Opportunity Mapping Factor Analysis Memo

Arthur Gailes

July 7, 2019

Contents

Recommendation

The recommended model was developed using a sparse, non-negative principal component analysis (PCA). I recommend either using the loading scores provided in the diagram below or weighting each domain score by the number of indicators it contains. The domains are:

- Domain 1:
  - Median home values
  - Percentage on public assistance*
  - Bachelor's degrees
  - Insured rate
  - Employment Rate
  - % of students not receiving free or reduced-price lunch
  - Pre-K
  - School district revenue
  - Reading scores
    - Because math and readings scores are so highly correlated (0.96), I've removed math scores from the analysis. An alternate method would be to average the two inputs.
- Domain 2:
  - Job proximity
  - Healthcare proximity
  - Religious proximity
  - Parks
  - Ozone pollution*
  - Broadband
  - % registed to vote
  - Pesticides*

- Domain 3:
  - Grocery
  - Clubs
  - Diesel emissions*
  - PM2.5 emissions*
- Domain 4:
  - Crime rate
  - Lead *
  - High school graduation rate
- Domain 5:
  - % of residents with long commutes*
  - Toxic Release*
  - Onsite Toxic Release*

*: This indicator was adjusted to be positive - i.e. toxic release refers to the lack of toxins in a tract.

Model Details

The selected model is based on a non-negative PCA model. "Non-negative" means that the indicator loads positively, so that an indicators that are negatively correlate with each other are kept apart. To create the final model, I simply assigned the indicators to the domains they loaded upon most strongly.

Advantages

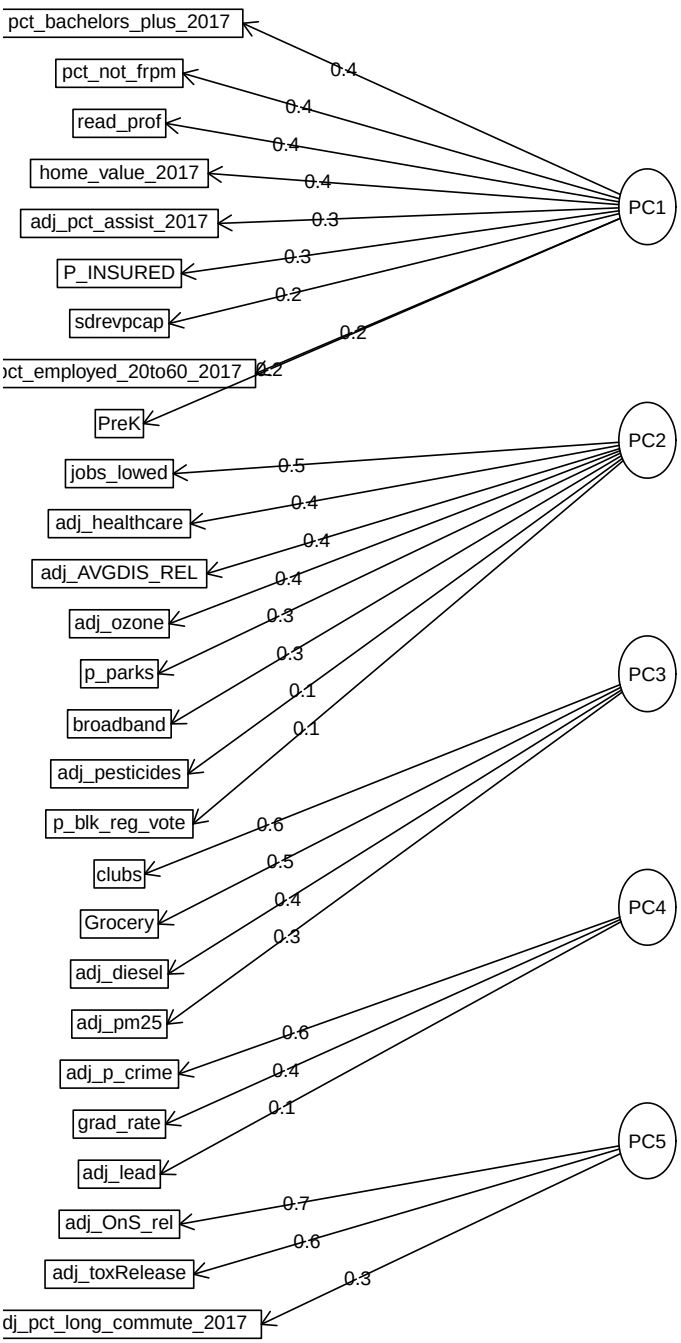
- Simple and easily interpreted by laypeople, especially compared to a standard PCA.
- Based primarily on data correlation, rather than intuition.
- Able to be empirically validated.
- Keeps all indicators.
- The non-negative constraint means that all of the indicators contribute positively to the final index score.

Disadvantages

- Explains slightly less of the variance than a standard PCA model (47% vs 50%)
- Some domains are unintuitive.

Diagram:

```
set.seed(3)
sparse.pca.result <- nsprcomp(inds, ncomp=5, nneg = T)
sparse2 <- nsprcomp(sparse.pca.result$x, ncomp=1,nneg=T)
pctVariance <- sum(peav(inds, sparse.pca.result$rotation))
fa.diagram(sparse.pca.result$rotation, cut = 0, simple = T, main = 'Sparse PCA')
```



Factor Loadings

Alternatively, we could use the loadings as presented in the table below. This will explain the most variance.

- Data note: this can be retrieved via sparse.pca.result$x.

Table 1: Full Sparse PCA Model

	PC1	PC2	PC3	PC4	PC5
adj_ozone	0	0.391	0.165	0	0
adj_pm25	0	0	0.314	0	0
adj_diesel	0.030	0	0.448	0.400	0
adj_toxRelease	0.160	0.125	0	0	0.633
adj_pesticides	0.015	0.105	0.016	0	0
adj_lead	0	0.126	0.008	0.131	0
grad_rate	0.211	0	0.058	0.419	0.027
pct_not_frpm	0.380	0	0.005	0.152	0
read_prof	0.373	0	0	0.107	0
jobs_lowed	0.079	0.474	0	0	0.039
broadband	0	0.257	0.123	0	0
Grocery	0.086	0	0.546	0	0.005
P_INSURED	0.333	0	0	0	0
clubs	0	0	0.566	0	0.060
adj_p_crime	0.035	0	0.140	0.573	0.160
adj_healthcare	0.027	0.427	0	0.337	0
adj_AVGDIS_REL	0	0.421	0	0.413	0
adj_pct_assist_2017	0.335	0	0.014	0	0
sdrevpcap	0.209	0	0.089	0	0.154
adj_OnS_rel	0.026	0.083	0.029	0	0.687
p_blk_reg_vote	0	0.075	0.013	0	0
p_parks	0	0.258	0.079	0	0
PreK	0.188	0.064	0	0	0
pct_employed_20to60_2017	0.195	0.099	0	0	0
home_value_2017	0.362	0.108	0	0	0.007
pct_bachelors_plus_2017	0.385	0.110	0	0	0
adj_pct_long_commute_2017	0.098	0.195	0.030	0	0.270

Alternative Models

Intuitive model

A more intuitive model, informed primarily by theory, and secondarily by correlations. Unfortunately, I was unable to evaluate this option empirically.

Model

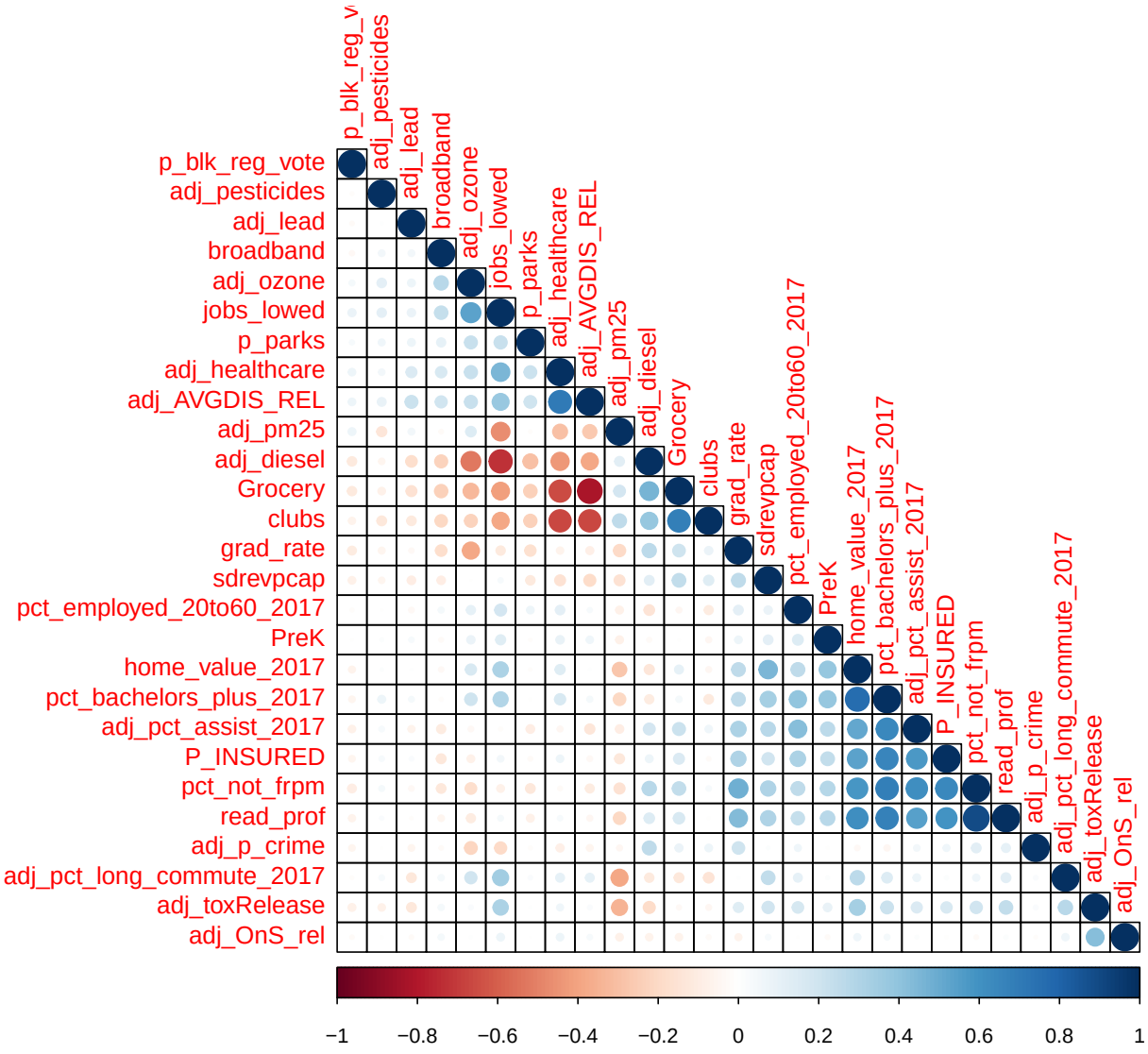
This model follows a domain structure of economics, education, environment, social, and proximity, modified slightly by taking into account the correlations below

- Economics:
  - Median home values
  - Percentage on public assistance*
  - Insured rate

  - Employment Rate
  - School district revenue
- Education:
  - High school graduation rate
  - Bachelor's degrees
  - % of students not receiving free or reduced-price lunch
  - Pre-K
  - Reading scores
- Environment:
  - Toxic Release*
  - Onsite Toxic Release*
  - Diesel emissions*
  - PM2.5 emissions*
  - Lead*
  - Pesticides*
  - Ozone pollution*
- Social:
  - Crime rate
  - Religious proximity
  - Clubs
  - % registed to vote
- Proximity to Services:
  - Grocery
  - % of residents with long commutes*
  - Job proximity
  - Healthcare proximity
  - Parks
  - Broadband

Correlation Plot

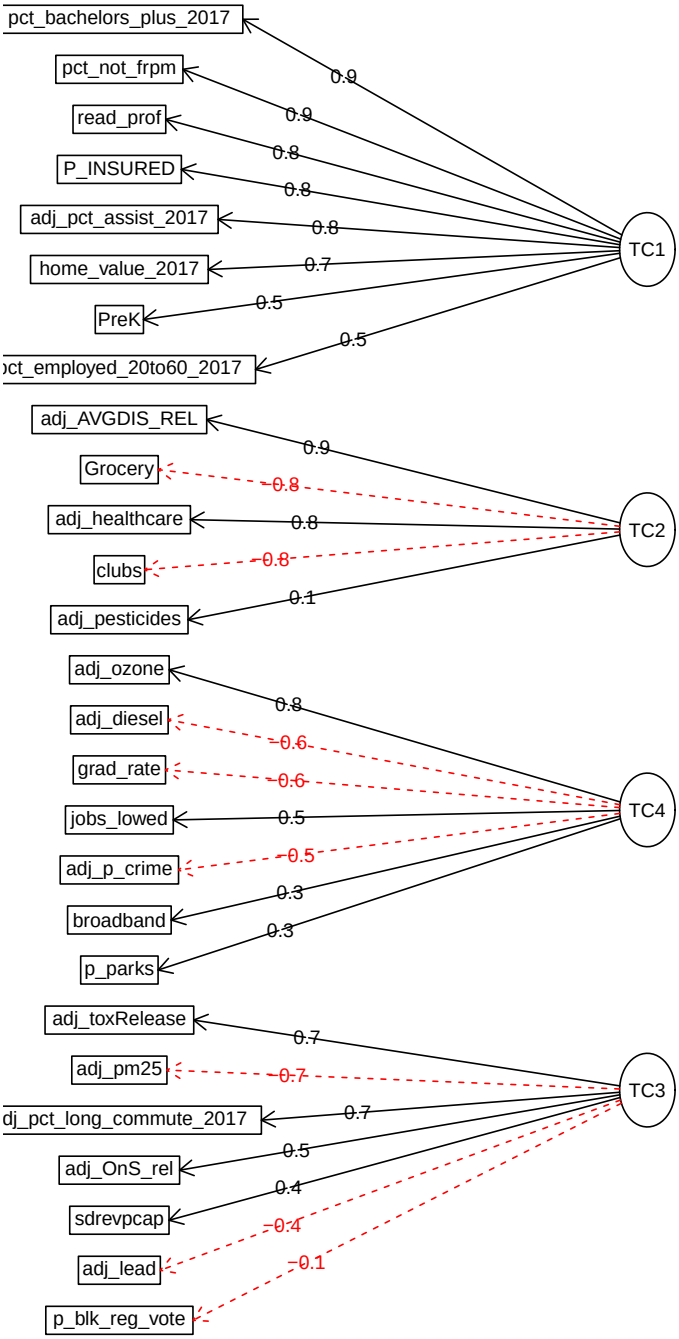
```
cor_inds <- cor(inds)
corrplot(cor_inds, type = 'lower', order='hclust', addgrid.col = T)
```



PCA

```
pca.result <- principal(inds, nfactors=4, rotate='oblimin')
fa.diagram(pca.result$loadings, cut = 0, simple = T, main='PCA')
```

PCA



The standard PCA model performs slightly better than the non-negative PCA above, but has two major disadvantages: it is harder to explain and some loadings are primarily negative, contrasting our theoritical interpretation of the scores.

Factor Analysis

I chose a PCA analysis over a factor analysis because factor analysis for three reasons:

1) The EFA models had difficulty including indicators that didn't correlate well with any other indicators. This is because the nature of factor analysis lies in finding underlying factors that effect groups of indicators similarly.

2) Factor analysis is geared towards understanding latent variables. Latent variables, in theory, cause the observed indicators in the analysis. For example, the standard factor analysis is of IQ. In an IQ test, a person's spatial reasoning score is dictated by their answers to spatial reasoning questions.

   The key here is that a person's spatial reasoning *causes* people's scores on that section of the IQ test. Whereas in opportunity mapping, it's unclear that a tract's opportunity causes there to be high home values, incomes, etc. It's more the opposite case, where the influx of people with high incomes, education, etc that allows that tract to flourish.

3) As with the standard PCA, the presence of negative loadings causes a theoretical dilemma, where some indicators will load negatively onto their domains, causing a score where opportunity increases when some of our indicators perform worse.

Exploratory Factor Analysis

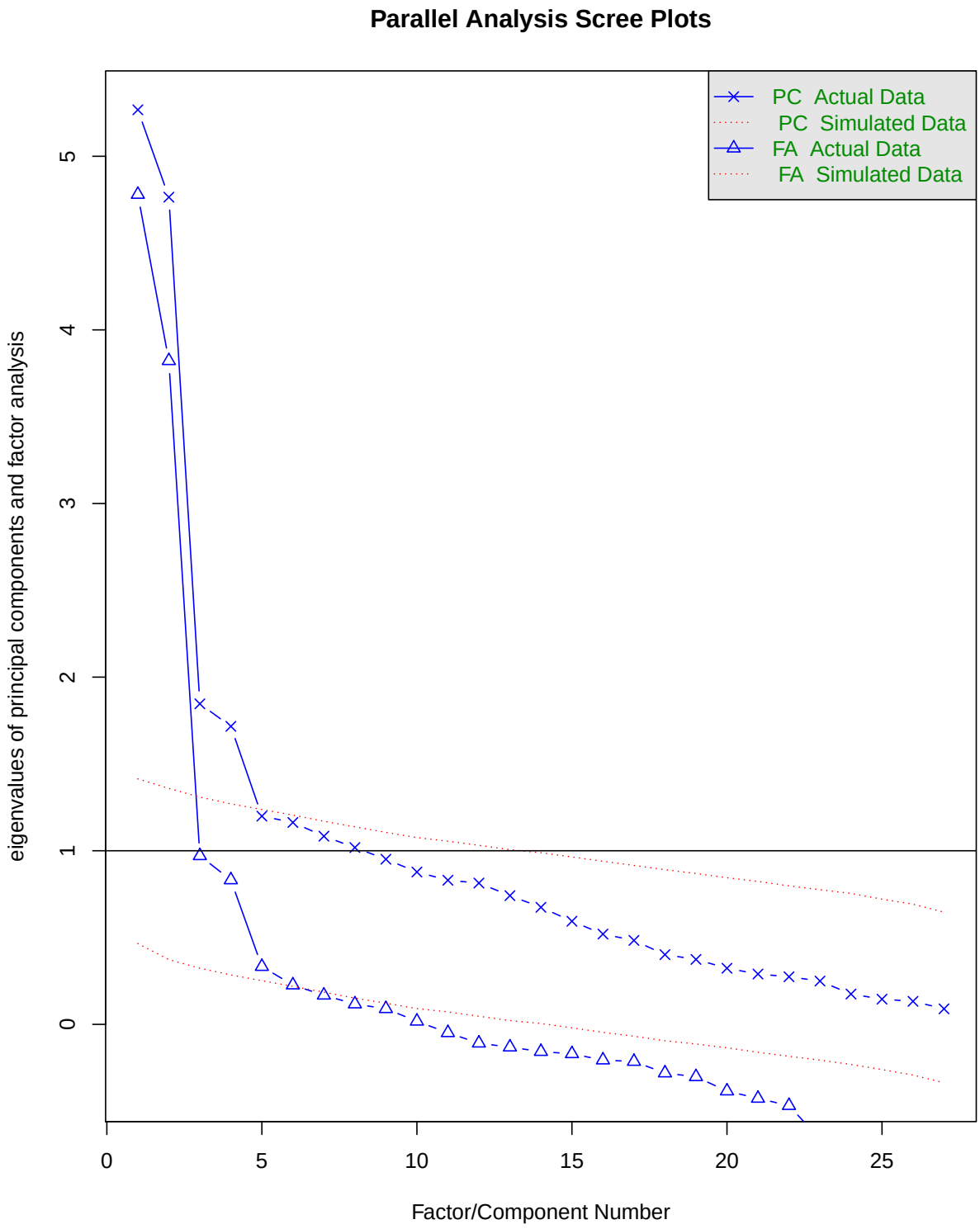
The primary benefit of exploratory factor analysis is that it evaluates the data without any theoretical preconceptions. This is useful for uncovering insights about the data may not be apparent from theory.

The disadvantage is that the EFA doesn't necessarily use all of the variables. This is because some variables don't correlate well with anything (e.g. pesticides).

Eigevalues and Scree plot

The scree plot is a preliminary analysis of the necessary number of factors.

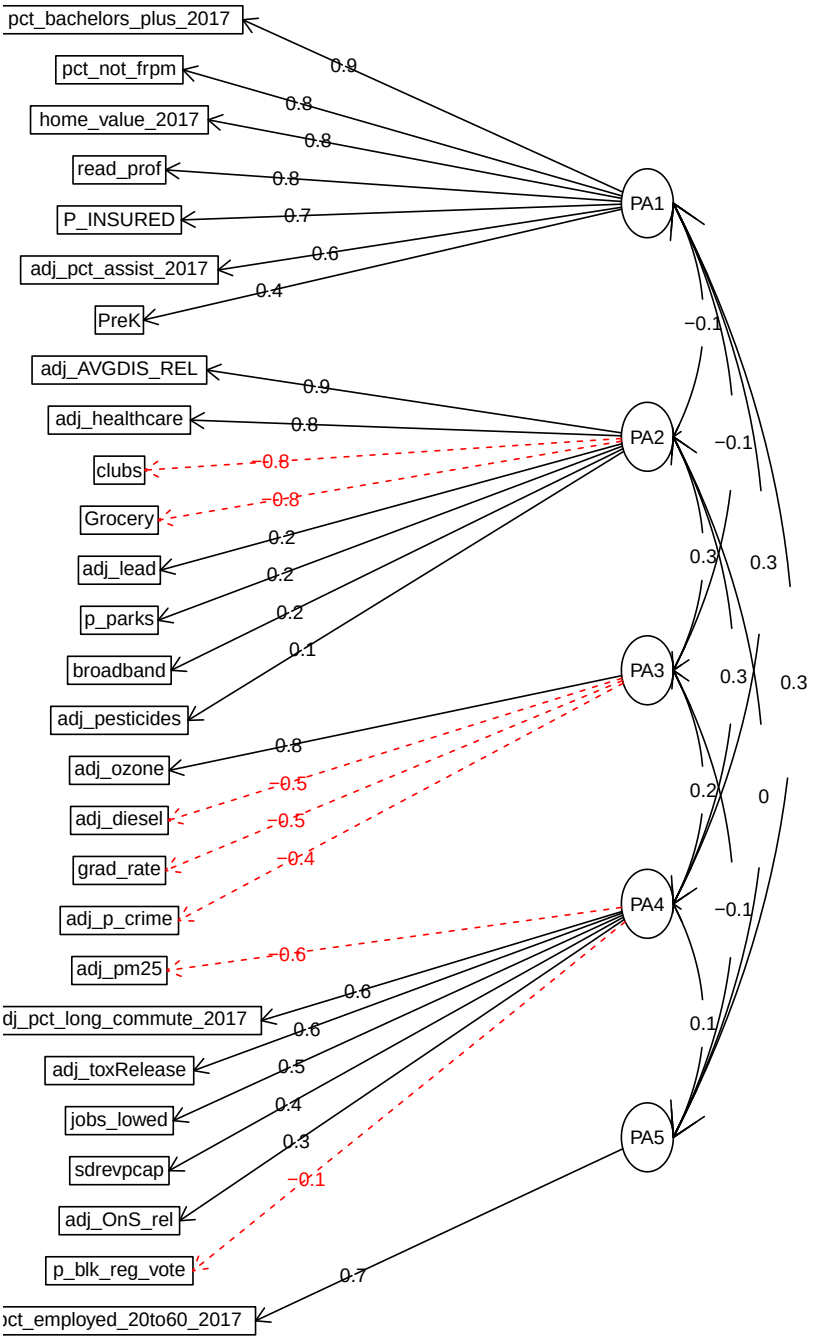
```
EFA_ind <- inds[indices_EFA,] #random half of the dataset
cor_EFA <- cor(EFA_ind)
scree <- fa.parallel(cor_EFA, n.obs = nrow(EFA_ind))
```




```
## Parallel analysis suggests that the number of factors =  5  and the number of components =  4
```

EFA Model

```
EFA_mod <- fa(EFA_ind, nfactors = 5, covar = F, fm = 'pa')
fa.diagram(EFA_mod, cut=0)
```

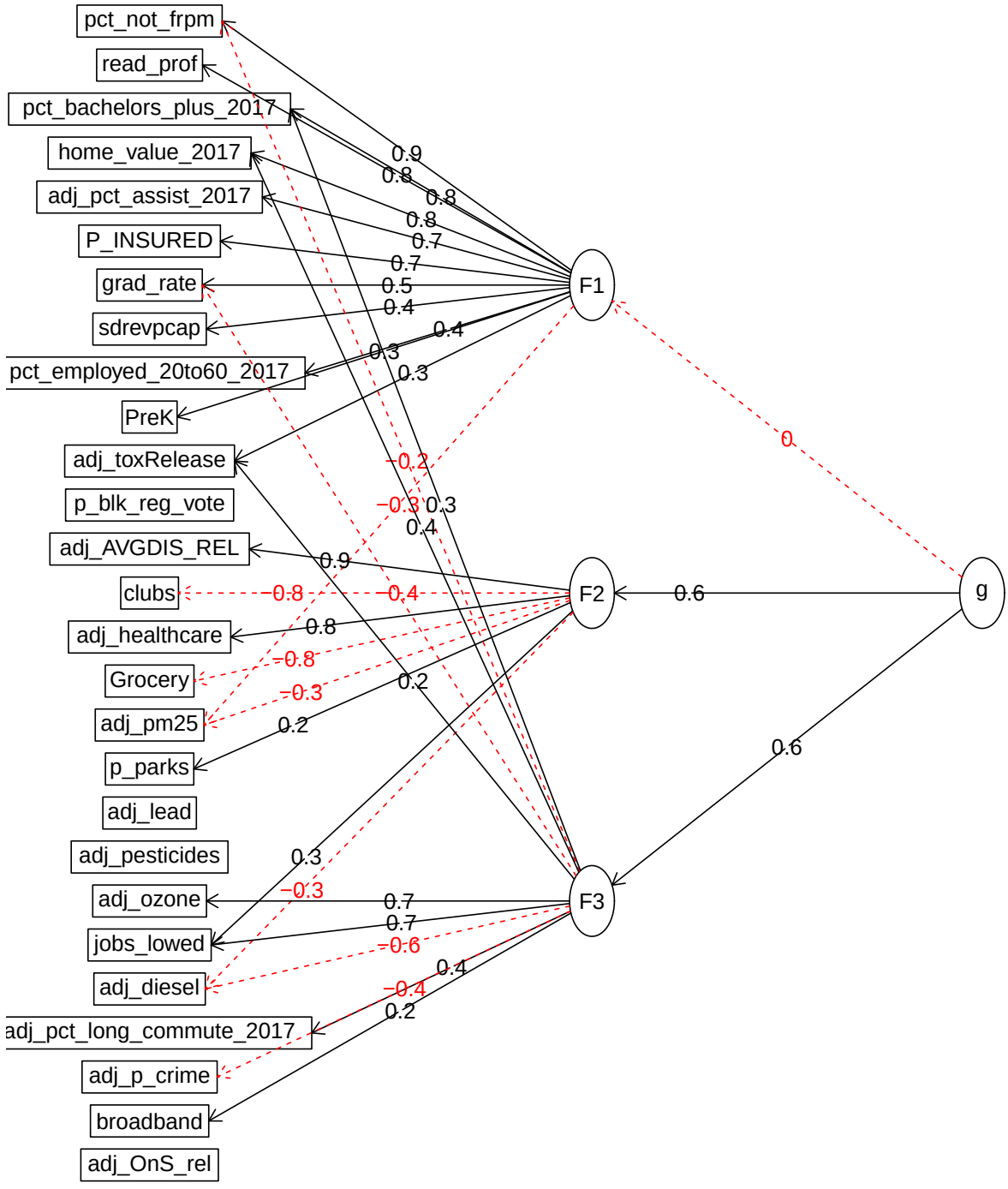


Hierarchical Factor Analysis

This model adds the general factor, and is useful for calculating the most efficient possible final opportunity index. Unfortunately, the output of this model is theoretically poor.

```
EFA_hier <- omega(EFA_ind, sl=F, fm='pa')
```

Omega


pct_not_frpm
read_prof
pct_bachelors_plus_2017
home_value_2017
adj_pct_assist_2017
P_INSURED
grad_rate
sdrevpcap
pct_employed_20to60_2017
PreK
adj_toxRelease
p_blk_reg_vote
adj_AVGDIS_REL
clubs
adj_healthcare
Grocery
adj_pm25
p_parks
adj_lead
adj_pesticides
adj_ozone
jobs_lowed
adj_diesel
adj_pct_long_commute_2017
adj_p_crime
broadband
adj_OnS_rel
F1
F2
F3
g
0.9
0.8
0.8
0.8
0.7
0.7
0.5
0.4
0.4
0.3
0.3
0
−0.2
−0.3
0.3
0.4
0.9
−0.8
−0.4
0.6
0.8
−0.8
−0.3
0.2
0.2
0.6
0.3
−0.3
0.7
0.7
−0.6
0.4
−0.4
0.2

Appendix: Code for Confirmatory Factor Analysis.

CFA was unsuccessful except for the PCA and non-hierarchical EFA models. The code below is provided primarily for reference.

Create lavaan-appropriate syntax for all models.

```
sparse_syn <- ("
  PC1 =~  + pct_not_frpm + read_prof + P_INSURED + adj_pct_assist_2017 + sdrevpcap + PreK + pct_employe
  PC2 =~  + adj_ozone + adj_pesticides + jobs_lowed + broadband + adj_healthcare + adj_AVGDIS_REL + p_bl
  PC3 =~  + adj_pm25 + adj_diesel + Grocery + clubs
  PC4 =~  + adj_lead + grad_rate + adj_p_crime
  PC5 =~  + adj_toxRelease + adj_OnS_rel + adj_pct_long_commute_2017
  g =~ PC1 + PC2 + PC3 + PC4 + PC5
  ")

intuit_syn <- ('
  Economics =~  + home_value_2017 + adj_pct_assist_2017 + pct_employed_20to60_2017  + P_INSURED
  Education =~  + PreK + grad_rate + pct_not_frpm + read_prof + sdrevpcap + pct_bachelors_plus_2017
  Environment =~ + adj_pm25 + adj_ozone + adj_pesticides + adj_OnS_rel +adj_lead + adj_diesel
  Social =~  + clubs + p_blk_reg_vote + adj_p_crime + p_blk_reg_vote + adj_AVGDIS_REL
  Proximity =~ + adj_healthcare + Grocery + jobs_lowed + p_parks + adj_pct_long_commute_2017 + broadban
  index =~ Economics + Education + Environment + Social + Proximity
                ')

EFA_mod_syn <- structure.diagram(EFA_mod, errors = T)$lavaan
EFA_hier_syn <- EFA_hier$omegaSem$model$lavaan
pca_syn <- structure.diagram(pca.result$loadings, errors = T)$lavaan
#the sparse PCA model, rather than my intuitive extrapolation
sparseMod_syn <- structure.diagram(sparse.pca.result$rotation, errors = T)$lavaan
```

Perform CFA for all models

```
# the sparse and hierarchical models don't evaluate properly.
# CFA_sparse <- lavaan::cfa(sparse_syn, data = CFA_ind)

CFA_intuit <- lavaan::cfa(intuit_syn, data = CFA_ind)

CFA_EFA <- lavaan::cfa(EFA_mod_syn, data = CFA_ind)

# CFA_EFA_hier <- lavaan::cfa(EFA_hier_syn, data = CFA_ind)

CFA_pca <- lavaan::cfa(pca_syn, data = CFA_ind)

CFA_sparseMod <- lavaan::cfa(sparseMod_syn, data = CFA_ind)
```

Evalutate and Compare Models

The model with the lowest BIC is the best-performing. Note that only the EFA and pca models were calculated in CFA without errors, so these results are speculative.

```
anova(CFA_EFA, CFA_pca)
```

	Df	AIC	BIC	Chisq	Chisq diff	Df diff	Pr(>Chisq)
CFA_EFA	177	30492.73	30730.61	1919.685	NA	NA	NA
CFA_pca	221	33508.56	33750.85	2102.699	183.0145	44	0

```
anova(CFA_EFA, CFA_intuit)
```

	Df	AIC	BIC	Chisq	Chisq diff	Df diff	Pr(>Chisq)
CFA_EFA	177	30492.73	30730.61	1919.685	NA	NA	NA
CFA_intuit	319	41502.29	41762.20	3936.959	2017.274	142	0

```
anova(CFA_EFA, CFA_sparseMod)
```

	Df	AIC	BIC	Chisq	Chisq diff	Df diff	Pr(>Chisq)
CFA_sparseMod	78	21352.69	21537.71	1365.511	NA	NA	NA
CFA_EFA	177	30492.73	30730.61	1919.685	554.1738	99	0

The non-negative PCA model explains slightly less of the indicator variance than the standard pca model.

```
# pca variance
sum(peav(inds, sparse.pca.result$rotation))
```

```
## [1] 0.4707907
```

```
# non-negative pca variance
pca.result$Vaccounted
```

```
##                            TC1       TC2        TC4        TC3
## SS loadings           4.864050 3.7081201 2.67658144 2.28946138
## Proportion Var        0.180150 0.1373378 0.09913265 0.08479487
## Cumulative Var        0.180150 0.3174878 0.41662044 0.50141530
## Proportion Explained  0.359283 0.2739003 0.19770567 0.16911105
## Cumulative Proportion 0.359283 0.6331833 0.83088895 1.00000000
```