Response to Review Comments

Paul Harmon

April 29, 2019

Introduction:

One of the points brought up by the reviewer is that the model is too sensitive to the size of the institutions. The reviewer notes the following in the feedback:

"Adopting the K&S variables cannot be imposed on these authors; however, a parallel re-analysis using those 10 variables and a comparison with their current findings is HIGHLY DESIRABLE. If the results are quite similar (as expected?), this would remove an unimportant source of noise resulting from the small and variable number of PhD's produced *each year* in the 3 boxes. To be explicit, in point 4 above, produce a second HGHMM set of scores using the K&S per capita variables. The final two scores would be (1) the original (biased towards large schools andd (2) a second (biased towards small schools via per capita."

We were unable to get the model to fit as a copy of the Carnegie version, but I interpret this comment as fitting the per-capita features into the SEM as we structured it.

Model Results:

The results from our original model:

```
library(dplyr);library(ggplot2);library(ggthemes);library(mclust);library(ggforce);library(shinyjs)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Package 'mclust' version 5.4.1
## Type 'citation("mclust")' for citing this R package in publications.
```

```
##
## Attaching package: 'shinyjs'
```

```
## The following objects are masked from 'package:methods':
##
##     removeClass, show
```

```
cc2015 <- filter(read.csv("data/CC2015data.csv",header = TRUE),BASIC2015 %in%c(15,16,17))
#dataset that we want to use
cc2015Ps<-
  na.omit(cc2015[,c("NAME","BASIC2010","BASIC2015","FACNUM","HUM_RSD","OTHER_RSD","SOCSC_RSD","STEM_RSD
minrank <- function(x){rank(x, ties.method = "min")}
#calculate the ranked data
cc2015.r <- data.frame(cc2015Ps[,1:3],sapply(cc2015Ps[,-c(1:3)],minrank))
cc2015percap <- cc2015Ps[,c("PDNFRSTAFF","S.ER.D","NONS.ER.D")]/cc2015Ps$FACNUM
```

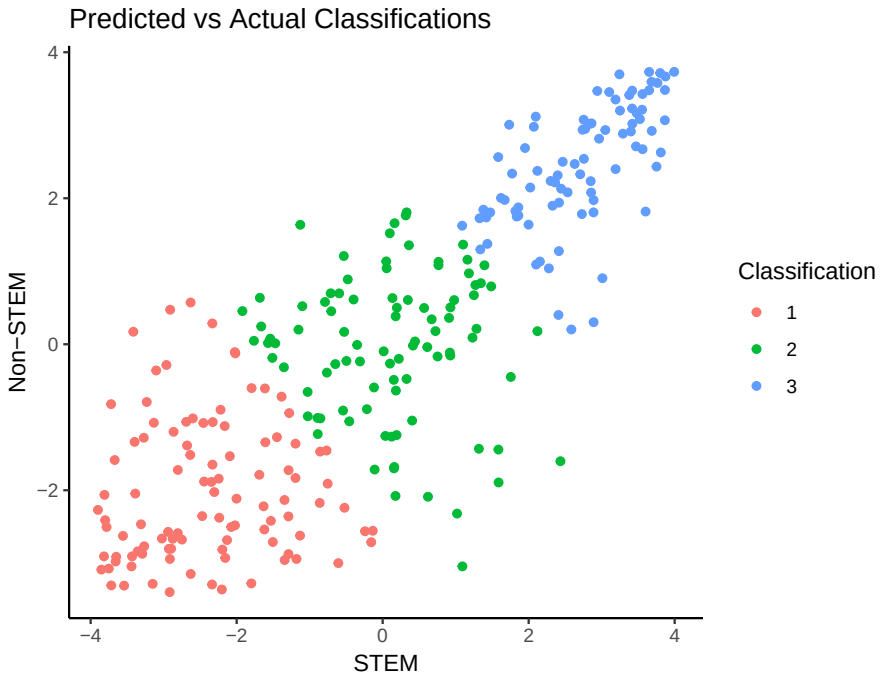
```r
colnames(cc2015percap) <- c("PDNRSTAFF_PC", "S.ER.D_PC", "NONS.ER.D_PC")
cc2015percap.r<-data.frame(sapply(cc2015percap,minrank))

#sem using raw data
cc2015_new <- cbind(cc2015.r, cc2015percap)
#cc2015_new <- cc2015.r

model_alt <- '
#latent factors
STEM=~STEM_RSD+PDNFRSTAFF+S.ER.D + FACNUM
HUM=~HUM_RSD + OTHER_RSD + SOCSC_RSD + NONS.ER.D + FACNUM
#factor of factors
Overall=~STEM+HUM
'
lavaan_sem_r_alternate <- lavaan::sem(model_alt, data=cc2015_new, std.lv=TRUE, orthogonal=FALSE, se="rob


#predicts the scores
CCScores_r_cov <- as.data.frame(lavaan::predict(lavaan_sem_r_alternate))
CCScores_r_cov_new <- apply(CCScores_r_cov[,c(1,2)], 2, scale)

mcres<-Mclust(CCScores_r_cov$Overall)
#summary(mcres)
Classifications <- mcres$classification
    #rownames(Classifications) <- cc2015Ps$NAME
#creates a plot and colors by Carnegie Classification Colors
ggplot(CCScores_r_cov) + geom_point(aes(x = STEM, y = HUM, color = factor(Classifications))) +
ggtitle("Predicted vs Actual Classifications") + theme_bw() + coord_fixed(ratio = 1)+
theme_classic() + guides(shape = FALSE) + guides(size = FALSE) +
labs(color = "Classification")+xlab("STEM") + ylab("Non-STEM")
```



```r
Scores1 <- CCScores_r_cov
Class1 <- Classifications
```

```r
lavaan::summary(lavaan_sem_r_alternate, fit.measures=TRUE)
```

```
## lavaan (0.6-1) converged normally after 128 iterations
##
##   Number of observations                           276
##
##   Estimator                                         ML
##   Model Fit Test Statistic                     110.024
##   Degrees of freedom                                17
##   P-value (Chi-square)                           0.000
##
## Model test baseline model:
##
##   Minimum Function Test Statistic             2223.162
##   Degrees of freedom                                28
##   P-value                                        0.000
##
## User model versus baseline model:
##
##   Comparative Fit Index (CFI)                    0.958
##   Tucker-Lewis Index (TLI)                       0.930
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)             -11847.548
##   Loglikelihood unrestricted model (H1)     -11792.536
##
##   Number of free parameters                         19
##   Akaike (AIC)                               23733.096
##   Bayesian (BIC)                             23801.883
##   Sample-size adjusted Bayesian (BIC)        23741.638
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.141
##   90 Percent Confidence Interval          0.116  0.166
##   P-value RMSEA <= 0.05                          0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.045
##
## Parameter Estimates:
##
##   Information                                 Observed
##   Observed information based on                Hessian
##   Standard Errors                   Robust.huber.white
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   STEM =~
```

```
##     STEM_RSD         33.562    9.334    3.596    0.000
##     PDNFRSTAFF       34.448    8.871    3.883    0.000
##     S.ER.D           33.529    9.122    3.676    0.000
##     FACNUM           13.886    6.094    2.279    0.023
##   HUM =~
##     HUM_RSD          38.108   10.291    3.703    0.000
##     OTHER_RSD        24.120    6.937    3.477    0.001
##     SOCSC_RSD        37.677   10.073    3.740    0.000
##     NONS.ER.D        27.306    7.356    3.712    0.000
##     FACNUM           18.010    6.969    2.584    0.010
##   Overall =~
##     STEM              2.068    0.688    3.007    0.003
##     HUM               1.885    0.649    2.905    0.004
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .STEM_RSD        796.821  117.031    6.809    0.000
##    .PDNFRSTAFF      629.148  162.047    3.883    0.000
##    .S.ER.D          415.725   88.612    4.692    0.000
##    .FACNUM         1908.187  223.671    8.531    0.000
##    .HUM_RSD        2612.532  366.089    7.136    0.000
##    .OTHER_RSD      3835.029  334.063   11.480    0.000
##    .SOCSC_RSD      1403.253  236.802    5.926    0.000
##    .NONS.ER.D      2992.509  305.049    9.810    0.000
##     STEM               1.000
##     HUM                1.000
##     Overall            1.000
```

Alternate Model

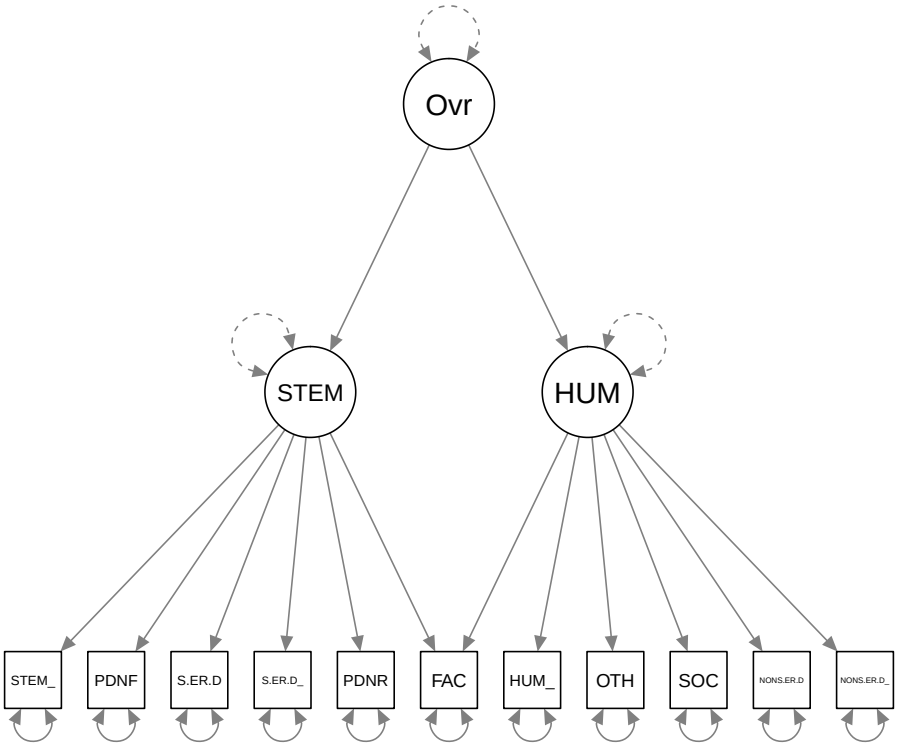
The model with Per-capita included looks a little different. A path diagram is given below.

```
model_alt2 <- '
#latent factors
STEM=~STEM_RSD+PDNFRSTAFF+S.ER.D + FACNUM + S.ER.D_PC + PDNRSTAFF_PC
HUM=~HUM_RSD + OTHER_RSD + SOCSC_RSD + NONS.ER.D + FACNUM + NONS.ER.D_PC
#factor of factors
Overall=~STEM+HUM
'


lavaan_sem_r_alternate <- lavaan::sem(model_alt2, data=cc2015_new, std.lv=TRUE, orthogonal=FALSE, se="r
```

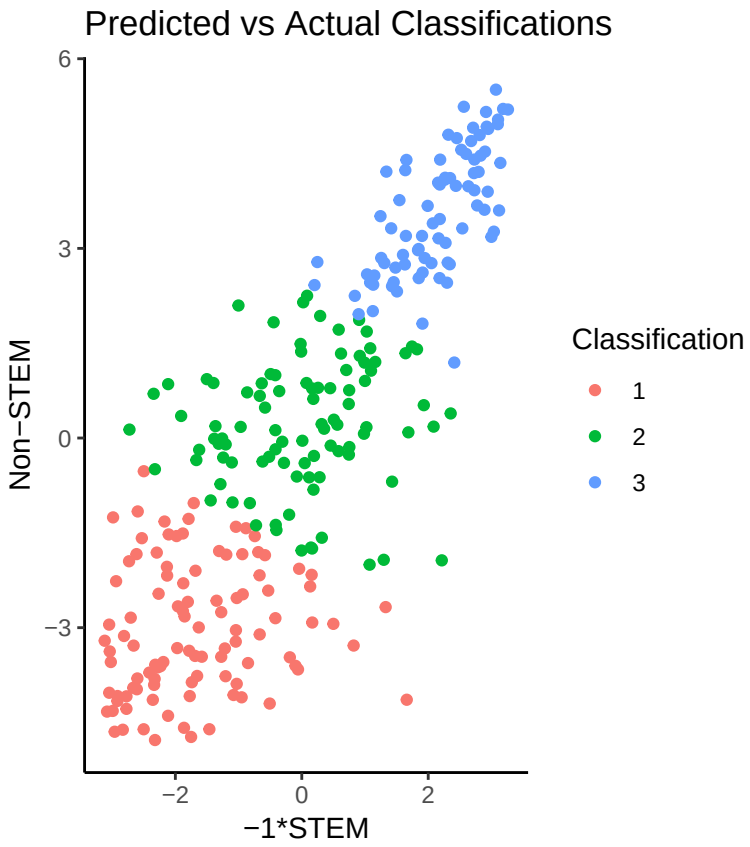
```
## Warning in lav_data_full(data = data, group = group, cluster = cluster, :
## lavaan WARNING: some observed variances are (at least) a factor 1000 times
## larger than others; use varTable(fit) to investigate
```

```
semPaths(lavaan_sem_r_alternate)
```




```
#predicts the scores
CCScores_r_cov <- as.data.frame(lavaan::predict(lavaan_sem_r_alternate))
CCScores_r_cov_new <- apply(CCScores_r_cov[,c(1,2)], 2, scale)

mcres<-Mclust(CCScores_r_cov$Overall)
#summary(mcres)
Classifications <- mcres$classification
    #rownames(Classifications) <- cc2015Ps$NAME
#creates a plot and colors by Carnegie Classification Colors
ggplot(CCScores_r_cov) + geom_point(aes(x = -STEM, y = HUM, color = factor(Classifications))) +
ggtitle("Predicted vs Actual Classifications") + theme_bw() + coord_fixed(ratio = 1)+
theme_classic() + guides(shape = FALSE) + guides(size = FALSE) +
labs(color = "Classification")+xlab("-1*STEM") + ylab("Non-STEM")
```




```
Scores2 <- CCScores_r_cov
Scores2$Name <- cc2015_new$NAME
Class2 <- Classifications
```

The summary of the model is here:

```
lavaan::summary(lavaan_sem_r_alternate, fit.measures=TRUE)
```

```
## lavaan (0.6-1) converged normally after 161 iterations
##
##   Number of observations                           276
##
##   Estimator                                         ML
##   Model Fit Test Statistic                     767.579
##   Degrees of freedom                                41
##   P-value (Chi-square)                           0.000
##
## Model test baseline model:
##
##   Minimum Function Test Statistic             3032.998
##   Degrees of freedom                                55
##   P-value                                        0.000
##
## User model versus baseline model:
##
##   Comparative Fit Index (CFI)                    0.756
##   Tucker-Lewis Index (TLI)                       0.673
```

```
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)             -15048.664
##   Loglikelihood unrestricted model (H1)     -14664.874
##
##   Number of free parameters                         25
##   Akaike (AIC)                               30147.328
##   Bayesian (BIC)                             30237.838
##   Sample-size adjusted Bayesian (BIC)        30158.567
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.253
##   90 Percent Confidence Interval          0.238  0.269
##   P-value RMSEA <= 0.05                          0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.121
##
## Parameter Estimates:
##
##   Information                                 Observed
##   Observed information based on                Hessian
##   Standard Errors                   Robust.huber.white
##
## Latent Variables:
##                    Estimate   Std.Err   z-value  P(>|z|)
##   STEM =~
##     STEM_RSD         -41.680     7.633   -5.460    0.000
##     PDNFRSTAFF       -43.041     7.871   -5.468    0.000
##     S.ER.D           -41.951     7.686   -5.458    0.000
##     FACNUM           -16.851     6.952   -2.424    0.015
##     S.ER.D_PC        -74.077    13.473   -5.498    0.000
##     PDNRSTAFF_PC      -0.125     0.027   -4.621    0.000
##   HUM =~
##     HUM_RSD           27.207    13.944    1.951    0.051
##     OTHER_RSD         17.293     9.557    1.810    0.070
##     SOCSC_RSD         26.909    13.750    1.957    0.050
##     NONS.ER.D         20.239    10.612    1.907    0.056
##     FACNUM            13.133     9.248    1.420    0.156
##     NONS.ER.D_PC       1.571     0.834    1.883    0.060
##   Overall =~
##     STEM              -1.546     0.405   -3.815    0.000
##     HUM                2.794     1.628    1.716    0.086
##
## Variances:
##                    Estimate   Std.Err   z-value  P(>|z|)
##    .STEM_RSD         849.536   118.816    7.150    0.000
##    .PDNFRSTAFF       608.781   159.167    3.825    0.000
##    .S.ER.D           380.097    83.681    4.542    0.000
##    .FACNUM          1957.969   223.188    8.773    0.000
##    .S.ER.D_PC      60621.416 49274.307    1.230    0.219
```

```
##     .PDNRSTAFF_PC       0.243     0.141    1.726    0.084
##     .HUM_RSD         2706.103   374.165    7.232    0.000
##     .OTHER_RSD       3850.351   338.156   11.386    0.000
##     .SOCSC_RSD       1489.965   253.081    5.887    0.000
##     .NONS.ER.D       2780.313   333.417    8.339    0.000
##     .NONS.ER.D_PC     154.388    32.925    4.689    0.000
##      STEM               1.000
##      HUM                1.000
##      Overall            1.000
```

Results

These are very preliminary results, but the shape hasn't changed drastically. A couple of institutions get pulled away from the rest of the cluster - the bottom one on the stem scale is interesting. Including per-capita features does not seem to cause problems in the SEM model, as the model converges.

However, the sign on the STEM term has flipped - loadings are now negative, and I had to reverse the order of the STEM factor to get the plot to point in the direction we expected to see. (This does not have an effect on the groupings but is interesting nonetheless). Looking at the estimated coefficients for the loadings and their associated p-values, we see that the evidence in favor of an effect of the humanities latent factor is moderate at best; nearly all the p-values for the model with per-capita features indicate diminished strength of evidence.

Looking at some of the individual school classifications, we see that Cal Tech remains in the middle-size category and that .

```
x <- tibble(Scores2$Name, Scores1$Overall, Scores2$Overall, Class1, Class2)
names(x) <- c("Name","Score1","Score2", "M1Class", "M2Class")
#x$DifferenceSq <- (x$Score1 - x$Score2)^2
x$AbsDiff <- abs(x$Score1 - x$Score2)

#biggest difference schools
arrange(x, desc(AbsDiff)) %>% head(15) %>% pander()
```

Name	Score1	Score2	M1Class	M2Class	AbsDiff
Yeshiva University	-0.03212	-0.4842	2	1	0.4521
Wake Forest University	-0.2564	-0.7058	2	1	0.4494
Claremont Graduate University	-0.5808	-0.149	1	2	0.4318
The New School	-0.7633	-0.3431	1	2	0.4202
Fordham University	-0.4942	-0.07932	1	2	0.4148
Rockefeller University	-0.3937	-0.8038	2	1	0.4101
California Institute of Technology	0.2285	-0.1769	2	2	0.4054
American University	-0.353	0.02492	2	2	0.3779
Indiana University-Purdue University-Indianapolis	0.06348	-0.3015	2	2	0.365
Rensselaer Polytechnic Institute	-0.3005	-0.6645	2	1	0.364
Dartmouth College	0.003213	-0.3517	2	2	0.3549
Missouri University of Science and Technology	-0.7822	-1.118	1	1	0.336
Augusta University	-0.5772	-0.9129	1	1	0.3357
Ball State University	-0.8033	-0.4762	1	1	0.3271
Indiana University of Pennsylvania-Main Campus	-1.046	-0.7255	1	1	0.3209