2019-01-27 - LCA data modeling Seth-Josh

Getting data from Google Sheets

```
library(googlesheets)
g <- gs_title("Observations_segment_Units_1-7_2013-14-with-duplicates-identified")
d <- gs_read(g)

g1 <- gs_title("Observations_segment_Units_1-7_2012-13-with-duplicates-identified")
d1 <- gs_read(g1)
```

1. Loading, setting up

```
library(tidyverse)
library(poLCA)
library(readxl)

# d <- read_excel("Observations_segment_Units_1-7_2013-14_updated01192015.xlsx")
#c <- count(d, `Teacher::TeacherID`, `ClassObservation::Date`, `ClassObservation::Unit`, SegNum)
# d %>% left_join(c) %>% rename(number_of_segments = n) %>%  write_csv("Observations_segment_Units_1-7_

#f1 <- "Observations_segment_Units_1-7_2012-13.csv"
#d1 <-read_csv(f1)
#c <- count(d1, `Teacher::TeacherID`, `ClassObservation::Date`, `ClassObservation::Unit`, SegNum)
# d1 %>% left_join(c) %>% rename(number_of_segments = n) %>%  write_csv("Observations_segment_Units_1-7
```

```
d <- read_csv("Observations_segment_Units_1-7_2013-14-with-duplicates-identified-fin.csv")
d1 <- read_csv("Observations_segment_Units_1-7_2012-13-with-duplicates-identified-fin.csv")
```

2. Preparing data with a few teacher and student variables

Checking for dups

```
d %>% count(`Teacher::TeacherID`, `ClassObservation::Date`, SegNum) %>% count(n)
```

```
## # A tibble: 3 x 2
##       n    nn
##   <int> <int>
## 1     1  1799
## 2     2   504
## 3     3     2
```

```
d1 %>% count(`Teacher::TeacherID`, `ClassObservation::Date`, SegNum) %>% count(n)
```

```
## # A tibble: 2 x 2
##       n    nn
##   <int> <int>
## 1     1  1334
## 2     2    96
```

```
d %>% count(`Duplicate Condition`)
```

```
## # A tibble: 3 x 2
##   `Duplicate Condition`     n
##   <chr>                 <int>
## 1 d                        97
## 2 D                       383
## 3 <NA>                   2333
```

```
d1 %>% count(`Duplicate Condition`)
```

```
## # A tibble: 2 x 2
##   `Duplicate Condition`     n
##   <chr>                 <int>
## 1 d                        96
## 2 <NA>                   1430
```

For `d`, we should remove 506 segments - but we remove 480 (should we identify 26 more?) For `d1`, we should remove 96.

For a totle of 602 to remove.

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

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

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

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

```
dds <- filter(dd, `Duplicate Condition` != "D" & `Duplicate Condition` != "d")

nrow(dd) - nrow(dds)
```

```
## [1] 576
```

Need to remove 26 more.

3. Choosing the number of classes/profiles

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

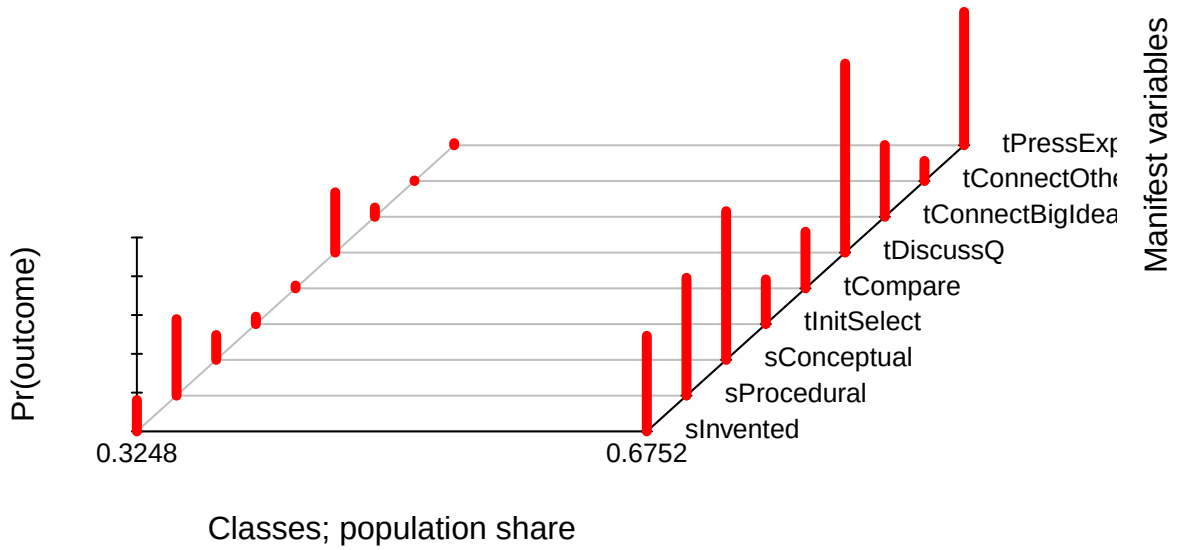
```
f <- cbind(sInvented, sProcedural, sConceptual, tInitSelect, tCompare, tDiscussQ, tConnectBigIdeas, tCo

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

```
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
```

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 or 5 class solutions are preferred.

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

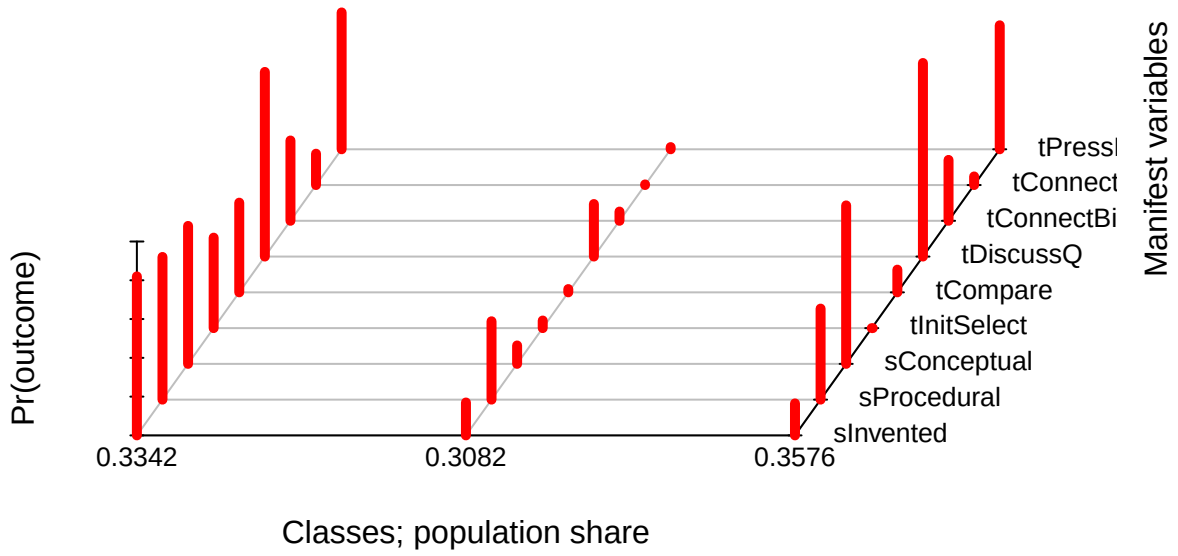
```
f <- cbind(sInvented, sProcedural, sConceptual, tInitSelect, tCompare, tDiscussQ, tConnectBigIdeas, tCo

m2 <- poLCA(f, dd, 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:  0.8380 0.1620
## class 2:  0.5072 0.4928
##
## $sProcedural
##            Pr(1)  Pr(2)
## class 1:  0.6055 0.3945
## class 2:  0.3924 0.6076
##
## $sConceptual
##            Pr(1)  Pr(2)
## class 1:  0.8719 0.1281
## class 2:  0.2332 0.7668
##
## $tInitSelect
##            Pr(1)  Pr(2)
## class 1:  0.9616 0.0384
## class 2:  0.7696 0.2304
##
## $tCompare
##            Pr(1)  Pr(2)
## class 1:  0.9855 0.0145
```

```
## class 2:  0.7083 0.2917
##
## $tDiscussQ
##            Pr(1)  Pr(2)
## class 1:  0.6895 0.3105
## class 2:  0.0253 0.9747
##
## $tConnectBigIdeas
##            Pr(1)  Pr(2)
## class 1:  0.9522 0.0478
## class 2:  0.6298 0.3702
##
## $tConnectOthers
##            Pr(1)  Pr(2)
## class 1:  0.9971 0.0029
## class 2:  0.8963 0.1037
##
## $tPressExplain
##            Pr(1)  Pr(2)
## class 1:  0.9889 0.0111
## class 2:  0.3121 0.6879
##
## Estimated class population shares
##  0.3248 0.6752
##
## Predicted class memberships (by modal posterior prob.)
##  0.3155 0.6845
##
## =========================================================
## Fit for 2 latent classes:
## =========================================================
## number of observations: 4339
## number of estimated parameters: 19
## residual degrees of freedom: 492
## maximum log-likelihood: -19914.63
##
## AIC(2): 39867.26
## BIC(2): 39988.39
## G^2(2): 2261.49 (Likelihood ratio/deviance statistic)
## X^2(2): 5442.154 (Chi-square goodness of fit)
##
```

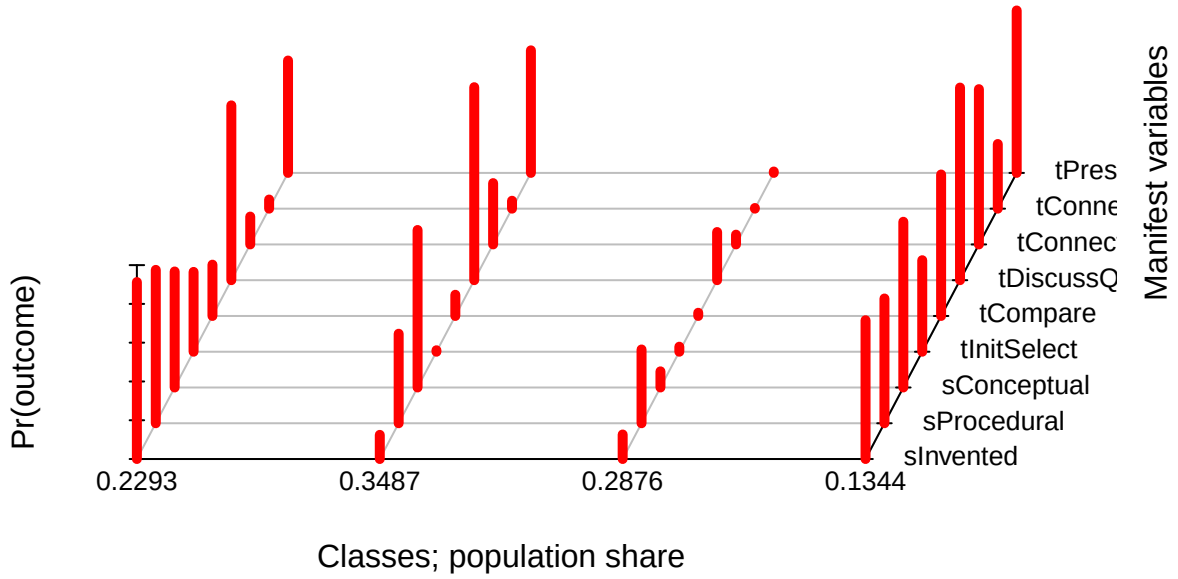
```
m3 <- poLCA(f, dd, nclass = 3, maxiter = 5000, graphs = TRUE)
```




```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
## $sInvented
##            Pr(1)  Pr(2)
## class 1:  0.1800 0.8200
## class 2:  0.8303 0.1697
## class 3:  0.8349 0.1651
##
## $sProcedural
##            Pr(1)  Pr(2)
## class 1:  0.2641 0.7359
## class 2:  0.5959 0.4041
## class 3:  0.5305 0.4695
##
## $sConceptual
##            Pr(1)  Pr(2)
## class 1:  0.2884 0.7116
## class 2:  0.9050 0.0950
## class 3:  0.1827 0.8173
##
## $tInitSelect
##            Pr(1)  Pr(2)
## class 1:  0.5334 0.4666
## class 2:  0.9608 0.0392
## class 3:  1.0000 0.0000
```

```
##
## $tCompare
##            Pr(1)  Pr(2)
## class 1:  0.5374 0.4626
## class 2:  0.9830 0.0170
## class 3:  0.8829 0.1171
##
## $tDiscussQ
##            Pr(1)  Pr(2)
## class 1:  0.0475 0.9525
## class 2:  0.7289 0.2711
## class 3:  0.0015 0.9985
##
## $tConnectBigIdeas
##            Pr(1)  Pr(2)
## class 1:  0.5857 0.4143
## class 2:  0.9523 0.0477
## class 3:  0.6859 0.3141
##
## $tConnectOthers
##            Pr(1)  Pr(2)
## class 1:  0.8382 0.1618
## class 2:  0.9967 0.0033
## class 3:  0.9556 0.0444
##
## $tPressExplain
##            Pr(1)  Pr(2)
## class 1:  0.2941 0.7059
## class 2:  0.9886 0.0114
## class 3:  0.3605 0.6395
##
## Estimated class population shares
##  0.3342 0.3082 0.3576
##
## Predicted class memberships (by modal posterior prob.)
##  0.3252 0.3121 0.3628
##
## =========================================================
## Fit for 3 latent classes:
## =========================================================
## number of observations: 4339
## number of estimated parameters: 29
## residual degrees of freedom: 482
## maximum log-likelihood: -19421.47
##
## AIC(3): 38900.94
## BIC(3): 39085.83
## G^2(3): 1275.171 (Likelihood ratio/deviance statistic)
## X^2(3): 2279.245 (Chi-square goodness of fit)
##
```

```
m4 <- poLCA(f, dd, nclass = 4, maxiter = 10000, graphs = TRUE)
```

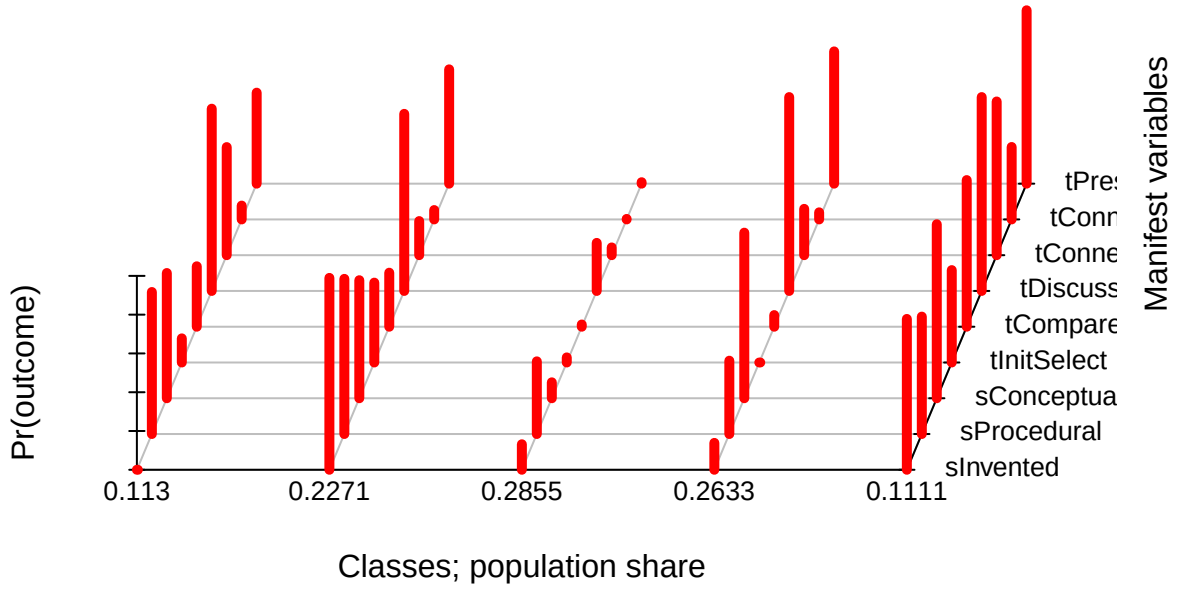



```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
## $sInvented
##            Pr(1)  Pr(2)
## class 1:  0.0872 0.9128
## class 2:  0.8753 0.1247
## class 3:  0.8735 0.1265
## class 4:  0.2841 0.7159
##
## $sProcedural
##            Pr(1)  Pr(2)
## class 1:  0.2088 0.7912
## class 2:  0.5382 0.4618
## class 3:  0.6194 0.3806
## class 4:  0.3565 0.6435
##
## $sConceptual
##            Pr(1)  Pr(2)
## class 1:  0.4016 0.5984
## class 2:  0.1876 0.8124
## class 3:  0.9167 0.0833
## class 4:  0.1452 0.8548
```

```
##
## $tInitSelect
##            Pr(1)  Pr(2)
## class 1:  0.5881 0.4119
## class 2:  0.9924 0.0076
## class 3:  0.9739 0.0261
## class 4:  0.5280 0.4720
##
## $tCompare
##            Pr(1)  Pr(2)
## class 1:  0.7353 0.2647
## class 2:  0.8906 0.1094
## class 3:  0.9834 0.0166
## class 4:  0.2703 0.7297
##
## $tDiscussQ
##            Pr(1)  Pr(2)
## class 1:  0.0982 0.9018
## class 2:  0.0045 0.9955
## class 3:  0.7515 0.2485
## class 4:  0.0064 0.9936
##
## $tConnectBigIdeas
##            Pr(1)  Pr(2)
## class 1:  0.8562 0.1438
## class 2:  0.6836 0.3164
## class 3:  0.9494 0.0506
## class 4:  0.1990 0.8010
##
## $tConnectOthers
##            Pr(1)  Pr(2)
## class 1:  0.9527 0.0473
## class 2:  0.9590 0.0410
## class 3:  0.9964 0.0036
## class 4:  0.6663 0.3337
##
## $tPressExplain
##            Pr(1)  Pr(2)
## class 1:  0.4203 0.5797
## class 2:  0.3678 0.6322
## class 3:  0.9926 0.0074
## class 4:  0.1622 0.8378
##
## Estimated class population shares
##  0.2293 0.3487 0.2876 0.1344
##
## Predicted class memberships (by modal posterior prob.)
##  0.2312 0.3517 0.2899 0.1272
##
## =========================================================
## Fit for 4 latent classes:
## =========================================================
## number of observations: 4339
## number of estimated parameters: 39
```

```
## residual degrees of freedom: 472
## maximum log-likelihood: -19221.76
##
## AIC(4): 38521.52
## BIC(4): 38770.16
## G^2(4): 875.7458 (Likelihood ratio/deviance statistic)
## X^2(4): 1840.229 (Chi-square goodness of fit)
##
```

```
m5 <- poLCA(f, dd, nclass = 5, maxiter = 10000, graphs = TRUE)
```




```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
## $sInvented
##            Pr(1)  Pr(2)
## class 1:  1.0000 0.0000
## class 2:  0.0103 0.9897
## class 3:  0.8683 0.1317
## class 4:  0.8612 0.1388
## class 5:  0.2220 0.7780
##
## $sProcedural
##            Pr(1)  Pr(2)
## class 1:  0.2675 0.7325
```

```
## class 2:  0.1998 0.8002
## class 3:  0.6252 0.3748
## class 4:  0.6218 0.3782
## class 5:  0.3945 0.6055
##
## $sConceptual
##            Pr(1)  Pr(2)
## class 1:  0.3540 0.6460
## class 2:  0.3911 0.6089
## class 3:  0.9178 0.0822
## class 4:  0.1459 0.8541
## class 5:  0.1026 0.8974
##
## $tInitSelect
##            Pr(1)  Pr(2)
## class 1:  0.8756 0.1244
## class 2:  0.5882 0.4118
## class 3:  0.9738 0.0262
## class 4:  1.0000 0.0000
## class 5:  0.5235 0.4765
##
## $tCompare
##            Pr(1)  Pr(2)
## class 1:  0.6882 0.3118
## class 2:  0.7213 0.2787
## class 3:  0.9885 0.0115
## class 4:  0.9400 0.0600
## class 5:  0.2437 0.7563
##
## $tDiscussQ
##            Pr(1)  Pr(2)
## class 1:  0.0599 0.9401
## class 2:  0.0865 0.9135
## class 3:  0.7519 0.2481
## class 4:  0.0000 1.0000
## class 5:  0.0000 1.0000
##
## $tConnectBigIdeas
##            Pr(1)  Pr(2)
## class 1:  0.4424 0.5576
## class 2:  0.8239 0.1761
## class 3:  0.9599 0.0401
## class 4:  0.7609 0.2391
## class 5:  0.2072 0.7928
##
## $tConnectOthers
##            Pr(1)  Pr(2)
## class 1:  0.9286 0.0714
## class 2:  0.9516 0.0484
## class 3:  0.9973 0.0027
## class 4:  0.9634 0.0366
## class 5:  0.6264 0.3736
##
## $tPressExplain
```

```
##            Pr(1)  Pr(2)
## class 1:  0.5306 0.4694
## class 2:  0.4103 0.5897
## class 3:  0.9925 0.0075
## class 4:  0.3178 0.6822
## class 5:  0.1058 0.8942
##
## Estimated class population shares
##  0.113 0.2271 0.2855 0.2633 0.1111
##
## Predicted class memberships (by modal posterior prob.)
##  0.089 0.2489 0.2885 0.2851 0.0885
##
## =========================================================
## Fit for 5 latent classes:
## =========================================================
## number of observations: 4339
## number of estimated parameters: 49
## residual degrees of freedom: 462
## maximum log-likelihood: -19170.92
##
## AIC(5): 38439.84
## BIC(5): 38752.23
## G^2(5): 774.0631 (Likelihood ratio/deviance statistic)
## X^2(5): 1403.946 (Chi-square goodness of fit)
##
```