LCA$Nobs)+1)*LCA$npar, digits=0),
  "\n aBIC:", round(-2*LCA$llik+log((LCA$Nobs +2)/24)*LCA$npar, digits=0),
  "\n G^2:", round(LCA$Gsq, digits=2), "(Likelihood ratio/deviance statistic)",
  "\n CHI^2:", round(LCA$Chisq, digits=2),"(Chi-square goodness of fit)",
  "\n CHI^2 emp. p-value:", if(exists("empP")) round(empP, digits=3) else "no bootstrap
      performed", "\n"
  ))

cat(c(
  "\n \n", paste0(rep("-", times=42), collapse=""), "\n Expected class prevalence for each
      class \n", paste0(rep("-", times=42), collapse=""), "\n"
))

class_prevalence <- data.frame(lapply(LCA$P, round, digits=3))
colnames(class_prevalence) <- c(paste("class", 1:nclasses))
print(class_prevalence, row.names=FALSE)

cat(c(
  "\n", paste0(rep("-", times=51), collapse=""), "\n Conditional item-response (column)
      probabilities,",
  "\n by outcome variable, for each class (row) \n", paste0(rep("-", times=51), collapse=""),
      "\n \n"
  ))

lapply(LCA$probs, round, digits=3)

cat(c(
  "\n", paste0(rep("-", times=23), collapse=""), "\n Hit Rate \n", paste0(rep("-", times=23),
      collapse=""), "\n"
))

print(round(hitrate, digits=3), row.names=FALSE)

cat(c(
  "\n \n", paste0(rep("-", times=23), collapse=""), "\n Hit Rate SD \n", paste0(rep("-", times
      =23), collapse=""), "\n"
))

print(round(hitrateSD, digits=3), row.names=FALSE)

sink()
```