NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Assignment #2: Exploratory Factor Analysis

MSDS 411

Data: The data for this assignment comes from the International Personality Item Pool (ipip.ori.org) as part of the Synthetic Aperture Personality Assessment (SAPA) web based personality assessment project. The BFI data consists of the 25 personality self reported items (i.e. survey questions) obtained from 2800 subjects. Three additional demographic variables (sex, education, and age) are also included. This data is freely available in the PSYCH package of the R-Project system.

You can use the following code to obtain, load, and see the original data:

```
install.packages("psych")
library(psych)
bfi_data=bfi
bfi_data
```

The personality variables in the BFI data set are all Likert type variables measured on a scale from 1 to 6. Each variable is based on a statement, where the values for the variable are: 1 = not at all like me, and 6=totally like me. The statements and codes associated with each variable are:

A1	Am indifferent to the feelings of others.	N1	Get angry easily.
A2	Inquire about others' well-being.	N2	Get irritated easily.
A3	Know how to comfort others.	N3	Have frequent mood swings.
A4	Love children.	N4	Often feel blue.
A5	Make people feel at ease.	N5	Panic easily.
C1	Am exacting in my work.	O1	Am full of ideas.
C2	Continue until everything is perfect.	O2	Avoid difficult reading material.
C3	Do things according to a plan.	O3	Carry the conversation to a higher level.
C4	Do things in a half-way manner.	O4	Spend time reflecting on things.
C5	Waste my time.	O5	Will not probe deeply into a subject.
E1	Don't talk a lot.	Demographic variables:	
E2	Find it difficult to approach others.		
E3	Know how to captivate people.	Gender (Males = 1, Females =2)	

E4	Make friends easily.	Education (1 = HS, 2 = finished HS, 3 = some college, 4 = college graduate 5 = graduate degree)
E5	Take charge.	Age (age in years)

Source

The items are from the ipip (Goldberg, 1999). The data are from the SAPA project (Revelle, Wilt and Rosenthal, 2010), collected Spring, 2010 (https://sapa-project.org).

References

Goldberg, L.R. (1999) A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In Mervielde, I. and Deary, I. and De Fruyt, F. and Ostendorf, F. (eds) Personality psychology in Europe. 7. Tilburg University Press. Tilburg, The Netherlands.

Revelle, W., Wilt, J., and Rosenthal, A. (2010) Individual Differences in Cognition: New Methods for examining the Personality-Cognition Link In Gruszka, A. and Matthews, G. and Szymura, B. (Eds.) Handbook of Individual Differences in Cognition: Attention, Memory and Executive Control, Springer.

Revelle, W, Condon, D.M., Wilt, J., French, J.A., Brown, A., and Elleman, L.G. (2016) Web and phone based data collection using planned missing designs. In Fielding, N.G., Lee, R.M. and Blank, G. (Eds). SAGE Handbook of Online Research Methods (2nd Ed), Sage Publcations.

Assignment Tasks:

(0) Conduct a basic Exploratory Data Analysis of this data. You will notice that there are missing values indicated by NA's. To make things simple, only retain the data points that have complete information.

```
#Remove rows with missing values and keep only complete cases
bfi_data=bfi_data[complete.cases(bfi_data),]
```

Is there enough data to conduct a basic Exploratory Factor Analysis on this data? Use the 20 times number of variables rule of thumb to decide.

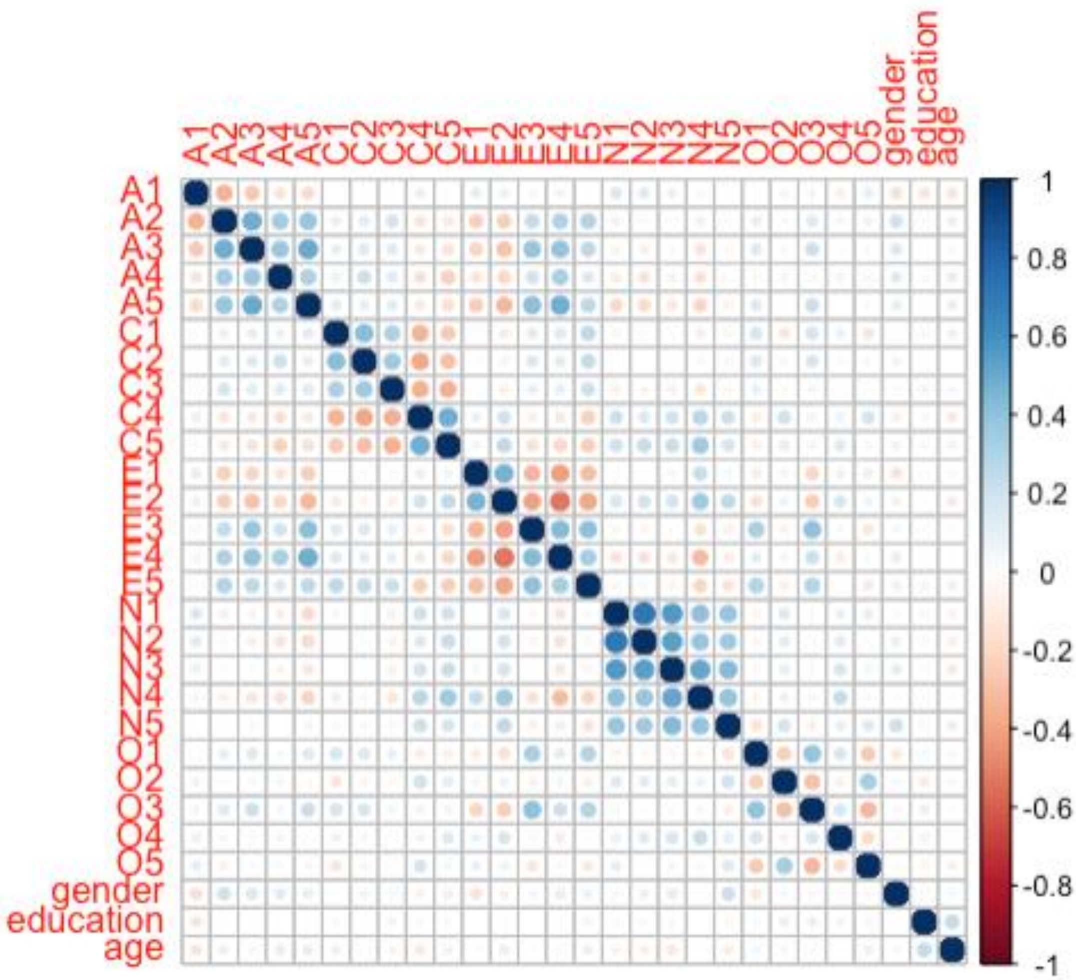
20*(# Variables)=20*28=560

There are enough items in the data to conduct this study.

Obtain the correlation matrix for the 25 personality variables. What do you notice about the correlations? Are there any discernable patterns just looking at the correlation matrix?

P

	A1	A2	A3	A4	A5	C1	C2	C3	C4	C5	E1	E2	E3	E4	E5	N1	N2	N3	N4	N5	O1	O2	O3	O4	O5	gender	education	age
A1		0.0000	0.0000	0.0000	0.0000	0.3458	0.5221	0.6554	0.0000	0.2539	0.0000	0.0000	0.0362	0.0012	0.3980	0.0000	0.0000	0.0000	0.0365	0.7073	0.9051	0.0017	0.0032	0.0000	0.0000	0.0000	0.0000	0.0000
A2	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0687	0.3342	0.0000	0.2538	0.0000	0.1251	0.0000	0.0108	0.0002	0.0000	0.3319	0.0000
A3	0.0000	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0012	0.0000	0.1770	0.0000	0.0563	0.0000	0.1451	0.0000	0.2526	0.0785	0.0000	0.8864	0.0457
A4	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0016	0.0000	0.9083	0.0445	0.0160	0.0494	0.0029	0.0972	0.0000	0.3591	0.0000
A5	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0000	0.9431	0.0000	0.8651	0.0433	0.0000	0.4706	0.0000
C1	0.3458	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000	0.0000	0.1750	0.0000	0.0000	0.0000	0.0000	0.0052	0.1936	0.5789	0.0000	0.0183	0.0000	0.0000	0.0000	0.0001	0.0000	0.8307	0.0602	0.0001
C2	0.5221	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000	0.2674	0.0012	0.0000	0.0000	0.0000	0.4409	0.9526	0.6796	0.0802	0.0194	0.0000	0.0163	0.0000	0.1401	0.0075	0.0040	0.5985	0.9750
C3	0.6554	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.3659	0.0000	0.0000	0.0000	0.0000	0.0002	0.0027	0.0005	0.0000	0.0937	0.0000	0.1243	0.0068	0.9939	0.8308	0.0379	0.0048	0.0157
C4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0009	0.0007	0.0000	0.0004	0.0956	0.0000
C5	0.2539	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0016	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0019	0.0000	0.0206	0.0000	0.0516	0.0006
E1	0.0000	0.0000	0.0000	0.0000	0.0000	0.1750	0.2674	0.3659	0.0000	0.0016		0.0000	0.0000	0.0000	0.0000	0.6715	0.7397	0.0185	0.0000	0.0349	0.0000	0.0085	0.0000	0.0000	0.0000	0.0000	0.9436	0.1189
E2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0012	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0000	0.0000	0.0002	0.0003	0.5233	0.0000
E3	0.0362	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0419	0.0076	0.5647	0.0000	0.0000	0.0000	0.0005	0.0000	0.0640	0.0000	0.0154	0.7575	0.2437
E4	0.0012	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0150	0.0000	0.0000	0.0357	0.0000	0.2139	0.5218
E5	0.3980	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0631	0.0316	0.0050	0.0000	0.0000	0.0000	0.0000	0.0000	0.3546	0.0000	0.0001	0.0069	0.0000
N1	0.0000	0.0002	0.0012	0.0000	0.0000	0.0052	0.4409	0.0002	0.0000	0.0000	0.6715	0.0000	0.0419	0.0000	0.0631		0.0000	0.0000	0.0000	0.0000	0.0135	0.0000	0.1290	0.0000	0.0000	0.0973	0.0367	0.0006
N2	0.0000	0.0687	0.0000	0.0000	0.0000	0.1936	0.9526	0.0027	0.0000	0.0000	0.7397	0.0000	0.0076	0.0000	0.0316	0.0000		0.0000	0.0000	0.0000	0.0292	0.0000	0.2649	0.0000	0.3808	0.0000	0.0765	0.0000
N3	0.0000	0.3342	0.1770	0.0016	0.0000	0.5789	0.6796	0.0005	0.0000	0.0000	0.0185	0.0000	0.5647	0.0000	0.0050	0.0000	0.0000		0.0000	0.0000	0.0279	0.0000	0.2201	0.0000	0.0099	0.0000	0.0597	0.0000
N4	0.0365	0.0000	0.0000	0.0000	0.0000	0.0000	0.0802	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.0064	0.0000	0.0051	0.0000	0.1128	0.3247	0.5029	0.2388
N5	0.7073	0.2538	0.0563	0.9083	0.0003	0.0183	0.0194	0.0937	0.0000	0.0000	0.0349	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0270	0.0000
O1	0.9051	0.0000	0.0000	0.0445	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0135	0.0292	0.0279	0.0064	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000	0.1016	0.0185
O2	0.0017	0.1251	0.1451	0.0160	0.9431	0.0000	0.0163	0.1243	0.0000	0.0000	0.0085	0.0003	0.0005	0.0150	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.0001	0.0000	0.0459	0.0000	0.0362
O3	0.0032	0.0000	0.0000	0.0494	0.0000	0.0000	0.0000	0.0068	0.0009	0.0019	0.0000	0.0000	0.0000	0.0000	0.0000	0.1290	0.2649	0.2201	0.0051	0.0000	0.0000	0.0000		0.0000	0.0000	0.0380	0.0000	0.2488
O4	0.0000	0.0108	0.2526	0.0029	0.8651	0.0001	0.1401	0.9939	0.0007	0.0000	0.0000	0.0000	0.0640	0.0000	0.3546	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000		0.0000	0.0910	0.0041	0.9634
O5	0.0000	0.0002	0.0785	0.0972	0.0433	0.0000	0.0075	0.8308	0.0000	0.0206	0.0000	0.0002	0.0000	0.0357	0.0000	0.0000	0.3808	0.0099	0.1128	0.0000	0.0000	0.0000	0.0000	0.0000		0.0804	0.0053	0.0000
gender	0.0000	0.0000	0.0000	0.0000	0.0000	0.8307	0.0040	0.0379	0.0004	0.0000	0.0000	0.0003	0.0154	0.0000	0.0001	0.0973	0.0000	0.0000	0.3247	0.0000	0.0000	0.0459	0.0380	0.0910	0.0804		0.7866	0.0263
education	0.0000	0.3319	0.8864	0.3591	0.4706	0.0602	0.5985	0.0048	0.0956	0.0516	0.9436	0.5233	0.7575	0.2139	0.0069	0.0367	0.0765	0.0597	0.5029	0.0270	0.1016	0.0000	0.0000	0.0041	0.0053	0.7866		0.0000
age	0.0000	0.0000	0.0457	0.0000	0.0000	0.0001	0.9750	0.0157	0.0000	0.0006	0.1189	0.0000	0.2437	0.5218	0.0000	0.0006	0.0000	0.0000	0.2388	0.0000	0.0185	0.0362	0.2488	0.9634	0.0000	0.0263	0.0000	



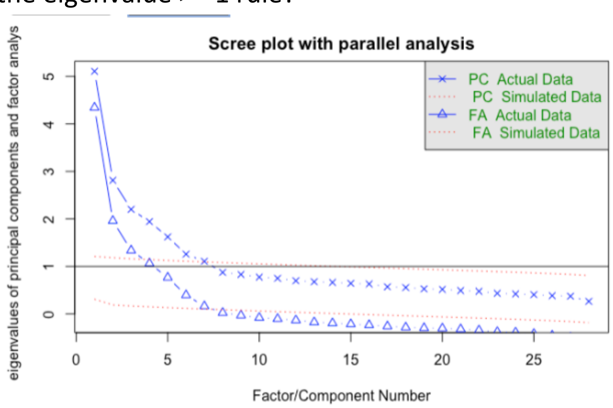
There appears to be similar correlations between the N questions to eachother, as that is the most prevalent square of blue dots visible here. Half of the E questions are positively correlated while the other half is negatively correlated. The same seems to go for the C questions. For the A questions, A1 seems to be negatively correlated against the other A questions while the others appear to be positively correlated.

You will want to save the correlations as a matrix. If you can figure out how to do this directly, great. If you don't know, or can't find it quickly, then you'll have to do it by hand. Here is some code to help. First, load the correlation matrix into R. How do we do that? Start with a vector of values, and then read that vector of values into a matrix object. Here is an example to help you figure out what to do for your data.

```
cor.values <- c(1.000,0.210,0.370,-0.32,0.000,-0.31,-0.26,0.090,-0.38,
                0.210,1.000,0.090,-0.29,0.120,-0.30,-0.14,0.010,-0.39,
                0.370,0.090,1.000,-0.31,-0.04,-0.30,-0.11,0.120,-0.39,
                -0.32,-0.29,-0.31,1.00,-0.16,0.25,-0.13,-0.14,0.900,
                0.00,0.120,-0.04,-0.16,1.000,-0.20,-0.03,-0.08,-0.38,
                -0.31,-0.30,-0.30,0.25,-0.20,1.000,-0.24,-0.16,0.180,
                -0.26,-0.14,-0.11,-0.13,-0.03,-0.24,1.000,-0.20,0.040,
                0.090,0.010,0.120,-0.14,-0.08,-0.16,-0.20,1.000,-0.24,
                -0.38,-0.39,-0.39,0.900,-0.38,0.180,0.040,-0.24,1.000
                );

# How do we put these correlation values into a correlation matrix?;
help(matrix)
cor.matrix <- matrix(cor.values,nrow=9,ncol=9,byrow=TRUE);

# Check that object is a matrix object;
is.matrix(cor.matrix)

# Check that matrix is symmetric;
# This check helps check for typos;
isSymmetric(cor.matrix)
```

We can check most data types in R using an is.* function. We type cast in R using an as.* function.

(1) Obtain the eigenvalues and eigenvectors of the correlation matrix. You can obtain this information in a number of different ways. You could use direct matrix functions or you could use the fa() function in the PSYCH package. Also, the Classroom may have other ways – check those resources. The goal for this task is to obtain a scree plot to go along with the eigenvalues. How many factors should you retain using the scree plot rule? How many factors should you retain to account for 90% of the overall variability? How many factors should you retain using the eigenvalue >= 1 rule?




```
You specified the number of subjects, implying a correlation matrix, but do not have a
correlation matrix, correlations found Parallel analysis suggests that the number of factors =  7
and the number of components =  6
```

According to the scree plot, Parallel Analysis suggests 7 number of factors and 6 components.

To account for 90% of the overall variability, we should retain 22 factors.
According to the eigenvalue >=1 rule, we should retain 4 factors.

(2) Use the eigenvalue >= 1 rule for the number of factors to retain.
Estimate a factor model for the number of factors with eigenvalues greater than 1. Use maximum likelihood factor analysis with a VARIMAX rotation. Report the factor loadings table and interpret each factor. What cutoff value did you use for deciding which loadings were sufficiently large for interpretation? What proportion of overall variability is explained by this model? Is that sufficient to you? You can use the fa() function of the PSYCH package or factanal() from the base STAT system.

```
factors_data <- fa(r = cor_matrix, nfactors = 6)

factors_data <- factanal(covmat=cor_matrix, n.obs=1442,
factors=3, rotation='varimax');
names(f.1)
```

4 of the eigenvalues are greater than 1 as seen in the above graph, hence 4 factors would be retained according to this rule.

```
Call:
factanal(factors = 4, covmat = corrmatrix, n.obs = 2236, rotation = "varimax")

Uniquenesses:
   A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1    E2    E3    E4    E5    N1    N2    N3    N4    N5    O1    O2    O3    O4
0.943 0.706 0.602 0.736 0.573 0.673 0.610 0.685 0.502 0.577 0.721 0.585 0.531 0.473 0.626 0.360 0.377 0.463 0.590 0.679 0.668 0.730 0.510 0.870
   O5 gender education   age
0.735  0.890     0.979 0.975

Loadings:
          Factor1 Factor2 Factor3 Factor4
A1        -0.207   0.119
A2         0.520           0.149
A3         0.621           0.111
A4         0.423           0.220  -0.173
A5         0.632  -0.148
C1                         0.526   0.208
C2         0.103           0.603
C3                         0.551
C4                 0.224  -0.659
C5        -0.172   0.272  -0.565
E1        -0.519
E2        -0.585   0.224  -0.102  -0.112
E3         0.609                   0.306
E4         0.708  -0.137
E5         0.465           0.305   0.250
N1                 0.796
N2                 0.786
N3                 0.729
N4        -0.262   0.556  -0.178
N5                 0.519          -0.216
O1         0.200           0.103   0.531
O2                 0.157  -0.122  -0.475
O3         0.326                   0.617
O4                 0.206           0.286
O5                                -0.505
gender     0.210   0.121   0.111  -0.198
education                          0.137
age               -0.106

               Factor1 Factor2 Factor3 Factor4
SS loadings      3.320   2.705   1.998   1.609
Proportion Var   0.119   0.097   0.071   0.057
Cumulative Var   0.119   0.215   0.287   0.344

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 3221.31 on 272 degrees of freedom.
The p-value is 0
```

Models:
M1=0.52*A2+0.621*A3+0.632*A5+0.609*E3+0.708*E4
M2=0.796*N1+0.786*N2+0.729*N3+0.556*N4+0.519*N5
M3=0.526*C1+0.603*C2+0.551*C3
M4=0.531*O1+0.617*O3

Cutoff value:
0.5

Proportion of variance explained:

About 34.4% of the variance is explained in the 4 factors which is quite low. We need to add more factors to account for more of the variance.

Does the statistical inference for the maximum likelihood factor analysis suggest that you have the correct number of factors to describe this correlation matrix? What is the null hypothesis for the chi-square test statistic? Do we reject or fail to reject this null hypothesis? Note that this hypothesis cannot be expressed in statistical notation like most hypotheses tests in Predict 410. (Hint: See Section 11.5 of Everitt.)

The statistical inference for the maximum likelihood factor analysis suggests that we do not have the correct number of factors to describe the correlation matrix since such a small amount of the variance is described in there 4 factors. It is still a lot of variances accounted for being only 4 factors out of the original 25, but more factors would be helpful.

Null:

(3) The VARIMAX factor rotation is an example of an orthogonal factor rotation. We also have oblique factor rotations. One example of an oblique factor rotation is the PROMAX rotation. Fit the same model from Task 2) but this time use the PROMAX rotation using maximum likelihood factor analysis.

```
Call:
factanal(factors = 4, covmat = corrmatrix, n.obs = 2236, rotation = "promax",     fm = "ml")

Uniquenesses:
    A1     A2     A3     A4     A5     C1     C2     C3     C4     C5     E1     E2     E3     E4     E5     N1     N2     N3     N4     N5     O1     O2     O3     O4
 0.943  0.706  0.602  0.736  0.573  0.673  0.610  0.685  0.502  0.577  0.721  0.585  0.531  0.473  0.626  0.360  0.377  0.463  0.590  0.679  0.668  0.730  0.510  0.870
    O5  gender education    age
 0.735  0.890  0.979  0.975

Loadings:
          Factor1 Factor2 Factor3 Factor4
A1        -0.216   0.103
A2         0.526
A3         0.640
A4         0.414           0.169  -0.190
A5         0.650
C1                         0.552   0.157
C2                 0.174   0.646
C3                         0.592
C4                 0.121  -0.702
C5                 0.175  -0.584
E1        -0.551           0.135  -0.114
E2        -0.580   0.161          -0.112
E3         0.607                   0.326
E4         0.737
E5         0.403   0.137   0.229   0.240
N1                 0.800
N2                 0.792
N3                 0.730
N4        -0.211   0.516  -0.122
N5                 0.518          -0.188
O1         0.135                   0.535
O2         0.156   0.149  -0.106  -0.460
O3         0.277                   0.635
O4                 0.198           0.304
O5                                -0.502
gender     0.225   0.160   0.102  -0.200
education                          0.135
age

               Factor1 Factor2 Factor3 Factor4
SS loadings      3.285   2.610   2.077   1.614
Proportion Var   0.117   0.093   0.074   0.058
Cumulative Var   0.117   0.211   0.285   0.342

Factor Correlations:
        Factor1 Factor2 Factor3 Factor4
Factor1   1.000 -0.1861  -0.412 -0.1213
Factor2  -0.186  1.0000   0.236  0.0864
Factor3  -0.412  0.2358   1.000  0.2168
Factor4  -0.121  0.0864   0.217  1.0000

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 3221.31 on 272 degrees of freedom.
The p-value is 0
```

a. Does this model have better interpretability than the Task 2 Model with the VARIMAX rotation?

Models:

M1=0.526*A2+0.64*A3+0.65*A5+0.607*E3+0.737*E4
M2=0.8*N1+0.792*N2+0.73*N3+0.516*N4+0.518*N5
M3=0.552*C1+0.646*C2+0.592*C3
M4=0.535*O1+0.635*O3
The interpretability seems to be almost identical with only the weights on each variable slightly changed.

b. Does the statistical inference for this maximum likelihood factor analysis suggest that this model has the correct number of factors to describe this correlation matrix? Should the factor rotation affect the statistical inference for the number of factors?
This still suggests that four factors are low to describe a majority of the variance of the data. About the same percentage of variance is still described. Adding more factors should increase the variance accounted for.

(4) Can we find the correct number of factors to describe this correlation matrix? Fit factor models using a VARIMAX rotation for k=1 through max (number of factors to retain from task 1 computations). For each factor model fit, use the factor loadings to interpret the individual factors.
What cutoff value did you use for deciding which loadings were sufficiently large for interpretation?
0.5 (same as before)
Some of these will be easier to interpret than others.
Which model is the easiest to interpret?
The first model (M1) is what is easiest to interpret due to how many coefficients that model has. In the prior models, the M1 had 5 coefficients within, accounting for a lot of the variance in just one model. However, it is still not enough of the variance accounted for as it was still only 11% of the variance, which is a lot for one model.
Even when run with only 1 factor, it comes up with the greatest number of coefficients (6), but still does not account for most of the variance.
Do any of these models represent the correct number of factors based on the inference results?

```
Call:
factanal(factors = 15, covmat = corrmatrix, n.obs = 2236, rotation = "varimax",     fm = "ml")
```

Uniquenesses:

A1	A2	A3	A4	A5	C1	C2	C3	C4	C5	E1	E2	E3	E4	E5	N1	N2	N3	N4	N5	O1	O2	O3	O4	O5
0.702	0.263	0.358	0.561	0.487	0.574	0.281	0.656	0.407	0.320	0.453	0.395	0.421	0.278	0.509	0.285	0.156	0.387	0.416	0.480	0.639	0.642	0.462	0.005	0.632

Loadings:

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9	Factor10	Factor11	Factor12	Factor13	Factor14	Factor15
A1	0.108					-0.505				0.116					
A2		0.142	-0.157	0.225		0.734				0.278	0.112		0.103		
A3		0.102	-0.123	0.585		0.464					0.166				0.136
A4		0.155	-0.109	0.260		0.264		-0.111			0.489				
A5	-0.157	0.109	-0.221	0.571		0.279									
C1		0.605			-0.163										
C2		0.732									0.215			0.330	
C3		0.535						-0.166		0.100					
C4	0.167	-0.595	0.123		0.196	-0.131		0.245	0.141	0.148					0.129
C5	0.226	-0.439	0.105					0.626			-0.122				
E1			0.697	-0.120		-0.125									
E2	0.181		0.644	-0.234				0.129	0.182			-0.100			
E3			-0.332	0.596	-0.202					0.193					-0.103
E4	-0.125	0.115	-0.560	0.454							0.161		0.366		
E5		0.316	-0.384	0.245	-0.156				-0.151	0.351					
N1	0.809					-0.131									0.103
N2	0.855											-0.261			
N3	0.721								0.127			0.234			
N4	0.541	-0.116	0.305					0.175	0.170			0.318			
N5	0.474				0.192				0.482						
O1		0.121		0.263	-0.427		0.102			0.268					
O2	0.106				0.568										
O3			-0.185	0.329	-0.552					0.158				0.152	0.116
O4	0.134		0.122		-0.184		0.956								
O5					0.573	-0.106									

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9	Factor10	Factor11	Factor12	Factor13	Factor14	Factor15
SS loadings	2.632	1.978	1.764	1.707	1.397	1.263	0.986	0.572	0.403	0.403	0.403	0.263	0.194	0.169	0.098
Proportion Var	0.105	0.079	0.071	0.068	0.056	0.051	0.039	0.023	0.016	0.016	0.016	0.011	0.008	0.007	0.004
Cumulative Var	0.105	0.184	0.255	0.323	0.379	0.430	0.469	0.492	0.508	0.524	0.540	0.551	0.559	0.565	0.569

```
Test of the hypothesis that 15 factors are sufficient.
The chi square statistic is 22.23 on 30 degrees of freedom.
The p-value is 0.845
```

The model with 15 factors tends to be the best performing model, as it has the highest designated p-value at the end, meaning it has the greatest chance of the null hypothesis being rejected. It appears to account for the most variance for this model all while reducing the dimensionality of the data by 10 variables. Even in this model, we would only use 11 of the factors as models due to two of them being insignificant (factors 9, 10, 11,12,13,14,15).

Model:

N1*0.809+N2*0.855 +N3*0.721+N4*0.541

C1*0.605+C2*0.732 +C3*0.535

E1*0.697+E2*0.644-E4*0.560

A3*0.585+A5*0.571+E3*0.596

O2*0.568-O3*0.522+O5*0.573

-A1*0.505+A2*0.734

C5*0.626

A2*0.55+E5*0.629

O4*1.005

C5*0.722

N5*0.622

E4*0.583

A4*0.536

The p-value gets into the rejecting the null hypothesis region at 12 factors with 3 factors dropped , but 15 factors with two dropped increases the amount of variance accounted for within the models.

(5) The researchers who commissioned the BFI data collection had a theory about personalities. According to their theory, there are 5 factors contained in this data. They are: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness. The variable naming convention (A, C, E, N, O) indicates which variables should band together to measure the associated latent trait. How does your easiest to interpret or best fitting model from Task 4) compare to this structure?

My model does this well to a certain degree, as at many times it separates one variable into its own model to account for its variance on its own better. Still, the C variables, E variables, and the N variables are banded together quite well, while the A and O variables are quite separated into their own models. The O variables group together)2 and O5, but separate the others, while C1, C2, and C3 are all together. A3 and A5 are combined with E3.

C1*0.705+C2*1.481+C3*0.511-C4*0.714

N1*0.797+N2*1.049+C3*0.502

A3*0.648+A5*0.605+E3*0.635

A2*0.992

C2*1.044

E1*0.744+E2*0.62

O2*0.614+O5*0.544

A2*0.55+E5*0.629

O4*1.005

C5*0.722

N5*0.622

E4*0.583

A4*0.536

(6) Just to be certain, refit a 5-factor model using the VARIMAX rotation and maximum likelihood factor analysis. Save the Factor Scores as variables to the BFI dataset. Use the Factors and response variables to determine:

a. If there are gender differences

```
Call:
lm(formula = gender ~ f1 + f2 + f3 + f4, data = bfi_data2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.0224 -0.5563  0.2297  0.3388  0.7543 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 1.069799   0.057436  18.626 < 0.0000000000000002
f1          0.020946   0.002299   9.109 < 0.0000000000000002
f2          0.013031   0.004378   2.977             0.002942
f3          0.011862   0.003369   3.521             0.000438
f4          0.040167   0.005553   7.233    0.000000000000623

Residual standard error: 0.4545 on 2525 degrees of freedom
  (270 observations deleted due to missingness)
Multiple R-squared:  0.06426,	Adjusted R-squared:  0.06278 
F-statistic: 43.35 on 4 and 2525 DF,  p-value: < 0.00000000000000022
```

MSE= 0.2061455

AIC= 2774.562

BIC= 2814.549

There appears to be separation between genders as the MSE value in this model is very low at 0.2.

b. If personality is related to education

```
Call:
lm(formula = education ~ f1 + f2 + f3 + f4 + f5, data = bfi_data2)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4670 -0.3130 -0.1477  0.7741  2.1216 

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  2.971030   0.159992  18.570 < 0.0000000000000002
f1          -0.010979   0.005975  -1.837               0.0663
f2          -0.018420   0.011606  -1.587               0.1126
f3           0.001604   0.008863   0.181               0.8564
f4           0.004306   0.014796   0.291               0.7711
f5           0.067996   0.016502   4.121            0.0000392

Residual standard error: 1.107 on 2230 degrees of freedom
Multiple R-squared:  0.01013,	Adjusted R-squared:  0.007906 
F-statistic: 4.562 on 5 and 2230 DF,  p-value: 0.0003836

[1] 1.222768
[1] 6809.191
[1] 6849.178
```

The MSE value at 1.223 is still quite low, indicating personality is related quite significantly to education.

c. If personality types are related to age

```
Call:
lm(formula = age ~ f1 + f2 + f3 + f4 + f5, data = bfi_data2)

Residuals:
    Min      1Q  Median      3Q     Max 
-26.388  -8.182  -3.004   5.934  56.279 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 26.25386    1.52533  17.212 < 0.0000000000000002
f1          -0.21533    0.05696  -3.780             0.000161
f2          -0.07405    0.11065  -0.669             0.503426
f3           0.20510    0.08450   2.427             0.015291
f4           0.40301    0.14107   2.857             0.004318
f5           0.28981    0.15733   1.842             0.065595

Residual standard error: 10.56 on 2230 degrees of freedom
Multiple R-squared:  0.02229,	Adjusted R-squared:  0.02009 
F-statistic: 10.17 on 5 and 2230 DF,  p-value: 0.000000001207

[1] 111.1414
[1] 16892.85
[1] 16932.84
```

The MSE vales and the AIC and BIC values are quite high, indicating that there is not that high of a relationship between age and personality types.

What do you conclude?

Personality and education are related quote significantly, and there are visible differences in gender, however there is no significant relationship between personality types and age.

(7) Please write a reflection on your experiences.

This was a more open-ended assignment than the last, allowing us more freedom on how we are to do our analysis, especially at the end in how we determine whether or not relationships between variables exist or not. It is always great to work with simplifying data and finding more ways to do so as it is always more helpful to work with more simplified data. It was difficult to determine how to analyze relationships at the end. I wish there were some suggestions on how we could approach this and determine relationships efficiently. Overall, this was still a great and challenging assignment.

Assignment Document:

All assignment reports should answer each of the questions separately. Please be sure to clearly indicate which question is being addressed. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary R-code, intermediary computations, R-results, or non-essential information. The document should be submitted in pdf format. Name your file Assign2_LastName.pdf.