Homework 11

Christina Lee

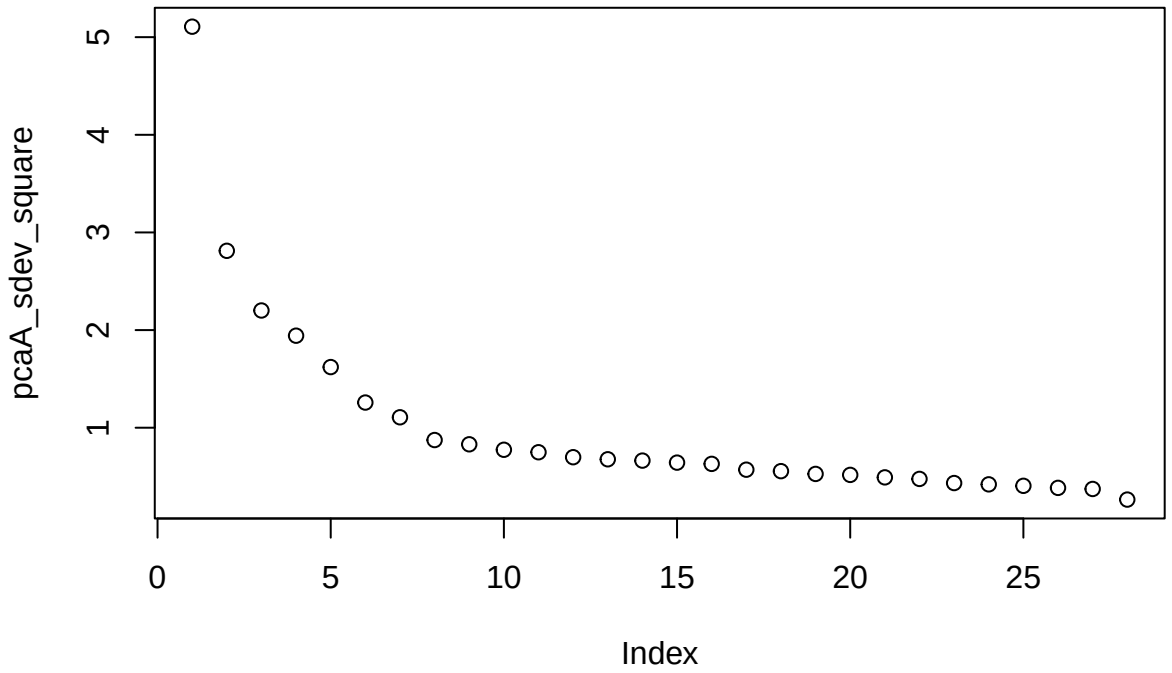
11/16/2021

First, load the data set from psych package:

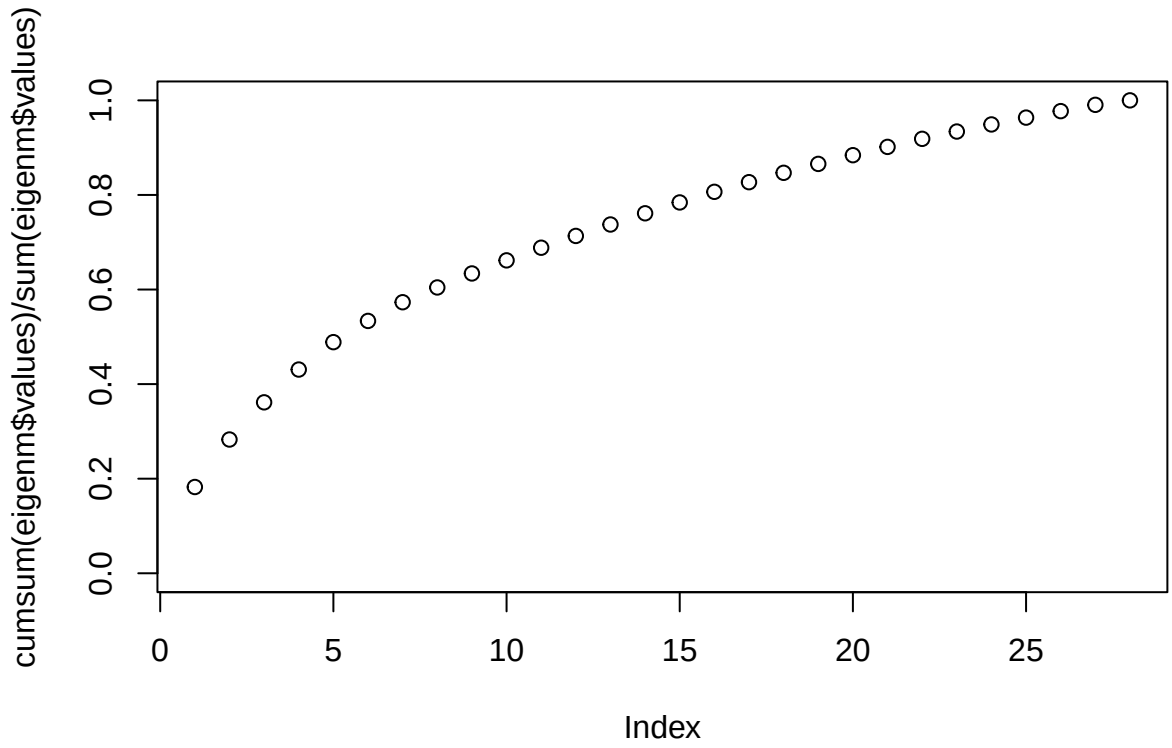
```
library(psych)
data(bfi)
bfi <- na.omit(bfi) # remove observations with NA items
bfi1 <- scale(bfi) # rescale
bfi1 <- data.frame(bfi1) # save as data frame
```

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?

```
# prcomp method using SVD
pcaA <- prcomp(bfi1)
pcaA1 <- pcaA$rotation[,1]


pcaA_sdev <- pcaA$sdev
pcaA_sdev_square <- pcaA_sdev^2


plot(pcaA_sdev_square)
```




```
covm <- cov(bfi1)
eigenm <- eigen(covm)
eigen1 <- eigenm$vectors[,1]

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



Based on the "elbow" test, my guess would be to retain 7 factors and it is capturing about 60% of the total variance.

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(bfi1, nfactors = 2)
fact1 <- fact$loadings[,1] # 1st factor


fact1[order(fact1)]
```

```
##          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 looking through the bfi data dictionary to understand what each variable mean, we can now analyze which variables go into the first factor by looking at the lowest and highest scoring variables. Therefore, while on the positive end is make friends easily, make people feel at ease, know how to captivate people, take charge, etc., the other negative side is, find it difficult to approach others, often feel blue, don't talk a lot, etc. Hence, the most important factor underlying these 25 personality items is more like introverts with both a negative and pessimistic attitude while performing certain tasks vs. extroverts with both a positive and optimistic attitude while performing certain tasks.

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


fact2[order(fact2)]
```

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

Now, we see a second dimension that is fairly different from the first: not introvert vs extrovert, but confident, organized, caring and loving person vs impatient, anxious 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(bfi1,centers=2,nstart = 25)

centroids <- kout$centers
```

```
topvar_centroid1 <- centroids[1,order(centroids[1,])]
topvar_centroid2 <- centroids[2,order(centroids[2,])]

tail(topvar_centroid1) # first cluster
```

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

The first cluster variables resembles a person that is an extrovert who knows how to interact with others, make friends easily, understands how to socialize and attract others' attention and understands how to take charge of things, which basically align with positive variables that were captured in the first principle component.

```
tail(topvar_centroid2) # second cluster
```

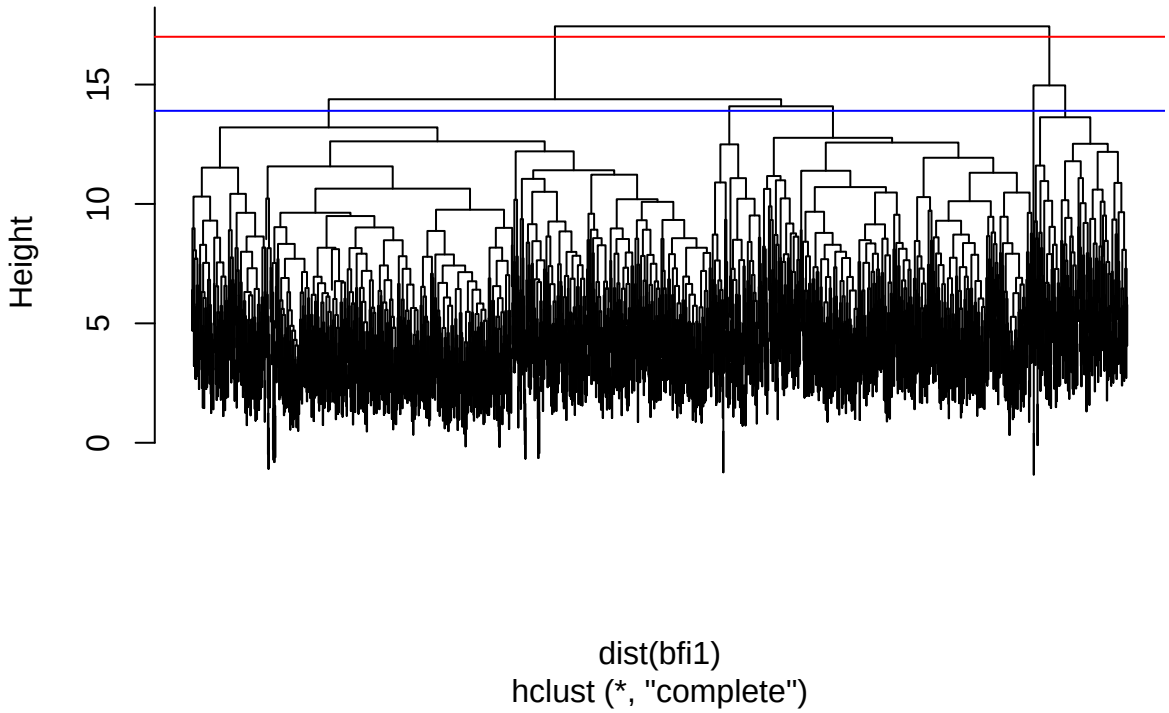
```
##        A2        E5        A3        E3        A5        E4
## 0.3526705 0.3746481 0.4078066 0.4181671 0.4501740 0.4568777
```

The second cluster variables resembles a person that is an introvert who does not know how to interact with others, doesn't talk much, find it difficult to approach others and does not understand how to take charge of things (often do things half-way manner), which again, matches with the negative variables that were captured in the first principle component.

As one can see, there is often a strong overlap between cluster and factors, however, a key difference is that factors are inherently dimensional and oppositional, such that there are two directions for every factor, and we often see clear oppositions at either end (Ex: negative/pessimistic introvert vs positive/optimistic extrovert for component 1 and confident/loving/organized vs impatient/anxious/irritable for component 2). Clusters are less oppositional, in which it only often talks about variables that score highly (as we did above), since it is less illuminating to look at variables that score weakly (those variables are not near the cluster).

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(bfi1),method= "complete")
plot(hout2, labels = FALSE) # generate cluster dendrogram

abline(a=17,b=0,col="red") # red cut
abline(a=13.9,b=0,col="blue") # blue cut
```



Here, I've made cuts with red and blue lines which it divides the tree into 2 (red) and 5 (blue) clusters.

```
cut <- as.vector(cutree(hout2,2))  #cutree at 2 clusters
clust_means <- aggregate(bfi1,by=list(cut), FUN=mean)
tail(unlist(sort(clust_means[1,names(clust_means)!= "Group.1"]))) # cluster 1
```

```
## Warning in xtfrm.data.frame(x): cannot xtfrm data frames
```

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

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

```
## Warning in xtfrm.data.frame(x): cannot xtfrm data frames
```

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

I've used the dendrogram to guide my choice of the number of clusters, which I ended up cutting the tree at 2 clusters. I then used "cutree" to generate a list of which clusters each observations belongs to and aggregate the data by cluster and examined those centers (aggregate means) as I did in Q3. It turns out that the outputs from the 2 clusters matches with the results from Q3 and very similar high scoring variables were observed. Such that, the first cluster seems to also have variables that resembles extrovert who knows how to interact with others, very social, know how to make friends, and understands how to take charge of tasks. In addition, the second cluster seems to also have variables that resembles introverts that does not know how to socialize, find it hard to approach others and does not know how to take charge of things. Therefore, I would say hierarchical clustering and the generated dendrogram do serve as a useful guide for the cluster analysis in Q3–that is to say, the suggested number of clusters to choose based on the generated dendrogram do make sense and produces similar high scoring variables as with the kmeans method.

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

In terms of factor analysis, the two underlying dimensions of this data set that may explain the majority of the variance in personality is first pessimistic and negative introverts vs optimistic and positive extroverts and the second dimension is confident, organized, caring and loving person vs impatient, anxious and irritable person. According the scree plot, we might be able to reduce the bfi dataset into 7 or so variables.

For cluster analysis, it suggests that there are two groups/clusters that might explain the majority of the variance in personality: the first cluster variables resembles an extrovert that know how to interact, socialize, make friends and understands how to take charge of things and the second cluster variables resembles an introvert that does not know how to interact with others, don't socialize, hard to approach others, and does not understand how to take charge of things (often do things half-way). As one can tell, although the results from cluster analysis are simpler, but the results are also fairly similar and agree to those variables seen in the first principle component of factor analysis. The dendrogram generated by hierarchical clustering also serves as a useful guide to determine how many clusters to choose (in our case 2 or 5, I picked 2), but which is the best choice is ultimately up to the judgement of the researcher based on his/her substantive knowledge of the results.