Problem Set 4

William Parker

Contents

For the following questions, use the world indicators data from class (`countries.csv`). *Be sure to prepare the data appropriately (e.g., standardize).*

```r
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------
## v ggplot2 3.2.1           v purrr   0.3.2
## v tibble  2.1.3           v dplyr   0.8.3
## v tidyr   0.8.99.9000     v stringr 1.4.0
## v readr   1.3.1           v forcats 0.4.0
## -- Conflicts ---------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(psych) # for fa function
```

```
##
## Attaching package: 'psych'
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
countries <- read_csv("countries.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
## Parsed with column specification:
## cols(
```

```
##   .default = col_double(),
##   X1 = col_character()
## )

## See spec(...) for full column specifications.
```

```
names <- countries %>% select(name = X1)

countries %>% skimr::skim()
```

```
## Skim summary statistics
##  n obs: 107
##  n variables: 22
##
## -- Variable type:character --------------------------------------------------------------------------
##  variable missing complete   n min max empty n_unique
##        X1       0      107 107   4  20     0      107
##
## -- Variable type:numeric ----------------------------------------------------------------------------
##    variable missing complete   n       mean         sd           p0
##     amnesty       0      107 107    2.66       1             1
##       autoc       0      107 107    2.09       2.93          0
##        cinc       0      107 107    0.0068     0.018         4.6e-05
##       democ       0      107 107    5.16       3.82          0
##   domestic9       0      107 107  651.79    1399.89          0
##      elecsd       0      107 107    1.11       0.86          0
##   gdp.pc.un       0      107 107 5110.17    8076.77        103.84
##  gdp.pc.wdi       0      107 107 5183.26    8196.74        128.64
##  idealpoint       0      107 107   -0.088      0.83         -1.68
##      milper       0      107 107  143.56     326.86          1
##  new_empinx       0      107 107    8.42       4.08          0
##     physint       0      107 107    4.32       2.16          0
##      polity       0      107 107    3.07       6.52        -10
##     polity2       0      107 107    3.07       6.52        -10
##     pop.wdi       0      107 107    4.7e+07    1.6e+08 564187
##      speech       0      107 107    1.07       0.72          0
##   statedept       0      107 107    2.48       1.09          1
##       unreg       0      107 107  147.6      137.96          2
##       wecon       0      107 107    1.33       0.56          0
##       wopol       0      107 107    1.85       0.55          0
##       wosoc       0      107 107    1.21       0.76          0
##          p25        p50        p75        p100     hist
##      2          3          3          5
##      0          0          4         10
##      0.00055    0.0015     0.0053     0.16
##      1          6          8.5       10
##      0          0        406       8687
##      0          1          2          2
##    568.64    1461.62    4803.93    37634.42
##    546.71    1461.02    5074.4     37299.64
##     -0.71      -0.36       0.73        1.74
##     13         51        138.5      2810
##      5          9         12         14
##      2.5        4          6          8
##     -3          6          8.5       10
```

```
##     -3           6           8.5       10
##   5e+06          1.1e+07 3e+07          1.3e+09
##      1           1           2          2
##      2           2           3          5
##      2         142         150        419
##      1           1           2          3
##      2           2           2          3
##      1           1           2          3
```

Looking over the variables, there are a bunch that clearly categorical variables. The factor analysis techniques we learned in class were using the Pearson's correlation matrix (assuming variables are continuous and follow multivariate normal distribution), so I'll select and standardize the continuous and normal (ish) variables only. Probably still violating assumptions. there seem to be extensions for ordinal variables (polychoric)

```
countries <- countries %>%
  select(idealpoint, polity, democ, unreg, physint, new_empinx, gdp.pc.wdi, pop.wdi, milper, cinc) %>%
  mutate_all(function(x) as.numeric(scale(x)))

countries
```

```
## # A tibble: 107 x 10
##    idealpoint polity  democ   unreg physint new_empinx gdp.pc.wdi pop.wdi
##         <dbl>  <dbl>  <dbl>   <dbl>   <dbl>      <dbl>      <dbl>   <dbl>
##  1     -0.421 -0.931 -1.09  -1.06    -1.54      -0.593     -0.558  -0.203
##  2      1.62   0.297  0.220  0.0174  -0.611      0.142     -0.489  -0.278
##  3     -0.655 -1.70  -1.35  -0.0406   1.24      -2.06       3.54   -0.278
##  4      0.434  0.297  0.220 -0.0406  -0.147     -0.593     -0.557  -0.278
##  5      1.25   1.06   1.27  -1.00     1.24       1.12       2.01   -0.177
##  6      0.184 -1.54  -1.35  -0.0406  -0.147     -1.08      -0.552  -0.247
##  7     -0.640 -0.624 -1.09  -1.06    -1.54      -0.838     -0.617  -0.255
##  8      1.65   1.06   1.27   0.0174   1.71       0.877      2.20   -0.233
##  9     -0.874 -0.931 -1.35  -1.06    -0.147      0.387     -0.605  -0.225
## 10     -0.888  0.450  0.220 -0.0406  -1.07      -0.348     -0.583   0.526
## # ... with 97 more rows, and 2 more variables: milper <dbl>, cinc <dbl>
```

Factor Analysis

1. How do CFA and EFA differ?

In confirmatory factor analysis, the researcher is using the structure of the factor model to test a specific, well defined hypothesis. For example, using test results on math, physics, English, and Latin, a researcher could fit a 2-factor model to test the specific hypothesis that there are 2 types of latent "intelligence" variables, one correlated (at a given pre-specified) with math and physics and one correlated with English and Latin.

In exploratory factor analysis, the researcher does not have a defined hypothesis and instead is using factor analysis to explore strength of correlations in the data between features and discover how many latent dimensions are required to represent the data. For example, if the k features are entirely independent, then EFA could reveal it take k latent dimensions to properly represent the data

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.

```r
fa_2 <- fa(countries, nfactors = 2)
```

```
## Loading required namespace: GPArotation

## 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.
```

```r
fa_2$loadings
```

```
##
## Loadings:
##            MR1    MR2
## idealpoint  0.717
## polity      0.904
## democ       0.978
## unreg       0.325
## physint     0.506 -0.171
## new_empinx  0.861 -0.123
## gdp.pc.wdi  0.464
## pop.wdi            0.922
## milper             0.965
## cinc               0.974
##
##                  MR1   MR2
## SS loadings    3.617 2.784
## Proportion Var 0.362 0.278
## Cumulative Var 0.362 0.640
```

```r
fa_3 <- fa(countries, nfactors = 3)
```

```
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate =
## rotate, : A loading greater than abs(1) was detected. Examine the loadings
## carefully.

## 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.
```

```r
fa_unrotated <- fa_3$loadings

fa_3$loadings
```

```
##
## Loadings:
##            MR1    MR2    MR3
## idealpoint  0.376         0.513
## polity      1.018
## democ       0.966
## unreg       0.467        -0.178
## physint           -0.130  0.684
```

```
## new_empinx  0.788 -0.135  0.125
## gdp.pc.wdi                 0.741
## pop.wdi            0.909 -0.123
## milper             0.956
## cinc               0.999  0.131
##
##                  MR1   MR2   MR3
## SS loadings    2.958 2.786 1.371
## Proportion Var 0.296 0.279 0.137
## Cumulative Var 0.296 0.574 0.711
```

```
fa_4 <- fa(countries, nfactors = 4)
```

```
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate =
## rotate, : A loading greater than abs(1) was detected. Examine the loadings
## carefully.
```

```
## 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.
```

```
fa_4$loadings
```

```
##
## Loadings:
##            MR1    MR2    MR3    MR4
## idealpoint  0.435         0.361  0.304
## polity      1.026        -0.102
## democ       0.980
## unreg       0.430               -0.253
## physint                   0.833
## new_empinx  0.768         0.229 -0.181
## gdp.pc.wdi                0.574  0.235
## pop.wdi            0.921
## milper             0.949
## cinc               0.999  0.106
##
##                  MR1   MR2   MR3   MR4
## SS loadings    2.981 2.766 1.238 0.258
## Proportion Var 0.298 0.277 0.124 0.026
## Cumulative Var 0.298 0.575 0.698 0.724
```

Inspecting the various number of factors, looks like the total proportion of variance explained jumps a lot from k = 2 to k =3, but not nearly as much when we add the 4 factor. This supports a 3 factor model.

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)?

I used the “varimax” rotation method

```
fa_3_rotate <- fa(countries,
                  nfactors = 3,
                  rotate = "varimax")

fa_rotated <- fa_3_rotate$loadings
```

```r
fa_rotated
```

```
##
## Loadings:
##             MR2    MR1    MR3
## idealpoint         0.442  0.631
## polity             0.945  0.268
## democ              0.919  0.385
## unreg              0.409
## physint    -0.168  0.172  0.696
## new_empinx -0.154  0.766  0.394
## gdp.pc.wdi         0.105  0.718
## pop.wdi     0.914        -0.136
## milper      0.958
## cinc        0.991
##
##                  MR2   MR1   MR3
## SS loadings    2.792 2.729 1.807
## Proportion Var 0.279 0.273 0.181
## Cumulative Var 0.279 0.552 0.733
```

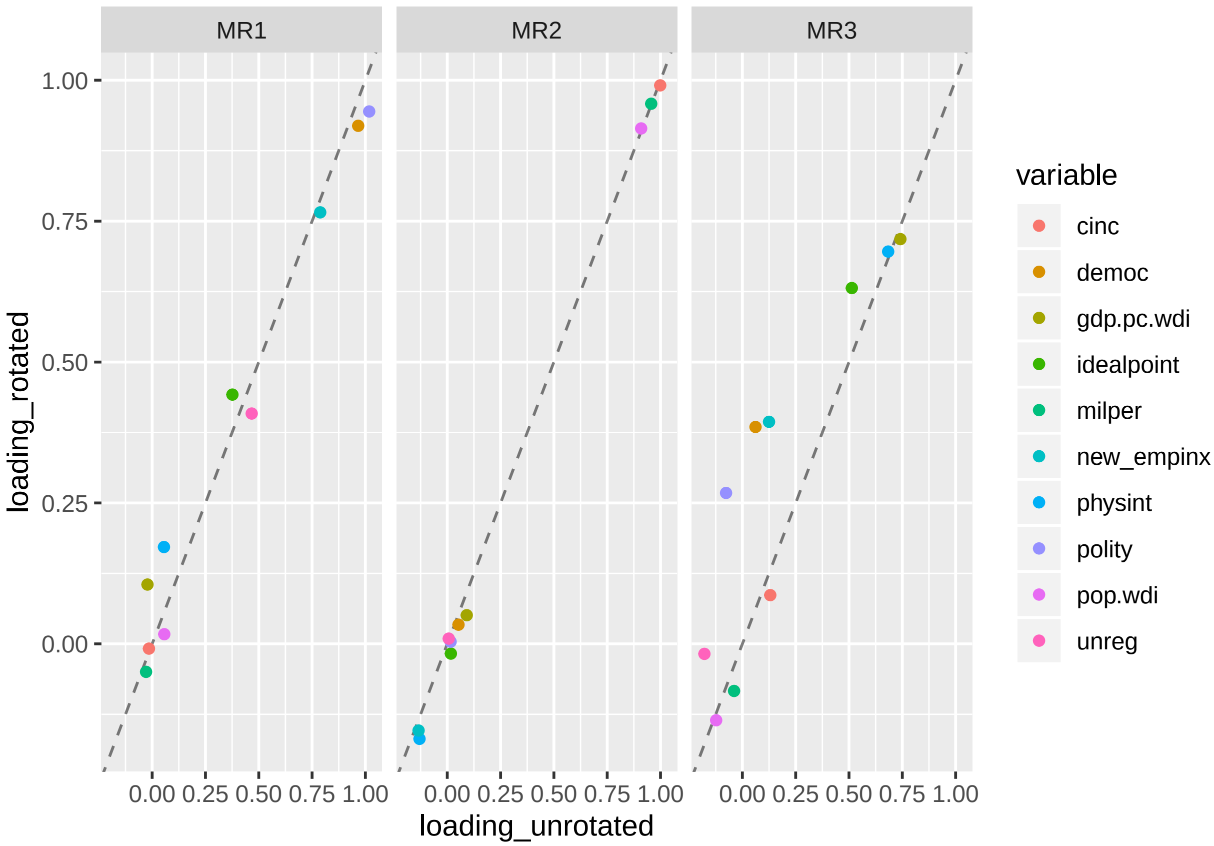
```r
compare_loadings <- as_tibble(fa_unrotated[,])%>%
  cbind(variable = names(countries)) %>%
  pivot_longer(cols = -variable, names_to = "factor", values_to = "loading_unrotated") %>%
  left_join(as_tibble(fa_rotated[,]) %>%
              cbind(variable = names(countries)) %>%
              pivot_longer(cols = -variable, names_to = "factor", values_to = "loading_rotated"))
```

```
## Joining, by = c("variable", "factor")
```

```r
compare_loadings
```

```
## # A tibble: 30 x 4
##    variable   factor loading_unrotated loading_rotated
##    <fct>      <chr>              <dbl>           <dbl>
##  1 idealpoint MR1               0.376           0.442
##  2 idealpoint MR2               0.0169         -0.0172
##  3 idealpoint MR3               0.513           0.631
##  4 polity     MR1               1.02            0.945
##  5 polity     MR2               0.0157          0.00383
##  6 polity     MR3              -0.0765          0.268
##  7 democ      MR1               0.966           0.919
##  8 democ      MR2               0.0527          0.0341
##  9 democ      MR3               0.0615          0.385
## 10 unreg      MR1               0.467           0.409
## # ... with 20 more rows
```

```r
compare_loadings %>%
  ggplot(aes(x = loading_unrotated, y = loading_rotated, color = variable)) +
    geom_abline(intercept = 0, slope = 1, color="black",
                linetype="dashed", alpha = 0.5) +
  geom_point() + facet_wrap(~factor)
```



Both of these factor analysis models explain the same proportion of variance in the data, so in that sense they are equivalent.

Also the magnitude of the loadings on each factor are roughly the same for most variables, so it doesn't appear this rotation has changed the interpretation of each factor significantly.

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.

Factor Analysis

Let's define $x_1, x_2, ..x_n$ to be our sample where $x_i \in R^d$, i.e. we have d features. Let the matrix X represent the whole sample with dimensions d x n (after scaling to mean zero for each feature).

In factor analysis, we represent

$X = LF + \varepsilon$

where L is a d x k of "loadings" and F is a k x n matrix representing the unobserved latent "factors" and ε is a d x k matrix of error terms. The loadings L here do not vary with n and represent the relationship between the unobserved factors F and the d input feature variables.

focusing on a single observation from the sample x_i is a d x 1 vector

$x_i = LF_i + \varepsilon_i$

where F_i is the ith column of F and is a k x 1 vector of weights.

PCA

In PCA, we minimize the following loss function

$Loss = \sum_{i=1}^{n} ||x_i - V\alpha_i||^2$

where V is an **orthogonal** d x k matrix (turns out to be the top k eigenvectors of the sample covariance $S = \frac{1}{n}\sum x_i x_i^t$) and the columns are called the principle component "vectors". α_i is a k x 1 vector representing the weight of each component in creating the estimate of x_i.

so V is analogous to the loading matrix L and α_i is analogous to the weighting vector F_i for observation x_i.

Differences

In factor analysis, there are many possible ways to optimize L and F with respect to various loss functions and restrictions on L. The orthogonality assumption of V and the specification of a particular loss function is what distinguishes PCA from (and makes it a a subtype of) factor analysis.

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?)?

```r
cov_matrix <- cov(countries)
country_eigen <- eigen(cov_matrix)
cov_matrix
```

```
##              idealpoint      polity        democ        unreg     physint
## idealpoint   1.00000000  0.60847112  0.666743593  0.118614302  0.51989079
## polity       0.60847112  1.00000000  0.973791515  0.351247476  0.32135851
## democ        0.66674359  0.97379151  1.000000000  0.389559842  0.39116398
## unreg        0.11861430  0.35124748  0.389559842  1.000000000  0.08409644
## physint      0.51989079  0.32135851  0.391163975  0.084096437  1.00000000
## new_empinx   0.57145463  0.83427786  0.828332534  0.341568008  0.48291871
## gdp.pc.wdi   0.48863673  0.29298378  0.400660785  0.025376495  0.51162824
## pop.wdi     -0.11225091 -0.01742067 -0.001144892  0.006114645 -0.22753618
## milper      -0.06805411 -0.06649447 -0.045782794 -0.002045234 -0.22176895
## cinc         0.02621179  0.01765334  0.046666228  0.017010333 -0.12219545
##             new_empinx  gdp.pc.wdi      pop.wdi       milper        cinc
## idealpoint   0.5714546  0.48863673 -0.112250911 -0.068054109  0.02621179
## polity       0.8342779  0.29298378 -0.017420669 -0.066494475  0.01765334
## democ        0.8283325  0.40066079 -0.001144892 -0.045782794  0.04666623
## unreg        0.3415680  0.02537649  0.006114645 -0.002045234  0.01701033
## physint      0.4829187  0.51162824 -0.227536176 -0.221768949 -0.12219545
## new_empinx   1.0000000  0.33095195 -0.170172200 -0.233045630 -0.11014391
## gdp.pc.wdi   0.3309520  1.00000000 -0.057906952 -0.033022399  0.13143354
## pop.wdi     -0.1701722 -0.05790695  1.000000000  0.889757944  0.89611332
## milper      -0.2330456 -0.03302240  0.889757944  1.000000000  0.93991436
## cinc        -0.1101439  0.13143354  0.896113320  0.939914363  1.00000000
```

```r
prcomp(countries)
```

```
## Standard deviations (1, .., p=10):
##  [1] 2.0023811 1.6754963 1.1183460 0.8558095 0.6887178 0.6282595 0.4105193
##  [8] 0.3286711 0.1985056 0.1227863
##
## Rotation (n x k) = (10 x 10):
##                    PC1         PC2         PC3         PC4          PC5
## idealpoint  0.38570844  0.09816232 -0.23984886  0.16062573 -0.013645561
## polity      0.43206099  0.14702064  0.25770448  0.30081381  0.071711315
## democ       0.44957098  0.16614736  0.18666281  0.18654515  0.115034249
## unreg       0.18458227  0.08473218  0.55253751 -0.77841744  0.057119640
## physint     0.32018588 -0.04100439 -0.44006227 -0.33977288 -0.707466938
## new_empinx  0.44303000  0.04151906  0.16469533  0.14122863 -0.160505013
## gdp.pc.wdi  0.26958511  0.09841296 -0.54984481 -0.32172775  0.654982898
## pop.wdi    -0.14749327  0.54337020  0.02687334  0.03864792 -0.122289230
## milper     -0.16013624  0.54840657 -0.03211145 -0.01063231 -0.088740592
## cinc       -0.09746299  0.56997664 -0.10515811 -0.04970939 -0.008501862
##                    PC6         PC7           PC8          PC9        PC10
## idealpoint -0.83945150 -0.17327544  0.1449991668  0.010805442  0.04726671
## polity      0.13819654  0.36426276 -0.1161845553 -0.169665215  0.65999576
## democ       0.08223761  0.39303949 -0.0192918236  0.067193974 -0.72199428
## unreg      -0.20025139 -0.01968772  0.0346503350 -0.007386582  0.05012329
## physint     0.16012685  0.24095099 -0.0323947972 -0.030125508  0.02312586
## new_empinx  0.34930853 -0.76397359 -0.0001792637  0.151538244 -0.02307973
## gdp.pc.wdi  0.24117348 -0.02424454  0.0534184589  0.112107251  0.09286596
## pop.wdi     0.12579423  0.06606377  0.7873382435  0.142748778  0.07318974
## milper     -0.08522130  0.02664312 -0.5271213601  0.610596364  0.07700754
## cinc        0.02122959 -0.18720446 -0.2494837585 -0.732722874 -0.13192766
```

```
summary(prcomp(countries))
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     2.0024 1.6755 1.1183 0.85581 0.68872 0.62826
## Proportion of Variance 0.4009 0.2807 0.1251 0.07324 0.04743 0.03947
## Cumulative Proportion  0.4009 0.6817 0.8067 0.87999 0.92743 0.96690
##                            PC7    PC8     PC9    PC10
## Standard deviation     0.41052 0.3287 0.19851 0.12279
## Proportion of Variance 0.01685 0.0108 0.00394 0.00151
## Cumulative Proportion  0.98375 0.9946 0.99849 1.00000
```

I used the `prcomp` function to get the proportion of explained variance quickly. Looks like 3 components characterize the data fairly well (80% proportion of variance), similar to what I found in the factor analysis.

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?

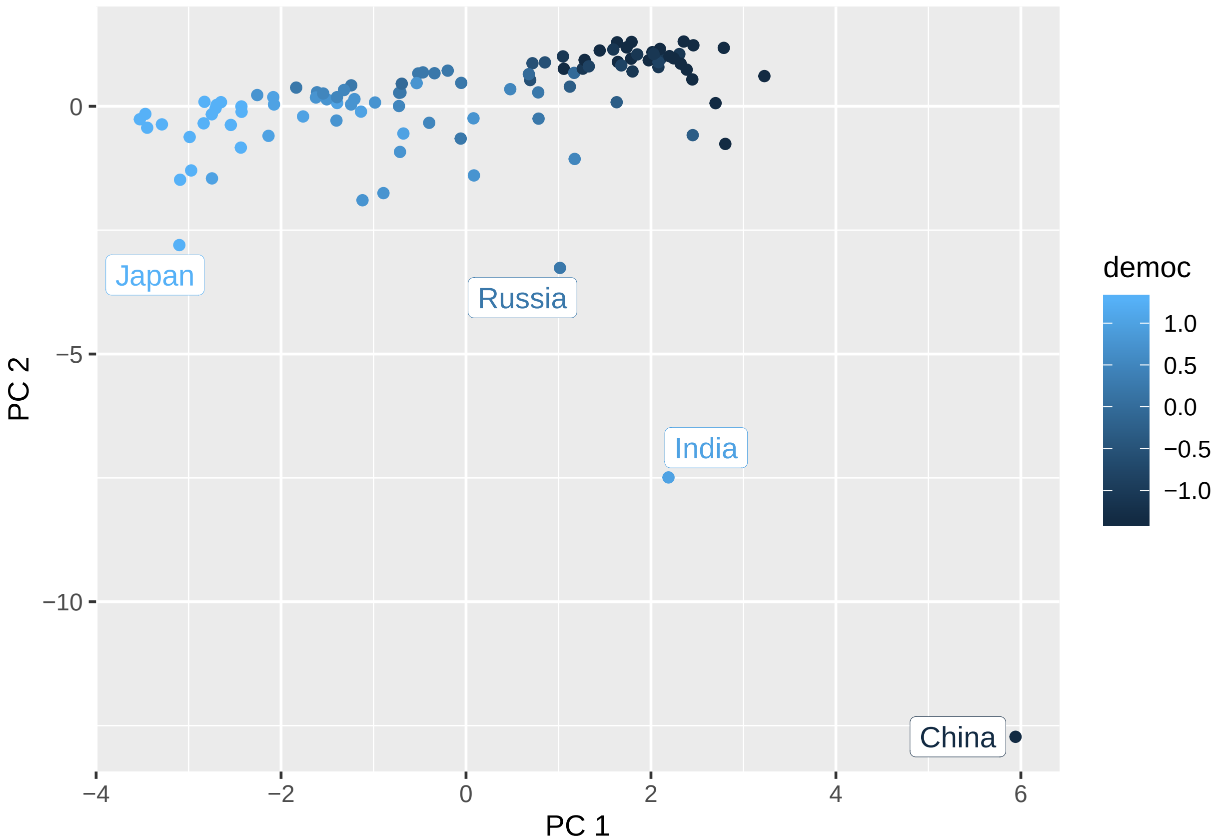
```
pca_project <- as.matrix(countries) %*% as.matrix(country_eigen$vectors) %>%
  as_tibble() %>%
  cbind(country = names$name) %>%
  cbind(countries)
```

```
## Warning: `as_tibble.matrix()` requires a matrix with column names or a `.name_repair` argument. Using
## This warning is displayed once per session.
```

```r
head(pca_project)
```

```
##           V1         V2         V3         V4          V5         V6
## 1  2.0806233  0.7904132 -0.0124199 -0.8961747  0.60785819  0.2637688
## 2 -0.7085624  0.2765544  0.3383516 -0.7709856  0.17602347  1.4870403
## 3  1.0581079  0.7562632 -3.3624566  2.6851470  1.55761350 -0.5244772
## 4 -0.0516031  0.4715476  0.3028321 -0.3865112 -0.06197356  0.6771370
## 5 -2.8369327 -0.3446551 -1.8082901 -0.6253694  0.45171444 -0.4777425
## 6  1.7857953  0.9594754 -0.4937915  0.5726803 -0.30252248  1.0246463
##            V7          V8          V9         V10              country
## 1  0.54265405 -0.01717142 -0.08625220 -0.03801203               Angola
## 2  0.29013679 -0.20233068 -0.01500247 -0.05593544              Albania
## 3 -0.77682357 -0.24652552 -0.23516081 -0.22090761 United Arab Emirates
## 4 -0.58850907 -0.02929139  0.15990128 -0.01446083              Armenia
## 5 -0.06357846 -0.07635098 -0.08807269  0.04853518            Australia
## 6  0.28399996 -0.20371857 -0.00798985  0.06181846           Azerbaijan
##   idealpoint     polity      democ       unreg    physint new_empinx
## 1 -0.4208015 -0.9308537 -1.0888860 -1.05539260 -1.5381108 -0.5930132
## 2  1.6171518  0.2968979  0.2202241  0.01741035 -0.6109116  0.1419568
## 3 -0.6545836 -1.6981985 -1.3507080 -0.04057900  1.2434868 -2.0629533
## 4  0.4335634  0.2968979  0.2202241 -0.04057900 -0.1473120 -0.5930132
## 5  1.2537340  1.0642426  1.2675123 -1.00465192  1.2434868  1.1219169
## 6  0.1836501 -1.5447295 -1.3507080 -0.04057900 -0.1473120 -1.0829933
##   gdp.pc.wdi    pop.wdi    milper         cinc
## 1 -0.5583914 -0.2030582 -0.1087941 -0.272653715
## 2 -0.4889105 -0.2781062 -0.2740010 -0.348194937
## 3  3.5409529 -0.2783501 -0.2403478 -0.251348075
## 4 -0.5565427 -0.2781872 -0.3137731 -0.340931144
## 5  2.0107811 -0.1773871 -0.2831792 -0.003759346
## 6 -0.5524347 -0.2470103 -0.2189321 -0.306973885
```

```r
pca_project %>%
  mutate(country = if_else(V2 < -2 | V1 > 4, as.character(country), ""))%>%
  ggplot(aes(x = V1, y = V2, label = country, color = democ)) +
  geom_point() + ggrepel::geom_label_repel(label.size = 0.05) +
  labs( x= "PC 1", y = "PC 2")
```



`democ` is correlated with PC1 (along with a few other features strongly correlated with democracy) and country size seems to be strongly correlated with PC 2.

so type of government is PC1, country size is PC2.