Northeastern University

ASSIGNMENT FRONT SHEET

Course Name: **ALY6040 Data Mining Applications**

Professor Name: **Nagadeepa Shanmuganathan**

Student Name: **Dong Quoc Tuong (Lukas)**

Student Class: **Fall 2019 CPS** **Term: Winter 2021**

Module 5: Factor Analysis with Personality test

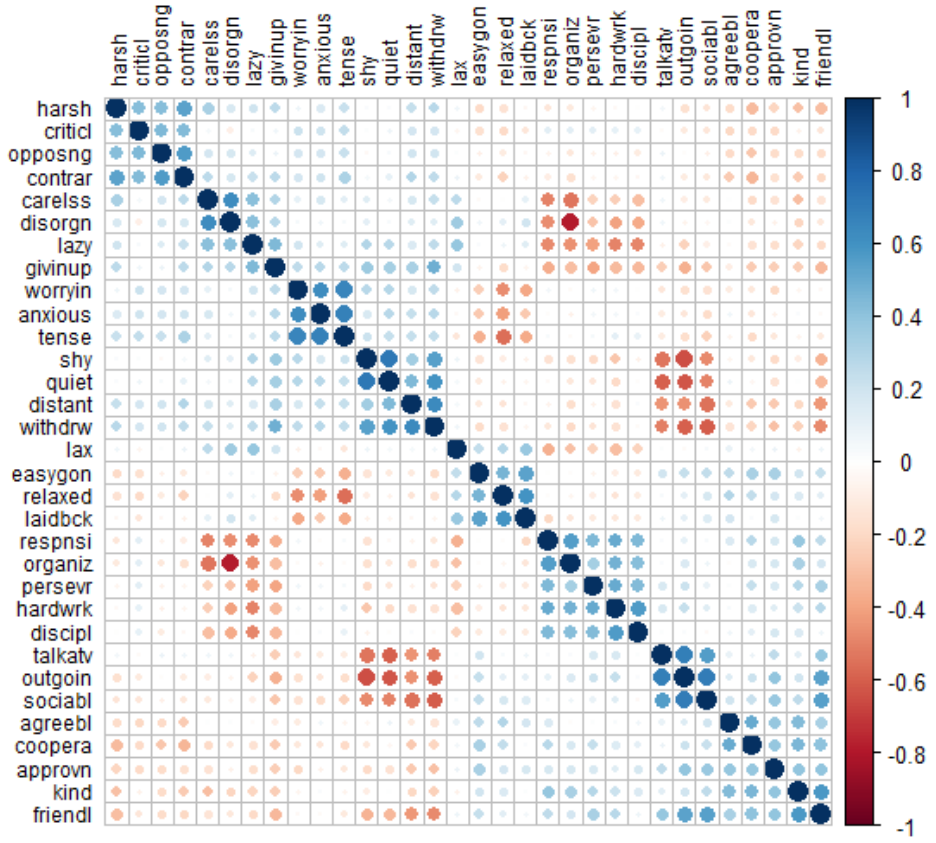
Completion Date: Feb 28t **Due Time:12:00am**

Statement of Authorship

I confirm that this work is my own. Additionally, I confirm that no part of this coursework, except where clearly quoted and referenced, has been copied from material belonging to any other person e.g. from a book, handout, another student. I am aware that it is a breach of Northeastern University's regulations to copy the work of another without clear acknowledgement and that attempting to do so renders me liable to disciplinary procedures. To this effect, I have uploaded my work onto Turnitin and have ensured that I have made any relevant corrections to my work prior to submission.

☒ **Tick here** to confirm that your paper version is identical to the version submitted through Turnitin

The dataset comprised self-ratings of 240 participants on 32 different personality traits on the scale from 1 to 10. Thus, it is a perfect exercise for factor analysis, which is the method of analyzing the covariation among the observed variables, as an outgoing individual is more often talkative compared to a reserved one.



After loading the data and necessary library, we plot correlation map. The corrplot shows the correlations between different variables in the dataset. Red indicates negative while blue illustrates positive correlation. The size and shade of the dot indicates how strong the correlation is. For example, “friendl” is insignificantly correlated with “criticl” in the downward trend and thus it the dot is nearly invisible.

Factor Analysis with no rotation

Factor analysis estimates a model which explains variance/ co variance between a set of observed variable by a set of fewer unobserved factors and weightings. (*Factor Analysis*, 2020) Thus, we compute variance (how much a random variables differs from its expected value) because they can allow us to find the unobserved factors that explain the variance. We used "factanal()" to look at the sum of squared (SS) loadings, these are the eigenvalues, or the variance in all variables which is accounted for by that factor, the higher SS loading, the more likely it will help us to explain the variances in the variables. Conventional wisdom is SS loadings must be greater than 1 (Kaiser Rule) to be considered and as we can see Factor 1-6 qualify for such rule

```
                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
SS loadings       6.706   3.071   2.554   2.283   1.979   1.105
Proportion Var    0.210   0.096   0.080   0.071   0.062   0.035
Cumulative Var    0.210   0.306   0.385   0.457   0.519   0.553
                Factor7 Factor8 Factor9 Factor10
SS loadings       0.867   0.689   0.393    0.382
Proportion Var    0.027   0.022   0.012    0.012
Cumulative Var    0.580   0.602   0.614    0.626
> |
```

For example, when we compute the eigenvalue of the first factor we have 6.71 and then for the proportion variances is estimated to be 0.21

```
> eigenv_fac1 = sum(loadings_fac1^2); eigenv_fac1
[1] 6.706499
> # Compute proportion variance
> eigenv_fac1/32
[1] 0.2095781
```

We are looking for uniqueness because Uniqueness is the variance that is 'unique' to the variable and not shared with other variables. (Hartmann et al., 2018) It is equal to 1 –communality

(variance that is shared with other variables). There are two ways to calculate Uniqueness; the first one is to use “uniqueness” function why the second method requires us to calculate the communality distant, then get the final result by 1 subtracting the communality distant. Both methods result in the “distant” variable’s uniqueness of 0.415 as seen below

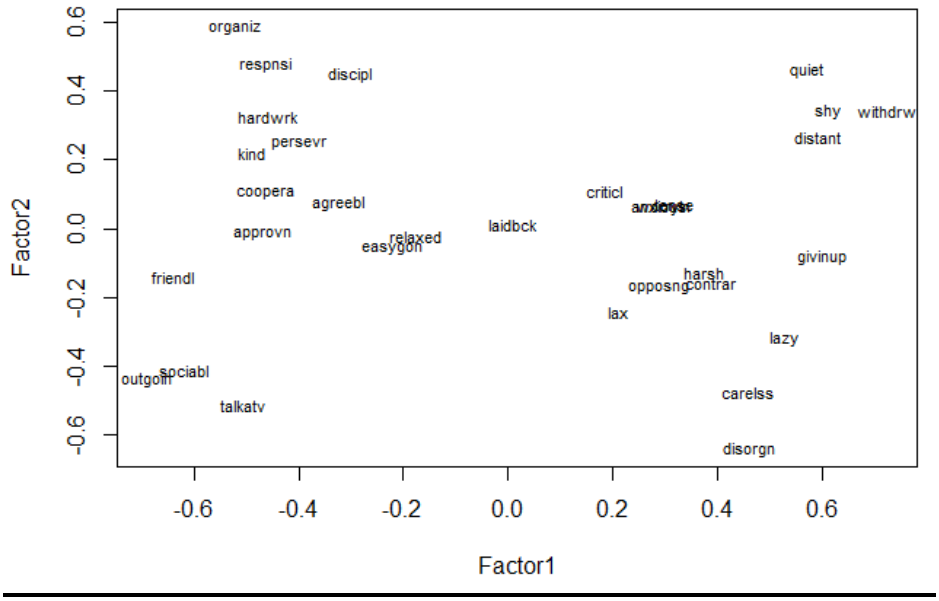
```
> res1b$uniquenesses
  distant   talkatv   carelss   hardwrk   anxious   agreebl
0.4155265 0.3662366 0.4332678 0.3765560 0.3414400 0.1085821
    tense      kind   opposng   relaxed   disorgn   outgoin
0.2479509 0.4473867 0.4135546 0.4137453 0.1676505 0.2036158
  approvn      shy   discipl     harsh   persevr   friendl
0.6424638 0.3227601 0.4613517 0.3813236 0.4950613 0.3363327
  worryin   respnsi   contrar   sociabl      lazy   coopera
0.3481255 0.3921460 0.4166332 0.3574583 0.4100754 0.5047471
    quiet   organiz   criticl       lax   laidbck   withdrw
0.2855130 0.2315167 0.5504645 0.6611637 0.0050000 0.2170508
  givinup   easygon
0.4644383 0.5503472
>
```

Method 1

```
> loadings_distant = res1b$loadings[1,]
> communality_distant = sum(loadings_distant^2); communality_distant
[1] 0.5844747
> uniqueness_distant = 1-communality_distant; uniqueness_distant
[1] 0.4155253
>
```

Method 2

Next we visualize the variables in the 2D space like for Factor 1 against Factor 2 below. Since we did not rotate the variables there could be overlaying like below where Factor 1 could be considered talkative & organized vs. quiet & disorganized, and Factor 2 as talkative & disorganized vs. organized and quiet. Hence, it is much better to rotate the factors first before we do any visualization



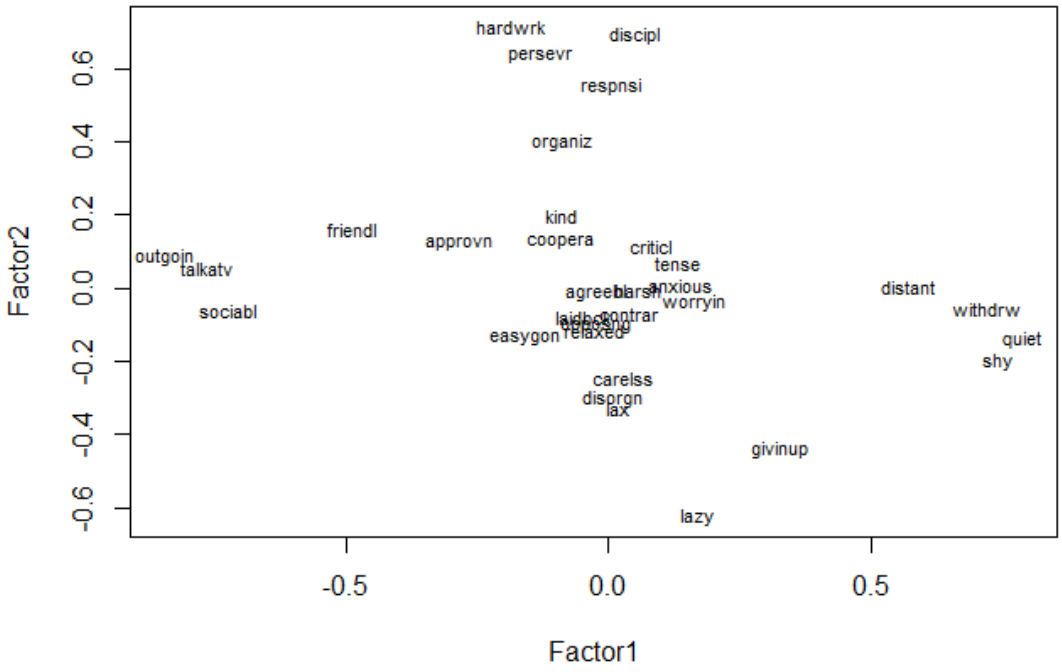
Factor Analysis with rotation

After the rotation, we can see the SS loadings become more evenly distributed (lower highest and higher lowest) compared to no rotation and even extended to include an additional Factor (7).

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
SS loadings	4.476	2.954	2.559	2.522	2.268	2.230
Proportion Var	0.140	0.092	0.080	0.079	0.071	0.070
Cumulative Var	0.140	0.232	0.312	0.391	0.462	0.532
	Factor7	Factor8	Factor9	Factor10		
SS loadings	1.591	0.531	0.485	0.415		
Proportion Var	0.050	0.017	0.015	0.013		
Cumulative Var	0.581	0.598	0.613	0.626		

>

After rotation, we can see that “sociabl” and “shy” load heavily on Factor 1 but have bery low loadings on Factors 2 We can define or label Factor 1 with terms like extraverted vs. introverted and Factor 2 as “conscientiousness”vs “hardwrk” vs. lazy).



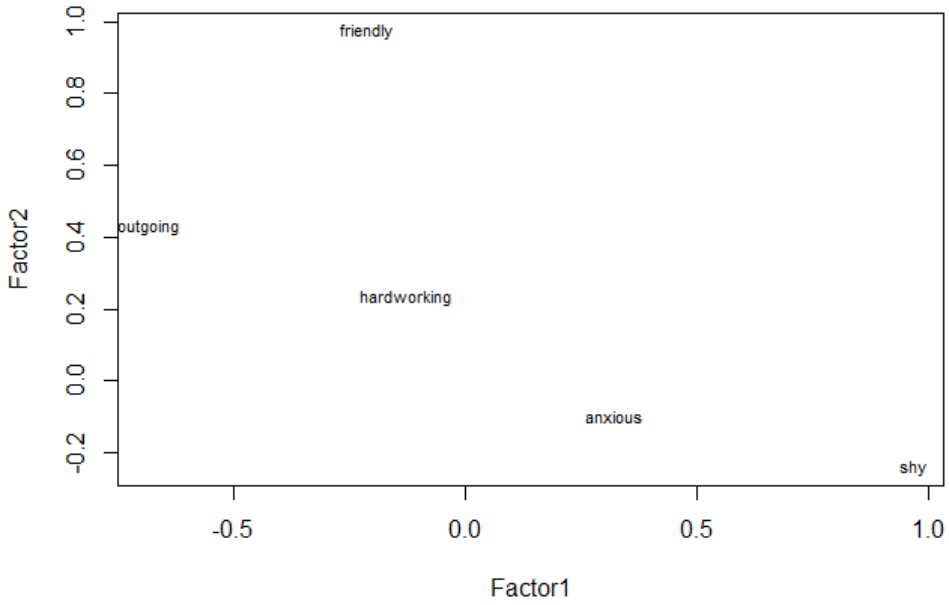
Creating composite variables

Now we look at the spurious factors that are introduced. We are going to reduce the dataset down to p=5 variables, so we cannot pull out k-10 as we did above, so k =2 in this case. After combining the different qualities down to 5 like shy is the main representation of distant, shy, withdrawn, quite, we have the Factor 1 and 2 SS loadings greater than 1 as below

```
Loadings:
            Factor1 Factor2
shy          0.968  -0.241
outgoing    -0.683   0.432
hardworking -0.126   0.233
friendly    -0.214   0.974
anxious      0.321

               Factor1 Factor2
SS loadings      1.568   1.258
Proportion Var   0.314   0.252
Cumulative Var   0.314   0.565
```

The visualization is also easier to see as well given to the fact for this one we only have solely 2 Factors



References

Factor Analysis. (2020, April 15). Statistics Solutions. https://www.statisticssolutions.com/factor-analysis-sem-factor-analysis/

Hartmann, K., Krois, J., Waske, B. (2018): E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin.