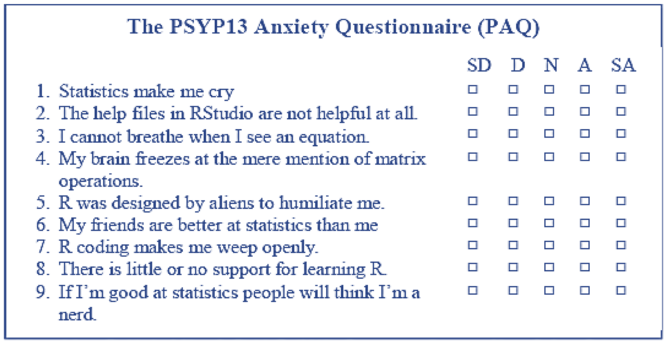
Assignment 1 – Principal Components Analysis

Introduction

Factor analysis is a broad expression used to describe methods by which one can reduce existing data in a dataset into a smaller more understandable structure. PCA is a kind of factor analysis. It is done to generate a hypothesis, not to test it. The goal of any PCA is to summarize the variance of all the variables into a smaller number of main components. A rotation transformation method is conducted which causes a maximization of the variance of the principal components, e.g. through Varimax or Oblimin methods (Eid, Gollwitzer & Schmitt 2011, Bortz & Schuster, 2010).

The fictional dataset given for this assignment consisted of 300 students who had taken a short questionnaire (seen in Figure 1 on the right) regarding their attitudes towards learning "R." The response anchors consisted of a 5-point Likert scale ranging from 'strongly disagree,' 'disagree,' 'neither agree nor disagree,' 'agree,' and 'strongly agree.' The assignment was to use a Principle Components Analysis (PCA) method to assess how and if the anxiety scale measures different aspects or not. The following parts will deal with the results of the data reduction as well as offer an interpretation.



The PSYP13 Anxiety Questionnaire (PAQ)

	SD	D	N	A	SA
1. Statistics make me cry	□	□	□	□	□
2. The help files in RStudio are not helpful at all.	□	□	□	□	□
3. I cannot breathe when I see an equation.	□	□	□	□	□
4. My brain freezes at the mere mention of matrix operations.	□	□	□	□	□
5. R was designed by aliens to humiliate me.	□	□	□	□	□
6. My friends are better at statistics than me	□	□	□	□	□
7. R coding makes me weep openly.	□	□	□	□	□
8. There is little or no support for learning R.	□	□	□	□	□
9. If I'm good at statistics people will think I'm a nerd.	□	□	□	□	□

Adapted from Field, A., Miles, J., Field, Z. (2012). *Discovering Statistics Using R*. Sage Publishing

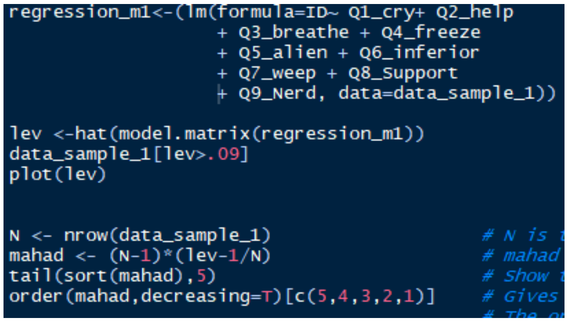
Figure 1 – Questionnaire PAQ: PSYP13 Anxiety Questionnaire

Result

A PCA is normally done in five steps, and this was also the approach used in this assignment.

1) Create a correlation matrix of the manifest variables.

To do this, the data set needed to be prepared. It was then analyzed using the *> summary()* function. The result yielded no irregularities on min/max values, means, etc. or missing data. The function *sum(is.na.data.frame())* confirmed the no missing's.



```
regression_m1<-(lm(formula=ID~ Q1_cry+ Q2_help
                    + Q3_breathe + Q4_freeze
                    + Q5_alien + Q6_inferior
                    + Q7_weep + Q8_Support
                    + Q9_Nerd, data=data_sample_1))

lev <-hat(model.matrix(regression_m1))
data_sample_1[lev>.09]
plot(lev)


N <- nrow(data_sample_1)                    # N is t
mahad <- (N-1)*(lev-1/N)                    # mahad
tail(sort(mahad),5)                         # Show t
order(mahad,decreasing=T)[c(5,4,3,2,1)]     # Gives
                                            # The o
```

Figure 2 – R-code syntax for the multivariate outliers

The following step, although strictly not required for this type of PCA, was to test the questions for normality. This was done using the following codes (Figure 2) for all the questions, as well as gender and age:

The result showed that all the variables failed to meet the normality assumption.

Following this, the data was checked for multivariate outliers. This was done using the code seen on the left.

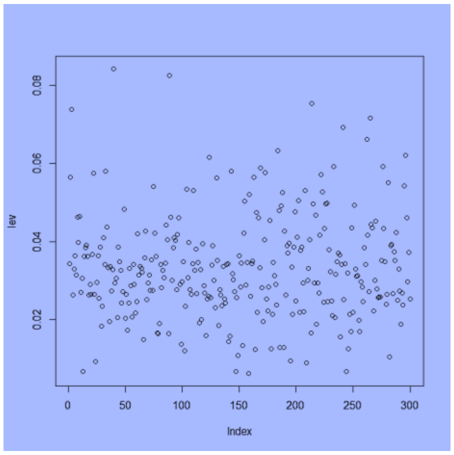


Figure 3 – Leverage plot over multivariate outliers.

In the first step, a regression expression is stated with the variable "ID" working as a dummy variable. The linear combination equation will always yield the "ID" as the outcome variable. As the next step the leverage, a measure similar to the Mahalanobis distance was used and subsequently plotted. The Mahad test = (df)*(leverage model-1)/N. The code *tail(sort(Mahad, 5)* showed the five highest Mahad scores which were then ordered using the order function. In this way, it is possible to find out the upper limit for the multivariate outliers. Again, no outliers were found as can be seen in Figure 2.

The first step of the PCA, creating a matrix, was then started using the following code: First, the columns were excluded which we did not need in our matrix: *subset1_PCA<-within(data_sample_1, rm(ID, age, sex))*. Then the matrix was saved as a correlation matrix: subset1_PCA<-cor(subset1_PCA)

2) Orthogonal Rotation transformation

The rotation is conducted using the code *PCA1<-princomp(covmat=d_cor).*

The summary function summary(PCA1,loadings=TRUE) will, in this case, generate a table with the importance of the components as measured by the standard deviation, the component proportion of variance and the cumulative proportion. The first results indicated that the first three factors might be the most important ones as their components had a standard deviation >1 (Kaiser-Guttman criteria).

The factor loadings and their interpretation depend on sample size. As our sample was large, anything above r.>0.25 is noteworthy while anything below is deemed unsure. If the sample is <200 the level of insecurity rises. This is equivalent to the coefficients in multiple regression (Steyer, 2002).

3) Deciding on the numbers of principal components

The decision can be made in a number of ways. In the R. Script, these methods were all explored, and they all returned similar results: Keep the first three components.

4) Renewed rotation in order to reach a "simple structure" (Thurstone, 1947)

As a decision on the number of components had been successfully made a renewed rotation was performed using the "oblimin" rotation transformation which does not require it factors to be independent of another.

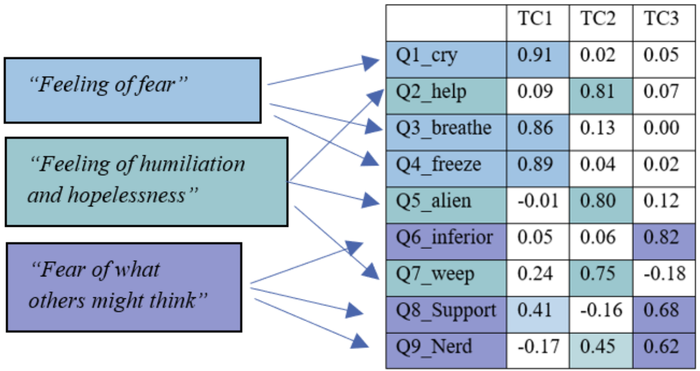
Conclusion and Discussion

5) Interpretation of the principal components as judged by the factor weights

The interpretation of the principal components found in figure 3 and the final result is that the first factor has high loadings on item 1, 3_breathe and 4. It also has a moderate loading of item 8. Roughly it appears to be measuring a construct related to "*the feeling of fear*".

The second component comes with high loading on item 2, 5, and 7. This component could then be hypothesized to measure a construct related to "*feelings of humiliation and helplessness*."

The third component comes with high loadings on the remaining items 6 and 9. However, it also shows a moderately high loading on item 8. It is difficult to summarize this component, but it appears to be related to a certain "*fear of what other people might think*." Item 8, however, remains diffuse.


"Feeling of fear"
"Feeling of humiliation and hopelessness"
"Fear of what others might think"

	TC1	TC2	TC3
Q1_cry	0.91	0.02	0.05
Q2_help	0.09	0.81	0.07
Q3_breathe	0.86	0.13	0.00
Q4_freeze	0.89	0.04	0.02
Q5_alien	-0.01	0.80	0.12
Q6_inferior	0.05	0.06	0.82
Q7_weep	0.24	0.75	-0.18
Q8_Support	0.41	-0.16	0.68
Q9_Nerd	-0.17	0.45	0.62

Figure 4 – A rough guess at the underlying constructs in the PAQ as judged by the factor loadings on the principal components

As always, it is difficult to come to any conclusion after just one type of analysis. This is after all exploratory and, performing a confirmatory factor analysis will be the only way to shed more light on the matter. However, as judged by the standardized loadings it is, in this case, safe to say that choosing three constructs is the best we can do as judged by the analyses made. Further information on this can be found in the extensive r-syntax.

Assignment 2 – Multivariate Scaling

Introduction

With multidimensional scaling (MDS), differences or similarities can be easily visualized on a two-dimensional plane. It is a form of data reduction, where MDS usually tries to show the information contained in a distance matrix. In the dataset done by Wish et al (1970), 18 students were given the assignment to rate the global similarity of different pairs of nations, e.g. 'Brazil and Egypt,' (12 countries in total). The answers were to be rated on a 9-point scale ranging from `1=very different' to `9=very similar'. The table in the text file “Nations.txt” showed the mean similarity ratings of the students.

The goal of the assignment was to visualize the relative perception of nations using MDS. In a first step, the mean similarity scores were converted to dissimilarity scores. In this way, differences and thus distances become maximized. A second step was performed using a non-metric multidimensional scaling method which reduced the dissimilarity data matrix to two dimensions. Finally, the corresponding stress value and possible interpretations of the coordinates were calculated. These will be discussed in the last sections.

Result

The first steps in the assignment were done similarly as in part one. No missings were found, nor was anything else amiss with the data. The dissimilarity matrix was used when performing the non-metric MDS procedure. Figure 1 was programmed using the code visible down below. The final result was a stress value of 20.01 which indicates a poor fit. Anything below 10 is considered good.

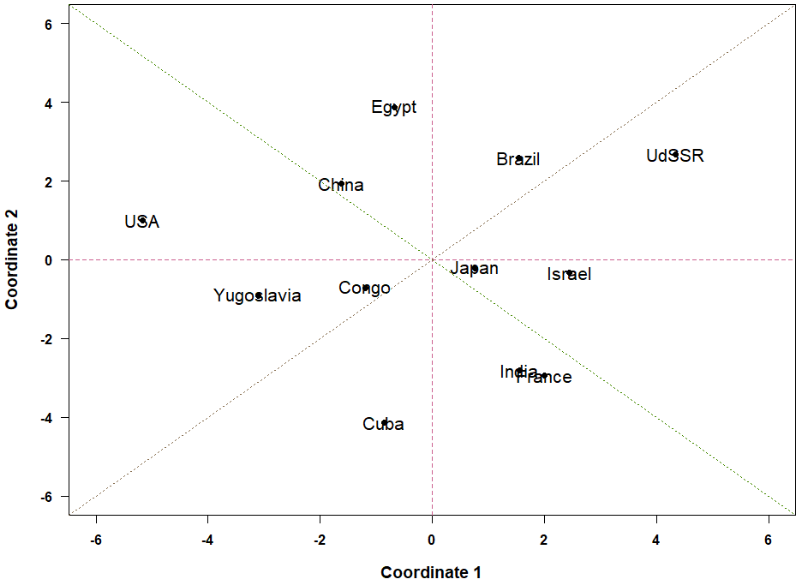
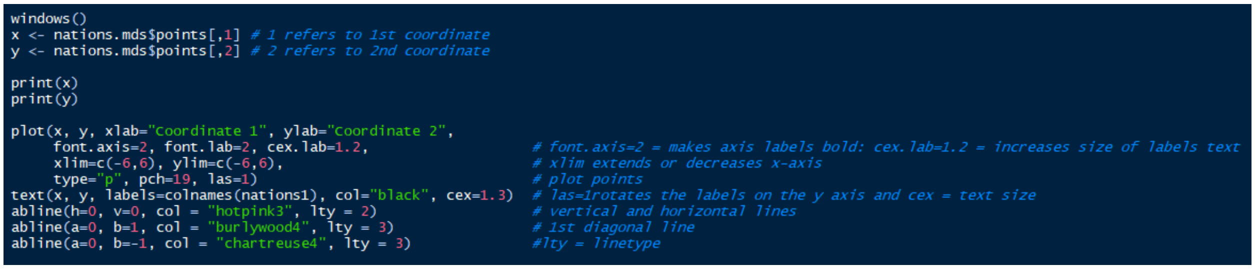


Figure 5 – The result of the MDS procedure. The two coordinates and the countries presented show a random pattern



```
windows()
x <- nations.mds$points[,1] # 1 refers to 1st coordinate
y <- nations.mds$points[,2] # 2 refers to 2nd coordinate

print(x)
print(y)

plot(x, y, xlab="Coordinate 1", ylab="Coordinate 2",
     font.axis=2, font.lab=2, cex.lab=1.2,                # font.axis=2 = makes axis labels bold: cex.lab=1.2 = increases size of labels text
     xlim=c(-6,6), ylim=c(-6,6),                          # xlim extends or decreases x-axis
     type="p", pch=19, las=1)                             # plot points
text(x, y, labels=colnames(nations1), col="black", cex=1.3) # las=1rotates the labels on the y axis and cex = text size
abline(h=0, v=0, col = "hotpink3", lty = 2)               # vertical and horizontal lines
abline(a=0, b=1, col = "burlywood4", lty = 3)             # 1st diagonal line
abline(a=0, b=-1, col = "chartreuse4", lty = 3)           #lty = linetype
```

Figure 6 – The r-code for the graph shown in figure 5.

Conclusion and Discussion

The high-stress score of 20.01 indicated a very poor fit of the model, implying that the forced reduction into the two dimensions causes a big information loss. A three-dimensional model may be better suited in order to show similarities and dissimilarities between nations and result in more comprehensible clusters of nations. However, given the small sample, this is highly unlikely. Due to this, it is difficult if not even impossible to interpret the coordinates. Neither coordinates could be said to visualize either rich/poor countries or conflict-torn/ peaceful countries. Nor could democratic vs. dictatorship be a possible explanation. The axes seem to be just the sum of the many implicit and explicit associations we all have lumped together into two dimensions.