Arroyo - Problem Set 4

Pedro Alberto Arroyo

11/7/2019

Factor Analysis

1. How do CFA and EFA differ?

Factor analysis is a method for revealing the latent dimensionality of a feature space; in other words, the goal of factor analysis is to produce insight into the underlying causal factors that shape a distribution of observations.

Factor analysis can be exploratory (EFA) or confirmatory (CFA).

EFA is 'theory-light'. That is, the analysis does not proceed on the assumption that the researcher already has knowledge about what set of factors structure the data. There is, however, a causal *a priori assumption* that the number of factors is less than the number of measured variables.

CFA is hypothesis-testing; that is, the researcher sets out to establish whether a hypothesized set of factors in fact predict or model a set of observations.

---

2. Fit three exploratory factor analysis models initialized at 2, 3, and 4 factors. Present the loadings from these solutions and discuss in substantive terms. How does each fit? What sense does this give you of the underlying dimensionality of the space? And so on.

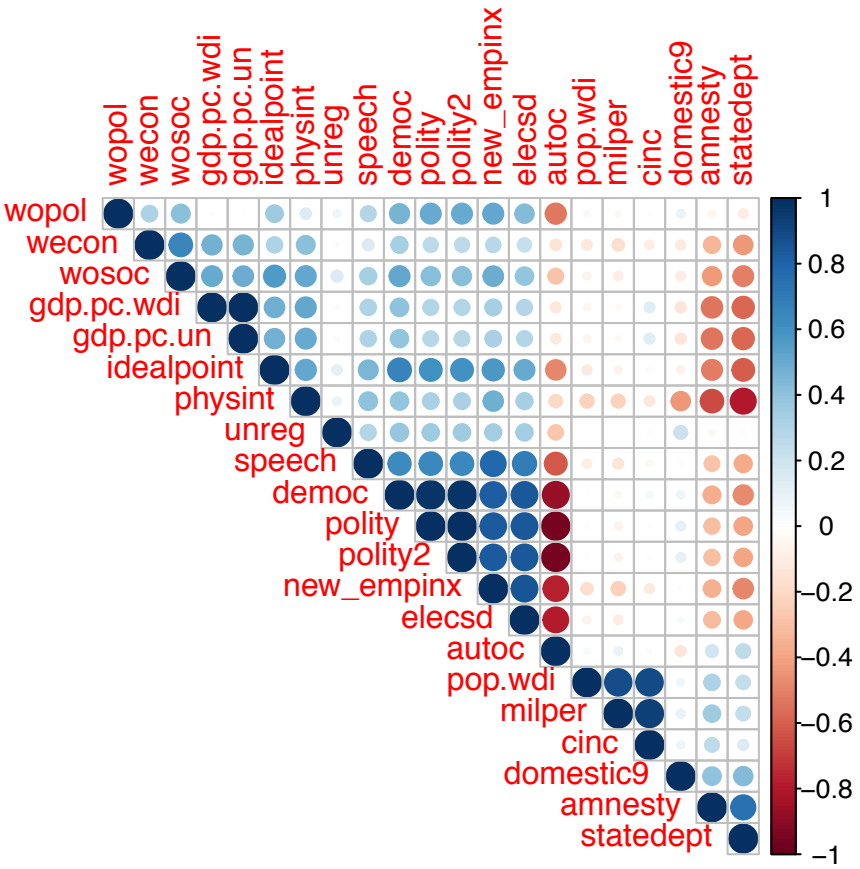
I began by creating a correlation matrix of the scaled values 'countries' and then fitting an EFA model initialized at 2, 3, and 4. That resulted in an error message:

In factor.scores, the correlation matrix is singular, an approximation is used... The estimated weights for the factor scores are probably incorrect. Try a different factor extraction method.

After some searching, it seems that the reason for this is high correlation between variables in the matrix.

To check this, I examined a correlation matrix of the variables.

```
countries <- read.csv("countries.csv", header = TRUE, row.names = 1)
corrplot(cor(scale(countries)), type = "upper", order = "hclust")
```



When two variables met the condition r > 0.8, I removed one of the two variables. In choosing which variable in the pair to remove, I opted to remove the variable that was most correlated with other variables in the set. I considered combining variables based on their scaled values. I did not do this because some variables were negatively correlated with each other, which made combination difficult. I reasoned that removal was still relevant since high negative correlation captures variation as well as high possitive correlation.

I'll admit that I remain a bit uncomfortable with this approach, but I justified it in the following way: the goal of EDA is to help the researcher detect latent structure and it is usually followed by a my structured data analysis. In the case of FA, interpretation is already a bit murky. Given that, I decided that removing noise, even if it also destroys data, might make it less likely that a researcher will be taken-in by a spurrious pattern. The flipside is to keep in mind that this approach might make a signal *clearer*, but also potentially *less* reliable. This then has to be factored into the subsequent analysis.

I also renamed variables to try to capture the choices that were made & aid with later interpretation. The FA loadings for the reduced dataset is below.

Variables

1. Polity, Polity2, democ, newempinex -> deleted in favor of elecsd (renamed SelfDeterm)
2. gdp.pc.wdi -> deleted in favor of gdp.pc.un (renamed GDP)
3. milper -> deleted in favor of cinc (renamed Military)
4. pop.wdi -> dropped from analysis (highly correlated with Military, not sure how to interpret)

```
c_sub1 <- countries %>% rename(GDP = gdp.pc.un, SelfDeterm = elecsd, Military = cinc, region = unreg,

c_sub <- scale(c_sub1[,-c(2,3,4,9,14,16,19)])

fa2 <- fa(c_sub, nfactors = 2)
```

```r
fa3 <- fa(c_sub, nfactors = 3)

fa4 <- fa(c_sub, nfactors = 4)

fa2$loadings
```

```
##
## Loadings:
##            MR1    MR2
## idealpoint  0.485  0.415
## autoc              -0.904
## region     -0.116  0.396
## physint     0.836
## speech      0.167  0.636
## weEcon      0.499
## wePol               0.544
## weSoc       0.532  0.293
## SelfDeterm          0.848
## GDP         0.636
## terrorAm   -0.809
## terrorST   -0.918
## Military   -0.143
## DomCon     -0.540  0.341
##
##                  MR1   MR2
## SS loadings    3.739 2.784
## Proportion Var 0.267 0.199
## Cumulative Var 0.267 0.466
```

```r
fa3$loadings
```

```
##
## Loadings:
##            MR1    MR2    MR3
## idealpoint -0.323  0.371  0.267
## autoc             -0.908
## region             0.400
## physint    -0.765         0.106
## speech     -0.213  0.700
## weEcon            -0.138  0.759
## wePol       0.268  0.404  0.415
## weSoc                     0.834
## SelfDeterm         0.867
## GDP        -0.428         0.337
## terrorAm    0.807
## terrorST    0.878
## Military    0.207         0.107
## DomCon      0.618  0.238  0.133
##
##                  MR1   MR2   MR3
## SS loadings    2.869 2.618 1.681
## Proportion Var 0.205 0.187 0.120
## Cumulative Var 0.205 0.392 0.512
```

```
fa4$loadings
```

```
##
## Loadings:
##            MR2    MR1    MR3    MR4
## idealpoint  0.382 -0.256  0.210  0.185
## autoc      -0.877                0.114
## region      0.422  0.149         0.119
## physint           -0.742  0.100
## speech      0.722 -0.147         0.106
## weEcon     -0.169 -0.116  0.703  0.128
## wePol       0.335  0.111  0.581 -0.266
## weSoc             -0.102  0.706  0.182
## SelfDeterm  0.866
## GDP                               0.876
## terrorAm           0.758
## terrorST   -0.110  0.820
## Military           0.359         0.304
## DomCon      0.237  0.706         0.185
##
##                  MR2   MR1   MR3   MR4
## SS loadings    2.591 2.578 1.422 1.109
## Proportion Var 0.185 0.184 0.102 0.079
## Cumulative Var 0.185 0.369 0.471 0.550
```

Discussion

The variance explained by the model increases with the number of factors included. This increase seems modest; from the 2-factor to the 3-factor model, we gain an additional ~10% of explained variance. My intuitive interpretation of this is that there are a few important factors driving the distribution but that there's also fair amount of other elements shaping it.

It's interesting to look at how the factors disaggregate as we increase the number of factors. The 2-factor model seems to indicate that there are two main drivers: political liberalism and physical safety. The 3-factor model keeps the physical safety dimension, but seems to seperate the other dimension into two: one that seems to capture political rights and another that seems to capture civil rights.

The 4-factor seems to provide more clarity. One factor seems to capture political rights, another seems to capture physical security, a third seems to capture women's rights, while a final factor seems to capture economic wellbeing.

My impressions is that, while the increase in factors provided for more of the variance being captured, another major benefit was that the factors seem to be easier to interpret.

---

3. Rotate the 3-factor solution using any oblique method you would like and present a visual of the unrotated and rotated versions side-by-side. How do these differ and why does this matter (or not)?

```
nonrotated.factors <- fa(cor(c_sub),
        fm = "pa",
        nfactors = 3,
        rotate = "none",
        residuals = TRUE)
```

```
nonrotated.factors$loadings
```

```
##
## Loadings:
##             PA1    PA2    PA3
## idealpoint   0.733
## autoc       -0.572 -0.663  0.140
## region       0.182  0.320
## physint      0.758 -0.349 -0.141
## speech       0.615  0.374 -0.223
## weEcon       0.552 -0.174  0.492
## wePol        0.385  0.392  0.263
## weSoc        0.735         0.497
## SelfDeterm   0.696  0.549 -0.164
## GDP          0.599 -0.249  0.107
## terrorAm    -0.698  0.383  0.210
## terrorST    -0.824  0.401  0.197
## Military             0.108  0.119
## DomCon      -0.237  0.472  0.201
##
##                  PA1   PA2   PA3
## SS loadings    4.915 1.900 0.829
## Proportion Var 0.351 0.136 0.059
## Cumulative Var 0.351 0.487 0.546
```

```
nonrot.pattern <- as.data.frame(nonrotated.factors$loadings[1:13,])

f1 <- xyplot(PA2 ~ PA1, data = nonrot.pattern,
       aspect = 1,
       xlim = c(-1, 1.3),
       ylim = c(-.7, .7),
       panel = function (x, y) {
 #        panel.segments(c(0, 0), c(0, 0),
 #            c(1, 0), c(0, 1), col = "gray")
         panel.text(1, 0, labels = "Initial\n(unrotated)\nfactor 1",
                    cex = .65, pos = 3, col = "gray")
         panel.text(0, .7, labels = "Initial\n(unrotated)\nfactor 2",
                    cex = .65, pos = 4, col = "gray")
         panel.segments(rep(0, 8), rep(0, 8), x, y,
            col = "black")
         panel.text(x[-7], y[-7], labels = rownames(nonrot.pattern)[-7],
            pos = 4, cex = .75)
         panel.text(x[7], y[7], labels = rownames(nonrot.pattern)[7],
                    pos = 1, cex = .75)
       },
       main = "Unrotated Factor Pattern",
       xlab = "",
       ylab = "",
#       scales = list(x = list(at = c(0, 1)),
#                     y = list(at = c(-.4, 0, .6)))
)

varimax.factors <- fa(cor(c_sub),
                      fm = "pa",
```

```
                      nfactors = 3,
                      rotate = "varimax")

orthog.pattern <- as.data.frame(varimax.factors$loadings[1:13,])

f2 <- xyplot(PA2 ~ PA1, data = orthog.pattern,
       aspect = 1,
       xlim = c(-.8, 1.3),
#       ylim = c(-.5, .8),
       panel = function (x, y) {
#         panel.segments(c(0, 0), c(0, 0),
#                        c(1, 0), c(0, 1), col = "gray")
         panel.text(1, 0, labels = "Rotated\nfactor 1",
                    cex = .65, pos = 3, col = "gray")
         panel.text(0, .95, labels = "Rotated\nfactor 2",
                    cex = .65, pos = 4, col = "gray")
         panel.segments(rep(0, 8), rep(0, 8), x, y,
                        col = "black")
         panel.text(x[-7], y[-7], labels = rownames(orthog.pattern)[-7],
                    pos = 4, cex = .75)
         panel.text(x[7], y[7], labels = rownames(orthog.pattern)[7],
                    pos = 1, cex = .75)
       },
       main = "Varimax Rotated Factor Pattern",
       xlab = "",
       ylab = "",
#       scales = list(x = list(at = c(0, 1)),
#                     y = list(at = c(-.4, 0, .6)))
)
```

Discussion

(Plots on Next Page)

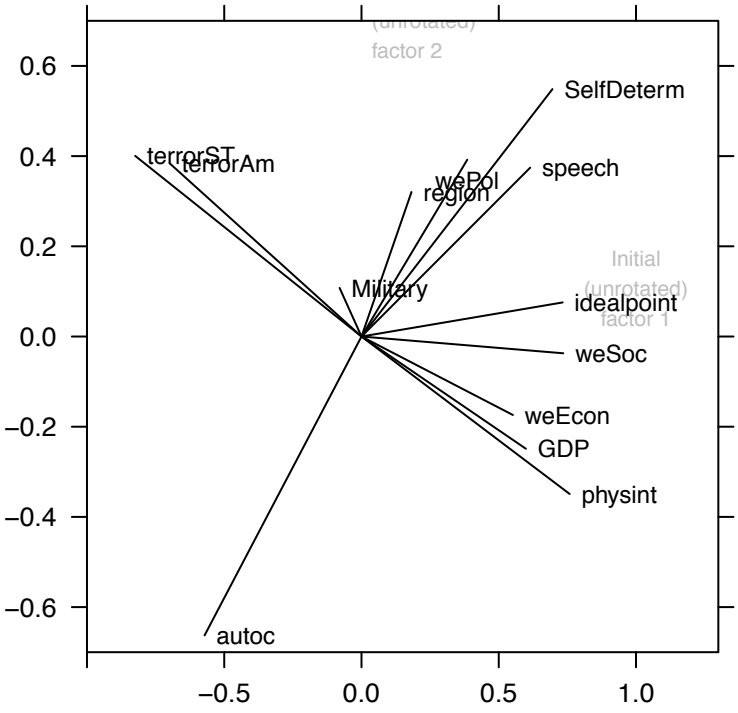
The comparison between the rotated and unrotated plots seems to illustrate a few things.

1. In both rotations, the autocracy variable is doing a lot of work.
2. Similarly, both terror variables seem to represent second clear dimension.
3. The physical integrity variable is important - but the rotated version seems to indicate a strong relationship with economic indicators. To me, that's a very interesting pattern and would make me want to look into ways of examining the relationship between personal physical safety and economic activity and/or property rights.
4. The fourth dimension is less organized in both plots.

The differences between the plots is idea-generating and, in that sense, useful. I am unsure that I would want to do much more with this sort of approach than generate hypotheses or illustrate data structure. Because of that, I find the differences between the rotated and un-rotated versions to be very useful.

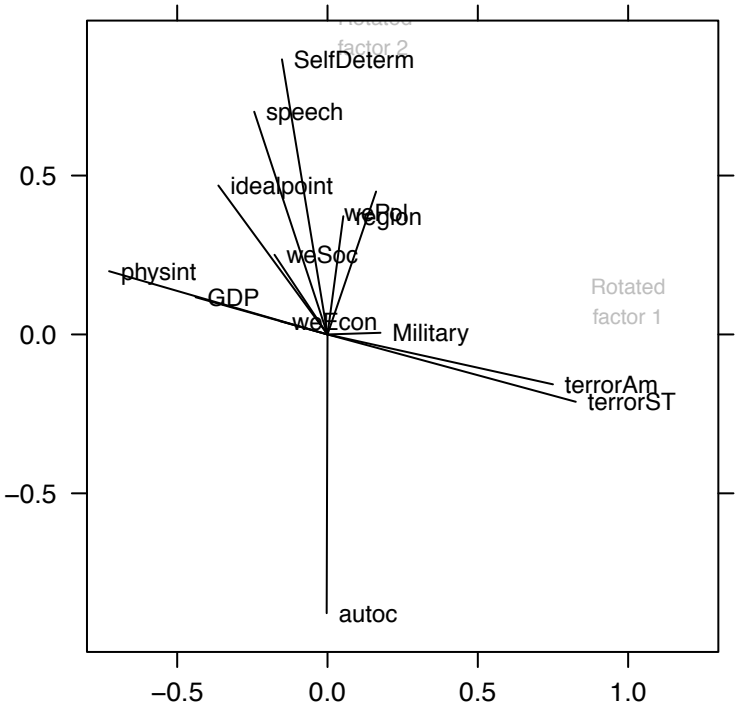
```
f1
```

Unrotated Factor Pattern




```
f2
```

Varimax Rotated Factor Pattern



Principal Components Analysis

1. What is the statistical difference between PCA and FA? Describe the basic construction of each approach using equations and then point to differences that exist across these two widely used methods for reducing dimensionality.

PCA and FA are both dimension-reduction techniques with strong surface resemblance. For this problem set, I spent some time googling for coding help and quickly learned that folks are not very disciplined in using each term appropriately. I'm sure the conflation both reflects and creates misunderstanding. A lot of sites provided description of one or both, but rarely in enough detail to help someone appreciate that there is a difference between the two.

Conceptually, PCA seems simpler. PCA begins with a set of observations on a series of variables and then combines these (after scaling and applying weights) to create a reduced set of dimensions. PCA is essentially looking for a lower-dimension interpretation of the observed variance.

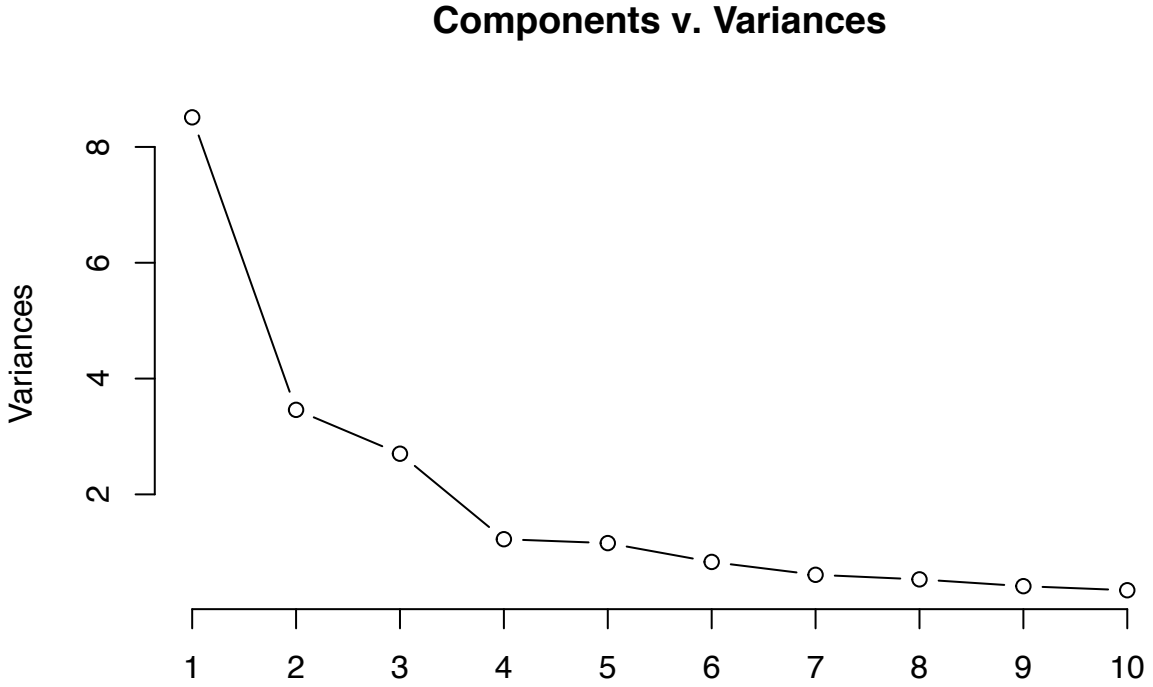
In some ways, FA does the opposite: while PCA seeks to parismoniously express the variance observed, FA tries to uncover the latent causal structure. Necessarily, FA makes stronger theoretical assumptions.

2. Fit a PCA model. Present the proportion of explained variance across the first 10 components. What do these values tell you substantively (e.g., how many components likely characterize these data?)?

```
countries <- read.csv("countries.csv", header = TRUE, row.names = 1)
p <- prcomp(countries, rank. = 10, scale = TRUE)
p2 <- prcomp(countries, center = TRUE, scale = TRUE)
summary(p)
```

```
## Importance of first k=10 (out of 21) 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
## Standard deviation     0.78181 0.72948 0.64421 0.58703
## Proportion of Variance 0.02911 0.02534 0.01976 0.01641
## Cumulative Proportion  0.88104 0.90638 0.92614 0.94255
```

```
plot(p2,
     type="l", # change to "b" for a barplot
     main = "Components v. Variances")
```



From the scree plot, it looks like the optimal number of components is between 4 and 8. Not surprisingly, the biggest gain comes in going from 1 to 2 comoponents, but the move to 3 and subsequent move to 4 also produce moderate gains. Beyond that, gains become minimal at every step, but moderate over the next 4-5 steps.

3. Present a biplot of the PCA fit from the previous question. Describe what you see (e.g., which countries are clustered together? Which input features are doing the bulk of the explaining? How do you know this?

autoplot(p, data = countries, colour = 'democ', label = FALSE, shape = 20,

```
main = "PCA Biplot - All Loadings",
label.repel = FALSE,
loadings = TRUE,
loadings.label.repel = TRUE,
loadings.label.size = 2,
loadings.colour = "cornsilk4",
loadings.label = TRUE
)
```

autoplot(p, data = countries, label = TRUE, shape = FALSE,

```
main = "PCA Biplot - Countries",
label.repel = TRUE,
label.size = 2,
label.colour ="tomato4"
)
```

(PLOTS ON NEXT TWO PAGES)

I ran the PCA model using both the full list of variables and the reduced list of variables I used in the earlier section and I found the results pretty fascinating. The first thing to say is that the countries moved around on the plots between the two models, though I'm not quite sure how to interpret that. The components don't map exaxtly to the factors, and the components in one PCA model don't map exactly to the components from the other PCA model. Those are things I would definitely want to think about more if I were going to build off of this analysis. Does the difference between the plots tell me anything about whether reducing the variable list was useful? Or does it just give me a different angle on the same thing? Or does it give me the same angle but in relation to two different feature spaces?

Again, the most striking thing is the continuity between the models. *Autocracy* is clearly important, as are measures of state terror (these are captured by *statedept* and *amnesty* in the full variable model, those correspond to the *terrorAm* and *terrorST* variables in the reduced model.) Also, we again see a relationship between measures of economic wellbeing and the measure on physical integrity. At this point, we should note that *physint* has a high negative correlation with the state terror measures. A look at the values in the dataset shows that these variables seem to be inversely coded: higher values are related with more state terror in *amnesty* and *statedept* while higher values are related with more physical integrity in *physint*.

As a first-order simplification, it looks like there are TWO major axes: one on which as physical autnomy increases, economic wellbeing increases; and another along which democratic structures are distributed. The autocracy feature seems to capture that end of the spectrum well, while the demcratic end of the spectrum is messier and harder to capture.

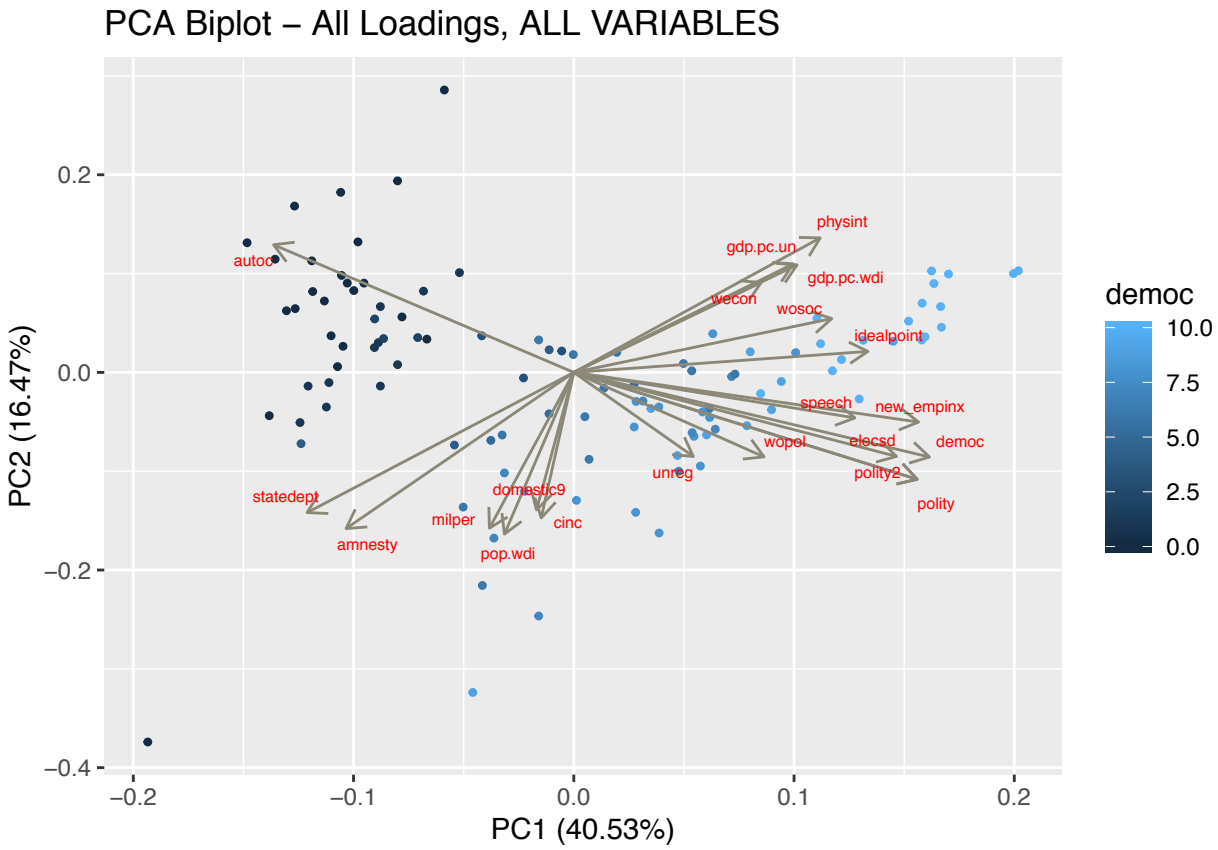
Lastly, it's interesting to look at the countries that cluster. There are clear regional clusters that capture a high proportion of, but not all regional members. My attention is drawn towads a few countries that seem to occupy meaningfully different positions from one plot to another.

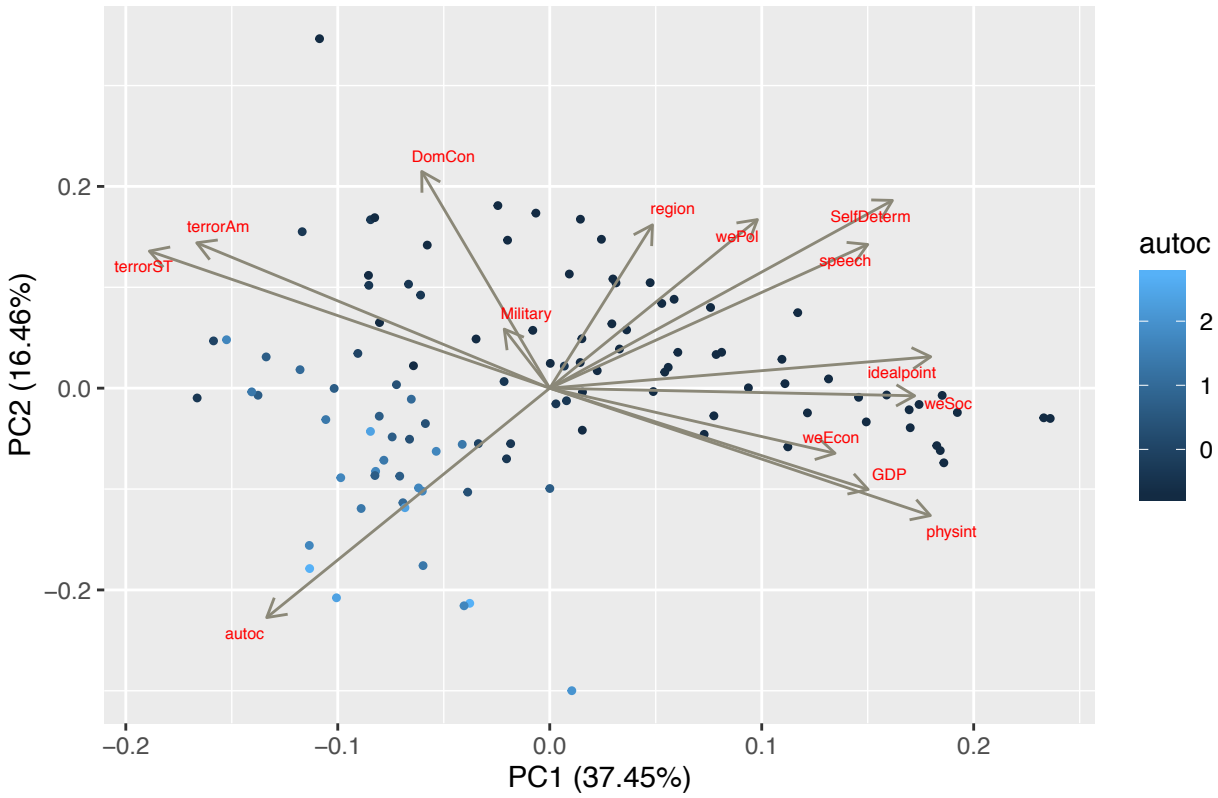
The first is Colombia, of personal interest. I am surprised by how much of an outlier Colombia represents in both models, particularly in the second model, which seems to indicate that the UAE and Colombia are

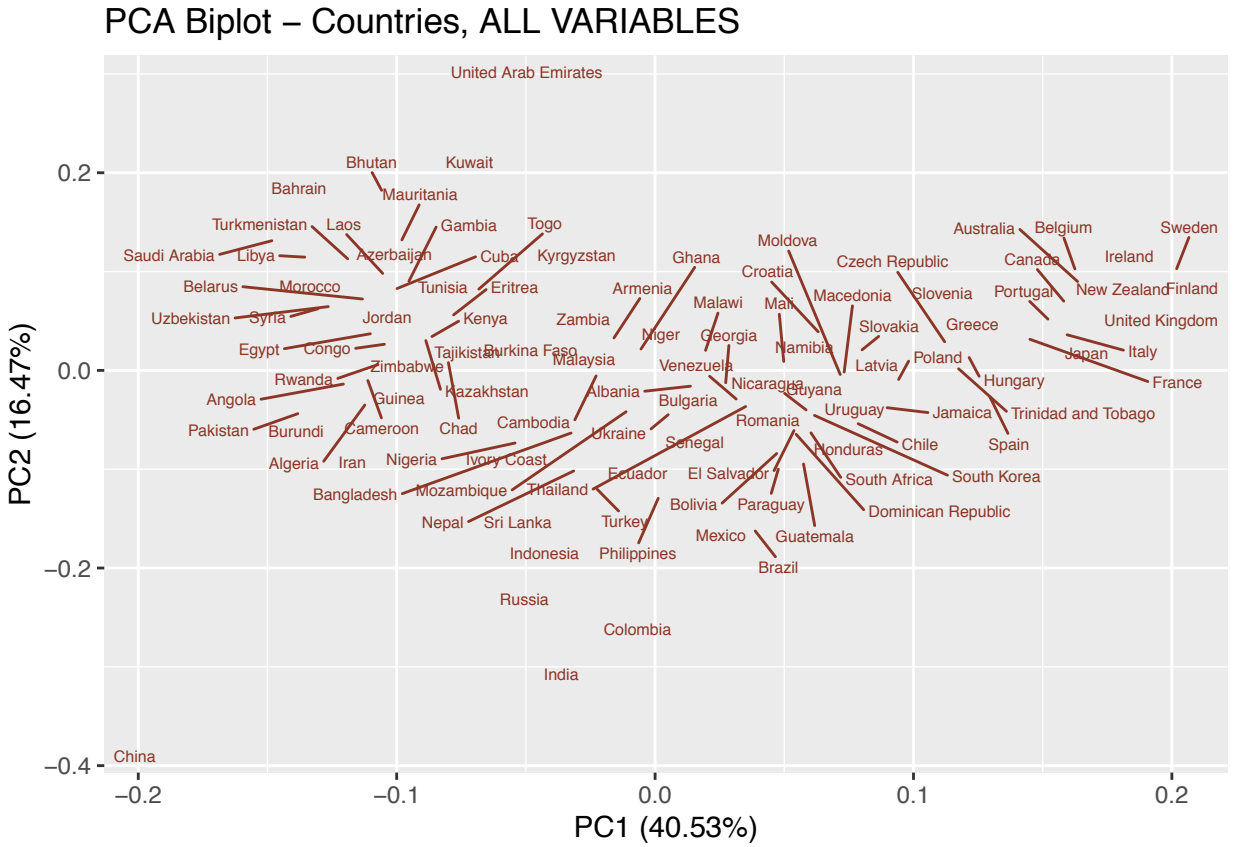
countries with strong, but different particularities. In trying to interpret these plots, I would explore those two countries and their profiles in more detail.

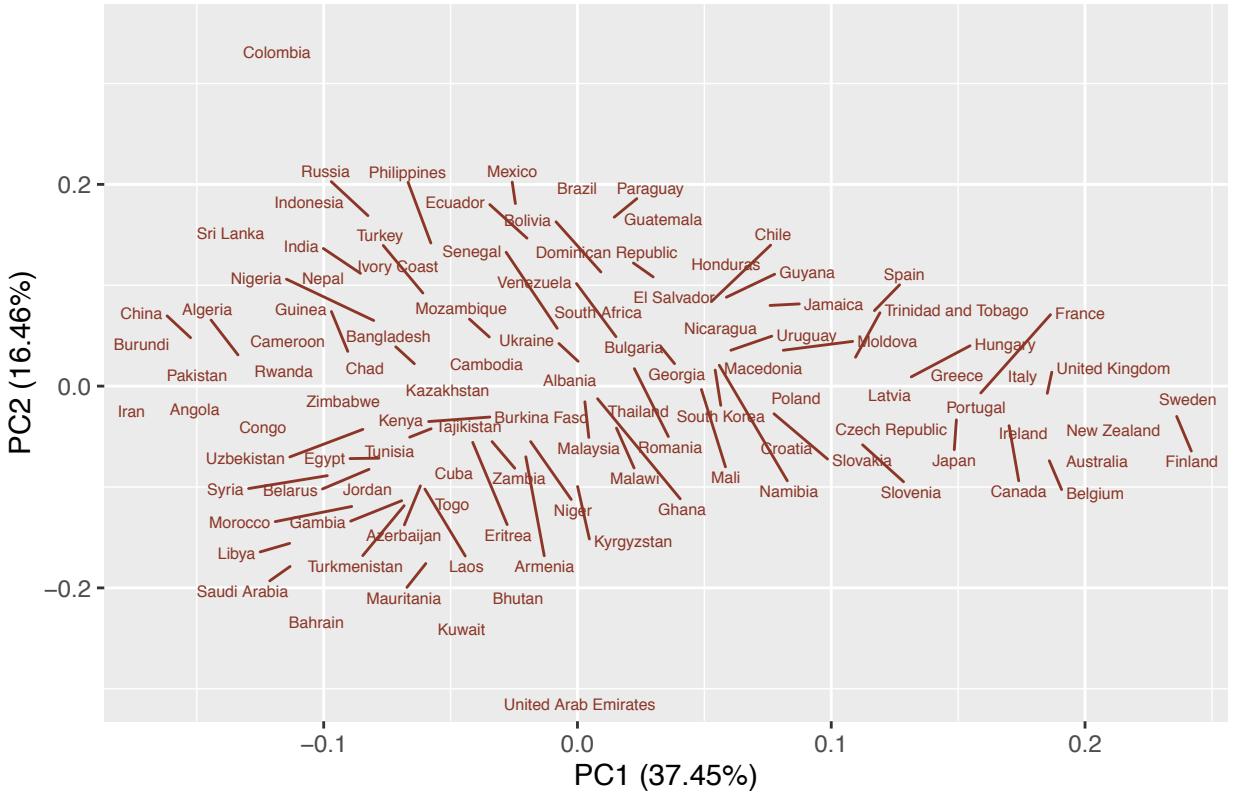
I am also interested by the fact that China , Russia, and India join the general cluster of countries in the second model as compared to the first model. In the case of India and Russia, this shift is interesting; in the case of China, it's dramatic. It's also not lost on me that China, Russia, and India are three very important countries with significant roles in the system of global relations. My takeaway is that the variable reduction did a decent job of capturing the similarities between most countries, but lost something critical to placing China into a global context.

This is a weird sort of finding. Whatever the missing dimension is between the two models seems important, but I'm not sure I would have noticed it looking at either model. I'm sure a more experienced researcher would have noticed this more parsimonously, but it lends credence to the value of just digging around when something seems odd.


PCA Biplot – All Loadings, ALL VARIABLES
PC2 (16.47%)
PC1 (40.53%)
0.2
0.0
−0.2
−0.4
−0.2
−0.1
0.0
0.1
0.2
physint
gdp.pc.un
gdp.pc.wdi
wecon
wosoc
idealpoint
speech
new_empinx
democ
elecsd
wopol
polity2
polity
unreg
autoc
statedept
amnesty
milper
domestic9
cinc
pop.wdi
democ
10.0
7.5
5.0
2.5
0.0


PCA Biplot – All Loadings, REDUCED VARIABLES
PC2 (16.46%)
PC1 (37.45%)
0.2
0.0
−0.2
−0.2
−0.1
0.0
0.1
0.2
DomCon
region
SelfDeterm
terrorAm
terrorST
wePol
speech
Military
idealpoint
weSoc
weEcon
GDP
physint
autoc
autoc
2
1
0


PCA Biplot – Countries, ALL VARIABLES
PC2 (16.47%)
PC1 (40.53%)
0.2
0.0
−0.2
−0.4
−0.2
−0.1
0.0
0.1
0.2
United Arab Emirates
Bhutan
Kuwait
Bahrain
Mauritania
Turkmenistan
Laos
Gambia
Togo
Saudi Arabia
Libya
Azerbaijan
Cuba
Kyrgyzstan
Belarus
Morocco
Tunisia
Eritrea
Uzbekistan
Syria
Jordan
Kenya
Egypt
Congo
Tajikistan
Burkina Faso
Rwanda
Zimbabwe
Angola
Guinea
Kazakhstan
Pakistan
Burundi
Cameroon
Chad
Algeria
Iran
Nigeria
Ivory Coast
Bangladesh
Mozambique
Nepal
Sri Lanka
Indonesia
Russia
Colombia
India
China
Armenia
Zambia
Ghana
Malaysia
Cambodia
Thailand
Albania
Ukraine
Bulgaria
Senegal
Ecuador
Bolivia
Turkey
Philippines
Venezuela
Niger
Malawi
Mali
Georgia
Moldova
Croatia
Macedonia
Namibia
Nicaragua
Guyana
Romania
El Salvador
Mexico
Brazil
Paraguay
Guatemala
Honduras
Uruguay
Dominican Republic
Australia
Belgium
Sweden
Canada
Ireland
Czech Republic
Slovenia
Portugal
New Zealand
Finland
Slovakia
Greece
United Kingdom
Latvia
Poland
Japan
Italy
Hungary
France
Jamaica
Trinidad and Tobago
Chile
Spain
South Africa
South Korea


PCA Biplot – Countries, REDUCED VARIABLES
PC2 (16.46%)
PC1 (37.45%)
0.2
0.0
−0.2
−0.1
0.0
0.1
0.2
Colombia
Russia
Philippines
Mexico
Brazil
Paraguay
Indonesia
Ecuador
Bolivia
Guatemala
Sri Lanka
Turkey
India
Senegal
Dominican Republic
Chile
Ivory Coast
Honduras
Guyana
Nigeria
Nepal
Venezuela
Spain
China
Algeria
Guinea
Mozambique
El Salvador
Jamaica
Trinidad and Tobago
France
Burundi
Cameroon
Bangladesh
South Africa
Nicaragua
Uruguay
Moldova
Ukraine
Bulgaria
Hungary
Pakistan
Rwanda
Chad
Cambodia
Macedonia
Greece
Italy
United Kingdom
Albania
Georgia
Iran
Angola
Zimbabwe
Kazakhstan
Poland
Latvia
Portugal
Sweden
Congo
Kenya
Tajikistan
Burkina Faso
Thailand
South Korea
Czech Republic
Ireland
New Zealand
Uzbekistan
Egypt
Tunisia
Malaysia
Romania
Croatia
Slovakia
Japan
Australia
Finland
Syria
Belarus
Jordan
Cuba
Zambia
Malawi
Mali
Namibia
Slovenia
Canada
Belgium
Morocco
Gambia
Togo
Niger
Ghana
Azerbaijan
Eritrea
Kyrgyzstan
Libya
Turkmenistan
Laos
Armenia
Saudi Arabia
Mauritania
Bhutan
Bahrain
Kuwait
United Arab Emirates

Bonus Question (5 points):

1. Fit a sparse PCA model and a probabilistic PCA model. Compare these results substantively. What does each tell you and why do these distinctions matter in terms of inference (or not)?