# Principal Components I

Mike Strube

September 10, 2018

## 1 Preliminaries

***In this section, the RStudio workspace and console panes are cleared of old output, variables, and other miscellaneous debris. Packages are loaded and any required data files are retrieved.***

```
options(replace.assign = TRUE, width = 65, digits = 4, scipen = 4, fig.width = 4,
    fig.height = 4)
# Clear the workspace and console.
rm(list = ls(all = TRUE))
cat("\f")
```

```
# Turn off showing of significance asterisks.
options(show.signif.stars = F)
# Set the contrast option; important for ANOVAs.
options(contrasts = c("contr.sum", "contr.poly"))
how_long <- Sys.time()
set.seed(123)
library(knitr)
```

### 1.1 Packages

```
library(psych)
library(ggplot2)

##
## Attaching package:  'ggplot2'
## The following objects are masked from 'package:psych':
##
##    %+%, alpha

library(factoextra)

## Warning:  package 'factoextra' was built under R version 3.5.1
## Welcome!  Related Books:  `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

library(FactoMineR)

## Warning:  package 'FactoMineR' was built under R version 3.5.1

library(reshape2)
library(GGally)
```

### 1.2 Data Files

***The data can be in standard wide format for principal component analyses. There are missing data in the original file. To keep things simple, the few cases with any missing data will be excluded from analyses (known as listwise deletion). This insures that the correlation matrix will not contain any impossible values (a problem that could arise with pairwise deletion). Alternative strategies, such as multiple imputation, would ordinarily be explored as well.***

```
# Get the data from the working directory.
setwd("C:\\Courses\\Psychology 516\\PowerPoint\\2018")
NC <- read.table("need_for_cognition.csv", sep = ",", header = TRUE)
NC <- as.data.frame(NC)
NC <- na.omit(NC)


# Reverse score items
NC$item_1 <- 6 - NC$item_1
NC$item_2 <- 6 - NC$item_2
NC$item_6 <- 6 - NC$item_6
NC$item_10 <- 6 - NC$item_10
NC$item_11 <- 6 - NC$item_11
NC$item_13 <- 6 - NC$item_13
NC$item_14 <- 6 - NC$item_14
NC$item_15 <- 6 - NC$item_15
NC$item_18 <- 6 - NC$item_18

NC$total <- NC$item_1 + NC$item_2 + NC$item_3 + NC$item_4 + NC$item_5 +
    NC$item_6 + NC$item_7 + NC$item_8 + NC$item_9 + NC$item_10 + NC$item_11 +
    NC$item_12 + NC$item_13 + NC$item_14 + NC$item_15 + NC$item_16 +
    NC$item_17 + NC$item_18

NC_long <- melt(NC, id.vars = "total")
NC_long <- as.data.frame(NC_long)
```

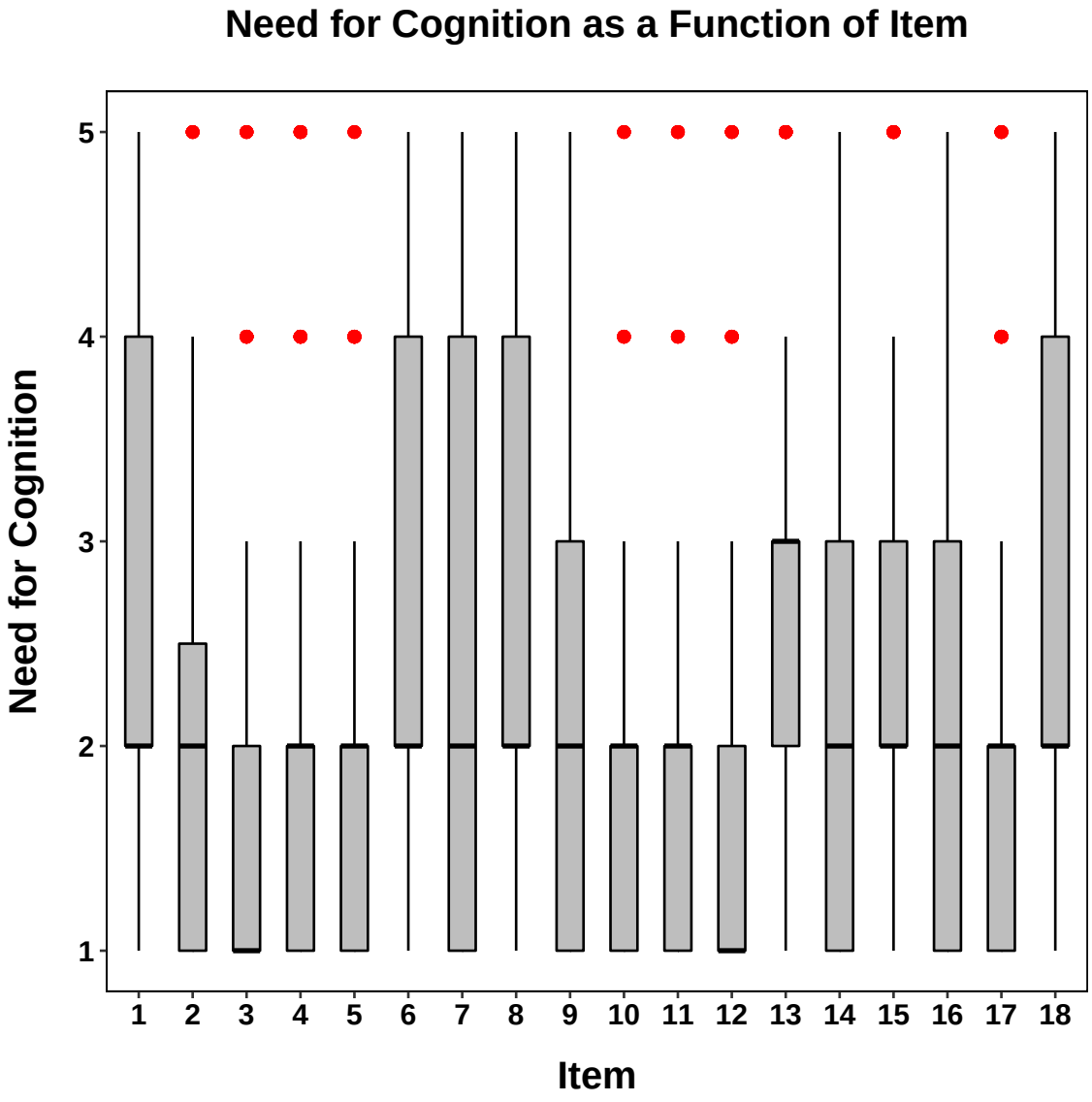
## 2 Graphical Display of Data

### 2.1 Item Scores

***Here are some boxplots of the items. The crude level of measurement is apparent for items. This is an important feature of the data. Variables can correlate more highly when their distributions have similar shapes. For this reason, and others, it is often better to conduct analyses on composite scores rather than individual items.***

```
NC_long$item <- factor(NC_long$variable, levels = c("item_1", "item_2",
    "item_3", "item_4", "item_5", "item_6", "item_7", "item_8", "item_9",
    "item_10", "item_11", "item_12", "item_13", "item_14", "item_15",
    "item_16", "item_17", "item_18"), labels = c("1", "2", "3", "4",
    "5", "6", "7", "8", "9", "10", "11", "12", "13", "14", "15", "16",
    "17", "18"))
ggplot(NC_long, aes(y = value, x = item)) + geom_boxplot(aes(y = value,
    x = item), color = "black", size = 0.5, width = 0.5, fill = "grey",
    outlier.colour = "red", outlier.shape = 19, outlier.size = 2,
```

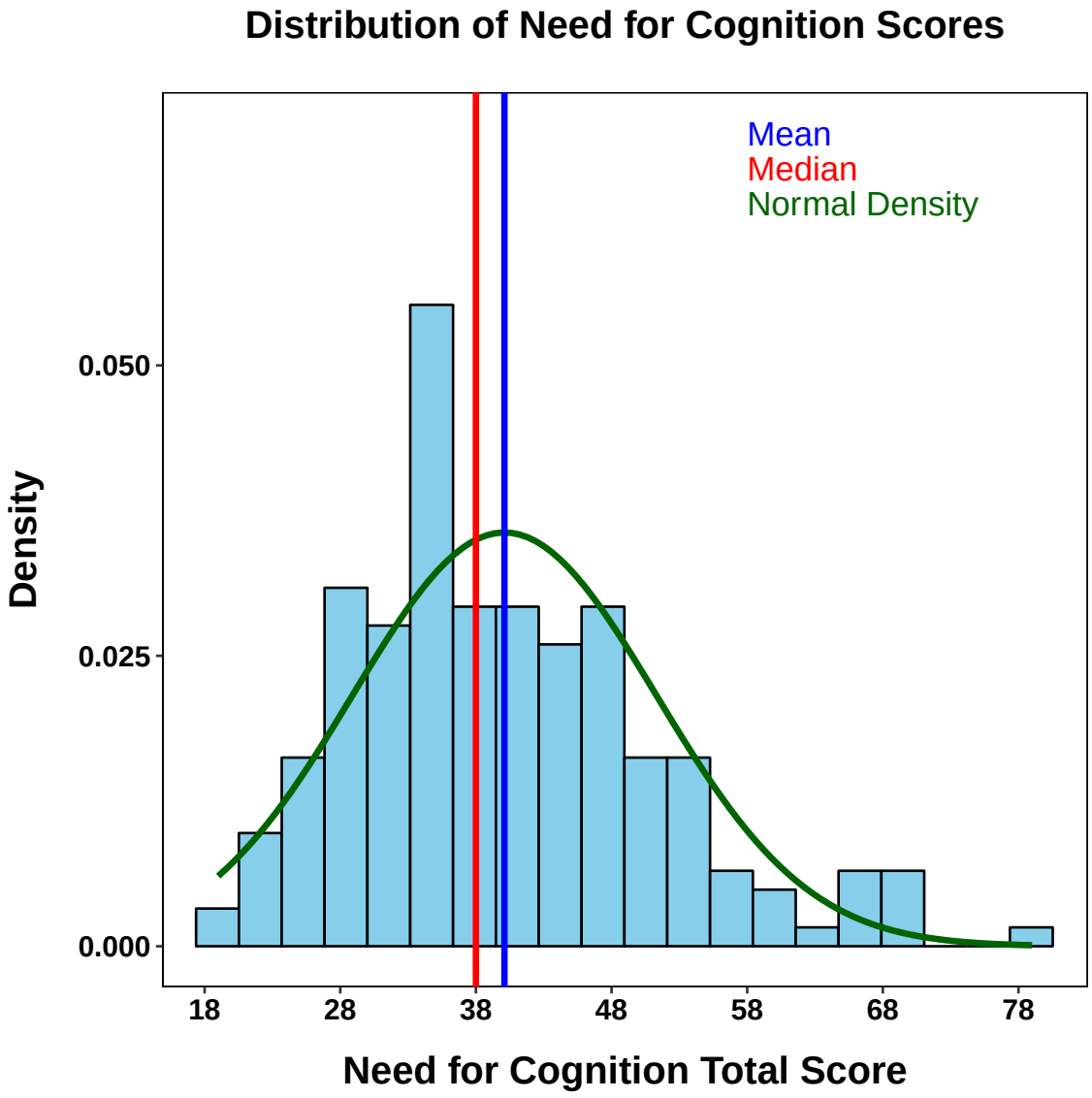
```
    notch = FALSE) + ylab("Need for Cognition") + xlab("Item") + theme(text = element_text(size = 14,
    family = "sans", color = "black", face = "bold"), axis.text.x = element_text(colour = "black",
    size = 12, face = "bold"), axis.text.y = element_text(colour = "black",
    size = 12, face = "bold"), axis.title.x = element_text(margin = margin(15,
    0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
    15, 0, 0), size = 16), axis.line.x = element_blank(), axis.line.y = element_blank(),
    plot.title = element_text(size = 16, face = "bold", margin = margin(0,
        0, 20, 0), hjust = 0.5), panel.background = element_rect(fill = "white",
        linetype = 1, color = "black"), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), plot.background = element_rect(fill = "white"),
    plot.margin = unit(c(1, 1, 1, 1), "cm")) + ggtitle("Need for Cognition as a Function of Item")
```



## 2.2 Total NC Score

***Here is a histogram for the total NC scores. As the number of scale levels increases, score distributions have a greater opportunity to approach normal (or whatever parent distribution underlies them).***

```
ggplot(NC, aes(x = total)) + geom_histogram(aes(y = ..density..),
    color = "black", fill = "skyblue", size = 0.5, na.rm = TRUE, bins = 20) +
    stat_function(fun = dnorm, args = list(mean = mean(NC$total, na.rm = TRUE),
        sd = sd(NC$total, na.rm = TRUE)), size = 1.25, color = "darkgreen") +
    coord_cartesian(xlim = c(18, 80), ylim = c(0, 0.07)) + scale_x_continuous(breaks = c(seq(18,
    80, 10))) + scale_y_continuous(breaks = seq(0, 0.07, 0.025)) +
    xlab("Need for Cognition Total Score") + ylab("Density") + theme(text = element_text(size = 14,
    family = "sans", color = "black", face = "bold"), axis.text.y = element_text(colour = "black",
    size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
    size = 12, face = "bold", angle = 0), axis.title.x = element_text(margin = margin(15,
    0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
    15, 0, 0), size = 16), axis.line.x = element_blank(), axis.line.y = element_blank(),
    plot.title = element_text(size = 16, face = "bold", margin = margin(0,
        0, 20, 0), hjust = 0.5), panel.background = element_rect(fill = "white",
        linetype = 1, color = "black"), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), plot.background = element_rect(fill = "white"),
    plot.margin = unit(c(1, 1, 1, 1), "cm"), legend.position = "bottom",
    legend.title = element_blank()) + geom_vline(xintercept = mean(NC$total),
    size = 1.25, color = "blue") + geom_vline(xintercept = median(NC$total),
    size = 1.25, color = "red") + annotate("text", x = 58, y = 0.07,
    label = "Mean", color = "blue", size = 5, hjust = 0) + annotate("text",
    x = 58, y = 0.067, label = "Median", color = "red", size = 5,
    hjust = 0) + annotate("text", x = 58, y = 0.064, label = "Normal Density",
    color = "darkgreen", size = 5, hjust = 0) + ggtitle("Distribution of Need for Cognition Scores")
```




## 3 Intercorrelations

*The PCA will attempt to summarize the data by finding linear combinations that contain most of the information in the original variables. The linear combinations correspond to items that correlate highly. This simplification can sometimes be viewed in the correlation matrix of the original variables, although as the number of variables increases, examination of correlations gets challenging.*

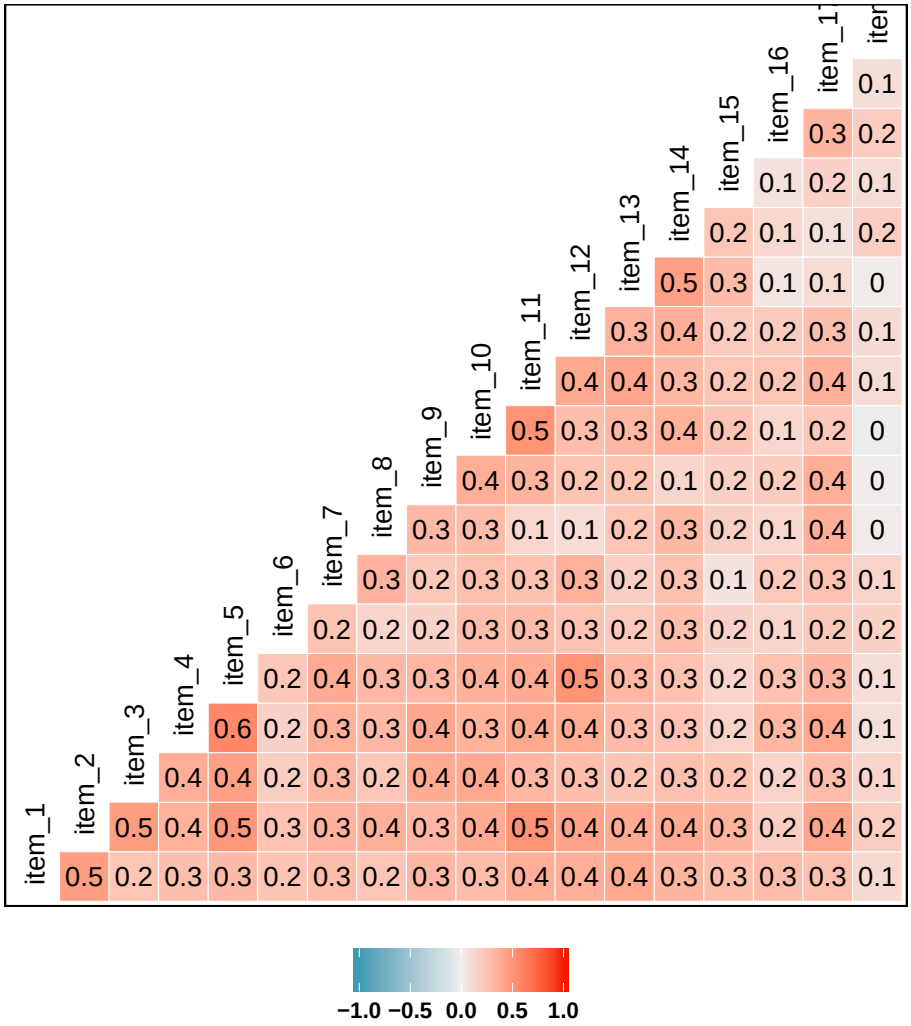
```
R <- cor(NC[1:18], use = "complete.obs")
round(R, 2)

##         item_1 item_2 item_3 item_4 item_5 item_6 item_7 item_8
## item_1    1.00   0.46   0.24   0.27   0.30   0.24   0.29   0.24
```

```
## item_2    0.46   1.00   0.46   0.35   0.50   0.26   0.34   0.36
## item_3    0.24   0.46   1.00   0.37   0.45   0.19   0.32   0.22
## item_4    0.27   0.35   0.37   1.00   0.57   0.17   0.35   0.31
## item_5    0.30   0.50   0.45   0.57   1.00   0.22   0.39   0.31
## item_6    0.24   0.26   0.19   0.17   0.22   1.00   0.25   0.15
## item_7    0.29   0.34   0.32   0.35   0.39   0.25   1.00   0.34
## item_8    0.24   0.36   0.22   0.31   0.31   0.15   0.34   1.00
## item_9    0.32   0.32   0.39   0.41   0.32   0.17   0.22   0.31
## item_10   0.29   0.38   0.40   0.34   0.35   0.28   0.27   0.29
## item_11   0.36   0.51   0.32   0.40   0.39   0.29   0.26   0.14
## item_12   0.36   0.43   0.28   0.38   0.52   0.25   0.35   0.10
## item_13   0.39   0.39   0.24   0.29   0.32   0.21   0.18   0.24
## item_14   0.33   0.39   0.27   0.26   0.25   0.28   0.26   0.31
## item_15   0.28   0.33   0.25   0.16   0.15   0.17   0.07   0.18
## item_16   0.28   0.19   0.16   0.32   0.26   0.15   0.21   0.14
## item_17   0.27   0.42   0.28   0.40   0.33   0.22   0.25   0.36
## item_18   0.12   0.19   0.14   0.09   0.09   0.17   0.15   0.03
##         item_9 item_10 item_11 item_12 item_13 item_14 item_15
## item_1    0.32    0.29    0.36    0.36    0.39    0.33    0.28
## item_2    0.32    0.38    0.51    0.43    0.39    0.39    0.33
## item_3    0.39    0.40    0.32    0.28    0.24    0.27    0.25
## item_4    0.41    0.34    0.40    0.38    0.29    0.26    0.16
## item_5    0.32    0.35    0.39    0.52    0.32    0.25    0.15
## item_6    0.17    0.28    0.29    0.25    0.21    0.28    0.17
## item_7    0.22    0.27    0.26    0.35    0.18    0.26    0.07
## item_8    0.31    0.29    0.14    0.10    0.24    0.31    0.18
## item_9    1.00    0.37    0.33    0.25    0.24    0.15    0.18
## item_10   0.37    1.00    0.52    0.28    0.32    0.35    0.24
## item_11   0.33    0.52    1.00    0.39    0.41    0.31    0.24
## item_12   0.25    0.28    0.39    1.00    0.34    0.36    0.20
## item_13   0.24    0.32    0.41    0.34    1.00    0.45    0.29
## item_14   0.15    0.35    0.31    0.36    0.45    1.00    0.23
## item_15   0.18    0.24    0.24    0.20    0.29    0.23    1.00
## item_16   0.21    0.14    0.24    0.20    0.05    0.13    0.07
## item_17   0.37    0.23    0.36    0.28    0.11    0.08    0.17
## item_18   0.01   -0.02    0.11    0.11    0.02    0.18    0.10
##         item_16 item_17 item_18
## item_1     0.28    0.27    0.12
## item_2     0.19    0.42    0.19
## item_3     0.16    0.28    0.14
## item_4     0.32    0.40    0.09
## item_5     0.26    0.33    0.09
## item_6     0.15    0.22    0.17
## item_7     0.21    0.25    0.15
## item_8     0.14    0.36    0.03
## item_9     0.21    0.37    0.01
## item_10    0.14    0.23   -0.02
## item_11    0.24    0.36    0.11
## item_12    0.20    0.28    0.11
## item_13    0.05    0.11    0.02
## item_14    0.13    0.08    0.18
## item_15    0.07    0.17    0.10
## item_16    1.00    0.33    0.20
## item_17    0.33    1.00    0.09
```

```
## item_18    0.20    0.09    1.00
```

***A heat map for the correlation matrix can help ease the pain of examining a large correlation matrix. Patterns in the data, if present, are easier to detect. The following heat map suggests fairly homogeneous correlations indicative of a single principal component.***

```
ggcorr(NC[, 1:18], label = TRUE, angle = 90, hjust = 0.1, size = 4) +
    theme(text = element_text(size = 14, family = "sans", color = "black",
        face = "bold"), axis.text.y = element_text(colour = "black",
        size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
        size = 12, face = "bold", angle = 0), axis.title.x = element_text(margin = margin(15,
        0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
        15, 0, 0), size = 16), axis.line.x = element_blank(), axis.line.y = element_blank(),
        plot.title = element_text(size = 16, face = "bold", margin = margin(0,
            0, 20, 0), hjust = 0.5), panel.background = element_rect(fill = "white",
            linetype = 1, color = "black"), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), plot.background = element_rect(fill = "white"),
        plot.margin = unit(c(1, 1, 1, 1), "cm"), legend.position = "bottom",
        legend.title = element_blank()) + ggtitle("Intercorrelations Among NC Items")
```




## 4 Should A PCA Be Conducted?

*Two tests can be used to determine if a PCA should be conducted (generally a good idea if the approach is exploratory). The Kaiser-Meyer-Olkin (KMO) factor adequacy test can range from 0 to 1 and roughly indicates the proportion of variance in the data that might be common factor variance. The KMO test has the following cut-offs for sampling adequacy: .90 and above (undeniable evidence for factorability), .80 to .89 (very strong evidence), .70 to .79 (modest evidence), .60 to .69 (weak evidence), .50 to .59 (very weak evidence), and below .50 (unacceptable for factoring). The Bartlett test for sphericity (not the same as in repeated measures ANOVA) should be highly significant, indicating that the correlation matrix departs noticeably from an identity matrix.*

```r
KMO(R)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = R)
## Overall MSA =  0.87
## MSA for each item =
##  item_1  item_2  item_3  item_4  item_5  item_6  item_7  item_8
##    0.91    0.89    0.88    0.89    0.86    0.91    0.90    0.75
##  item_9 item_10 item_11 item_12 item_13 item_14 item_15 item_16
##    0.89    0.87    0.86    0.87    0.87    0.83    0.89    0.84
## item_17 item_18
##    0.83    0.69
```

```r
cortest.bartlett(R = R, n = length(NC[, 1]))
```

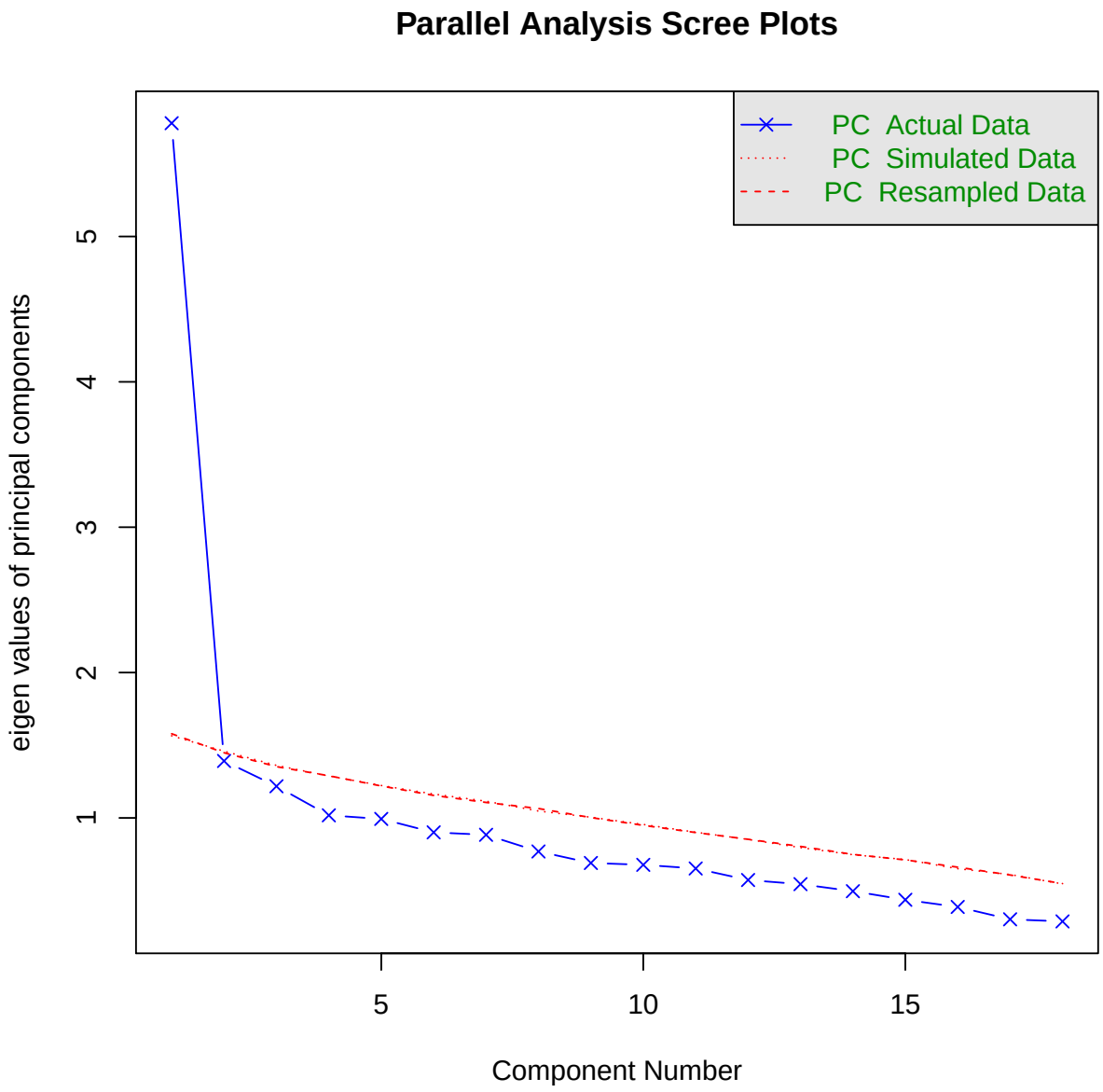
```
## $chisq
## [1] 1024
##
## $p.value
## [1] 8.209e-129
##
## $df
## [1] 153
```

# 5 How Many Components?

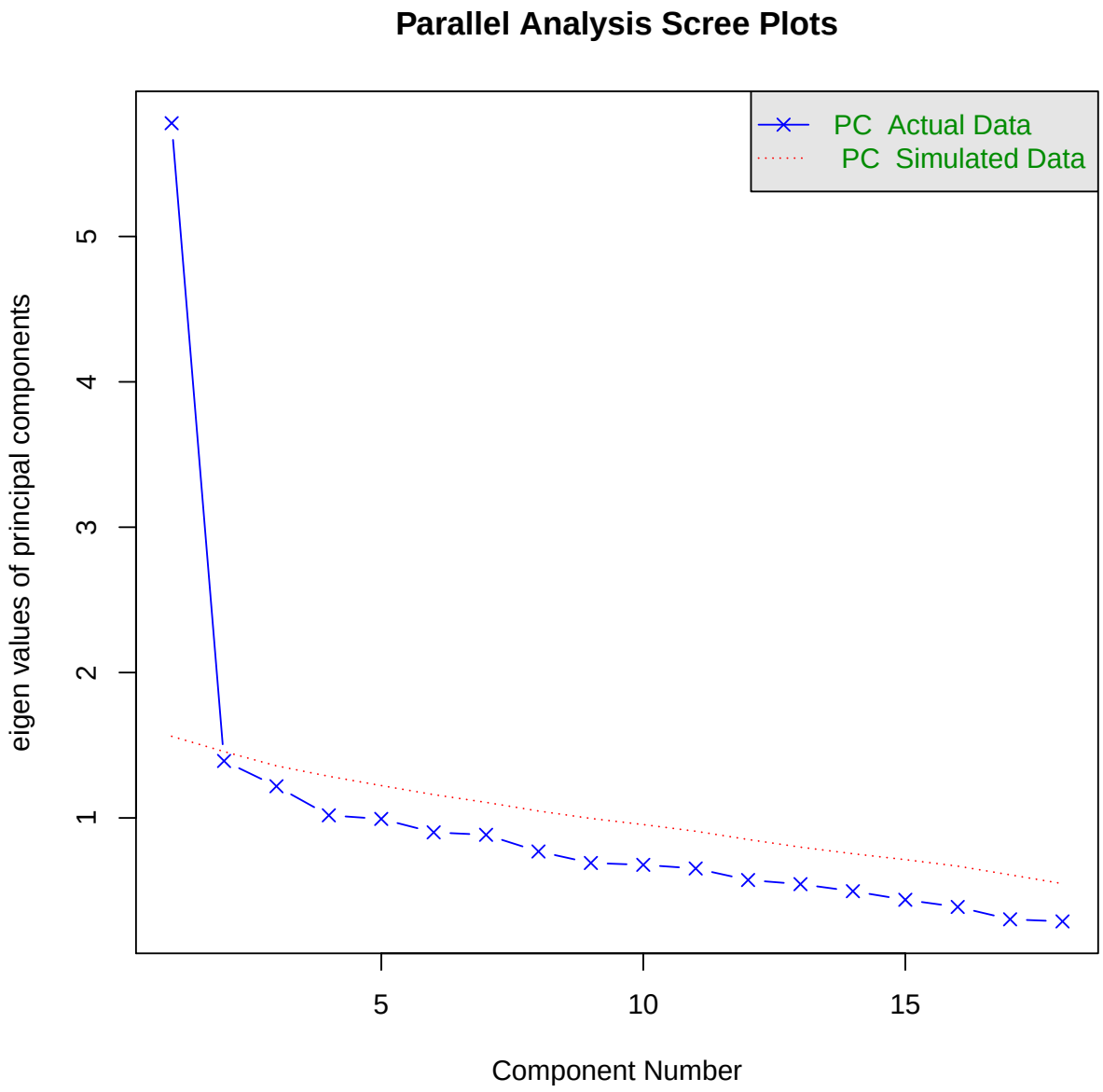
***If the correlation matrix is not singular, then as many components as there are variables or items can be extracted. But, only a few of them are likely to be meaningful or useful. The scree test is the most common way to determine how many components should be extracted. To make sure only meaningful departures from the scree are interpreted, a parallel analysis (Horn's procedure) or random selection of data points can be used.***

***Most of the PCA-related analyses can be conducted on raw data or on correlation matrices. If the latter are used, the number of cases needs to be specified as well. Note that the first option will give simulated and resampled parallel analyses. Analysis of the correlation matrix will not provide resampled results.***

```r
scree <- fa.parallel(NC[, c(1:18)], fa = "pc")
```




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

scree <- fa.parallel(R, n.obs = 195, fa = "pc")
```




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

## 6 PCA

***The scree test indicates a single component is the likely best solution (accounts for non-trivial amount of variance). One is requested here.***

```
PCA_1 <- principal(R, nfactors = 1, rotate = "none", n.obs = 195,
    residuals = TRUE)
PCA_1

## Principal Components Analysis
```

```
## Call: principal(r = R, nfactors = 1, residuals = TRUE, rotate = "none",
##     n.obs = 195)
## Standardized loadings (pattern matrix) based upon correlation matrix
##          PC1    h2   u2 com
## item_1  0.60 0.364 0.64   1
## item_2  0.74 0.553 0.45   1
## item_3  0.61 0.370 0.63   1
## item_4  0.66 0.436 0.56   1
## item_5  0.70 0.483 0.52   1
## item_6  0.44 0.194 0.81   1
## item_7  0.55 0.297 0.70   1
## item_8  0.49 0.245 0.76   1
## item_9  0.56 0.317 0.68   1
## item_10 0.62 0.387 0.61   1
## item_11 0.68 0.459 0.54   1
## item_12 0.62 0.388 0.61   1
## item_13 0.56 0.312 0.69   1
## item_14 0.55 0.305 0.69   1
## item_15 0.41 0.164 0.84   1
## item_16 0.39 0.156 0.84   1
## item_17 0.55 0.304 0.70   1
## item_18 0.21 0.044 0.96   1
##
##                 PC1
## SS loadings    5.78
## Proportion Var 0.32
##
## Mean item complexity =  1
## Test of the hypothesis that 1 component is sufficient.
##
## The root mean square of the residuals (RMSR) is  0.08
##  with the empirical chi square  350.4  with prob <  4.5e-21
##
## Fit based upon off diagonal values = 0.93
```

# 7 Examination of Residuals

***A residual matrix gives the variances in the main diagonal and correlations in the off-diagonals. This can be converted to a correlation matrix, which can then be examined using the KMO and Bartlett tests to determine if additional components should be extracted.***

```
# Create a correlation matrix of the residuals by replacing the
# main diagonal with ones.
R1 <- diag(PCA_1$residual)
R2 <- diag(R1)
R3 <- PCA_1$residual - R2
R4 <- diag(18) + R3

# Assess the factorability of the residual correlation matrix.
KMO(R4)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = R4)
## Overall MSA =  0.35
## MSA for each item =
##  item_1  item_2  item_3  item_4  item_5  item_6  item_7  item_8
##    0.30    0.28    0.27    0.38    0.35    0.28    0.30    0.36
##  item_9 item_10 item_11 item_12 item_13 item_14 item_15 item_16
##    0.34    0.33    0.31    0.31    0.41    0.43    0.34    0.41
## item_17 item_18
##    0.42    0.46

cortest.bartlett(R = R4, n = 195)

## $chisq
## [1] 190.1
##
## $p.value
## [1] 0.02247
##
## $df
## [1] 153
```

***Mixed evidence, but probably not worth additional extraction. The KMO test indicates no additional common component variance is present. Bartlett's test is significant however. In situations like this it is best to mistrust the significance test, which will be increasingly powerful as sample size increases, perhaps identifying trivial evidence for additional components.***