Multivariate Statistics

Group #27

Assignment 1: PCA & EFA

Adam Binst r0252193
Natalia Eremeeva r0648321
Rana Cansu Kebabcı r0772826
Ömer Yiğit r0767950

Prof. Vandebroek Martina 2019-2020

PRINCIPAL COMPONENT ANALYSIS (PCA):

1) Problem Statement:

The dataset "*politics*" contains data obtained from participants from 28 countries and 9 variables about political attitudes. The variables are country code, trust in the country's parliament, trust in the legal system, trust in politicians, trust in political parties, trust in the European Parliament, trust in the united nations, immigration bad or good for country's economy, the country's cultural life undermined or enriched by immigrants, immigrants make country worse or better place to live.

In this analysis we are looking at a relative large number of variables and we strongly expect there will be redundancy among them. In this case, redundancy means that some of the variables are correlated with each other, because we suspect they are measuring the same construct. Because of this redundancy, we assume it should be possible to reduce the observed variables into a smaller number of principal components that will account for most of the variance in the observed variables. Intuitively, the variables may be categorized into two groups, related to "political trust" and "immigration". In order to reach more precise conclusions, principal component analysis (PCA) should be applied to the data and the results should be interpreted.

2) Descriptive Statistics:

Before conducting PCA, let us look at descriptive statistics, to have a better understanding whether PCA is applicable in our case and if our assumptions are right.

The first assumption we should check is whether observed variables have high correlations with each other. You can find correlation matrix below. Mainly, the variables related to trust seem to have high correlation between each other and have low correlation between the variables related to immigration. It means that the dataset can be projected into a lower dimensional space while retaining the most of the variance explained.

	trstprl	trstlgl	trstplt	trstprt	trstep	trstun	imbgeco	imueclt	imwbcnt
trstprl	1.000	0.9039	0.9249	0.9198	0.3816	0.5122	0.4495	0.4159	0.3859
trstlgl	0.9039	1.000	0.8509	0.8482	0.2901	0.4967	0.3801	0.3836	0.3438
trstplt	0.9249	0.8509	1.000	0.9946	0.5293	0.6793	0.5593	0.5070	0.4878
trstprt	0.9198	0.8482	0.9946	1.000	0.5241	0.7025	0.5713	0.5340	0.5229
trstep	0.3816	0.2901	0.5293	0.5241	1.000	0.6437	0.5093	0.4224	0.4530
trstun	0.5122	0.4967	0.6793	0.7025	0.6437	1.000	0.7292	0.7811	0.7071
imbgeco	0.4495	0.3801	0.5593	0.5713	0.5093	0.7292	1.000	0.8525	0.9052
imueclt	0.4159	0.3836	0.5070	0.5340	0.4224	0.7811	0.8525	1.000	0.8831
imwbcnt	0.3859	0.3438	0.4878	0.5229	0.4530	0.7071	0.9052	0.8831	1.000

Table 1: Correlation Matrix

Outlier detection is also an important before conducting the PCA method. The variables are standardized. We decided that there is no outlier point, because all values of the standardized variables are within ± 3 standard deviation.

Next step is to look at Kaiser's Measure of Sampling Adequacy to look at partial correlations and make sure the correlation matrix can be factored. If Kaiser's MSA is larger than .8 the covariance matrix can be factored. In our case is the Overall MSA = 0.78081716. We conclude that this is close enough to 0.8 to continue.

trstprl	trstlgl	trstplt	trstprt	trstep	trstun	imbgeco	imueclt	imwbcnt
0.8613	0.8742	0.7330	0.7247	0.8005	0.7961	0.7791	0.8344	0.6995

Table 2: Kaiser's Measure of Sampling Adequacy

3) Testing the Assumptions:

In order to be able to reach satisfying conclusions, the following three assumptions about the data should be checked.

- **Linearity**: The relationship between the variables should be linearly related. This is satisfied in the data, because all variables are measured in a 1-10 scale.
- **Random sampling:** Each participant contributes one score on each observed variable. These sets of scores represents a random sample drawn from the population of interest.
- **Bivariate normal distribution:** Each pair of observed variables displays a bivariate normal distribution (looks like an elliptical scattergram when plotted).

4) Conducting the Method:

PCA method is conducted in SAS, using *proc princomp*. At the start, we need to decide how many principal components to retain. In order to do so, we can use several methods:

Eigenvalues above one: Count how many eigenvalues of correlation matrix exceed value of 1. From eigenvalues of correlation matrix we see that only the first two eigenvalues (5.93 and 1.67) are above 1. So we are suggested to take two PC-s from this method.

	Eigenvalue	Difference	Proportion	Cumulative
1	5.93800283	4.26126921	0.6598	0.6598
2	1.67673362	0.95940482	0.1863	0.8461
3	0.71732880	0.44750244	0.0797	0.9258
4	0.26982637	0.11728492	0.0300	0.9558
5	0.15254145	0.04584116	0.0169	0.9727
6	0.10670029	0.01864451	0.0119	0.9846
7	0.08805578	0.04045248	0.0098	0.9944
8	0.04760330	0.04439573	0.0053	0.9996
9	0.00320757		0.0004	10.000

Table 3: Eigenvalues of the Correlation Matrix

Explained variance: Looking at cumulative proportion of the variance explained and analyzing how many PC-s are enough to explain at least 70% of the variance. From the *Table 3*, we can conclude that by retaining 2 PC-s we are aiming at explaining 85% of variance of the original variables.

Scree plot: By looking at this plot of eigenvalues we should notice where it 'flattens' down and take the amount of PC-s just before the 'flattening' point.

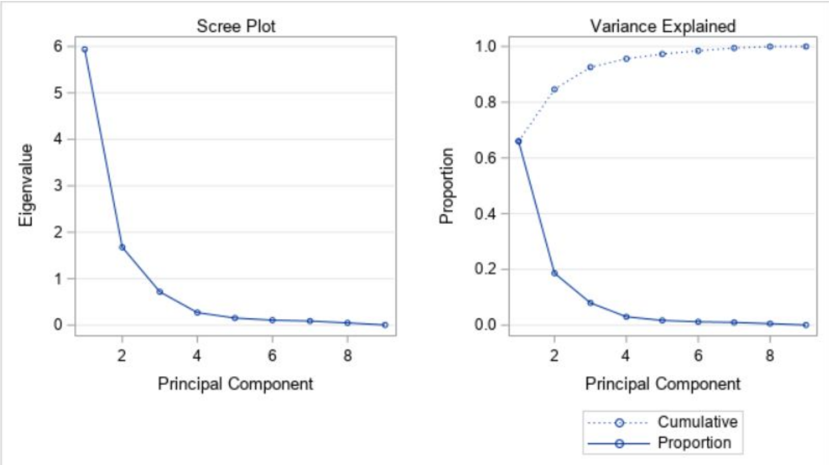

Figure 1: Scree Plot

From scree plot it is easy to see that after the 2nd eigenvalue there is no sharp 'fall' in the plot anymore, therefore retaining 2 PC-s is suggested.

After we determined with the help of 3 methods above that we are going to perform PCA on 2 principal components, we run analysis with 2 PC-s and checked first unrotated loadings of observed variables on factors and then we used two rotations, the results will be discussed in the next section.

5) Interpreting the Solution:

Resulting from the above conducted methods of determining number of PC-s to keep, we are left with 2 principal components. If we perform PCA without rotating PC-s, we do not get simple structure of the solution (we saw cases where variables were loading on both principal components). In order to make solution interpretable, we used varimax rotation.

	Factor1		Factor2	
trstprl	82	*	-52	*
trstlgl	76	*	-54	*
trstplt	91	*	-38	
trstprt	92	*	-35	
trstep	64	*	17	
trstun	86	*	25	
imbgeco	81	*	47	
imueclt	79	*	50	
imwbcnt	77	*	53	*

Table 4: Factor Pattern without Rotation

	Factor 1		Factor 2	
trstprl	22		95	*
trstlgl	16		92	*
trstplt	38		91	*
trstprt	40		90	*
trstep	57	*	33	
trstun	79	*	42	
imbgeco	91	*	24	
imueclt	91	*	20	
imwbcnt	93	*	17	

Table 5: Factor Pattern with Varimax Rotation

The threshold to find the loading on the factor high was taken at 0.5. Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.5 are flagged by an '*'. First factor (PC1) will be labelled as "trust in immigration and international policies". The second factor (PC2) will be labelled as "trust in domestic politics".

6) Comparing the Alternatives:

Two different rotation methods are tested as alternative solutions, namely *varimax* and *quartimax*. The proc factor procedure with NFACT = 2 is used in SAS.

The difference with the rotations can be seen with the variance explained by factors (PCs). After the rotations, two PCs have relatively similar amount of variance explained from original variables.

	Varimax	Quartimax
Factor 1	3.8287398	4.1163501
Factor 2	3.7859967	3.4983864

Table 6: Explained Variances After Transformations with Varimax and Quartimax

Rotated factor patterns are helpful to label the first two PCs selected. 0.5 is selected as the flag threshold. These patterns are given below in *Table 7* and *Table 8*. First factor (PC) includes high values of structural loadings for the variables trstep (trust in the European Parliament), trstun (trust in the United Nations), imbgeco (immigration bad or good for country's economy), imueclt (country's cultural life undermined or enriched by immigrants) and imwbcnt (immigrants make country worse or better place to live); whereas the second factor (PC) includes high values of structural loading for the variables trstprl (trust in the country's parliament), trstlgl (trust in legal system), trstplt (trust in politicians) and trstprt (trust in political parties). This is consistent with the labelling we did above.

	Factor1		Factor2	
trstprl	22		95	*
trstlgl	16		92	*
trstplt	38		91	*
trstprt	40		90	*
trstep	57	*	33	
trstun	79	*	42	
imbgeco	91	*	24	
imueclt	91	*	20	
imwbcnt	93	*	17	

Table 7: Factor Pattern with Varimax Rotation

	Factor 1		Factor 2	
trstprl	28		92	*
trstlgl	22		91	*
trstplt	44		88	*
trstprt	46		87	*
trstep	59	*	29	
trstun	81	*	37	
imbgeco	92	*	17	
imueclt	92	*	14	
imwbcnt	93	*	10	

Table 8: Factor Pattern with Quartimax Rotation

7) Conclusion:

Due to the redundancy in variables in the dataset, PCA method is conducted to reduce the number of observed variables into a smaller number of principal components that accounted for most of the variance in the original variables. According to the outputs in the PCA method, it is decided to use two PCs for further analysis. Then, two rotation methods are tested to obtain alternative solutions. According to the result from rotated solutions, the two PCs selected can be classified as "trust in immigration and international policies" and "trust in domestic politics".

EXPLORATORY FACTOR ANALYSIS (EFA):[1]

1) Problem Statement:

In this analysis we are looking at the dataset "humanvalues.txt" from the 2016 ESS based on 21 questions asked to 1766 Belgian individuals. This subset of questions is focusing on Human Values. In short, by performing a factor analysis on responses to this questionnaire, we are able to determine the number of constructs measured by this questionnaire (three) as well as the nature of those constructs.

2) Descriptive Statistics:

The data provided to us contains the responses of 1766 participants towards questions asked about human values.The scale they could rate their answers by varies between 1 and 6, with an additional score of 8 (In research about political sciences there is often a category which refers to "not applicable to me" and we assume the number 8 refers to this answer) We decided to keep theses participants as they are not influencing our interpretation nor conclusion.

The first question we had to ask ourselves was if we had enough data to pursue the analysis and make a meaningful conclusion There are 2 rules of thumb:

- Have at least 100 participants
- There have to me more than 10 times as many participants as variables

In our case both are fulfilled. 1766 participants with 21 variables (21*10 = 210). So we can pursue the research.

We first standardized the data and looked at the correlation matrix. Surprisingly, correlations between variables were rarely over .35, so at first glance it is hard to explain the construct with latent factors. But we still suspect that we are using the right technique in order to explore the dataset and will reach the results we are aiming at. Therefore did we look at the Measure of Sampling Adequacy to look at partial correlations and make sure the correlation matrix can be factored. If Kaiser's MSA is larger than .8 the covariance matrix can be factored. This is the case for all variables except one, which has an MSA of .76, so we consider the matrix to be acceptable to be factored.

3) Testing the Assumptions:

- **Interval-level measurement and linearity:** As mentioned above the answers to the questionnaire have been given on a linear scale (1 "*Very much like me*" to 6 "*Not like me at all*").
- **Random sampling**: Each participant contributed one score for each observed variable. These sets of scores represent a random sample drawn from the Belgian population.
- **Multivariate normality:**[2] Responses obtained from the participants demonstrate an approximate multivariate normal distribution.

[1] O'Rourke, Norm, and Larry Hatcher. 2013. A Step-by-Step Approach to Using SAS® for Factor Analysis and Structural Equation Modeling, Second Edition. Cary, NC: SAS Institute Inc.

[2] Bivariate normal distribution. Each pair of observed variables should display a bivariate normal distribution (e.g., they should form an elliptical scattergram when plotted). When the maximum likelihood method is used to extract factors, the output provides a significance test for the null hypothesis that the number of factors retained in the current analysis is sufficient to explain the observed correlations. The following assumption should be met for the probability value associated with this test to be valid.

4) Conducting the Method:

We start by extracting the factors. In order to determine the number of latent variables that we are going to keep we need to look at the eigenvalues of the correlation matrix that are above 1, look at the proportion of variance explained, analyse the scree plot, perform maximum likelihood and conduct the Chi-squared test in order to see how many factors should be retained. At first we calculated the eigenvalues of the correlation matrix and looked at how many of them were above 1.

Eigenvalues of the Reduced Correlation Matrix: Total = 6.55731718 Average = 0.3122532				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.00726542	2.23146613	0.6111	0.6111
2	1.77579929	0.49323510	0.2708	0.8819
3	1.28256418	0.86254693	0.1956	1.0775
4	0.42001725	0.11397380	0.0641	1.1416

Table 9: Eigenvalues of the reduced correlation matrix

From the table above we concluded that only the first 3 eigenvalues are above the threshold, so we are suggested to use 3 latent variables in the analysis. moreover we reach more than 100% looking at the cumulative proportion of variance explained with just 3 factors. This happens when using ml because of estimates of squared multiple correlation as a starting point. At first glance would we go with 3 factors, but to be sure we continued the methods.

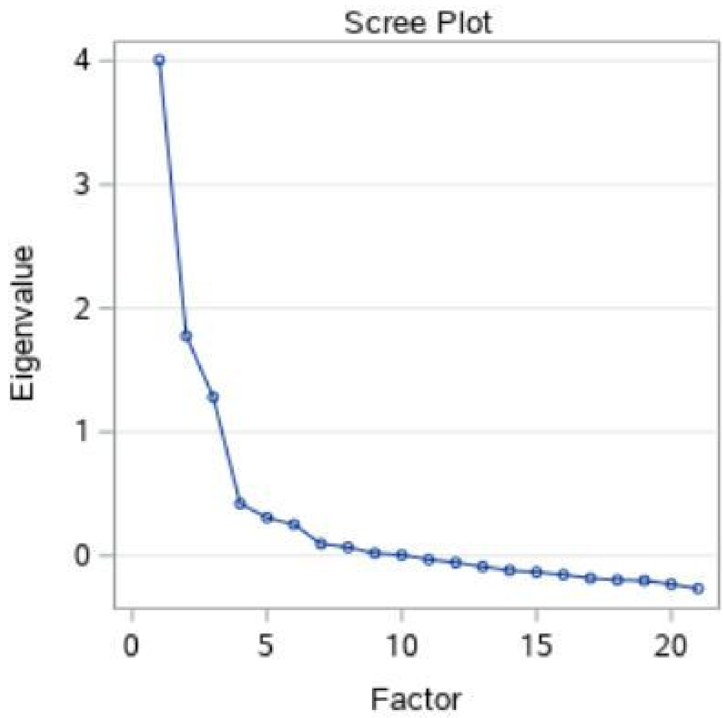

Figure 2: Scree plot of Eigenvalues

Looking at the scree plot we notice that the graph flattens after the third eigenvalue, so we again are suggested to retain 3 factors.

Re-running the analysis using maximum likelihood and looking at the Chi-squared test we notice the following:

Significance Tests Based on 1766 Observations			
Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	210	8954.3817	<.0001
HA: At least one common factor			
H0: 3 Factors are sufficient	150	1178.0279	<.0001
HA: More factors are needed			

Table 10: Significance Tests

This test suggest us to use 4 latent variables. Although we are concluding to use 3 factors in our analysis, as suggested by the eigenvalues larger than 1, the proportion of variance explained and the scree plot (3 out of 4 tests suggest 3 factors). Once we have determined the number of latent variables (3), we run the EFA with 3 factors and unrotated for factor patterns in order to check for a simple structure. Simple structure means that the pattern possesses two characteristics:

- Most of the variables have relatively high factor loadings on only one component and near zero loadings on the other components
- Most components have relatively high loadings for some variables and near-zero loadings for the remaining variables.

We have discovered that some of the variables were loading on more than 1 factor, which makes it hard to interpret. In order to make the solution interpretable we used factor rotation *promax* (oblique). Although this is for correlated factors this rotation method can still be used in uncorrelated cases, as ours (correlation was around .25 for all 3 the factors. Alternatives like *varimax* or *quartimax* could also have been used. After the rotation we received the following result with a simple structure.

Factor Structure (Correlations)						
Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.4 are flagged by an '*'.						
	Factor1		Factor2		Factor3	
ipcrtiv	44	*	31		9	
imprich	41	*	-20		33	
ipeqopt	15		50	*	18	
ipshabt	55	*	-1		38	
impsafe	6		24		53	*
impdiff	58	*	31		5	
ipfrule	5		16		56	*
ipudrst	18		60	*	18	
ipmodst	-4		44	*	29	
ipgdtim	56	*	23		12	
impfree	44	*	36		11	
iphlppl	22		60	*	21	
ipsuces	62	*	5		46	*
ipstrgv	16		33		54	*
ipadvnt	62	*	6		2	
ipbhprp	1		35		51	*
iprspot	34		3		53	*
iplylfr	26		55	*	19	
impenv	20		55	*	23	
imptrad	11		25		46	*
impfun	53	*	21		2	

Table 11: Factor structure

Each of the variables only loads on one of the factors.

5) Interpreting the Solution:

Resulting from the above conducted methods, we are left with 3 factors. Each referring to a specific personal attitude towards human values. The table below gives an overview of each of the variables and the factor it refers to.

Factor 1 can be described as increasing self worth. All the variables relate to increasing the status of one's self. Factor 2 is being good to the world around one's self. Being kind and caring towards others. Factor 3 is being a law abiding citizen. Following rules and traditions with the goal of having a safe environment for one's self.

Factor 1 ***Increasing self worth***	**Factor 2** ***Being good to the world around***	**Factor 3** ***Being law abiding***
Ipcrtiv: being creative	Ipeqopt: treating people equal	Impsafe: living in safe surroundings
Imprich: being rich	Ipudrst: understanding different people	Ipfrule: following the rules
Ipshabt: being admired	Ipmodst: being honest and modest	Ipstrgv: having a strong government
Impdiff: trying new things	Iphlppl: caring for others	Ipbhprp: behaving properly
Ipgdtim: having a good time	Iplyfr: being loyal to friends	Iprspot: getting respect from others
Impfree: being free	Impenv: caring for nature and environment	Imptrad : following traditions and customs
Ipsuces: being successful		
Ipadvnt: seeking adventure		
Impfun: having fun		

Table 12: Factor interpretation table

6) Comparing the Alternatives:

We looked at several possible options to perform this analysis. First we tried to look at unrotated factor solutions, this turned out to be hard to interpret. Secondly we tried *promax* rotation and obtained the solution we described in the previous section. Third, we tried two other rotation methods (*varimax* and *quartimax*). The results of those 2 rotation methods were very similar to the one we chose.

Earlier we mentioned that while using the ml method to determine the number of factors to be retained, we were suggested to use more that 3 factors. We redid our analysis using 4 factors. Which had no significant use to our analysis, creating a fourth factor only containing one variable.

7) Conclusion:

Concluding our analysis of the dataset “human values” from the 2016 ESS. We retain 3 factors. The first one referring to values which are related to increasing of self worth. The second one referring to the attitude towards others and nature, being kind. And the final one referring to the values of being law abiding and security.

APPENDIX A - SAS CODES FOR PCA

```
/*Reading and printing data*/
data politics;
infile      "C:\Users\tejksedopc\Desktop\Classes\Multivariate
Statistics\Assignment 1\politics.txt";
input  cntry$  trstprl  trstlgl  trstplt  trstprt  trstep  trstun
imbgeco imueclt imwbcnt;
run;
proc print data=politics;
run;

/*Correlation*/
proc corr data=politics;
run;

/*Outliers*/
proc standard data=politics
                 mean=0
                 std=1
                 out=std_politics;
proc print data=std_politics;
run;

/*Principal Component Analysis (PCA)*/
proc princomp out = politics_pca;
run;

/*Alternative solutions and rotations*/
proc factor data=politics
               simple
               method=prin
               priors=one
               mineigen=1
               plots=scree
               rotate=varimax,quartimax
               nfact=2
               round
               flag=0.50;
var  trstprl  trstlgl  trstplt  trstprt  trstep  trstun  imbgeco
imueclt imwbcnt;
run;
```

APPENDIX B - SAS CODES FOR EFA

```
proc factor data = work.import
               simple
               msa
               corr
               method=prin
               priors=smc
               plots=scree
               rotate=promax
               round
               flag=.4;
var ipcrtiv imprich ipeqopt ipshabt impsafe impdiff ipfrule
ipudrst ipmodst ipgdtim impfree iphlppl ipsuces ipstrgv
ipadvnt ipbhprp iprspot iplylfr impenv imptrad impfun;
run;

proc factor data = WORK.import
               simple
               msa
               simple
               method=ml
               priors=smc
               plots=scree
               rotate=promax
               round
               flag=.4;
var ipcrtiv imprich ipeqopt ipshabt impsafe impdiff ipfrule
ipudrst ipmodst ipgdtim impfree iphlppl ipsuces ipstrgv
ipadvnt ipbhprp iprspot iplylfr impenv imptrad impfun;
run;

proc factor data = WORK.import
               simple
               method=prin
               priors=smc
               nfact=3 /*nfact = 4*/
               plots=scree
               rotate=promax
               round
               flag=.4;
var ipcrtiv imprich ipeqopt ipshabt impsafe impdiff ipfrule
ipudrst ipmodst ipgdtim impfree iphlppl ipsuces ipstrgv
ipadvnt ipbhprp iprspot iplylfr impenv imptrad impfun;
run;
```