Exploratory Factor Analysis

Serra Uzun, MSDS_411 FALL 2020
10/04/2020

Introduction

Large datasets often make it difficult to determine the structure of all variables and relationships between variables. Exploratory Factor Analysis (EFA) is useful in uncovering the underlying connections and loadings in a large dataset with complex structures. To determine the underlying relationships in the SAPA (Synthetic Aperture Personality Assessment) survey results data, we will be conducting an EFA, following an exploratory analysis of the dataset and look into the correlations between all and selected variables. We aim to compute the factor loadings, discover underlying connections, and determine if variables can be banded together to measure the associated latent trait.

The Dataset *(Question #0)*

The SAPA dataset (BFI dataset) was obtained through the International Personality Item Pool and is built into R with the PSYCH package. The dataset has 2,800 observations and 28 variables. The 25 of the 28 total variables are Likert type personality variables on a scale to 1 to 6, whereas the additional three variables are demographic related, that are gender, education and age. The 25 personality trait variables are in 5 sub-groups as Agreeable, Conscientiousness, Extraversion, Neuroticism, and Openness (A, C, E, N, O). All variables types are integer. Below, in Table 1, are the variables and their descriptions:

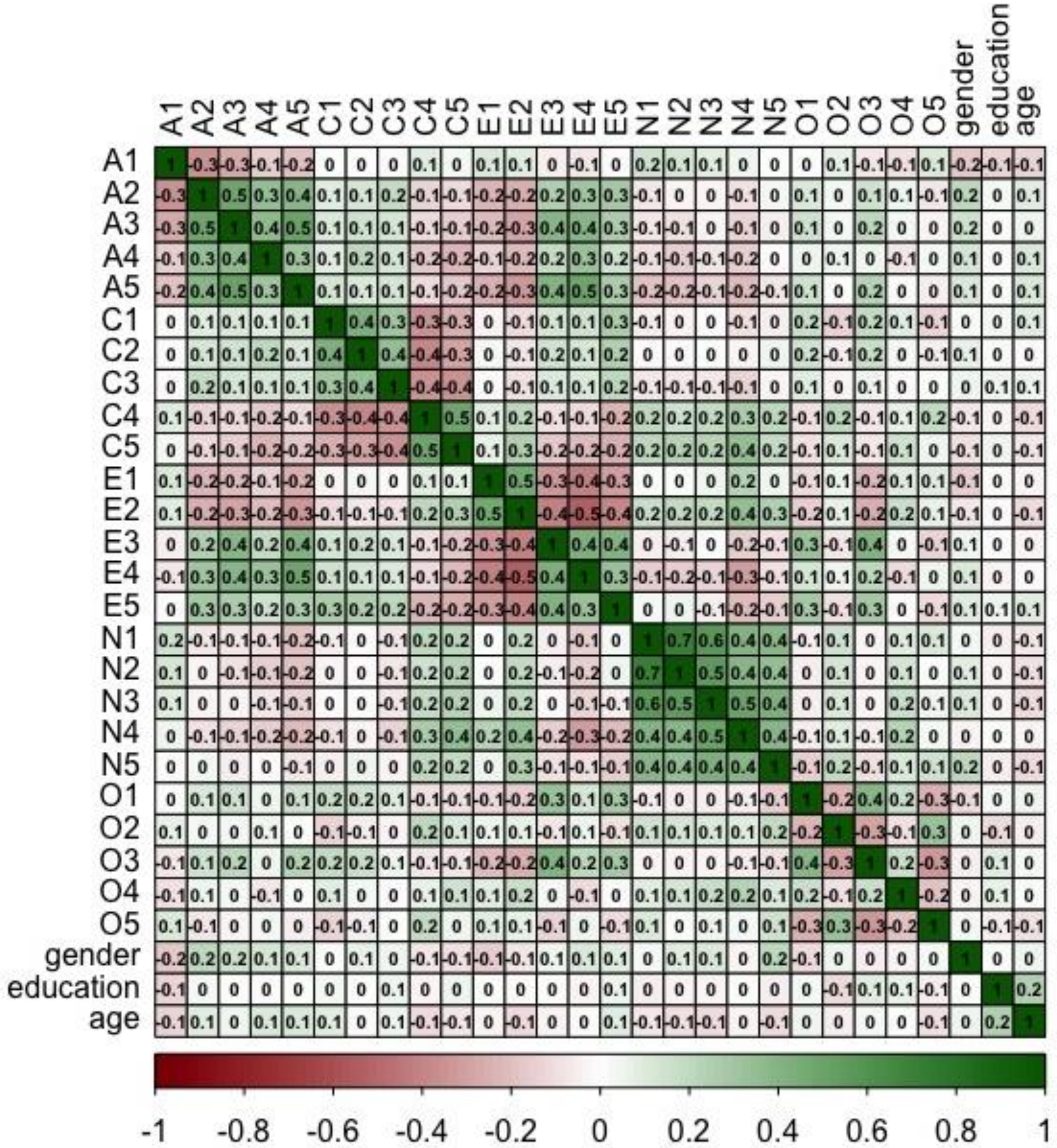
Variables	Description	Variables	Description
A1	Am indifferent to the feelings of others.	N1	Get angry easily.
A2	Inquire about others' well-being.	N2	Get irritated easily.
A3	Know how to comfort others.	N3	Have frequent mood swings.
A4	Love children.	N4	Often feel blue.
A5	Make people feel at ease.	N5	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.	Gender	(Males = 1, Females =2)
E2	Find it difficult to approach others.	Education	(1 = HS, 2 = finished HS, 3 = some college, 4 = college graduate 5 = graduate degree)
E3	Know how to captivate people.		
E4	Make friends easily.	Age	(age in years)
E5	Take charge.		

Table 1: BFI Dataset Variables and Descriptions

Upon cleaning the dataset from missing values, the number of observations went from 2,800 down to 2,236, and 564 observations are removed. The 2236 observation dataset is a sufficient size dataset for the EFA and meets the rule of thumb for EFAs where the total number of observations need to be at least 20 times the number of variables.

Correlation Analysis for Returns *(Question #0 continued)*

The correlation matrix and the plot have given us insight into the weak and strong relationships between and within each personality trait subgroup. Our threshold for a significant correlation is set to 0.5 or higher, and with that, we see from Figure 1 that the majority of our variables have less than 0.5 correlation with one another.



	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	1	-0.3	-0.3	-0.1	-0.2	0	0	0	0.1	0	0.1	0.1	0	-0.1	0	0.2	0.1	0.1	0	0	0	0.1	-0.1	-0.1	0.1	-0.2	-0.1	-0.1
A2	-0.3	1	0.5	0.3	0.4	0.1	0.1	0.2	-0.1	-0.1	-0.2	-0.2	0.2	0.3	0.3	-0.1	0	0	-0.1	0	0.1	0	0.1	0.1	-0.1	0.2	0	0.1
A3	-0.3	0.5	1	0.4	0.5	0.1	0.1	0.1	-0.1	-0.1	-0.2	-0.3	0.4	0.4	0.3	-0.1	-0.1	0	-0.1	0	0.1	0	0.2	0	0	0.2	0	0
A4	-0.1	0.3	0.4	1	0.3	0.1	0.2	0.1	-0.2	-0.2	-0.1	-0.2	0.2	0.3	0.2	-0.1	-0.1	-0.1	-0.2	0	0	0.1	0	-0.1	0	0.1	0	0.1
A5	-0.2	0.4	0.5	0.3	1	0.1	0.1	0.1	-0.1	-0.2	-0.2	-0.3	0.4	0.5	0.3	-0.2	-0.2	-0.1	-0.2	-0.1	0.1	0	0.2	0	0	0.1	0	0.1
C1	0	0.1	0.1	0.1	0.1	1	0.4	0.3	-0.3	-0.3	0	-0.1	0.1	0.1	0.3	-0.1	0	0	-0.1	0	0.2	-0.1	0.2	0.1	-0.1	0	0	0.1
C2	0	0.1	0.1	0.2	0.1	0.4	1	0.4	-0.4	-0.3	0	-0.1	0.2	0.1	0.2	0	0	0	0	0	0.2	-0.1	0.2	0	-0.1	0.1	0	0
C3	0	0.2	0.1	0.1	0.1	0.3	0.4	1	-0.4	-0.4	0	-0.1	0.1	0.1	0.2	-0.1	-0.1	-0.1	-0.1	0	0.1	0	0.1	0	0	0	0.1	0.1
C4	0.1	-0.1	-0.1	-0.2	-0.1	-0.3	-0.4	-0.4	1	0.5	0.1	0.2	-0.1	-0.1	-0.2	0.2	0.2	0.2	0.3	0.2	-0.1	0.2	-0.1	0.1	0.2	-0.1	0	-0.1
C5	0	-0.1	-0.1	-0.2	-0.2	-0.3	-0.3	-0.4	0.5	1	0.1	0.3	-0.2	-0.2	-0.2	0.2	0.2	0.2	0.4	0.2	-0.1	0.1	-0.1	0.1	0	-0.1	0	-0.1
E1	0.1	-0.2	-0.2	-0.1	-0.2	0	0	0	0.1	0.1	1	0.5	-0.3	-0.4	-0.3	0	0	0	0.2	0	-0.1	0.1	-0.2	0.1	0.1	-0.1	0	0
E2	0.1	-0.2	-0.3	-0.2	-0.3	-0.1	-0.1	-0.1	0.2	0.3	0.5	1	-0.4	-0.5	-0.4	0.2	0.2	0.2	0.4	0.3	-0.2	0.1	-0.2	0.2	0.1	-0.1	0	-0.1
E3	0	0.2	0.4	0.2	0.4	0.1	0.2	0.1	-0.1	-0.2	-0.3	-0.4	1	0.4	0.4	0	-0.1	0	-0.2	-0.1	0.3	-0.1	0.4	0	-0.1	0.1	0	0
E4	-0.1	0.3	0.4	0.3	0.5	0.1	0.1	0.1	-0.1	-0.2	-0.4	-0.5	0.4	1	0.3	-0.1	-0.2	-0.1	-0.3	-0.1	0.1	0.1	0.2	-0.1	0	0.1	0	0
E5	0	0.3	0.3	0.2	0.3	0.3	0.2	0.2	-0.2	-0.2	-0.3	-0.4	0.4	0.3	1	0	0	-0.1	-0.2	-0.1	0.3	-0.1	0.3	0	-0.1	0.1	0.1	0.1
N1	0.2	-0.1	-0.1	-0.1	-0.2	-0.1	0	-0.1	0.2	0.2	0	0.2	0	-0.1	0	1	0.7	0.6	0.4	0.4	-0.1	0.1	0	0.1	0.1	0	0	-0.1
N2	0.1	0	-0.1	-0.1	-0.2	0	0	-0.1	0.2	0.2	0	0.2	-0.1	-0.2	0	0.7	1	0.5	0.4	0.4	0	0.1	0	0.1	0	0.1	0	-0.1
N3	0.1	0	0	-0.1	-0.1	0	0	-0.1	0.2	0.2	0	0.2	0	-0.1	-0.1	0.6	0.5	1	0.5	0.4	0	0.1	0	0.2	0.1	0.1	0	-0.1
N4	0	-0.1	-0.1	-0.2	-0.2	-0.1	0	-0.1	0.3	0.4	0.2	0.4	-0.2	-0.3	-0.2	0.4	0.4	0.5	1	0.4	-0.1	0.1	-0.1	0.2	0	0	0	0
N5	0	0	0	0	-0.1	0	0	0	0.2	0.2	0	0.3	-0.1	-0.1	-0.1	0.4	0.4	0.4	0.4	1	-0.1	0.2	-0.1	0.1	0.1	0.2	0	-0.1
O1	0	0.1	0.1	0	0.1	0.2	0.2	0.1	-0.1	-0.1	-0.1	-0.2	0.3	0.1	0.3	-0.1	0	0	-0.1	-0.1	1	-0.2	0.4	0.2	-0.3	-0.1	0	0
O2	0.1	0	0	0.1	0	-0.1	-0.1	0	0.2	0.1	0.1	0.1	-0.1	0.1	-0.1	0.1	0.1	0.1	0.1	0.2	-0.2	1	-0.3	-0.1	0.3	0	-0.1	0
O3	-0.1	0.1	0.2	0	0.2	0.2	0.2	0.1	-0.1	-0.1	-0.2	-0.2	0.4	0.2	0.3	0	0	0	-0.1	-0.1	0.4	-0.3	1	0.2	-0.3	0	0.1	0
O4	-0.1	0.1	0	-0.1	0	0.1	0	0	0.1	0.1	0.1	0.2	0	-0.1	0	0.1	0.1	0.2	0.2	0.1	0.2	-0.1	0.2	1	-0.2	0	0.1	0
O5	0.1	-0.1	0	0	0	-0.1	-0.1	0	0.2	0	0.1	0.1	-0.1	0	-0.1	0.1	0	0.1	0	0.1	-0.3	0.3	-0.3	-0.2	1	0	-0.1	-0.1
gender	-0.2	0.2	0.2	0.1	0.1	0	0.1	0	-0.1	-0.1	-0.1	-0.1	0.1	0.1	0.1	0	0.1	0.1	0	0.2	-0.1	0	0	0	0	1	0	0
education	-0.1	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0.1	0	0	0	0	0	0	-0.1	0.1	0.1	-0.1	0	1	0.2
age	-0.1	0.1	0	0.1	0.1	0.1	0	0.1	-0.1	-0.1	0	-0.1	0	0	0.1	-0.1	-0.1	-0.1	0	-0.1	0	0	0	0	-0.1	0	0.2	1

Figure 1: Correlation Plot of BFI Dataset

Firstly, from the correlation plot of all variables in BFI Dataset presented in Figure 1, we can clearly see that demographic related variables that are gender, education, and age have no significant correlation with any of the personality trait variables. The maximum correlation we got for these three variables is 0.2, under the threshold we have set, which is 0.5. Secondly, we observe some strong correlations between the variables that are related to the same personality trait. For example, if we look at variables N1 through N5, the correlations between the N variables are generally on the higher side (0.4 and above). Also, the correlation within the N subgroup of variables is all positive. In contrast, within some of the other groups, we observe some strong negative correlations, such as subgroup E. Subgroup of variables related to Openness (O) seem to have the least correlation with each other compared to other personality trait variable subgroups. We will not include the demographic variables because our aim is to understand if we can and should band together with the same personality trait variables. The demographic variables at this stage do not play a crucial role in that.

(Question #1)
To attain 90% of the overall variability, we will be looking to keep all factors with an eigenvalue above 1.0. The correlation matrix's computed eigenvalues show that the first seven factors account for 90% of the BFI dataset variability. Also presented in Table 2, we observe that eigenvalue drops below 1.0 as it gets to factor 8.

Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7
5.0685	2.7625	2.1526	1.8923	1.5175	1.0788	0.8309

Table 2: PC Eigen Values of BFI Dataset Correlation

The Scree Plots with Parallel analysis graph in Figure 2 demonstrates that while we have four factors that have eigenvalues above 1.0, there are six factors that we should be taken into account that fall above the FA Simulated Data line.

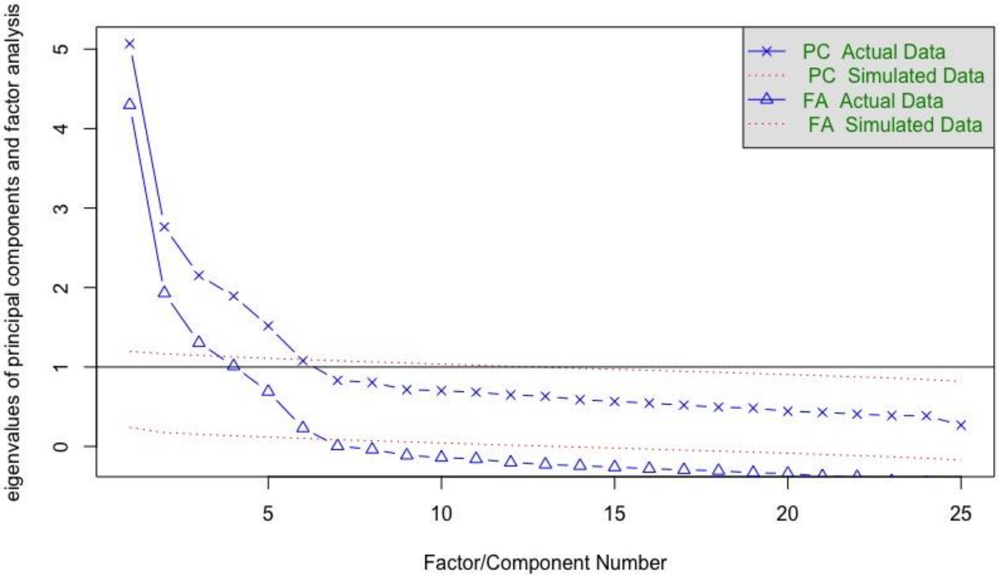


Figure 2: Scree Plots with Parallel Analysis

(Question #2)
Upon our computation of eigenvalues and scree plot, we have determined that our FA model moving forward should have six factors to attain at least 90% of the overall variability in the dataset. We will initially choose 'varimax' as our rotation and model type as 'maximum likelihood' for our factor analysis model.

	ML1	ML2	ML3	ML5	ML4	ML6
A1	0.03	0.1	0.05	-0.53	-0.11	0.12
A2	0.26	0.04	0.13	0.65	0.04	-0.01
A3	0.38	0.01	0.13	0.57	0.03	0.15
A4	0.24	-0.07	0.24	0.39	-0.15	0.1
A5	0.45	-0.14	0.11	0.43	0.02	0.23
C1	0.08	0	0.55	-0.01	0.19	0.08
C2	0.04	0.07	0.67	0.05	0.09	0.16
C3	0.04	-0.04	0.55	0.1	-0.01	0
C4	-0.06	0.22	-0.63	-0.1	-0.12	0.31
C5	-0.18	0.27	-0.55	-0.04	0.03	0.14

E1	-0.57	0.03	0.06	-0.13	-0.07	0.18
E2	-0.67	0.23	-0.09	-0.09	-0.05	0.12
E3	0.59	0	0.11	0.14	0.25	0.23
E4	0.68	-0.14	0.11	0.21	-0.11	0.14
E5	0.51	0.05	0.31	0.07	0.2	-0.06
N1	0.05	0.82	-0.05	-0.16	-0.07	-0.12
N2	0.01	0.8	-0.04	-0.12	0	-0.19
N3	-0.07	0.71	-0.05	-0.01	-0.01	0.08
N4	-0.34	0.56	-0.16	0	0.07	0.2
N5	-0.15	0.52	-0.04	0.09	-0.17	0.16
O1	0.23	-0.02	0.13	-0.01	0.49	0.16
O2	0.01	0.17	-0.09	0.05	-0.5	0.14
O3	0.34	0.02	0.08	0.04	0.58	0.18
O4	-0.16	0.21	-0.03	0.12	0.35	0.17
O5	-0.01	0.06	-0.05	-0.08	-0.58	0.15

Table 3: Factor Analysis Loading Table with Varimax Rotation

The Factor Analysis loadings presented in Table 3 suggest that N subgroup personality trait variables are the highest factor loadings overall and specifically in ML2. The loading of each N subgroup variable is above 0.5 in ML2, showing us that these variables could be the same output indicator, yet none of the other personality trait subgroups that we expected to observe high loadings present collective loadings over 0.5 under any Factor. In addition, for any variables aside from the N subgroup, the model has detected negative loadings. As the negative loadings are to be subtracted from positive loadings in the given factor, the other subgroup loadings presented in Table 3 do not suggest considerable insight.

	ML1	**ML2**	**ML3**	**ML5**	**ML4**	**ML6**
SS Loadings	2.73	2.72	2.05	1.56	1.54	0.62

Table 4: SS Eigen Value of Factors of the FA with Varimax Rotation

Per the results presented in Table 4, we can determine that all factors got eigenvalue above 1.0 except for ML6. This is an expected result as ML 6 has no loadings above 0.31. We also observe that the eigenvalues of ML1 and ML2 are almost the same as they have somewhat similar total loadings if we look at Table 3. In the case of our FA for the BFI dataset, our null hypothesis was that the variables of each personality trait to have high loading under each separate factor, which would indicate that these variables could potentially be banded together for dimensional reduction. Yet, through our first FA model presented above, we are rejecting this null hypothesis.

(Question #3)

Following the FA model where we used 'varimax' rotation, we use the 'promax' rotation to assess and observe improvement in our overall FA results. The results of this analysis are presented in Table 5.

	ML1	**ML2**	**ML3**	**ML4**	**ML5**	**ML6**
A1	0.1	0.05	0.11	0.14	-0.61	0.12
A2	0.18	0.11	0.05	-0.03	0.63	-0.01
A3	0.3	0.02	0.03	0.02	0.47	0.17
A4	0.15	-0.04	0.2	0.19	0.31	0.07
A5	0.37	-0.16	-0.01	0.03	0.29	0.26
C1	-0.04	0.03	0.59	-0.13	-0.09	0.01
C2	-0.11	0.07	0.74	-0.02	-0.06	0.06
C3	-0.08	0.02	0.6	0.05	0.05	-0.1
C4	0.05	0.03	-0.66	0.11	-0.14	0.45

C5	-0.08	0.14	-0.55	-0.06	0	0.26
E1	-0.67	-0.14	0.2	0.06	-0.11	0.15
E2	-0.73	0.07	0.07	0.02	-0.01	0.12
E3	0.59	0	-0.02	-0.18	-0.03	0.29
E4	0.68	-0.08	-0.03	0.17	0.05	0.16
E5	0.53	0.18	0.22	-0.15	0	-0.08
N1	0.22	0.91	0.01	0.07	-0.1	-0.11
N2	0.17	0.91	0.03	-0.01	-0.03	-0.18
N3	0.01	0.69	0.02	0.02	0.01	0.11
N4	-0.32	0.42	-0.06	-0.07	0.03	0.25
N5	-0.13	0.44	0.04	0.18	0.08	0.18
O1	0.2	-0.06	0.07	-0.45	-0.1	0.21
O2	0.03	0.12	-0.06	0.51	0.01	0.13
O3	0.33	-0.02	-0.02	-0.54	-0.06	0.26
O4	-0.21	0.11	-0.01	-0.34	0.12	0.22
O5	0.01	0.01	0	0.6	-0.14	0.12

Table 5: Factor Analysis Loading Table with Promax Rotation

Our FA model's 'promax' rotation gave us more insightful interpretability than our model, where we used 'varimax' rotation. Firstly, we see that the N subgroup variables are now more weighted in three loadings than the five we had previously. Secondly, overall, we see an increase in the negative loadings, especially in ML1, where E1's and E2's loadings have changed from -0.57 and -0.67 to -0.67 and -0.73, respectively. In ML3, we observe an increase in positive loadings and in negative loadings as well. While ML4 loadings are highest around O subgroup variables, the negative and positive loadings somewhat cancel each other out, not giving any solid results. Finally, in ML5, we observe an increase in negative loading, yet a decrease in positive loading; therefore, we can say that ML5 also does not give us any valuable results.

	ML1	**ML2**	**ML3**	**ML4**	**ML5**	**ML6**
SS Loadings	2.82	2.62	2.02	1.49	1.43	0.84

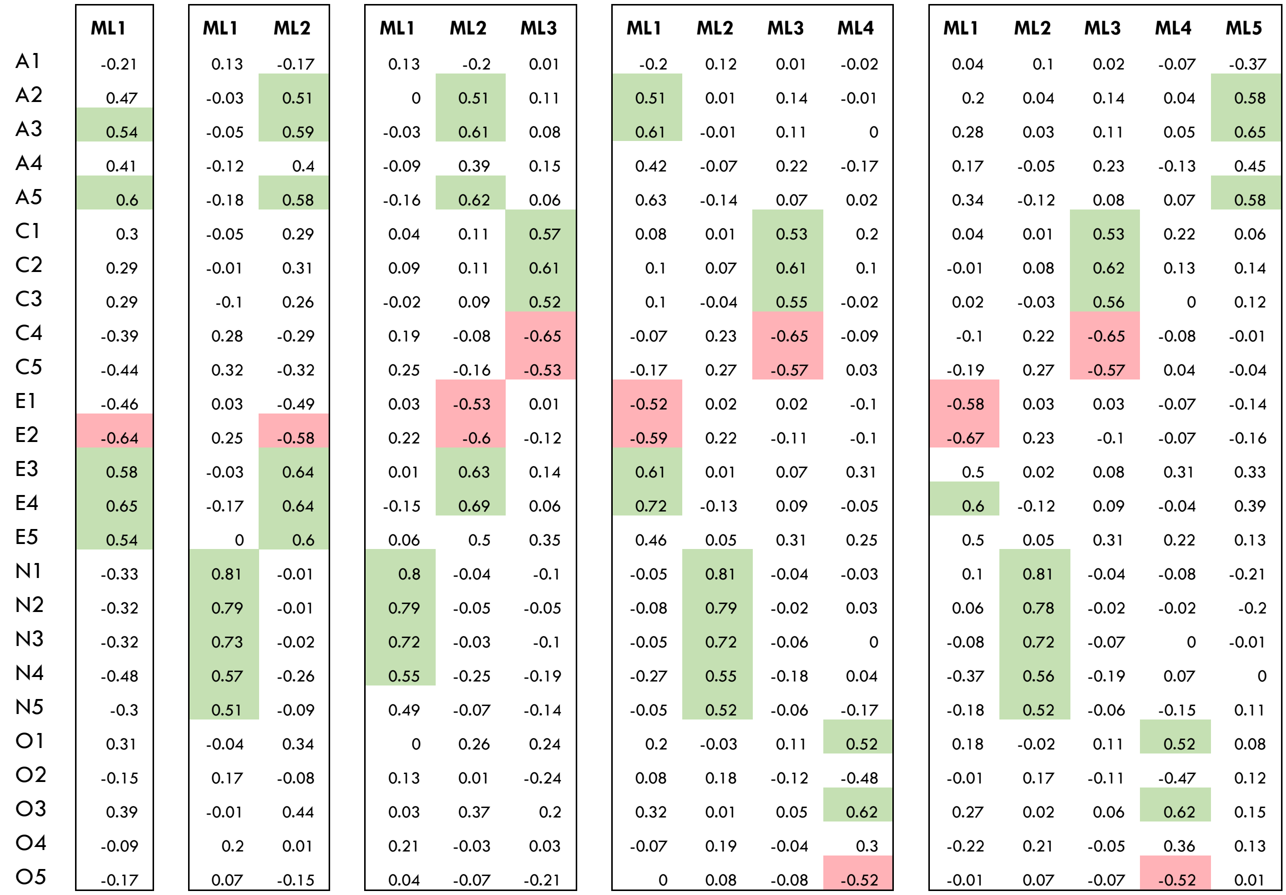
Table 6: SS Eigen Value of Factors of the FA with Promax Rotation

Each factor's eigenvalues had slightly changed when we modified the rotation from 'varimax' to 'promax'. Also seen in Table 6, the eigenvalues of ML1 and ML2, which used to be almost identical, are now moderately different from each other. Similar to the 'varimax' rotated model, our 'promax' rotated FA model suggests that we do not need six factors. Still, five factors would be sufficient to uncover any potential underlying relationships between BFI dataset variables.

(Question #4)

Per our results presented and interpreted above, we use five factors instead of 6, like the earlier analyses for our incremental factor analysis. To observe the changes that are the cause of the number of factors used in the model, we incrementally increase the number of factors from min 1 to max 5. The cut-off values of 0.5 and -0.5 were used to assess these various iterations so that we can compare the interpretation of the previous results and outcomes we obtained.

The loadings for each FA model that was run with nfactor 1 through 5 presented in Table 7 show that with one-factor model, we only observe positive and negative loadings over our cut-off value of 0.5 for E some for A subgroup variables. As we move along with our incrementally increased models, we see that loadings are being added to other variables in various personality trait subgroups. The loadings, in general, stabilize as we move from 1 to 5-factor model. In addition to stabilizing, we see overall an increase in some loadings under various variable subgroups, as the loading of the certain subgroup accumulates more within one loading than the others. While we see from Table 7 that the 5-factor model is the correct number of variables because it provides the loadings for at least one of each personality trait variable subgroup that we aim to study, the fa model with three factors seems to be the one that is easiest to interpret. The 3-factor model covers 90% of the information obtained from the model with five factors, giving us all the essential loadings that may be viable.



	ML1	ML1	ML2	ML1	ML2	ML3	ML1	ML2	ML3	ML4	ML1	ML2	ML3	ML4	ML5
A1	-0.21	0.13	-0.17	0.13	-0.2	0.01	-0.2	0.12	0.01	-0.02	0.04	0.1	0.02	-0.07	-0.37
A2	0.47	-0.03	0.51	0	0.51	0.11	0.51	0.01	0.14	-0.01	0.2	0.04	0.14	0.04	0.58
A3	0.54	-0.05	0.59	-0.03	0.61	0.08	0.61	-0.01	0.11	0	0.28	0.03	0.11	0.05	0.65
A4	0.41	-0.12	0.4	-0.09	0.39	0.15	0.42	-0.07	0.22	-0.17	0.17	-0.05	0.23	-0.13	0.45
A5	0.6	-0.18	0.58	-0.16	0.62	0.06	0.63	-0.14	0.07	0.02	0.34	-0.12	0.08	0.07	0.58
C1	0.3	-0.05	0.29	0.04	0.11	0.57	0.08	0.01	0.53	0.2	0.04	0.01	0.53	0.22	0.06
C2	0.29	-0.01	0.31	0.09	0.11	0.61	0.1	0.07	0.61	0.1	-0.01	0.08	0.62	0.13	0.14
C3	0.29	-0.1	0.26	-0.02	0.09	0.52	0.1	-0.04	0.55	-0.02	0.02	-0.03	0.56	0	0.12
C4	-0.39	0.28	-0.29	0.19	-0.08	-0.65	-0.07	0.23	-0.65	-0.09	-0.1	0.22	-0.65	-0.08	-0.01
C5	-0.44	0.32	-0.32	0.25	-0.16	-0.53	-0.17	0.27	-0.57	0.03	-0.19	0.27	-0.57	0.04	-0.04
E1	-0.46	0.03	-0.49	0.03	-0.53	0.01	-0.52	0.02	0.02	-0.1	-0.58	0.03	0.03	-0.07	-0.14
E2	-0.64	0.25	-0.58	0.22	-0.6	-0.12	-0.59	0.22	-0.11	-0.1	-0.67	0.23	-0.1	-0.07	-0.16
E3	0.58	-0.03	0.64	0.01	0.63	0.14	0.61	0.01	0.07	0.31	0.5	0.02	0.08	0.31	0.33
E4	0.65	-0.17	0.64	-0.15	0.69	0.06	0.72	-0.13	0.09	-0.05	0.6	-0.12	0.09	-0.04	0.39
E5	0.54	0	0.6	0.06	0.5	0.35	0.46	0.05	0.31	0.25	0.5	0.05	0.31	0.22	0.13
N1	-0.33	0.81	-0.01	0.8	-0.04	-0.1	-0.05	0.81	-0.04	-0.03	0.1	0.81	-0.04	-0.08	-0.21
N2	-0.32	0.79	-0.01	0.79	-0.05	-0.05	-0.08	0.79	-0.02	0.03	0.06	0.78	-0.02	-0.02	-0.2
N3	-0.32	0.73	-0.02	0.72	-0.03	-0.1	-0.05	0.72	-0.06	0	-0.08	0.72	-0.07	0	-0.01
N4	-0.48	0.57	-0.26	0.55	-0.25	-0.19	-0.27	0.55	-0.18	0.04	-0.37	0.56	-0.19	0.07	0
N5	-0.3	0.51	-0.09	0.49	-0.07	-0.14	-0.05	0.52	-0.06	-0.17	-0.18	0.52	-0.06	-0.15	0.11
O1	0.31	-0.04	0.34	0	0.26	0.24	0.2	-0.03	0.11	0.52	0.18	-0.02	0.11	0.52	0.08
O2	-0.15	0.17	-0.08	0.13	0.01	-0.24	0.08	0.18	-0.12	-0.48	-0.01	0.17	-0.11	-0.47	0.12
O3	0.39	-0.01	0.44	0.03	0.37	0.2	0.32	0.01	0.05	0.62	0.27	0.02	0.06	0.62	0.15
O4	-0.09	0.2	0.01	0.21	-0.03	0.03	-0.07	0.19	-0.04	0.3	-0.22	0.21	-0.05	0.36	0.13
O5	-0.17	0.07	-0.15	0.04	-0.07	-0.21	0	0.08	-0.08	-0.52	-0.01	0.07	-0.07	-0.52	0.01

Table 7: Loadings for FA Models Fitted with Factors 1 through 5

(Question #5)

While we do not see a distinct indicator of a possible grouping of variables of different personality trait subgroups, through the results presented in Table 7, the model we found easiest to interpret is the model with nfactor = 3. We can suggest that the research could further look into banding together with the A (Agreeableness) and E (Extraversion) personality trait variables for their study.

(Question #6)

As the final step of our Exploratory Factor Analysis, we have assigned variables to our EFA model factor score to add them to the BFI dataset and assess potential relationships between the factors and our demographic variables. Per our plot, also seen in Figure x, there are no correlations that are above our 0.5 thresholds, between ML1 through ML6 and demographic variables. The highest correlation we observe is between ML5 and 'gender' wit 0.3. In the previous sections, we determined that ML5 has the most loadings of A subgroup variables related to a person's Agreeableness. While 0.3 correlation does not indicate a strong relationship, this assessment and our correlation plot inform us to keep our eye out in the following steps of the model/analysis for a potential correlation between Agreeableness and gender, and they are to be further studied.

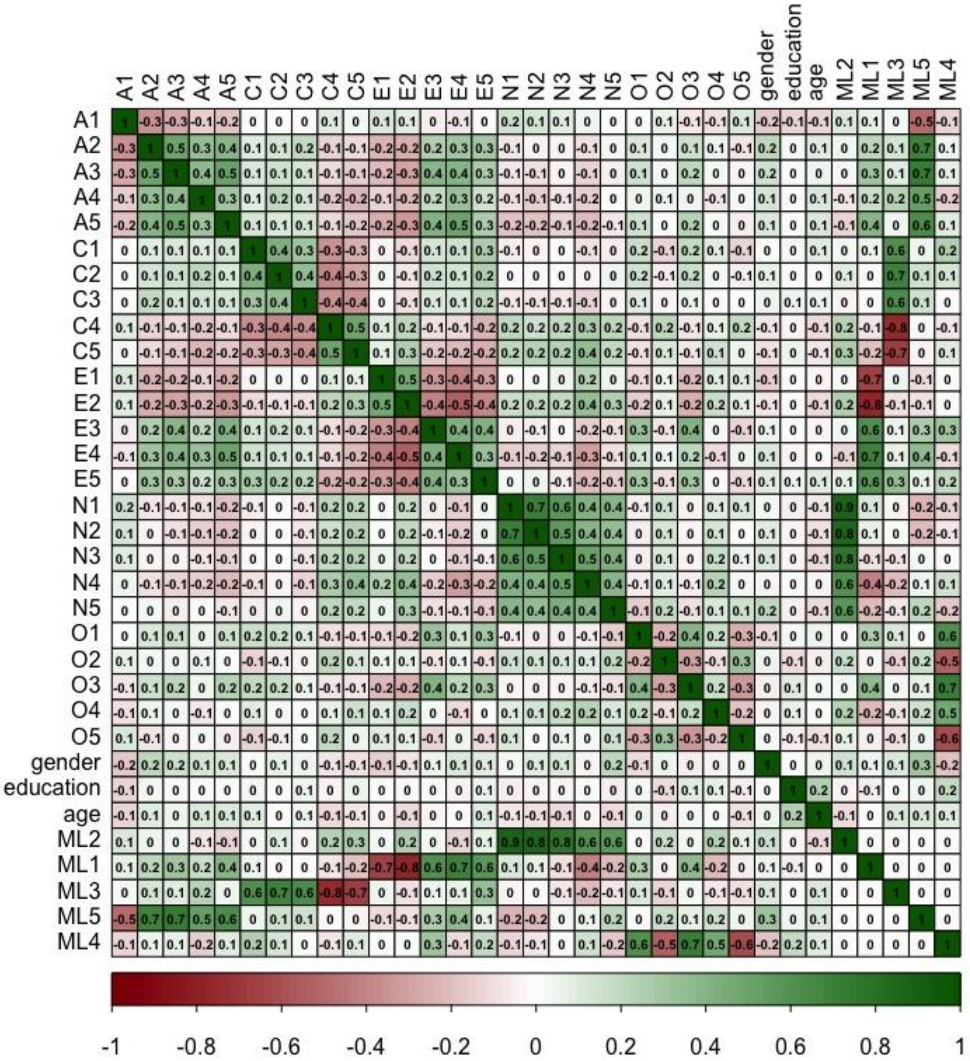


Figure 3: Correlation Plot of BFI Dataset including FA scores obtained through FA Model

Conclusion

In conclusion, throughout the Exploratory Factor Analysis we have conducted on the BFI dataset, we can state that (a) Neuroticism (N) personality trait subgroups variables are highly correlated and are high in positive loading in all models we conducted, (b) there is a potential opportunity to band together with the Agreeableness (A) and Extraversion (E) personality trait variables, and (c) Agreeableness (A) personality trait variables show slight but noticeable correlation with 'gender', one of our demographic variables. We obtained these potential underlying relationships and characteristics throughout Exploratory Factor Analysis to create an informed starting point for the more detailed and supervised learning models that may follow.

Reflection

After completing this week's assignment, I am confident that I will frequently be using Exploratory Factor Analysis in the future, both for professional and academic settings. While the BFI dataset was interesting to study, I personally look forward to conducting the FA model for various datasets such as sales and HR. Another path I look forward to exploring following this week's exercise is running an FA model for a dataset with mixed variable types. As we complete week 3, from the models I have practiced in this class so far, I would prefer EFA as it provides dimensional reduction and is the easiest to interpret and understand the underlying correlations/relationships. Compared to PCA and t-SNE, EFA provides a more thorough understating through loadings at a more granular level.

The R Code for Assignment #1

```
install.packages("psych")
install.packages("corrplot")
install.packages("GPArotation")
library(GPArotation)
library(psych)
library(corrplot)
library(RColorBrewer)
library(readxl)
library(corrplot)

bfi_data <- bfi
bfi_data

str(bfi_data)
dim(bfi_data)
head(bfi_data)
summary(bfi_data)
```

```
#Remove rows with missing values and keep only complete cases
sapply(bfi_data, function(x) sum(is.na(x)))
bfi_data  <- bfi_data[complete.cases(bfi_data),]

dim(bfi_data)
2800 - 2236
#564 observations were removed due to having missing values
# we have observation count over variable count * 20 so it is enough data to work with

bfi_cor <- cor(bfi_data)
bfi_cor
corrplot(bfi_cor, method = "color", outline = T, addrect = 4, rect.col = "black", rect.lwd = 3,cl.pos = "b", tl.col =
"black", tl.cex = 1, cl.cex = 1, addCoef.col = "black", number.digits = 1, number.cex = 0.60, col =
colorRampPalette(c("darkred","white","darkgreen"))(100))

is.matrix(bfi_cor)
isSymmetric(bfi_cor)

###
bfi_small <- bfi_data[,c(1:25)]
bfi_small_cor <- cor(bfi_small)
bfi_small_cor
colnames(bfi_small_cor)

Z <- eigen(bfi_small_cor)
Z$values


#scree plot
fa.parallel(bfi_small,n.obs = 2236, fa = "both", n.iter = 100, show.legend = TRUE, main = 'Scree Plots with Parallel
Analysis')

#Question 2
fa_varimax <- fa(bfi_small, nfactors = 6, rotate = "varimax", fm = "ml")
fa_varimax

#Question 3
fa_promax <- fa(bfi_small, nfactors = 6, rotate = "promax", fm = "ml")
fa_promax

#Question 4
fa_varimax_1 <- fa(bfi_small, nfactors = 1, rotate = "varimax", fm = "ml")
fa_varimax_1

fa_varimax_2 <- fa(bfi_small, nfactors = 2, rotate = "varimax", fm = "ml")
fa_varimax_2

fa_varimax_3 <- fa(bfi_small, nfactors = 3, rotate = "varimax", fm = "ml")
fa_varimax_3

fa_varimax_4 <- fa(bfi_small, nfactors = 4, rotate = "varimax", fm = "ml")
fa_varimax_4

fa_varimax_5 <- fa(bfi_small, nfactors = 5, rotate = "varimax", fm = "ml")
fa_varimax_5

#Question 6
fa_varimax_all <- fa(bfi_data, nfactors = 5, rotate = "varimax", fm = "ml")
fs <- factor.scores(bfi_data, fa_varimax_all)
fs <- fs$scores
bfi_data_comb <- cbind(bfi_data,fs)

bfi_all_cor <- cor(bfi_data_comb)
corrplot(bfi_all_cor, method = "color", outline = T, addrect = 4, rect.col = "black", rect.lwd = 3,cl.pos = "b", tl.col =
"black", tl.cex = 1, cl.cex = 1, addCoef.col = "black", number.digits = 1, number.cex = 0.60, col =
colorRampPalette(c("darkred","white","darkgreen"))(100))
```