PS#4 - Lilly R.

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

##Factor Analysis 1. How do CFA and EFA differ?

# Confirmatory Factor Analysis is an a priori hypotheses that defines factors and accordingly tests hypotheses. However, Exploratory Factor Analysis doesn't require a priori information and rather fits the model to a given number of factos. CFA is used to test a proposed model applicability whereas EFA tries to model the data without any input. CFA also requires much more information about the "science" of the dataset's behaviour.

1. 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.

library(lattice)  
library(psych)  
library(GPArotation)  
library(ggplot2)

##   
## Attaching package: 'ggplot2'

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

df <- read.csv('/Users/lillyreich/Desktop/countries.csv')

mydata=df[, -c(1,3,4)]  
mydatacor <- (cor(mydata))

mydata\_scaled <- scale(mydata)  
fit2 <- factanal(mydata, 2, rotation="varimax")  
fit2

##   
## Call:  
## factanal(x = mydata, factors = 2, rotation = "varimax")  
##   
## Uniquenesses:  
## idealpoint democ autoc unreg physint speech   
## 0.759 0.839 0.980 0.999 0.689 0.892   
## new\_empinx wecon wopol wosoc elecsd gdp.pc.wdi   
## 0.854 0.753 1.000 0.741 0.906 0.005   
## gdp.pc.un pop.wdi amnesty statedept milper cinc   
## 0.005 0.146 0.595 0.605 0.063 0.031   
## domestic9   
## 0.971   
##   
## Loadings:  
## Factor1 Factor2  
## idealpoint 0.491   
## democ 0.402   
## autoc -0.135   
## unreg   
## physint 0.532 -0.165   
## speech 0.318   
## new\_empinx 0.348 -0.158   
## wecon 0.484 -0.115   
## wopol   
## wosoc 0.508   
## elecsd 0.304   
## gdp.pc.wdi 0.994   
## gdp.pc.un 0.994   
## pop.wdi -0.142 0.913   
## amnesty -0.567 0.288   
## statedept -0.601 0.183   
## milper -0.122 0.960   
## cinc 0.984   
## domestic9 -0.148   
##   
## Factor1 Factor2  
## SS loadings 4.230 2.943  
## Proportion Var 0.223 0.155  
## Cumulative Var 0.223 0.378  
##   
## Test of the hypothesis that 2 factors are sufficient.  
## The chi square statistic is 1121.09 on 134 degrees of freedom.  
## The p-value is 1.93e-155

# Since the p-value is low for the 2 factor EFA, we can conclude that more factors are needed to explain the dataset. The uniqueness is very high in a number of variables which indicates a bad fit as well. Examining the loadings shows that some variables aren't even explained by either factor as well as factors explaining multiple variables which also isn't useful. SS loadings indicates that all factors used are useful and should be kept. The cumulative variance is relatively low at 0.378 indicating that much of the covariance in our system is yet to be explained by more factors.  
fit3 <- factanal(mydata, 3, rotation="varimax")  
fit3

##   
## Call:  
## factanal(x = mydata, factors = 3, rotation = "varimax")  
##   
## Uniquenesses:  
## idealpoint democ autoc unreg physint speech   
## 0.507 0.101 0.187 0.852 0.628 0.455   
## new\_empinx wecon wopol wosoc elecsd gdp.pc.wdi   
## 0.153 0.734 0.686 0.622 0.194 0.005   
## gdp.pc.un pop.wdi amnesty statedept milper cinc   
## 0.005 0.145 0.569 0.532 0.062 0.028   
## domestic9   
## 0.960   
##   
## Loadings:  
## Factor1 Factor2 Factor3  
## idealpoint 0.531 0.459   
## democ 0.881 0.347   
## autoc -0.897   
## unreg 0.384   
## physint 0.283 0.520 -0.147   
## speech 0.681 0.278   
## new\_empinx 0.861 0.299 -0.132   
## wecon 0.169 0.476   
## wopol 0.560   
## wosoc 0.377 0.486   
## elecsd 0.861 0.252   
## gdp.pc.wdi 0.990 0.114   
## gdp.pc.un 0.991 0.115   
## pop.wdi -0.163 0.910   
## amnesty -0.204 -0.562 0.271   
## statedept -0.310 -0.587 0.163   
## milper -0.141 0.956   
## cinc 0.986   
## domestic9 -0.156   
##   
## Factor1 Factor2 Factor3  
## SS loadings 4.678 3.997 2.904  
## Proportion Var 0.246 0.210 0.153  
## Cumulative Var 0.246 0.457 0.609  
##   
## Test of the hypothesis that 3 factors are sufficient.  
## The chi square statistic is 528.67 on 117 degrees of freedom.  
## The p-value is 1.24e-53

# The analysis for the two factor EFA also applies here and it appears we need more factors. Yet, the p-value is still small. The cumulative var is 0.609 which shows significant improvement but is also still small. All of the data is covered to an extension (at least beyond a cutoff of 0.1) by the three factors. The uniqueness is still very high for a number of variables and doesn't seem to decrease meaningfully.  
fit4 <- factanal(mydata, 4, rotation="varimax")  
fit4

##   
## Call:  
## factanal(x = mydata, factors = 4, rotation = "varimax")  
##   
## Uniquenesses:  
## idealpoint democ autoc unreg physint speech   
## 0.442 0.089 0.153 0.838 0.258 0.465   
## new\_empinx wecon wopol wosoc elecsd gdp.pc.wdi   
## 0.162 0.713 0.689 0.581 0.200 0.005   
## gdp.pc.un pop.wdi amnesty statedept milper cinc   
## 0.005 0.145 0.350 0.137 0.063 0.026   
## domestic9   
## 0.620   
##   
## Loadings:  
## Factor1 Factor2 Factor3 Factor4  
## idealpoint 0.519 0.345 0.412   
## democ 0.899 0.276 0.161   
## autoc -0.919   
## unreg 0.395   
## physint 0.223 0.325 -0.107 0.758   
## speech 0.677 0.207 0.169   
## new\_empinx 0.853 0.199 -0.131 0.230   
## wecon 0.170 0.403 -0.111 0.290   
## wopol 0.554   
## wosoc 0.365 0.390 0.364   
## elecsd 0.867 0.187 0.117   
## gdp.pc.wdi 0.106 0.963 0.239   
## gdp.pc.un 0.964 0.236   
## pop.wdi 0.914 -0.118   
## amnesty -0.169 -0.393 0.251 -0.636   
## statedept -0.258 -0.382 0.126 -0.797   
## milper 0.957 -0.118   
## cinc 0.106 0.978   
## domestic9 0.167 -0.592   
##   
## Factor1 Factor2 Factor3 Factor4  
## SS loadings 4.683 2.914 2.842 2.625  
## Proportion Var 0.246 0.153 0.150 0.138  
## Cumulative Var 0.246 0.400 0.549 0.688  
##   
## Test of the hypothesis that 4 factors are sufficient.  
## The chi square statistic is 367.91 on 101 degrees of freedom.  
## The p-value is 5.18e-32

# The analyis for the two factor EFA also applies in this instance since we also neeed to include more factors. Here, the p-value is higher but still small. The cummulative variance is 0.688 which is higher than 0.609 for the 3-factor analysis but not significantly. The uniqueness has decreased for many of the variables, but is still close to 1 for many. Finally, we can conclude that the dimensionality of the system is higher than 4.

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(mydata),  
 fm = "pa", # communalities along the diagonal (total variation across features)  
 nfactors = 3,  
 rotate = "none",  
 residuals = TRUE)  
  
promax.factors <- fa(cor(mydata),  
 fm = "pa",  
 nfactors = 3,  
 rotate = "promax")

## 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.

#smc = TRUE)  
  
nonrot.pattern <- as.data.frame(nonrotated.factors$loadings[1:19,])  
nonorthog.pattern <- as.data.frame(promax.factors$loadings[1:19,])

plot(PA2 ~ PA1, data = nonrot.pattern,   
 aspect = 1,  
 xlim = c(-1, 1),  
 ylim = c(-1, 1),  
 main = "Unrotated Factor Pattern",  
 xlab = "",  
 ylab = "",  
 scales = list(x = list(at = c(0, 1)),  
 y = list(at = c(-.4, 0, .6)))  
)

## Warning in plot.window(...): "aspect" is not a graphical parameter

## Warning in plot.window(...): "scales" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "aspect" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "scales" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "aspect" is  
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "scales" is  
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "aspect" is  
## not a graphical parameter

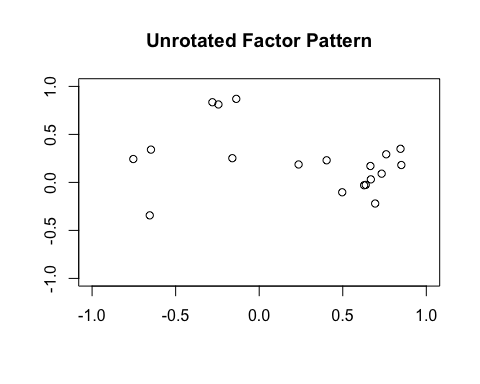
## Warning in axis(side = side, at = at, labels = labels, ...): "scales" is  
## not a graphical parameter

## Warning in box(...): "aspect" is not a graphical parameter

## Warning in box(...): "scales" is not a graphical parameter

## Warning in title(...): "aspect" is not a graphical parameter

## Warning in title(...): "scales" is not a graphical parameter



plot(PA2 ~ PA1, data = nonorthog.pattern,   
 aspect = 1,  
 xlim = c(-1, 1),  
 ylim = c(-1, 1),  
 # 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, 19), rep(0, 19), x, y,   
 # col = "black")  
 #panel.text(x[-18], y[-18], labels = rownames(nonorthog.pattern)[-18],  
 # pos = 4, cex = .75)  
 #panel.text(x[18], y[18], labels = rownames(nonorthog.pattern)[18],  
 # pos = 1, cex = .75)  
 # },  
 main = "Non-Orthogonal Rotated Factor Pattern",  
 xlab = "",  
 ylab = "",  
 scales = list(x = list(at = c(0, 1)),  
 y = list(at = c(-.4, 0, .6)))  
)

## Warning in plot.window(...): "aspect" is not a graphical parameter

## Warning in plot.window(...): "scales" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "aspect" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "scales" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "aspect" is  
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "scales" is  
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "aspect" is  
## not a graphical parameter

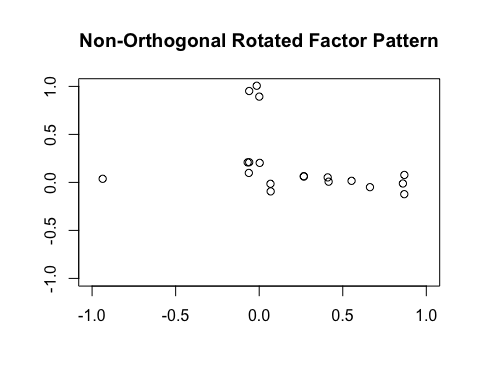
## Warning in axis(side = side, at = at, labels = labels, ...): "scales" is  
## not a graphical parameter

## Warning in box(...): "aspect" is not a graphical parameter

## Warning in box(...): "scales" is not a graphical parameter

## Warning in title(...): "aspect" is not a graphical parameter

## Warning in title(...): "scales" is not a graphical parameter



#Compare i.e. why does this matter

# The nonrotated plot has its factors all over the plot and there seems to be a lot of correlation between the two factors since the data is at a combination of intermediate values of the two. In the Promax rotated plot, we see more of the data lying around the x and y axes which indicates there is far more independence in the two factors. This matters because it helps us in defining better factors with more independence by simply rotating the data without any mathematical variation.

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.

# In Principal Component Analysis, we are calculating the weighted contribution of each observed variable to n-component. This will give us an indication of different components at play and how each observation plays a part in their construction. Alternatively, Factor Analysis defines different factors and weights that contributes to different observation values and also accounts for a summed term called the "uniqueness".  
# For example, for a two factor/component (F or C) approach of 4 measured variables (Y), an FA would look like the following:  
# Y1 = b11 \* F1 + b12 \* F2 + U1  
# Y2 = b21 \* F1 + b22 \* F2 + U2  
# Y3 = b31 \* F1 + b32 \* F2 + U3  
# Y4 = b41 \* F1 + b42 \* F2 + U4  
# and for PCA, we would have:  
# C1 = w11 \* Y1 + w12 \* Y2 + w13 \* Y3 + w14 \* Y4   
# C2 = w21 \* Y1 + w22 \* Y2 + w23 \* Y3 + w24 \* Y4   
# Clearly, FA includes a uniqueness and PCA doesn't. There are less equations to fit in PCA and both include weights that are defined in different ways. For example, in PCA, the weights (wij) are the individual contributions of the observed variable to the component. On the other hand, in FA, the weights are the individual contributions of the factors to the different observed variables. In a sense, it is an inverted approach.

1. 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?)?

PCAfit <- prcomp(mydata, center = TRUE,scale. = TRUE)  
summary(PCAfit)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 2.6655 1.7598 1.6189 1.10564 1.07615 0.90051  
## Proportion of Variance 0.3739 0.1630 0.1379 0.06434 0.06095 0.04268  
## Cumulative Proportion 0.3739 0.5369 0.6749 0.73922 0.80017 0.84285  
## PC7 PC8 PC9 PC10 PC11 PC12  
## Standard deviation 0.77811 0.67396 0.64080 0.58629 0.54539 0.49340  
## Proportion of Variance 0.03187 0.02391 0.02161 0.01809 0.01566 0.01281  
## Cumulative Proportion 0.87472 0.89863 0.92024 0.93833 0.95398 0.96680  
## PC13 PC14 PC15 PC16 PC17 PC18  
## Standard deviation 0.43617 0.39932 0.32738 0.2892 0.23963 0.18033  
## Proportion of Variance 0.01001 0.00839 0.00564 0.0044 0.00302 0.00171  
## Cumulative Proportion 0.97681 0.98520 0.99084 0.9952 0.99827 0.99998  
## PC19  
## Standard deviation 0.01990  
## Proportion of Variance 0.00002  
## Cumulative Proportion 1.00000

# 37.39% of the data is explained by the first component alone. 16.30% by the second or 53.69 of the data is explained by only the first and second components combined. 67.48 by 3. 73.91 by 4. 80 by 5. 84.27 by 6. 87.464 by 7. 89.85 by 8. 92.02 by 9. 93.824 by 10. Essentially, the 93.82% represents a summation of the proportion of variance explained by the first 10 components. It seems that only 5 components are enough to explain 80% of the covariance. Interestingly, there is a small increase in proportion of variance with 5 components added when more components are added. This indicates that five components might be sufficient to explain all of the data. This might change if the required accuracy is higher.

1. 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?

library(devtools)

## Loading required package: usethis

#install\_github("vqv/ggbiplot")  
library(ggbiplot)

## Loading required package: plyr

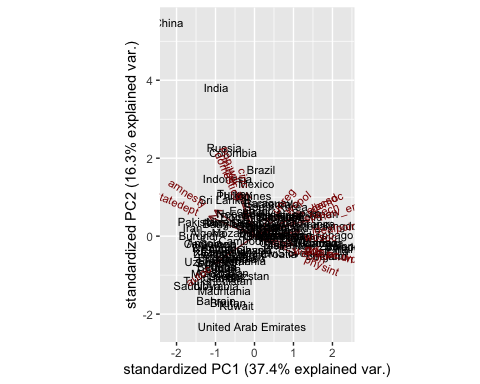
## Loading required package: scales

##   
## Attaching package: 'scales'

## The following objects are masked from 'package:psych':  
##   
## alpha, rescale

## Loading required package: grid

par(pin=c(1.9,1.9))  
ggbiplot(PCAfit, labels=df$X)



# The biplot graphs the contributions of the different measured values to the first two components. For example, gdp.pc.wdi contributes mostly to component 1. We can clearly see that countries cluster in the plot based on their geographic locations. For example, European contries are lumped to the right of the plot around the axis for PC2. This is a clear indication that PC1 is related to prosperity and democracy. There are more clusterings around the plot according to geographic locations; Gulf and Middle Eastern countries around the bottom left, African countries around the center, South American Countries around the center right. cinc and domestic9 seem to have the most impact on PC2 whereas amnesty, statedept, speech, gdp.pc.wdi, gdp.pc.un and wosoc seem to contribute most to PC1. The direction of the arrows indicates their contribution to the components. For example, wosoc is almost directly pointing east which indicates it contributes to PC1 positively and PC2 negligibly.

## Bonus Question (5 points):t

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)?

#install.packages("nsprcomp")  
library(nsprcomp)  
spca <- nsprcomp(mydata, center = TRUE,scale. = TRUE)  
summary(spca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 2.6654 1.7583 1.6206 1.10512 1.07638 0.90084  
## Proportion of Variance 0.3739 0.1627 0.1382 0.06428 0.06098 0.04271  
## Cumulative Proportion 0.3739 0.5366 0.6749 0.73915 0.80013 0.84284  
## PC7 PC8 PC9 PC10 PC11 PC12  
## Standard deviation 0.77797 0.67282 0.64195 0.5864 0.54545 0.49361  
## Proportion of Variance 0.03185 0.02383 0.02169 0.0181 0.01566 0.01282  
## Cumulative Proportion 0.87470 0.89852 0.92021 0.9383 0.95397 0.96679  
## PC13 PC14 PC15 PC16 PC17 PC18  
## Standard deviation 0.43576 0.39985 0.32724 0.28939 0.23961 0.18038  
## Proportion of Variance 0.00999 0.00841 0.00564 0.00441 0.00302 0.00171  
## Cumulative Proportion 0.97679 0.98520 0.99084 0.99524 0.99827 0.99998  
## PC19  
## Standard deviation 0.01990  
## Proportion of Variance 0.00002  
## Cumulative Proportion 1.00000

ppca <- pca(mydata, ncomp=19, method="ppca")  
summary(ppca)

##   
## Factor analysis with Call: principal(r = r, nfactors = nfactors, residuals = residuals,   
## rotate = rotate, n.obs = n.obs, covar = covar, scores = scores,   
## missing = missing, impute = impute, oblique.scores = oblique.scores,   
## method = method, ncomp = 19)  
##   
## Test of the hypothesis that 1 factor is sufficient.  
## The degrees of freedom for the model is 152 and the objective function was 17.45   
## The number of observations was 107 with Chi Square = 1713.33 with prob < 3.6e-262   
##   
## The root mean square of the residuals (RMSA) is 0.19

# A probabilistic PCA's (PPCA's) results look more like a FA with results such as uniqueness and residual calculation. PPCA implements a maximum likelihood approach to improve the choice of components which will yield better results. It sort of does automatic model selection by fitting the maximum likelihood.

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.