LCA data modeling Seth-Josh

1. Loading, setting up

```
library(tidyverse)
library(poLCA)

f <- "obs-segment_units1-7_2013-2014.csv"

d <- read_csv(f)
```

2. Preparing data with a few teacher and student variables

None of the unit-specific variables included.

```
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)
```

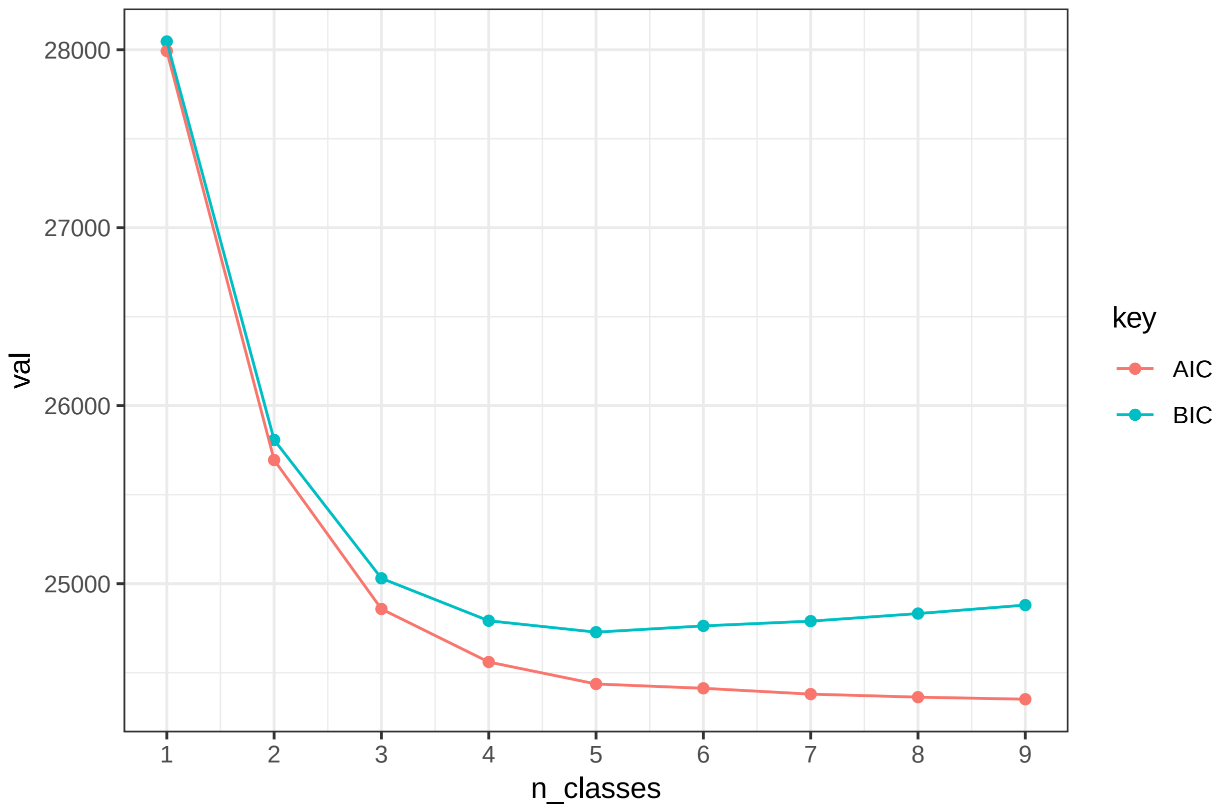
3. Choosing the number of classes/profiles

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

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

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

saveRDS(od, "2013-2014-dat-compare.rds")
```

```
od <- read_rds("2013-2014-dat-compare.rds")
```

```
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 con
```

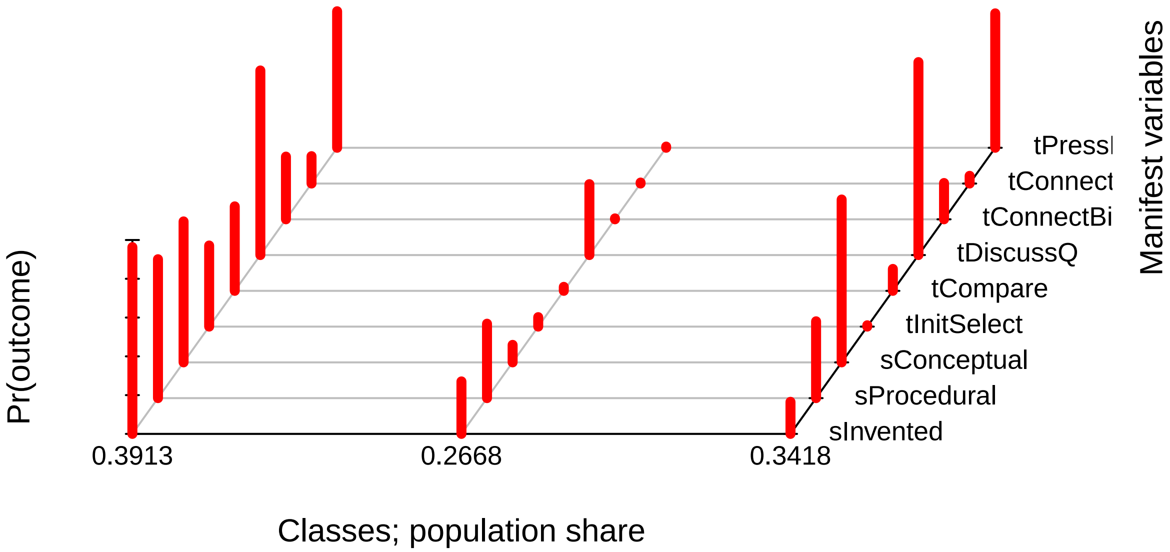


ver values of the AIC & BIC suggest preferred model(s); generally, BIC is more conservative than AIC

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 3, 4, and 5 class solutions

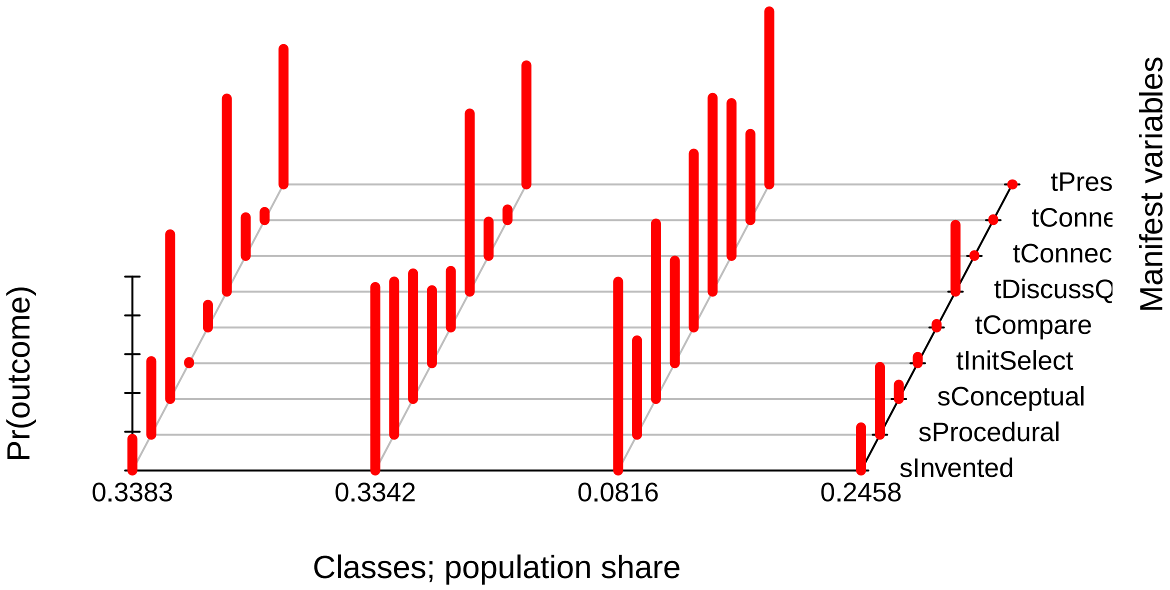
```
f <- cbind(sInvented, sProcedural, sConceptual, tInitSelect, tCompare, tDiscussQ, tConnectBigIdeas, tCo

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:  0.0366 0.9634
## class 2:  0.7292 0.2708
## class 3:  0.8345 0.1655
##
## $sProcedural
##            Pr(1)  Pr(2)
## class 1:  0.2835 0.7165
## class 2:  0.6160 0.3840
## class 3:  0.6037 0.3963
##
## $sConceptual
##            Pr(1)  Pr(2)
## class 1:  0.2729 0.7271
## class 2:  0.9097 0.0903
## class 3:  0.1600 0.8400
##
## $tInitSelect
##            Pr(1)  Pr(2)
## class 1:  0.5817 0.4183
## class 2:  0.9511 0.0489
## class 3:  0.9930 0.0070
```

```
##
## $tCompare
##            Pr(1)  Pr(2)
## class 1:  0.5643 0.4357
## class 2:  0.9793 0.0207
## class 3:  0.8868 0.1132
##
## $tDiscussQ
##            Pr(1)  Pr(2)
## class 1:  0.0478 0.9522
## class 2:  0.6341 0.3659
## class 3:  0.0046 0.9954
##
## $tConnectBigIdeas
##            Pr(1)  Pr(2)
## class 1:  0.6765 0.3235
## class 2:  0.9951 0.0049
## class 3:  0.8126 0.1874
##
## $tConnectOthers
##            Pr(1)  Pr(2)
## class 1:  0.8586 0.1414
## class 2:  0.9947 0.0053
## class 3:  0.9601 0.0399
##
## $tPressExplain
##            Pr(1)  Pr(2)
## class 1:  0.2956 0.7044
## class 2:  0.9937 0.0063
## class 3:  0.3065 0.6935
##
## Estimated class population shares
##  0.3913 0.2668 0.3418
##
## Predicted class memberships (by modal posterior prob.)
##  0.4042 0.273 0.3228
##
## =========================================================
## 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)
##
```

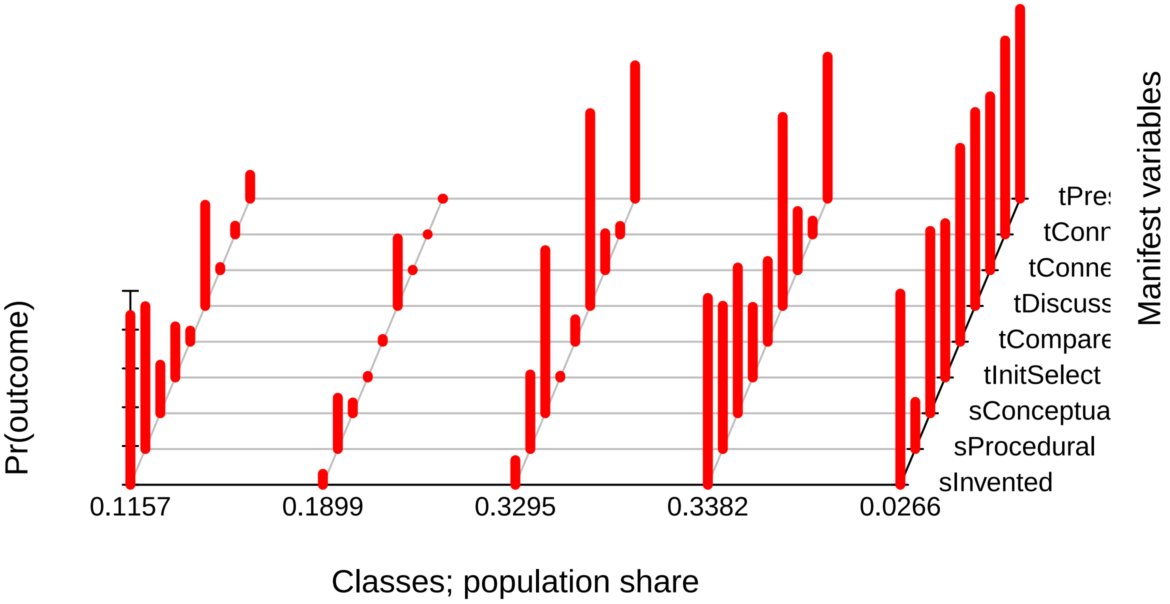
```
m4 <- poLCA(f, ds, nclass = 4, maxiter = 5000, graphs = TRUE)
```



```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
## 
## $sInvented
##            Pr(1)  Pr(2)
## class 1:  0.8366 0.1634
## class 2:  0.0535 0.9465
## class 3:  0.0260 0.9740
## class 4:  0.7775 0.2225
## 
## $sProcedural
##            Pr(1)  Pr(2)
## class 1:  0.6202 0.3798
## class 2:  0.2098 0.7902
## class 3:  0.5133 0.4867
## class 4:  0.6502 0.3498
## 
## $sConceptual
##            Pr(1)  Pr(2)
## class 1:  0.1508 0.8492
## class 2:  0.3523 0.6477
## class 3:  0.0952 0.9048
## class 4:  0.9260 0.0740
```

```
##
## $tInitSelect
##            Pr(1)  Pr(2)
## class 1:  0.9940 0.0060
## class 2:  0.6234 0.3766
## class 3:  0.4710 0.5290
## class 4:  0.9672 0.0328
##
## $tCompare
##            Pr(1)  Pr(2)
## class 1:  0.8830 0.1170
## class 2:  0.7082 0.2918
## class 3:  0.1035 0.8965
## class 4:  0.9821 0.0179
##
## $tDiscussQ
##            Pr(1)  Pr(2)
## class 1:  0.0040 0.9960
## class 2:  0.0810 0.9190
## class 3:  0.0000 1.0000
## class 4:  0.6552 0.3448
##
## $tConnectBigIdeas
##            Pr(1)  Pr(2)
## class 1:  0.8006 0.1994
## class 2:  0.8229 0.1771
## class 3:  0.2116 0.7884
## class 4:  0.9961 0.0039
##
## $tConnectOthers
##            Pr(1)  Pr(2)
## class 1:  0.9575 0.0425
## class 2:  0.9447 0.0553
## class 3:  0.5548 0.4452
## class 4:  0.9952 0.0048
##
## $tPressExplain
##            Pr(1)  Pr(2)
## class 1:  0.3013 0.6987
## class 2:  0.3861 0.6139
## class 3:  0.1076 0.8924
## class 4:  1.0000 0.0000
##
## Estimated class population shares
##  0.3383 0.3342 0.0816 0.2458
##
## Predicted class memberships (by modal posterior prob.)
##  0.3374 0.3466 0.075 0.241
##
## =========================================================
## Fit for 4 latent classes:
## =========================================================
## number of observations: 2813
## number of estimated parameters: 39
## residual degrees of freedom: 472
## maximum log-likelihood: -12241.14
##
## AIC(4): 24560.28
## BIC(4): 24792.02
## G^2(4): 655.7447 (Likelihood ratio/deviance statistic)
## X^2(4): 958.0743 (Chi-square goodness of fit)
##
```

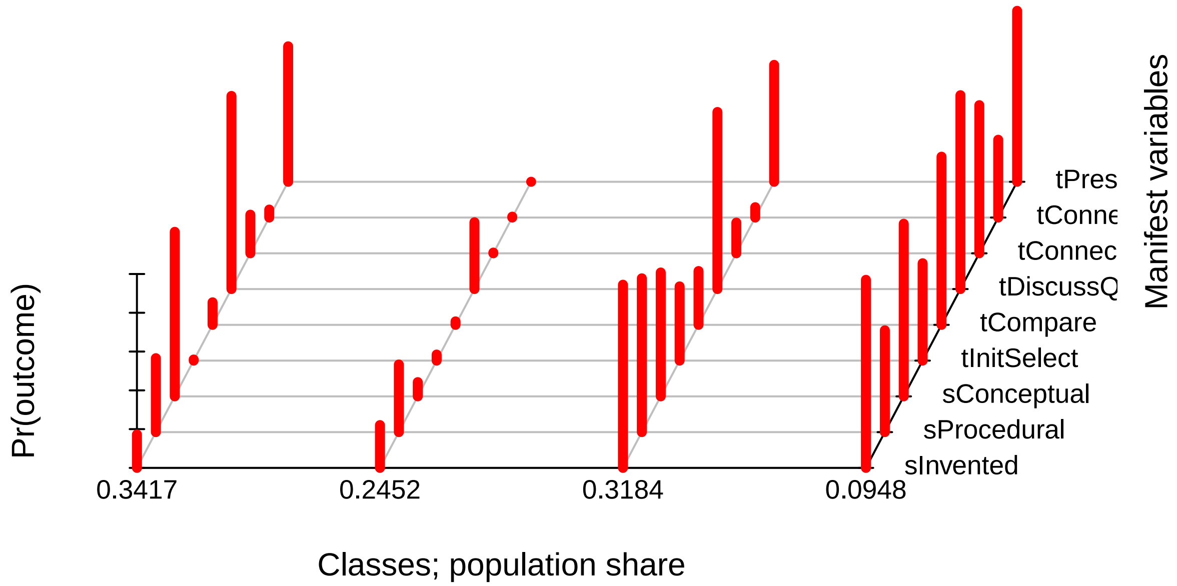
```
m5 <- poLCA(f, ds, nclass = 5, maxiter = 5000, graphs = TRUE)
```



```
## 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:  0.8732 0.1268
## class 4:  0.0366 0.9634
## class 5:  0.0141 0.9859
##
## $sProcedural
##            Pr(1)  Pr(2)
## class 1:  0.2635 0.7365
## class 2:  0.7358 0.2642
## class 3:  0.6157 0.3843
## class 4:  0.2614 0.7386
## class 5:  0.7570 0.2430
##
## $sConceptual
##            Pr(1)  Pr(2)
## class 1:  0.7499 0.2501
## class 2:  0.9445 0.0555
## class 3:  0.1590 0.8410
## class 4:  0.2486 0.7514
## class 5:  0.0597 0.9403
##
## $tInitSelect
##            Pr(1)  Pr(2)
## class 1:  0.7365 0.2635
## class 2:  0.9920 0.0080
## class 3:  0.9913 0.0087
## class 4:  0.6360 0.3640
## class 5:  0.2043 0.7957
##
## $tCompare
##            Pr(1)  Pr(2)
## class 1:  0.9429 0.0571
## class 2:  0.9861 0.0139
## class 3:  0.8845 0.1155
## class 4:  0.5838 0.4162
## class 5:  0.0000 1.0000
##
## $tDiscussQ
##            Pr(1)  Pr(2)
## class 1:  0.4783 0.5217
## class 2:  0.6520 0.3480
## class 3:  0.0055 0.9945
## class 4:  0.0251 0.9749
## class 5:  0.0000 1.0000
##
## $tConnectBigIdeas
##            Pr(1)  Pr(2)
## class 1:  0.9839 0.0161
## class 2:  0.9984 0.0016
## class 3:  0.8097 0.1903
## class 4:  0.6948 0.3052
## class 5:  0.1034 0.8966
##
## $tConnectOthers
##            Pr(1)  Pr(2)
## class 1:  0.9550 0.0450
## class 2:  0.9993 0.0007
## class 3:  0.9563 0.0437
## class 4:  0.9290 0.0710
## class 5:  0.0000 1.0000
##
## $tPressExplain
##             Pr(1)  Pr(2)
## class 1:  0.8770 0.1230
## class 2:  1.0000 0.0000
## class 3:  0.3124 0.6876
## class 4:  0.2681 0.7319
## class 5:  0.0214 0.9786
##
## Estimated class population shares
##  0.1157 0.1899 0.3295 0.3382 0.0266
##
## Predicted class memberships (by modal posterior prob.)
##  0.1013 0.1927 0.3057 0.3726 0.0277
##
## =========================================================
## 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): 761.1198 (Chi-square goodness of fit)
##
```

5. Examining the effects of a covariate (fGroups)

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

m4cov <- poLCA(fcov, 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:  0.8274 0.1726
## class 2:  0.7792 0.2208
## class 3:  0.0552 0.9448
## class 4:  0.0300 0.9700
##
## $sProcedural
##            Pr(1)  Pr(2)
## class 1:  0.6184 0.3816
## class 2:  0.6510 0.3490
## class 3:  0.2066 0.7934
## class 4:  0.4747 0.5253
##
## $sConceptual
##            Pr(1)  Pr(2)
## class 1:  0.1513 0.8487
## class 2:  0.9265 0.0735
## class 3:  0.3611 0.6389
## class 4:  0.1096 0.8904
##
## $tInitSelect
##            Pr(1)  Pr(2)
## class 1:  0.9942 0.0058
## class 2:  0.9686 0.0314
## class 3:  0.6171 0.3829
## class 4:  0.4978 0.5022
##
## $tCompare
##            Pr(1)  Pr(2)
## class 1:  0.8837 0.1163
## class 2:  0.9821 0.0179
## class 3:  0.7224 0.2776
## class 4:  0.1327 0.8673
##
## $tDiscussQ
##            Pr(1)  Pr(2)
## class 1:  0.0038 0.9962
## class 2:  0.6552 0.3448
## class 3:  0.0865 0.9135
## class 4:  0.0000 1.0000
##
## $tConnectBigIdeas
##            Pr(1)  Pr(2)
## class 1:  0.8007 0.1993
## class 2:  0.9961 0.0039
## class 3:  0.8415 0.1585
## class 4:  0.2358 0.7642
##
## $tConnectOthers
##            Pr(1)  Pr(2)
## class 1:  0.9587 0.0413
## class 2:  0.9953 0.0047
## class 3:  0.9464 0.0536
## class 4:  0.5986 0.4014
##
## $tPressExplain
##            Pr(1)  Pr(2)
## class 1:  0.3012 0.6988
## class 2:  0.9997 0.0003
## class 3:  0.3970 0.6030
## class 4:  0.1182 0.8818
##
## Estimated class population shares
##  0.3417 0.2452 0.3184 0.0948
##
## Predicted class memberships (by modal posterior prob.)
##  0.3359 0.241 0.3381 0.085
##
## =========================================================
## Fit for 4 latent classes:
## =========================================================
## 2 / 1
##             Coefficient  Std. error  t value  Pr(>|t|)
## (Intercept)    -0.27228     0.19090   -1.426     0.154
## fGroups        -0.04025     0.11801   -0.341     0.733
## =========================================================
## 3 / 1
##             Coefficient  Std. error  t value  Pr(>|t|)
## (Intercept)     0.19533     0.20089    0.972     0.331
## fGroups        -0.18191     0.11812   -1.540     0.124
## =========================================================
## 4 / 1
##             Coefficient  Std. error  t value  Pr(>|t|)
## (Intercept)    -0.04788     0.27026   -0.177     0.859
## fGroups        -0.89593     0.19105   -4.690     0.000
## =========================================================
## number of observations: 2813
## number of estimated parameters: 42
## residual degrees of freedom: 469
## maximum log-likelihood: -12227.46
##
## AIC(4): 24538.93
## BIC(4): 24788.49
## X^2(4): 959.7363 (Chi-square goodness of fit)
##
## ALERT: estimation algorithm automatically restarted with new initial values
##
```

- What covariates do we want to add?

6. Plot of segments over time for the four class solution

```
d$m4_class <- m4$predclass

t <- d %>%
    dplyr::select(seg_num = SegNum, unit = `ClassObservation::Unit`, teacher = `Teacher::TeacherID`) %>
    arrange(teacher, unit, seg_num) %>%
    group_by(unit) %>%
    summarize(max_seg_num = max(seg_num),
              max_unit = max(unit))

the_seqqer <- function(x) {
    seq(1, t$max_seg_num[x])
}

l <- list()
for (i in seq(t$max_seg_num)) {
    l[[i]] <- seq(1, t$max_seg_num[i])
}

t$seq_l <- l

dtm <- dplyr::select(d, `Teacher::TeacherID`, `Teacher::Condition`) %>% distinct() %>% arrange(`Teacher

d$m4_class <- as.factor(d$m4_class)

levels = c("Discussing Ideas", "Inventing & Connecting", "Inventing & Discussing", "Low Activity")
```

```
d$m4_class_name <- forcats::fct_recode(d$m4_class,
                                       `Discussing Ideas` = "1",
                                       `Inventing & Connecting` = "2",
                                       `Inventing & Discussing` = "3",
                                       `Low Activity` = "4")
```

```
teacher_ID = dtm$`Teacher::TeacherID`
condition = dtm$`Teacher::Condition`
map2(teacher_ID, condition, f, d)
```

```
## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
##
## [[5]]
## NULL
##
## [[6]]
## NULL
##
## [[7]]
## NULL
##
## [[8]]
## NULL
##
## [[9]]
## NULL
##
## [[10]]
## NULL
##
## [[11]]
## NULL
##
## [[12]]
## NULL
##
## [[13]]
## NULL
##
## [[14]]
## NULL
##
## [[15]]
## NULL
##
## [[16]]
## NULL
##
## [[17]]
## NULL
##
## [[18]]
## NULL
##
## [[19]]
## NULL
##
## [[20]]
## NULL
##
## [[21]]
## NULL
##
## [[22]]
## NULL
##
## [[23]]
## NULL
##
## [[24]]
## NULL
##
## [[25]]
## NULL
##
## [[26]]
## NULL
##
## [[27]]
## NULL
##
## [[28]]
## NULL
##
## [[29]]
## NULL
##
## [[30]]
## NULL
##
## [[31]]
## NULL
##
## [[32]]
## NULL
##
## [[33]]
## NULL
##
## [[34]]
## NULL
##
## [[35]]
## NULL
##
## [[36]]
## NULL
##
## [[37]]
## NULL
##
## [[38]]
## NULL
##
## [[39]]
## NULL
##
## [[40]]
## NULL
##
## [[41]]
## NULL
##
## [[42]]
## NULL
##
## [[43]]
## NULL
##
## [[44]]
## NULL
##
## [[45]]
## NULL
##
## [[46]]
## NULL
##
## [[47]]
## NULL
```

7. Adding extra data

```
post_probs <- m4$posterior %>% as.data.frame() %>% setNames(paste0("C", 1:4, "_prob"))
ds <- bind_cols(d, post_probs)
ds <- rename(ds, class = m4_class)
write_csv(ds, "data-with-class-probs.csv")
```

8. Other ideas

- Examine unit-specific indicators by class?
- Examine control vs. intervention mean proportions?