cao_problem_set_4

Steven Cao

11/11/2019

Part 1, Problem 1

Confirmatory Factor Analysis is a test of *whether* the specified number of latent variables can account (well enough) for the observed covariation among the indicator variables. This particular approach is good for testing theories/hypothesises about the nature of the data; e.g. if one has good reason to assume that a set of indicators are really being caused by two hidden variables, then one would perform a test (via CFA) to see if two variables would in fact capture enough of the covariance within the data.

Exploratory Factor Analysis, however, is more hypothetical in nature: it asks, "*suppose* there are some latent variables which are the true cause(s) of these observations - if there are 'this' many of such latent variables, how much of covariation in the observations would be accounted for and attributable to such latent variables?" In the sense that it poses hypothetical situations and explores what their outcomes would be like (e.g. what the factor loadings would be, which can in turn elucidate the structure of the data), it is exploratory.

Part 1, Problem 2

```
factorAnalysis_2factor$loadings
```

```
##
## Loadings:
##            MR1    MR2
## idealpoint  0.449  0.429
## polity      0.995
## polity2     0.995
## democ       0.931
## autoc      -0.969  0.159
## unreg       0.412 -0.131
## physint            0.782
## speech      0.631  0.154
## new_empinx  0.802  0.197
## wecon              0.509
## wopol       0.551
## wosoc       0.286  0.497
## elecsd      0.852
## gdp.pc.wdi         0.673
## gdp.pc.un          0.671
## pop.wdi     0.204 -0.476
## amnesty           -0.821
## statedept         -0.849
## milper      0.158 -0.468
## cinc        0.211 -0.366
## domestic9   0.288 -0.479
##
```

```
##                    MR1   MR2
## SS loadings    6.523 4.527
## Proportion Var 0.311 0.216
## Cumulative Var 0.311 0.526
```

```
factorAnalysis_3factor$loadings
```

```
##
## Loadings:
##            MR1    MR2    MR3
## idealpoint  0.432  0.468
## polity      0.992
## polity2     0.992
## democ       0.910  0.144
## autoc      -0.994  0.191
## unreg       0.413 -0.129
## physint            0.737 -0.136
## speech      0.646  0.128
## new_empinx  0.840  0.131 -0.125
## wecon              0.518
## wopol       0.552
## wosoc       0.263  0.547
## elecsd      0.858
## gdp.pc.wdi         0.856  0.158
## gdp.pc.un          0.853  0.157
## pop.wdi                   0.892
## amnesty           -0.715  0.243
## statedept         -0.803  0.144
## milper                    0.949
## cinc                      0.999
## domestic9   0.269 -0.443
##
##                    MR1   MR2   MR3
## SS loadings    6.466 4.275 2.881
## Proportion Var 0.308 0.204 0.137
## Cumulative Var 0.308 0.512 0.649
```

```
factorAnalysis_4factor$loadings
```

```
##
## Loadings:
##            MR1    MR3    MR4    MR2
## idealpoint  0.467         0.214 -0.294
## polity      0.995
## polity2     0.995
## democ       0.922         0.127
## autoc      -0.986         0.146
## unreg       0.405                0.165
## physint     0.119               -0.761
## speech      0.658               -0.109
## new_empinx  0.855               -0.145
## wecon       0.105         0.390 -0.170
## wopol       0.555
```

```
## wosoc       0.300         0.350 -0.239
## elecsd      0.865
## gdp.pc.wdi                0.986
## gdp.pc.un                 0.979
## pop.wdi           0.923
## amnesty           0.177 -0.197  0.602
## statedept  -0.137        -0.139  0.783
## milper            0.965
## cinc              0.981  0.111
## domestic9   0.247         0.204  0.757
##
##                   MR1   MR3   MR4   MR2
## SS loadings    6.605 2.811 2.426 2.370
## Proportion Var 0.315 0.134 0.116 0.113
## Cumulative Var 0.315 0.448 0.564 0.677
```

In all 3 cases (i.e. 2-factor, 3-factor, and 4-factor loads), the first factor tends to weigh in for the most variables. Particularly with this first factor, it contributes the most to variables like `polity`, `polity2`, `democ`, `autoc`, `speech`, `elecsd`, and others, intuitively suggesting that this factor measures political climate (furthermore, `polity` and `polity2` are quite similar, while `democ` and `autoc` are opposite of each other).

In the 3-factor case, the second factor loads most heavily on both gdp variables in addition to `amnesty` and `statedept`, which may hint at an underlying common relationship between them. The third factor loads most heavily on population-related variables such as `pop.wdr` and `cinc`.

As for the 4-factor case, these "specialties" associated with each factor are preserved, with the fourth factor (MR2) not weighing in as heavily for any given variable as the others (which have loading values above 0.9 for some variables). This suggests that a 3-factor model might be more powerful than a 4-factor model (i.e. that it can be adequately captured in terms of 3 latent variables) - this can be verified by the scree plot, which shows a sizable difference in eigenvalue between n=3 and n=4.

Part 1, Problem 3

```
## Warning in cor.smooth(R): Matrix was not positive definite, smoothing was
## done
```

```
## Warning in cor.smooth(R): Matrix was not positive definite, smoothing was
## done
```

```
## Warning in cor.smooth(R): Matrix was not positive definite, smoothing was
## done
```

```
## Warning in cor.smooth(r): Matrix was not positive definite, smoothing was
## done
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs
## = np.obs, : The estimated weights for the factor scores are probably
## incorrect. Try a different factor extraction method.
```

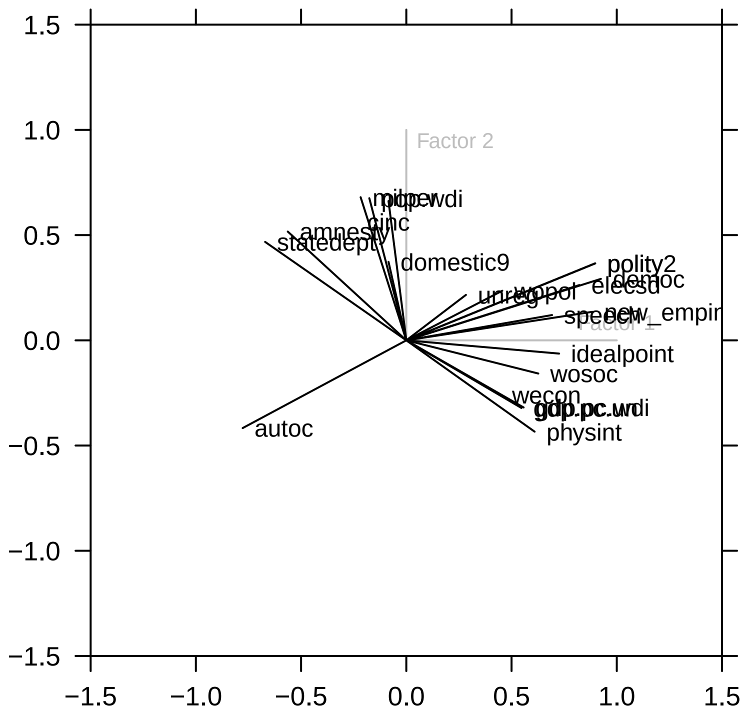
```
## In factor.scores, the correlation matrix is singular, an approximation is used
```

```
## Warning in cor.smooth(r): Matrix was not positive definite, smoothing was
## done
```

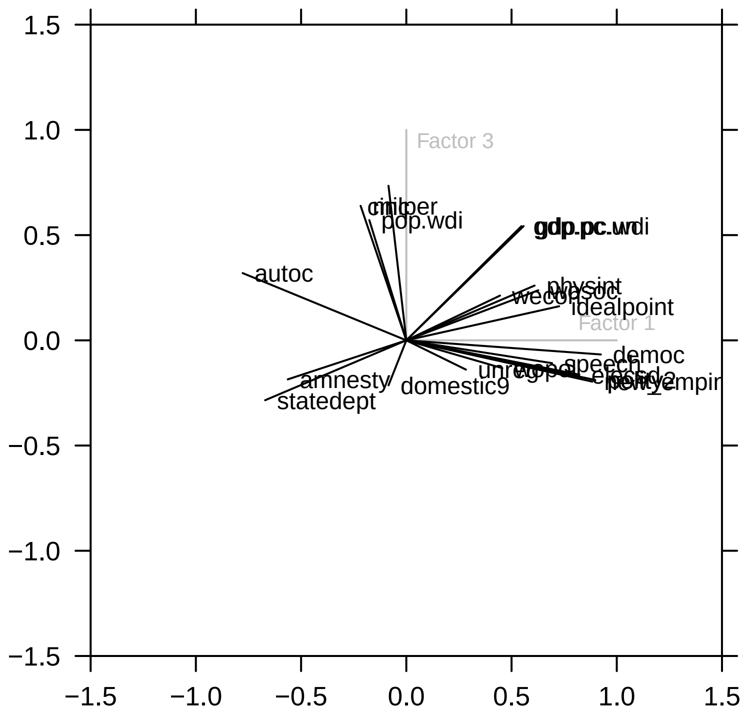
```
##
## Factor analysis with Call: fa(r = countries, nfactors = 3, rotate = "none")
##
## Test of the hypothesis that 3 factors are sufficient.
## The degrees of freedom for the model is 150  and the objective function was  46.65
## The number of observations was  107  with Chi Square =  4486.65  with prob <  0
##
## The root mean square of the residuals (RMSA) is  0.06
## The df corrected root mean square of the residuals is  0.07
##
## Tucker Lewis Index of factoring reliability =  0.06
## RMSEA index =  0.549  and the 10 % confidence intervals are  0.509 NA
## BIC =  3785.72
```

```
##
## 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
```

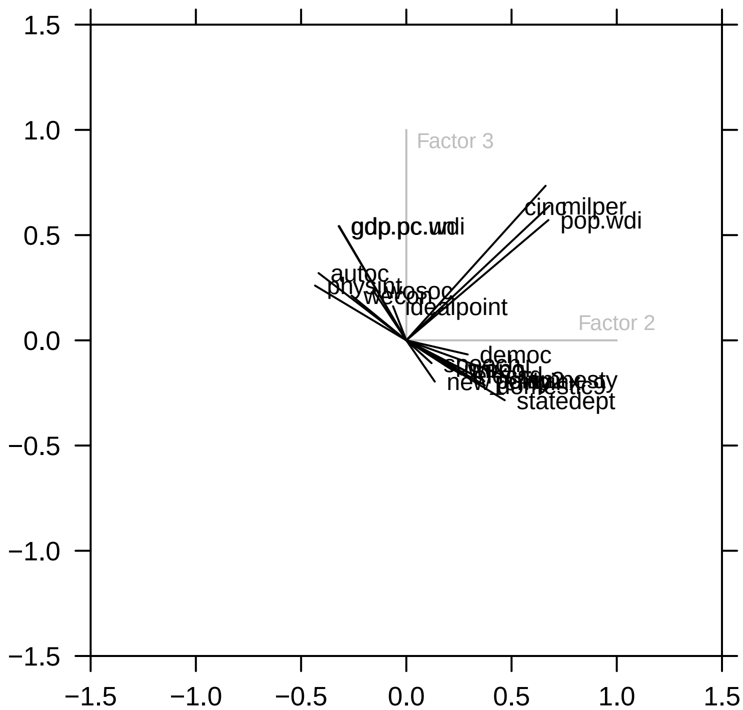
Non−Rotated Factor Pattern (2~1)



Non−Rotated Factor Pattern (3~1)



Non−Rotated Factor Pattern (3~2)

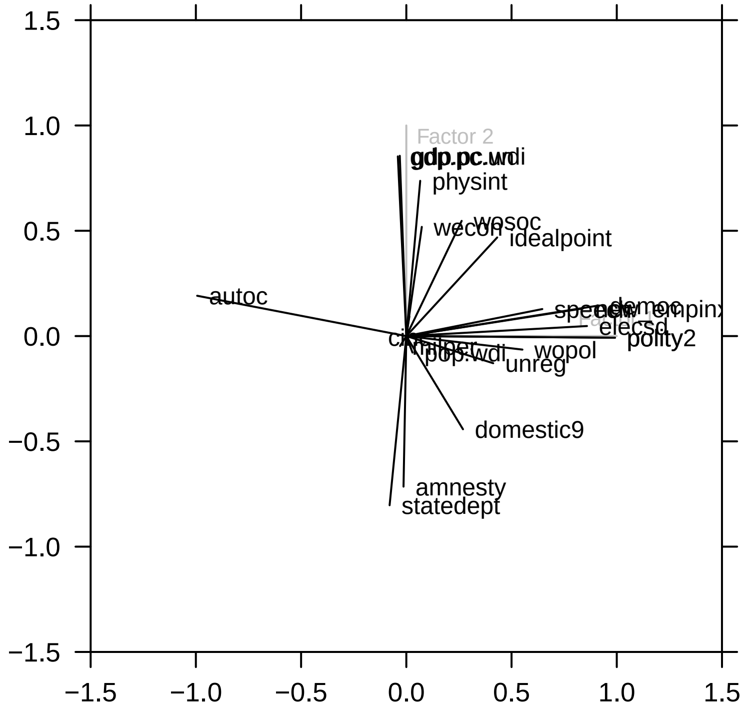


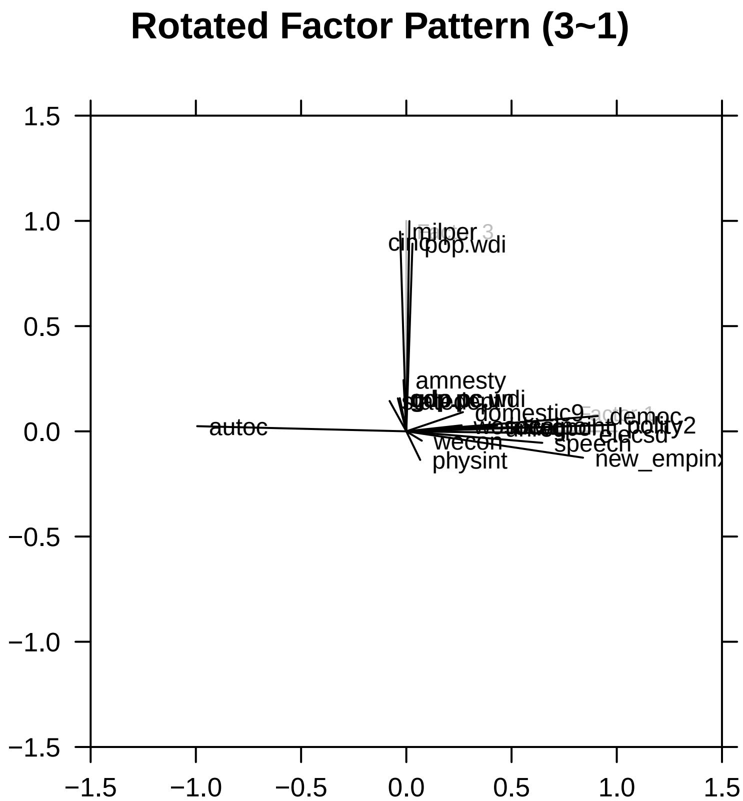

```
## Warning in cor.smooth(R): Matrix was not positive definite, smoothing was
## done

## Warning in cor.smooth(R): Matrix was not positive definite, smoothing was
## done

## Warning in cor.smooth(R): Matrix was not positive definite, smoothing was
## done

## Warning in cor.smooth(r): Matrix was not positive definite, smoothing was
## done

## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs
## = np.obs, : The estimated weights for the factor scores are probably
## incorrect. Try a different factor extraction method.

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

## Warning in cor.smooth(r): Matrix was not positive definite, smoothing was
## done

##
## Factor analysis with Call: fa(r = countries, nfactors = 3, rotate = "oblimin")
##
```

```
## Test of the hypothesis that 3 factors are sufficient.
## The degrees of freedom for the model is 150  and the objective function was  46.65
## The number of observations was  107  with Chi Square =  4486.65  with prob <  0
##
## The root mean square of the residuals (RMSA) is  0.06
## The df corrected root mean square of the residuals is  0.07
##
## Tucker Lewis Index of factoring reliability =  0.06
## RMSEA index =  0.549  and the 10 % confidence intervals are  0.509 NA
## BIC =  3785.72
##  With factor correlations of
##       MR1   MR2   MR3
## MR1  1.00  0.38 -0.05
## MR2  0.38  1.00 -0.12
## MR3 -0.05 -0.12  1.00
```

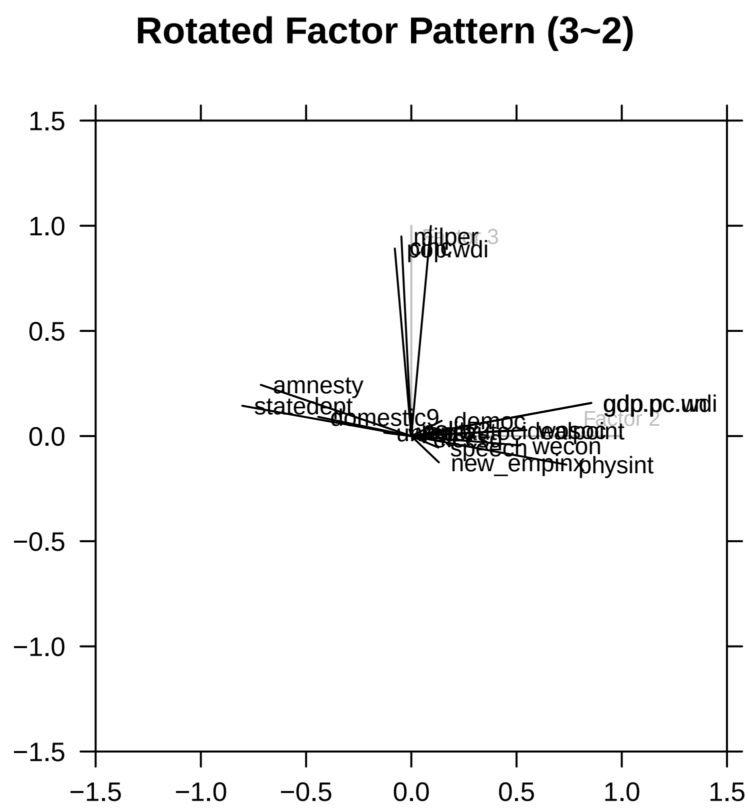
```
##
## Loadings:
##            MR1    MR2    MR3
## idealpoint  0.432  0.468
## polity      0.992
## polity2     0.992
## democ       0.910  0.144
## autoc      -0.994  0.191
## unreg       0.413 -0.129
## physint            0.737 -0.136
## speech      0.646  0.128
## new_empinx  0.840  0.131 -0.125
## wecon              0.518
## wopol       0.552
## wosoc       0.263  0.547
## elecsd      0.858
## gdp.pc.wdi         0.856  0.158
## gdp.pc.un          0.853  0.157
## pop.wdi                   0.892
## amnesty           -0.715  0.243
## statedept         -0.803  0.144
## milper                    0.949
## cinc                      0.999
## domestic9   0.269 -0.443
##
##                  MR1   MR2   MR3
## SS loadings    6.466 4.275 2.881
## Proportion Var 0.308 0.204 0.137
## Cumulative Var 0.308 0.512 0.649
```

Rotated Factor Pattern (2~1)





Rotated Factor Pattern (3~1)
milper
Factor 3
cinc
pop.wdi
amnesty
gdp.pc.wdi
domestic9
Factor 1
democ
polity2
elecsd
autoc
speech
wecon
new_empinx
physint
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5



Comparing the non-rotated factor patterns with their obliquely-rotated counterparts, the difference in the ease of interpretability is clear. With the non-rotated case, the various factors - while they tend to clump in the direction in which they point - are not clear in their relationship to the factors themselves. In the rotated case, however, the various indicators clump very tightly around the axis of a particular factor, which makes the relationship much more clear - namely, that the factors are doing a good job of accounting for the indicator variables.

Part 2, Problem 1

The main difference between Principal Components Analysis and Factor Analysis is the "direction of inference". Factor Analysis is intended to show what number of latent variables (and which ones) are *causing* the observations to have the values that they do. PCA, on the other hand, is more like curve-fitting, in the sense that it is just trying to grab the best coefficients that will account for as much of the covariation in the data as possible. This makes PCA much more atheoretical in nature than FA. The essence of this is described in the difference in semantics between the following two equations:

$$X_1 = b_1F + d_1U_1$$

$$Comp_1 = L_1X_1 + L_2X_2 + ... + L_kX_k$$

FA is also computed using the correlation matrix, whereas PCA is computing using the covariance matrix.

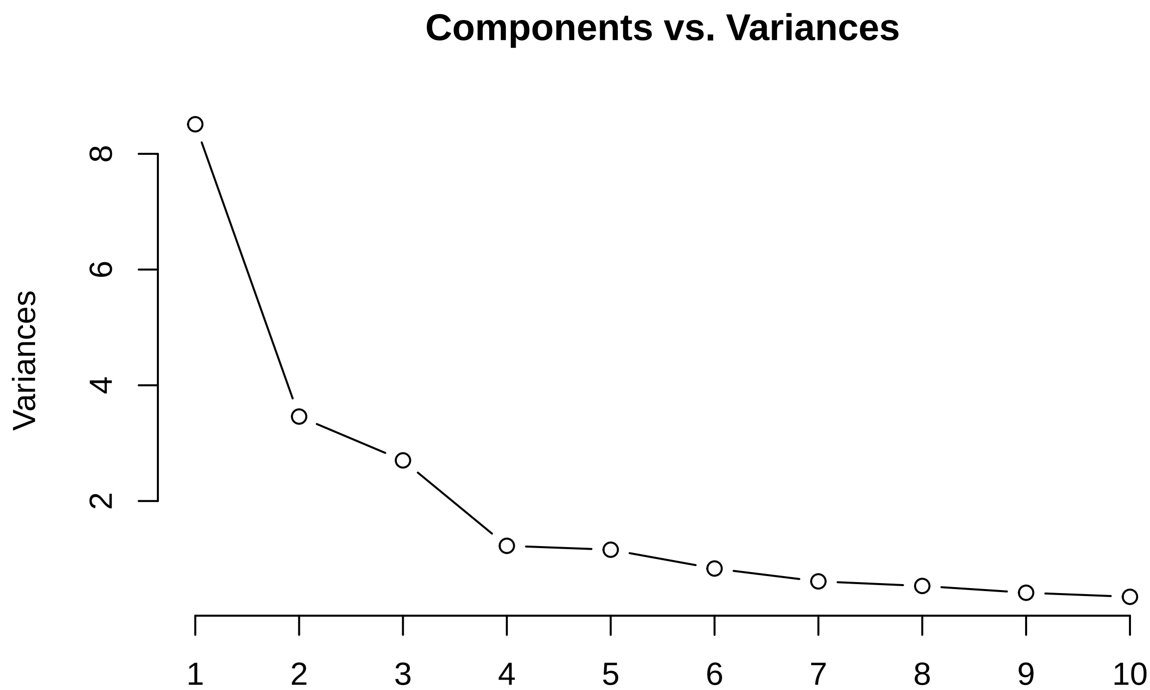
Part 2, Problem 2

```
## 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 8.602e-16 2.409e-16
## Proportion of Variance 0.00002 0.000e+00 0.000e+00
## Cumulative Proportion  1.00000 1.000e+00 1.000e+00
```

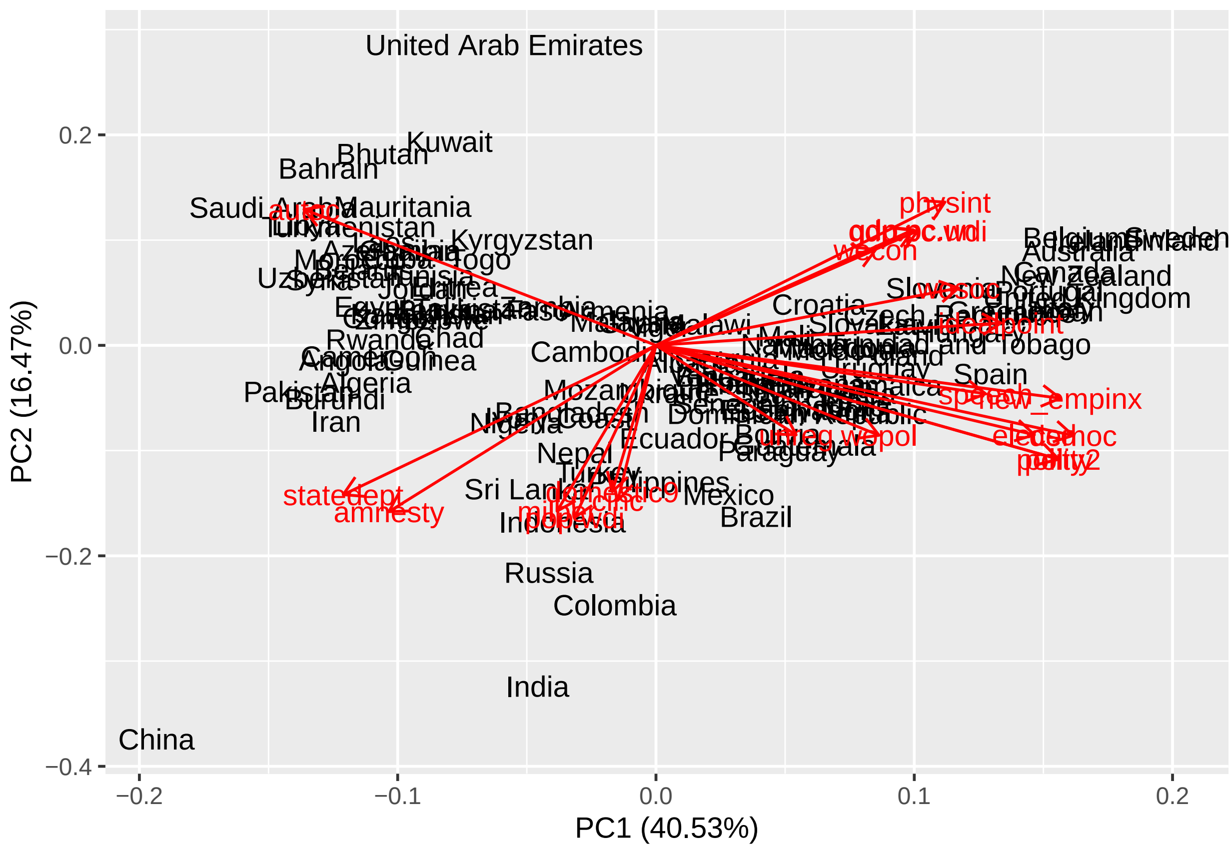
```
##                   PC1        PC2         PC3         PC4         PC5
## idealpoint  0.2590509  0.0411834 -0.09502442 -0.00940566 -0.04815128
## polity      0.3024428 -0.2099553  0.07988403  0.03264443 -0.01088367
## polity2     0.3024428 -0.2099553  0.07988403  0.03264443 -0.01088367
## democ       0.3132604 -0.1661158  0.01740663  0.01322431  0.04334737
## autoc      -0.2643837  0.2505050 -0.15504195 -0.05538732  0.08075072
```

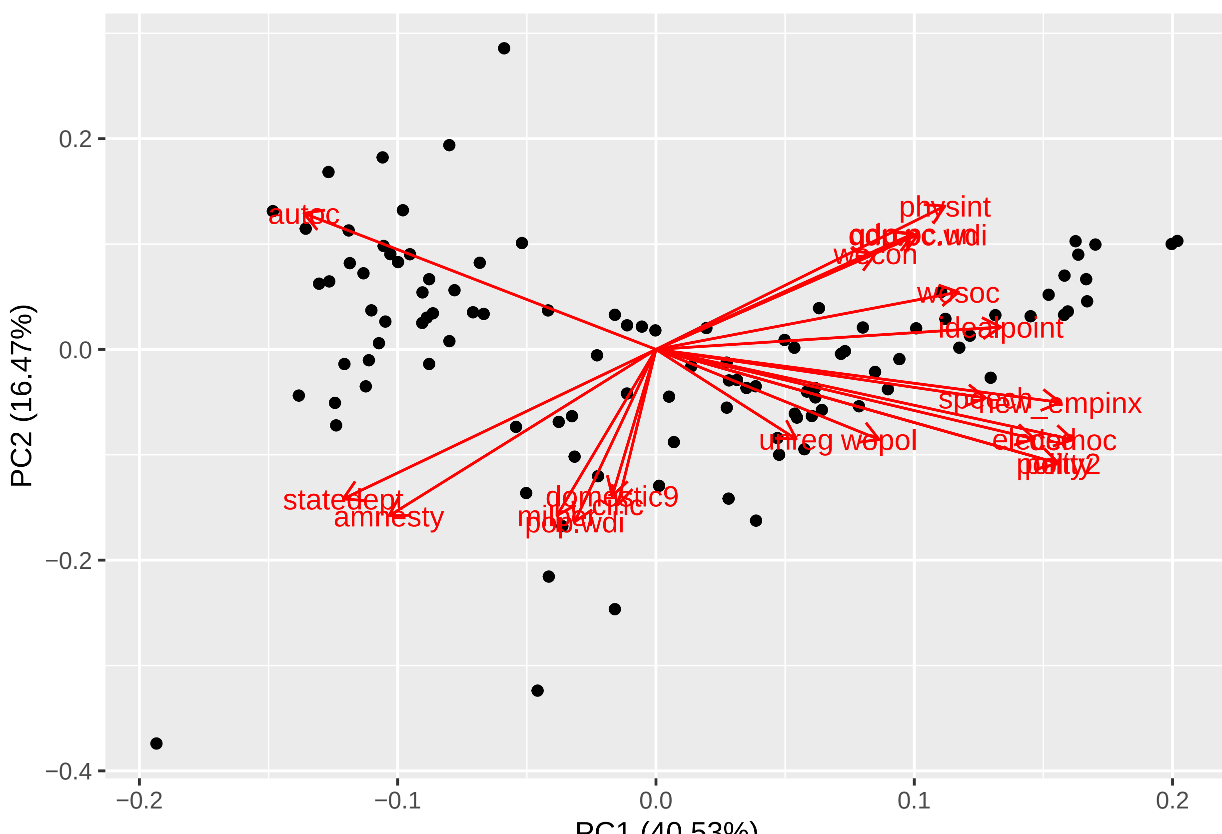
```
##                             PC1        PC2        PC3         PC4
## Angola               -3.6399214 -0.2672129  0.6179432 -1.15914878
## Albania               0.4130631 -0.3079309  0.7922239  0.44603842
## United Arab Emirates -1.7735942  5.4981470 -3.6187606 -0.03256528
## Armenia              -0.4803014  0.6320023  0.5323238  1.27958406
## Australia             4.9340546  1.7301635 -1.5588599 -0.60720269
##                              PC5
## Angola               -0.68409947
## Albania              -0.60670923
## United Arab Emirates  3.59482373
## Armenia               0.11368204
## Australia            -0.01285816
```



The scree plot suggests that, just like in the previous scree plot, 3 components/factors are adequate in capturing most of the variance in the data, i.e. that reducing to 3 dimensions results in minimal information loss. (It is also reassuring that the models for both the FA and PCA approaches are in agreement.)

Part 2, Problem 3





Middle-eastern countries tend to fall on the opposite spectrum from western (largely European) countries. Asian and South American countries tend to fall in the middle. The variables, `autoc` and `democ` seem to form a dichotomy which accounts for most of the clustering along the PC1 axis. The PC2 axis seems to resonate strongly with population-related variables, as we can see India and China extend far along the PC2 axis.