# Q-STEP R ‘HOW TO’ GUIDES:

## ROBUSTNESS CHECKS WITH SURVEY RESEARCH 2: PRINCIPAL COMPONENTS ANALYSIS

Creator: Dr James Weinberg

In most social science disciplines, quantitative researchers will work with survey research to develop, build or test theories. In order for such survey research to be as rigorous as possible, it is important that we conduct robustness checks on our results in order to assess the validity of our theoretical claims. Therefore, this guide (the second in a series of 3) focuses on principal components analysis with questionnaire items.

In this guide, you will be given a simple contextual description of principal components analysis and when/why it should be used, as well as an example worked through in Rstudio. This guide assumes a basic competency in R from the start - for example users should already be comfortable with assigning and calling objects.

The example used in this guide is based on a dataset of people’s Basic Human Values. Basic values are a personality characteristic that can be measured by psychometric surveys. In this instance, a 20 item questionnaire was administered to 107 people, with two items each tapping one of the ten basic values in the theory. These ten values can be clustered further into 4 higher order values. This guide will use principal components analysis to assess the reliability of the questionnaire items used to test this theory.

Q-Step
A step-change in quantitative social science skills
Funded by the Nuffield Foundation, ESRC and HEFCE
Nuffield Foundation

Sheffield Methods Institute.

The University Of Sheffield.

# Principal Components Analysis

## What is it and when to use it?

Principal component analysis (PCA) is a method of data reduction or compression that is used to turn a conceivably large dataset of (potentially) correlated variables (or questionnaire items) into a smaller number of uncorrelated variables known as principal components.

PCA is performed on a square symmetric matrix. This can be a SSCP matrix (pure sums of squares and cross products), Covariance matrix (scaled sums of squares and cross products), or Correlation matrix (sums of squares and cross products from standardized data). You should only use a correlation matrix if the variances differ substantially across indiviudal items or they are measured in different units.

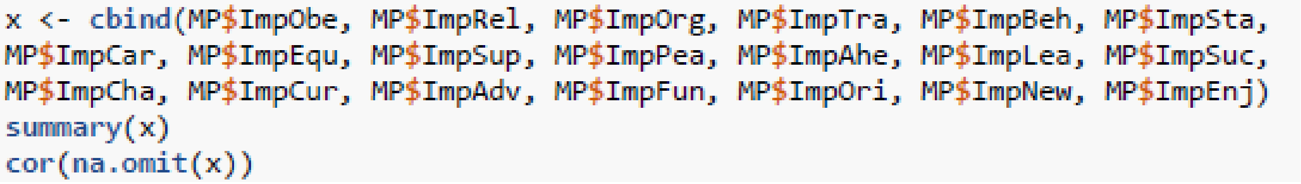
PCA will reduce the dataset to a series of principal components. The first principal component accounts for as much variabality in the data as possible, and each successive component accounts for as much of the remaining variability as possible. You should use PCA to determine the fewest possible dimensions you can statistically use in further analysis of your data.

### Example:

Start by setting your working directory and reading the data file containing your questionnaire responses. You can conduct PCA using the “psych” package.

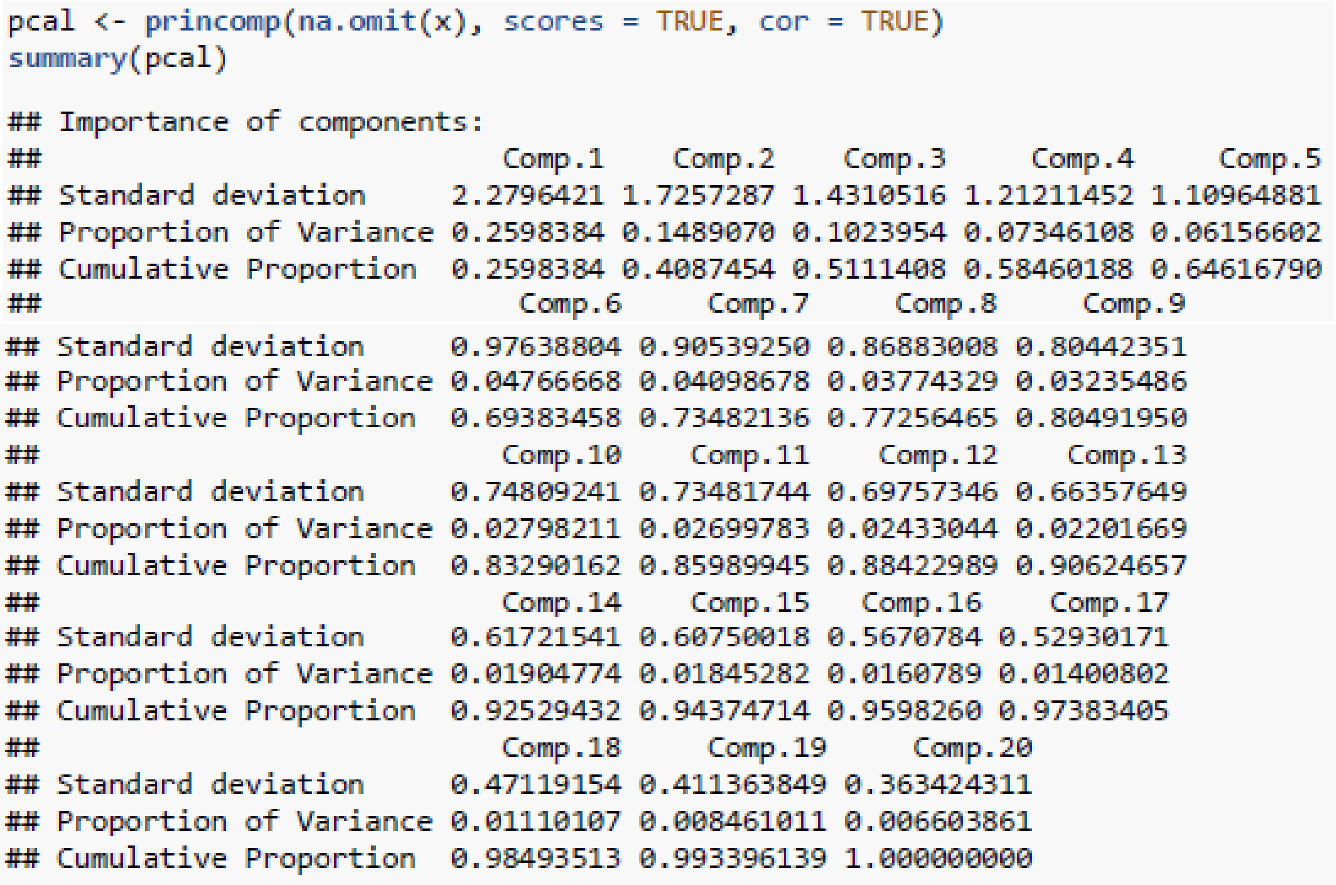
It is likely that your survey contains a lot more items than you need for this analysis (i.e. socio-demographic data or another item battery). For example, in my dataset I have 165 variables but for the purpose of the current test, I am only interested in the 20 questions related to respondents’ basic values. Therefore, you need to isolate these data as a new matrix in your global environment (top right panel in Rstudio).

I am going to organise my survey battery for basic values into a new matrix (x) using the cbind function. Once you have created this, you can use the summary() function to check the descriptive statistics for each item. You can also use the cor() function check the inter-item correlations before you conduct PCA. Your correlations should be higher between items that you think tap the same latent (unobservable) constructs. If you have any missing values in your dataset, remember to remove these beforehand or use the na.omit() function to tell R that they should be ignored.



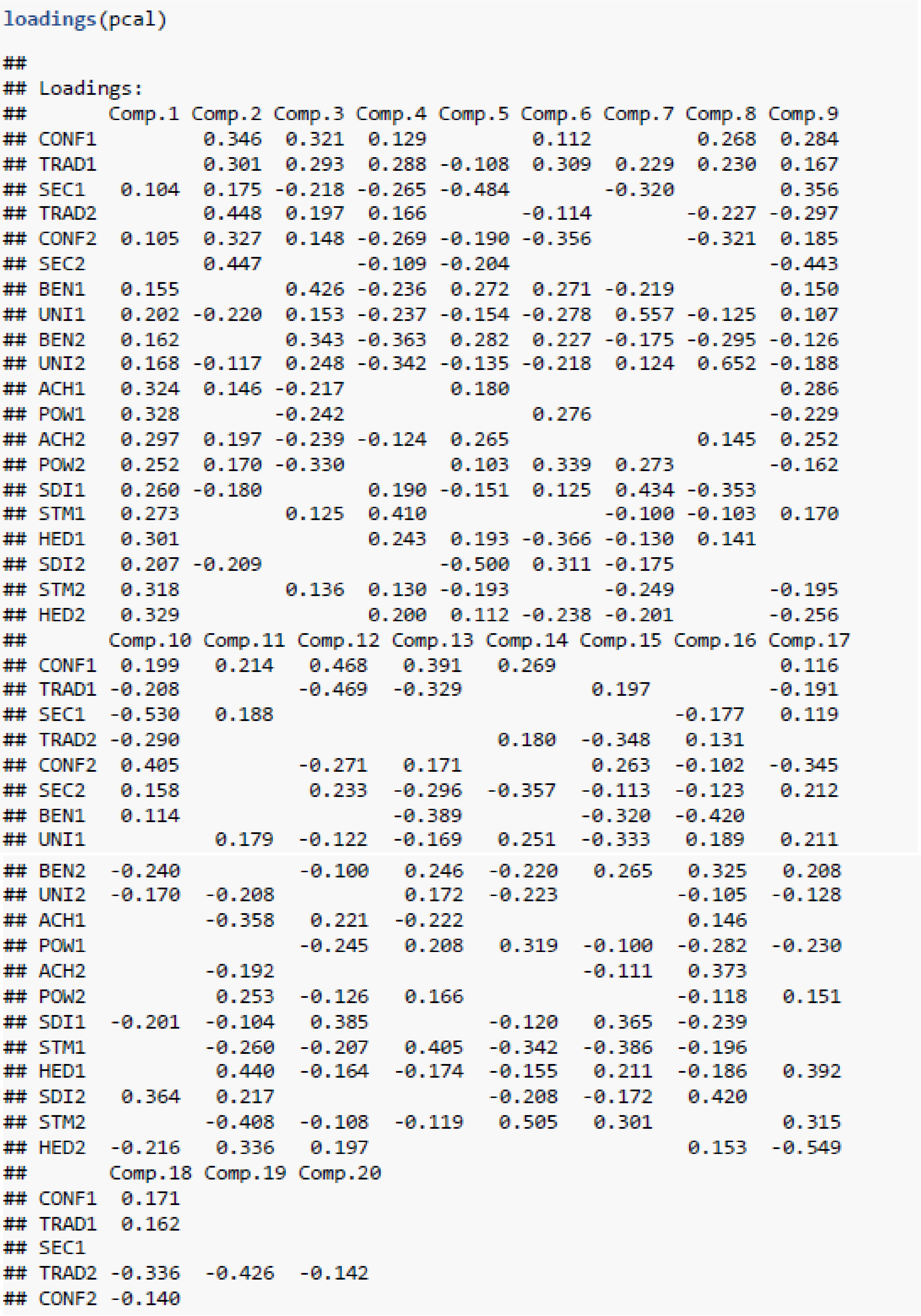
```
x <- cbind(MP$ImpObe, MP$ImpRel, MP$ImpOrg, MP$ImpTra, MP$ImpBeh, MP$ImpSta,
MP$ImpCar, MP$ImpEqu, MP$ImpSup, MP$ImpPea, MP$ImpAhe, MP$ImpLea, MP$ImpSuc,
MP$ImpCha, MP$ImpCur, MP$ImpAdv, MP$ImpFun, MP$ImpOri, MP$ImpNew, MP$ImpEnj)
summary(x)
cor(na.omit(x))
```

Once you've re-organised your data, you are ready to conduct PCA. You can do this using the 'princomp' function in R. Remember to omit any missing values in your dataset. Here I have used 'scores = TRUE' to tell R that the score on each principal component should be calculated. The 'cor' argument is used to indicate whether the calculation should use the correlation matrix or the covariance matrix. Once you have executed the command, your can use the summary() and loadings() commands to check your results.



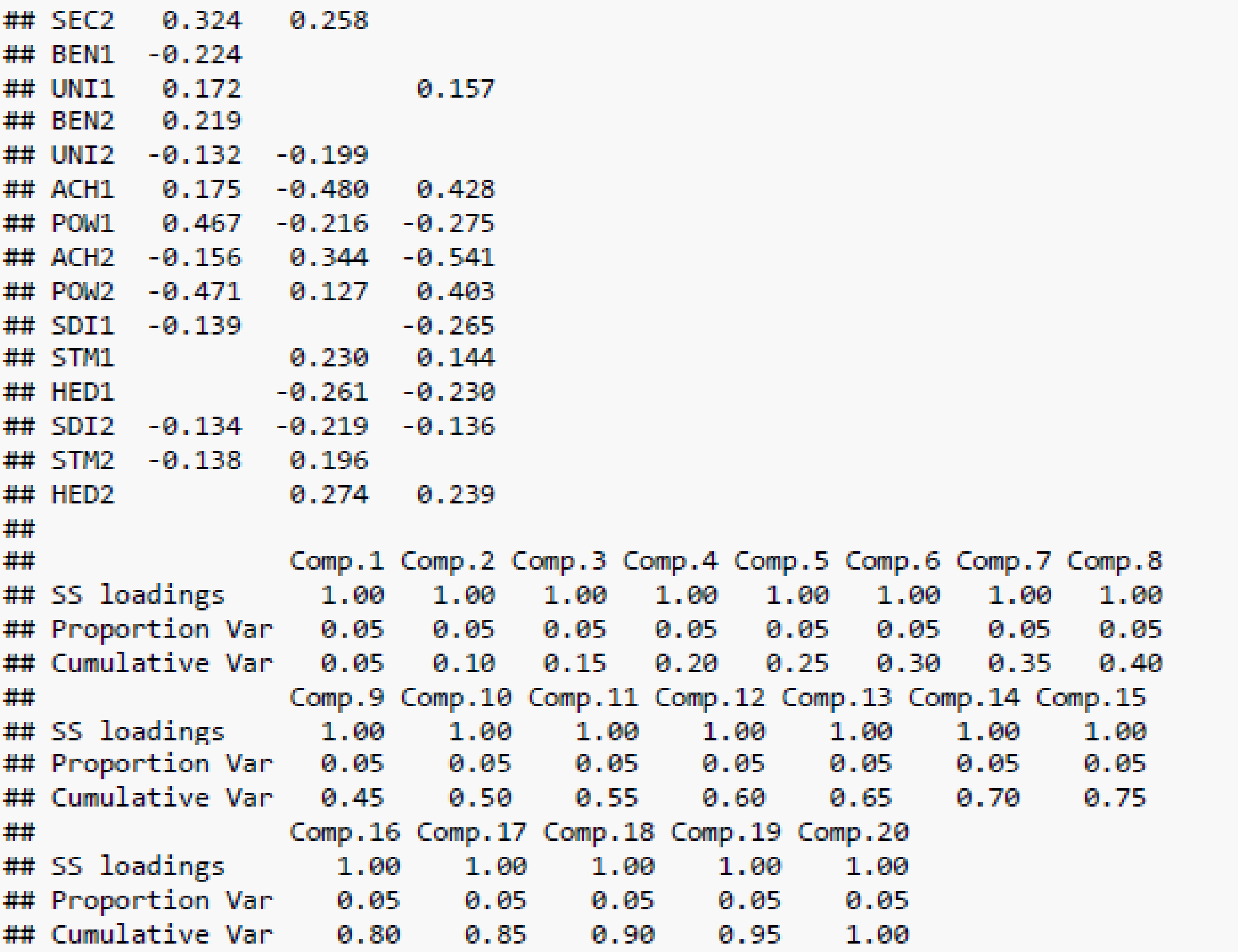
```
pcal <- princomp(na.omit(x), scores = TRUE, cor = TRUE)
summary(pcal)

## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.2796421 1.7257287 1.4310516 1.21211452 1.10964881
## Proportion of Variance 0.2598384 0.1489070 0.1023954 0.07346108 0.06156602
## Cumulative Proportion  0.2598384 0.4087454 0.5111408 0.58460188 0.64616790
##                            Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     0.97638804 0.90539250 0.86883008 0.80442351
## Proportion of Variance 0.04766668 0.04098678 0.03774329 0.03235486
## Cumulative Proportion  0.69383458 0.73482136 0.77256465 0.80491950
##                           Comp.10    Comp.11    Comp.12    Comp.13
## Standard deviation     0.74809241 0.73481744 0.69757346 0.66357649
## Proportion of Variance 0.02798211 0.02699783 0.02433044 0.02201669
## Cumulative Proportion  0.83290162 0.85989945 0.88422989 0.90624657
##                           Comp.14    Comp.15   Comp.16    Comp.17
## Standard deviation     0.61721541 0.60750018 0.5670784 0.52930171
## Proportion of Variance 0.01904774 0.01845282 0.0160789 0.01400802
## Cumulative Proportion  0.92529432 0.94374714 0.9598260 0.97383405
##                           Comp.18     Comp.19     Comp.20
## Standard deviation     0.47119154 0.411363849 0.363424311
## Proportion of Variance 0.01110107 0.008461011 0.006603861
## Cumulative Proportion  0.98493513 0.993396139 1.000000000
```



```
loadings(pcal)

##
## Loadings:
##       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## CONF1         0.346  0.321  0.129         0.112         0.268  0.284
## TRAD1         0.301  0.293  0.288 -0.108  0.309  0.229  0.230  0.167
## SEC1   0.104  0.175 -0.218 -0.265 -0.484        -0.320         0.356
## TRAD2         0.448  0.197  0.166        -0.114        -0.227 -0.297
## CONF2  0.105  0.327  0.148 -0.269 -0.190 -0.356        -0.321  0.185
## SEC2          0.447        -0.109 -0.204                      -0.443
## BEN1   0.155         0.426 -0.236  0.272  0.271 -0.219         0.150
## UNI1   0.202 -0.220  0.153 -0.237 -0.154 -0.278  0.557 -0.125  0.107
## BEN2   0.162         0.343 -0.363  0.282  0.227 -0.175 -0.295 -0.126
## UNI2   0.168 -0.117  0.248 -0.342 -0.135 -0.218  0.124  0.652 -0.188
## ACH1   0.324  0.146 -0.217         0.180                       0.286
## POW1   0.328        -0.242                0.276               -0.229
## ACH2   0.297  0.197 -0.239 -0.124  0.265                0.145  0.252
## POW2   0.252  0.170 -0.330         0.103  0.339  0.273        -0.162
## SDI1   0.260 -0.180         0.190 -0.151  0.125  0.434 -0.353
## STM1   0.273         0.125  0.410               -0.100 -0.103  0.170
## HED1   0.301                0.243  0.193 -0.366 -0.130  0.141
## SDI2   0.207 -0.209               -0.500  0.311 -0.175
## STM2   0.318         0.136  0.130 -0.193        -0.249        -0.195
## HED2   0.329                0.200  0.112 -0.238 -0.201        -0.256
##       Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17
## CONF1  0.199   0.214   0.468   0.391   0.269                   0.116
## TRAD1 -0.208          -0.469  -0.329           0.197          -0.191
## SEC1  -0.530   0.188                                  -0.177   0.119
## TRAD2 -0.290                           0.180  -0.348   0.131
## CONF2  0.405          -0.271   0.171           0.263  -0.102  -0.345
## SEC2   0.158           0.233  -0.296  -0.357  -0.113  -0.123   0.212
## BEN1   0.114                  -0.389          -0.320  -0.420
## UNI1           0.179  -0.122  -0.169   0.251  -0.333   0.189   0.211
## BEN2  -0.240          -0.100   0.246  -0.220   0.265   0.325   0.208
## UNI2  -0.170  -0.208           0.172  -0.223          -0.105  -0.128
## ACH1          -0.358   0.221  -0.222                   0.146
## POW1                  -0.245   0.208   0.319  -0.100  -0.282  -0.230
## ACH2          -0.192                          -0.111   0.373
## POW2           0.253  -0.126   0.166                  -0.118   0.151
## SDI1  -0.201  -0.104   0.385          -0.120   0.365  -0.239
## STM1          -0.260  -0.207   0.405  -0.342  -0.386  -0.196
## HED1           0.440  -0.164  -0.174  -0.155   0.211  -0.186   0.392
## SDI2   0.364   0.217                  -0.208  -0.172   0.420
## STM2          -0.408  -0.108  -0.119   0.505   0.301           0.315
## HED2  -0.216   0.336   0.197                           0.153  -0.549
##       Comp.18 Comp.19 Comp.20
## CONF1  0.171
## TRAD1  0.162
## SEC1
## TRAD2 -0.336  -0.426  -0.142
## CONF2 -0.140
```

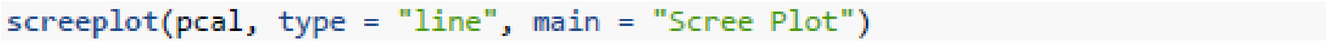


```
## SEC2   0.324   0.258
## BEN1  -0.224
## UNI1   0.172           0.157
## BEN2   0.219
## UNI2  -0.132  -0.199
## ACH1   0.175  -0.480   0.428
## POW1   0.467  -0.216  -0.275
## ACH2  -0.156   0.344  -0.541
## POW2  -0.471   0.127   0.403
## SDI1  -0.139          -0.265
## STM1           0.230   0.144
## HED1          -0.261  -0.230
## SDI2  -0.134  -0.219  -0.136
## STM2  -0.138   0.196
## HED2           0.274   0.239
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings      1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
## Proportion Var   0.05   0.05   0.05   0.05   0.05   0.05   0.05   0.05
## Cumulative Var   0.05   0.10   0.15   0.20   0.25   0.30   0.35   0.40
##                Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15
## SS loadings      1.00    1.00    1.00    1.00    1.00    1.00    1.00
## Proportion Var   0.05    0.05    0.05    0.05    0.05    0.05    0.05
## Cumulative Var   0.45    0.50    0.55    0.60    0.65    0.70    0.75
##                Comp.16 Comp.17 Comp.18 Comp.19 Comp.20
## SS loadings       1.00    1.00    1.00    1.00    1.00
## Proportion Var    0.05    0.05    0.05    0.05    0.05
## Cumulative Var    0.80    0.85    0.90    0.95    1.00
```

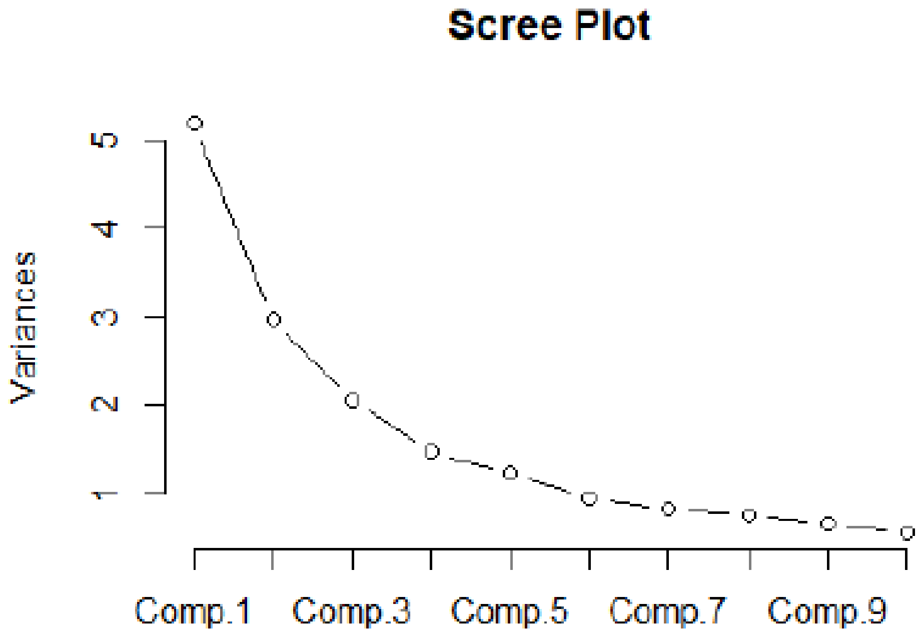
You should obtain as many principal components as you had columns in your data matrix. Each component explains a percentage of the variation in the dataset. For example, in this dataset we can see that the first component exaplins 26% of the total variance, which means that roughly a quarter of the information in the dataset (20 variables) can be contained by that one principal component. Comp.2 and Comp.3 account for a further 15% and 10% of the remaining variance respectively, meaning that the first three components contain roughly 50% of the information in the entire set of variables. Your second set of outputs, produced by the loadings() command, shows you the correlation coefficients between the variables (rows) and factors (columns).

### Visualising PCA

Your next step is to plot the results of your PCA. The best place to start is with a screeplot. A Scree Plot is a simple line segment plot that shows the fraction of total variance in the data as explained or represented by each component.

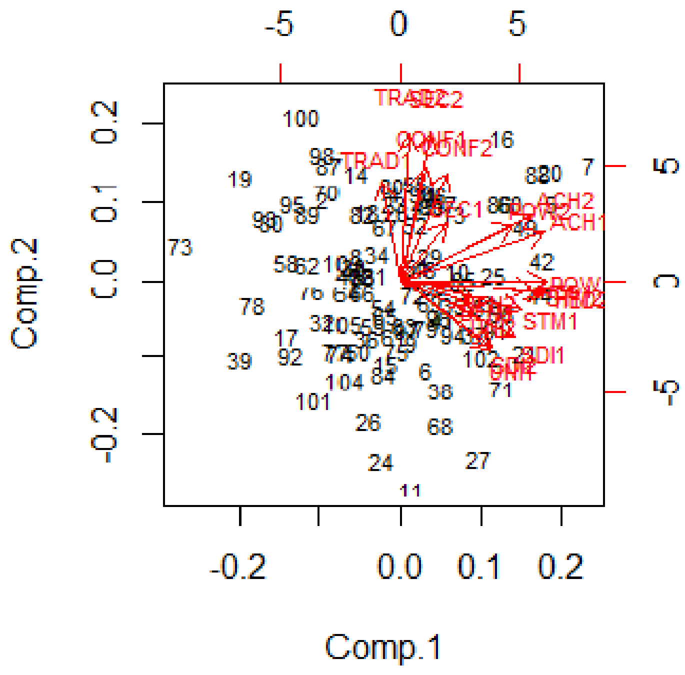


```
screeplot(pca1, type = "line", main = "Scree Plot")
```



If you want to explore your results in more detail, you can use a biplot. A biplot is a type of plot that will allow you to visualize how the sample (participants in this case) relate to one another in our PCA and will reveal how each variable (questionnaire items in this case) contributes to each principal component. You can do this simply using the biplot() function.

BIPLOT of Questionnaire Items on Basic Human Values (n – 107)



The axes in the biplot orginate from the center point and the arrows show the ways in which each variable (questionnaire item in this instance) contribute to the first two principal components. In this example you can see that variables TRAD1, TRAD2, CONF1, CONF2, SEC1 AND SEC2 all contribute to component 2, with participants (represented as numbers on the biplot) who scored highly for these variables moving to the top on this plot. Given that all of these questionnaire items are theoretically supposed to measure the underlying value Conservation, this biplot suggests that this might be the case. If you are working with anonymous data, then it is best to keep your rownames as numbers for the purpose of reporting these analyses in your written work.