LCA data modeling Seth-Josh

1. Loading, setting up

```r
library(tidyverse)
library(poLCA)

f <- "obs-segment_units1-7_2013-2014.csv"
d <- read_csv(f)

f1 <- "Observations_segment_Units_1-7_2012-13.csv"
d1 <-read_csv(f1)
```

2. Preparing data with a few teacher and student variables

None of the unit-specific variables included.

```r
add_one <- function(x) {
    x + 1
}

ds <- d %>%
    dplyr::select(sInvented, sProcedural, sConceptual, tInitSelect, tCompare, tDiscussQ, tConnectBigIdea
    map_df(replace_na, 0) %>%
    map_df(add_one)

ds1 <- d1 %>%
    dplyr::select(sInvented, sProcedural, sConceptual, tInitSelect, tCompare, tDiscussQ, tConnectBigIdea
    map_df(replace_na, 0) %>%
    map_df(add_one)

dd <- bind_rows(ds, ds1)
```

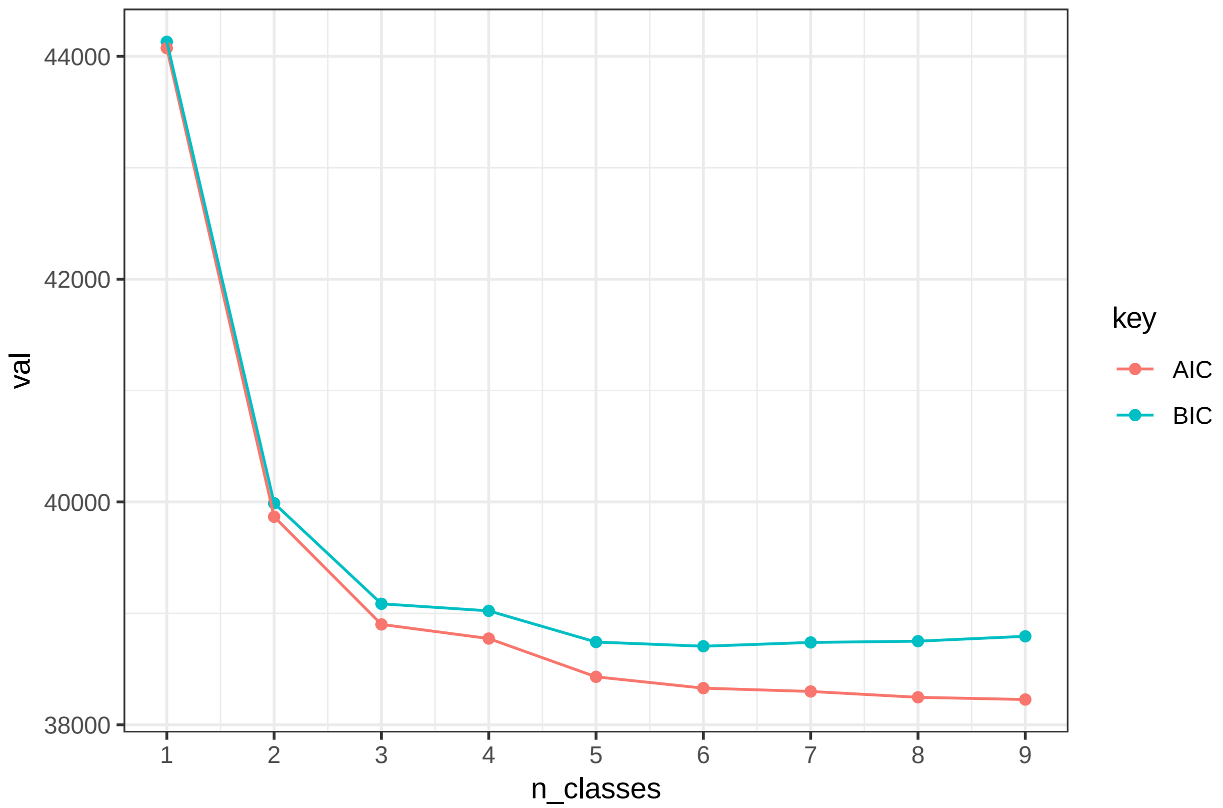
3. Choosing the number of classes/profiles

Using latent class analysis through the **poLCA** R package.

```r
f <- cbind(sInvented, sProcedural, sConceptual, tInitSelect, tCompare, tDiscussQ, tConnectBigIdeas, tCon

od <- map(1:9, poLCA, formula = f, data = dd, maxiter = 5000, verbose = FALSE, graphs = FALSE) %>%
    map_df(broom::glance)
```

```r
od %>%
    mutate(n_classes = 1:9) %>%
    gather(key, val, BIC, AIC) %>%
    ggplot(aes(x = n_classes, y = val, color = key, group = key)) +
    geom_point() +
    geom_line() +
    scale_x_continuous(breaks = 1:9, labels = 1:9) +
    theme_bw() +
    labs(caption = "Lower values of the AIC & BIC suggest preferred model(s); generally, BIC is more co
```



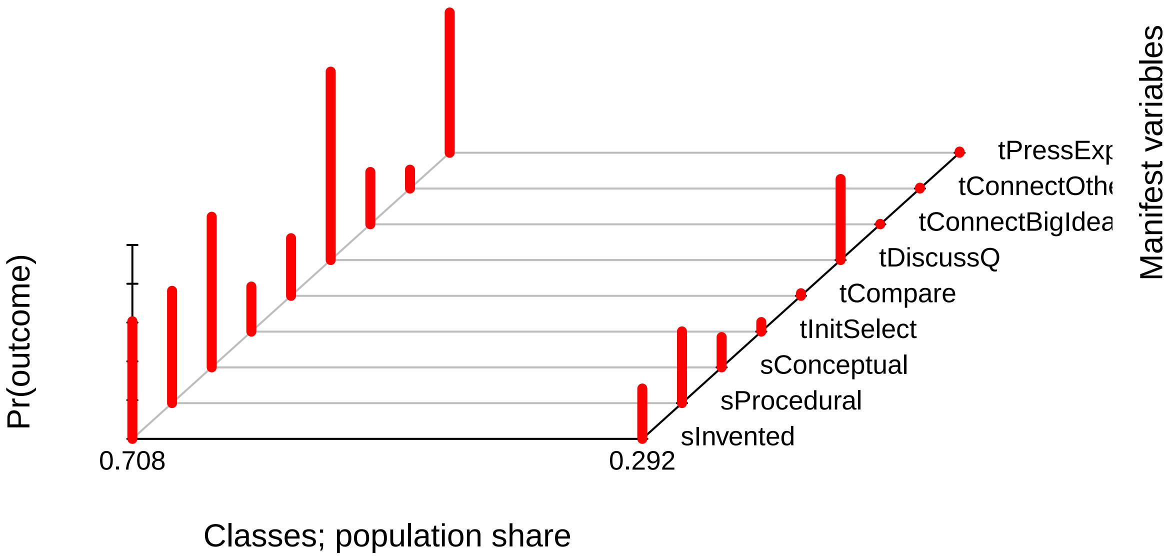
wer values of the AIC & BIC suggest preferred model(s); generally, BIC is more conservative than AIC

Based on this fit statistic–the Bayesian Information Criteria, which is just a transformation of the log-likelihood, and is usually recommended along with the AIC as one criterion for model selection–it looks like 3 and especially 4 or 5 class solutions seem reasonable.

4. Examining 2, 3, 4, and 5 class solutions

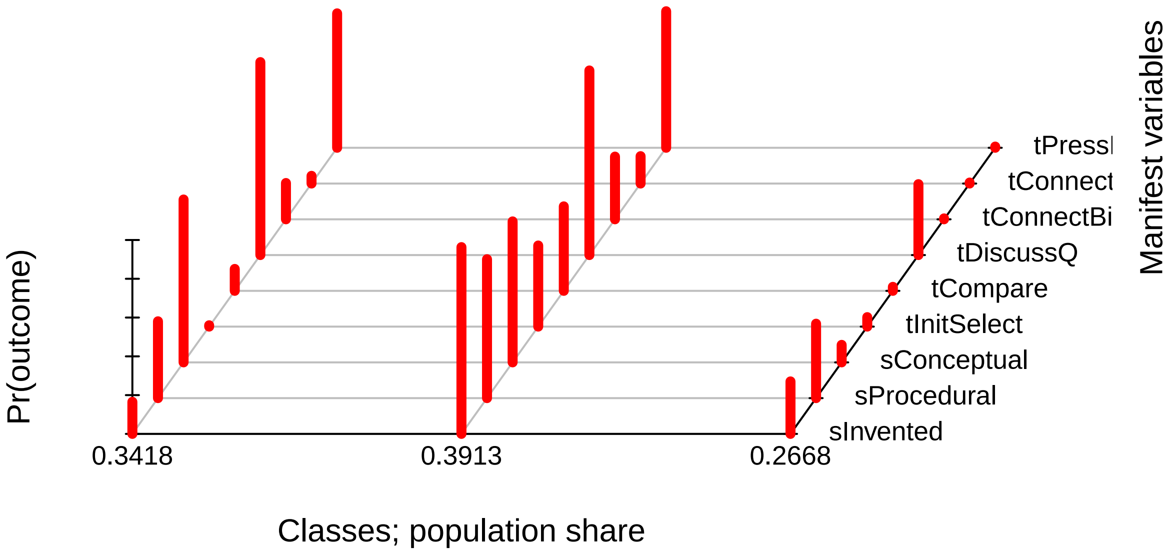
```
f <- cbind(sInvented, sProcedural, sConceptual, tInitSelect, tCompare, tDiscussQ, tConnectBigIdeas, tCo

m2 <- poLCA(f, ds, nclass = 2, maxiter = 5000, graphs = TRUE)
```



```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
## $sInvented
##           Pr(1)  Pr(2)
## class 1:  0.3927 0.6073
## class 2:  0.7400 0.2600
##
## $sProcedural
##           Pr(1)  Pr(2)
## class 1:  0.4206 0.5794
## class 2:  0.6299 0.3701
##
## $sConceptual
##           Pr(1)  Pr(2)
## class 1:  0.2230 0.7770
## class 2:  0.8436 0.1564
##
## $tInitSelect
##           Pr(1)  Pr(2)
## class 1:  0.7674 0.2326
## class 2:  0.9505 0.0495
##
## $tCompare
##           Pr(1)  Pr(2)
## class 1:  0.7024 0.2976
## class 2:  0.9863 0.0137
##
## $tDiscussQ
##            Pr(1)  Pr(2)
## class 1:  0.0278 0.9722
## class 2:  0.5815 0.4185
##
## $tConnectBigIdeas
##            Pr(1)  Pr(2)
## class 1:  0.7298 0.2702
## class 2:  0.9977 0.0023
##
## $tConnectOthers
##            Pr(1)  Pr(2)
## class 1:  0.9025 0.0975
## class 2:  0.9953 0.0047
##
## $tPressExplain
##            Pr(1)  Pr(2)
## class 1:  0.2762 0.7238
## class 2:  0.9935 0.0065
##
## Estimated class population shares
##  0.708 0.292
##
## Predicted class memberships (by modal posterior prob.)
##  0.7295 0.2705
##
## =========================================================
## Fit for 2 latent classes:
## =========================================================
## number of observations: 2813
## number of estimated parameters: 19
## residual degrees of freedom: 492
## maximum log-likelihood: -12828.56
##
## AIC(2): 25695.12
## BIC(2): 25808.02
## G^2(2): 1830.582 (Likelihood ratio/deviance statistic)
## X^2(2): 5480.487 (Chi-square goodness of fit)
##
```

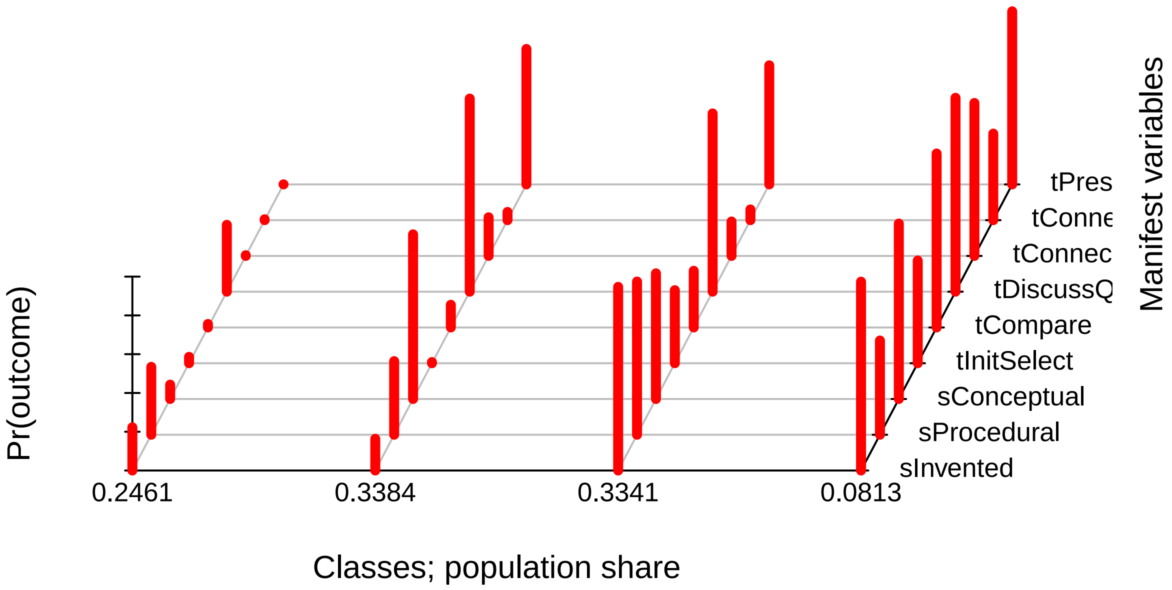
```r
m3 <- poLCA(f, ds, nclass = 3, maxiter = 5000, graphs = TRUE)
```



```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
## $sInvented
##            Pr(1)  Pr(2)
## class 1:  0.8345 0.1655
## class 2:  0.0366 0.9634
## class 3:  0.7292 0.2708
##
## $sProcedural
##            Pr(1)  Pr(2)
## class 1:  0.6037 0.3963
## class 2:  0.2835 0.7165
## class 3:  0.6160 0.3840
##
## $sConceptual
##            Pr(1)  Pr(2)
## class 1:  0.1600 0.8400
## class 2:  0.2729 0.7271
## class 3:  0.9097 0.0903
##
## $tInitSelect
##            Pr(1)  Pr(2)
## class 1:  0.9930 0.0070
## class 2:  0.5817 0.4183
## class 3:  0.9511 0.0489
```

```
##
## $tCompare
##            Pr(1)  Pr(2)
## class 1:  0.8868 0.1132
## class 2:  0.5643 0.4357
## class 3:  0.9793 0.0207
##
## $tDiscussQ
##            Pr(1)  Pr(2)
## class 1:  0.0046 0.9954
## class 2:  0.0478 0.9522
## class 3:  0.6341 0.3659
##
## $tConnectBigIdeas
##            Pr(1)  Pr(2)
## class 1:  0.8126 0.1874
## class 2:  0.6765 0.3235
## class 3:  0.9951 0.0049
##
## $tConnectOthers
##            Pr(1)  Pr(2)
## class 1:  0.9601 0.0399
## class 2:  0.8586 0.1414
## class 3:  0.9947 0.0053
##
## $tPressExplain
##            Pr(1)  Pr(2)
## class 1:  0.3065 0.6935
## class 2:  0.2956 0.7044
## class 3:  0.9937 0.0063
##
## Estimated class population shares
##  0.3418 0.3913 0.2668
##
## Predicted class memberships (by modal posterior prob.)
##  0.3228 0.4042 0.273
##
## =========================================================
## Fit for 3 latent classes:
## =========================================================
## number of observations: 2813
## number of estimated parameters: 29
## residual degrees of freedom: 482
## maximum log-likelihood: -12400.05
##
## AIC(3): 24858.1
## BIC(3): 25030.41
## G^2(3): 973.5587 (Likelihood ratio/deviance statistic)
## X^2(3): 2279.987 (Chi-square goodness of fit)
##
```

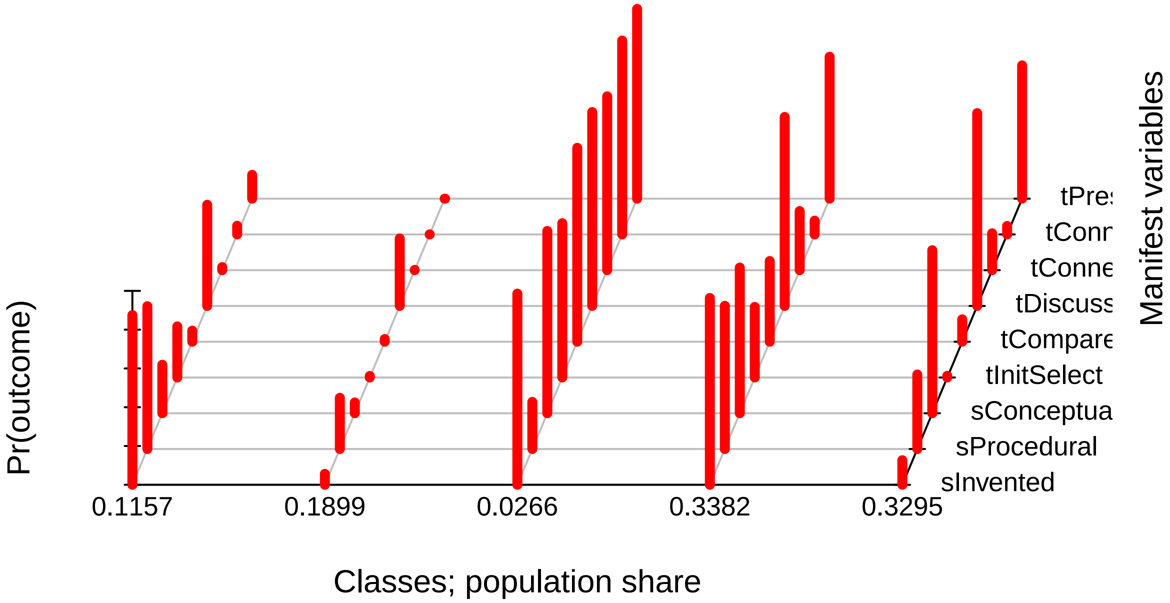
```
m4 <- poLCA(f, ds, nclass = 4, maxiter = 5000, graphs = TRUE)
```



```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
## $sInvented
##            Pr(1)  Pr(2)
## class 1:  0.7772 0.2228
## class 2:  0.8365 0.1635
## class 3:  0.0528 0.9472
## class 4:  0.0259 0.9741
##
## $sProcedural
##            Pr(1)  Pr(2)
## class 1:  0.6498 0.3502
## class 2:  0.6201 0.3799
## class 3:  0.2098 0.7902
## class 4:  0.5140 0.4860
##
## $sConceptual
##            Pr(1)  Pr(2)
## class 1:  0.9259 0.0741
## class 2:  0.1507 0.8493
## class 3:  0.3518 0.6482
## class 4:  0.0951 0.9049
```

```
##
## $tInitSelect
##             Pr(1)  Pr(2)
## class 1:  0.9670 0.0330
## class 2:  0.9938 0.0062
## class 3:  0.6232 0.3768
## class 4:  0.4704 0.5296
##
## $tCompare
##             Pr(1)  Pr(2)
## class 1:  0.9820 0.0180
## class 2:  0.8829 0.1171
## class 3:  0.7076 0.2924
## class 4:  0.1027 0.8973
##
## $tDiscussQ
##             Pr(1)  Pr(2)
## class 1:  0.6550 0.3450
## class 2:  0.0039 0.9961
## class 3:  0.0806 0.9194
## class 4:  0.0000 1.0000
##
## $tConnectBigIdeas
##             Pr(1)  Pr(2)
## class 1:  0.9961 0.0039
## class 2:  0.8006 0.1994
## class 3:  0.8224 0.1776
## class 4:  0.2109 0.7891
##
## $tConnectOthers
##             Pr(1)  Pr(2)
## class 1:  0.9952 0.0048
## class 2:  0.9575 0.0425
## class 3:  0.9446 0.0554
## class 4:  0.5536 0.4464
##
## $tPressExplain
##             Pr(1)  Pr(2)
## class 1:  0.9995 0.0005
## class 2:  0.3013 0.6987
## class 3:  0.3858 0.6142
## class 4:  0.1073 0.8927
##
## Estimated class population shares
##  0.2461 0.3384 0.3341 0.0813
##
## Predicted class memberships (by modal posterior prob.)
##  0.241 0.3374 0.3466 0.075
##
## =========================================================
## Fit for 4 latent classes:
## =========================================================
## number of observations: 2813
## number of estimated parameters: 39
## residual degrees of freedom: 472
## maximum log-likelihood: -12241.13
##
## AIC(4): 24560.26
## BIC(4): 24792
## G^2(4): 655.7218 (Likelihood ratio/deviance statistic)
## X^2(4): 959.0916 (Chi-square goodness of fit)
##
```

```
m5 <- poLCA(f, ds, nclass = 5, maxiter = 5000, graphs = TRUE)
```



```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
## $sInvented
##            Pr(1)  Pr(2)
## class 1:  0.1246 0.8754
## class 2:  0.9436 0.0564
## class 3:  0.0141 0.9859
## class 4:  0.0366 0.9634
## class 5:  0.8732 0.1268
##
## $sProcedural
##            Pr(1)  Pr(2)
## class 1:  0.2635 0.7365
## class 2:  0.7358 0.2642
## class 3:  0.7570 0.2430
## class 4:  0.2614 0.7386
## class 5:  0.6157 0.3843
##
## $sConceptual
##             Pr(1)  Pr(2)
## class 1:  0.7499 0.2501
## class 2:  0.9445 0.0555
## class 3:  0.0597 0.9403
## class 4:  0.2486 0.7514
## class 5:  0.1590 0.8410
##
## $tInitSelect
##             Pr(1)  Pr(2)
## class 1:  0.7365 0.2635
## class 2:  0.9920 0.0080
## class 3:  0.2043 0.7957
## class 4:  0.6360 0.3640
## class 5:  0.9913 0.0087
##
## $tCompare
##             Pr(1)  Pr(2)
## class 1:  0.9429 0.0571
## class 2:  0.9861 0.0139
## class 3:  0.0000 1.0000
## class 4:  0.5838 0.4162
## class 5:  0.8845 0.1155
##
## $tDiscussQ
##             Pr(1)  Pr(2)
## class 1:  0.4783 0.5217
## class 2:  0.6520 0.3480
## class 3:  0.0000 1.0000
## class 4:  0.0251 0.9749
## class 5:  0.0055 0.9945
##
## $tConnectBigIdeas
##             Pr(1)  Pr(2)
## class 1:  0.9839 0.0161
## class 2:  0.9984 0.0016
## class 3:  0.1034 0.8966
## class 4:  0.6948 0.3052
## class 5:  0.8097 0.1903
##
## $tConnectOthers
##             Pr(1)  Pr(2)
## class 1:  0.9550 0.0450
## class 2:  0.9993 0.0007
## class 3:  0.0000 1.0000
## class 4:  0.9290 0.0710
## class 5:  0.9563 0.0437
##
## $tPressExplain
##            Pr(1)  Pr(2)
## class 1:  0.8770 0.1230
## class 2:  1.0000 0.0000
## class 3:  0.0214 0.9786
## class 4:  0.2681 0.7319
## class 5:  0.3124 0.6876
##
## Estimated class population shares
##  0.1157 0.1899 0.0266 0.3382 0.3295
##
## Predicted class memberships (by modal posterior prob.)
##  0.1013 0.1927 0.0277 0.3726 0.3057
##
## =========================================================
## Fit for 5 latent classes:
## =========================================================
## number of observations: 2813
## number of estimated parameters: 49
## residual degrees of freedom: 462
## maximum log-likelihood: -12169.39
##
## AIC(5): 24436.79
## BIC(5): 24727.95
## G^2(5): 512.2494 (Likelihood ratio/deviance statistic)
## X^2(5): 761.12 (Chi-square goodness of fit)
##
```