Final Project Report: The Study of Factor Analysis and Its Applications

04/23/2017

BANA 7047

Surineni, Sriharsha

This project involved discovering the math and processes behind Factor Analysis as a Data Mining Tool. The objective was to dive deep into the understanding of this topic and to be able to use R/Rstudio to apply Factor Analysis to a real-world dataset. This case example uses data displaying the results from a survey about the anxiety students face learning a new computer software, and we were able to determine a breakdown of the specific reasons this anxiety is caused.

1. Executive Summary

For this final project we chose to develop a stronger understanding of the topic of Factor Analysis, and then be able to use this knowledge to conduct factor analysis with a real-world dataset. As a team, we have used various resources, found in the references section, to learn what factor analysis is, the math behind the process, compared it to Principle Component Analysis and then finally conduct factor analysis in R. We tried to apply factor analysis in SAS Enterprise Miner but were unable to do to software restrictions. We then were able to take what we have learned and then teach the class about the purpose of factor analysis.

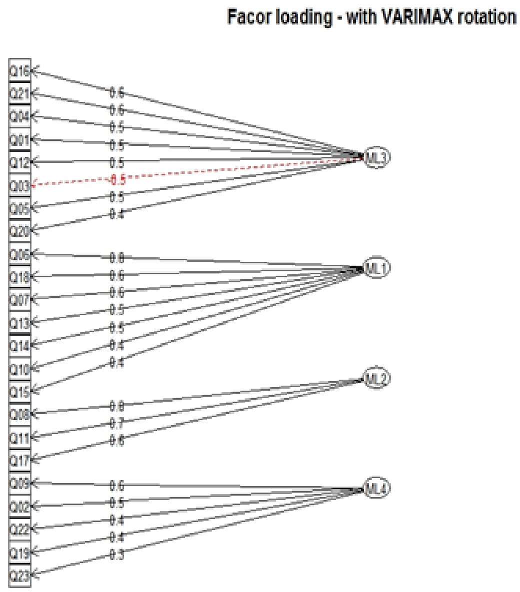
Factor Analysis is used particularly in science fields to capture latent features that cannot be captured underlying in a high dimensional dataset. We found that a process called principle component analysis that is incredibly similar to factor analysis and proceeded to determine the differences of each and how to determine which process is best used.

Factor Analysis	**Principal Component Analysis**
Factor Analysis is based on the assumption that there are some factors affecting all of the variables that we cannot measure directly	There is no such assumption here. The main aim is to form some index variables or principal components based on the measure variables
Factors account for the underlying common variances between the variables, and it leaves an explained part of the variance unique to each variable	PCA accounts for the total variance in the data
We can think of this as a regression model, where the variables are dependent on the factors, and error terms (which is the variance not explained by the factors)	The components can be represented as a simple linear combination of the variables. This is a deterministic model.

Afterwards we found a dataset to use to conduct factor analysis on. The data set that was be used to understand and explain factor analysis was found online from a website that offers open data. Further research in the dataset found that this set is a dataset used for the textbook Discovering Statistics Using SPSS by Andy Field. The set is named SAQ.sv and it is the results of a psychological survey conducted to measure a trait that the author creates called “SPSS Anxiety.” Field generated 23 questions for the survey based on interviews with anxious and non-anxious students that measured various aspects of students’ anxiety towards learning the software.

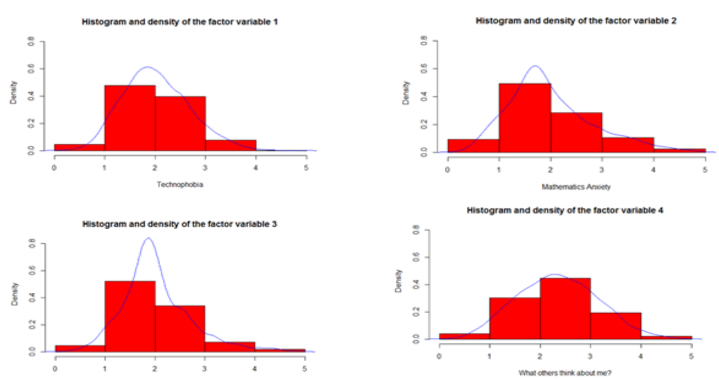
The questions are written as statements that followed by a five-point scale ranging from ‘strongly disagree’ to ‘strongly agree.’ The main objective of the questionnaire is to predict how anxious a student is about learning SPSS. Field also wanted to use this data to explore if the “SPSS Anxiety” could be broken down into specific forms of anxiety.

The results from the Factor Analysis are as follows:



Technophobia	Mathematics	Statistics	What my friends think
I have little experience of computers	I have never been good at mathematics	Statistics makes me cry	My friends will think I m stupid for not being able to cope with SPSS
All computers hate me	I did badly at mathematics at school	Standard deviations excite me	My friends are better at statistics than me
Computers are useful only for playing games	I slip into a coma whenever I see an equation	I dream that Pearson is attacking me with correlation coefficients	Everybody looks at me when I use SPSS
I worry that I will cause irreparable damage because of my incompetence with computers		I don t understand statistics	My friends are better at SPSS than I am
Computers have minds of their own and deliberately go wrong whenever I use them		People try to tell you that SPSS makes statistics easier to understand but it doesn't	If I am good at statistics my friends will think I m a nerd
Computers are out to get me		I weep openly at the mention of central tendency	
SPSS always crashes when I try to use it		I can t sleep for thoughts of eigen vectors	
		I wake up under my duvet thinking that I am trapped under a normal distribution	

Results were found for both Oblimin and Varimax rotation, however Varimax rotation resulted in coherent structure of variables, segregating all the question into above four categories which made more sense intuitively.



From four factor model with varimax rotation, it is observed that density plot of the factor representing peer pressure is the only one which is not skewed. Density plots of Technophobia, Mathematics anxiety and Statistics anxiety are left skewed which suggests most of the students are below normal level on anxiety scale.

2 Factor Analysis

Background & Important Information

Factor Analysis is particularly useful in social science fields such as psychology, socio-economics in capturing latent features which cannot be captured underlying in a high dimensional dataset of survey questions or other measures involved in the study. Factor analysis can be thought of as modeling as these latent features being manifested into measurable parameters involved in our study. For example, in measuring “mental ability of a person”, there is no single measure of level of mental ability, instead the level of mental ability manifests into various other measurable parameters such as “Quantitative aptitude score”, “mental age” etc., Factor analysis tries to group together all the parameters changing together throughout our dataset and explain the group of parameters with the help of one or more factors.

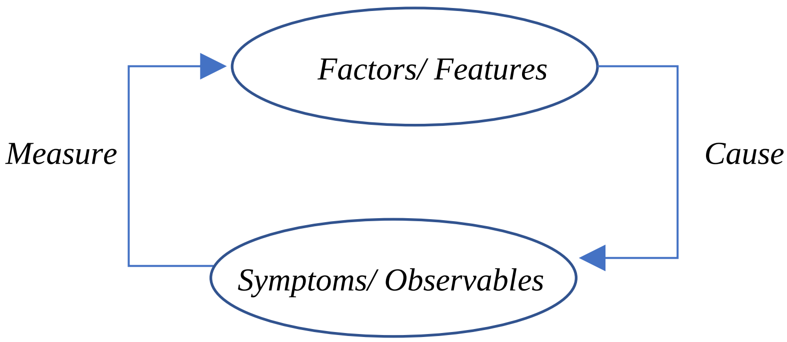


Figure 1: Factor Analysis Flow Chart

Mathematics of Factor Analysis:

Factor Analysis is similar to Principal Component Analysis and it differs in only one way. We change the basis of representation to eigen vector space of variance-covariance matrix of the observations in PCA, but in for Factor Analysis, basis of representation is changed to eigen vector space of common variance matrix of the observations (variance matrix of the observations can be separated into common and specific variance matrices). Observations, X are modeled as:

$$X=\mu+\lambda F+\delta$$

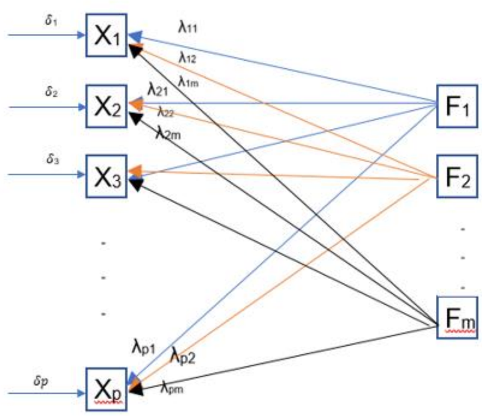


Figure 2: Explaining model of Factor Analysis

where X is represented as three components, 1) constant, 2) linear combination of factors/ features and 3) unexplained error terms. Assumptions of the model are:

a. $E(X) = \mu$
b. $E(F) = 0$
c. $Cov(\delta)=0$
d. $Cov(F, \delta) = 0$ and
e. $Cov(F) = 0$ (Only for Orthogonal Factor Analysis, where factors are independent)

Model:

$$X-\mu=\lambda F+\delta$$

From the above assumptions, covariance matrix of X, turns out to be:

$$\Sigma=\lambda\lambda T+\phi$$

Where, $\lambda\lambda T$ represents the variance of factors / common variance generally called as communality and ϕ represents the specific variance.

Non-uniqueness of the solution:

There are p(1+p)/2 degrees of freedom available to estimate [(pm+p) – (½)m(m-1)] parameters. Let s = [p(1+p)/2] - [(pm+p) – (½)m(m-1)]].

- If s=0, unique solution
- If s>0, several solutions
- If s<0, no solution

Methods of Estimation:

- Principal Component Method:

Covariance matrix of the observations is decomposed into eigen value – eigen vector decomposition similar to the case of Principal Component Analysis(PCA)

- Principal Factor Method:

$$\Sigma - \phi = \lambda\lambda^T$$

Covariance matrix minus specific variance matrix is decomposed in Principal Factor Method. As the method suggests we pick principal factors by decomposing communality into eigen value – eigen vector decomposition. As, covariance matrix is estimated from the sample and there is no way to know specific variance part of it, analysis starts with some random initialization of specific variance (diagonal matrix) and goes on iteratively replacing those values with estimated values after choosing m factors to represent the substantial variance. This process converges when the change in the specific variance goes below some set threshold (0.001 usually)

- Maximum Likelihood Estimation:

In this method we obtain estimates which are most *likely* given the observed data. In other words, we assume a probability distribution function and write the probability of observing the data as a function of the unknown parameters. This function called the likelihood function is maximized to obtain the parameter estimates which are called the maximum likelihood estimates. One advantage of the maximum likelihood method is the ability to test hypothesis about number of common factors:

$$H_o : \mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi}$$

$$H_a : \mathbf{\Sigma} \text{ is any positive definite matrix}$$

Under the ML restriction $tr\left(\hat{\mathbf{\Sigma}}^{-1}\mathbf{S}_n\right) = p$. So, $-2\ln\Lambda$ can be written as:

$$n\ln\frac{|\mathbf{\Sigma}_0|}{|\mathbf{S}_n|} = n\ln\left(\frac{|\hat{\mathbf{L}}\hat{\mathbf{L}}^T + \hat{\mathbf{\Psi}}|}{|\mathbf{S}_n|}\right)$$

with degrees of freedom equal to

$$\nu-\nu_0 = (1/2)p(p+1) - \left[mp + p - (1/2)m(m-1)\right] = 1/2\left[\left(p-m\right)^2 - p - m\right]$$

where p=number of variables and m=number of factors. The test is rejected if

$$-2\ln\Lambda > \chi^2_{\nu-\nu_0,\alpha}$$

Difference between Factor Analysis and Principal Component Analysis (PCA) :

Both Factor Analysis, and Principal Component Analysis are dimensionality reduction techniques. They are similar, and this always leads to confusion as to which method should be used. However, there are some fundamental differences between the two.

Factor Analysis	Principal Component Analysis
Factor Analysis is based on the assumption that there are some factors affecting all of the variables that we cannot measure directly	There is no such assumption here. The main aim is to form some index variables or principal components based on the measure variables
Factors account for the underlying common variances between the variables, and it leaves an explained part of the variance unique to each variable	PCA accounts for the total variance in the data
We can think of this as a regression model, where the variables are dependent on the factors, and error terms (which is the variance not explained by the factors)	The components can be represented as a simple linear combination of the variables. This is a deterministic model.

Figure 3: Table explaining differences of PCA and Factor Analysis

Principal Component Analysis can be used effectively when the main aim is data reduction. It makes more sense to use Factor Analysis when we have some theoretical concepts we want to test, such as measuring factors that cannot be directly measured.

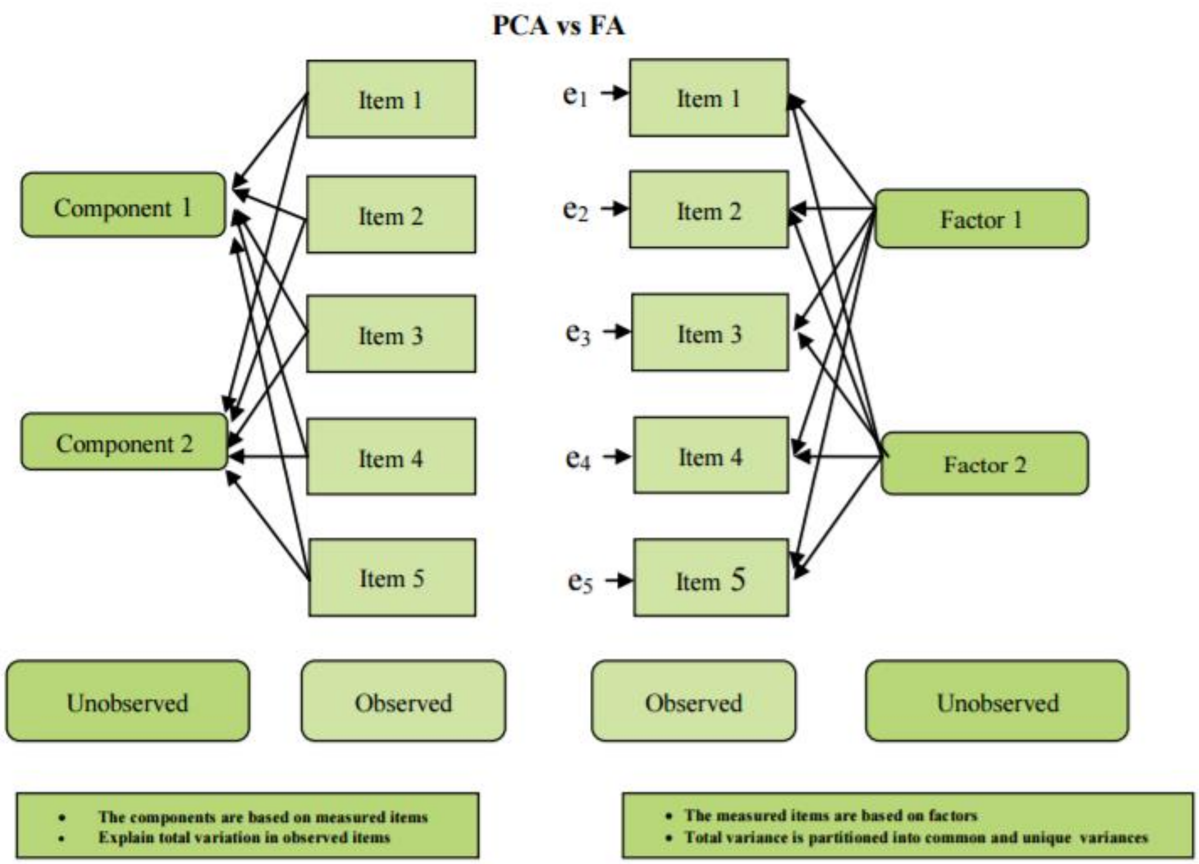


Figure 4: Difference in the Mathematical Model of Factor Analysis and PCA (Source: https://d1pbog36rugm0t.cloudfront.net/-/media/ualberta/faculties-and-programs/centres-institutes/community-university-partnership/research/ecmap-reports/comparisonpca.pdf)

Objective of Project:

The objective of our project is to dive deep into the topic of factor analysis as a data mining tool as well as fully understand how and why factor analysis is used. Main objectives can be further broken down into following objectives.

1. Understanding basic mathematics behind factor analysis
2. Study on practical applications of factor analysis
3. Conducting factor analysis on the real dataset

Then we will be using an example dataset to test what we have learned. Afterwards our goal is to be able to take our topic of factor analysis, teach the class what we have learned, and walk through our example set and results to further explain how the process of factor analysis works.

Practical Applications:

1. Dimensionality reduction

Dimensionality reduction is the process of reducing number of random variables under consideration. PCA, LDA, QDA are some of the other methods used to reduce dimensions.

Uses:

i) Reducing time and storage capacity
ii) Improving model performance through reduction of multicollinearity
iii) Easier exploratory analysis

2. Identification of groups of inter-related variables

Factor analysis can be used to find groups of variables that are in a way have a common attribute affecting each one of them. One can arrive at crucial findings by observing such variables.

For example:

Carroll used factor analysis to build his Three Stratum Theory. He found that a factor called "broad visual perception" relates to how good an individual is at visual tasks. He also found a "broad auditory perception" factor, relating to auditory task capability. Furthermore, he found a global factor, called "g" or general intelligence, that relates to both "broad visual perception" and "broad auditory perception". This means someone with a high "g" is likely to have both a high "visual perception" capability and a high "auditory perception" capability, and that "g" therefore explains a good part of why someone is good or bad in both of those domains (Source: Wiki)

3. Estimating abstract quantities

Factor analysis can also be used to estimate immeasurable quantities such as Intelligence, Poverty, Aptitude, Strength

For example: To estimate the athletic ability, we can arrive it by combining various variables such as running speed, maximum lifted weight, reflex time, Highest jump etc. For estimating poverty of country, we might use GDP, Income of lowest 10 percent, Property of lowest 10 percent etc.

Factor analysis is used in a variety of fields such as

1. Investing
2. Human Resources
3. Insurance
4. Restaurants
5. Education

Data Set

Background:

The data set that will be used to understand and explain factor analysis was found online from a website that offers open data. Further research in the dataset found that this set is a dataset used for the textbook Discovering Statistics Using SPSS by Andy Field. The set is named SAQ.sv and it is the results of a psychological survey conducted to measure a trait that the author creates called “SPSS Anxiety.” Field generated 23 questions for the survey based on interviews with anxious and non-anxious students that measured various aspects of students’ anxiety towards learning the software.

The questions are written as statements that followed by a five-point scale ranging from ‘strongly disagree’ to ‘strongly agree.’ The main objective of the questionnaire is to predict how anxious a student is about learning SPSS. Field also wanted to use this data to explore is the “SPSS Anxiety” could be broken down into specific forms of anxiety.

Variable Explanation:

The SPSS anxiety results has 2571 observations and 23 variables, with each variable labeled indicating its corresponding question. A print out of the entire survey and its questions can be found in Figure 1. Each possible answer and categorized from ‘strongly disagree’ to ‘strongly agree’ and each response is given a number. There are two types of questions on the survey. The first type of question has a score of strongly disagree=1, disagree=2, neither=3, agree=4, and strongly agree=5. These are any question that indicates that statistics is bad in any way. An example of this is question one, “Statistics makes me cry.” Additionally, there are other questions such as question three “Standard deviations excite me” that imply that statistics are good. These are scored on an opposite scale with strongly disagree=5, disagree=4, neither=3, agree=2, and strongly agree=1. The switching of scales depending on the question ensures that someone who is anxious about learning SPSS will score highly on the SAQ index, and alternatively if they love statistics they will have a low score on the SAQ index. Field explains within the textbook that "These reverse-phrased items are important for reducing response bias; participants will actually have to read the items in case they are phrased the other way around. For factor analysis, this reverse phrasing doesn’t matter"(Field 675). He delves deeper explaining all that would happen in this case is that one would get a negative factor for any reverse phrase items.

SD = Strongly Disagree, D = Disagree, N = Neither, A = Agree, SA = Strongly Agree

		SD	D	N	A	SA
1	Statistics make me cry	O	O	O	O	O
2	My friends will think I'm stupid for not being able to cope with SPSS	O	O	O	O	O
3	Standard deviations excite me	O	O	O	O	O
4	I dream that Pearson is attacking me with correlation coefficients	O	O	O	O	O
5	I don't understand statistics	O	O	O	O	O
6	I have little experience of computers	O	O	O	O	O
7	All computers hate me	O	O	O	O	O
8	I have never been good at mathematics	O	O	O	O	O
9	My friends are better at statistics than me	O	O	O	O	O
10	Computers are useful only for playing games	O	O	O	O	O
11	I did badly at mathematics at school	O	O	O	O	O
12	People try to tell you that SPSS makes statistics easier to understand but it doesn't	O	O	O	O	O
13	I worry that I will cause irreparable damage because of my incompetence with computers	O	O	O	O	O
14	Computers have minds of their own and deliberately go wrong whenever I use them	O	O	O	O	O
15	Computers are out to get me	O	O	O	O	O
16	I weep openly at the mention of central tendency	O	O	O	O	O
17	I slip into a coma whenever I see an equation	O	O	O	O	O
18	SPSS always crashes when I try to use it	O	O	O	O	O
19	Everybody looks at me when I use SPSS	O	O	O	O	O
20	I can't sleep for thoughts of eigenvectors	O	O	O	O	O
21	I wake up under my duvet thinking that I am trapped under a normal distribution	O	O	O	O	O
22	My friends are better at SPSS than I am	O	O	O	O	O
23	If I am good at statistics people will think I am a nerd	O	O	O	O	O

Figure 5: The SPSS Anxiety Questionnaire

Factor Analysis

Factor Analysis of SPSS Anxiety data set:

- Data Preparation:

As mentioned earlier, dataset consists of 23 survey questions corresponding to different characteristics of anxiety among students each rated along the scale of 1 to 5 (1 – Strongly Disagree, 2- Disagree, 3- Neither, 4- Agree and 5- Strongly Agree). The dataset is clean with no missing values. Factor variables are converted to numeric based on the above-mentioned scale.

Viability Check:

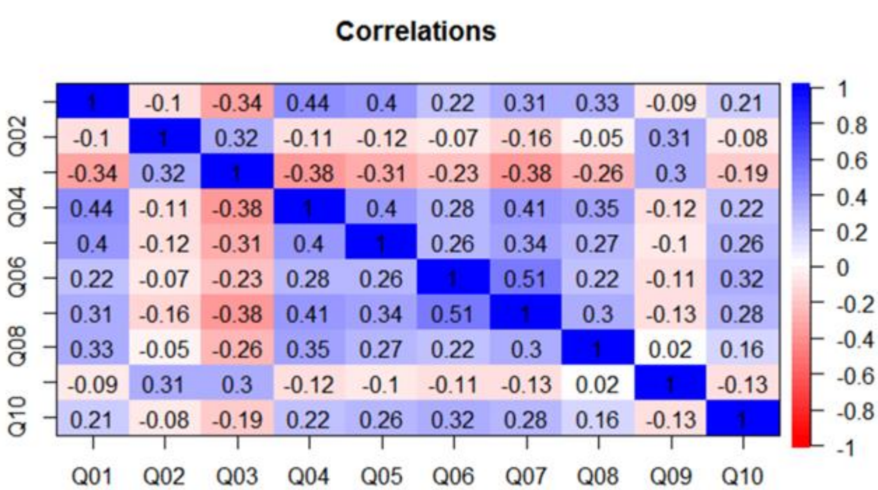


Figure 6: The SPSS Anxiety Questionnaire Variable Correlations

Visual check of correlation matrix suggests substantial number of pair-wise correlations are greater than 0.3. Bartlett's Sphericity test resulted in p-value of 3.075166e-41 which is significant at 1% level. So, we reject the null hypothesis that the correlation matrix is identity matrix, suggesting non-zero correlations among the variables.

- Number of Factors:

There are various guidelines and ways to choose the optimal number of factors explaining the common hidden construct present in the data. Four such methods are used here and three of them resulted in 4 factors and Acceleration Factor method resulted in 1 optimal factor.

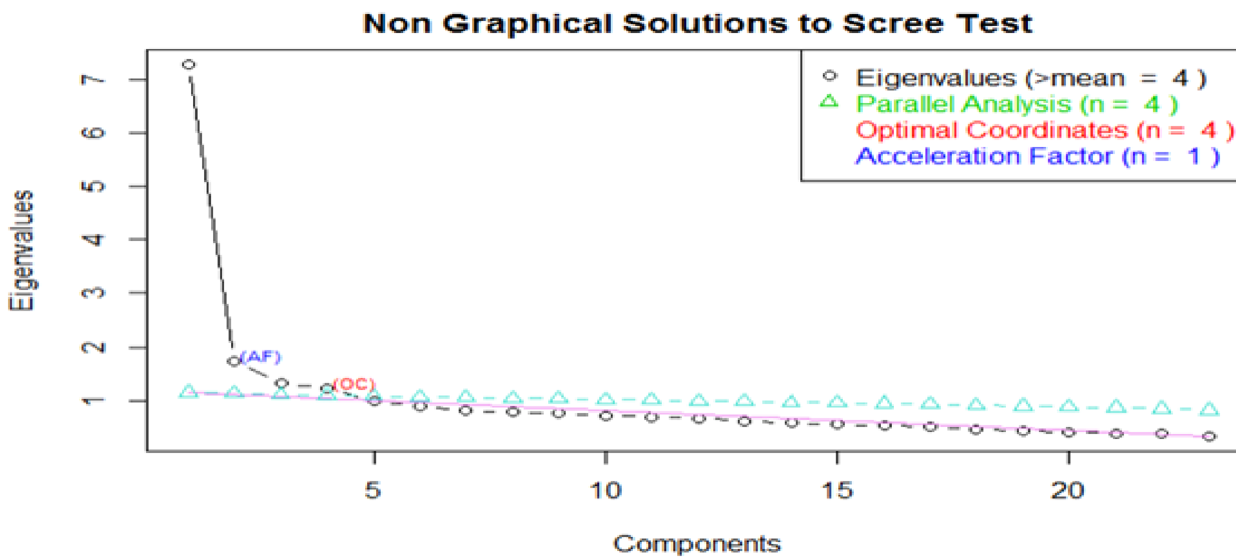


Figure 7: Scree Test Chart

Factor Analysis is carried out for both 1 factor and 4 factors setups using maximum likelihood estimation. Maximum likelihood is chosen to test for the goodness of fit of the assumed model.

Model Estimation:

2. One factor model:

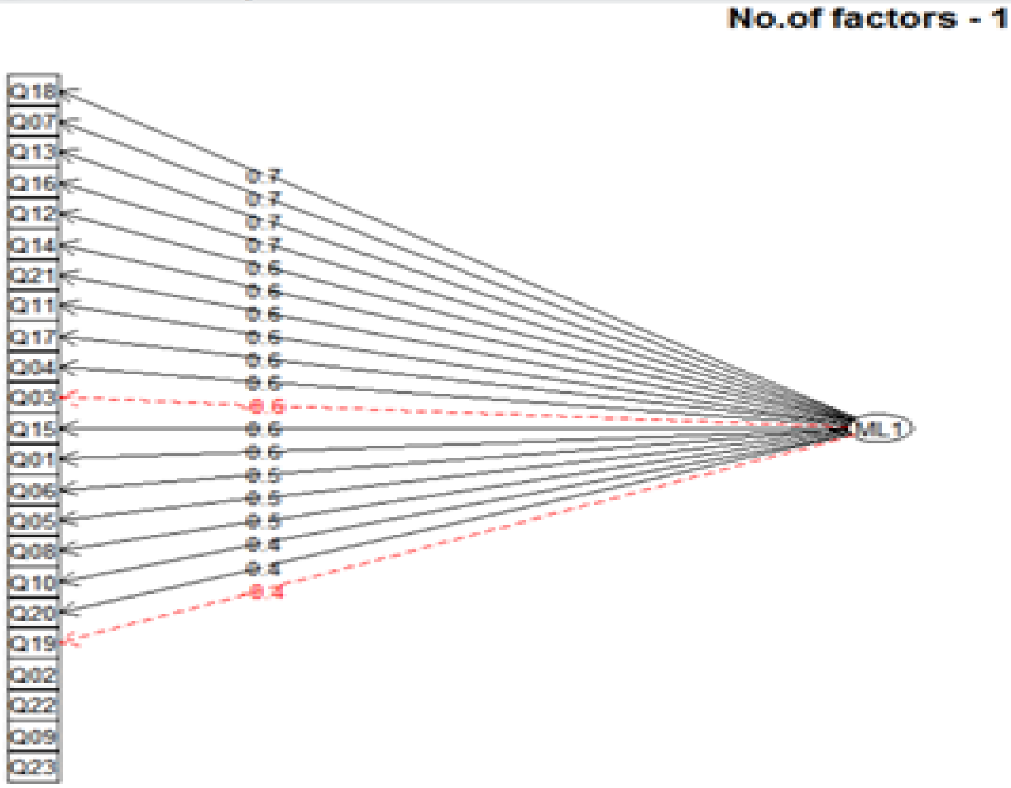


Figure 8: Factor Analysis with 4 Factors

One factor model with "VARIMAX" rotation captured all the questions related to anxiousness as one category.

Four factor model:

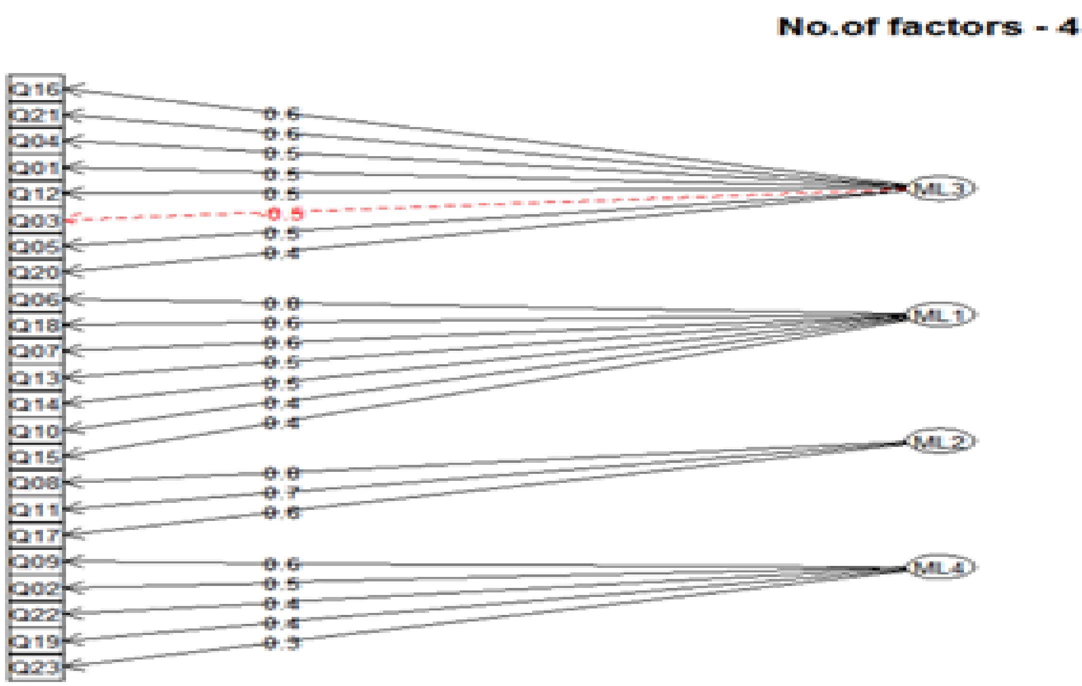


Figure 9: Factor Analysis with 4 Factors

Four factor model clearly separated the questions into 4 different categories.

Goodness of fit:

	RMSEA		RMSR	
No. of Factors	Value	Acceptable Range	Value	Acceptable Range
1	0.0072	<0.06	0.065	<0.08
4	0.0032	<0.06	0.026	<0.08

Figure 10: Factor Analysis Goodness of Fit

Goodness of fit criteria used here are RMSEA and RMSR which are measures of deviation of estimated correlation matrix of model from the sample correlation matrix. RMSE and RMSR of both One factor model and Four factor model are in the acceptable ranges as shown in the table, suggesting reasonably good estimation of the underlying construct.

Rotation:

As mentioned earlier, rotation is possible and sometimes useful since there is no unique solution of factor analysis for any chosen number of factors. If at all the estimated factors does not clearly segregate the variables into different categories, rotational transformation can be applied to obtain separate clusters of variables. There are two main categories of rotations possible:

Orthogonal Rotation:

Results in orthogonal or independent factor variables. Some examples are: varimax, quartimax and equimax. Varimax is considered the best orthogonal rotation and consequently is used the most often in psychology research. An advantage of orthogonal rotation is its simplicity and conceptual clarity, although there are several disadvantages. In the social sciences, there is often a theoretical basis for expecting constructs to be correlated, therefore orthogonal rotations may not be very realistic because it ignores this possibility. Also, because orthogonal rotations require factors to be uncorrelated, they are less likely to produce solutions with simple structure.

Oblique Rotation:

Results in correlated factors. Direct oblimin and promax are two such rotations. If factors do not correlate these rotations may produce solutions like orthogonal rotation. An advantage of oblique rotation is that it produces solutions with better simple structure because it allows factors to correlate, and produces estimates of correlations among factors.

Results:

4 – factor model loading with and without rotation:

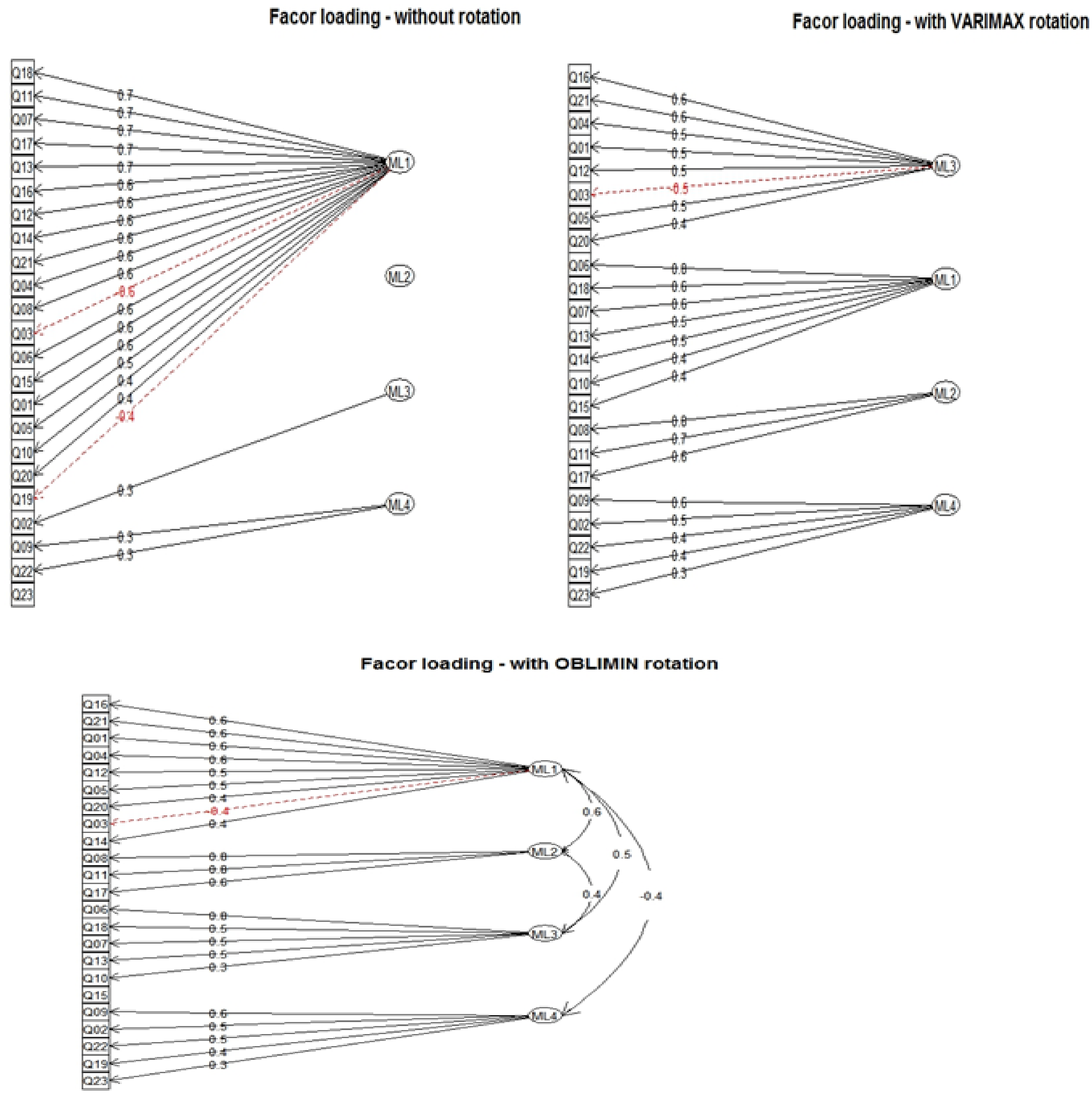


Figure 11: Factor Loading with no rotation, Varimax Roataion, and Oblimin Rotation

For four factor model without rotation, factors are not able to form a coherent structure of variables as shown in the above factor loading matrix. Varimax rotation resulted in more coherent separate categories of variables. Oblimin resulted in correlated factors.

Four factor model with "VARIMAX" rotation:

Technophobia	Mathematics	Statistics	What my friends think
I have little experience of computers	I have never been good at mathematics	Statistics makes me cry	My friends will think I m stupid for not being able to cope with SPSS
All computers hate me	I did badly at mathematics at school	Standard deviations excite me	My friends are better at statistics than me
Computers are useful only for playing games	I slip into a coma whenever I see an equation	I dream that Pearson is attacking me with correlation coefficients	Everybody looks at me when I use SPSS
I worry that I will cause irreparable damage because of my incompetence with computers		I don t understand statistics	My friends are better at SPSS than I am
Computers have minds of their own and deliberately go wrong whenever I use them		People try to tell you that SPSS makes statistics easier to understand but it doesn't	If I am good at statistics my friends will think I m a nerd
Computers are out to get me		I weep openly at the mention of central tendency	
SPSS always crashes when I try to use it		I can t sleep for thoughts of eigen vectors	
		I wake up under my duvet thinking that I am trapped under a normal distribution	

Figure 12: Four Factor Model with Varimax rotation

Varimax rotation resulted in coherent structure of variables, segregating all the question into above four categories which makes a lot of sense intuitively.

Four factor model with "OBLIMIN" rotation:

Statistics	Mathematics	Technophobe	What my friends think	NA
Statiscs makes me cry	I have never been good at mathematics	I have little experience of computers	My friends will think I'm stupid for not being able to cope with SPSS	Computers are out to get me
negative(Standard deviations excite me)	I did badly at mathematics at school	All computers hate me	My friends are better at statistics than me	
I dream that Pearson is attacking me with correlation coefficients	I slip into a coma whenever I see an equation	Computers are useful only for playing games	negative (Everybody looks at me when I use SPSS)	
I don't understand statistics		I worry that I will cause irreparable damage because of my incompetenece with computers	My friends are better at SPSS than I am	
People try to tell you that SPSS makes statistics easier to understand but it doesn't		SPSS always crashes when I try to use it	If I'm good at statistics my friends will think I'm a nerd	
Computers have minds of their own and deliberately go wrong whenever I use them				
I weep openly at the mention of central tendency				
I can't sleep for thoughts of eigen vectors				
I wake up under my duvet thinking that I am trapped under a normal distribtion				

Figure 13: Four Factor Model with Oblimin rotation

Oblimin rotation resulted in same conclusion about the survey as Varimax rotation with minor deviations, which suggests there may not be any correlation among the factors.

Conclusions:

Factor scores are estimated from the obtained models for one factor model and four factor model with varimax rotation to analyze the level of anxiety among students.

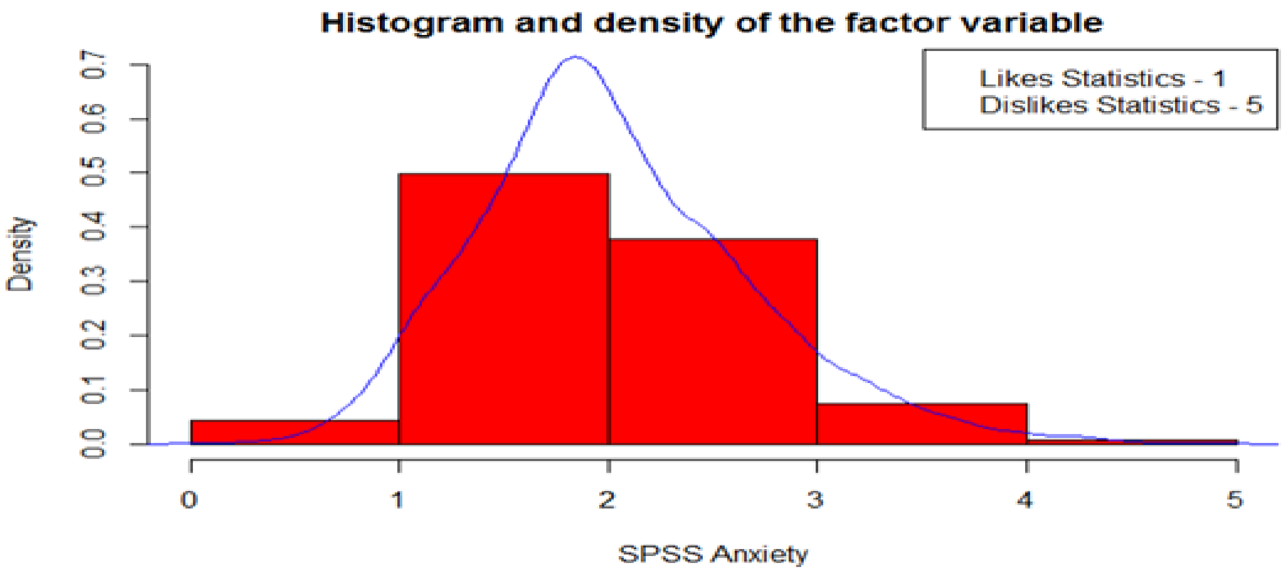


Figure 14: Entire Model Histogram and Density Plot

From one factor model, over all anxiety level of students is as shown in the above density plot. Most of the students are not that anxious

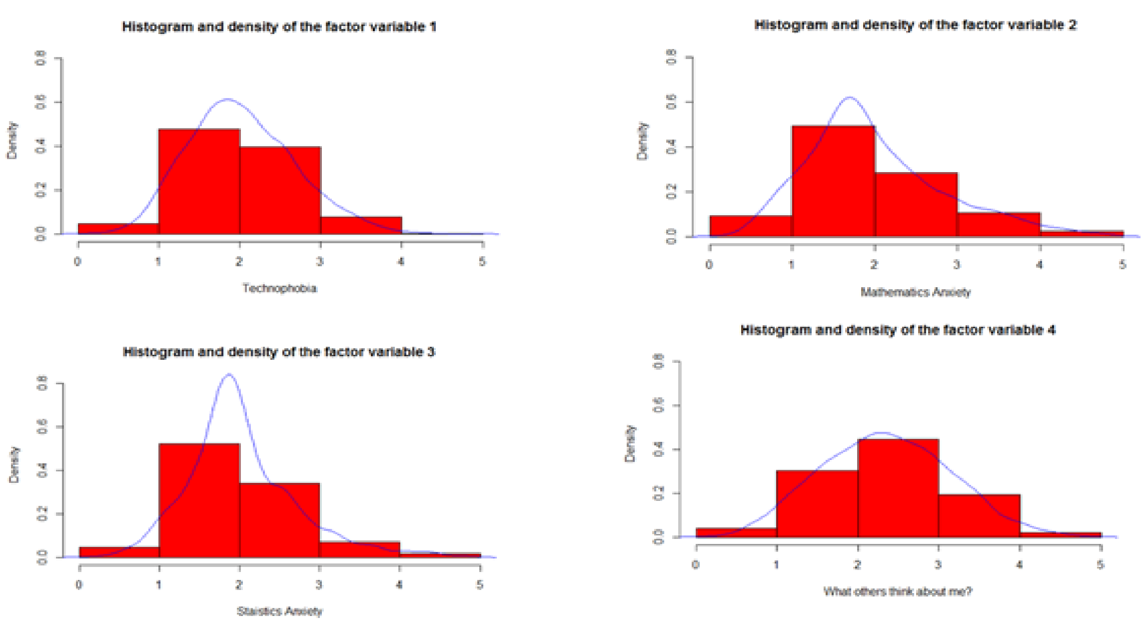


Figure 15: Individual Factor Histogram and Density Plots

From four factor model with varimax rotation, it is observed that density plot of the factor representing peer pressure is the only one which is not skewed. Density plots of Technophobia, Mathematics anxiety and Statistics anxiety are left skewed which suggests most of the students are below normal level on anxiety scale.

References

"Factor Analysis." Factor Analysis Adapted from Ewart Thomas's "Factor Analysis and Cluster Analysis". Stanford, n.d. Web. 2 Apr. 2017. <https://web.stanford.edu/class/psych253/tutorials/FactorAnalysis.html>.

Field, Andy P. Discovering Statistics Using SPSS. Third ed. London: SAGE Publications, 2009. Print.

Mod-01 Lec-33 Factor Analysis. Perf. Dr J Maiti. Applied Multivariate Statistical Modeling. Department of Management, IIT Kharagpur, 9 May 2014. Web. 2 Apr. 2017. <https://www.youtube.com/watch?v=n3y3xLNoPk4&list=PLbMVogVj5nJRt-ZxRG1KRjxNoy7J_IaW2&index=33>.

Walker, Ian. "Statistics for Psychology." Data Files. University of Bath, 2008. Web. 02 Apr. 2017.

A Comparison of Principal Components Analysis and Factor Analysis for Uncovering the Early Development Instrument (EDI) Domains (https://d1pbog36rugm0t.cloudfront.net/-/media/ualberta/faculties-and-programs/centres-institutes/community-university-partnership/research/ecmap-reports/comparisonpca.pdf)

The Fundamental Difference Between Principal Component Analysis and Factor Analysis (http://www.theanalysisfactor.com/the-fundamental-difference-between-principal-component-analysis-and-factor-analysis/)