Principal Component Analysis and Factor Analysis

Probability and Statistical Inference

Bojan Božić

TU Dublin

Winter 2020

Inferential (Predictive) Models

- Used to explore the relationship between one outcome variable and a set of independent variables (predictors).
- You should have a sound theoretical or conceptual reason for exploring the relationship and the order of the variables entering the model.
- For variables that have not been previously shown to contribute to the variation in the outcome variable try to establish statistical evidence for their inclusion in advance.

Inferential Model - What does it allow you to do?

- Prediction:
  - Really what we are looking at is the variance in the outcome variable and how much of the variance could be considered to be explained by the predictor variables.
- Questions it allows you to answer:
  - How well a set of variables is able to predict an outcome variable?
  - Which variable in a set is the best predictor?
  - Whether a variable is still able to predict an outcome when controlling for other variables ?

Choosing variables for an inferential model

- Inferential Model
  - Pseudo Prediction - if we use a linear equation.

$$y = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n + \epsilon_i$$

- b_0 is the intercept.
  - The intercept is the value of the Y variable when all Xs $= 0$.
  - This is the point at which the regression plane crosses the Y-axis (vertical).
- b_1 to b_n are the regression coefficients for variable 1 to n.

Choosing variables for an inferential model

- Inferential Model:
  - Pseudo Prediction - if we use a linear equation.
- You must:
  - Have a sound theoretical basis for including your variables.
  - Explore the data using univariate and bivariate analysis to establish statistical evidence.
  - Only include variables that have potentially informative results or which serve as controls.
- You need to understand not only how your variables relate to your outcome but also to each other.

Dimension Reduction

- The statistical researcher strives to avoid two major obstacles:
  - Using a single score/measure from a multi-item measure in which there is great heterogeneity OR
  - Using several scores/measures that are highly correlated or unreliable.
- The goal is to obtain the "right" number of scores (and it might be one).

Collinearity

- Occurs when two or more independent variables contain strongly redundant information.
  - If variables are collinear they are essentially measuring the same thing, there is not enough distinct information in these variables for inferential statistics to operate.
- If we conduct inferential statistics with collinear variables then the model will produce unreliable results.
- Need to check for collinearity by examining a correlation matrix that compares your independent variables with each other.
- A correlation coefficient above 0.8 suggests collinearity might be present.

Collinearity – Why is it a problem?

- You are introducing bias:
  - Over-representing a concept.
- Increasing possibility of a Type I error.

Collinearity - What can we do?

- Delete all but one of the collinear variables from the model.
  - Which variables are most representative of concept of interest.
  - It will act as a surrogate or proxy for the concept:
  - E.g. number of previous criminal offences as a predictor of reoffending.
  - May weaken the fit of your model.
- Combine them into an index mathematically (e.g. multiplying, adding etc.) .
  - This must make sense theoretically.
  - E.g. Select a range of items:
  - The Retail Price Index is calculated by taking a sample of goods and services a typical household buys and items weighted according to the amount spent on them.
- Estimate a latent variable using an appropriate data reduction technique.
  - Factor analysis (FA)/Principal Component Analysis (PCA):
  - Constructing indices by methodologically assessing contributions of a range of variables so that a combined measure can be extracted and used.

Dimension Reduction

- We know that we are using variables as proxy measures for real world concepts.
- We know that multicollinearity can cause problem.
- We aim for parsimony in our models (fewest number of predictor variables).
- Want to avoid two major obstacles:
  - Using a single variable created from a multi-item measure in which there is great heterogeneity; or
  - Using several variables that are highly correlated or unreliable.
- The goal is to obtain the "right" number of variables to represent a concept.
  - (and it might be one).

Everything that you need to know about Factor Analysis/Dimension Reduction

- The basic issue is the degree of correlation among a set of items.
- We want to test for clusters of variables or measures.
- Or to see whether different measures are tapping aspects of a common dimension.
- We expect to find "clumps" of items sometimes, and these are called **factors or components**.
  - If all of the items correlate highly with each other, then you have a simple outcome:
  - A single factor/component.
  - However, we often find that we don't have a single homogeneous factor/component.
  - Rather we have several clumps or factors/components.

Why?

- Many phenomena cannot be measured directly - no measurement scale exists.
  - E.g. social deprivation, cost of living, consumer confidence etc.
  - These are **latent** variables.
- To work with these latent variables we must use **manifest** variables and construct a measurement scale from these.
  - E.g. using measures of intelligence as an indicator of ability to perform on particular tests.

What are we doing?

- We are theorizing that there is a single individual concept, or a small number of concepts, which lead to different manifestations (which is captured in the variables).
- We are trying to distil the underlying concept(s) from its manifestations.
  - But how many dimensions does a concept have?
  - How should they be represented?

When do we use it?

- We may be interested in creating an agreed measure that will allow us to track changes over time.
- We may be interested in exploring particular dimensions of a concept.
- We may, pragmatically, be interested in reducing the number of variables we need to work with in an analysis.
- All the above are not mutually exclusive.

Approaches?

- Factor Analysis:
  - Exploratory:
    - Used in early stages of research to gather information about interrelationships between variables e.g. you have created an instrument with multiple questions and you want to test how well they measure concepts of interest
  - Confirmatory:
    - Used later in research to confirm hypotheses about the structure underlying a set of variables.
- Principal Components Analysis:
  - Used when working with a set of uncorrelated variables which it has been well established that they measure the same underlying component (s).

What are we looking at?

- Start off with a set of variables.
- By the end of the process have a smaller number 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.
- The variance (how different each measurement is from the mean) and co-variance (measure of how much two random variables change together) (i.e. correlation).

What are we looking at?

The reduction might be:

- By discovering that a particular linear combination 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'.

Factors/Components?

- A factor/component in this context is equivalent to what is known as a **latent** variable (also called a **construct**).
- A latent variable is a variable that cannot be measured directly but is measured indirectly through several observable variables (called **manifest** variables).
- **construct = latent variable = factor/component**

Factors?

- So how do we identify factors?
- Statisticians refer to "extraction" as the way to identify factors.
- There are at least 7 methods for extracting factors: factor analysis is one of these.

Example (From Andy Field)

- Factor analysis could be used to identify the "core" characteristics (out of a potentially large number of characteristics) that make a person popular.
- Candidate characteristics:
  - Time spent talking about the other person (Talk 1 - a relatively positive trait).
  - Level of social skills (Social Skills).
  - How interesting a person is to others (Interest).
  - Time spent talking about themselves (Talk 2 - a relatively negative trait).
  - Selfishness (Selfish).
  - The person's propensity to lie about themselves (Liar).

Getting started

- Need to be working with interval/ratio or ordinal data.
- Produce basic descriptive statistics look for missing or outlier values and take appropriate action.
- All the variables should measure the construct in the same direction.
  - The direction does not matter the important thing is that all the questions score in the same direction.
  - If they differ, then recode to ensure that the scales are the same.

Getting Started

Looking at how the concepts relate to each other.

	Talk 1	Social Skills	Interest	Talk 2	Selfish	Liar
Talk 1	1.000					
Social Skills	.772	1.000				
Interest	.646	.879	1.000			
Talk 2	.074	−.120	.054	1.000		
Selfish	−.131	.031	−.101	.441	1.000	
Liar	.068	.012	.110	.361	.277	1.000

Factor 1

Factor 2

- Look for correlations in the Correlation matrix.
- In Factor Analysis and PCA we look to reduce the R-matrix into smaller set of uncorrelated dimensions.
- Looking for correlations above 0.3.
- We seem to have 2 clusters.

Getting Started

Looking at what concepts relate to sociability and consideration traits in people.

	Talk 1	Social Skills	Interest	Talk 2	Selfish	Liar
Talk 1	1.000					
Social Skills	.772	1.000				
Interest	.646	.879	1.000			
Talk 2	.074	−.120	.054	1.000		
Selfish	−.131	.031	−.101	.441	1.000	
Liar	.068	.012	.110	.361	.277	1.000

Factor 1

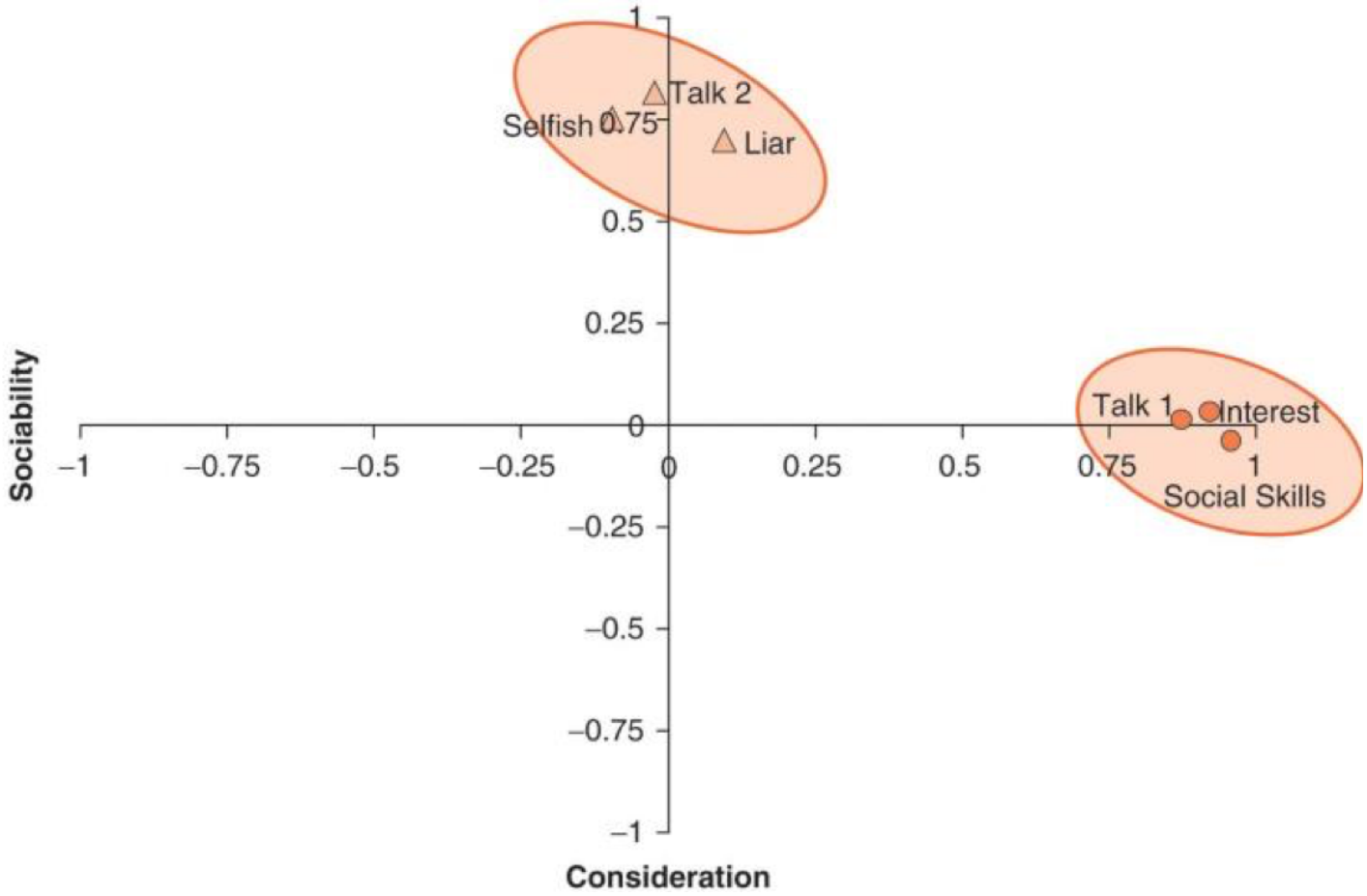
Factor 2

- Look for correlations in the Correlation matrix.
- Social Skills strongly correlated with Time spent talking about the other person, How interesting a person is to others is highly correlated with both of these. These are positive traits.
- Selfishness is moderately correlated with Time spent talking about themselves. Propensity to Lie is also moderately correlated with Time spent talking about themselves. The correlation between propensity to lie and selfishness is a weak correlation but it is approaching moderate. These are negative traits.

What is a Factor/Component?

- Factors/Component can be viewed as classification axes along which the individual variables can be plotted.
- The greater the **loading** of variables onto a factor/component, the more the factor/component explains relationships among those variables.
    - Loading can be viewed as the correlation between the latent variable and the manifest variable.
- Ideally, variables should be strongly related to (or load onto) only one factor/component.

Graphical Representation of a Factor Plot



- Note that each variable loads primarily on only one factor.
- Loadings tell use about the relative contribution that a variable makes to a factor.

Principal Component Analysis (PCA)

- A transformation of the data that estimates as many components as there are manifest variables.
- PCA reduces:
  - A set of highly correlated manifest variables.
  - To a set of unrelated components.
  - Arranged in descending order of importance.
- If PCA results in the first few components explaining a large amount of the variation we can without losing much information replace our manifest variables with our components.

Factor Analysis vs Principal Components Analysis

- FA attempts to achieve parsimony by explaining the maximum amount of common variance in a correlation matrix using the smallest number of explanatory constructs.
  - These 'explanatory constructs' are called factors.
- PCA tries to explain the maximum amount of total variance in a correlation matrix.
  - It does this by transforming the original variables into a set of linear components.

Assumptions

A linear relationship exists between the latent variable and the manifest variables.

$$F_i = b_1 X_{1_i} + b_2 X_{2_i} + ... + b_n X_{n_i} + \epsilon_i$$

$$Component_i \text{ or } Factor_i = b_1 Variable_{1_i} + b_2 Variable_{2_i} + ... + b_n Variable_{n_i}$$

$$F_i = b_1 X_{1_i} + b_2 X_{2_i} + ... + b_n X_{n_i}$$

$$Sociability_i = b_1 Talk1_i + b2 SocialSkills_i + b_3 Interest_i + b_4 Talk2_i + b_5 Selfish_i + b_6 Liar_i$$

$$Consideration_i = b_1 Talk1_i + b2 SocialSkills_i + b_3 Interest_i + b_4 Talk2_i + b_5 Selfish_i + b_6 Liar_i$$

Mathematical Equation

$$F_i = b_1 X_{1_i} + b_2 X_{2_i} + ... + b_n X_{n_i} + \epsilon_i$$

$$Component_i \text{ or } Factor_i = b_1 Variable_{1_i} + b_2 Variable_{2_i} + ... + b_n Variable_{n_i}$$

- Note that there is no b_0.
- The b values are called factor loadings.
- They can be interpreted as the correlation between the manifest variable and the factor.
- As with all correlations they range from -1 to +1.
- We should standardise our Xs so that our interpretation of the levels of correlation make sense.

Mathematical Equation

$$F_i = b_1 X_{1_i} + b_2 X_{2_i} + ... + b_n X_{n_i} + \epsilon_i$$

$$Component_i \text{ or } Factor_i = b_1 Variable_{1_i} + b_2 Variable_{2_i} + ... + b_n Variable_{n_i}$$

- For every case in our dataset we have the values of X but not of F.
- Therefore we cannot fit a regression line as we have done previously.
- But we can transform the relationships between the variables using a method of extraction to retrieve factors from the data

Loadings

Both FA and PCA are linear models in which loadings are used as weights.

- These loadings can be expressed as a matrix
- This matrix is called the factor matrix or component matrix (if doing PCA).
- The assumption of FA(but not PCA) is that these algebraic factors represent real-world dimensions.

Variables	Factors	
	Sociability	Consideration
Talk 1	0.87	0.01
Social Skills	0.96	-0.03
Interest	0.92	0.04
Talk 2	0.00	0.82
Selfish	-0.10	0.75
Liar	0.09	0.70

Loadings

- 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.
- **The higher the value the better.**

Variables	Factors	
	Sociability	Consideration
Talk 1	0.87	0.01
Social Skills	0.96	-0.03
Interest	0.92	0.04
Talk 2	0.00	0.82
Selfish	-0.10	0.75
Liar	0.09	0.70

Definition

There is no definition of what constitutes a high or low factor loading. The convention is that a loading of magnitude above 0.3 (irrespective of sign) are considered high.

Mathematical Representation of a Factor Plot

- In our example there are two factors underlying the *popularity* construct: **general sociability** and **consideration**.
- We can construct equations that describe each factor in terms of the variables that have been measured.

$$Sociability_i = b_1 Talk1_i + b2SocialSkills_i + b_3 Interest_i + b_4 Talk2_i + b_5 Selfish_i + b_6 Liar_i + \epsilon_i$$

$$Consideration_i = b_1 Talk1_i + b2SocialSkills_i + b_3 Interest_i + b_4 Talk2_i + b_5 Selfish_i + b_6 Liar_i + \epsilon_i$$

Mathematical Representation of a Factor Plot

The values of the "b's" in the two equations differ, depending on the relative importance of each variable to a particular factor.

$$Sociability_i = 0.87\,Talk1_i + 0.96\,SocialSkills_i + 0.92\,Interest_i + 0\,Talk2_i - 0.1\,Selfish_i + 0.09\,Liar_i + \epsilon_i$$

$$Consideration_i = 0.01\,Talk1_i - 0.03\,SocialSkills_i + 0.04\,Interest_i + 0.82\,Talk2_i + 0.75\,Selfish_i + 0.7\,Liar_i + \epsilon_i$$

Replace values of b with the co-ordinate of each variable on the graph.

Ideally, variables should have very high *b*-values for one factor and very low *b*-values for all other factors.

Doing Factor Analysis: An Example (Andy Field)

- Students often become stressed about statistics (RAQ) and the use of computers and/or R to analyze data.
- Suppose we develop a questionnaire to measure this propensity (see sample items on the following slides; the data can be found in RAQ.dat).
- Does the questionnaire measure a single construct?
- Or is it possible that there are multiple aspects comprising students' anxiety toward R?

RAQ Questionnaire

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 R	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 R 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	R always crashes when I try to use it	O	O	O	O	O
19	Everybody looks at me when I use R	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 R than I am	O	O	O	O	O

Initial Considerations

- The quality of the data.
- Sample size is important!
  - A sample of 300 or more will likely provide a stable factor solution, but depends on the number of variables and factors identified.
  - Sample to variable ratio N:p where N is no of cases and p is no of variables (aim for 10:1).
- Correlations among the items should not be too low (less than .3) or too high (greater than .8), but the pattern is what is important.
- Factors that have four or more loadings greater than 0.6 are likely to be reliable regardless of sample size.
- Screen the correlation matrix, eliminate any variables that obviously cause concern.

Step 1: Generate a correlation matrix (lots of ways)

```
#load data
raqData<-read.delim("raq.dat", header = TRUE)
#create a correlation matrix
raqMatrix<-cor(raqData)
round(raqMatrix, 2)

#Showing significance levels also
Hmisc::rcorr(as.matrix(raqData))

#Visualisation of correlations using circles
corrplot::corrplot(raqMatrix, method="circle")

#Visualisation using numbers
corrplot::corrplot(raqMatrix, method="number")

#Visualisation of significance levels at 0.05
res1 <- corrplot::cor.mtest(raqMatrix, conf.level = .95)
corrplot::corrplot(raqMatrix, p.mat = res1$p, sig.level = .05)

#Showing p-value for non-significant results
corrplot(raqMatrix, p.mat = res1$p, insig = "p-value")
```

Step 1: Generate a correlation matrix (lots of ways)

```
##Using ggcorrplot
p.mat <- ggcorrplot::cor_pmat(raqData)
ggcorrplot::ggcorrplot(raqMatrix, title =
"Correlation matrix for RAQ data")

#Showing Xs for non-significant correlations
ggcorrplot::ggcorrplot(raqMatrix, title =
"Correlation matrix for RAQ data",
p.mat = p.mat, sig.level = .05)

#Showing lower diagonal
ggcorrplot::ggcorrplot(raqMatrix, title =
"Correlation matrix for RAQ data",
p.mat = p.mat, sig.level = .05, type="lower")

#Showing the co-efficients
ggcorrplot::ggcorrplot(raqMatrix, lab=TRUE, title =
"Correlation matrix for RAQ data", type="lower")
```

Screen Correlation Matrix

- The quality of analysis depends upon the quality of the data.
- Test variables should correlate quite well:
  - $r > .3$
- Avoid multicollinearity:
  - several variables highly correlated, $r > .80$
- Avoid singularity:
  - some variables perfectly correlated, $r = 1$
- Screen the correlation matrix, eliminate any variables that obviously cause concern.

Step 2: Generate Relevant Statistics

- Checking for multicollinearity:
  - Inspect determinant and check to see if it is greater than 0.00001.
- The Bartlett Test of Sphericity:
  - Tests that your variables are correlated.
  - 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 - all correlations are perfect).
  - You are looking for a statistically significant p value which suggests that they are related and a PCA/FA would be useful.
  - Highly unlikely for us to have obtained the observed correlation matrix from a population with zero correlation.

Step 2: Generate Relevant Statistics

- Kaiser-Meyer-Olkin Measure of Sampling Adequacy (MSA)
  - Indicates the proportion of variance in your variables that might be caused by an underlying factor.
  - Doesn't produce a p-value.
  - High values close to 1 suggest that PCA/FA might be useful.
  - Values of 0.8 or over are considered strong.
  - Anything less than 0.5 suggests that PCA/FA won't be useful.
- Kaiser put together the following descriptors:
  - 0.00 to 0.49 unacceptable.
  - 0.50 to 0.59 miserable.
  - 0.60 to 0.69 mediocre.
  - 0.70 to 0.79 middling.
  - 0.80 to 0.89 meritorious.
  - 0.90 to 1.00 marvellous.

Step 2: Generate Relevant Statistics

```
#Bartlett's test

psych::cortest.bartlett(raqData)
psych:: cortest.bartlett(raqMatrix, n = 2571)

#Measure of Sampling Accuracy (execute one of these):
psych::kmo(raqData)
REdaS::KMOS(raqData)

#Determinant (execute one of these from the base package):
det(raqMatrix)
det(cor(raqData))
```

Initial Considerations – Is it possible?

- Check the Determinant:
  - Indicator of multicollinearity.
  - Should be > 0.00001.
- Check Kaiser-Meyer-Olkin (KMO):
  - Measures sampling adequacy (how well suited is your data to the dimension reduction).
  - Should be $> .5$
  - Indicates the proportion of variance in your variables that might be caused by an underlying factor.
- Check Bartlett's test of Sphericity:
  - Tests whether the R-matrix is an identity matrix.
  - Should be significant at $p < .05$
  - We can see the correlations ourselves, this just confirms that it is possible to combine the variables because they are correlated.

Our Example RAQ

```
Determinant: 0.0005271037

KMO ReDAs: 0.9302245
KMO psych: 0.93

Bartlett:chi-square=19334.49, p=0, df=253
```

Step 3: Do The Dimension Reduction

RAQ:

- In this case we are using all the variables so Principal Component Analysis is the function we need to apply.
- We also create a Scree plot which we will use later.

FA/PCA

- Principal Components Analysis (PCA)
  - This is a data reduction method, not a true factor analysis (as it analyses all of the variance of the variables) so attempts to extract same number of components as there are variables.
- Factor Analysis
  - Need to choose a method of factor extraction.
  - There are six extraction methods : unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring and image factoring.
  - The general rule of thumb is:
    - If your data are normally distributed then use maximum likelihood (ML).
    - If your data are not normally distributed then use Principal Axis Factoring (PAF).

Step 3: Do the Dimension Reduction

```
#Principal Component Analysis
#On raw data using principal components analysis,
#nfactors is set to the number of variables we expect
#to get out which is equal to the number going in.
pc1 <-  principal(raqData, nfactors = 23,
rotate = "none")
pc1 <-  principal(raqData, nfactors = length(raqData),
rotate = "none")
plot(pc1$values, type = "b") #scree plot
```

Step 3: Do the Dimension Reduction

```
#Factor Analysis
#Principal Axis Factoring
pc3 <- fa(raqMatrix, nfactors=4, obs=NA, n.iter=1,
rotate="varimax", fm="pa")
psych::print.psych(pc3,cut=0.3, sort=TRUE)
fa.graph(pc3)
fa.sort(pc3$loading)
#create a diagram showing the factors and how
#the manifest variables load
fa.diagram(pc3)
plot(pc3$values, type = "b") #scree plot
```

How many variables do we want?

- How many latent variables do we want to keep?
  - The common method is to select factors with Eigenvalues greater than 1 option.
- *Eigenvalue* = the variance of a component or factor.
- Kaiser (1960) recommends retaining all factors with eigenvalues greater than 1.
  - Based on the idea that eigenvalues represent the amount of variance explained by a factor and that an eigenvalue of 1 represents a substantial amount of variation.
  - Jolliffe (1972; 1986) reported that Kaiser's criterion is too strict and recommended retaining factors with eigenvalues more than 0.7.

Extracting the Factors

- If we have n manifest variables in our analysis, PCA transforms the data such that the total variance of the components n will be redistributed among the components.
  - The first will have the largest eigenvalue.
  - The next the next highest etc.
- If the eigenvalue is divided by n the result is the proportion of variance explained by that factor.
- A rule of thumb is to retain in the final solution all factors with an Eigenvalue greater than one.

Retrieving our eigenvalues

```
pc1$values #output eigenvalues

#Another way to look at Eigen values plus
#variance explained
pcf=princomp(raqData)
factoextra::get_eigenvalue(pcf)
```

Example

```
pc1$values #output eigenvalues
```

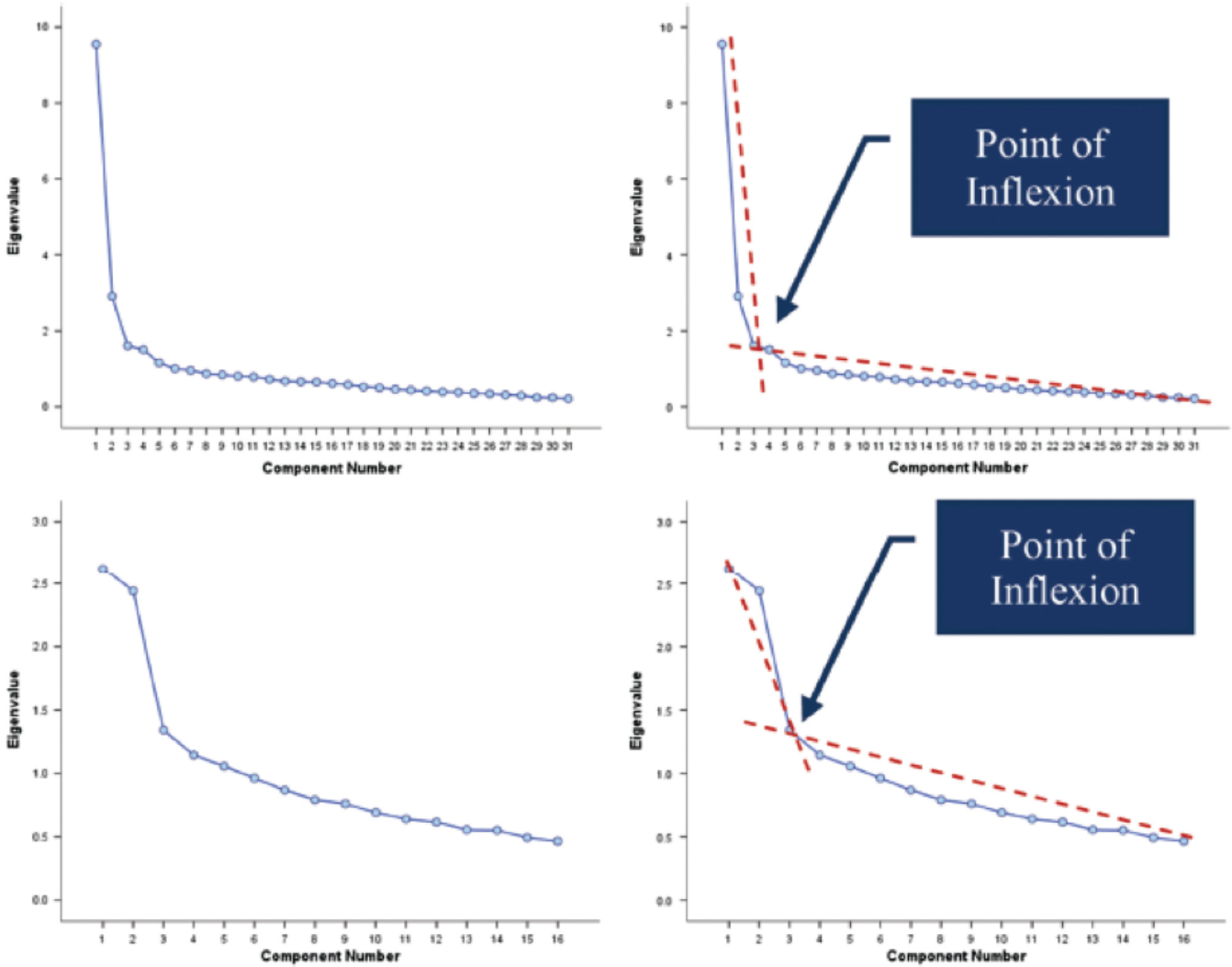
```
##  [1] 7.2900471 1.7388287 1.3167515 1.2271982 0.9878779 0.8953304 0.8055604
##  [8] 0.7828199 0.7509712 0.7169577 0.6835877 0.6695016 0.6119976 0.5777377
## [15] 0.5491875 0.5231504 0.5083962 0.4559399 0.4238036 0.4077909 0.3794799
## [22] 0.3640223 0.3330618
```

```
#Another way to look at eigen values plus variance explained
pcf=princomp(raqData)
factoextra::get_eigenvalue(pcf)
```

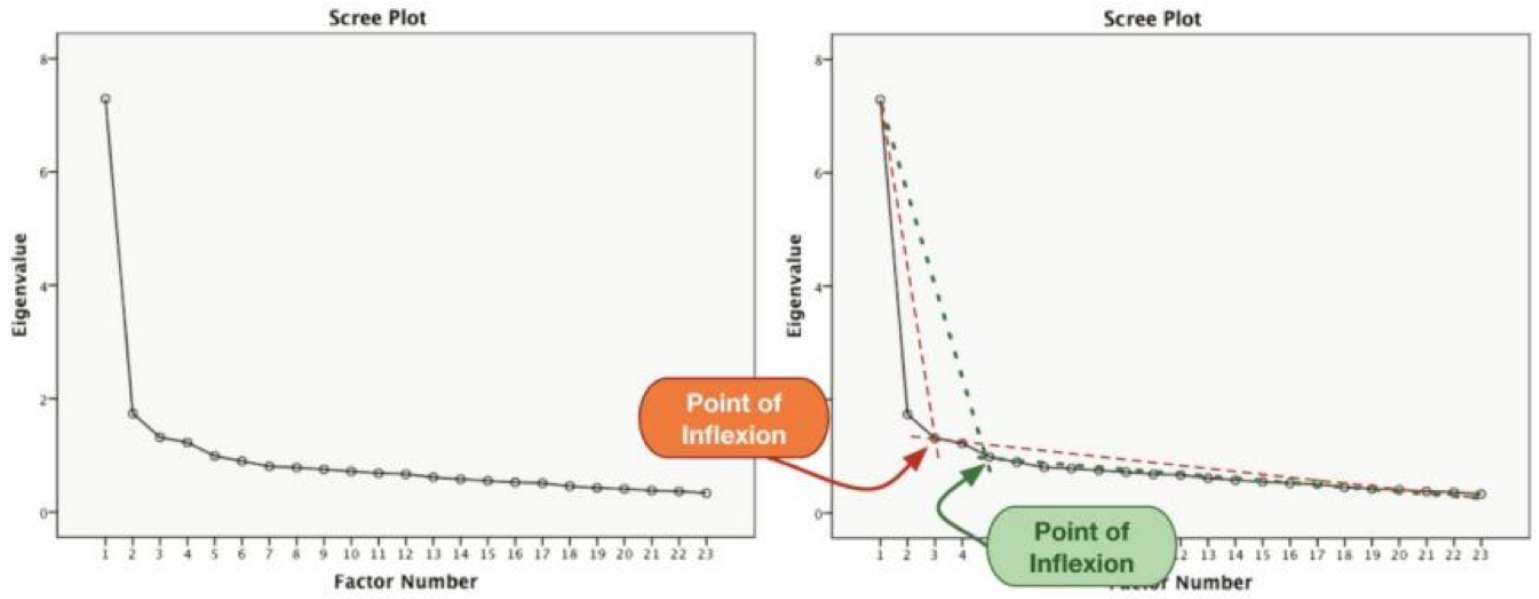
```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1   7.0430039        31.059169                      31.05917
## Dim.2   1.9919543         8.784383                      39.84355
## Dim.3   1.3755696         6.066168                      45.90972
## Dim.4   1.1415644         5.034221                      50.94394
## Dim.5   1.0453716         4.610018                      55.55396
## Dim.6   0.8929067         3.937658                      59.49162
## Dim.7   0.8530615         3.761943                      63.25356
## Dim.8   0.8136079         3.587955                      66.84152
## Dim.9   0.7489525         3.302830                      70.14435
## Dim.10  0.7054281         3.110890                      73.25524
## Dim.11  0.6894466         3.040413                      76.29565
## Dim.12  0.6195139         2.732014                      79.02766
## Dim.13  0.5624056         2.480171                      81.50783
## Dim.14  0.5500250         2.425573                      83.93341
## Dim.15  0.5309084         2.341270                      86.27468
```

Finding Factors

Cattell (1966) suggests using the 'point of inflexion' of the scree plot to decide how many factors to extract.



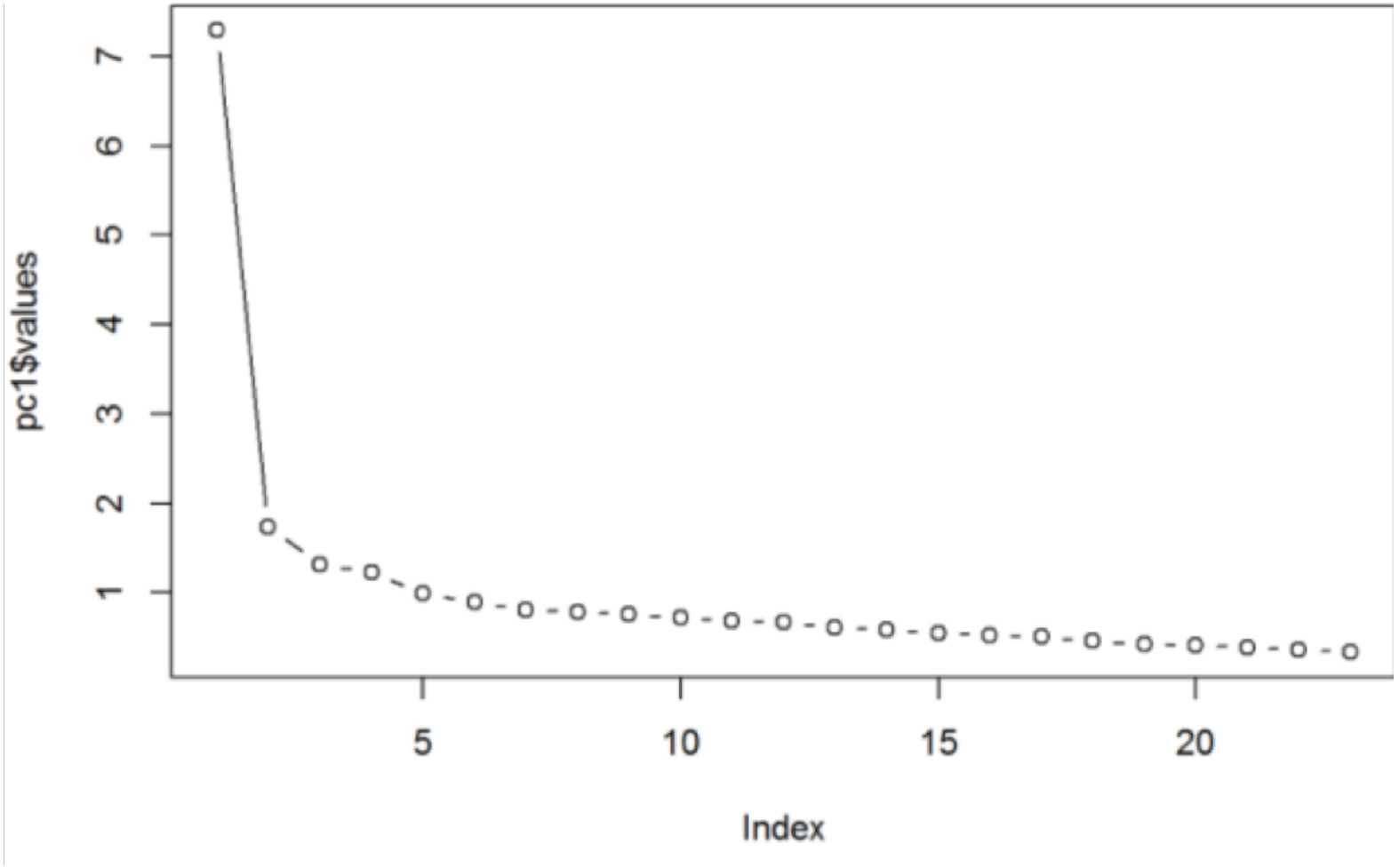
Scree Plot



OUTPUT 17.6

Look for change in shape of the plot - points of inflexion, only factors/components above this point are retained.

Example



Factor/Component Extraction

Which Rule?

- Use Kaiser's Extraction when
  - less than 30 variables, communalities after extraction > 0.7.
  - sample size > 250 and mean communality ≥ 0.6.
- Scree plot is good if sample size is > 200.

Total Variance Explained

Indicates how much of the variability in the data has been modelled by the extracted factors.

Variance Explained

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1   7.0430039        31.059169                    31.05917
## Dim.2   1.9919543         8.784383                    39.84355
## Dim.3   1.3755696         6.066168                    45.90972
## Dim.4   1.1415644         5.034221                    50.94394
## Dim.5   1.0453716         4.610018                    55.55396
## Dim.6   0.8929067         3.937658                    59.49162
## Dim.7   0.8530615         3.761943                    63.25356
## Dim.8   0.8136079         3.587955                    66.84152
## Dim.9   0.7489525         3.302830                    70.14435
## Dim.10  0.7054281         3.110890                    73.25524
## Dim.11  0.6894466         3.040413                    76.29565
## Dim.12  0.6195139         2.732014                    79.02766
## Dim.13  0.5624056         2.480171                    81.50783
## Dim.14  0.5500250         2.425573                    83.93341
```

- **variance.percent**: Calculated by dividing Eigenvalue by the number of factors/components.
- **cumulative.variance.percent**: Indicates how much of the variability in the data has been modelled by the extracted factors. Factor/Component 11 explains 31.69%, Factor/Component 1+2 explain 39.26% etc. Together the 4 uncorrelated factors summarise 50.32% of the information in the original inter-related variables.

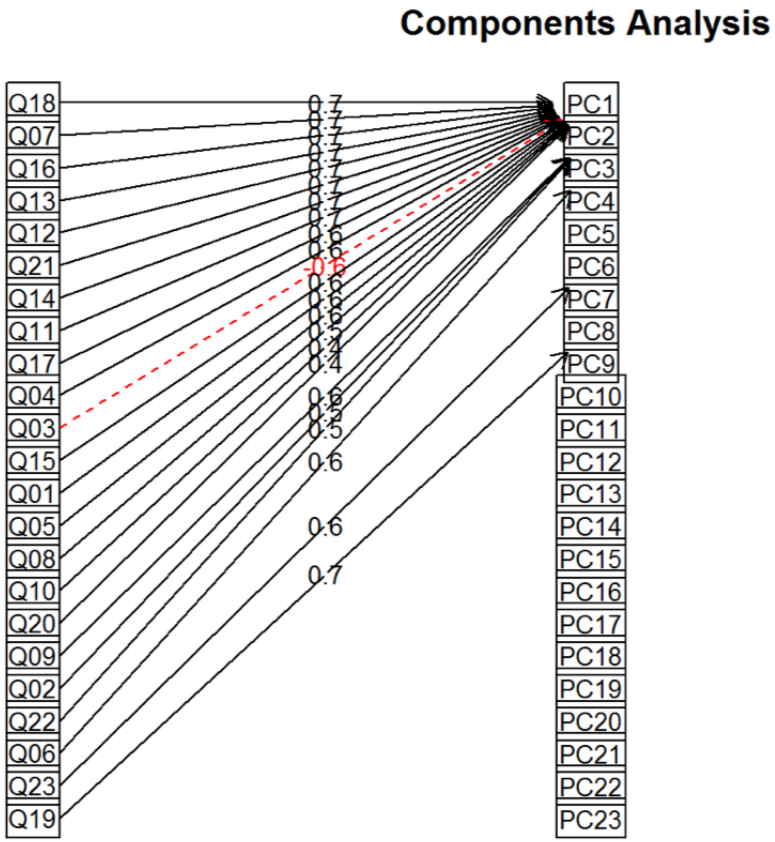
In R

*pc*1$*Vaccounted*#*Variance accounted for*

Deciding on Factors/Components to Retain

- Need to look at both loadings and communalities.
- Strong data:
  - Uniformly high communalities (> 0.8 is strong but it is more likely to be in the region of $0.4 - 0.7$).
  - Item loadings above 0.3 is acceptable.
  - None, or few, cross loadings (a variable loads higher than 0.32 on two or more factors).
  - Several variables loading on to each factor/component (factor loadings of 0.3 or greater).
  - A factor/component with fewer than 3 items is typically weak.

Loadings (PA1=Dim1=Component 1)



Communalities

- b is the correlation between the manifest variable and the factor or component.
- The square of the correlation is the proportion of variance in the manifest variable that can be explained by the factor.
- If we square and sum all the factor loadings for a manifest variable across the factor equations we have the communality.
  - How much of the variance in the manifest variable is accounted for by the combination of factors.

Communalities

- In the initial PCA because we have as many components/factors as manifest variables this will be 1.
- If we drop components it will be less than 1.
- This will then provide an indication of how much each manifest variable contributes to our reduced solution.

Factor/Component Matrix - Loadings

	Factor			
	1	2	3	4
Statistics makes me cry	.557	.112	-.090	.205
My friends will think I'm stupid for not being able to cope with SPSS	-.281	.372	.185	.097
Standard deviations excite me	-.605	.249	.204	-.050
I dream that Pearson is attacking me with correlation coefficients	.607	.082	-.046	.204
I don't understand statistics	.522	.039	-.011	.157
I have little experience of computers	.552	.023	.489	-.223
All computers hate me	.663	-.025	.221	.013
I have never been good at mathematics	.546	.483	-.282	-.188
My friends are better at statistics than me	-.266	.460	.142	.190
Computers are useful only for playing games	.404	-.007	.157	-.094
I did badly at mathematics at school	.646	.313	-.219	-.257
People try to tell you that SPSS makes statistics easier to understand but it doesn't	.643	-.099	.070	.160
I worry that I will cause irreparable damage because of my incompetence with computers	.650	.022	.212	-.078
Computers have minds of their own and deliberately go wrong whenever I use them	.628	-.037	.163	.051
Computers are out to get me	.559	-.005	.040	-.088
I weep openly at the mention of central tendency	.653	-.017	-.094	.149
I slip into a coma whenever I see an equation	.632	.359	-.170	-.136
SPSS always crashes when I try to use it	.684	-.039	.273	.014
Everybody looks at me when I use SPSS	-.397	.268	.116	.049
I can't sleep for thoughts of eigenvectors	.407	-.153	-.208	.183
I wake up under my duvet thinking that I am trapped under a normal distribution	.633	-.084	-.076	.232
My friends are better at SPSS than I am	-.280	.291	.073	.281
If I'm good at statistics my friends will think I'm a nerd	-.131	.183	.100	.236

Extraction Method: Principal Axis Factoring.

a. 4 factors extracted. 11 iterations required.

Called component matrix for PCA

Definition

Correlation between a specific manifest/observed variable and a specific factor. These are the b's for the factor equations.

Communalities

Communalities

	Initial	Extraction
Statistics makes me cry	.373	.373
My friends will think I'm stupid for not being able to cope with SPSS	.188	.260
Standard deviations excite me	.398	.472
I dream that Pearson is attacking me with correlation coefficients	.385	.419
I don't understand statistics	.291	.299
I have little experience of computers	.427	.594
All computers hate me	.470	.489
I have never been good at mathematics	.490	.646
My friends are better at statistics than me	.220	.339
Computers are useful only for playing games	.197	.197
I did badly at mathematics at school	.530	.629
People try to tell you that SPSS makes statistics easier to understand but it doesn't	.424	.453
I worry that I will cause irreparable damage because of my incompetence with computers	.451	.474
Computers have minds of their own and deliberately go wrong whenever I use them	.393	.425
Computers are out to get me	.344	.322
I weep openly at the mention of central tendency	.463	.458
I slip into a coma whenever I see an equation	.494	.575
SPSS always crashes	.492	.544

Communality = sum of the square of the loading of a variable across a component/factor.

Definition

How much variance in the manifest/observed variables is accounted for by the factors found?

Communalities

	1	2	3	4
I wake up under my duvet thinking that I am trapped under a normal distribution.	0.594	0.264	0.149	0.15

Table: Factor Matrix

	Extraction
I wake up under my duvet thinking that I am trapped under a normal distribution.	.467

Table: Communalities

Take Q1 Statistics Make Me Cry
Communality=$.594^2 + .264^2 + .149^2 + .15^2 = .467$

Step 4: Factor Rotation

- Given the communalities, there are multiple solutions for loadings that result in the same communality and correlation.
  - "Correct" answer is dependent upon interpretability.
- To aid interpretation it is possible to maximise the loading of a variable on one factor while minimising its loading on all other factors.
- The aim of rotation is to clarify the data structure.
  - You cannot apply a rotation if only one factor emerges.
  - Aim to have non-zero loadings on each factor for only some of the variables.
- Rotation reallocates the variable variance among the factors.
  - Shifts variance from earlier factors to later ones.
  - Interpretation shifts from unrotated factors delineating the most comprehensive data patterns to factors delineating the distinct groups of interrelated data.

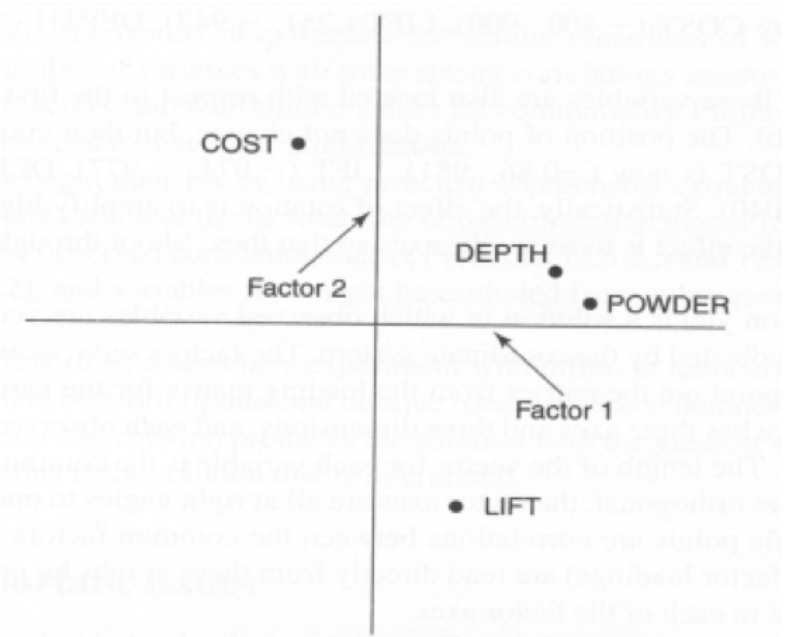
Rotation - Why?

- The factor patterns define decreasing amounts of variation in the data.
- Each pattern may involve all or most of the variables and the variables may have moderate or high loadings for several factor patterns.
- To uncover the first pattern, a factor is fitted to the data to account for the greatest regularity;
  - Each successive factor is fitted to best define the remaining regularity.
  - The result of this is that the first unrotated factor may be located between independent clusters of interrelated variables.
  - These clusters cannot be distinguished in terms of their loadings on the first factor, although they will have loadings different in sign on the second and subsequent factors.
  - Each factor is rotated until it defines a distinct cluster of interrelated variables.

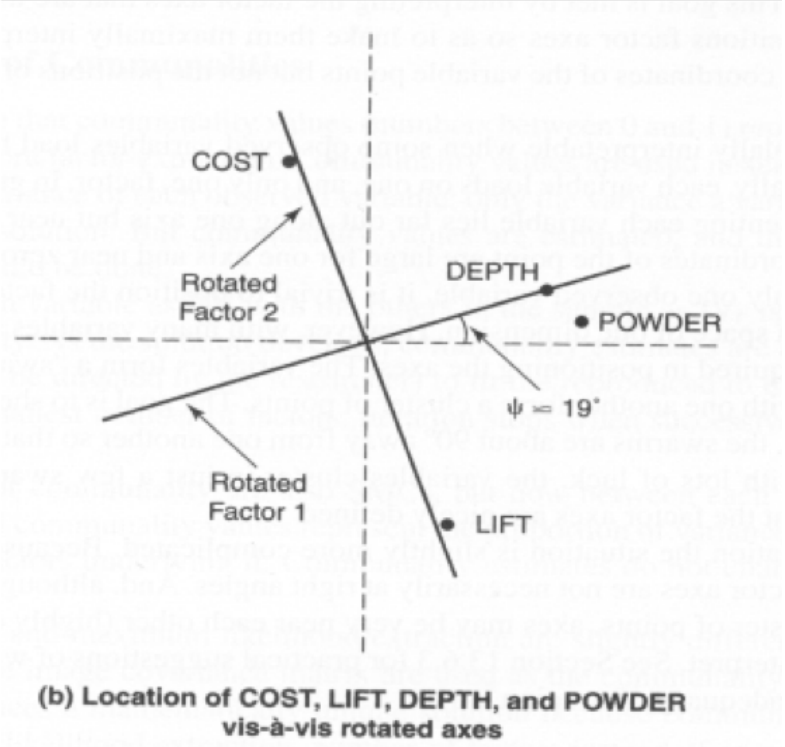
Geometric Rotation

- Factor extraction equivalent to coordinate planes.
- Factors are the axes.
- Length of the line from the origin to the variable coordinates is equal to the communality for that variable.
- Orthogonal Factors are at right angles.
- Factor loadings are found by dropping a line from the variable coordinates to the factor at a right angle.
- Repositioning the axes changes the loadings on the factor but keeps the relative positioning of the points the same.

Geometric Rotation

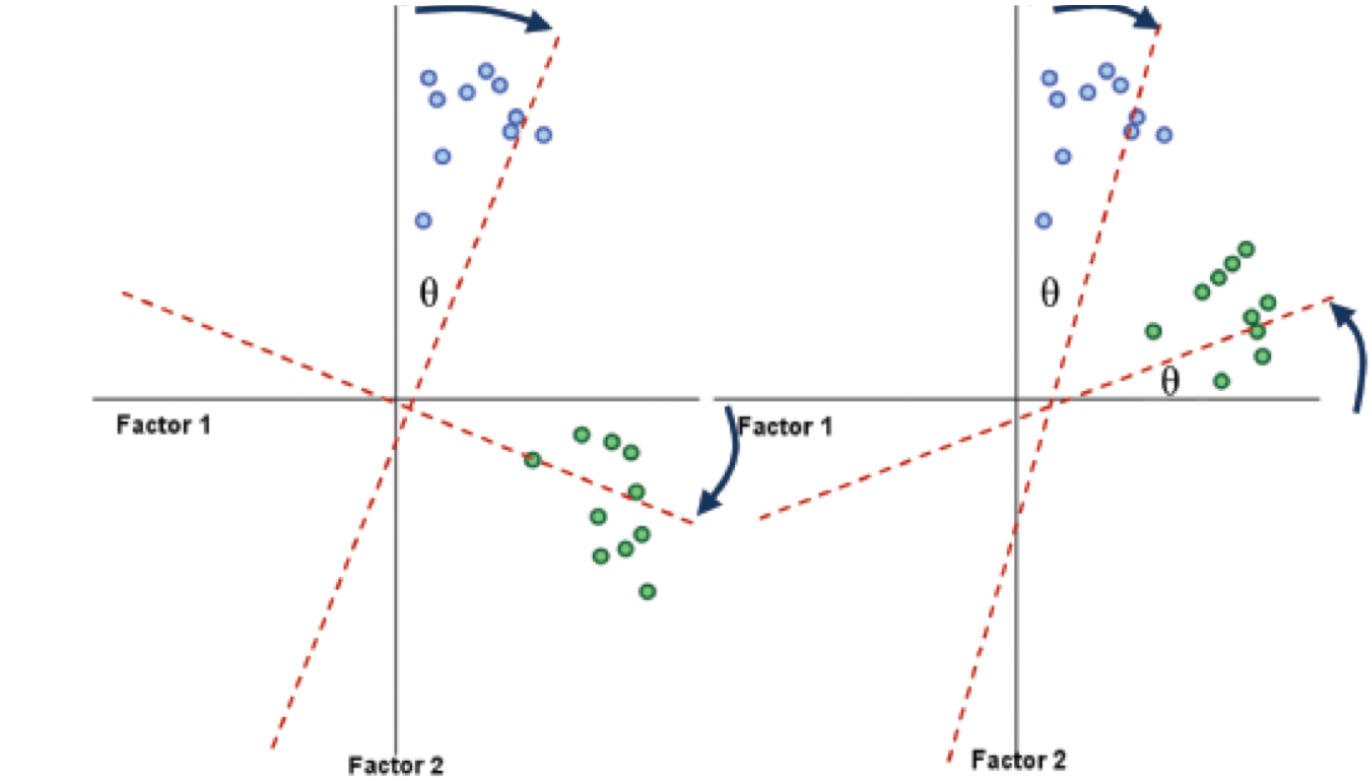


(a) Location of COST, LIFT, DEPTH, and POWDER after extraction, before rotation



(b) Location of COST, LIFT, DEPTH, and POWDER vis-à-vis rotated axes

Geometric Rotation



Orthogonal Rotation **Oblique Rotation**

Rotating the axes - Aim is to make the location of the axes fit the actual data points better. Move to a position that encompasses the actual data points better overall. This will hopefully make them more interpretable.

Rotation

```
#Rotation PCA
pc2 <-  principal(raqData, nfactors = 4,
rotate = "varimax")
#Extracting 4 factors
print.psych(pc2, cut = 0.3, sort = TRUE)
#Get the output as you would for initial solution
```

Rotated Factor Matrix (FA)

Check the factor loadings after rotation.

Need to consider rotated factor loadings before making a final decision.

Step 4: Reliability Analysis

- A Cronbach's alpha measurement of consistency tells you how consistently the variables behave as a scale.
  - Internal consistency refers to the general agreement between multiple items (often Likert scale items) that make-up a composite score of a survey measurement of a given construct.
  - This agreement is generally measured by the correlation between items.
- If it's high (say .80 or .90), then we probably have one factor/component.

Interpreting Cronbach's Alpha

- Depends on the number of items.
  - More questions = bigger α.
  - Low numbers of items (< 10) can result in low levels of alpha.
- Treat Subscales separately.
- Remember to reverse score reverse phrased items!
  - If not, α is reduced and can even be negative.

Reliability

Cronbach's Alpha:

- Ranges from 0 (no reliability) to 1 (complete reliability).
- Should be .7 or greater to be considered "reliable".
- The closer Cronbach's alpha coefficient is to 1.0 the greater the internal consistency of the items in the scale.
- George and Mallery provide the following rules of thumb:
  - $>$ .9 - Excellent
  - $>$ .8 - Good
  - $>$ .7 - Acceptable
  - $>$ .6 - Questionable
  - $>$ .5 - Poor and
  - $<$ .5 - Unacceptable

Item-Total Statistics: Fear of Computers sub-scale

```
 raw_alpha std.alpha G6(smc) average_r S/N    ase mean   sd median_r
      0.82      0.82    0.81       0.4 4.6 0.0052  3.4 0.71     0.39

 lower alpha upper     95% confidence boundaries
0.81 0.82 0.83

 Reliability if an item is dropped:
    raw_alpha std.alpha G6(smc) average_r S/N alpha se  var.r med.r
Q06      0.79      0.79    0.77      0.38 3.7   0.0063 0.0081  0.38
Q07      0.79      0.79    0.77      0.38 3.7   0.0063 0.0079  0.36
Q10      0.82      0.82    0.80      0.44 4.7   0.0053 0.0043  0.44
Q13      0.79      0.79    0.77      0.39 3.8   0.0062 0.0081  0.38
Q14      0.80      0.80    0.77      0.39 3.9   0.0060 0.0085  0.36
Q15      0.81      0.81    0.79      0.41 4.2   0.0056 0.0095  0.44
Q18      0.79      0.78    0.76      0.38 3.6   0.0064 0.0058  0.38

 Item statistics
       n raw.r std.r r.cor r.drop mean   sd
Q06 2571  0.75  0.74  0.68   0.62  3.8 1.12
Q07 2571  0.75  0.73  0.68   0.62  3.1 1.10
Q10 2571  0.54  0.57  0.44   0.40  3.7 0.88
Q13 2571  0.72  0.73  0.67   0.61  3.6 0.95
Q14 2571  0.70  0.70  0.64   0.58  3.1 1.00
Q15 2571  0.64  0.64  0.54   0.49  3.2 1.01
Q18 2571  0.76  0.76  0.72   0.65  3.4 1.05
```

Alpha above .7 is strong, .6 is acceptable.

Item-Total Statistics: Fear of Statistics sub-scale

```
 raw_alpha std.alpha G6(smc) average_r S/N  ase mean  sd median_r
      0.61      0.64    0.71      0.18 1.8 0.01  3.1 0.5     0.34

 lower alpha upper     95% confidence boundaries
0.59 0.61 0.63

 Reliability if an item is dropped:
    raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
Q01      0.52      0.56    0.64      0.15 1.3   0.0128 0.123  0.33
Q03      0.80      0.80    0.79      0.37 4.1   0.0061 0.007  0.40
Q04      0.50      0.55    0.64      0.15 1.2   0.0133 0.119  0.33
Q05      0.52      0.57    0.66      0.16 1.3   0.0127 0.129  0.33
Q12      0.52      0.56    0.65      0.15 1.3   0.0131 0.120  0.33
Q16      0.51      0.55    0.63      0.15 1.2   0.0133 0.116  0.33
Q20      0.56      0.60    0.68      0.18 1.5   0.0118 0.133  0.39
Q21      0.50      0.55    0.63      0.15 1.2   0.0136 0.117  0.30
```

Look for items that would improve alpha if removed.

In R

```
#Reliability analysis

computerFear<-raqData[,c(6, 7, 10, 13, 14, 15, 18)]
statisticsFear <- raqData[, c(1, 3, 4, 5, 12, 16, 20, 21)]
mathFear <- raqData[, c(8, 11, 17)]
peerEvaluation <- raqData[, c(2, 9, 19, 22, 23)]


alpha(computerFear)
alpha(statisticsFear, keys = c(1, -1, 1, 1, 1, 1, 1, 1))
alpha(mathFear)
alpha(peerEvaluation)
alpha(statisticsFear) #for illustrative purposes
```

Reporting the Results

A principal component analysis (PCA) was conducted on the 23 items with orthogonal rotation (varimax). Bartlett's test of sphericity, $X_2(253) = 19334.49$, $p < .001$, indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. Four components had eigenvalues over Kaiser's criterion of 1 and in combination explained 50.94% of the variance. The scree plot was slightly ambiguous and showed inflexions that would justify retaining either 2 or 4 factors.

Given the large sample size, and the convergence of the scree plot and Kaiser's criterion on four components, four components were retained in the final analysis. Component 1 represents a fear of computers, component 2 a fear of statistics, component 3 a fear of math, and component 4 peer evaluation concerns.

The fear of computers, and fear of math subscales of the RAQ all had high reliabilities, all Cronbach's $\alpha = .82$. The fear of statistics had an acceptable reliability of Cronbach's $\alpha = .61$. However, the fear of negative peer evaluation subscale had a relatively low reliability, Cronbach's $\alpha = .57$.

The End?

- Describe Factor Structure/Reliability
- What items should be retained?
  - What items did you eliminate and why?
- Application
  - How does it fit in with your theory?
  - Where will your questionnaire be used? (if you are this is your research focus)

Exploratory Factor Analysis (EFA)

- Is a variable reduction technique which identifies the number of latent constructs and the underlying factor structure of a set of variables.
- Hypothesizes an underlying construct, a variable not measured directly.
- Estimates factors which influence responses on observed variables.
- Allows you to describe and identify the number of latent constructs (factors).
- Includes unique factors, error due to unreliability in measurement.

Exploratory Factor Analysis (EFA)

- Traditionally has been used to explore the possible underlying factor structure of a set of measured variables without imposing any preconceived structure on the outcome.
- Since measurement is not perfect, error or unreliability is estimated and specified explicitly in the diagram.
- Factor loadings (parameter estimates) help interpret factors.
- Loadings are the correlation between observed variables and factors are standardized regression weights if variables are standardized (weights used to predict variables from factor), and are path coefficients in path analysis.
- Standardized linear weights represent the effect size of the factor on variability of observed variables.

Principal Component Analysis (PCA)

- Is a large sample procedure.
- Is a variable reduction technique.
- Is used when variables are highly correlated.
- Reduces the number of observed variables to a smaller number of principal components which account for most of the variance of the observed variables.
- The total amount of variance in PCA is equal to the number of observed variables being analyzed.
- In PCA, observed variables are standardized,
  - e.g., mean=0, standard deviation=1, diagonals of the matrix are equal to 1.

Principal Component Analysis (PCA)

- The amount of variance explained is equal to the trace of the matrix (sum of the diagonals of the decomposed correlation matrix).
- The number of components extracted is equal to the number of observed variables in the analysis.
- The first principal component identified accounts for most of the variance in the data.
- The second component identified accounts for the second largest amount of variance in the data and is uncorrelated with the first principal component and so on.
- Components accounting for maximal variance are retained while other components accounting for a trivial amount of variance are not retained.
- Eigenvalues indicate the amount of variance explained by each component.
- Eigenvectors are the weights used to calculate components scores.

Principal Component Analysis (PCA)

- A transformation of the data that estimates as many components as there are manifest variables.
- Constraints
  - The sum of squared factor loadings is constrained to be 1.

$$(b_1^2 + b_2^2 + ... + b_n^2) = 1$$

  - The components are to be uncorrelated.
    - The coefficient must be 0.
    - Referred to in factor analysis as being orthogonal (at right angles to each other).
  - The first component must account for the maximum amount of variance possible.
  - The second for the largest amount of the variance left after the first component and so on.

PCA and EFA Requirements

- Measurement scale is interval or ratio level (or ordinal).
- Random sample
  - At least 5 observations per observed variable and at least 100 observations in total required.
- Larger sample sizes are recommended for more stable estimates, 10-20 observations per observed variable.
- Over sample to compensate for missing values.
- Linear relationship between observed variables.

PCA

- Not based on any causative relationship between the components and the manifest variables.
- It is not proposing that the components are latent variables.
- It is simply rearranging the data under the given constraints.

Factor Analysis vs Principal Component Analysis

- Common variance
  - variance shared with other variables (communality)
- Unique variance
  - variance associated with a specific variable and not explained by correlation with other variables
- Error variance
  - variance also unexplained by correlation with other variables due to factors such as measurement error or unreliability in data gathering

Factor Analysis vs Principal Component Analysis

- Both attempt to produce a smaller number of linear combinations of original variables that captures or accounts for most of the variability in the pattern of correlations.
- In PCA
  - Assumes all the variance is common.
  - There is little unique or error variance – it is a very small part of the total.
  - Objective is to identify minimum number of factors to capture the total amount of variation.
- In FA
  - Factors are estimated using a mathematical model.
  - Only the shared/common variance is used.
  - Objective is to identify latent constructs represented in the original variables.

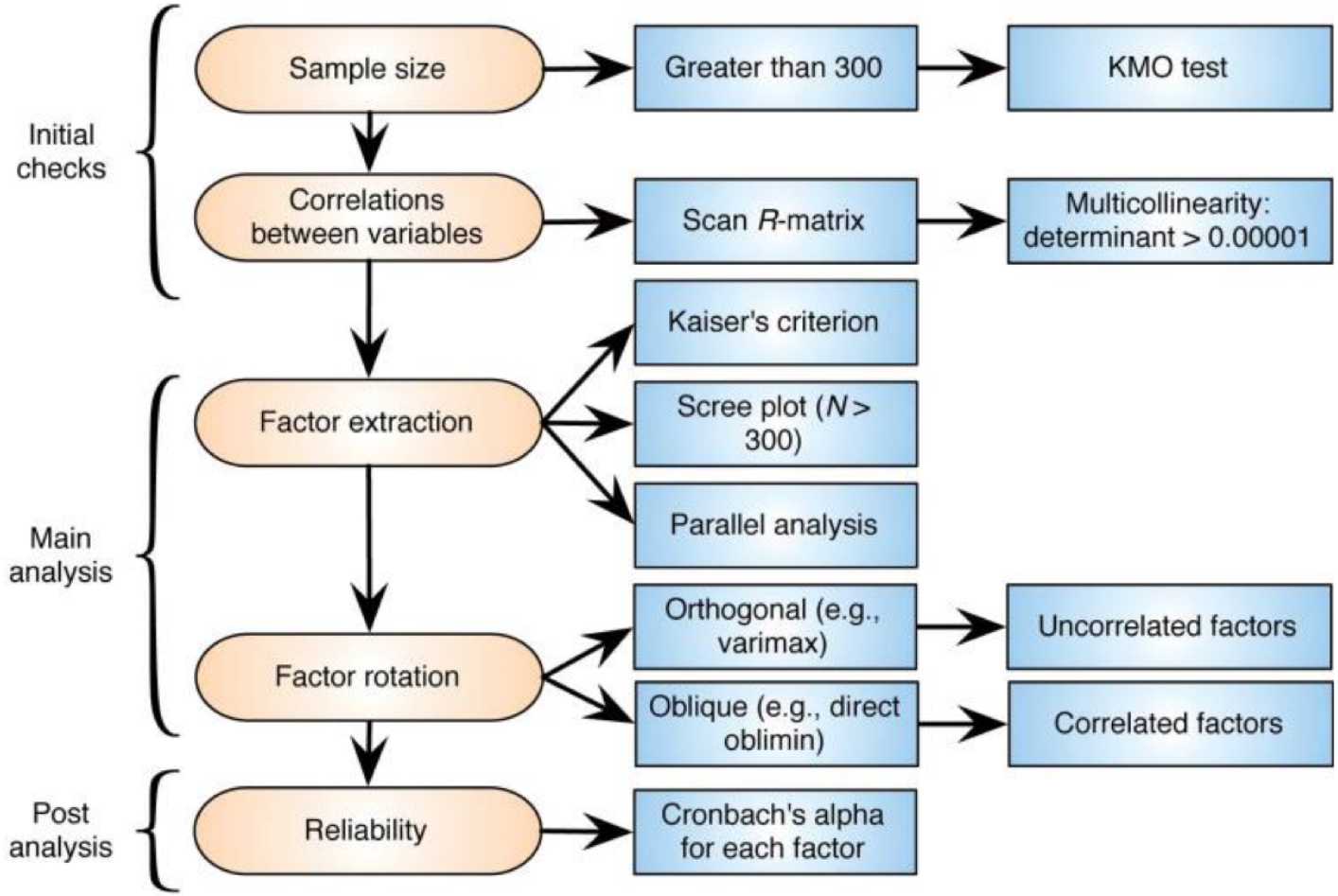
Factor Analysis vs Principal Component Analysis

- Results from PCA and EFA may be similar.
  - When number of variables exceeds 30 or
  - Communalities exceed .6 for most variables.
- Confirmatory factor analysis.
  - Not an exploratory technique.
  - Requires hypotheses regarding factors and variables.

PCA vs EFA

- Similarities
  - Correlated observed variables.
  - Sample size requirements.
  - Measures of adequacy.
  - Analysis of factors requires decisions regarding factor retention, rotation and interpretation.
- Differences
  - Variance
    - PCA Models total variance.
    - EFA Models common variance only.
  - What is modelled?
    - PCA Models factors in terms of observed variables.
    - EFA Models observed variables in terms of factors.
  - Objective
    - PCA Objective is typically data reduction.
    - EFA Objective is typically to identify latent factors.

Procedure for factor analysis & PCA



Lab Exercise

- Run the code from this lecture and try to reproduce all steps.