An introduction to

Principal Component Analysis & Factor Analysis

Using SPSS 19 and R (psych package)

Robin Beaumont
robin@organplayers.co.uk

Monday, 23 April 2012

Acknowledgment:
The original version of this chapter was written several years ago by Chris Dracup

Contents

1 Learning outcomes 3

2 Introduction 4

2.1 Hozinger & Swineford 1939 5

3 Overview of the process 6

3.1 Data preparation 6

3.2 Do we have appropriate correlations to carry out the factor analysis? 6

3.3 Extracting the Factors 8

3.4 Giving the factors meaning 9

3.5 Reification 10

3.6 Obtaining factor scores for individuals 11

3.6.1 Obtaining the factor score coefficient matrix 11

3.6.2 Obtaining standardised scores 11

3.6.3 The equation 11

3.7 What do the individual factor scores tell us? 12

4 Summary - to Factor analyse or not 13

5 A typical exam question 14

5.1 Data layout and initial inspection 14

5.2 Carrying out the Principal Component Analysis 15

5.3 Interpreting the output 16

5.4 Descriptive Statistics 16

5.5 Communalities 16

5.6 Eigenvalues and Scree Plot 17

5.7 Unrotated factor loadings 17

5.8 Rotation 18

5.9 Naming the factors 19

5.10 Summary 19

6 PCA and factor Analysis with a set of correlations or covariances in SPSS 20

7 PCA and factor analysis in R 21

7.1 Using a matrix instead of raw data 23

8 Summary 24

9 Reference 24

1 Learning outcomes

Working through this chapter, you will gain the following knowledge and skills. After you have worked through it you should come back to these points, ticking off those with which you feel happy.

Learning outcome	**Tick box**
Be able to set out data appropriately in SPSS to carry out a Principal Component Analysis and also a basic Factor analysis.	
Be able to assess the data to ensure that it does not violate any of the assumptions required to carry out a Principal Component Analysis/ Factor analysis.	
Be able to select the appropriate options in SPSS to carry out a valid Principal Component Analysis/factor analysis.	
Be able to select and interpret the appropriate SPSS output from a Principal Component Analysis/factor analysis.	
Be able explain the process required to carry out a Principal Component Analysis/Factor analysis.	
Be able to carry out a Principal Component Analysis factor/analysis using the psych package in R.	
Be able to demonstrate that PCA/factor analysis can be undertaken with either raw data or a set of correlations	

After you have worked through this chapter and if you feel you have learnt something not mentioned above please add it below:

2 Introduction

This chapter provides details of two methods that can help you to restructure your data specifically by reducing the number of variables; and such an approach is often called a "data reduction" or "dimension reduction" technique. What this basically means is that we start off with a set of variables, say 20, and then by the end of the process we have a smaller number but which still reflect a large proportion of the information contained in the original dataset. The way that the 'information contained' is measured is by considering the variability within and co-variation across variables, that is the variance and co-variance (i.e. correlation). Either the reduction might be by discovering that a particular linear componation of our variables accounts for a large percentage of the total variability in the data or by discovering that several of the variables reflect another 'latent variable'.

This process can be used in broadly three ways, firstly to simply discover the linear combinations that reflect the most variation in the data. Secondly to discover if the original variables are organised in a particular way reflecting another a 'latent variable' (called **Exploratory Factor Analysis – EFA**) Thirdly we might want to confirm a belief about how the original variables are organised in a particular way (**Confirmatory Factor Analysis – CFA**). It must not be thought that EFA and CFA are mutually exclusive often what starts as an EFA becomes a CFA.

I have used the term Factor in the above and we need to understand this concept a little more.

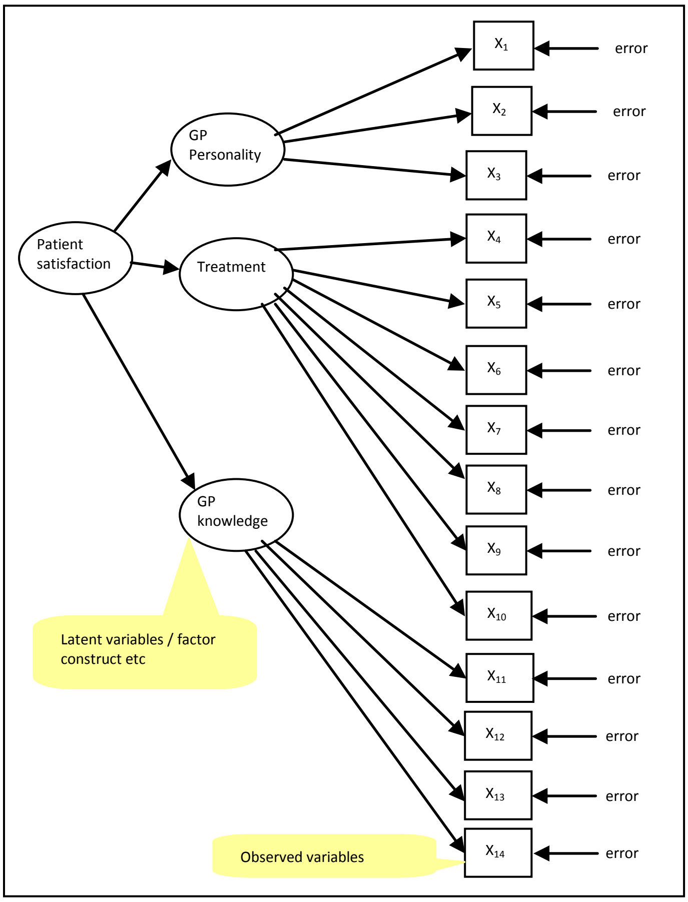
A factor in this context (its meaning is different to that found in Analysis of Variance) is equivalent to what is known as a **Latent** variable which is also called a **construct**.

construct = latent variable = factor

A latent variable is a variable that cannot be measured directly but is measured indirectly through several observable variables (called **manifest** variables). Some examples will help, if we were interested in measuring intelligence (=latent variable) we would measure people on a battery of tests (=observable variables) including short term memory, verbal, writing, reading, motor and comprehension skills etc.

Similarly we might have an idea that patient satisfaction (=latent variable) with a person's GP can be measured by asking questions such as those used by Cope et al (1986), and quoted in Everitt & Dunn 2001 (page 281). Each question being presented as a five point option from strongly agree to strongly disagree (i.e. Likert scale, scoring 1 to 5):

1. My doctor treats me in a friendly manner
2. I have some doubts about the ability of my doctor
3. My doctor seems cold and impersonal
4. My doctor does his/her best to keep me from worrying
5. My doctor examines me as carefully as necessary
6. My doctor should treat me with more respect
7. I have some doubts about the treatment suggested by my doctor
8. My doctor seems very competent and well trained
9. My doctor seems to have a genuine interest in me as a person
10. My doctor leaves me with many unanswered questions about my condition and its treatment
11. My doctor uses words that I do not understand
12. I have a great deal of confidence in my doctor
13. I feel a can tell my doctor about very personal problems
14. I do not feel free to ask my doctor questions



You might be thinking that you could group some of the above variables (manifest variables) above together to represent a particular aspect of patient satisfaction with their GP such as personality, knowledge and treatment. So now we are not just thinking that a set of observed variables relate to one latent variable but that specific subgroups of them relate to specific aspects of a single latent variable each of which is itself a latent variable.

Two other things to note; firstly often the observable variables are questions in a questionnaire and can be thought of as **items** and consequently each subset of items represents a **scale**.

Secondly you will notice in the diagram above that besides the line pointing towards the observed variable X_i from the latent variable, representing its degree of correlation to the latent variable, there is another line pointing towards it labelled error. This error line represents the unique contribution of the variable, that is that portion of the variable that cannot be predicted from the remaining variables. This uniqueness value is equal to $1-R^2$ where R^2 is the standard multiple R squared value. We will look much more at this in the following sections considering a dataset that has been used in many texts concerned with factor analysis, using a common dataset will allow you to compare this exposition with that presented in other texts.

2.1 Hozinger & Swineford 1939

In this chapter we will use a subset of data from the Holzinger and Swineford (1939) study where they collected data on 26 psychological tests from seventh – eighth grade children in a suburban school district of Chicago (file called grnt_fem.sav). Our subset of data consists of data from 73 girls from the Grant-White School. The six variables represent scores from seven tests of different aspects of educational ability, Visual perception, Cube and lozenge identification, Word meanings, sentence structure and paragraph understanding.

Descriptive Statistics (produced in SPSS)

	N	Minimum	Maximum	Mean	Std. Deviation
VISPERC	73	11.00	45.00	29.3151	6.91592
CUBES	73	9.00	37.00	24.6986	4.53286
LOZENGES	73	3.00	36.00	14.8356	7.91099
PARAGRAP	73	2.00	19.00	10.5890	3.56229
SENTENCE	73	4.00	28.00	19.3014	5.05438
WORDMEAN	73	2.00	41.00	18.0137	8.31914

Correlations

	wordmean	sentence	paragrap	lozenges	cubes	visperc
wordmean	1.000					
sentence	.696	1.000				
paragrap	.743	.724	1.000			
lozenges	.369	.335	.326	1.000		
cubes	.184	.179	.211	.492	1.000	
visperc	.230	.367	.343	.492	.483	1.000

Exercise 1.

Consider how you might use the above information to assess the data concerning:

The shape of the various distributions

Any relationships that may exist between the variables

Any missing / dodgy(!) values

Could some additional information help?

3 Overview of the process

There are many varieties of factor analysis involving a multitude of different techniques, however the common characteristic is that factor analysis is carried out using a computer although the early researchers in this area were not so lucky, with the first paper introducing factor analysis being published in 1904 by C. Spearman of Spearman's rank correlation coefficient fame, long before the friendly PC was available.

Factor analysis works only on interval/ratio data, and ordinal data at a push. If you want to carry out some type of variable reduction process on nominal data you have to use other techniques or substantially adapt the factor analysis see Bartholomew, Steele, Moustaki & Galbraith 2008 for details.

3.1 Data preparation

Any statistical analysis starts with standard data preparation techniques and factor analysis is no different. Basic descriptive statistics are produced to note any missing/abnormal values and appropriate action taken. Also in addition to this two other processes are undertaken:

1. Any **computed variables** (slickly speaking only linear transformations) are **excluded** from the analysis. These are easily identified as they will have a correlation of 1 with the variable from which they were calculated.
2. **All the variables should measure the construct in the same direction**. Considering the GP satisfaction scale we need all the 14 items to measure satisfaction in the same direction where a score of 1 represents high satisfaction and 5 the least satisfaction or the other way round. The direction does not matter the important thing is that all the questions score in the same direction. Taking question 1: *My doctor treats me in a friendly manner and question,* this provides the value 1 when the respondent agrees, representing total satisfaction and 5 when the respondent strongly disagrees and is not satisfied. However question three is different: *My doctor seems cold and impersonal.* A patient indicating strong agreement to this statement would also provide a value of 1 but this time it indicates a high level of dissatisfaction. The solution is to reverse score all these negatively stated questions.

Considering our Holzinger and Swineford dataset we see that we have 73 cases and from the descriptive statistics produced earlier there appears no missing values and no out of range values. Also the correlation matrix does not contain any '1''s except the expected diagonals.

3.2 Do we have appropriate correlations to carry out the factor analysis?

The starting point for all factor analysis techniques is the correlation matrix. All factor analysis techniques try to clump subgroups of variables together based upon their correlations and often you can get a feel for what the factors are going to be just by looking at the correlation matrix and spotting clusters of high correlations between groups of variables.

	wordmean	sentence	paragrap	lozenges	cubes	visperc
wordmean	1.000					
sentence	.696	1.000				
paragrap	.743	.724	1.000			
lozenges	.369	.335	.326	1.000		
cubes	.184	.179	.211	.492	1.000	
visperc	.230	.367	.343	.492	.483	1.000

Looking at the matrix from the Holzinger and Swineford dataset we see that Wordmean, sentence and paragraph seem to form one cluster and lozenges, cubes and visperc tests the other cluster.

Norman and Streiner (p 197) quote Tabachnick & Fidell (2001) saying that if there are few correlations above 0.3 it is a waste of time carrying on with the analysis, clearly we do not have that problem.

Besides looking at the correlations we can also consider any number of other matrixes that the various statistical computer programs produce. I have listed some below and filled in some details.

Exercise 2.

Considering each of the following matrixes complete the table below:

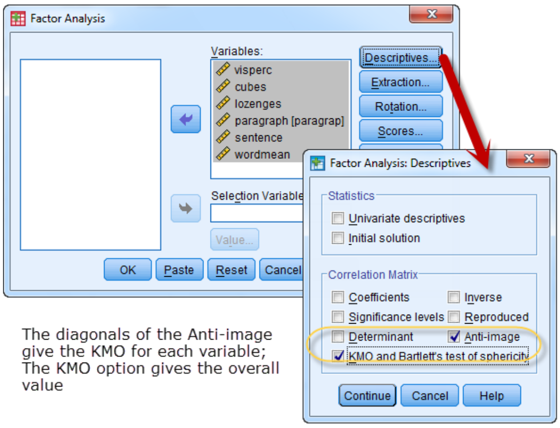
Name of the matrix	Elements are:	Good signs	Bad signs
Correlation 'R'	correlations	Many above 0.3 and possible clustering	Few above 0.3
Partial correlation		Few above 0.3 and possible clustering	Many above 0.3
Anti-image correlation	Partial correlations - reversed	Few above 0.3 and possible clustering	Many above 0.3

While eyeballing is a valid method of statistical analysis (!) obviously some type of statistic, preferably with an associated probability density function to produce a p value, would be useful to help us make this decision. Two such statistics are the Bartlett test of Sphericity and the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (usually called the MSA).

The Bartlett Test of Sphericity compares the correlation matrix with a matrix of zero correlations (technically called the identity matrix, which consists of all zeros except the 1's along the diagonal). From this test we are looking for a small p value indicating that it is highly unlikely for us to have obtained the observed correlation matrix from a population with zero correlation. However there are many problems with the test – a small p value indicates that you should not continue but a large p value does not guarantee that all is well (Norman & Streiner p 198).

The MSA does not produce a P value but we are aiming for a value over 0.8 and below 0.5 is considered to be miserable! Norman & Streiner p 198 recommend that you consider removing variables with a MSA below 0.7

In SPSS we can obtain both the statistics by selecting the menu option Analyse-> dimension reduction and then placing the variables in the variables dialog box and then selecting the descriptives button and selecting the Anti-image option to show the MSA for each variable and the KMO and Bartlett's test for the overall MSA as well:



The diagonals of the Anti-image give the KMO for each variable; The KMO option gives the overall value

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.763
Bartlett's Test of Sphericity	Approx. Chi-Square	180.331
	df	15
	Sig.	.000

Anti-image Matrices

		visperc	cubes	lozenges	paragraph	sentence	wordmean
Anti-image Covariance	visperc	.613	-.204	-.177	-.065	-.101	.091
	cubes	-.204	.676	-.210	-.017	.042	-.008
	lozenges	-.177	-.210	.615	.022	-.012	-.100
	paragraph	-.065	-.017	.022	.354	-.145	-.176
	sentence	-.101	.042	-.012	-.145	.399	-.133
	wordmean	.091	-.008	-.100	-.176	-.133	.371
Anti-image Correlation	visperc	.734[a]	-.317	-.289	-.140	-.204	.191
	cubes	-.317	.732[a]	-.326	-.034	.082	-.015
	lozenges	-.289	-.326	.780[a]	.047	-.025	-.209
	paragraph	-.140	-.034	.047	.768[a]	-.385	-.486
	sentence	-.204	.082	-.025	-.385	.803[a]	-.346
	wordmean	.191	-.015	-.209	-.486	-.346	.743[a]

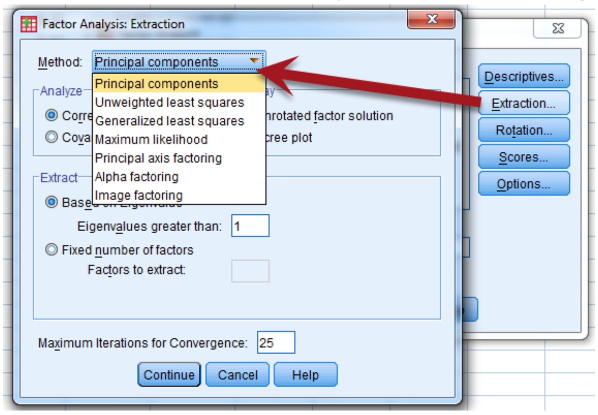
a. Measures of Sampling Adequacy(MSA)

We can see that we have good values for all variables for the MSA but the overall value is a bit low at 0.763, however Bartlett's Test of Sphericity has an associated P value (sig in the table) of <0.001 as by default SPSS reports p values of less than 0.001 as 0.000! So from the above results we know that we can now continue and perform a valid factor analysis.

Finally I mentioned that we should exclude variables that are just simple derivations of another in the analysis, say variable A = variable B + 4. A similar problem occurs with variables that are very highly correlated (this is called **multicollinearity**) and when this occurs the computer takes a turn and can't produce valid factor loading values. A simple way of assessing this is to inspect a particular summary measure of the correlation matrix called the **determinant** and check to see if it is greater than 0.00001 (Field 2012 p771). Clicking on the determinant option in the above dialog box produces a determinant value of 0.0737 for our dataset.

3.3 *Extracting the Factors*

There are numerous ways to do this, and to get an idea you just need to look at the pull down list box in SPSS shown opposite.



There are two common methods, the Principal components and the Principal axis factoring extraction methods and strictly speaking the Principal components method is not a type of factor analysis but it often gives very similar results. Let's try both and see what we get.

However there is one other thing we need to consider first. How many latent variables do we want or do we want the computer to decide for use using some criteria – the common method is to let the computer decide for use by simply selecting the *Eigenvalues greater than 1* option however there are several reasons why this is not an altogether good idea both Norman & Streiner 2008 and Field 2012 discuss them in detail. For now I'll use the dodgy eigenvalue >1 approach.

I have run both a Principal Axis and also a Principal Component Analysis below.

Principal Axis (PA)

Factor Matrix[a]

	Factor 1	Factor 2
visperc	.555	.423
cubes	.452	.553
lozenges	.585	.401
paragraph	.819	-.307
sentence	.785	-.270
wordmean	.778	-.330

Extraction Method: Principal Axis Factoring.
a. 2 factors extracted. 11 iterations required.

Communalities

	Extraction
visperc	.487
cubes	.511
lozenges	.504
paragraph	.764
sentence	.689
wordmean	.714

Extraction Method: Principal Axis Factoring.

Total Variance Explained

Factor		Extraction Sums of Squared Loadings: Total	% of Variance	Cumulative %
dimension0	1	2.747	45.775	45.775
	2	.923	15.382	61.157

Extraction Method: Principal Axis Factoring.

Principal component (PCA)

Component Matrix[a]

	Component 1	Component 2
visperc	.641	.490
cubes	.526	.660
lozenges	.670	.448
paragraph	.822	-.388
sentence	.811	-.374
wordmean	.794	-.427

Extraction Method: Principal Component Analysis (PCA).
a. 2 components extracted.

Communalities

	Extraction
visperc	.650
cubes	.712
lozenges	.650
paragraph	.826
sentence	.797
wordmean	.812

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component		Extraction Sums of Squared Loadings: Total	% of Variance	Cumulative %
dimension0	1	3.099	51.648	51.648
	2	1.349	22.478	74.126

Extraction Method: Principal Component Analysis.

You will notice that both methods extracted 2 factors. However the factor loadings (or strictly speaking the component loadings for the PCA) for the PCA are larger in absolute values as are the communalities and as a consequence the total variance explained is also greater. Here are a few pointers to help you interpret the above:

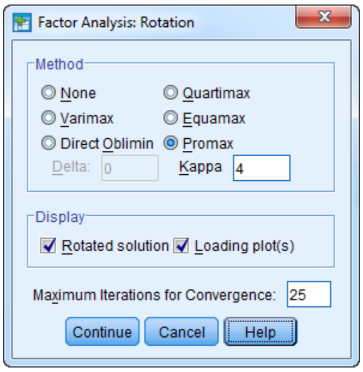
Factor loadings for the PA = correlation between a specific observed variable and a specific factor. Higher values mean a closer relationship. They are equivalent to standardised regression coefficients (β weights) in multiple regression. **Higher the value the better.**

Communality for the PA = Is the total influence on a single observed variable from all the factors associated with it. It is equal to the sum of all the squared factor loadings for all the factors related to the observed variable and this value is the same as R^2 in multiple regression. The value ranges from zero to 1 where 1 indicates that the variable can be fully defined by the factors and has no uniqueness. In contrast a value of 0 indicates that the variable cannot be predicted at all from any of the factors. The communality can be derived for each variable by taking the sum of the squared factor loadings for each of the factors associated with the variable. So for visperc = $0.555^2 + 0.423^2 = 0.4869$ and for cubes $=0.452^2 + 0.553^2 = 0.510$ These values can be interpreted the same way as R squared values in multiple regression that is they represent the % of variability attributed to the model, inspecting the total variance explained table in the above analyses you will notice that this is how the % of variance column is produced. Because we are hoping that the observed dataset is reflected in the model we **want** this value to be as **high** as possible, nearer to one the better.

Uniqueness for each observed variable it is that portion of the variable that cannot be predicted from the other variables (i.e. the latent variables). It's value is 1-communality. So for wordmean we have 1-0.714 = 0.286 and as the communality can be interpreted as the % of the variablility that is predicted by the model we can say this is the % variability in a specific observed variable that is NOT predicted by the model. This means that we **want** this value for each observed variable to be as **low** as possible. On pgae 3 referring to the diagram it is the 'error' arrow.

Total variance explained this indicates how much of the variability in the data has been modelled by the extracted factors. You might think that given that the PCA analysis models 74% of the variability compared to just 61% for the PA analysis we should go for the PCA results. However why the estimate is higher is because in the PCA analysis the initial estimates for the communalities are all set to 1 which is higher than for the PA analysis which uses an estimate of the R^2 value also whereas the PCA makes use of all the variability available in the dataset in the PA analysis the unique variability for each observed variable is disregarded as we are only really interested in how each relates to the latent variable(s). What is an acceptable level of variance explained by the model? Well one would hope for the impossible which would be 100% often analyses are reported with 60-70%.

Besides using the eigenvalue >1 criteria we could have inspected a scree pot and worked out where the factors levelled off – we will look at this approach latter.



Now we have our factors we need to find a way of interpreting them – to enable this we carry out a process called factor rotation.

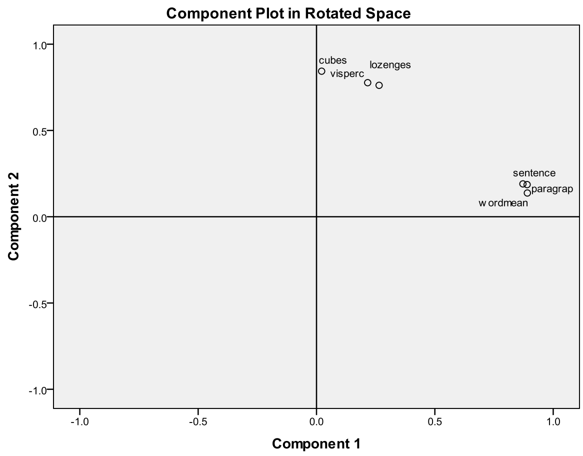
3.4 Giving the factors meaning

Norman & Streiner provide an excellent discussion as to the reasons for rotation in giving factors meaning. To select a rotation method in SPSS you select the Rotation button in the factor analysis dialog box. We will consider two types Varimax and Promax. First Varimax:

Varimax rotation from the PCA extraction method

Rotated Component Matrix[a]		
	Component	
	1	2
visperc	.216	.777
cubes	.022	.843
lozenges	.264	.762
paragraph	.890	.186
sentence	.872	.190
wordmean	.891	.137

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

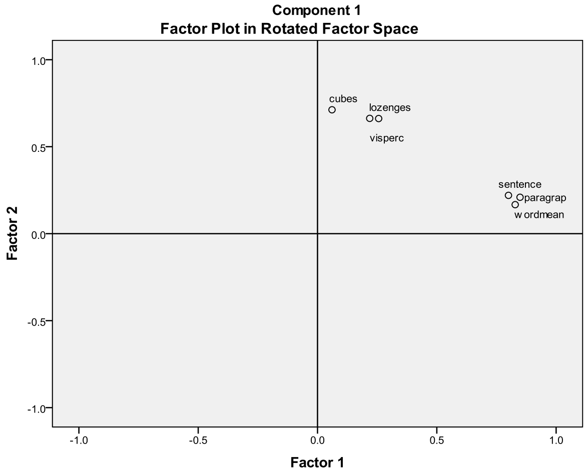


Varimax rotation from the PA extraction method

Rotated Factor Matrix[a]		
	Factor	
	1	2
visperc	.219	.663
cubes	.061	.712
lozenges	.256	.662
paragraph	.849	.209
sentence	.800	.220
wordmean	.829	.167

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.



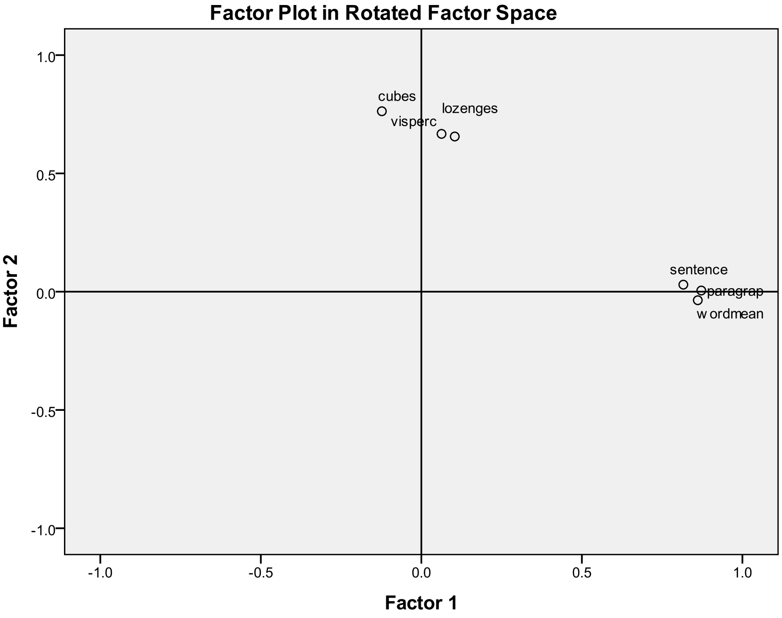
We can see from both of the above set of results that they are pretty similar. Paragraph, sentence and Wordmean load heavily on the first factor/component and the other three on the second factor/component.

By selecting the Varimax rotation option I have demanded that the factors are uncorrelated (technically orthogonal). However, this might not be the case and we can use a rotation that allows for correlated factors and such a one is Promax.

PA extraction with Promax rotation

Structure Matrix		
	Factor	
	1	2
visperc	.368	.696
cubes	.226	.706
lozenges	.404	.704
paragraph	.874	.404
sentence	.830	.403
wordmean	.845	.358

Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.



Factor Correlation Matrix			
Factor		1	2
dimension0	1	1.000	.458
	2	.458	1.000

Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.

So by allowing the two latent variables to correlate which has resulted in a correlation of 0.458 the factor loading have changed little.

The next thing we do is to disregard those loading below a certain threshold on each factor often this is something like 0.3 or 0.4 but Norman and Streiner suggest a significant test (page 205) but for now I'll use the quick and dirty approach. Looking at the above I have highlighted the high loadings for each factor and we can see immediately it makes sense, that is they seem to appear logical.

Possibly you might be asking yourself why we spent to all this time and effort when we have come to pretty much the same conclusion that we had when we eyeballed the correlation matrix at the start of the procedure, and some people agree. However factor analysis does often offer more than can be achieved by merely eyeballing a set of correlations along with some level of statistical rigor (although statisticians argue this point).

Exercise 3.

Made some suggestions for the names of the two latent variables (factors) identified.

3.5 Reification

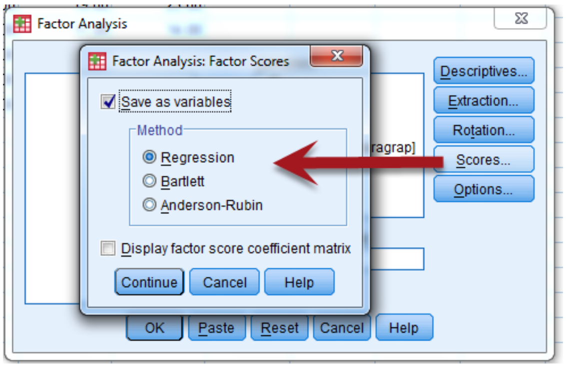
Although the computer presents us with what appears a lovely organised se of variables that make or a factor there is no reason that this subset of variable should equate to something in reality. This is also called the fallacy of misplaced concreteness. Basically it is assuming something exists because it appears so for example a Latent variable.

Exercise 4.

Do a Google search on reification – a good place to start is the Wikipedia article.

3.6 Obtaining factor scores for individuals

We can obtain the factor scores for each individual (case) and then compare them. In SPSS we select the Score button from the factor analysis options dialog box as shown below.



The result is that two additional columns are added to the dataset each representing the factor score for each factor for each individuals standardised scores:

	visperc	cubes	lozenges	paragrap	sentence	wordmean	FAC1_1	FAC2_1
1	33.00	22.00	17.00	8.00	17.00	10.00	-.70053	-.04886
2	30.00	25.00	20.00	10.00	23.00	18.00	.17741	.30324
3	36.00	33.00	36.00	17.00	25.00	41.00	2.07901	2.05252
4	28.00	25.00	9.00	10.00	18.00	11.00	-.45305	-.30760
5	30.00	25.00	11.00	11.00	21.00	8.00	-.26468	-.08457

The above shows the estimated factor scores for the FA analysis and opposite for the PCA analysis.

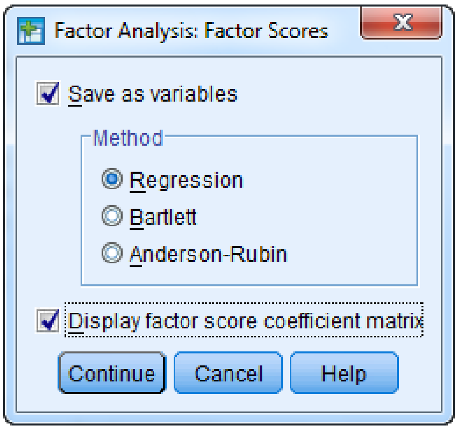
FAC1_1	FAC2_1
-.49029	.63343
.32003	.13068
2.56909	.43067
-.51469	.10434
-.26285	.16235
-.60119	-.71579
-1.89247	-.15043
[illegible]	1.27513
[illegible]	.22118
[illegible]	-.12900

from the PCA

How are the above factor scores for each case calculated? The answer is that an equation is used where the dependent variable is the predicted factor score and the independent variables are the observed variables. We can check this but to do this we need two more pieces of information the factor score coefficient matrix and the standardised scores for the observed variables.

3.6.1 Obtaining the factor score coefficient matrix

You obtain the factor score coefficient matrix by checking the Display factor score coefficient matrix option in the factor scores dialog box.



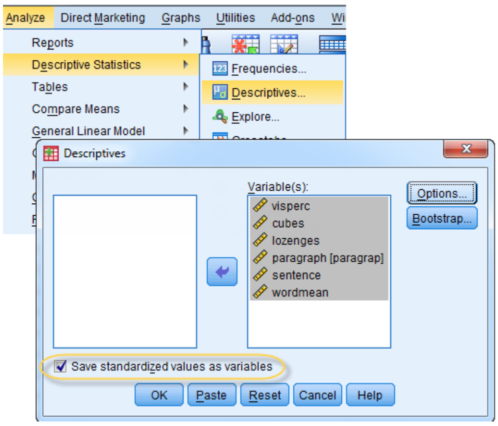
Component Score Coefficient Matrix

	Component 1	Component 2
visperc	.207	.363
cubes	.170	.489
lozenges	.216	.332
paragraph	.265	-.288
sentence	.262	-.277
wordmean	.256	-.317

Extraction Method: Principal Component Analysis. Component Scores.

3.6.2 Obtaining standardised scores

Standardised scores can easily obtained in SPSS using the Analyse -> descriptive statistics menu option.



Zvisperc	Zcubes	Zlozenges	Zparagrap	Zsentence	Zwordmean
.53282	-.59535	.27359	-.72679	-.45532	-.96328
.09904	.06649	.65281	-.16535	.73177	-.00165
.96660	1.83138	2.67532	1.79968	1.12746	2.76306
-.19015	.06649	-.73766	-.16535	-.25747	-.84308
.09904	.06649	-.48485	.11536	.33607	-1.20369
-1.34690	.06649	-1.11688	-.44607	.33607	-.24206

3.6.3 The equation

For a Principle components analysis you can check the individual factor score values produced by SPSS values by plugging the standardised variable scores for the individual into the equation below, however this does not work for the other types of factor extraction as we have lost some of the variance in the extraction process, you can't go back and in these cases the factor scores produced by SPSS are estimates rather than exact values.

So returning back to the PCA factors. As we have two factors we have two factor equations:

Using the values from the component/factor score coefficient matrix:

FS_1 = (0.207)visperc + (0.170)cubes + (0.216)lozenges + (0.265)paragraph + (0.262)sentence + (0.256)wordmean

FS_2 = (0.363)visperc + (0.489)cubes + (0.332)lozenges + (-0.288)paragraph + (-0.277)sentence + (-0.317)wordmean

Now considering the first case that is the first row in the SPSS datasheet, we can also plug in their standardised scores:

$FS_{1subject1}$ = (0. 207)0.53282 + (0. 170)(-0.59535) + (0. 216)0.27359 + (0. 265)(-0.72679) + (0. 262)(-0.45532) + (0. 256)(-0.96328)

In R

```
Answer <- (0.207)*0.53282 + (0.170)*(-0.59535) + (0.216)*0.27359  + (0.265)*(-0.72679) + (0.262)*(-0.45532) + (0.256)*(-0.96328)
```

= -0.4903132

Which is the same as the answer produced by SPSS, shown again opposite.

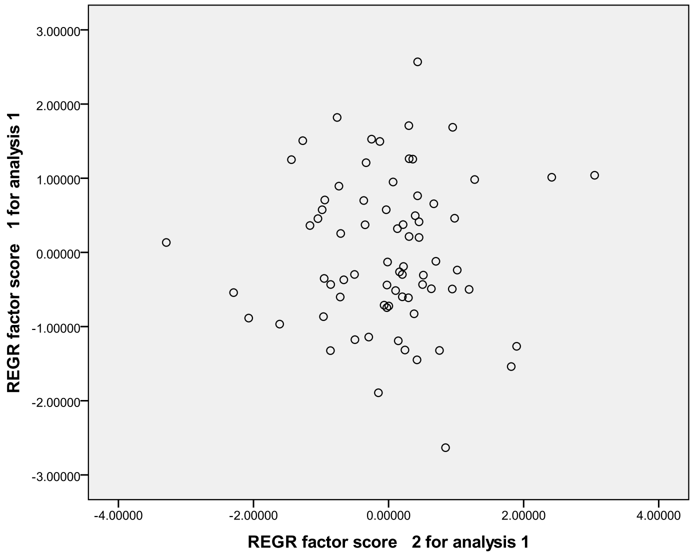
FAC1_1	FAC2_1
-.49029	.63343
.32003	.13068
2.56909	.43067
-.51469	.10434
-.26285	.16235
-.60119	-.71579
-1.89247	-.15043
[illegible]	1.27513
[illegible]	.22118
[illegible]	-.12900

from the PCA

Obviously you do not need to go through this process of checking how SPSS produced the individual factor score values I just did it to show you how they are produced.

3.7 What do the individual factor scores tell us?

What do these factor scores tell us about them, well as the first factor is concerned with reading/writing and the second one is concerned with visual comprehension, we can see how the individual has scored for each of these two latent variables.



It is of interest to carry out some basic descriptive statistics on these new variables. Opposite is a simple scatterplot of the factor score from the PCA. While the degree of correlation is as expected we can see that the values range from around -3 to 3 for the first factor score (reading/writing ability) and around -4 to 4 for the second factor (visual comprehension).

We can also see that the mean for each is zero and the standard deviation is 1 in other words they are standardised variables.

Descriptive Statistics

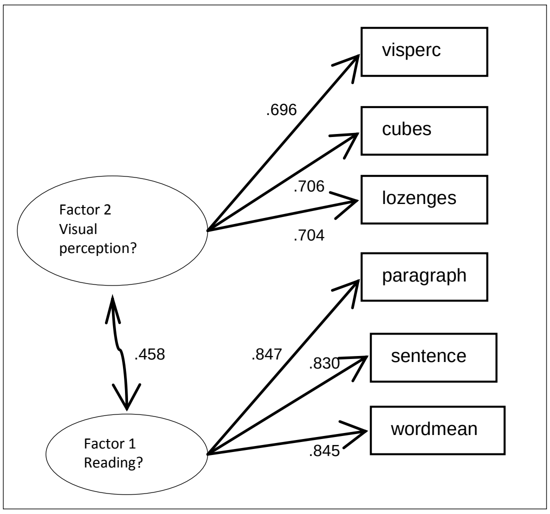
	N	Minimum	Maximum	Mean	Std. Deviation
REGR factor score 1 for analysis 1	73	-2.63	2.56	.00	1.00
REGR factor score 2 for analysis 1	73	-3.29	3.05	.00	1.00
Valid N (listwise)	73				

Exercise 5.

1. Why would we expect the factors not to be correlated in the above scatterplot?
2. Would you expect the scatterplot to always show uncorrelated variables for any type of factor extraction/rotation strategy? Hint: you may need to run the analysis several times with different extraction/rotation methods to see what you get to confirm your suspicions.
3. Given that the factors are 'standardised' scores assuming that they are also normally distributed what would a score of around -2 for the first factor suggest (refer back to the first statistics course concerning Z values if necessary)

4 Summary - to Factor analyse or not

We can use the information from the analysis along with the diagramming technique we introduced earlier to summarize our results in the diagram below. Notice that I have left out the lines where the loadings were below 0.69 and I have used the results from the PA extraction with Promax rotation, showing the correlation between the two factors I could also have put the uniqueness values in (see page 3) but CPA does not take these into account compared to factor analysis. We followed a clearly defined set of stages:



1. Data preparation (most of it was already been done in this example)
2. Observed correlation matrix inspection
3. Statistics to assess suitability of dataset for basis of PCA (KMO, Bartlett's and determinant measures)
4. Factor extraction - PCA
5. Factor rotation – to allow interpretation
6. Factor name attribution
7. Factor score interpretation

We have barely scratched the surface in this short introduction and there has always a hot debate at to the benefits of CPA and factor analysis. This is easily seen as analysis of each stage I is not described in the previous pages shows there are many ways of interpreting the results at that stage and also a multitude of ways of carrying out the next stage, not only this but different authorities suggest that the analysis stops at different points in the analysis and also different authors give different interpretations to the various results. Quoting Everitt & Dunn 2001 page 288:

Hills (1977) has gone as far as to suggest that factor analysis is not worth the time necessary to understand it and carry it out. And Chatfield and Collins (1980) recommend that factor analysis should not be used in most practical situations. Such criticisms go too far. Factor analysis is simply an additional, and at times very useful, tool for investigating particular features of the structure of multivariate observations. Of course, like many models used in analysing data, the one used in factor analysis is likely to be only a very idealized approximation to the truth in the situations in which it is generally applied. Such an approximation may, however, prove a valuable starting point for further investigations.

Hills M 1977 Book review. Applied statistics 26 339-340

Chatfield C, Collins A J 1980 Introduction to Multivariate Analysis. Chapman & Hall. London.

Loehlin 2004 p.230-6 provides an excellent in depth criticism of Latent Variable modelling (of which factor analysis is one example). In contrast to these criticisms a more positive approach can be found in many books about factor analysis for example chapter 7 entitled factor analysis in Bartholomew, Steele, Moustaki & Galbraith 2008 also the conference proceedings held in 2004 entitled 100 years of factor analysis are available at http://www.fa100.info/

Factor analysis forms the basis of a more complex technique called Structural Equation modelling (SEM) and all we have done here, and much more, can be achieved using SEM. SEM provides much more sophistication than the traditional exploratory factor analysis, although a traditional EFA often is the first step to a full SEM analysis notably we can compare models and also analyse the overall fit of a model this will be discussed in a chapter re-analysing this data using a SEM framework.

The next section provides a worked example of a typical PCA/factor analysis exam question. I have also provided two practical sections one describing how to carry out a PCA/factor analysis using a correlation matrix as the basis rather than raw data and also how to carry out the equivalent analysis in R.

5 A typical exam question

Kinnear and Gray (2004, page 429) provide the following example which is suitable for Principal Component Analysis (though the sample size is completely inadequate):

Ten participants are given a battery of personality tests, comprising the following items: Anxiety; Agoraphobia; Arachnophobia; Adventure; Extraversion; and Sociability (with a scoring range of 0 to 100). The purpose of this project is to ascertain whether the correlations among the six variables can be accounted for in terms of comparatively few latent variables or factors.

Part	Anx	Agora	Arach	Adven	Extra	Socia
1	71	68	80	44	54	52
2	39	30	41	77	90	80
3	46	55	45	50	46	48
4	33	33	39	57	64	62
5	74	75	90	45	55	48
6	39	47	48	91	87	91
7	66	70	69	54	44	48
8	33	40	36	31	37	36
9	85	75	93	45	50	42
10	45	35	44	70	66	78

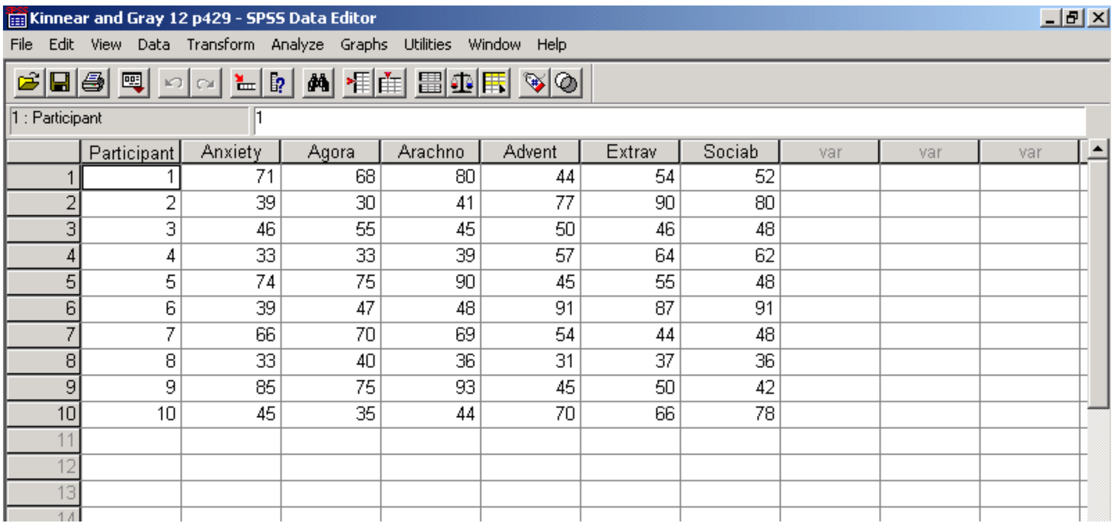
In this section I will provide an answer to a typical exam question based on this data.

The exam question

Conduct a principal component analysis to determine how many important components are present in the data. To what extent are the important components able to explain the observed correlations between the variables? Rotate the components in order to make their interpretation more understandable in terms of a specific theory. Which tests have high loadings on each of the rotated components? Try to identify and name the rotated components.

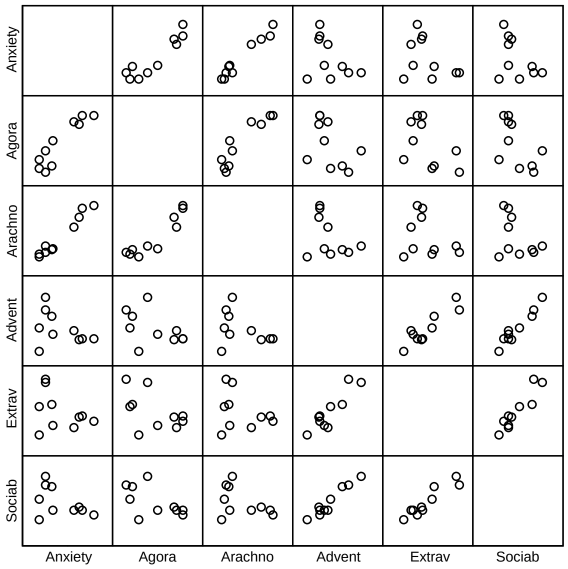
5.1 Data layout and initial inspection

The data are put into appropriately named SPSS variable columns:



	Participant	Anxiety	Agora	Arachno	Advent	Extrav	Sociab	var	var	var
1	1	71	68	80	44	54	52			
2	2	39	30	41	77	90	80			
3	3	46	55	45	50	46	48			
4	4	33	33	39	57	64	62			
5	5	74	75	90	45	55	48			
6	6	39	47	48	91	87	91			
7	7	66	70	69	54	44	48			
8	8	33	40	36	31	37	36			
9	9	85	75	93	45	50	42			
10	10	45	35	44	70	66	78			
11										
12										
13										

It is possible, as we have seen before, to look at the scatterplots of all the variables with one another, as I did before we are looking for significant correlations, and possibly clusters of them. Also we want to check that there are no perfectly correlated variables (which would need removing). The following output was generated by the Graphs, Scatterplot, Matrix command.



We can also produce a correlation matrix verifying our suspicions from the scatterplot. Most of the correlations are well above 0.3 (a good indication that we will obtain a result) and there appears to be two groups of variables –highlighted in yellow below. Anxiety, Agoraphobia, and Arachnophobia in one, and Adventure, Extraversion, and Sociability in the other

Correlations

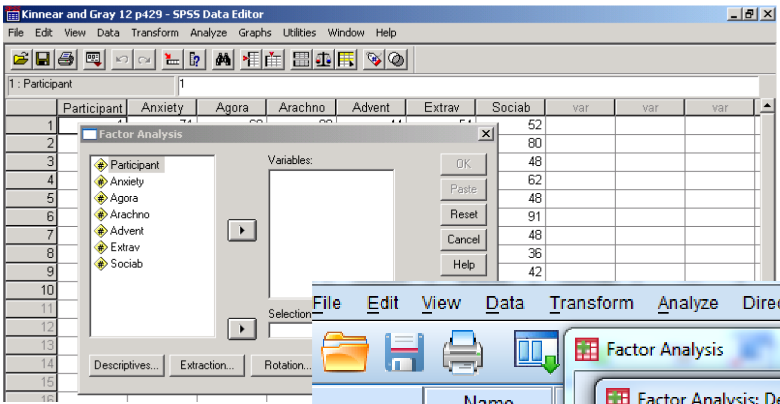
		anx	agora	arach	adven	extra	social
anx	Pearson Correlation	1					
agora	Pearson Correlation	.921**	1				
arach	Pearson Correlation	.979**	.921**	1			
adven	Pearson Correlation	-.389	-.461	-.366	1		
extra	Pearson Correlation	-.365	-.508	-.301	.905**	1	**
social	Pearson Correlation	-.462	-.569	-.425	.967**	.934**	1

**. Correlation is significant at the 0.01 level (2-tailed).

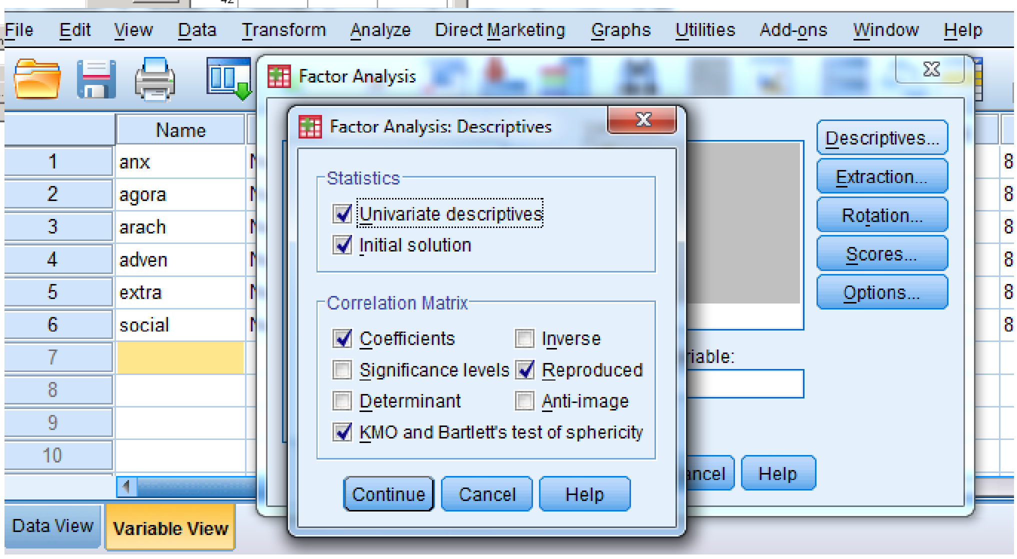
So we'll go ahead with the Principal Component Analysis.

5.2 *Carrying out the Principal Component Analysis*

Click on Analyze, Dimension Reduction, Factor, to open the Factor Analysis dialogue box:

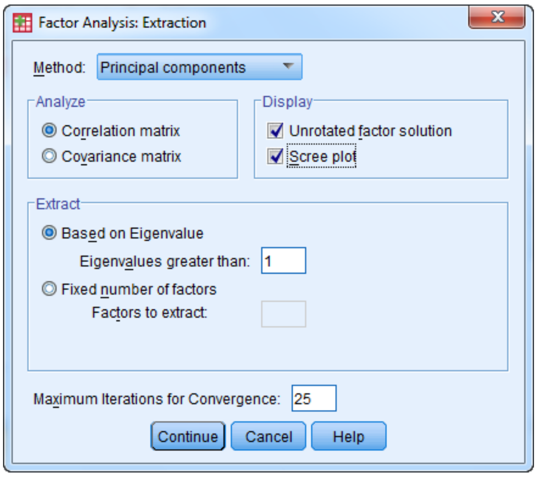


Move the six variables over to the Variables: box. Click on Descriptives… and select Univariate Descriptives, Coefficients, and Reproduced:

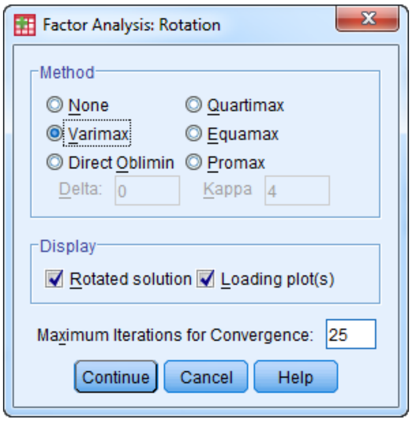


Click on Continue, and then on Extraction where you should select Scree Plot, after making sure that the method chosen is Principal Components, that the analysis is to be carried out on the Correlation matrix[1], that we want the un-rotated factor solution to be displayed, and that we want factors with eigenvalues over 1 to be extracted:

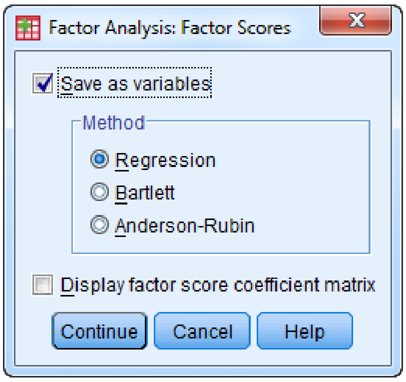
[1] If Covariance matrix is selected, more weight is given to variables with higher standard deviation. With Correlation matrix, all the variables are given equal weight (by standardising them).



Click on Continue and then on Rotation where you should select Varimax rotation and Loading plots:



Click on Continue and then on Scores to select which type of factor scores you want to save in the data set, select regression:



Click on Continue and the on OK (the Options subcommand isn't relevant here). The output is as follows:

Exercise 6.

Add some notes below about some of the various options in the dialogue boxes shown above.

5.3 Interpreting the output

5.4 Descriptive Statistics

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Anxiety	53.10	19.041	10
Agora	52.80	18.085	10
Arachno	58.50	22.237	10
Adv ent	56.40	17.989	10
Extrav	59.30	17.695	10
Sociab	58.50	18.447	10

The table opposite simply shows the means, standard deviations and sample size for each variable. It appears that the average score for all the tests is very similar and all have a similar spread.

Next is the observed correlation matrix, which we have already commented on.

Correlation Matrix

		Anxiety	Agora	Arachno	Adv ent	Extrav	Sociab
Correlation	Anxiety	1.000	.921	.979	-.389	-.365	-.462
	Agora	.921	1.000	.921	-.461	-.508	-.569
	Arachno	.979	.921	1.000	-.366	-.301	-.425
	Adv ent	-.389	-.461	-.366	1.000	.905	.967
	Extrav	-.365	-.508	-.301	.905	1.000	.934
	Sociab	-.462	-.569	-.425	.967	.934	1.000

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.550
Bartlett's Test of Sphericity	Approx. Chi-Square	73.582
	df	15
	Sig.	.000

The KMO value indicates that we have is pretty poor – just above miserable, however Bartlett's test of sphericity with an associated p value of <0.001 indicates that we can proceed.

5.5 Communalities

Next is a table of estimated communalities (i.e. estimates of that part of the variability in each variable that is shared with others, and which is not due to measurement error or latent variable influence on the observed variable). The initial values can be ignored.

Communalities

	Initial	Extraction
Anxiety	1.000	.976
Agora	1.000	.942
Arachno	1.000	.982
Adv ent	1.000	.954
Extrav	1.000	.942
Sociab	1.000	.980

Extraction Method: Principal Component Analysis.

5.6 Eigenvalues and Scree Plot

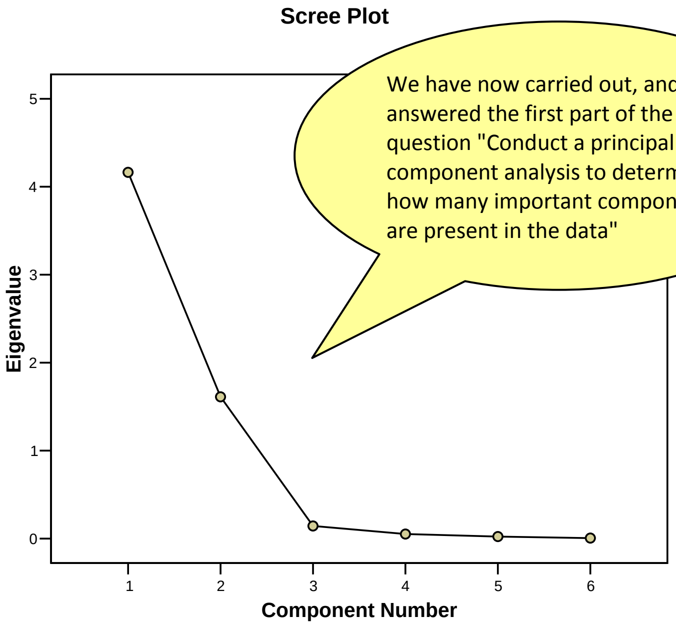
Next comes a table showing the importance of each of the six principal components. Only the first two have eigenvalues over 1.00, and together these explain over 96% of the total variability in the data. This leads us to the conclusion that a two factor solution will probably be adequate.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.164	69.397	69.397	4.164	69.397	69.397	2.895	48.251	48.251
2	1.612	26.862	96.259	1.612	26.862	96.259	2.881	48.008	96.259
3	.144	2.396	98.655						
4	.052	.867	99.522						
5	.023	.383	99.905						
6	.006	.095	100.000						

Extraction Method: Principal Component Analysis.

This conclusion is supported by the scree splot (which is actually simply displaying the same data visually):



We have now carried out, and answered the first part of the question "Conduct a principal component analysis to determine how many important components are present in the data"

5.7 Unrotated factor loadings

The unrotated factor loadings are presented next. These show the expected pattern, with high positive and high negative loadings on the first factor:

Component Matrix[a]

	Component	
	1	2
Anxiety	.824	.545
Agora	.878	.415
Arachno	.799	.586
Advent	-.818	.533
Extrav	-.804	.544
Sociab	-.873	.467

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

The next table shows the extent to which the original correlation matrix can be reproduced from two factors:

Reproduced Correlations

		Anxiety	Agora	Arachno	Advent	Extrav	Sociab
Reproduced Correlation	Anxiety	.976[b]	.949	.978	-.383	-.365	-.464
	Agora	.949	.942[b]	.944	-.497	-.479	-.572
	Arachno	.978	.944	.982[b]	-.341	-.323	-.423
	Advent	-.383	-.497	-.341	.954[b]	.948	.963
	Extrav	-.365	-.479	-.323	.948	.942[b]	.956
	Sociab	-.464	-.572	-.423	.963	.956	.980[b]
Residual[a]	Anxiety		-.028	.002	-.006	.000	.002
	Agora	-.028		-.023	.036	-.028	.003
	Arachno	.002	-.023		-.025	.022	-.002
	Advent	-.006	.036	-.025		-.042	.003
	Extrav	.000	-.028	.022	-.042		-.022
	Sociab	.002	.003	-.002	.003	-.022	

Extraction Method: Principal Component Analysis.

a. Residuals are computed between observed and reproduced correlations. There are 0 (.0%) nonredundant residuals with absolute values greater than 0.05.

b. Reproduced communalities

The small residuals show that there is very little difference between the reproduced correlations and the correlations actually observed between the variables. The two factor solution provides a very accurate summary of the relationships in the data.

We have now carried out, and answered the second part of the question "To what extent are the important components able to explain the observed correlations between the variables?"

5.8 Rotation

The next table shows the factor loadings that result from Varimax rotation:

Rotated Component Matrix[a]

	Component	
	1	2
Anxiety	-.200	.967
Agora	-.330	.913
Arachno	-.154	.979
Adv ent	.956	-.199
Extrav	.953	-.181
Sociab	.948	-.284

Extraction Method: Principal Component Analy sis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation conv erged in 3 iterations.

We have now carried out, and answered the third part of the question "Which tests have high loadings on each of the rotated components?"

These two rotated factors are just as good as the initial factors in explaining and reproducing the observed correlation matrix (see the table below). In the rotated factors, Adventure, Extraversion and Sociability all have high positive loadings on the first factor (and low loadings on the second), whereas Anxiety, Agoraphobia, and Arachnophobia all have high positive loadings on the second factor (and low loadings on the first).

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulativ e %	Total	% of Variance	Cumulativ e %	Total	% of Variance	Cumulativ e %
1	4.164	69.397	69.397	4.164	69.397	69.397	2.895	48.251	48.251
2	1.612	26.862	96.259	1.612	26.862	96.259	2.881	48.008	96.259
3	.144	2.396	98.655						
4	.052	.867	99.522						
5	.023	.383	99.905						
6	.006	.095	100.000						

Extraction Method: Principal Component Analy sis.

Same overall % but very different division (⇕)

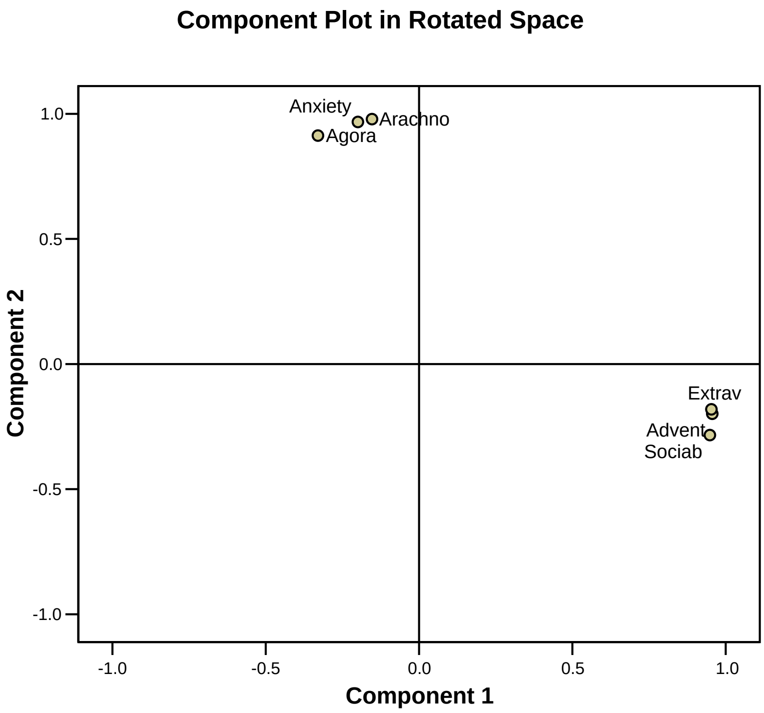
Above, is the table showing the eigenvalues and percentage of variance explained again. The middle part of the table shows the eigenvalues and percentage of variance explained for just the two factors of the initial solution that are regarded as important. Clearly the first factor of the initial solution is much more important than the second. However, in the right hand part of the table, the eigenvalues and percentage of variance explained for the two rotated factors are displayed. Whilst, taken together, the two rotated factors explain just the same amount of variance as the two factors of the initial solution, the division of importance between the two rotated factors is very different. The effect of rotation is to spread the importance more or less equally between the two rotated factors. You will note in the above table that the eigenvalues of the rotated factor are 2.895 and 2.881, compared to 4.164 and 1.612 in the initial solution. I hope that this makes it clear how important it is that you extract an appropriate number of factors. If you extract more than are needed, then rotation will ensure that the variability explained is more or less evenly distributed between them. If the data are really the product of just two factors, but you extract and rotate three, the resulting solution is not likely to be very informative.

The next table gives information about the extent to which the factors have been rotated. In this case, the factors have been rotated through 45 degrees. (The angle can be calculated by treating the correlation coefficient as a cosine. The cosine of 45 degrees is .707.)

Component Transformation Matrix

Component	1	2
1	-.709	.705
2	.705	.709

Extraction Method: Principal Component Analy sis.
Rotation Method: Varimax with Kaiser Normalization.

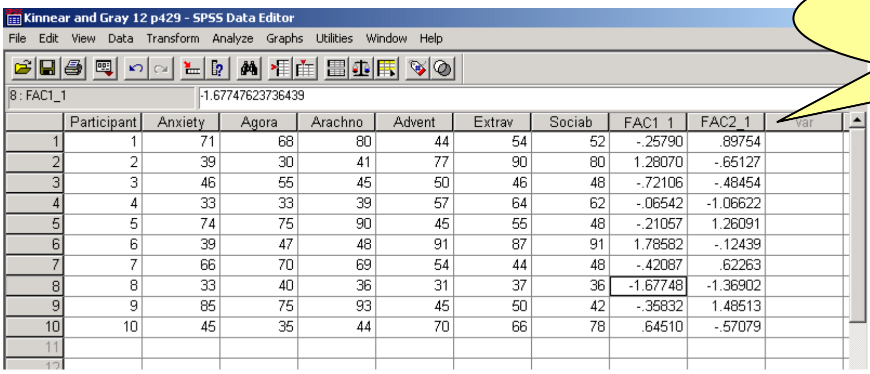


5.9 Naming the factors

SPSS now produces a decent plot of the six variables on axes representing the two rotated factors:

It seems reasonable to tentatively identify the first rotated factor as “Outgoingness”, as Extraversion, Adventure, and Sociability all have high loadings on it. The second rotated factors looks rather like “Neuroticism”, as Anxiety and the two phobias all have high loadings on it.

The Saved Factor scores have been added to the data, as you will see overleaf. These are standardized scores, obtained by applying the rotated factor loadings to the standardized score of each participant on each of the variables (just like making a prediction using a regression equation). Participant 8 has a low standardized score on the first rotated factor (-1.68) and can therefore be said to be low in “Outgoingness”. The same participant also has a low standardized score on the second rotated factor (-1.37) and can therefore be said to be low in “Neuroticism”. Participant 6, on the other hand, scores high (1.79) on “Outgoingness”, but has a score close to average (-.12) on “Neuroticsm”.


Kinnear and Gray 12 p429 - SPSS Data Editor

8 : FAC1_1 -1.67747623736439

	Participant	Anxiety	Agora	Arachno	Advent	Extrav	Sociab	FAC1_1	FAC2_1	var
1	1	71	68	80	44	54	52	-.25790	.89754	
2	2	39	30	41	77	90	80	1.28070	-.65127	
3	3	46	55	45	50	46	48	-.72106	-.48454	
4	4	33	33	39	57	64	62	-.06542	-1.06622	
5	5	74	75	90	45	55	48	-.21057	1.26091	
6	6	39	47	48	91	87	91	1.78582	-.12439	
7	7	66	70	69	54	44	48	-.42087	.62263	
8	8	33	40	36	31	37	36	-1.67748	-1.36902	
9	9	85	75	93	45	50	42	-.35832	1.48513	
10	10	45	35	44	70	66	78	.64510	-.57079	
11										

We have now carried out, and answered the fourth and final part of the question "Try to identify and name the rotated components"

5.10 Summary

In answering the question requiring us to conduct a principal component analysis we went through a series of clearly defined stages:

1. Data preparation (most of it was already been done in this example)
2. Observed correlation matrix inspection
3. Statistics to assess suitability of dataset for basis of PCA (KMO, Bartlett’s and determinant measures)
4. Factor extraction - PCA
5. Factor rotation – to allow interpretation
6. Factor name attribution
7. Factor score interpretation

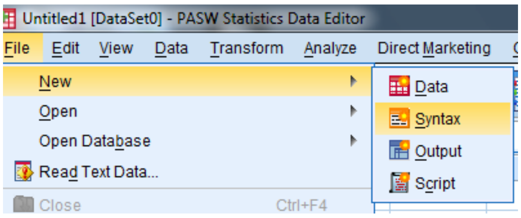
Exercise 7.

For each of the above stages add a sentence stating its purpose along with another giving the finding(s) from the analysis above.

---------------------------------- end of exam answer ---

6 PCA and factor Analysis with a set of correlations or covariances in SPSS

Often we wish to carry out a PCA or factor analysis in SPSS when we do not have the raw data but a set of correlations or covariances. You can achieve this using SPSS syntax, to do this you first need to open a new syntax window in SPSS assuming we are repeating the previous analysis we just type in now:



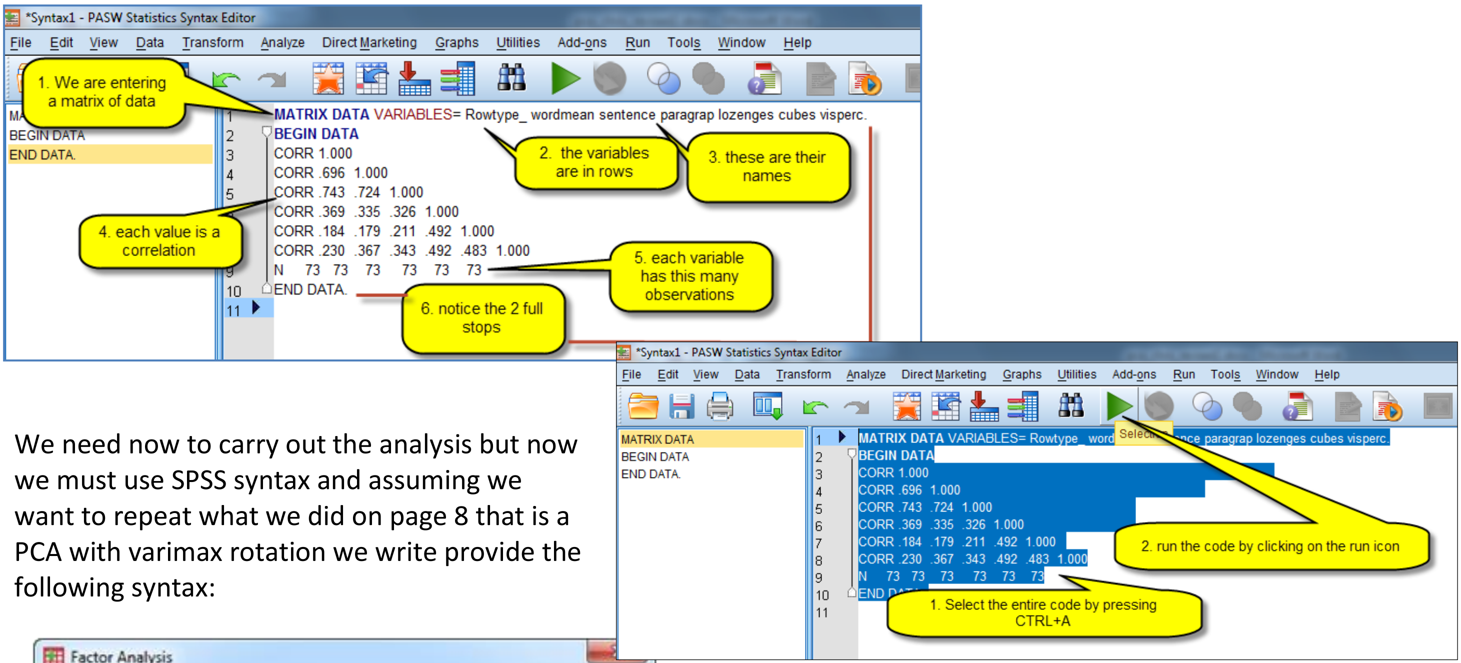

```
MATRIX DATA VARIABLES = Rowtype_ wordmean sentence paragrap lozenges cubes visperc.
BEGIN DATA
CORR 1.000
CORR .696  1.000
CORR .743  .724  1.000
CORR .369  .335  .326  1.000
CORR .184  .179  .211  .492  1.000
CORR .230  .367  .343  .492  .483  1.000
N    73  73   73    73   73   73
END DATA.
```

The first thing to notice is the set of correlations which you will find on page 4.

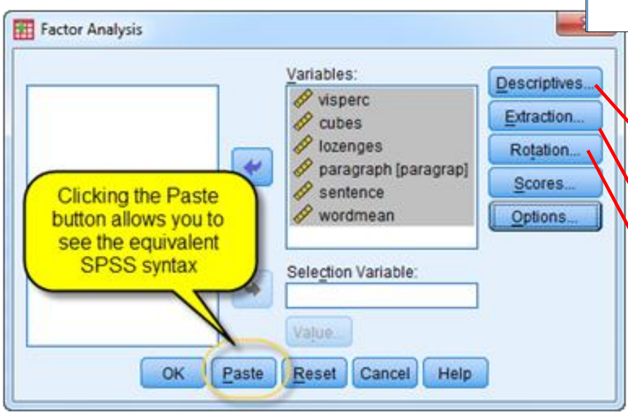
A quick explanation is given below:

Warning once you have used SPSS syntax to define the data as a correlation matrix you cannot use the SPSS dialog boxes to carry out any subsequent analysis. If you do you just get rubbish out.

To actually run the SPSS syntax you need to highlight the code you wish to run and then click on the run button (below right).



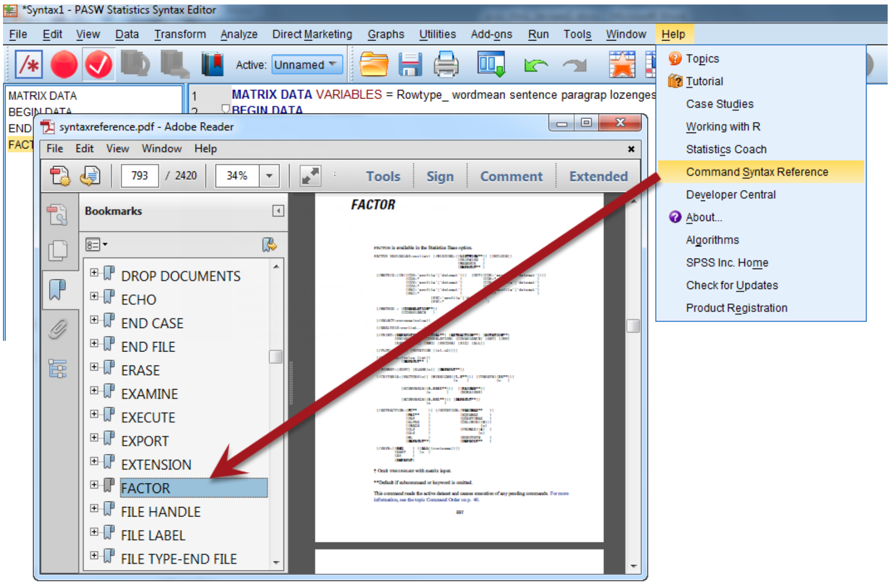
We need now to carry out the analysis but now we must use SPSS syntax and assuming we want to repeat what we did on page 8 that is a PCA with varimax rotation we write provide the following syntax:




```
FACTOR
/MATRIX = IN (CORR=*)
/PRINT KMO EXTRACTION ROTATION
 /PLOT EIGEN ROTATION
 /CRITERIA MINEIGEN(1) ITERATE(25)
 /EXTRACTION PC
 /CRITERIA ITERATE(25)
 /ROTATION VARIMAX.
```

You can always use the dialog boxes to get a rough idea of what the syntax should look like, by clicking on the Paste button, before trying to write it yourself. For Example the syntax above for the Factor analysis was largely generated from the factor analysis dialog boxes except the line /MATRIX= IN (CORR=*) which instructs SPSS to use the correlation matrix as the data which we previously defined also using SPSS syntax.

You can get help about the SPSS syntax various ways but I personally prefer looking up the entry in the Command Syntax Guide which is a pdf file accessed from the help menu - don't try printing this out as the FACTOR entry along is nearly 20 pages long, providing numerous examples, and often it is just a case of adapting one of them to your own particular needs.



7 PCA and factor analysis in R

There is a wealth of information about using R to carry out PCA and factor analysis in R. For PCA there is an excellent youtube video given by Edward Boone who is associate professor at Virginia Commonwealth University at: http://youtu.be/Heh7Nv4qimU you can see his personal page at: www.people.vcu.edu/~elboone2/

There are many ways of carrying out factor analysis in R and the R site, quick-R (http://www.statmethods.net/), provides not only general advice about R but also detailed information about carrying out various types of factor analysis with links to sources of additional information all of which can be found at http://www.statmethods.net/advstats/factor.html

The R psych package, mentioned above, aids factor analysis and the developer of the package maintains an excellent online book including a very detailed chapter on factor analysis (http://personality-project.org/r/book/Chapter6.pdf). Also the factor analysis chapter in Andy fields Discovering Statistics using R (2012) makes use of the package to carry out a PCA analysis.

R offers many more options than SPSS for both PCA and factor analysis. One very interesting option (in the psych package) is the ability to create "Parallel Analysis Scree plots". This is where R produces a random data matrix besides the dataset you are working with and then plots the Eigen values from both on a scree plot allowing you to assess the difference between what your dataset has produced against a random dataset. For more details see the fa.parallel() entry in the psych package manual.

The R code on the next page repeats most of the analysis carried above previously in SPSS, and I have added numerous comments to aid understanding. Notice I have used various procedures from the psych package:

```
install.packages("psych", dependencies=TRUE)
library(psych)
hozdata <-read.delim(file=file.choose()) #the required file is available to download named grnt_fem.dat
# put the data into a matrix of correlations
hozdatamatrix <- cor(hozdata)
# print out the correlation matrix but ask for numbers to 4 decimal places
round(hozdatamatrix,4)
# bartlett test - want a small p value here to indicate c0rrelation matrix not zeros
cortest.bartlett(hozdata)
# unable to calculate the kmo - see field 2012 p776
# but can do the determinant need it to be above 0.00001
# to be able to continue
det(hozdatamatrix)
# appropriate value therefore can continue
#  do a pca analysis use the principal function in the psych package
model1<- principal(hozdata, nfactors = 6, rotate = "none")
model1
# get the scree plot
plot(model1$values, type = "b")
# now know how many components we want to extract = 2
# rerun the anylsis specifying this
model2 <- principal(hozdata, nfactors = 2, rotate = "none")
model2
# can find the reproduced correlations and the communalities (the diagonals)
factor.model(model2$loadings)
# can also find the differences between the observed and model estimated correlations
# the diagonals represent the uniqueness values (1- R squared):
residuals <- factor.residuals(hozdatamatrix, model2$loadings)
residuals
# nice to plot the residuals to check there are normally distributed
hist(residuals)
# now to the rotation
model3 <- principal(hozdata, nfactors = 2, rotate = "varimax")
model3
# can get the loading matrix to stop printing out loading below
# a specific value say 0.3 cna also get it sorted by size of loading
# h2 is the communality; u2 is the uniqueness
print.psych(model3, cut = 0.3 , sort = TRUE)
# now to do a principal axis factor analysis
# fa means factoring method; rotate options=none/varimax/blimin/promax etc.
model4 <- fa(hozdata, nfactors = 2,fm = "pa", rotate = "none")
model4
# repeat the analysis with a varimax rotation
model5 <-  fa(hozdata, nfactors = 2,fm = "pa", rotate = "varimax")
model5
# repeat the analysis with a promax rotation (correlated factors)
model6 <-  fa(hozdata, nfactors = 2,fm = "pa",rotate="promax")
model6
# the factor loadings in the above are not the same as that in SPSS
# this is because SPSS scales the values using something called Kaiser normalisation
# the psych package provides a function to do this
# best to input the non rotated form into the function (info. from help file)
model7 <- kaiser(model4, rotate="promax")
model7
#the above output is slight more like that of SPSS
# to obtain factor scores
# the for PCA we just add scores = true
model3a <- principal(hozdata, nfactors = 2, rotate = "varimax", scores = TRUE)
model3a
# to print out all the scores:
model3a$scores
# to print out just the top 10 scores:
head(model3a$scores, 10)
# to save the above values we need to add them to a dataframe
factorscores <- cbind(model3a$scores)
# then we can produce a plot of the scores:
 plot(factorscores)
```

If you run the above script you will notice that there are many more statistics than that produced by SPSS this is because R takes a different approach to the factoring modelling process, considering of primary importance how well the model (i.e. the model correlations) fits the observed correlations. This is in the spirit of the SEM approach which I will discuss in another chapter.

7.1 Using a matrix instead of raw data

As in SPSS you can either provide raw data or a matrix of correlations as input to the CPA/factor analysis. The R code below provides an equivalent analysis to that described above but using a correlation matrix as input.

```
library(psych)
hozdatamatrix <-  matrix(c(
1.000, .696, .743, .369, .184, .230,
.696, 1.000, .724, .335, .179, .367,
.743, .724, 1.000, .326, .211, .343,
.369, .335, .326, 1.000, .492, .492,
.184, .179,.211, .492, 1.000,  .483,
.230, .367, .343, .492, .483, 1.000), ncol = 6, byrow = TRUE)
# now give the columns and rows names
colnames(hozdatamatrix)<- c("wordmean",  "sentence",     "paragrap",     "lozenges",     "cubes", "visperc")
rownames(hozdatamatrix)<- c("wordmean",  "sentence",     "paragrap",     "lozenges",     "cubes", "visperc")
######### you can NOT use the standard cor function i.e. cor(hozdatamatrix)
#  With a correlaion matrix as it produces correlations or the correlations.
# you can produce a correlation plot by using the cor.plot function
# the darker the shading for the cell the highter the correlation
cor.plot(hozdatamatrix)
# According to the psych package manual you should be able to use the function
#below to obtain p values for the associated correlation matrix
# but these values appear very different to those produced by SPSS
# corr.p(r = hozdatamatrix, n= 73)
# In contrast using the determinant function det()
#  gives same answer as using the raw data 0.07374609
det(hozdatamatrix)
# above gives same answer as using using raw data 0.07374609
# To carry out a PCA analysis using a correlation matrix need to
#  tell the principal function how many observations formed the
#  basis oc the correlations specifying a value for the  the n.obs parameter
model1<- principal(hozdatamatrix, nfactors = 6, n.obs = 73,  rotate = "none")
model1
# produces the same output as with the previous analysis with the raw data
# to carry out a factor analysis using a correlation matrix
# adapt the fa function in a similar way:
modelb <- fa(hozdatamatrix, nfactors = 2, fm = "pa", n.obs = 73, rotate = "none")
modelb

# The above shows how easy it is to adapt to either using raw data or the correlation  matric
# in R for PCA and factor analysis in R
```

You can also use the table2matrix() function in the psych package to convert a R table to a matrix. Also in the psych package is various read.clipboard() functions which allow you to copy and paste a matrix of correlations in something like Excel or word and then paste directly into R (see the psych package manual for details.

Optional Exercise 8.

Returning back to the patients' satisfaction with their GP discussed on page 4 here are the correlations for the 14 items discussed on that page.

Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	1.00													
2	0.56	1.00												
3	0.63	0.58	1.00											
4	0.64	0.46	0.35	1.00										
5	0.52	0.44	0.50	0.52	1.00									
6	0.70	0.51	0.49	0.52	0.54	1.00								
7	0.45	0.48	0.28	0.34	0.38	0.63	1.00							
8	0.61	0.68	0.44	0.43	0.56	0.64	0.49	1.00						
9	0.79	0.58	0.66	0.55	0.66	0.64	0.34	0.70	1.00					
10	0.57	0.63	0.40	0.55	0.54	0.58	0.65	0.62	0.62	1.00				
11	0.32	0.27	0.33	0.21	0.13	0.26	0.22	0.24	0.17	0.25	1.00			
12	0.55	0.72	0.51	0.49	0.63	0.62	0.47	0.75	0.70	0.67	0.31	1.00		
13	0.69	0.51	0.60	0.54	0.51	0.73	0.44	0.50	0.66	0.53	0.24	0.65	1.00	
14	0.62	0.42	0.33	0.47	0.38	0.58	0.51	0.49	0.53	0.56	0.23	0.51	0.56	1.00

Quoting Everitt and Dunn 2001 p.283 "The results [Principal factor analysis + varimax rotation] suggests that we should use a three-factor solution. The rotated factors might be labelled 'trust in doctor', 'confidence in doctor's ability' and 'confidence in recommended treatment'.   Carry out an appropriate analysis and demonstrate that this is indeed the case.

You might wish to try using one of the read.clipboard() functions.

8 Summary

This chapter has provided a large amount of information, focusing on the practical aspects of carrying out a PCA or factor analysis in SPSS or R. The first section focused on interpreting the output at each stage and then we considered a typical exam question and finally the matrix input approach using both SPSS and R.

9 Reference

Bartholomew D J, Steele F, Moustaki I, Galbraith J I. 2008 (2nd ed.) Analysis of multivariate Social Science data. CRC press.

Everitt B S, Hothorn T. 2010 (2nd ed.) A handbook of Statistical Analyses using R. CRC Press.

Everitt B S, Hothorn T. 2011 An introduction to applied multivariate analysis with R. Springer.

Everitt B S, Dunn G. 2001 (2nd ed.) Applied Multivariate Data Analysis. Arnold

Field A 2012 Discovering statistics using R.

Kinnear, P.R. and Gray, C.D. (2004) *SPSS 12 Made Simple.* Hove: Psychology Press.

Loehlin J C. 2004 (4th ed.) Latent Variable Models: an introduction to factor, path, and structural equation analysis. Erlbaum.