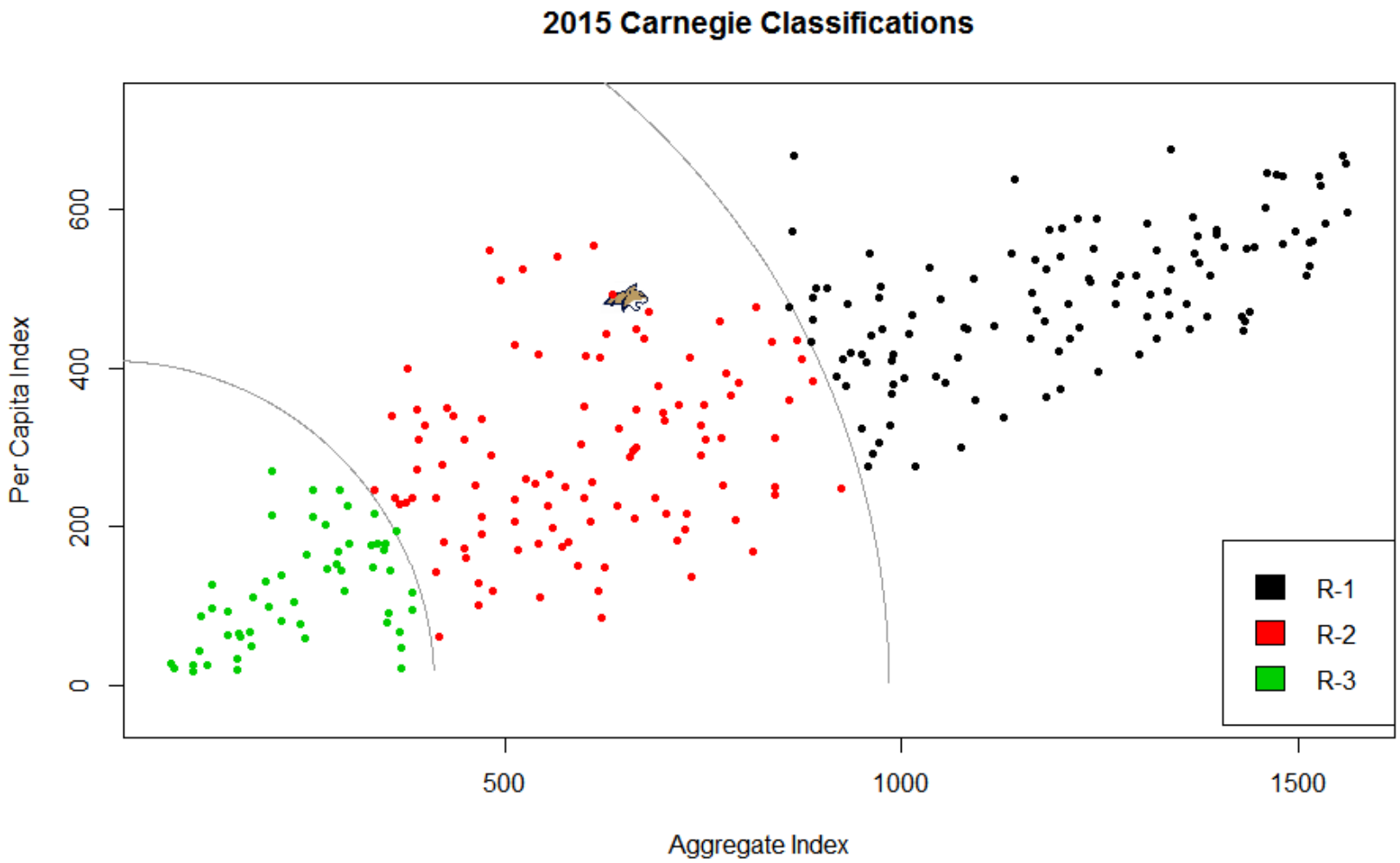
Carnegie SEM

Paul Harmon and Sarah McKnight

November 1, 2017

The Carnegie Classifications

- Based on PCA on correlation matrix of ranked institutional data
- Two Scales: Aggregate (x-axis) and Per-Capita (y-axis) _ Three categories of PhD-granting institutions: R1 to R3



How are Scores Calculated?

The Carnegie Classifications are based on ranked data (because some schools are vastly larger on certain metrics than the rest). The ranked data are partitioned into two datasets and a PCA is generated on the both datasets. The first PC score is then used as an index for each trait.

Aggregate and Per-Capita Indices

The classifications are calculated based on two indices of institutional output. The first is based on a weighted average of the number of PhDs awarded by the institution; the second is based on a per-capita measurement of research expenditures and research staff. **Aggregate Index:**

$$Ag.Index_i = HumanitiesPhD_i + StemPhD_i + SocialSciencePhD_i + OtherPhD$$

Per Capita Index:

$$PC.Index_i = \frac{ResearchStaff_i + StemExpenditures_i + NonStemExpenditure}{FacultySize_i}$$

PCA Plot - Arbitrary Boundaries

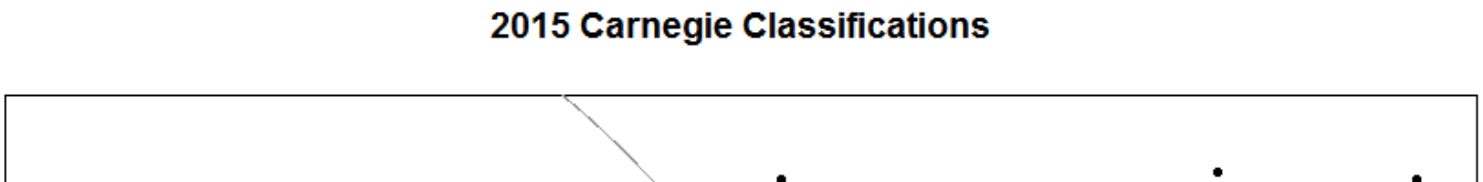
From here, the methodology of the Carnegie Classifications is to produce a plot of the two correlated indices. Lines are arbitrarily drawn to create three partitions of the space. The following plot is produced.

[The 2015 Classifications with lines]

Problems with PCA

Many of the decisions made in this process are data driven and are outside the control of researchers:

- Aggregate and Per-Capita Indices were constructed so that they are correlated, but how correlated they are can change from year to year
- Attributes of the PCAs themselves change from year to year, meaning important policy drivers in one year might not be important in the next year (Social Science PhDs)
- This can affect policy decisions made based on the classifications
- Data are rarely grouped into three neat clusters, and lines are arbitrarily drawn
- Both PCAs throw out roughly 30 percent of the variation in underlying variables



Structural Equation Models

GoaL: Model two latent traits, one for the **aggreagate scale** and the other for the **per-capita scale**, and then use those to generate a "super-latent" single trait.

We could then separate institutions into groups via univariate model-based clustering or something as simple as breaking the single variable into several groups.

Problems Solved by the SEM

The SEM is constructed a priori because we set the latent variables and correlation structure beforehand. This leads to a solution that is less succeptible to change when the data are updated.

- We don't have an arbitray curve slicing up groups of schools - it's either cut based on a model-based clustering algorithm or sliced into proportional groups.

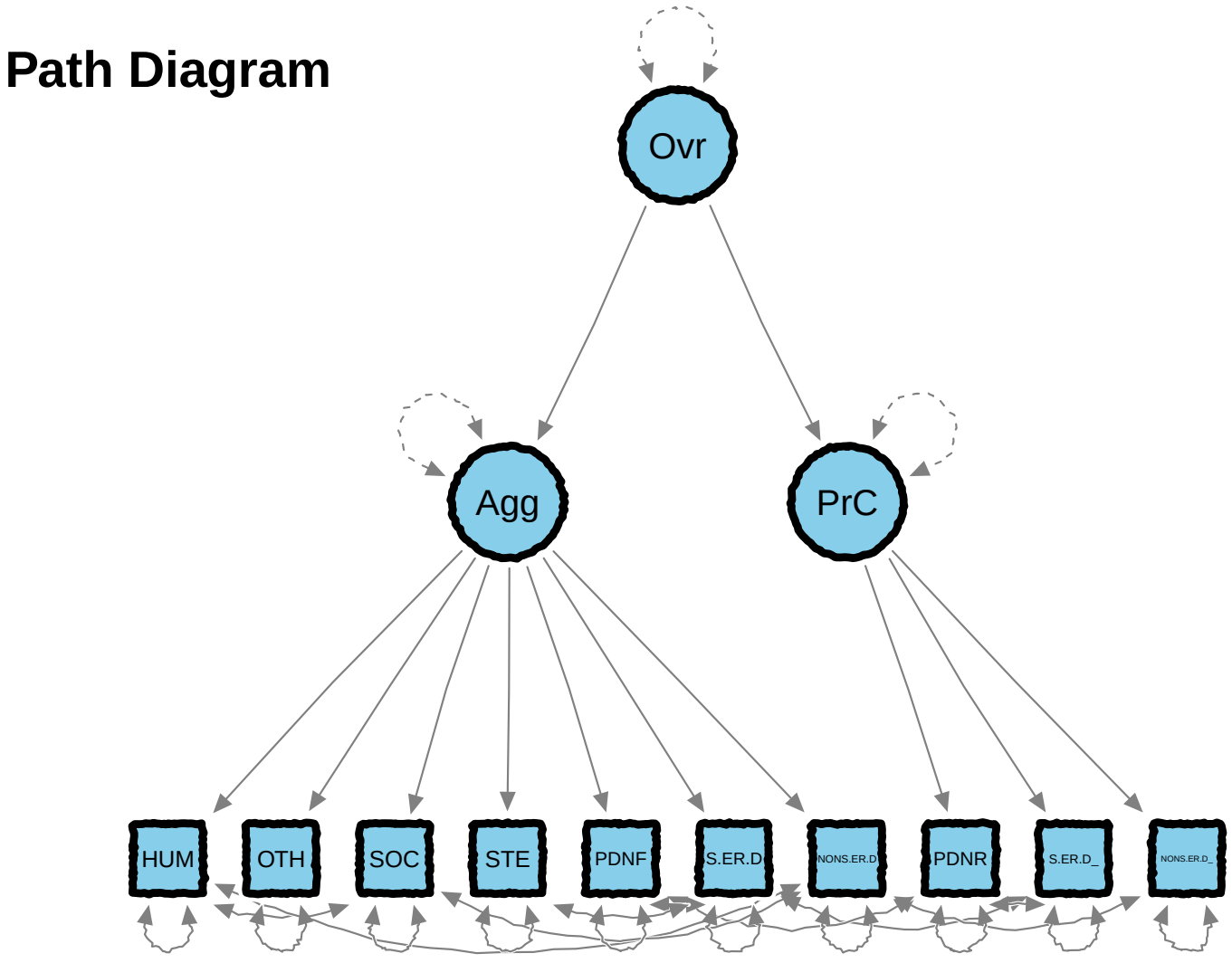
SEM Model

Output from SEM Model is below. The RMSEA was initially close to 0.9 but after modeling covariances, we were able to get it down to 0.1. This model had the lowest AIC value of any we fit and the tests for the manifest variables and correlations were found to be *statistically significant*.

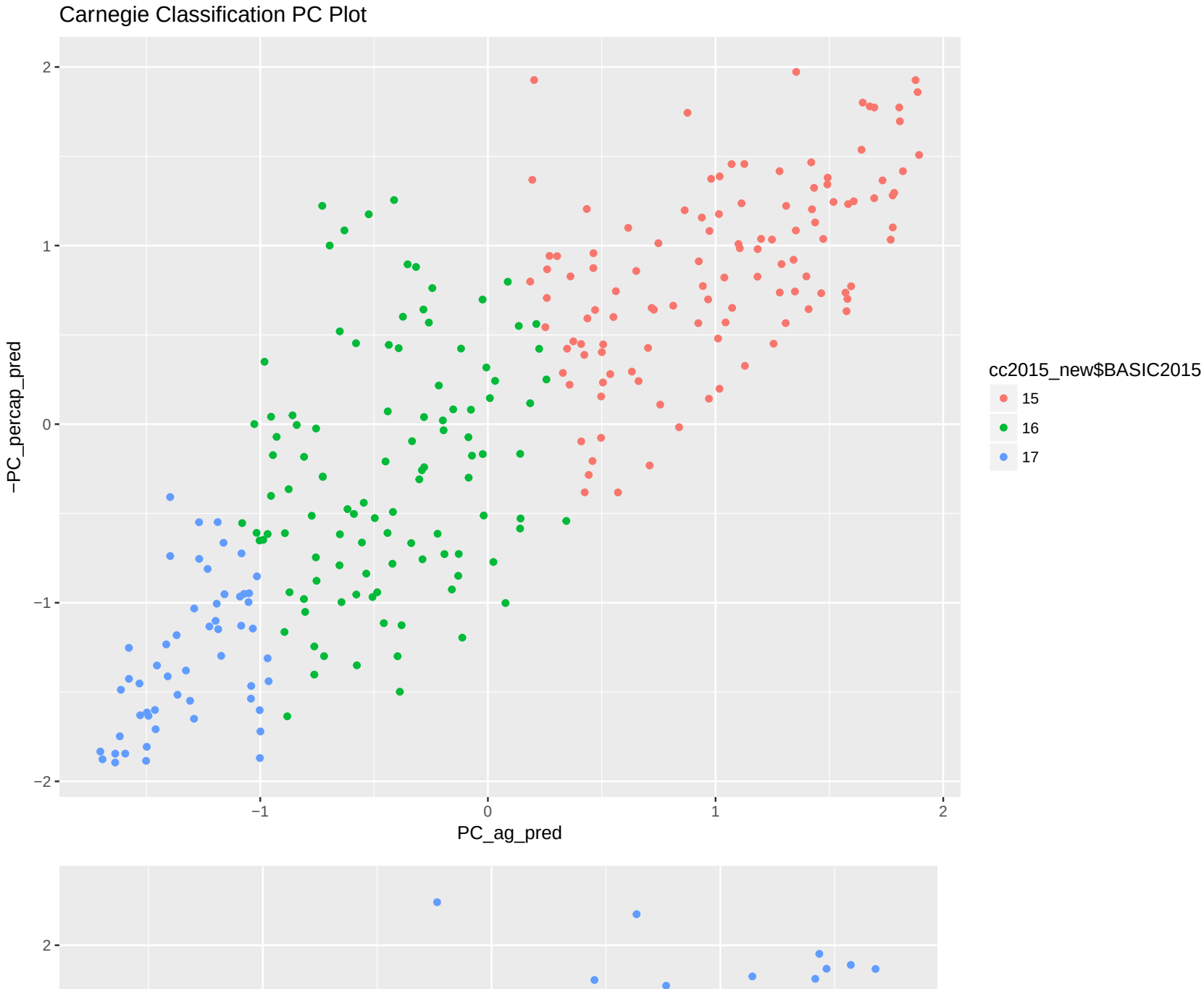
```
## Warning in lav_object_post_check(object): lavaan WARNIN
##                 variables (theta) is not positive defin
##                 use inspect(fit,"theta") to investigate

## lavaan (0.5-23.1097) converged normally after 499 iterat
##
##   Number of observations                           276
##
##   Estimator                                         ML
##   Minimum Function Test Statistic              253.377
##   Degrees of freedom                                24
##   P-value (Chi-square)                           0.000
##
```

Path Diagram



PCA vs. SEM


Carnegie Classification PC Plot
−PC_percap_pred
PC_ag_pred
2
1
0
−1
−2
−1
0
1
2
cc2015_new$BASIC2015
15
16
17

Questions For Laura

1. What is the official term for "super latent trait"?
2. Could we improve the fit by doing something other than just modeling covariances?
3. Comparing scores from PCA to SEM - is this reasonable?