Project 6: Principal Components Analysis

DA 410

Marjorie Blanco

```
#load data
probe <- read_table("book_data/T3_6_PROBE.DAT",
                    col_names = c("Subject","y1", "y2", "y3", "y4", "y5"))
```

```
kable(probe)  %>%
  kable_styling(bootstrap_options = "striped")
```

Subject	y1	y2	y3	y4	y5
1	51	36	50	35	42
2	27	20	26	17	27
3	37	22	41	37	30
4	42	36	32	34	27
5	27	18	33	14	29
6	43	32	43	35	40
7	41	22	36	25	38
8	38	21	31	20	16
9	36	23	27	25	28
10	26	31	31	32	36
11	29	20	25	26	25

Do a principle component analysis of the data in Table 3.6 (page 79)

Use R to solve this part (built-in function).

Make sure you include the commands and outputs, as well as the interpretations of the outputs.

```
# apply PCA
probe.pca <- prcomp(probe,
                    center = TRUE,
                    scale. = TRUE)
probe.pca
```

```
## Standard deviations (1, .., p=5):
## [1] 1.8483758 0.7838567 0.7564879 0.5207797 0.3543867
##
## Rotation (n x k) = (5 x 5):
##           PC1        PC2        PC3        PC4        PC5
## y1 -0.4418394  0.2006104 -0.6786078  0.2125365 -0.5087760
## y2 -0.4535595  0.4280646  0.3491277  0.6055405  0.3499642
## y3 -0.4727808 -0.3678765 -0.3754368 -0.2581448  0.6584479
## y4 -0.4536224  0.3934629  0.3345386 -0.7010073 -0.1899641
## y5 -0.4120276 -0.6974023  0.4058723  0.1734903 -0.3860467
```

```r
res.eig
```

```
## eigen() decomposition
## $values
## [1] 3.4164933 0.6144313 0.5722740 0.2712115 0.1255899
##
## $vectors
##            [,1]       [,2]       [,3]       [,4]       [,5]
## [1,] -0.4418394 -0.2006104  0.6786078  0.2125365  0.5087760
## [2,] -0.4535595 -0.4280646 -0.3491277  0.6055405 -0.3499642
## [3,] -0.4727808  0.3678765  0.3754368 -0.2581448 -0.6584479
## [4,] -0.4536224 -0.3934629 -0.3345386 -0.7010073  0.1899641
## [5,] -0.4120276  0.6974023 -0.4058723  0.1734903  0.3860467
```

The first PC was weakly negative correlated with variable y3, y4, y2, y1 and y5.

The second PC was strongly correlated with variables y5 (negative) and weakly correlated with variable y2, y4, y3 (negative), y4 and y1.

```r
probe.pca.summary <- summary(probe.pca)
probe.pca.summary
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.8484 0.7839 0.7565 0.52078 0.35439
## Proportion of Variance 0.6833 0.1229 0.1144 0.05424 0.02512
## Cumulative Proportion  0.6833 0.8062 0.9206 0.97488 1.00000
```

The result contain 5 principal components (PC1-5). Each of these explains a percentage of the total variation in the dataset. PC1 explains 68.3% of the total variance, which means that nearly two-thirds of the information in the dataset (5 variables) can be encapsulated by just that one Principal Component. PC2 explains 12.3% of the variance. PC3 explains 11.4% of the variance. PC4 explains 5.4% of the variance. PC5 explains 2.5% of the variance. PC1 and PC2 can explain explains 80.6% of the variance. Based on this it, we can forget about PC3 through PC5.

The sum of the eigenvalues from a correlation matrix will equal the number of variables. The sum of the eigenvalues will equal 5.

The first principal component accounts for the most variable variance (0.6833 / 1 = 68.3%) with the remaining components in lesser and lesser amounts.

Only PC1 contains SS loadings > 1 (Kaiser's criterion).

The first two principal components are:

Z_l = -0.4418394 y_1 + -0.4535595 y_2 + -0.4727808 y_3 + -0.4536224 y_4 + -0.4120276 y_5

Z_2 = -0.2006104 y_1 + -0.4280646 y_2 + 0.3678765 y_3 + -0.3934629 y_4 + 0.6974023 y_5

There are 5 components with descending eigenvalues (3.4164933, 0.6144313, 0.572274, 0.2712115, 0.1255899). The sum of these eigenvalues is equal to the sum of the variable variances in the variance-covariance matrix (S). The sum of the eigenvalues for the five principal components is 5, which is referred to as the trace of a matrix. The sum of the variable variances indicates the total amount of variance that is available to be partitioned across the 5 principal components.

Decide how many components to retain. Show your reason.

For the probe data set, I recomend to retain 2 components

Method 1: % of variance

An appropriate threshold percentage should be selected prior to starting the process. If we want to explain at least 70% of variance then we would select PC1 and PC2.

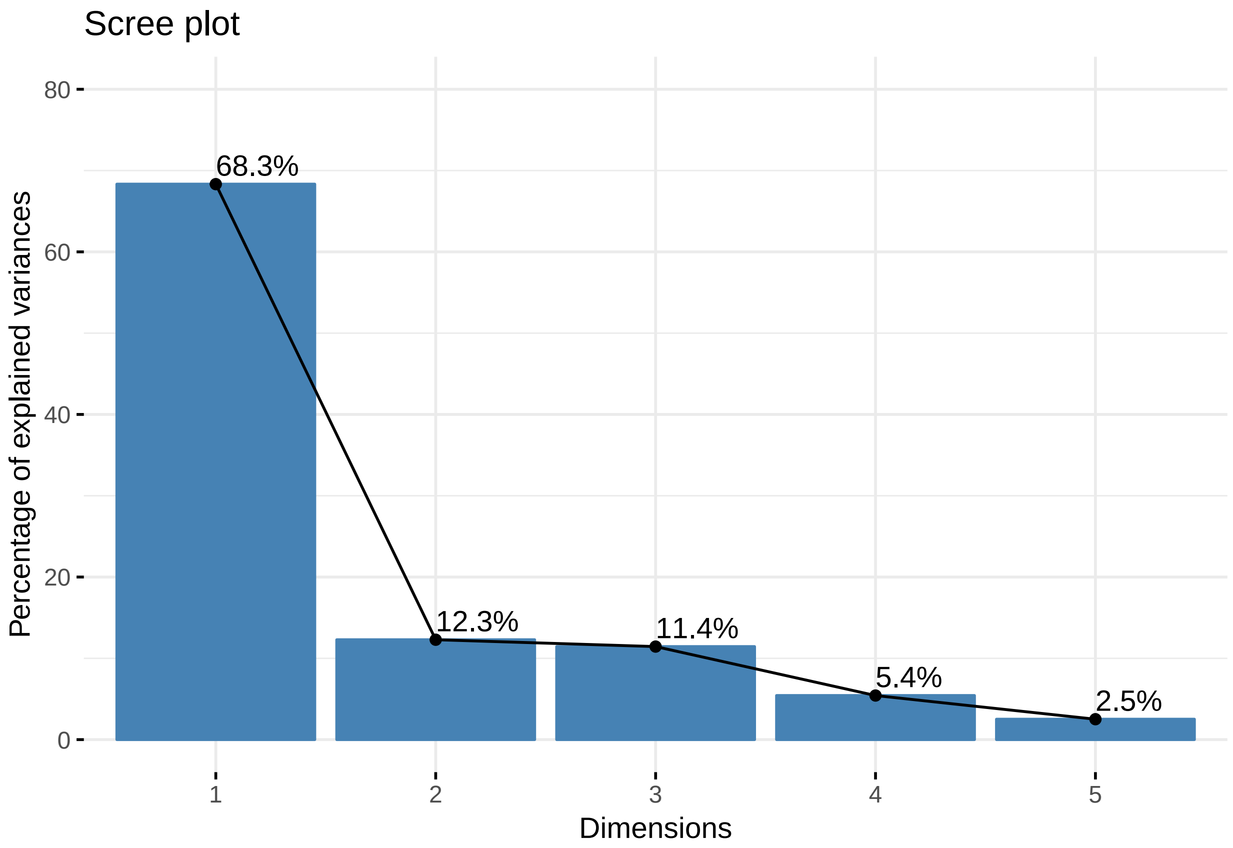
Method 2: Kaiser's criterion

Components with SS loadings > 1. In the probe data, retaining only PC1 is recomended. The SS loading for PC2 is < 1. Retaining PC2 is not recomended.

Method 3: Scree plot

The number of points after point of inflexion. For this plot, retaining PC1 and PC2 is recomended.

```r
fviz_eig(probe.pca, addlabels = TRUE, ylim = c(0, 80))
```



The scree plot suggest to keep 2 principal components.

Method 4: Significance of the "larger" components

Test H_{ok}: $\gamma_{p-k+1} = \ldots = \gamma_p$ using a likelihood ratio approach. Reject H_0 if $u \geq \chi^2_{\alpha,v}$

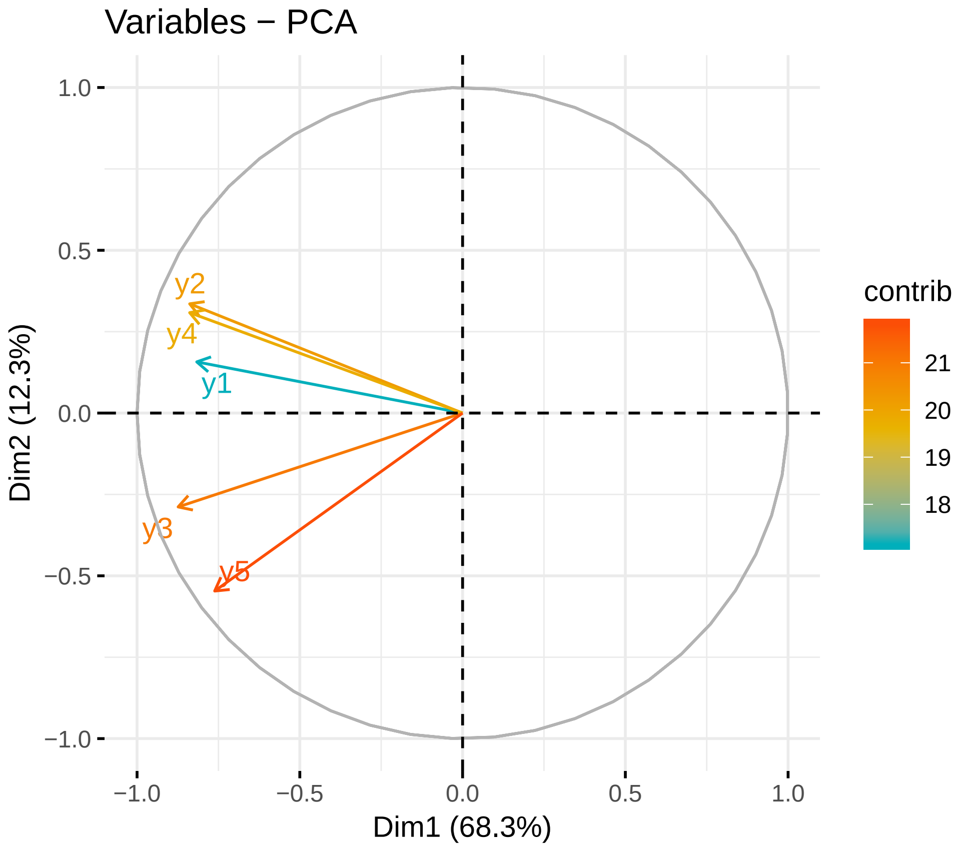
```
##      Eigenvalue k          u df Crit Value
## [1,]  3.4164933 5 23.971279 14  23.684791
## [2,]  0.6144313 4  5.386313  9  16.918978
## [3,]  0.5722740 3  4.107388  5  11.070498
## [4,]  0.2712115 2  1.084925  2   5.991465
## [5,]  0.1255899 1  0.000000  0   0.000000
```

H_{02}: $\gamma_{p-1} = \gamma_p$ $\gamma_4 = \gamma_5$ Reject null hypothesis since $u \geq \chi^2_{\alpha,v}$

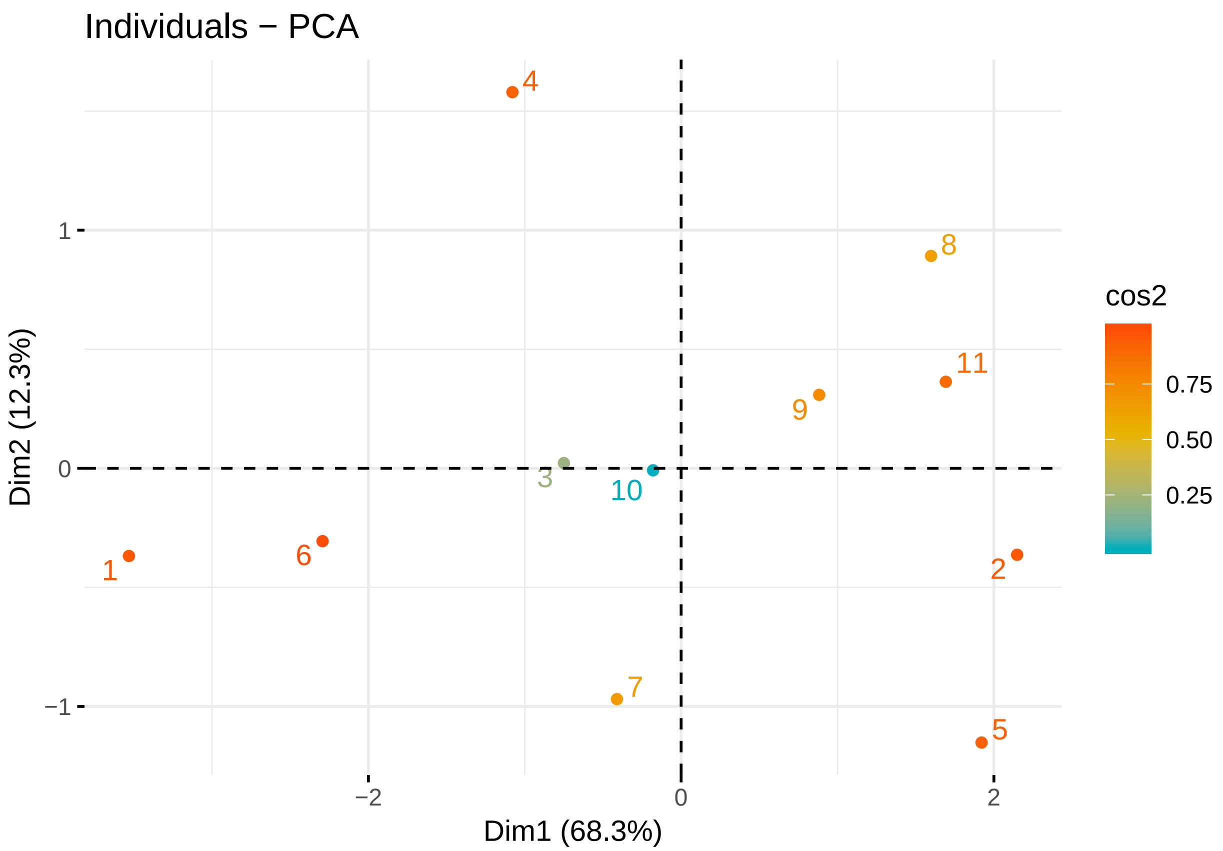
H_{03}: $\gamma_{p-2} = \gamma_{p-1}$ $\gamma_3 = \gamma_2$ Fail to reject null hypothesis since $u < \chi^2_{\alpha,v}$

The tests indicate that only the last four (population) eigenvalues are equal and we should retain PC1

Method 1, 2, and 4 recomend that only PC1 is retained. I would only include PC2 if the minimum explained variance % is selected to be at least 80%. Selecting PC3 should be avoided as it explain 92% and this could indicate possible overfitting. None of the methods recomended PC3-PC5 to be retained.



```r
fviz_pca_ind(probe.pca,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             )
```



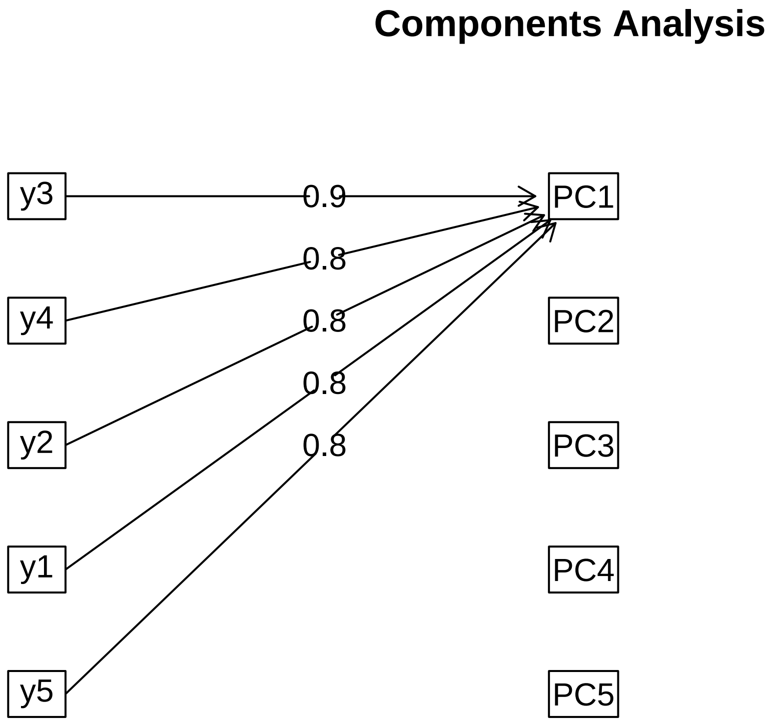
Individuals with a similar profile are grouped together.

```
## Principal Components Analysis
## Call: principal(r = pcacor, nfactors = 5, rotate = "none", scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##     PC1   PC2   PC3   PC4   PC5 h2      u2 com
## y1 0.82 -0.16 -0.51  0.11  0.18  1 0.0e+00 1.9
## y2 0.84 -0.34  0.26  0.32 -0.12  1 2.7e-15 1.9
## y3 0.87  0.29 -0.28 -0.13 -0.23  1 2.0e-15 1.7
## y4 0.84 -0.31  0.25 -0.37  0.07  1 1.7e-15 1.9
## y5 0.76  0.55  0.31  0.09  0.14  1 5.6e-16 2.3
##
##                        PC1  PC2  PC3  PC4  PC5
## SS loadings           3.42 0.61 0.57 0.27 0.13
## Proportion Var        0.68 0.12 0.11 0.05 0.03
## Cumulative Var        0.68 0.81 0.92 0.97 1.00
## Proportion Explained  0.68 0.12 0.11 0.05 0.03
## Cumulative Proportion 0.68 0.81 0.92 0.97 1.00
##
## Mean item complexity =  1.9
## Test of the hypothesis that 5 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0
##
## Fit based upon off diagonal values = 1

##
```

```
## Factor analysis with Call: principal(r = pcacor, nfactors = 5, rotate = "none", scores = TRUE)
##
## Test of the hypothesis that 5 factors are sufficient.
## The degrees of freedom for the model is -5  and the objective function was  0
##
## The root mean square of the residuals (RMSA) is  0
```

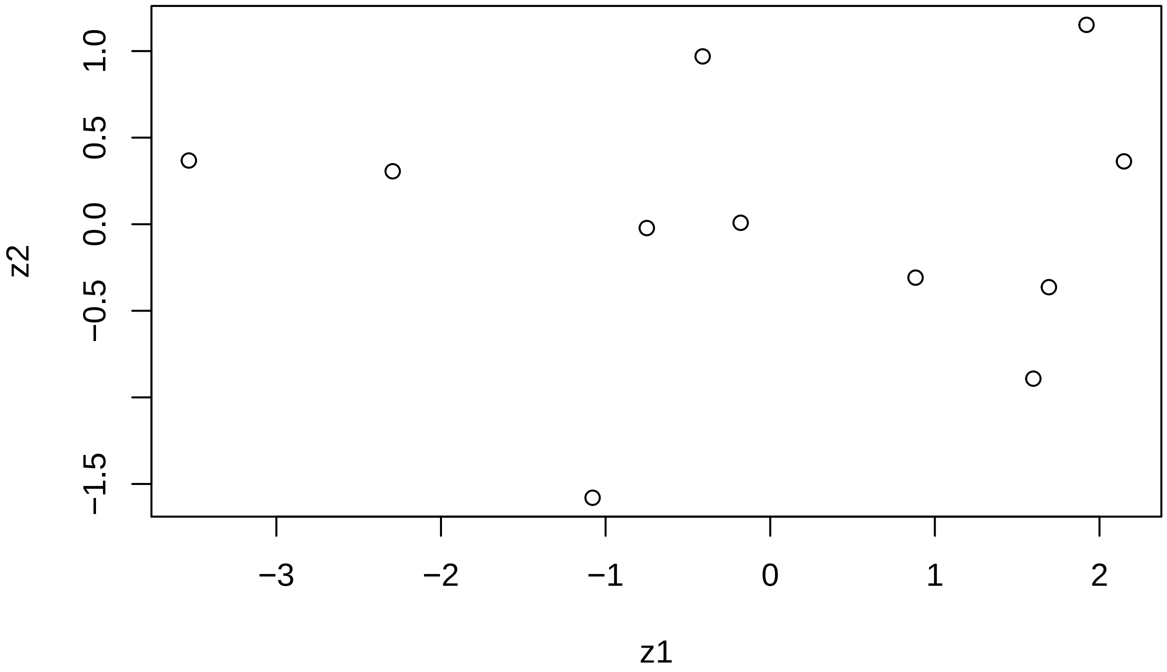
```
##
## Reliability analysis
## Call: alpha(x = pcacor)
##
##   raw_alpha std.alpha G6(smc) average_r S/N median_r
##       0.88      0.88     0.9       0.6 7.6     0.59
##
##  Reliability if an item is dropped:
##    raw_alpha std.alpha G6(smc) average_r S/N  var.r med.r
## y1      0.86      0.86    0.85      0.61 6.3 0.0083  0.58
## y2      0.85      0.85    0.85      0.59 5.9 0.0149  0.59
## y3      0.84      0.84    0.81      0.57 5.3 0.0123  0.56
## y4      0.85      0.85    0.87      0.60 5.9 0.0147  0.58
## y5      0.88      0.88    0.87      0.64 7.2 0.0081  0.61
##
##  Item statistics
##       r r.cor r.drop
## y1 0.81  0.78   0.70
## y2 0.84  0.79   0.74
## y3 0.87  0.86   0.79
## y4 0.84  0.78   0.74
## y5 0.77  0.71   0.64
```



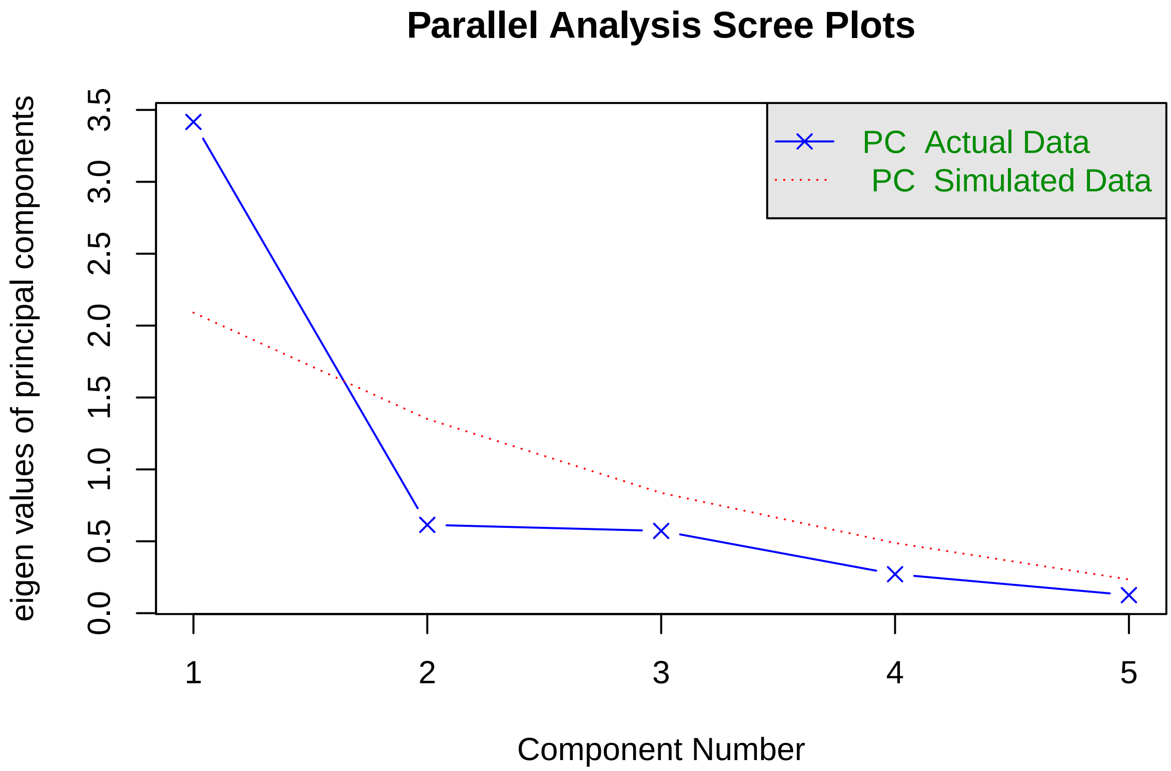
A component will summarize the five variable relations and yield 76% of the variable variance. The principal component equation to generate the scores is computed using the first set of weights.

First two principal components for the probe data

```
plot(as.matrix(probe.scaled) %*% res.eig$vectors[,1:2], xlab = "z1", ylab = "z2")
```



```
fa.parallel(probe.scaled, n.obs = nrow(probe.scaled), fm = "pa", fa = "pc")
```



```
## Parallel analysis suggests that the number of factors =  NA  and the number of components =  1
```

From this plot, we can see that the first component has the largest eigen value.

Check assumptions

The KMO test

```
## $KMO
## [1] 0.6914
##
## $MSA
##        MSA
## y1 0.65900
## y2 0.70881
## y3 0.64428
## y4 0.80664
## y5 0.65915
##
## $Bartlett
## [1] 23.971
##
## $Communalities
##    Initial Communalities Final Extraction
## y1                0.68304          0.58306
## y2                0.65501          0.62681
## y3                0.76358          0.72565
## y4                0.61642          0.62917
## y5                0.58793          0.46773
##
## $Factor.Loadings
##       [,1]
## y1 0.76358
## y2 0.79172
## y3 0.85185
## y4 0.79320
## y5 0.68391
##
## $RMS
## [1] 0.085206
```

The KMO test is close to 1 (KMO = 0.6914), so we would conclude that n = 11 with 5 variables is an adequate sample size. Bartlett = 23.97128), but requires another function to test for significance.

Test significance of the Bartlett test using correlation matrix

```
## $chisq
## [1] 23.971
##
## $p.value
## [1] 0.007677
##
## $df
## [1] 10
```

The Bartlett chi-square = 23.97128, df = 10, p = 0.00768 is significant indicating correlations in matrix are significant.

```
## [1] 0.040919
```

The determinant is positive. The three assumptions for conducting a principal components analysis have been met.