HW4

Arun Pandian

11/10/2019

Factor Analysis

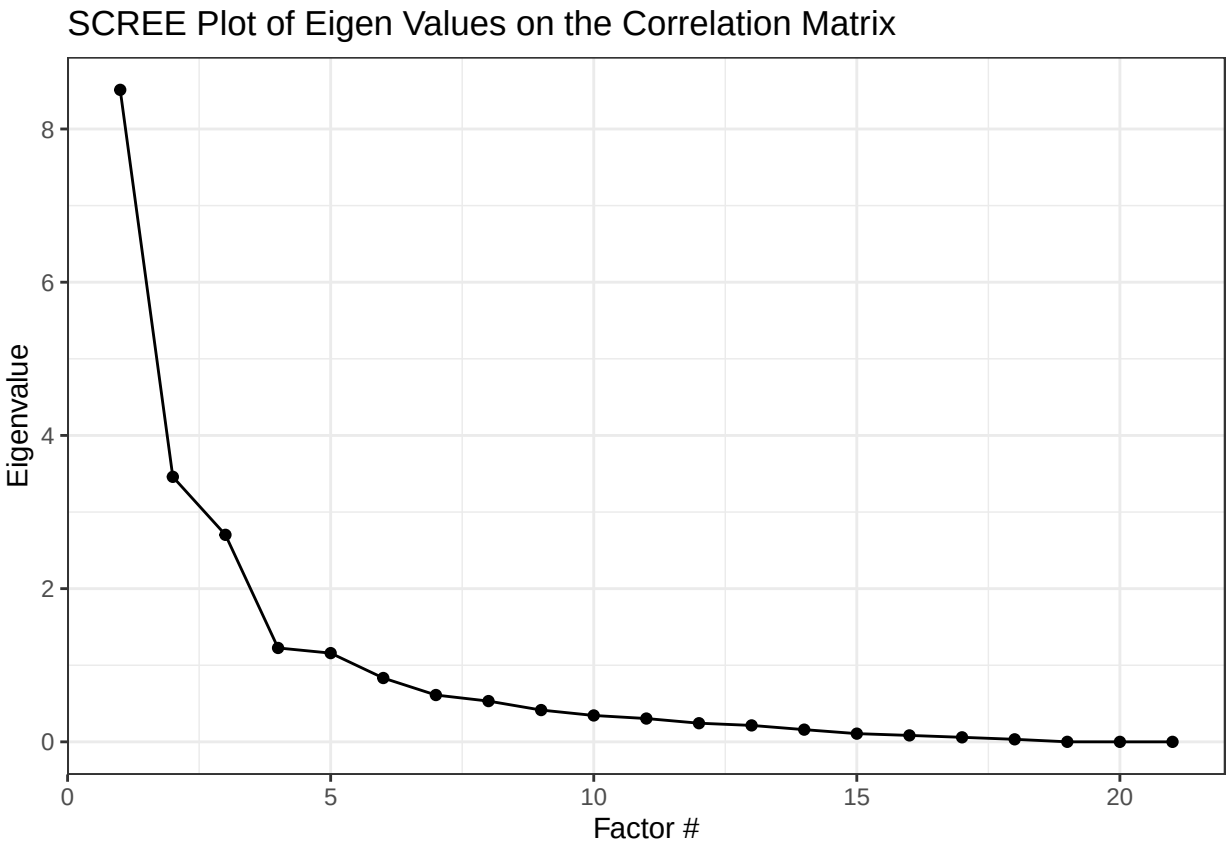
1.

Exploratory factor analysis is used to figure out how many latent dimensions are needed to account for (sometimes most of the) co-variation amongst the input features of the population based on sampling. We do not use domain knowledge to guess in advance how many factors are necessary.

Confirmatory factor analysis is used to confirm the number of factors in accordance to domain knowledge/existing theory. For instance, check if two factors can account for most of the co-variance in the input features because the theory suggests two hidden factors could account for the variation in the data.

2.

```
## In factor.scores, the correlation matrix is singular, an approximation is used
## In factor.scores, the correlation matrix is singular, an approximation is used
## In factor.scores, the correlation matrix is singular, an approximation is used
```



The elbow appears at 4 factors. Suggesting we should keep around 3 or 4 factors. The numerical evaluations below might shed more light.

```
## Loading of two factors:
```

```
##
## Loadings:
##             MR1    MR2
## idealpoint   0.726 -0.119
## polity       0.901  0.402
## polity2      0.901  0.402
## democ        0.928  0.295
## autoc       -0.777 -0.485
## unreg        0.285  0.250
## physint      0.610 -0.518
## speech       0.694  0.140
## new_empinx   0.883  0.176
## wecon        0.444 -0.318
## wopol        0.458  0.261
## wosoc        0.626 -0.230
## elecsd       0.824  0.293
## gdp.pc.wdi   0.542 -0.439
## gdp.pc.un    0.532 -0.441
## pop.wdi     -0.159  0.405
## amnesty     -0.564  0.576
## statedept   -0.671  0.559
## milper      -0.195  0.382
## cinc                0.333
## domestic9           0.440
##
##                  MR1   MR2
## SS loadings    8.229 2.994
## Proportion Var 0.392 0.143
## Cumulative Var 0.392 0.534
```

```
## Loading of three factors:
```

```
##
## Loadings:
##             MR1    MR2    MR3
## idealpoint   0.726          0.162
## polity       0.898  0.366 -0.189
## polity2      0.898  0.366 -0.189
## democ        0.925  0.292
## autoc       -0.778 -0.417  0.319
## unreg        0.283  0.216 -0.139
## physint      0.610 -0.434  0.260
## speech       0.693  0.120 -0.108
## new_empinx   0.884  0.135 -0.196
## wecon        0.445 -0.260  0.213
## wopol        0.456  0.236 -0.132
## wosoc        0.627 -0.158  0.238
## elecsd       0.822  0.263 -0.163
## gdp.pc.wdi   0.558 -0.319  0.543
## gdp.pc.un    0.547 -0.322  0.543
## pop.wdi     -0.176  0.675  0.572
## amnesty     -0.563  0.517 -0.186
## statedept   -0.671  0.468 -0.285
## milper      -0.217  0.680  0.639
## cinc                0.662  0.734
```

```
## domestic9          0.373 -0.214
##
##                  MR1   MR2   MR3
## SS loadings    8.258 3.203 2.512
## Proportion Var 0.393 0.153 0.120
## Cumulative Var 0.393 0.546 0.665
```

```
## Loading of four factors:
```

```
##
## Loadings:
##             MR1    MR2    MR3    MR4
## idealpoint  0.725          0.143
## polity      0.893  0.383 -0.170
## polity2     0.893  0.383 -0.170
## democ       0.922  0.304
## autoc      -0.773 -0.439  0.293
## unreg       0.282  0.226 -0.125
## physint     0.620 -0.452  0.221 -0.319
## speech      0.691  0.132 -0.106
## new_empinx  0.881  0.155 -0.195
## wecon       0.447 -0.263  0.184  0.101
## wopol       0.453  0.249 -0.124
## wosoc       0.627 -0.160  0.211
## elecsd      0.819  0.278 -0.151
## gdp.pc.wdi  0.579 -0.371  0.567  0.424
## gdp.pc.un   0.568 -0.373  0.565  0.419
## pop.wdi    -0.177  0.641  0.619 -0.158
## amnesty    -0.567  0.520 -0.137  0.177
## statedept  -0.681  0.487 -0.242  0.307
## milper     -0.218  0.639  0.684 -0.133
## cinc                0.610  0.767
## domestic9           0.408 -0.193  0.459
##
##                  MR1   MR2   MR3   MR4
## SS loadings    8.290 3.255 2.591 0.873
## Proportion Var 0.395 0.155 0.123 0.042
## Cumulative Var 0.395 0.550 0.673 0.715
```

In the two factor case, a input dimensions polity and the seemingly correlated polity2, as well as the democ(Institutionalized Democracy), new_empinx(CIRI Empowerment Rights Index), elecsd(CIRI Electoral Self determination) dimensions and speech(CIRI freedom of speech score) load onto component 1. The first component loading seems to capture measures of extent of democracy. Wecon(CIRI Women's Economic Rights), wesoc(CIRI Women's Social Rights), GDP.pc.wdi(GDP per capita: World Development Bank indicators), GDP.pc.un(GDP per capita: UN data) as well as physint (CIRI physical integrity score) load onto component 2 (The components are unordered, numbering is done for clarity). The second component loading consist of two groups, it seems like. Once is Women's Economic/Social Rights and the other is the two GDP measures which are probably highly correlated within themselves and possibly across each other. We see that about 52.6 of the cumulative variance captured by these two dimensions.

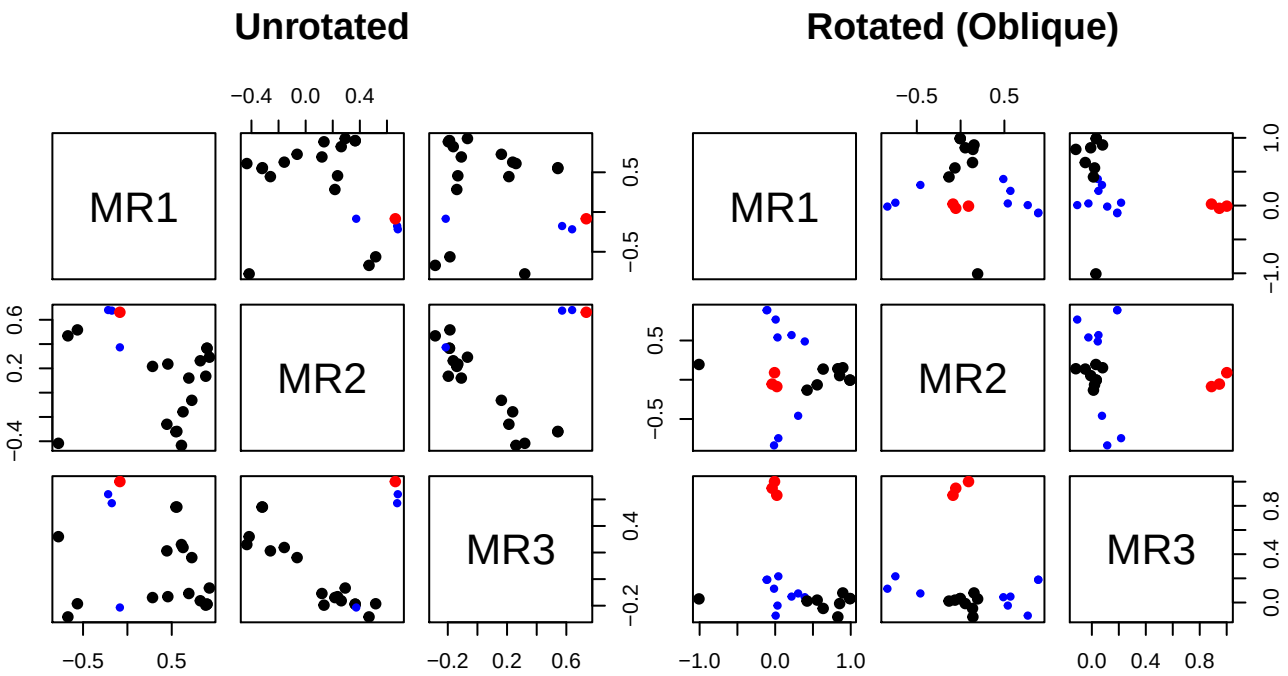
Moving to three factors, makes the GDP per capita, and two military (size and capability) dimensions load on the third factor, while preserving the factor structure otherwise. The cumulative variance explained is now about 65 percent. This addition of a third factor seems justified.

Moving to four factors, the factor structure of the three factor is more of less preserved and we get amnesty, statedept and domestic (domestic conflict/stability index) load on the fourth factor. Since some of the loading

onto factors decrease from the two factor case, the cumulative variance only goes up by about 2 percent from the 3 factor case. This may suggest, we don't need a fourth factor.

3.

```
## In factor.scores, the correlation matrix is singular, an approximation is used
```




```
##
##  Loading of three unrotated factors
```

```
##
## Loadings:
##             MR1    MR2    MR3
## idealpoint  0.726          0.162
## polity      0.898  0.366 -0.189
## polity2     0.898  0.366 -0.189
## democ       0.925  0.292
## autoc      -0.778 -0.417  0.319
## unreg       0.283  0.216 -0.139
## physint     0.610 -0.434  0.260
## speech      0.693  0.120 -0.108
## new_empinx  0.884  0.135 -0.196
## wecon       0.445 -0.260  0.213
## wopol       0.456  0.236 -0.132
## wosoc       0.627 -0.158  0.238
## elecsd      0.822  0.263 -0.163
## gdp.pc.wdi  0.558 -0.319  0.543
## gdp.pc.un   0.547 -0.322  0.543
## pop.wdi    -0.176  0.675  0.572
## amnesty    -0.563  0.517 -0.186
## statedept  -0.671  0.468 -0.285
## milper     -0.217  0.680  0.639
## cinc                0.662  0.734
## domestic9           0.373 -0.214
```

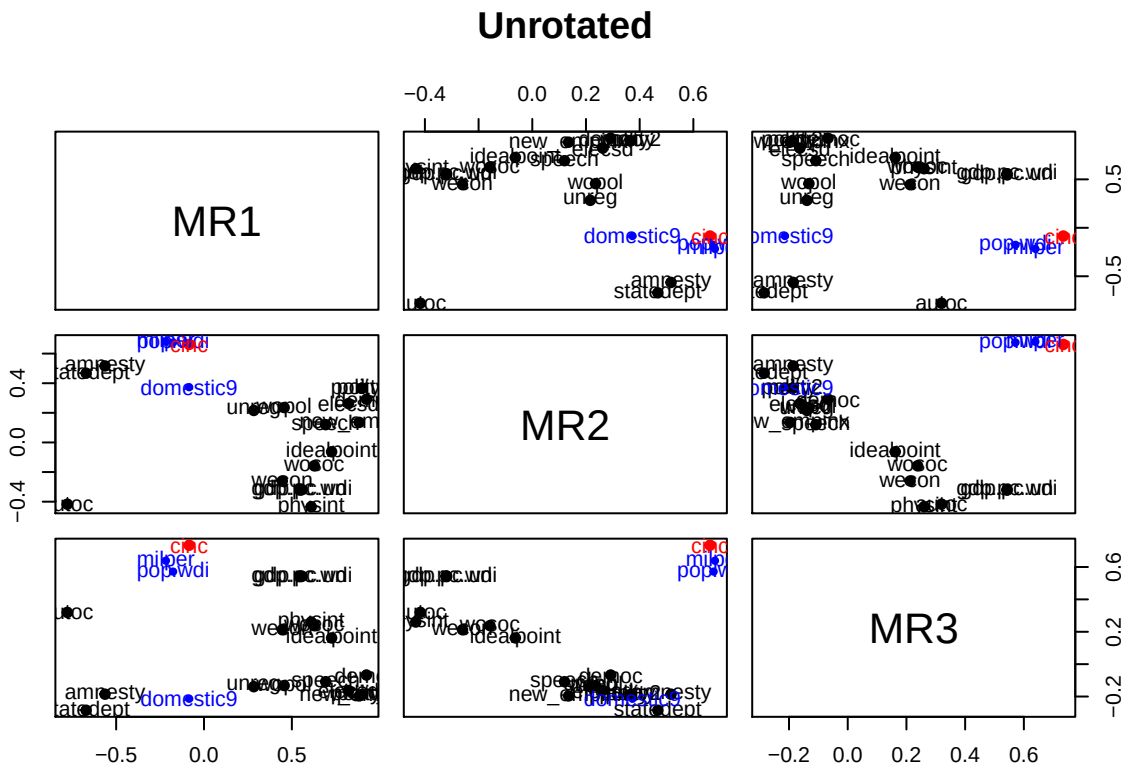
```
##
##                  MR1   MR2   MR3
## SS loadings    8.258 3.203 2.512
## Proportion Var 0.393 0.153 0.120
## Cumulative Var 0.393 0.546 0.665
```

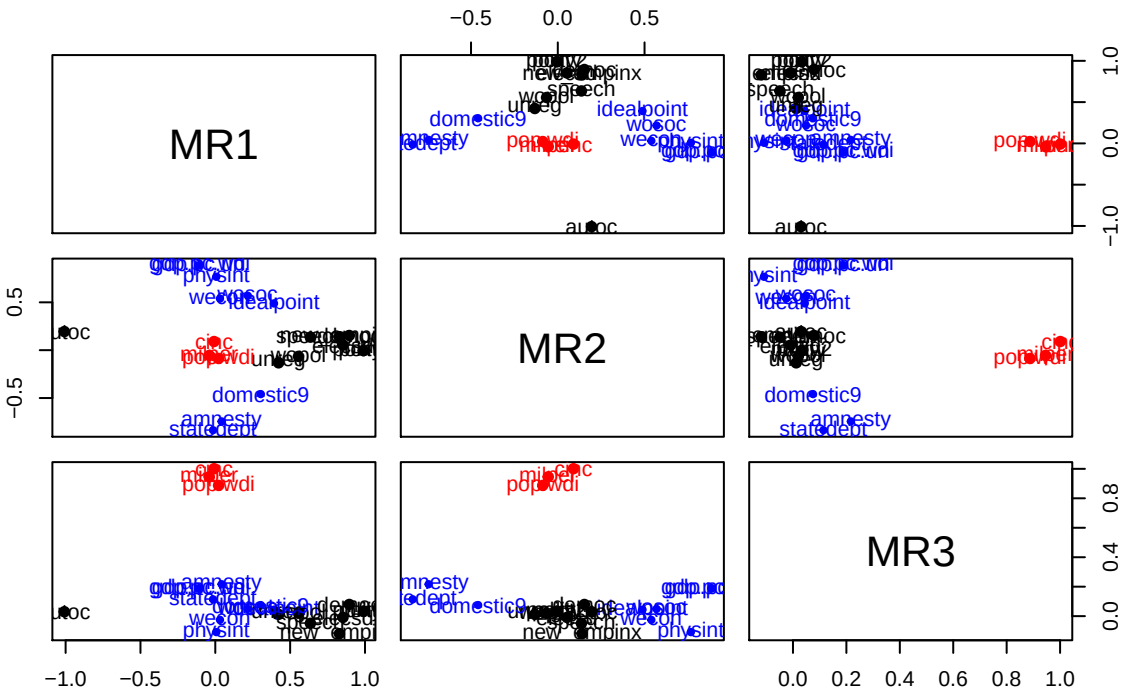
```
##
##  Loading of three obliquely rotated factors
```

```
##
## Loadings:
##             MR1    MR2    MR3
## idealpoint   0.392  0.488
## polity       0.990
## polity2      0.990
## democ        0.895  0.153
## autoc       -1.008  0.195
## unreg        0.423 -0.132
## physint             0.767 -0.108
## speech       0.635  0.136
## new_empinx   0.829  0.140 -0.119
## wecon               0.539
## wopol        0.556
## wosoc        0.217  0.569
## elecsd       0.853
## gdp.pc.wdi -0.104  0.888  0.189
## gdp.pc.un  -0.113  0.885  0.188
## pop.wdi                   0.887
## amnesty           -0.744  0.217
## statedept         -0.836  0.114
## milper                    0.946
## cinc                      1.001
## domestic9    0.304 -0.459
##
##                  MR1   MR2   MR3
## SS loadings    6.407 4.620 2.863
## Proportion Var 0.305 0.220 0.136
## Cumulative Var 0.305 0.525 0.661
```

The side by side versions can't be labelled legibly. Therefore, I have plotted the two plots one by one.

```
## [1] "......................................."
```





Since in the rotated version, we do not have orthogonal axes, the factor structure is slightly different. The total cumulative variance of three factors vary slightly. Also the ideal point loading is much less clearer in the rotated version (as opposed to clearly loading onto factor 1 in the unrotated version). Visually the rotated loading seem to be clearer overall along M3.

Overall three factors seems to cause the same amount of variation in the data in both the rotated and non-rotated cases.

Principal Component Analysis

1. In Factor Analysis, we are trying to find the Gaussian causal latent factors that can explain the co variance and variance in the data.

In PCA on the other hand, the components are just weighted combination of the variables that accounts for as much as observed variance as possible. That is C(1) = L(1).Dim(1)+....+L(n).Dim(n). There is no concept of a error component. PCA is more commonly used for dimensionality reduction, due to the lack of distributional assumptions.

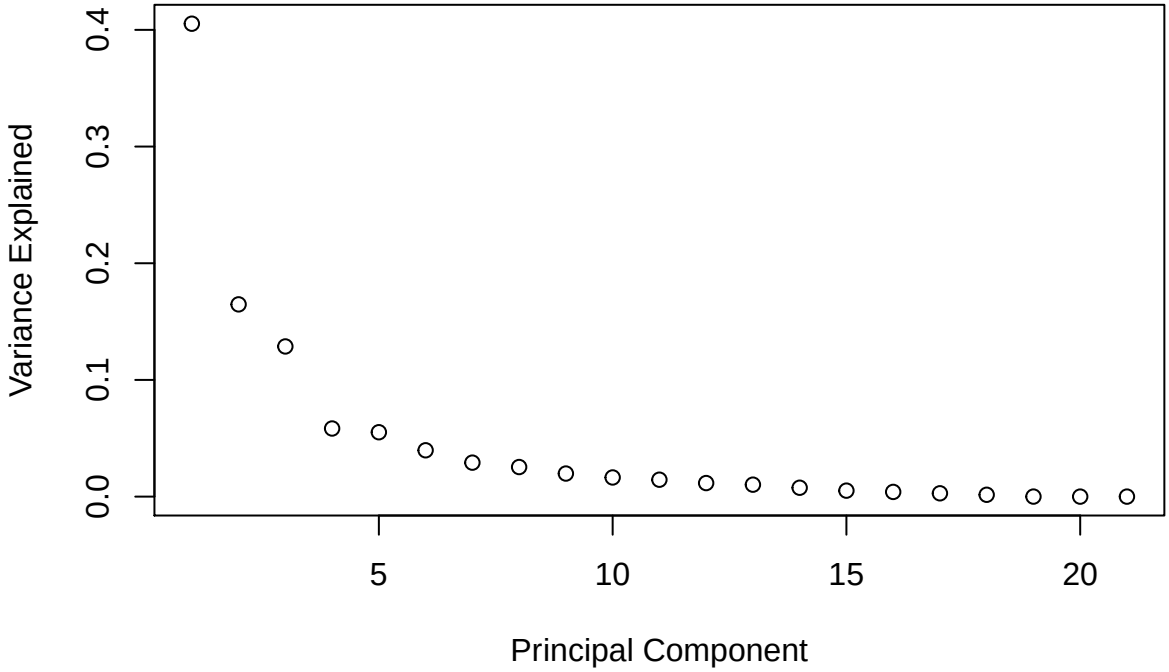
2.

```
## Warning: In prcomp.default(data, graph = FALSE) :
##  extra argument 'graph' will be disregarded
```

```
## Importance of components:
##                            PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     2.9173 1.8600 1.6439 1.10713 1.07631 0.91289
## Proportion of Variance 0.4053 0.1648 0.1287 0.05837 0.05516 0.03968
## Cumulative Proportion  0.4053 0.5700 0.6987 0.75708 0.81225 0.85193
##                            PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.78181 0.72948 0.64421 0.58703 0.55164 0.49341
## Proportion of Variance 0.02911 0.02534 0.01976 0.01641 0.01449 0.01159
## Cumulative Proportion  0.88104 0.90638 0.92614 0.94255 0.95704 0.96864
```

```
##                            PC13   PC14    PC15    PC16    PC17    PC18
## Standard deviation     0.46337 0.3995 0.32765 0.29011 0.24347 0.18215
## Proportion of Variance 0.01022 0.0076 0.00511 0.00401 0.00282 0.00158
## Cumulative Proportion  0.97886 0.9865 0.99157 0.99558 0.99840 0.99998
##                            PC19      PC20      PC21
## Standard deviation     0.01990 7.605e-16 2.858e-16
## Proportion of Variance 0.00002 0.000e+00 0.000e+00
## Cumulative Proportion  1.00000 1.000e+00 1.000e+00
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2901  0.5516  0.7253  0.9129  2.9173
```




```
##
## Proportions of variance explained by first 10 components:
```

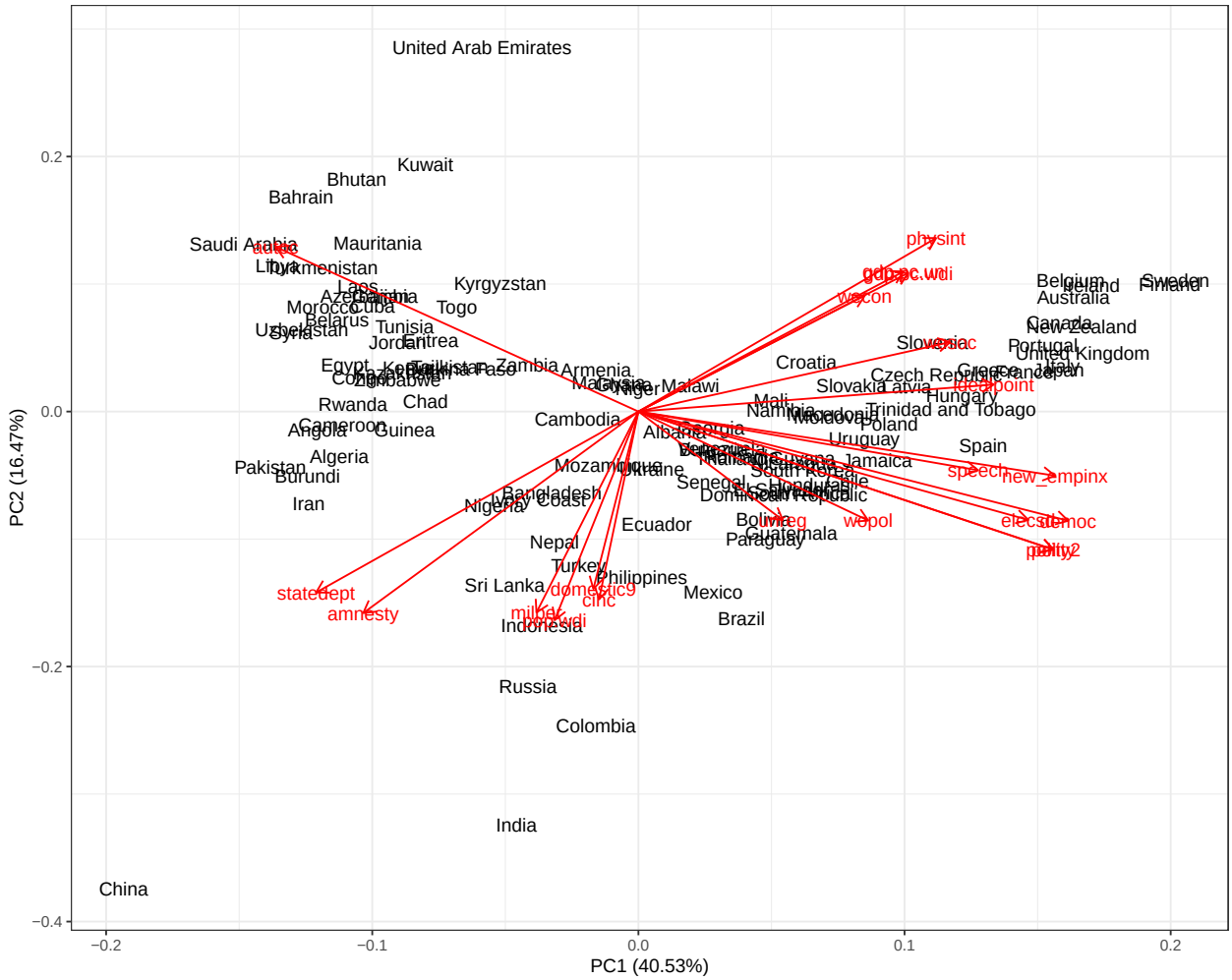
```
##  [1] 0.4053 0.1647 0.1287 0.0584 0.0552 0.0397 0.0291 0.0253 0.0198 0.0164
```

```
## Cumilative variances of first ten components:
```

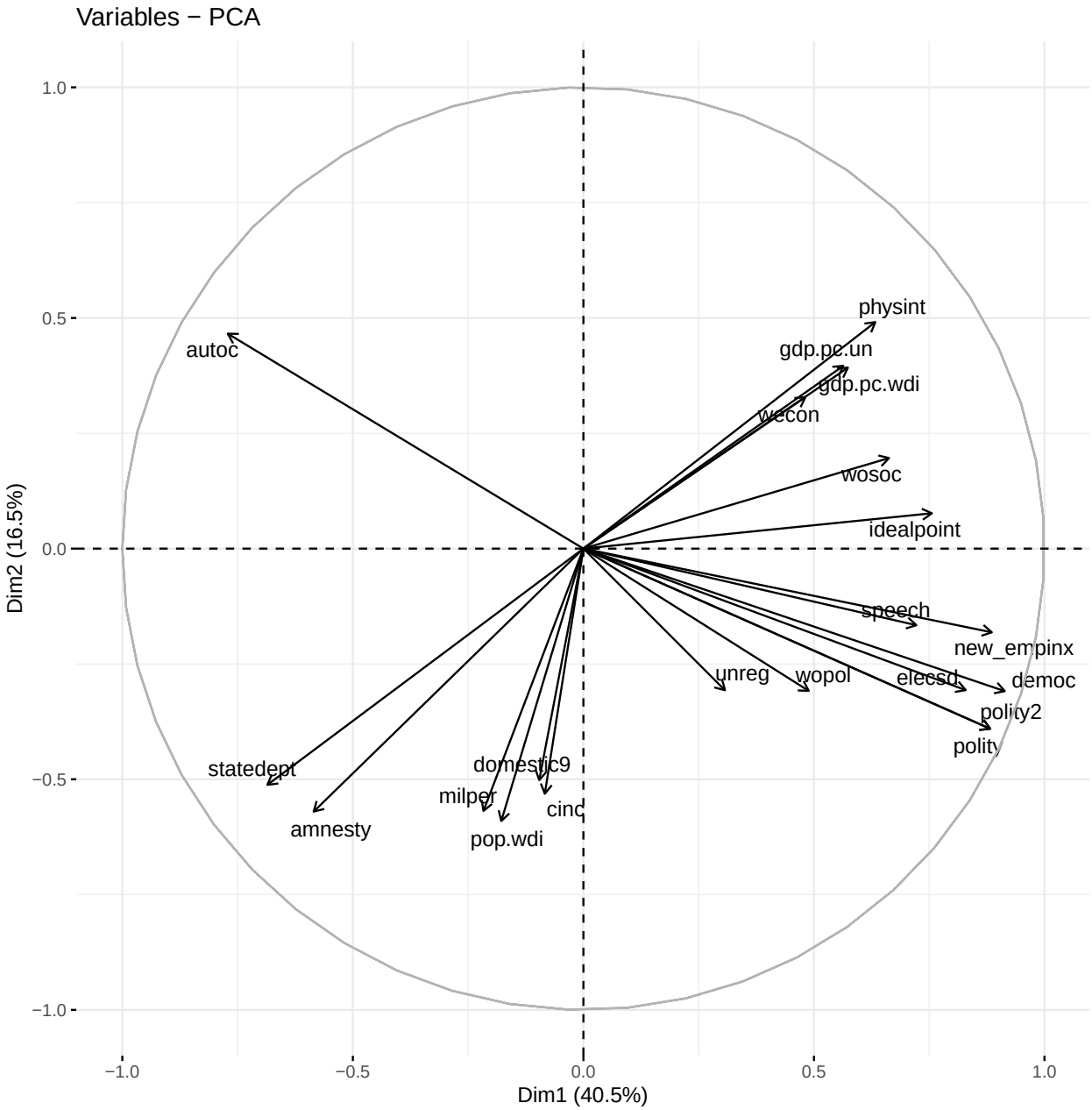
```
##  [1] 0.4053 0.5700 0.6987 0.7571 0.8123 0.8520 0.8811 0.9064 0.9262 0.9426
```

There is an elbow around n=7 in the scree plot. But looking at the fact that the proportion of variance captured by component 7 and beyond goes down relative to n<=6, I think 6 components capture the data well. 10 components capture 94 percent of the variation in the data. The first component accounts for 40 percent of the variation in the data.

3.



There seems to be some clustering of countries. The countries in the GCC seems to be clustered on the top left along the autocratic dimension which makes sense. The countries in the European union seem to be clustered along positive scores for women's rights and economic markers which also makes sense. Third world countries which are functionally democratic(Brazil, Bolivia, Paraguay, Ecuador) are clustered in the below the origin point but the reason is not clear. Looking at only the input dimensions and how they are distributed:



Autocracy lies along the upper left, economic and social rights are distributed in the upper right and democratic indicators are distributed in the lower right of the plot. This may explain the clustering of countries.

Bonus

Sparse PCA

```
## [1] "Iteration:    1, Objective: 6.75639e+01, Relative improvement Inf"
## [1] "Iteration:   11, Objective: 6.75532e+01, Relative improvement 1.05847e-05"
```

```
##                            PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8
## Explained variance       8.507 3.455 2.699 1.223 1.156 0.831 0.609 0.530
## Standard deviations      2.917 1.859 1.643 1.106 1.075 0.912 0.780 0.728
## Proportion of variance 0.405 0.165 0.129 0.058 0.055 0.040 0.029 0.025
```

```
## Cumulative proportion  0.405 0.570 0.698 0.756 0.811 0.851 0.880 0.905
##                           PC9  PC10
## Explained variance     0.413 0.343
## Standard deviations    0.643 0.586
## Proportion of variance 0.020 0.016
## Cumulative proportion  0.925 0.941
```

10 components here seem to capture about the same amount of variance.