HW_11_Gupta_S

Sumit Gupta

November 29, 2017

Loading the dataset from the psych package

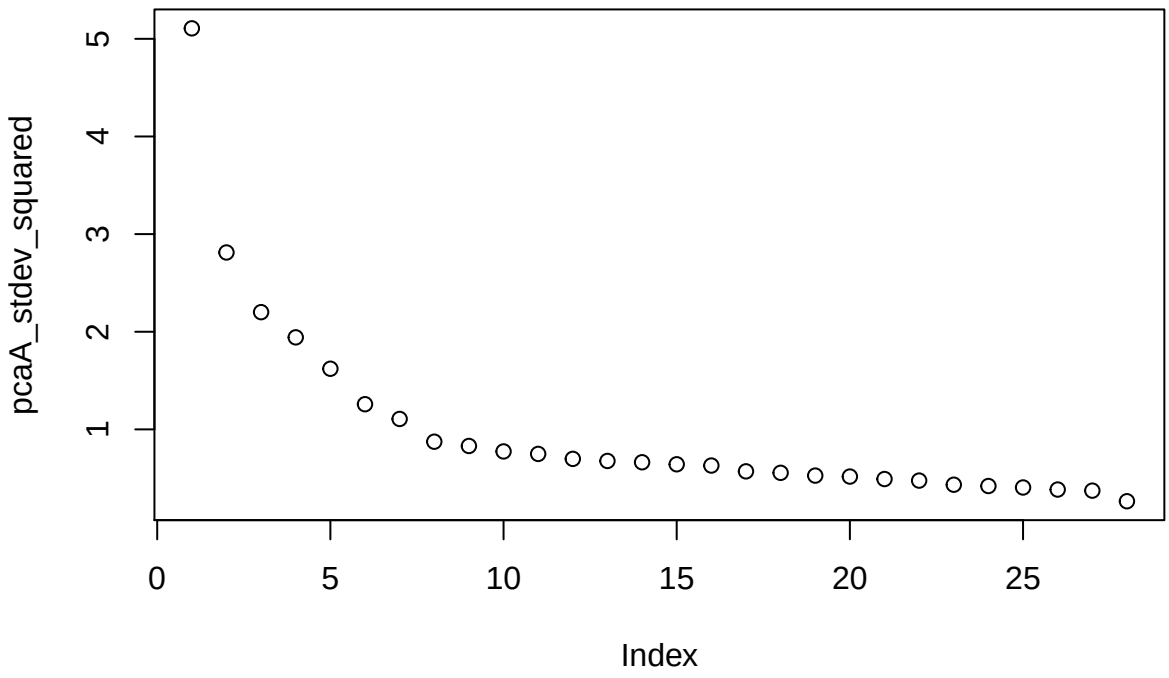
```
library(psych)
data(bfi)
bfi <- na.omit(bfi) # removing missing values
bfi_final<- scale(bfi) # scaling
bfi_final <- data.frame(bfi_final)
```

1. Examine the factor eigenvalues or variances (or the sdev or standard deviations as reported by prcomp or princomp, which you then need to square to get the variances). Plot these in a scree plot and use the "elbow" test to guess how many factors one should retain. What proportion of the total variance does your subset of variables explain?

Estimating the prinicpal components:

```
# prcomp method using SVD
pcaA<- prcomp(bfi_final)
pcaA1 <- pcaA$rotation[,1]

pcaA_stdev <- pcaA$sdev # Extracting std deviations
pcaA_stdev_squared<- pcaA_stdev^2 # Squaring to get variances
```

```
plot(pcaA_stdev_squared)
```

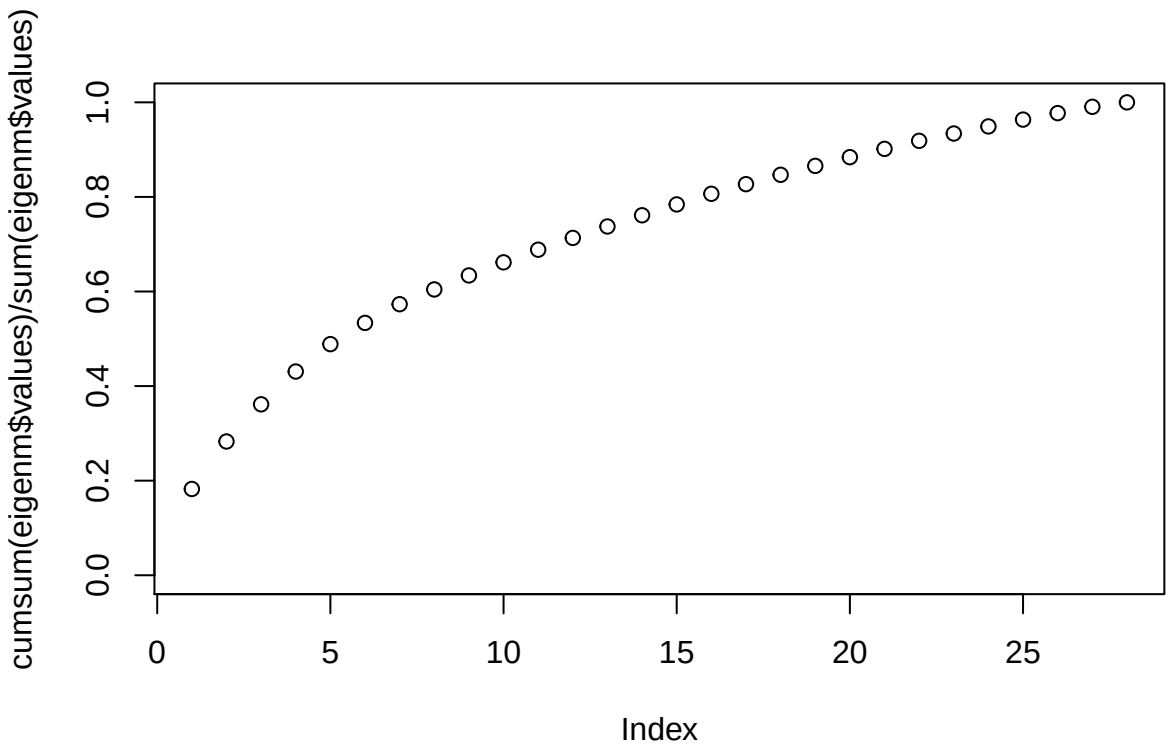


Around 7 components are capturing close to 90% of the total variance. So 7 factors can be retained.

Manual Approach:

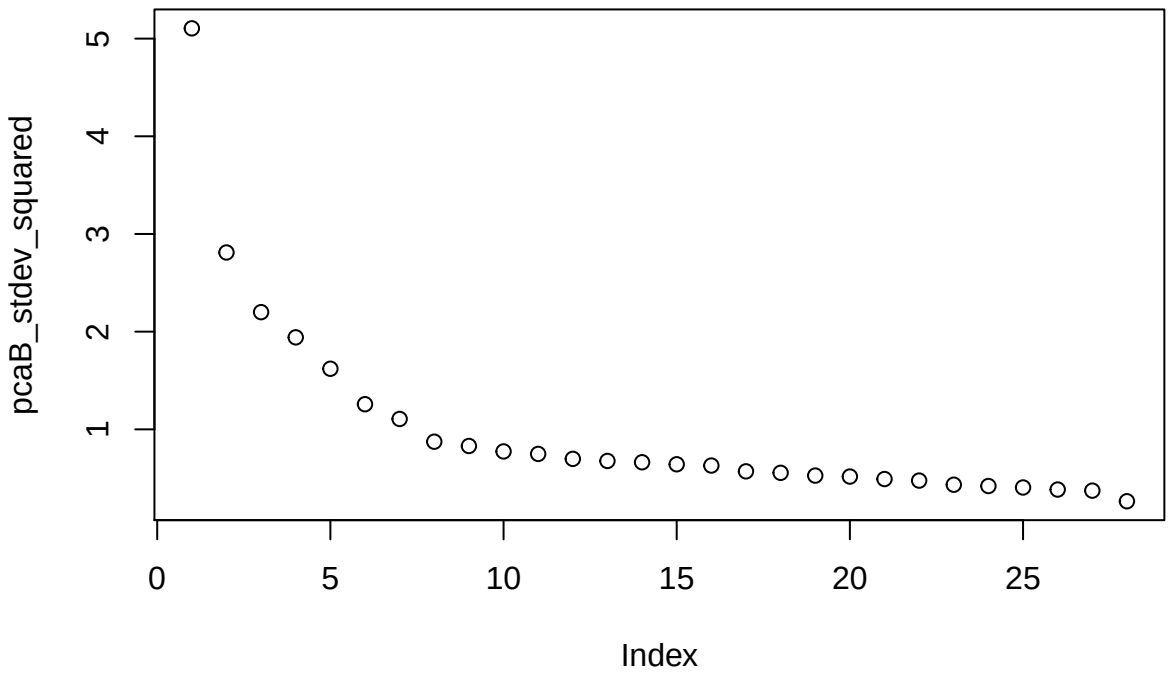
Using using cov() to estimate the covariance matrix and eigen() to get the eigenvectors:

```
# Direct eigen of cov
covm <- cov(bfi_final)
eigenm <- eigen(covm)
eigen1 <- eigenm$vectors[,1]

# Scree plot
plot(cumsum(eigenm$values)/sum(eigenm$values),ylim=c(0,1))
```



Thus, there is an elbow at around 7 and it is capturing about 55% of total variance. Approach3: Using princomp():

```
# Princomp method using eigen of cov
pcaB <- princomp(bfi_final)
pcaB1 <- pcaB$loadings[,1]
pcaB_stdev <- pcaB$sdev # Extracting std deviations
pcaB_stdev_squared<- pcaB_stdev^2 # Squaring to get variances

plot(pcaB_stdev_squared)
```



Similarly here, 7 factors can be retained as seen from the elbow.

2. Examine the loadings of the factors on the variables (sometimes called the "rotation" in the function output) - ie, the projection of the factors on the variables - focusing on just the first one or two factors. Sort the variables by their loadings, and try to interpret what the first one or two factors "mean." This may require looking more carefully into the dataset to understand exactly what each of the variables were measuring. You can find more about the data in the psych package using ?psych or visiting http://personality-project.org/ .

```
library(GPArotation)
fact <- fa(bfi_final, nfactors = 2)
fact1 <- fact$loadings[,1] # Extracting the loadings

# ordering the loadings in descending order
fact1[order(fact1)] # First factor
```

```
##          E2          E1          C5          C4          A1          N4
## -0.54279146 -0.46260890 -0.31131768 -0.29995670 -0.20204588 -0.19999938
##          O5          O2          N5          N1          N2   education
## -0.16858117 -0.09048724 -0.01966931  0.02749246  0.03410620  0.03987422
##          N3          O4         age      gender          C3          C1
##  0.05241653  0.05456320  0.09348815  0.21195808  0.28451834  0.32013686
##          O1          C2          A4          O3          A2          A5
##  0.33740992  0.34474604  0.41290110  0.44459566  0.55036780  0.58311640
##          E5          E4          A3          E3
##  0.60059910  0.60953996  0.61420838  0.63947970
```

After going through the data dictionary to understand the meanings of the variables, we can say that there

are two kinds of variables. The lowest scoring variables (denoted by negative sign) represent a person who is not a positive person, is introvert and doesnot do things in the right manner; whereas the higher scoring variables represent a positive person who is extrovert, educated, confident, cares about others, does things rightly and takes charge of things.

```
# 2nd factor
fact2 <- fact$loadings[,2]
fact2[order(fact2)]
```

```
##          age           C3           E4           A5    education
## -0.107866239 -0.084174938 -0.066036317 -0.056380169 -0.039684639
##           A4           C1           E1           O1           C2
## -0.038307029 -0.033609921 -0.022120392 -0.004569452  0.024798599
##           E5           O3           A1           O5           A3
##  0.046382289  0.053926991  0.054915879  0.061209959  0.074909133
##           E3           A2       gender           O2           E2
##  0.078721964  0.091195054  0.150477291  0.172771293  0.192402390
##           O4           C4           C5           N5           N4
##  0.226494621  0.280064018  0.314116581  0.554620629  0.583906854
##           N2           N1           N3
##  0.741035598  0.755598383  0.764963730
```

The 2nd factor looks shows a picture of 2 personalities: Consolidated, patient and energetic person Vs Impatient and Irritable person.

3. First use k-means and examine the centers of the first two or three clusters. How are they similar to and different from the factor loadings of the first couple factors?

```
kout <- kmeans(bfi_final,centers=2,nstart=25)
```

```
centroids <- kout$centers
topvars_centroid1 <- centroids[1,order(centroids[1,])]
topvars_centroid2 <- centroids[2,order(centroids[2,])]
tail(topvars_centroid1)
```

```
##        A2        E5        A3        E3        A5        E4
## 0.3526705 0.3746481 0.4078066 0.4181671 0.4501740 0.4568777
```

The first cluster suggests a negative personality who gets angry, depressed and wastes time. It captures variables which are negative on 1st PCA.

```
tail(topvars_centroid2)
```

```
##        N1        C4        E1        C5        N4        E2
## 0.4176598 0.4296434 0.4339739 0.4544198 0.5178946 0.6184365
```

The second cluster is representative of a positive perosnality who is friendly, helps, comforts and captivates others. It captures variables that are positive on 1st PCA.

Comparison with factor loadings of first two clusters:

Factors are dimensional and oppositional in nature i.e. each factor has two directions and in our case we see clear oppositions at both ends of the factors. In our case the first factor distincts between two personalities: One which is not positive, introvert and doesnot do things in right manner Vs a positive person who is confident, takes charge of things and cares about others.

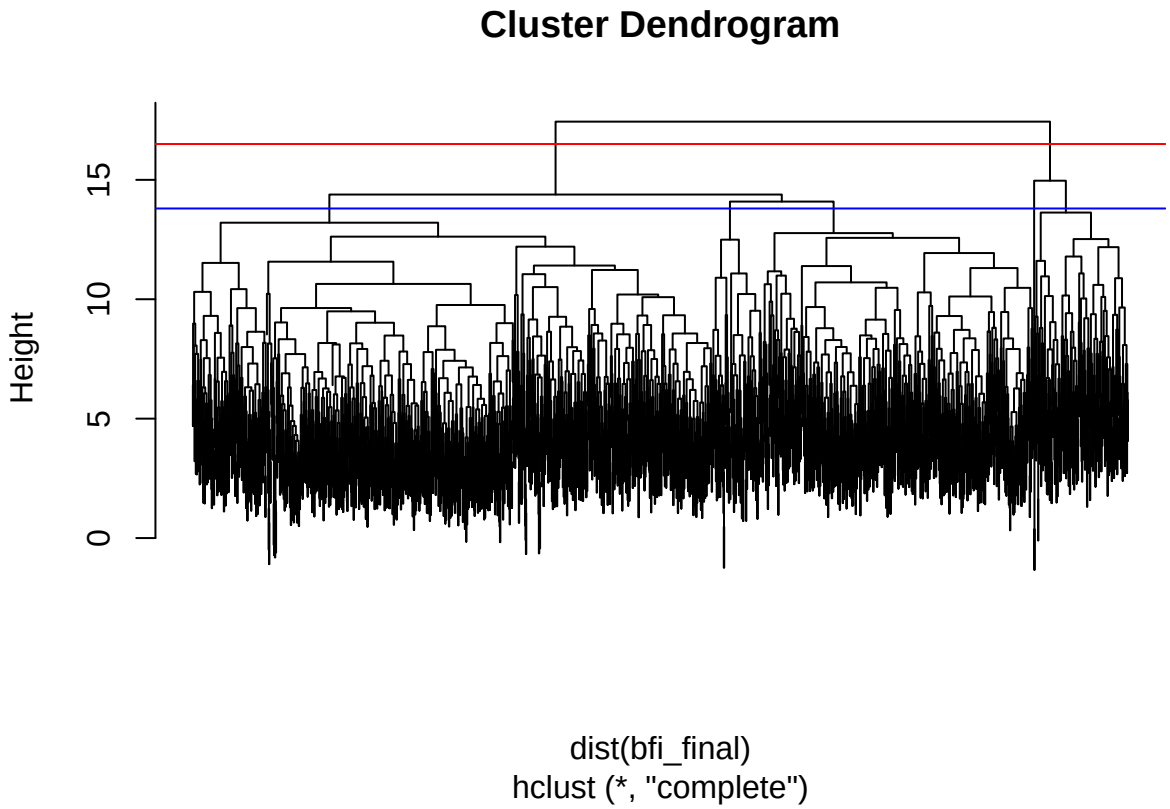
The second factor distincts between: A patient and energetic personality Vs an Pmatient and irritable personality

On the other hand Clusters are less oppositional ie. one directional and we focus more on variables that score highly ; the first cluster depicts a negative perosnality who gets angry, depressed and wastes time whereas

the second cluster depicts friendly perosnality, who is helpful and comforting to others.

4. Next use hierarchical clustering. Print the dendrogram, and use that to guide your choice of the number of clusters. Use cutree to generate a list of which clusters each observation belongs to. Aggregate the data by cluster and then examine those centers (the aggregate means) as you did in (3). Can you interpret all of them meaningfully using the methods from (3) to look at the centers?

```
hout2 <- hclust(dist(bfi_final),method="complete")
plot(hout2,labels=FALSE)
abline(a=16.5,b=0,col="red")
abline(a=13.8,b=0,col="blue")
```



We make the cuts with the red and blue lines to divide the tree into 2 and 5 clusters respectively.

```
cut <-as.vector(cutree(hout2,k=2 ))
clust_means <- aggregate(bfi_final, by=list(cut), FUN=mean)
tail(unlist(sort(clust_means[1, names(clust_means)!="Group.1"])))
```

```
##        E3        E5        A5        A2        A3        E4
## 0.1149700 0.1161775 0.1312102 0.1322361 0.1332015 0.1450269
```

```
tail(topvars_centroid2)
```

```
##        N1        C4        E1        C5        N4        E2
## 0.4176598 0.4296434 0.4339739 0.4544198 0.5178946 0.6184365
```

```
tail(unlist(sort(clust_means[2, names(clust_means)!="Group.1"])))
```

```
##        A1        C5        C4        N4        E1        E2
## 0.4304819 0.4569451 0.5025960 0.5993251 1.0223622 1.0901949
```

```
tail(topvars_centroid1)
```

```
##        A2        E5        A3        E3        A5        E4
## 0.3526705 0.3746481 0.4078066 0.4181671 0.4501740 0.4568777
```

Thus, we observe that the output matches the results from 3. Similar high scoring variables can be observed using kmeans and cut tree combined output.

5. From the factor and cluster analysis, what can you say more generally about what you have learned about your data?

Ans: Generally speaking, Factors are dimensional and oppositional in nature i.e. each factor has two directions and in our case we see clear oppositions at both ends of the factors.

On the other hand Clusters are less oppositional ie. one directional and we focus more on variables that score highly ; the first cluster depicts a negative personality who gets angry, depressed and wastes time whereas the second cluster depicts friendly personality, who is helpful and comforting to others. Also, choosing the number of factors/clusters is quite a judgement call.