

# Homework 3

## Applied Multivariate Analysis

Emorie Beck

September 22, 2018

## 1 Workspace

### 1.1 Packages

```
library(car)
library(knitr)
library(psych)
library(kableExtra)
library(multcomp)
library(lme4)
library(plyr)
library(tidyverse)
library(MVN)
```

### 1.2 data

The file, Set.3.csv, contains the data from a study in which 500 high school students completed a measure of scholastic aptitude: Grammar, Paragraph Comprehension, Vocabulary, Sentence Completion, Geometry, Algebra, Numerical Puzzles, Series Completion, Practical Problem Solving, Symbol Manipulation, Analytical Ability, and Formal Logic.

```
wd <- "https://github.com/emoriebeck/homeworks/raw/master/multivariate/homeworks/homework3"

dat <- sprintf("%s/Set_3.csv", wd) %>%
  read.csv(., stringsAsFactors = F)

head(dat)

##      ID      Grammar Paragraph_Comprehension Vocabulary Sentence_Completion
## 1  1  2.0298794          0.7009379    0.9224983          0.7783650
## 2  2  1.8460110          0.8176540    1.6230497          0.5595109
## 3  3 -0.5514456          0.1155194   -0.2451959          1.2206362
## 4  4 -1.3804105          0.2193181    0.5195521          0.3530657
## 5  5  0.4384477          1.5177577    0.4692875          1.4074032
## 6  6 -0.5984267         -0.8757810   -0.9889196         -1.4836151
##      Geometry      Algebra Numerical_Puzzles Series_Completion
## 1  0.7169340  0.649042462          0.1797594          0.521331792
## 2 -1.4336680  0.008714271         -0.2517458          0.000110179
## 3 -0.5504154 -0.776083508          0.8131658         -0.802679845
## 4  1.7218792  1.076026142          0.7711456         -0.381686114
```

```
## 5  0.7914582  1.541237112          0.4042484          0.825899485
## 6 -0.5157728 -0.441559349        -1.0049260        -2.612748945
##   Practical_Problem_Solving Symbol_Manipulation Analytical_Ability
## 1          1.3030926          1.3690616          1.5512126
## 2          0.9545397         -0.9592880          1.4883905
## 3         -1.5259042         -1.2038384         -0.7812775
## 4         -0.5231818          0.1203525         -0.3958278
## 5          1.2598039          2.6013012          0.9772288
## 6         -0.1936980         -0.1956773         -0.2486582
##   Formal_Logic
## 1    0.7546186
## 2   -0.3733971
## 3   -1.1192996
## 4    1.7353933
## 5    3.0419200
## 6   -0.8137567
```

Answer the following questions about these data:

## 2 Question 1

What evidence do you have that these data should be subjected to a principal components analysis?

```
R <- dat %>% select(-ID) %>% cor

(KMO1 <- KMO(R))

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = R)
## Overall MSA = 0.82
## MSA for each item =
##           Grammar Paragraph_Comprehension
##           0.83          0.84
##           Vocabulary Sentence_Completion
##           0.83          0.82
##           Geometry Algebra
##           0.82          0.83
##           Numerical_Puzzles Series_Completion
##           0.77          0.84
## Practical_Problem_Solving Symbol_Manipulation
##           0.84          0.81
##           Analytical_Ability Formal_Logic
##           0.82          0.83

(CB_1 <- cortest.bartlett(R=R,n=nrow(dat)))

## $chisq
## [1] 1794.866
##
## $p.value
## [1] 0
##
## $df
## [1] 66
```

The overall MSA is .82, and all but one of the MSA values are .8 (1 (Numerical Puzzles) is .77), which indicates very strong evidence for conducting a PCA.

In addition, the  $\chi^2$  value of the Bartlett test ( $\chi^2(66) = 1794.87$ ), which indicates that the correlation matrix departs significantly from an identity matrix (independence among indicators).

### 3 Question 2

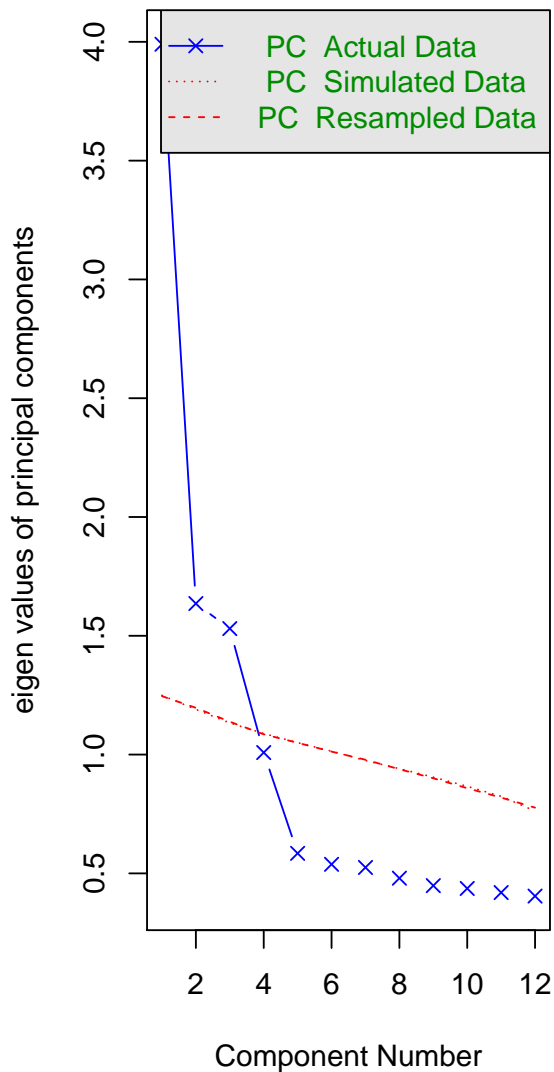
How many principal components should be extracted?

```
par(mfrow=c(1,2))
scree_1 <- fa.parallel(dat %>% select(-ID), fa="pc")

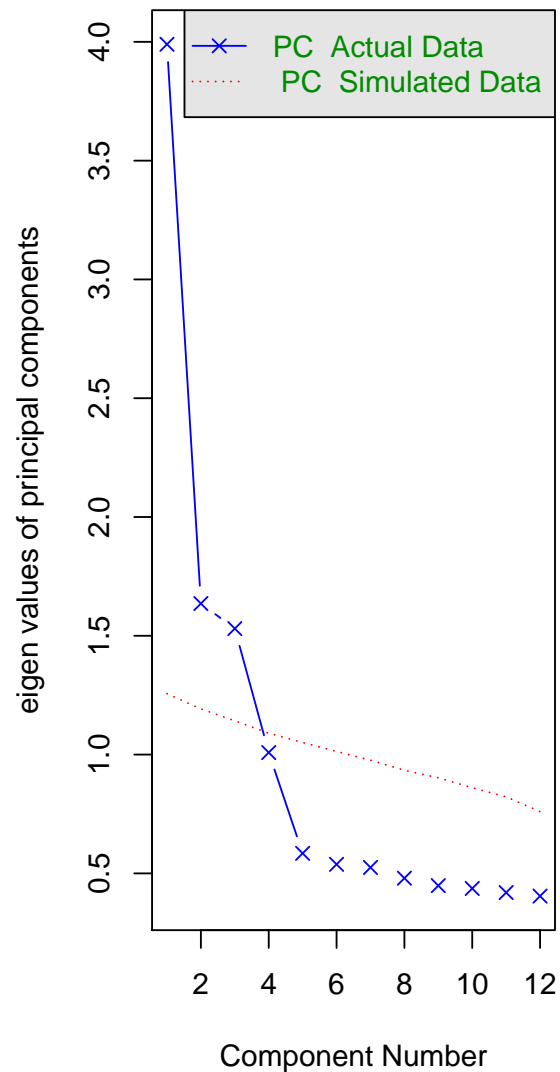
## Parallel analysis suggests that the number of factors = NA and the number of components = 3

scree_2 <- fa.parallel(R, fa = "pc", n.obs = nrow(dat))
```

Parallel Analysis Scree Plots



Parallel Analysis Scree Plots



```
## Parallel analysis suggests that the number of factors = NA and the number of components = 3
```

Parallel analysis suggests that 3 factors should be extracted from the data.

## 4 Question 3

How much variance do these extracted components account for in the original data?

```
pca_1 <- principal(R, nfactors = 3, rotate = "none", n.obs = nrow(dat), residuals = T)

pca_1$Vaccounted %>% data.frame %>% mutate(m = rownames(.)) %>%
  mutate_at(vars(PC1:PC3), funs(round(.,2))) %>%
  select(m, everything()) %>%
```

```
kable(., "latex", booktabs = T, escape = F)
```

m	PC1	PC2	PC3
SS loadings	3.99	1.64	1.53
Proportion Var	0.33	0.14	0.13
Cumulative Var	0.33	0.47	0.60
Proportion Explained	0.56	0.23	0.21
Cumulative Proportion	0.56	0.79	1.00

The three extracted components account for all of the variance.

## 5 Question 4

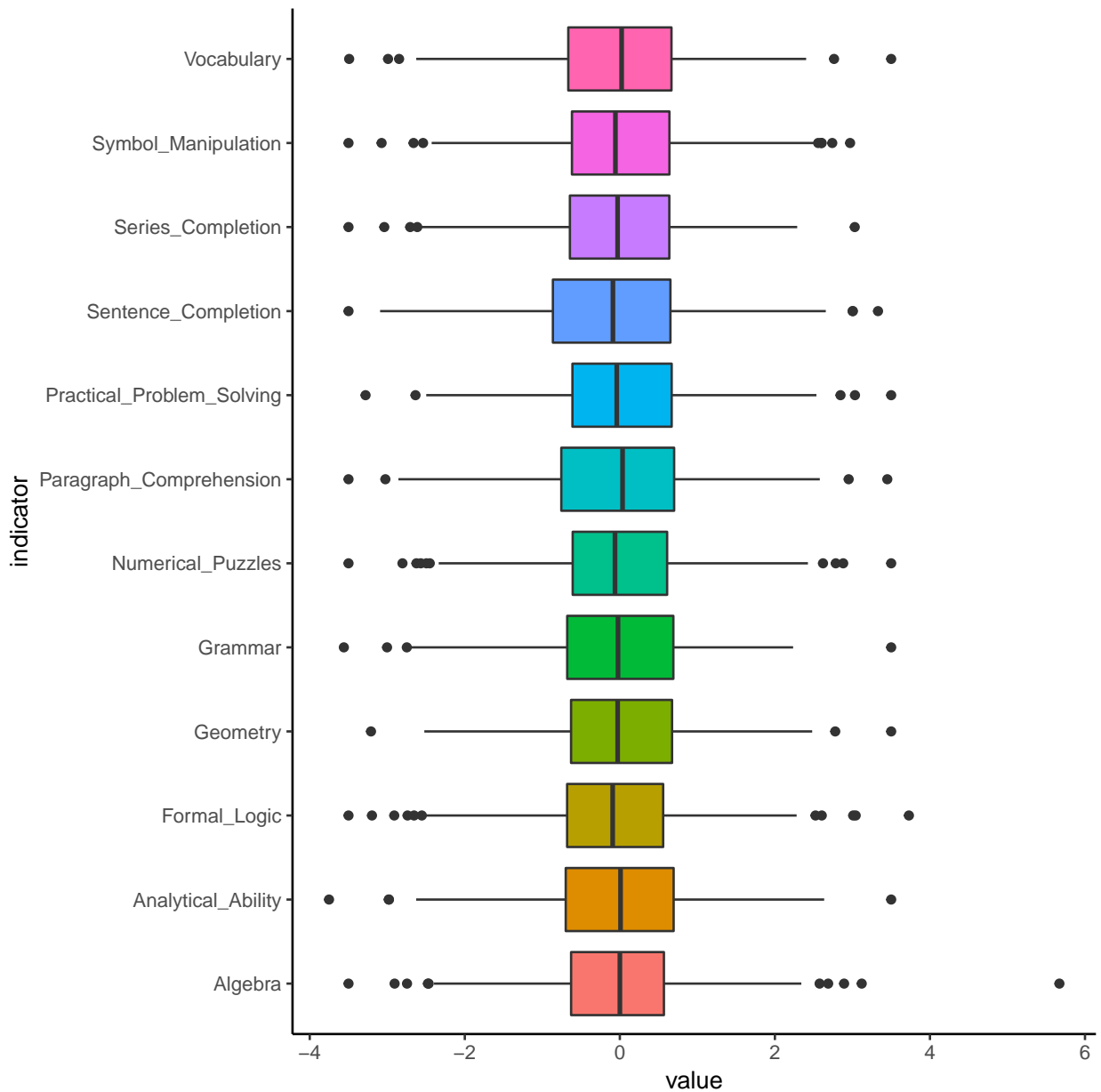
How much variance in the original Geometry variable is accounted for by these extracted components? The extracted components account for 57.85% of the variance in the original Geometry variable.

## 6 Question 5

Now screen the data for unusual cases and determine if your conclusions change when any such cases are excluded from the analysis.

If you believe there is more than one outlier in the data, follow a sequential approach to determining how many to exclude. This means that you will identify the worst offender, exclude that case, and then repeat your diagnostics to determine if other outliers are present. If so, again exclude the worst one, and repeat the diagnostics to determine if an additional outlier is present. Keep cycling through these steps until you are satisfied you have all outliers identified and excluded. Then conduct the principal components analysis. This iterative approach is necessary for multivariate diagnostics such as Mahalanobis distance because the presence of one outlier can influence the apparent presence of others via their joint influence on the covariance matrix. Removing them one at a time insures you dont miss any or mistakenly remove cases that are not really outliers.

```
dat %>%  
  gather(indicator, value, -ID) %>%  
  ggplot(aes(x = indicator, y = value, fill = indicator)) +  
    geom_boxplot() +  
    coord_flip() +  
    theme_classic() +  
    theme(legend.position = "none")
```



Visual inspection of the boxplot suggests there is one outlier in the algebra indicator, but let's check multivariate normality before making a decision.

```
(pca_2 <- principal(
  dat %>% select(-ID)
  , nfactors = ncol(dat)-1
  , rotate = "none"
  , residuals = T
  , scores = TRUE)
)
```

## Principal Components Analysis  
 ## Call: principal(r = dat %>% select(-ID), nfactors = ncol(dat) - 1,  
 ## residuals = T, rotate = "none", scores = TRUE)  
 ## Standardized loadings (pattern matrix) based upon correlation matrix

```

##          PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8
## Grammar      0.62 -0.16 -0.46  0.27  0.04 -0.07  0.32 -0.20
## Paragraph_Comprehension 0.61 -0.09 -0.48 -0.30 -0.01 -0.24 -0.10  0.17
## Vocabulary     0.60 -0.18 -0.44  0.30 -0.25  0.36 -0.05 -0.04
## Sentence_Completion 0.60 -0.14 -0.50 -0.27  0.21 -0.04 -0.14  0.02
## Geometry      0.51  0.55  0.09  0.23  0.39  0.34 -0.23  0.02
## Algebra       0.56  0.46  0.10 -0.38 -0.21  0.19  0.27  0.34
## Numerical_Puzzles 0.48  0.60  0.09  0.32  0.11 -0.37  0.13  0.11
## Series_Completion 0.57  0.50  0.17 -0.20 -0.32 -0.08 -0.13 -0.46
## Practical_Problem_Solving 0.59 -0.31  0.34  0.30 -0.31 -0.09 -0.31  0.24
## Symbol_Manipulation 0.55 -0.35  0.44 -0.28  0.10  0.18  0.25 -0.08
## Analytical_Ability 0.60 -0.32  0.38  0.32  0.07 -0.07  0.20  0.00
## Formal_Logic   0.61 -0.33  0.37 -0.26  0.22 -0.09 -0.18 -0.09
##          PC9  PC10  PC11  PC12 h2      u2 com
## Grammar      -0.08  0.14 -0.04 -0.36  1 -1.3e-15 4.4
## Paragraph_Comprehension 0.22 -0.37 -0.14 -0.09  1  3.3e-16 4.5
## Vocabulary     0.07  0.04 -0.20  0.29  1  1.2e-15 4.9
## Sentence_Completion -0.14  0.18  0.38  0.18  1  1.3e-15 4.5
## Geometry      0.05 -0.12  0.02 -0.17  1 -4.4e-16 4.8
## Algebra       -0.22  0.05 -0.03 -0.05  1  4.4e-16 5.4
## Numerical_Puzzles 0.18  0.19 -0.07  0.22  1 -4.4e-16 4.6
## Series_Completion -0.02 -0.08  0.09  0.00  1  1.1e-16 4.4
## Practical_Problem_Solving 0.05  0.14  0.15 -0.20  1 -2.2e-16 5.6
## Symbol_Manipulation 0.42  0.07  0.11  0.01  1 -2.2e-16 5.4
## Analytical_Ability -0.25 -0.37  0.13  0.17  1 -2.2e-16 5.1
## Formal_Logic   -0.21  0.17 -0.38  0.03  1  4.4e-16 5.0
##
##          PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9  PC10
## SS loadings      3.99 1.64 1.53 1.01 0.58 0.54 0.52 0.48 0.45 0.44
## Proportion Var    0.33 0.14 0.13 0.08 0.05 0.04 0.04 0.04 0.04 0.04
## Cumulative Var    0.33 0.47 0.60 0.68 0.73 0.77 0.82 0.86 0.89 0.93
## Proportion Explained 0.33 0.14 0.13 0.08 0.05 0.04 0.04 0.04 0.04 0.04
## Cumulative Proportion 0.33 0.47 0.60 0.68 0.73 0.77 0.82 0.86 0.89 0.93
##          PC11  PC12
## SS loadings      0.42 0.40
## Proportion Var    0.03 0.03
## Cumulative Var    0.97 1.00
## Proportion Explained 0.03 0.03
## Cumulative Proportion 0.97 1.00
##
## Mean item complexity = 4.9
## Test of the hypothesis that 12 components are sufficient.
##
## The root mean square of the residuals (RMSR) is 0
## with the empirical chi square 0 with prob < NA
##
## Fit based upon off diagonal values = 1

scores_2 <- pca_2$scores %>% data.frame

describe(scores_2)

##      vars    n mean sd median trimmed  mad   min   max range  skew
## PC1      1 500    0  1  0.01  0.01 0.97 -3.03  2.72  5.75 -0.12

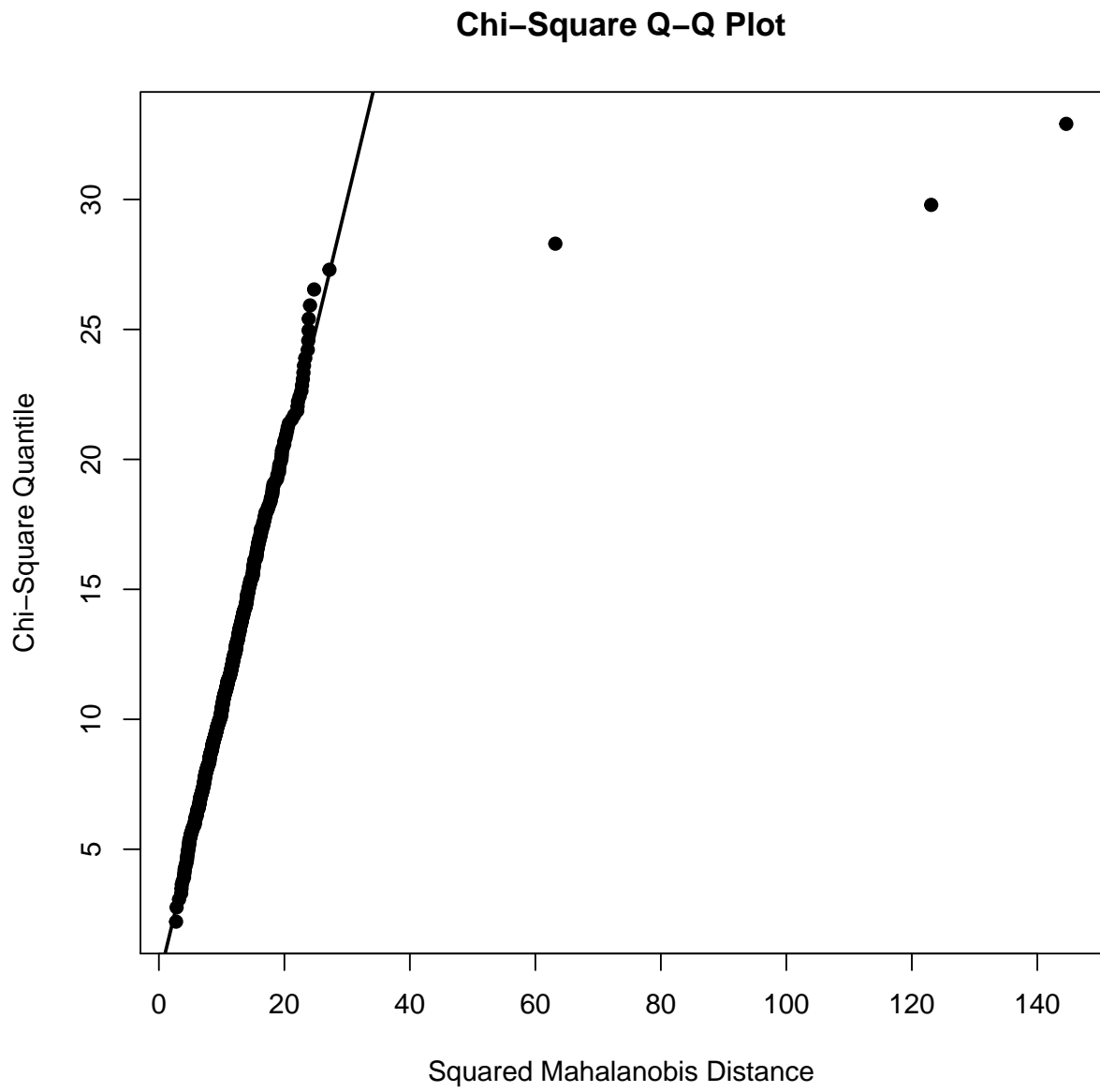
```

```
## PC2      2 500      0 1      0.00      -0.02 1.02      -2.93 3.00 5.93 0.15
## PC3      3 500      0 1      0.03      -0.01 0.97      -3.06 3.10 6.16 0.06
## PC4      4 500      0 1     -0.04      -0.01 0.67     -10.78 11.72 22.50 0.71
## PC5      5 500      0 1      0.02      0.00 1.01      -4.09 2.68 6.77 -0.11
## PC6      6 500      0 1      0.00      0.02 0.96      -2.74 3.11 5.85 -0.14
## PC7      7 500      0 1      0.05      0.03 0.96      -3.22 3.17 6.39 -0.31
## PC8      8 500      0 1     -0.04      -0.01 0.98      -2.54 3.59 6.13 0.13
## PC9      9 500      0 1      0.02      -0.01 1.07      -3.21 3.10 6.31 0.03
## PC10     10 500      0 1     -0.03      0.00 0.94      -3.44 3.34 6.77 0.04
## PC11     11 500      0 1      0.03      0.02 0.98      -3.00 3.33 6.33 -0.08
## PC12     12 500      0 1      0.01      0.01 1.00      -2.67 2.90 5.56 -0.03
##          kurtosis      se
## PC1          -0.26 0.04
## PC2           0.01 0.04
## PC3           0.14 0.04
## PC4          62.60 0.04
## PC5           0.17 0.04
## PC6          -0.04 0.04
## PC7           0.10 0.04
## PC8           0.06 0.04
## PC9          -0.17 0.04
## PC10          0.27 0.04
## PC11          0.09 0.04
## PC12         -0.20 0.04
```

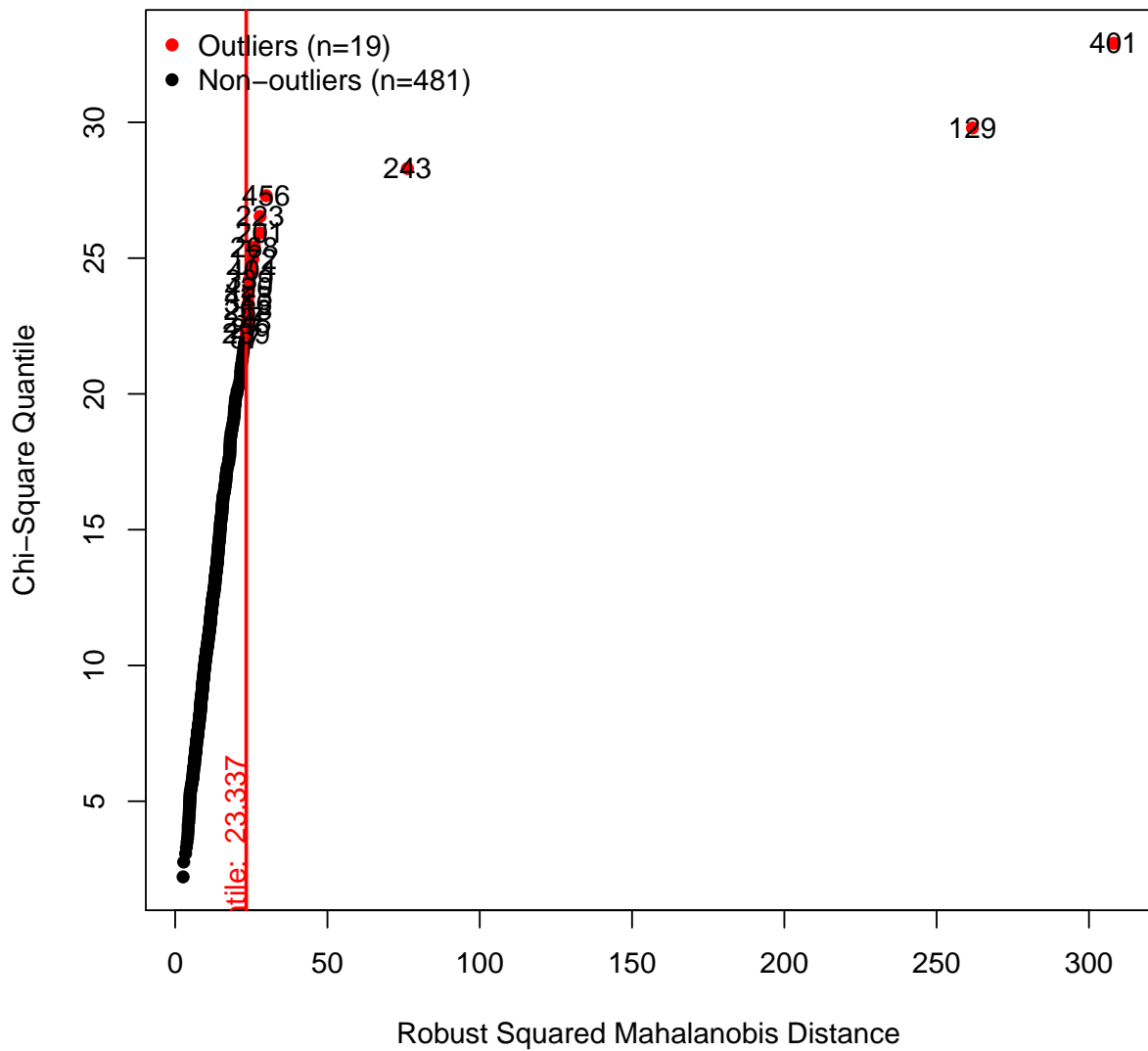
Checking the PCA suggests that there is a multivariate outlier in PCA 4  
 But let's check multivariate normality

```
dat2 <- dat %>% select(-ID) %>% data.frame
rownames(dat2) <- 1:nrow(dat2)
(mv <- mvn(dat2,mvnTest="mardia", multivariatePlot="qq",multivariateOutlierMethod="quan",showOutliers=T
```





### Chi-Square Q-Q Plot



```
## $multivariateNormality
##      Test      Statistic      p value Result
## 1 Mardia Skewness 468.99164354467 0.000162906731930752 NO
## 2 Mardia Kurtosis 37.8554992977048 0 NO
## 3 MVN <NA> <NA> NO
##
## $univariateNormality
##      Test      Variable Statistic p value Normality
## 1 Shapiro-Wilk Grammar 0.9967 0.4074 YES
## 2 Shapiro-Wilk Paragraph_Comprehension 0.9989 0.9922 YES
## 3 Shapiro-Wilk Vocabulary 0.9958 0.1996 YES
## 4 Shapiro-Wilk Sentence_Completion 0.9981 0.8464 YES
## 5 Shapiro-Wilk Geometry 0.9978 0.7646 YES
## 6 Shapiro-Wilk Algebra 0.9835 <0.001 NO
```

```

## 7 Shapiro-Wilk Numerical_Puzzles 0.9968 0.4251 YES
## 8 Shapiro-Wilk Series_Completion 0.9971 0.5203 YES
## 9 Shapiro-Wilk Practical_Problem_Solving 0.9967 0.4047 YES
## 10 Shapiro-Wilk Symbol_Manipulation 0.9956 0.1713 YES
## 11 Shapiro-Wilk Analytical_Ability 0.9977 0.7187 YES
## 12 Shapiro-Wilk Formal_Logic 0.9952 0.1221 YES
##
## $Descriptives
##           n           Mean   Std.Dev         Median
## Grammar          500 -0.0083641562 1.0016648 -0.0237048215
## Paragraph_Comprehension 500 -0.0070088967 1.0829739 0.0350785195
## Vocabulary        500 -0.0169538702 1.0169743 0.0232942325
## Sentence_Completion 500 -0.0903289108 1.1065777 -0.0893020655
## Geometry          500 0.0071552259 0.9973637 -0.0282354790
## Algebra           500 -0.0151778962 1.0149132 0.0003852515
## Numerical_Puzzles  500 -0.0153611723 1.0132508 -0.0624955155
## Series_Completion  500 -0.0077428597 0.9819436 -0.0291777090
## Practical_Problem_Solving 500 -0.0182512446 0.9584017 -0.0400207855
## Symbol_Manipulation 500 -0.0009142506 0.9922338 -0.0569430780
## Analytical_Ability  500 -0.0073984716 1.0009061 0.0090736190
## Formal_Logic       500 -0.0650934689 1.0367206 -0.0925315375
##           Min           Max          25th          75th
## Grammar          -3.560000 3.500000 -0.6789687 0.6903545
## Paragraph_Comprehension -3.500000 3.450000 -0.7551515 0.7002892
## Vocabulary        -3.491518 3.500000 -0.6651436 0.6660205
## Sentence_Completion -3.500000 3.330000 -0.8637548 0.6525117
## Geometry          -3.210000 3.500000 -0.6294255 0.6737671
## Algebra           -3.500000 5.670000 -0.6288277 0.5685107
## Numerical_Puzzles  -3.500000 3.500000 -0.6082066 0.6101796
## Series_Completion  -3.500000 3.030000 -0.6444230 0.6380559
## Practical_Problem_Solving -3.280000 3.500000 -0.6115184 0.6685177
## Symbol_Manipulation -3.500000 2.970000 -0.6179801 0.6397075
## Analytical_Ability  -3.751222 3.500000 -0.6981759 0.6938528
## Formal_Logic       -3.500000 3.728522 -0.6802409 0.5589411
##           Skew      Kurtosis
## Grammar          -0.05430769 0.13596613
## Paragraph_Comprehension -0.08800734 0.02507948
## Vocabulary        -0.20788125 0.16156136
## Sentence_Completion 0.02093275 -0.07335932
## Geometry          0.11181433 0.10313111
## Algebra           0.23409531 1.98559538
## Numerical_Puzzles 0.01651535 0.33457069
## Series_Completion -0.13807654 0.07180595
## Practical_Problem_Solving 0.13693839 0.33690319
## Symbol_Manipulation 0.05168770 0.34049898
## Analytical_Ability -0.11832265 0.22344750
## Formal_Logic       0.03731202 0.57661676
##
## $multivariateOutliers
##      Observation Mahalanobis Distance Outlier
## 401          401          308.127      TRUE
## 129          129          261.764      TRUE
## 243          243           76.265      TRUE

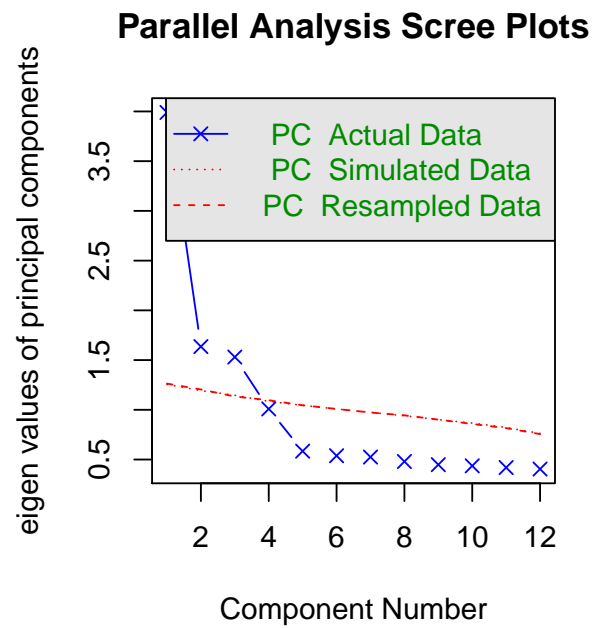
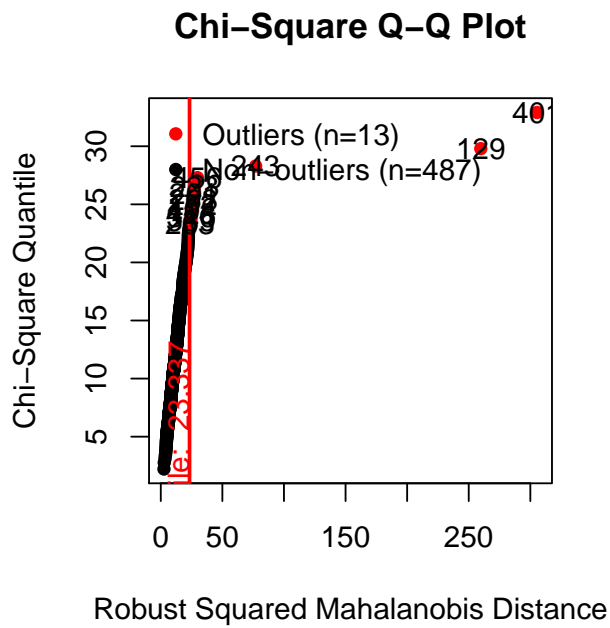
```

## 456	456	29.895	TRUE
## 223	223	27.886	TRUE
## 201	201	27.874	TRUE
## 268	268	26.026	TRUE
## 172	172	25.537	TRUE
## 404	404	25.216	TRUE
## 339	339	24.361	TRUE
## 420	420	24.155	TRUE
## 423	423	24.045	TRUE
## 316	316	24.008	TRUE
## 263	263	23.977	TRUE
## 37	37	23.946	TRUE
## 245	245	23.704	TRUE
## 27	27	23.397	TRUE
## 239	239	23.374	TRUE
## 87	87	23.350	TRUE

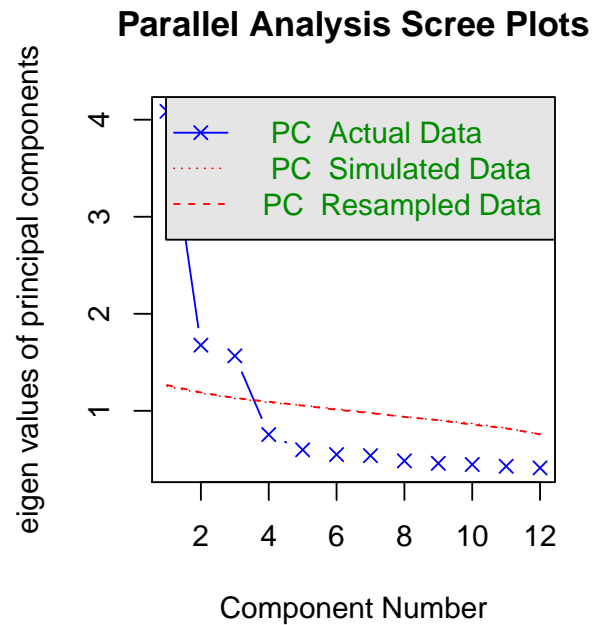
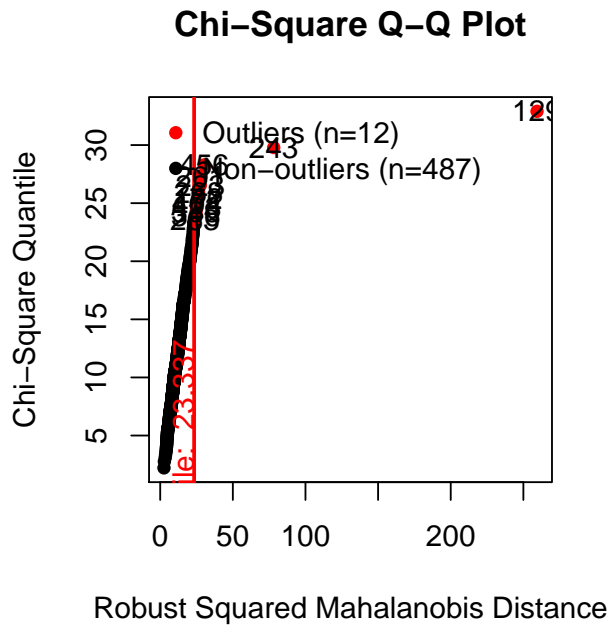
Based on the test of multivariate normality, there are 16(!) outliers. Let's remove the outlier and check again.

I'm going to use a while loop to do this rather than copying and pasting the code.

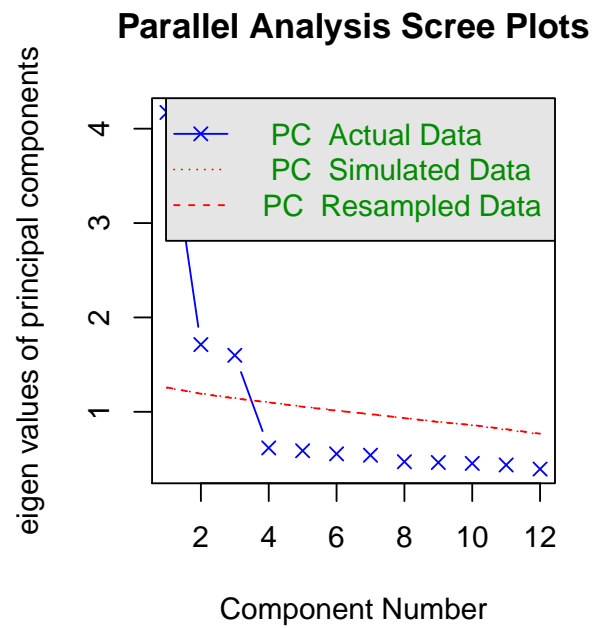
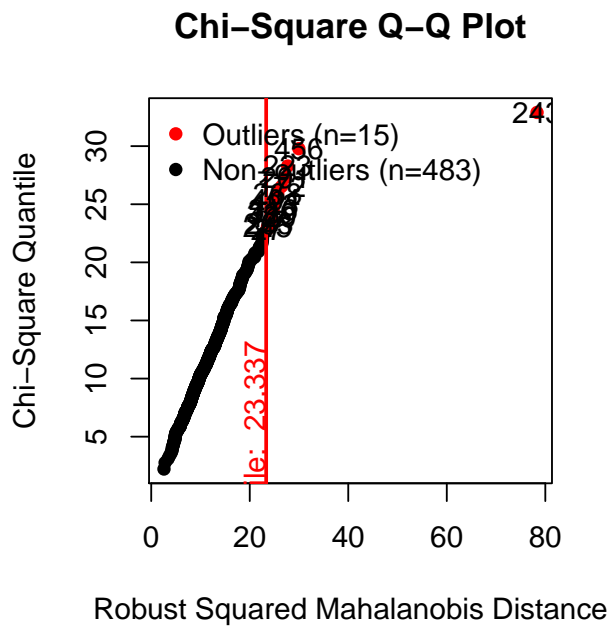
```
k <- 1
remove <- c()
par(mfrow = c(1,2))
while(nrow(mv$multivariateOutliers) > 0){
  if(k!=1){tmp <- dat2[-remove,]} else{tmp <- dat2}
  mv <- mvn(tmp, mvnTest="mardia", multivariatePlot="none",multivariateOutlierMethod="quan",showOutliers=TRUE)
  mv$multivariateOutliers
  remove <- c(remove, as.numeric(as.character(mv$multivariateOutliers$Observation[1])))
  sink("/dev/null")
  scree <- fa.parallel(tmp, fa="pc")
  sink()
  print(sprintf("Case %s removed. %s factors remain. This is the %s round", remove[k], scree$ncomp, k))
  if(nrow(mv$multivariateOutliers) == 0){break}
  k <- k + 1
}
```



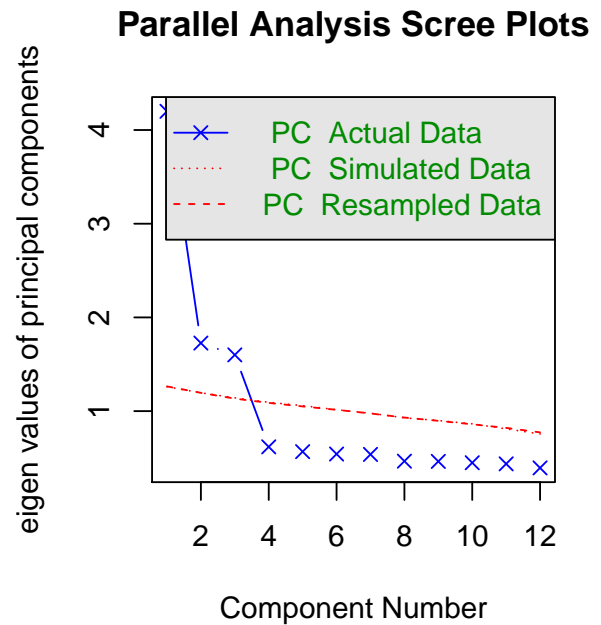
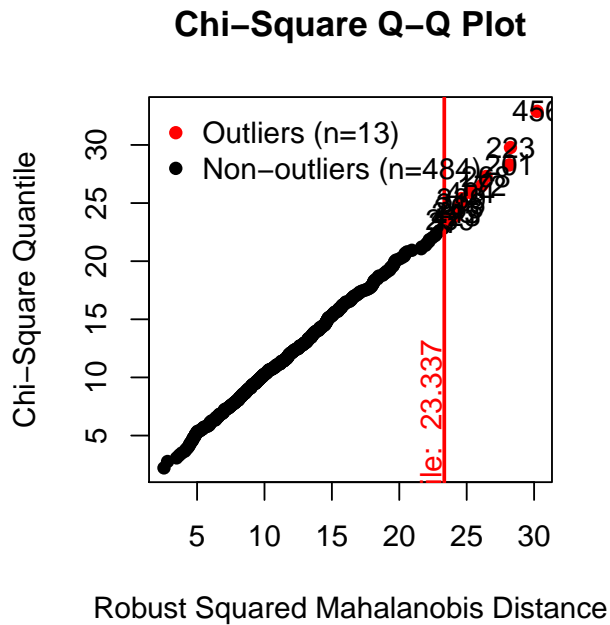
```
## [1] "Case 401 removed. 3 factors remain. This is the 1 round"
```



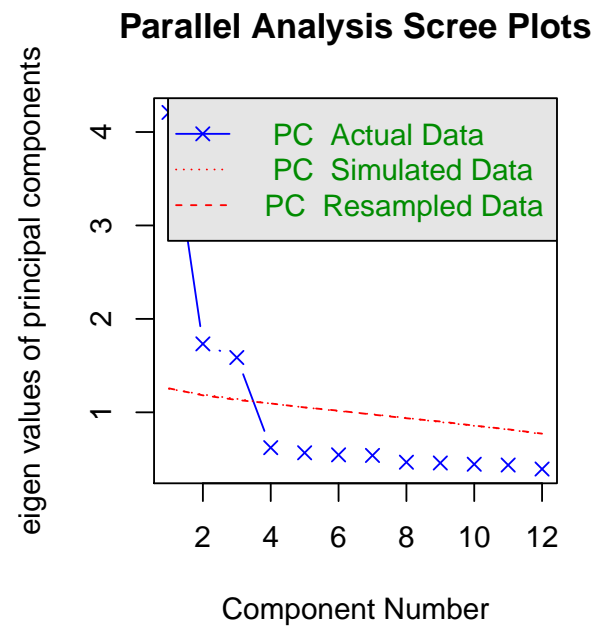
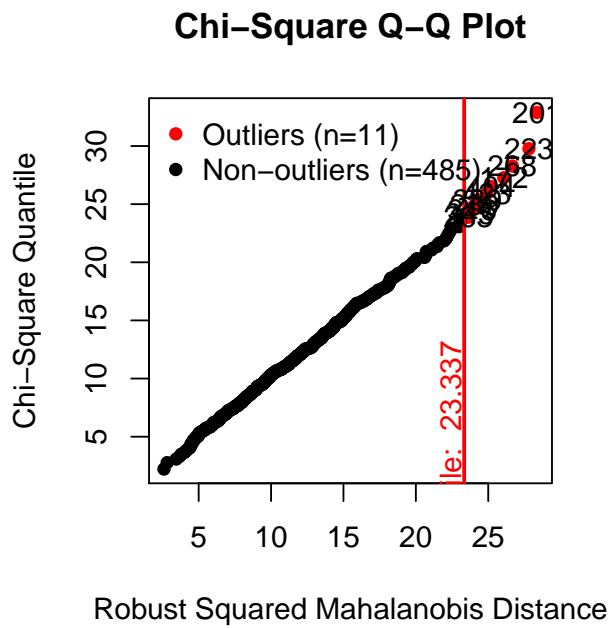
```
## [1] "Case 129 removed. 3 factors remain. This is the 2 round"
```



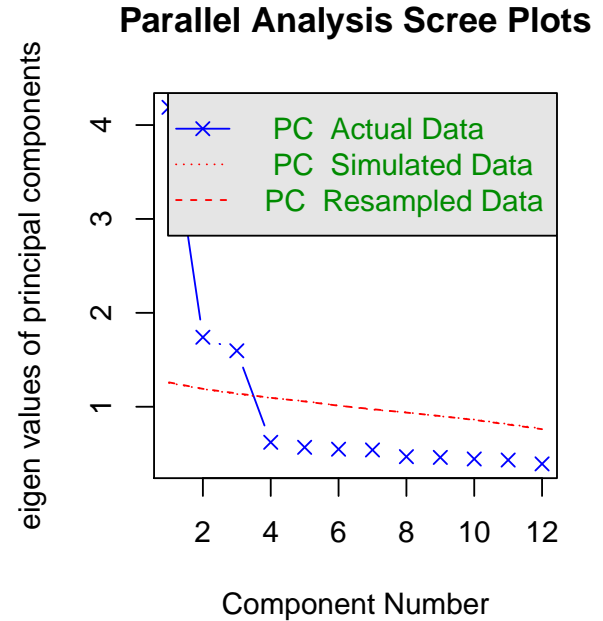
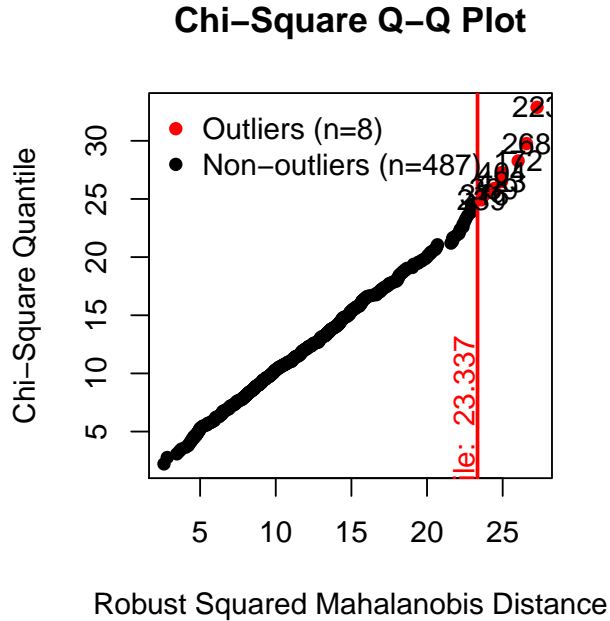
```
## [1] "Case 243 removed. 3 factors remain. This is the 3 round"
```



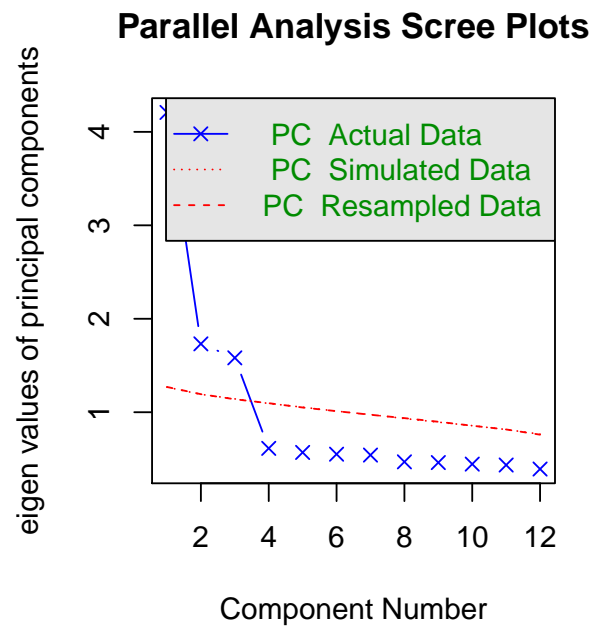
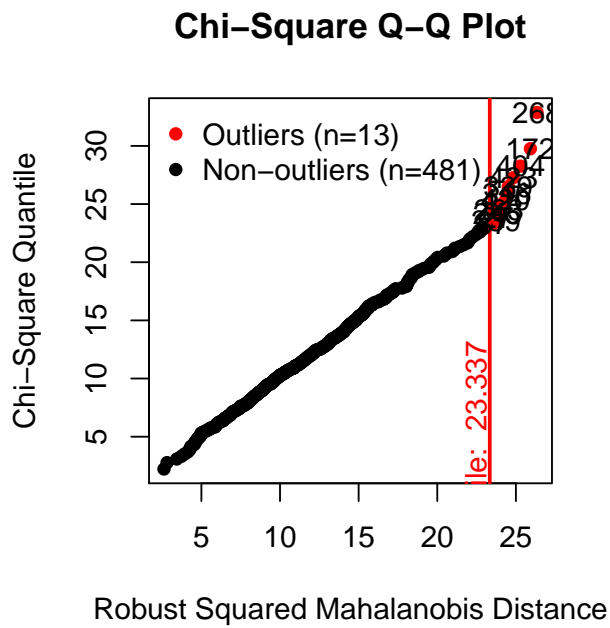
```
## [1] "Case 456 removed. 3 factors remain. This is the 4 round"
```



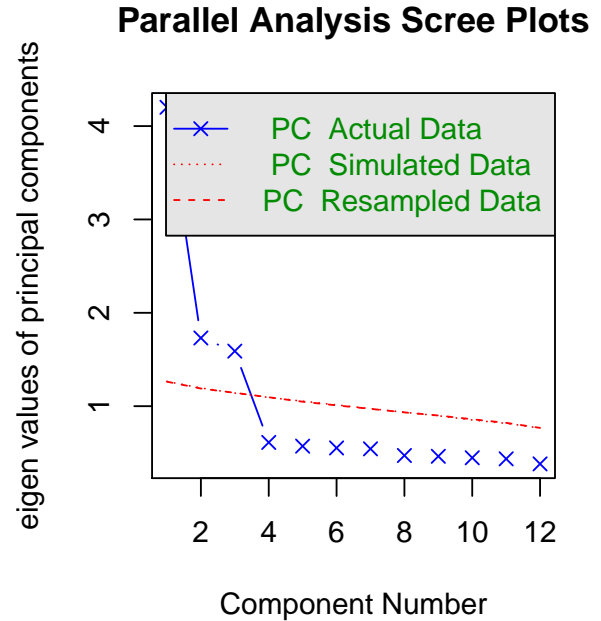
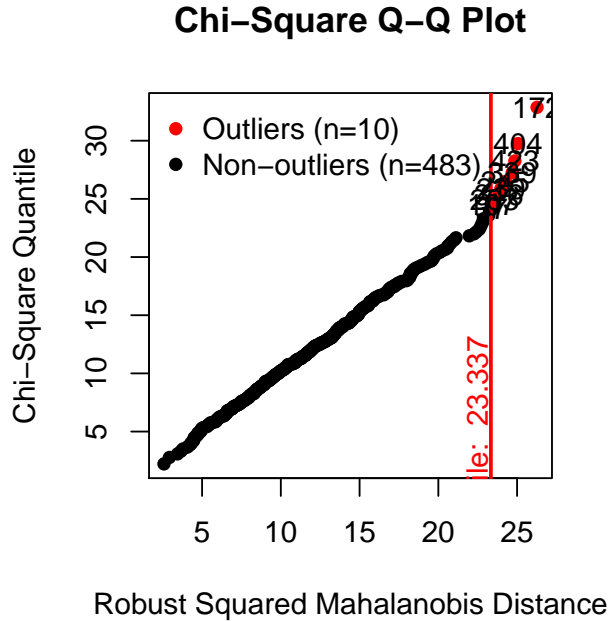
```
## [1] "Case 201 removed. 3 factors remain. This is the 5 round"
```



```
## [1] "Case 223 removed. 3 factors remain. This is the 6 round"
```

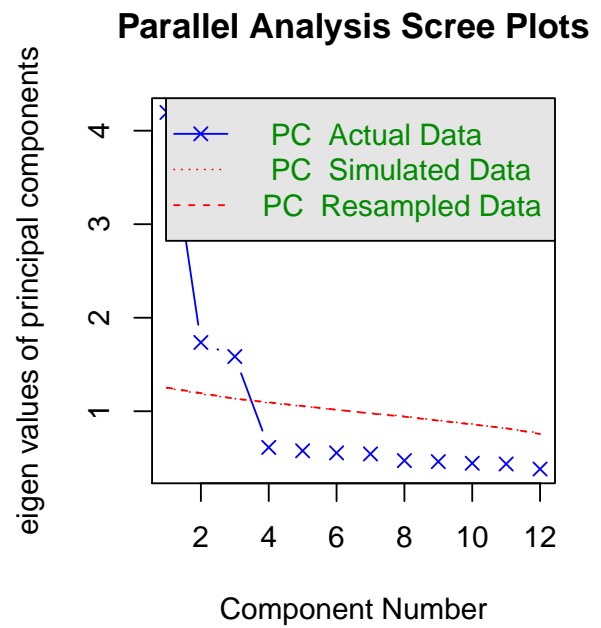
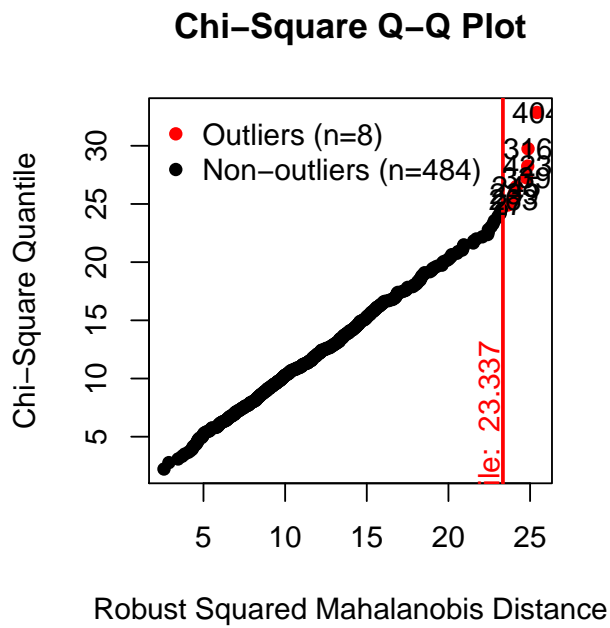


```
## [1] "Case 268 removed. 3 factors remain. This is the 7 round"
```

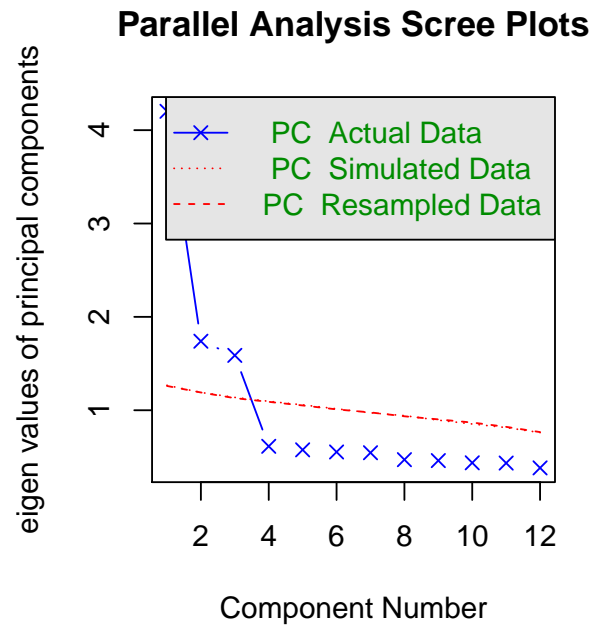
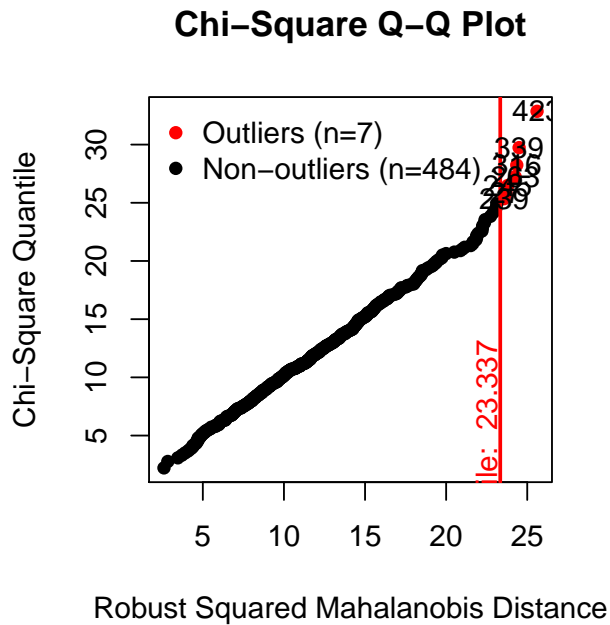


```
## [1] "Case 172 removed. 3 factors remain. This is the 8 round"
```



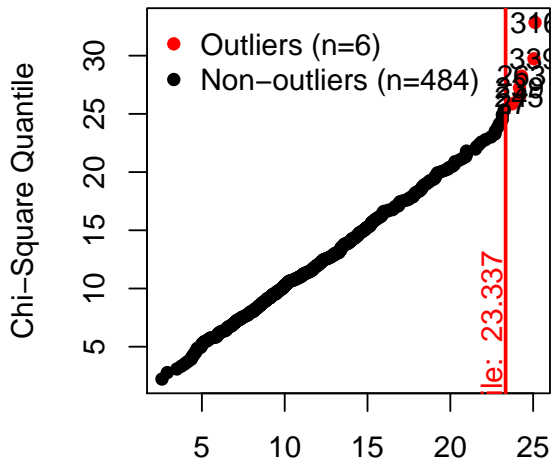


```
## [1] "Case 404 removed. 3 factors remain. This is the 9 round"
```

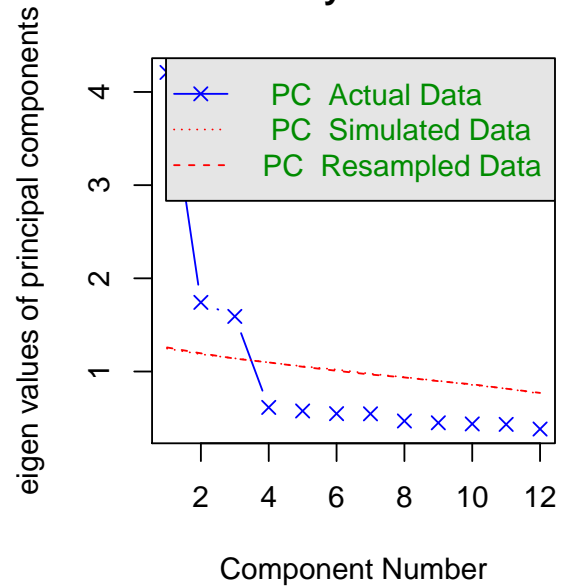


```
## [1] "Case 423 removed. 3 factors remain. This is the 10 round"
```

**Chi-Square Q-Q Plot**

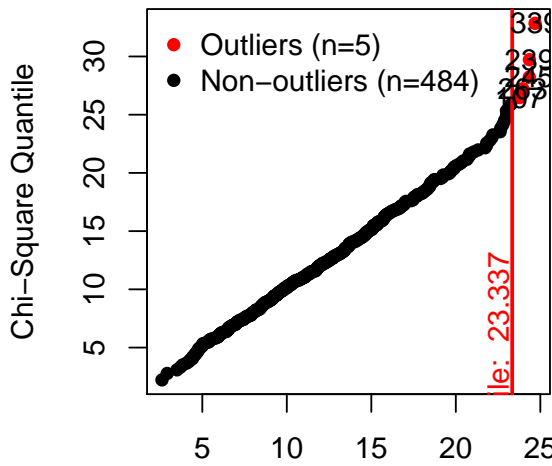


**Parallel Analysis Scree Plots**

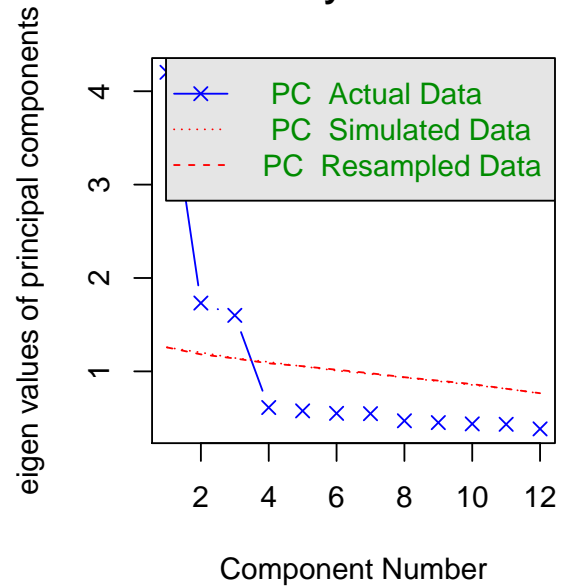


```
## [1] "Case 316 removed. 3 factors remain. This is the 11 round"
```

**Chi-Square Q-Q Plot**

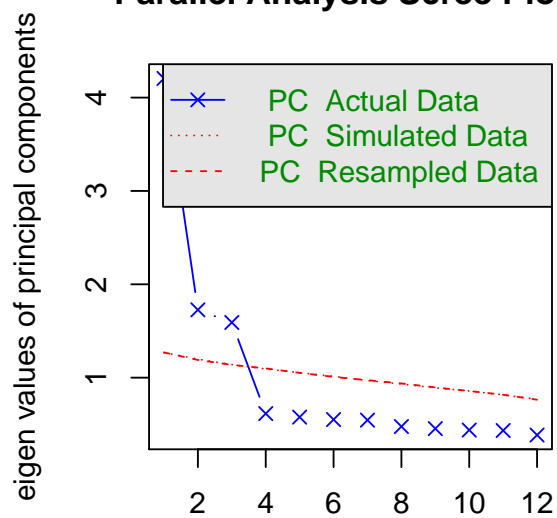


**Parallel Analysis Scree Plots**



```
## [1] "Case 339 removed. 3 factors remain. This is the 12 round"
```

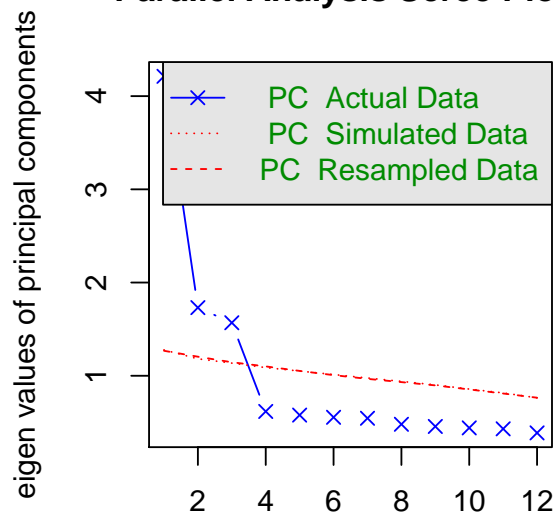
## Parallel Analysis Scree Plots



Component Number

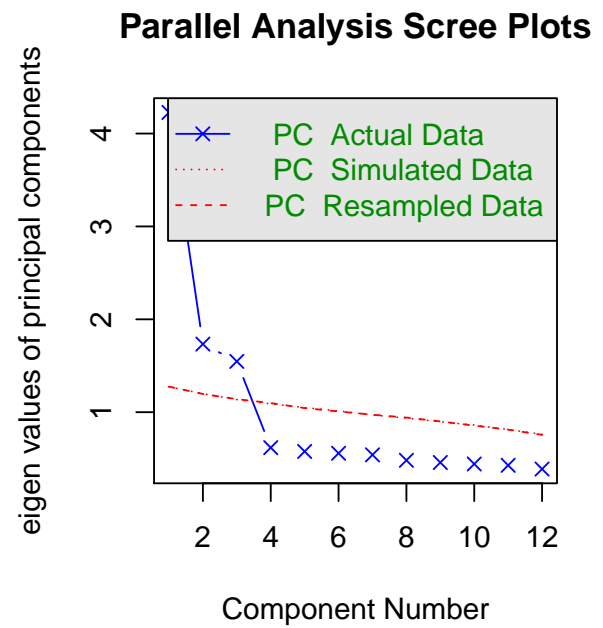
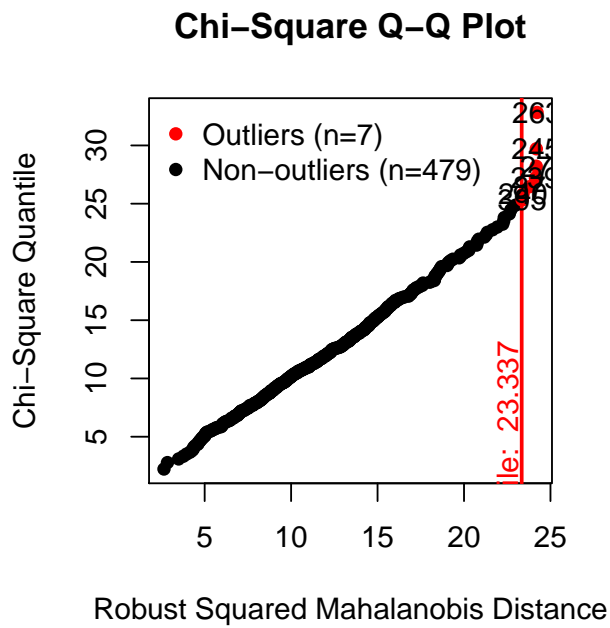
```
## [1] "Case 37 removed. 3 factors remain. This is the 13 round"
```

## Parallel Analysis Scree Plots

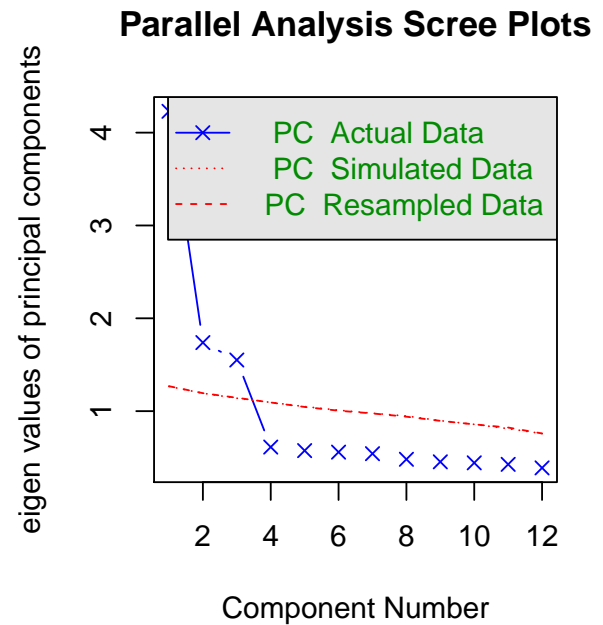
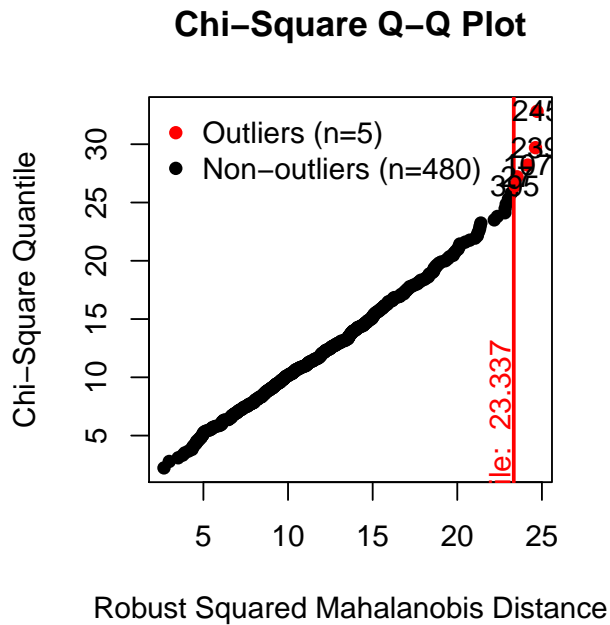


Component Number

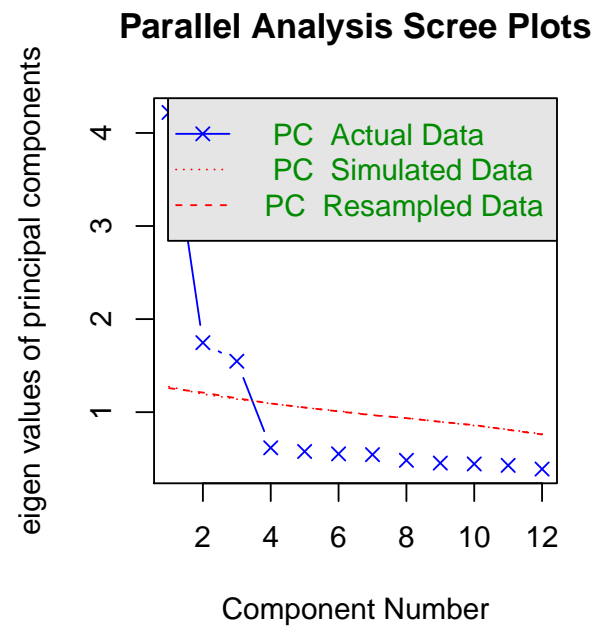
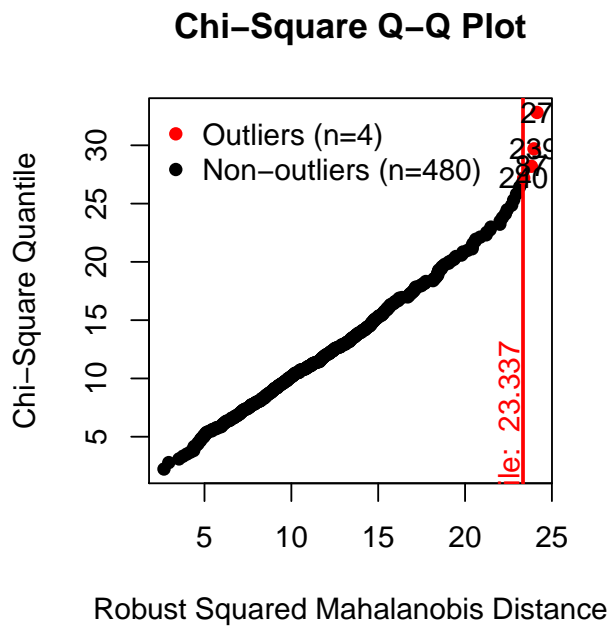
```
## [1] "Case 420 removed. 3 factors remain. This is the 14 round"
```



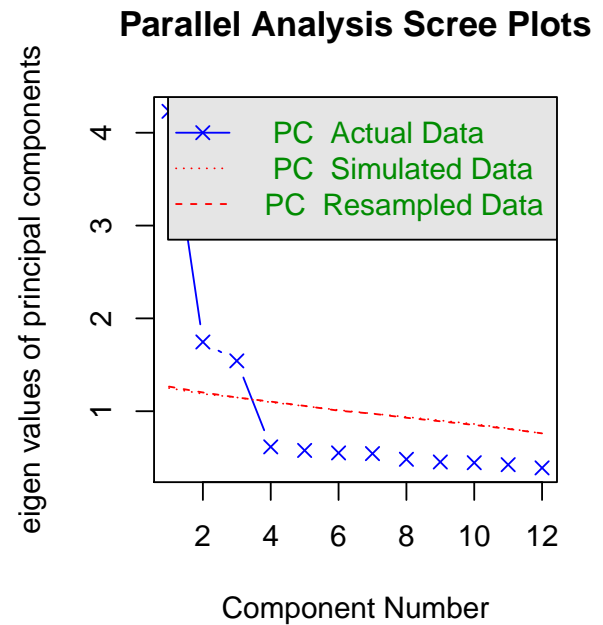
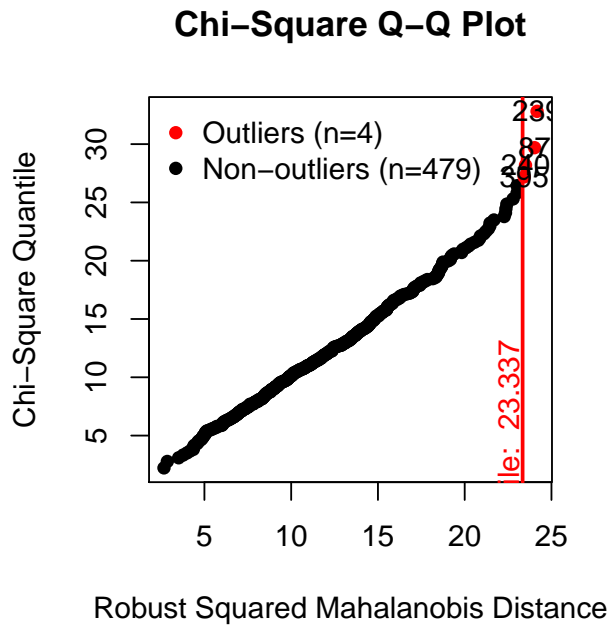
```
## [1] "Case 263 removed. 3 factors remain. This is the 15 round"
```



```
## [1] "Case 245 removed. 3 factors remain. This is the 16 round"
```

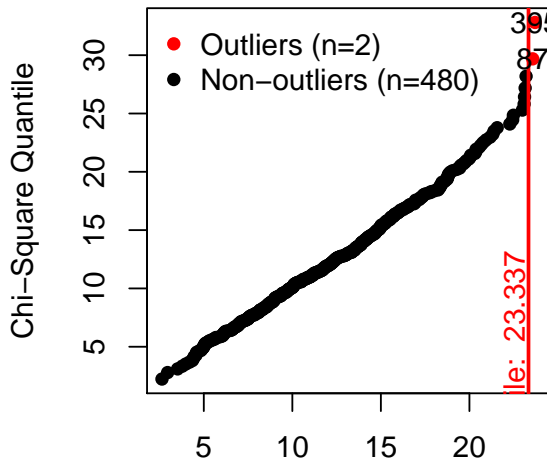


```
## [1] "Case 27 removed. 3 factors remain. This is the 17 round"
```



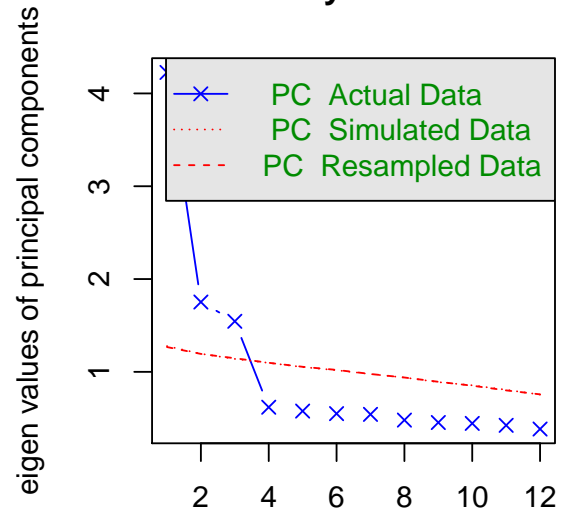
```
## [1] "Case 239 removed. 3 factors remain. This is the 18 round"
```

**Chi-Square Q-Q Plot**



Robust Squared Mahalanobis Distance

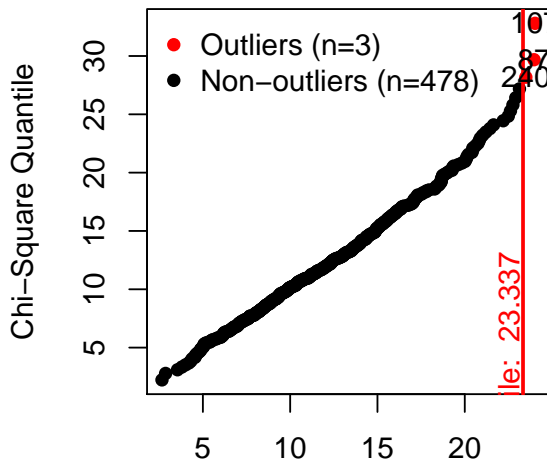
**Parallel Analysis Scree Plots**



Component Number

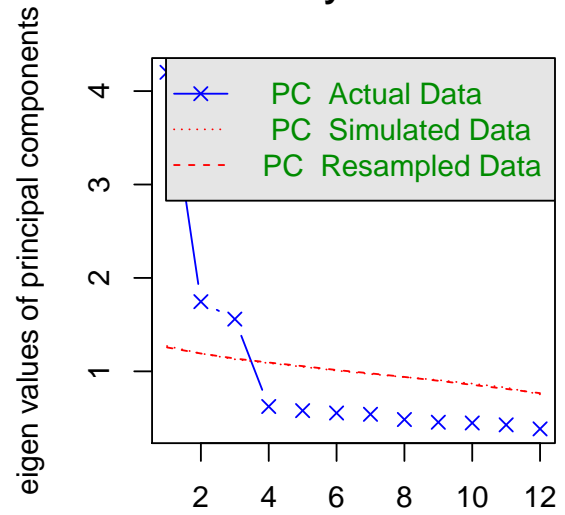
```
## [1] "Case 395 removed. 3 factors remain. This is the 19 round"
```

**Chi-Square Q-Q Plot**



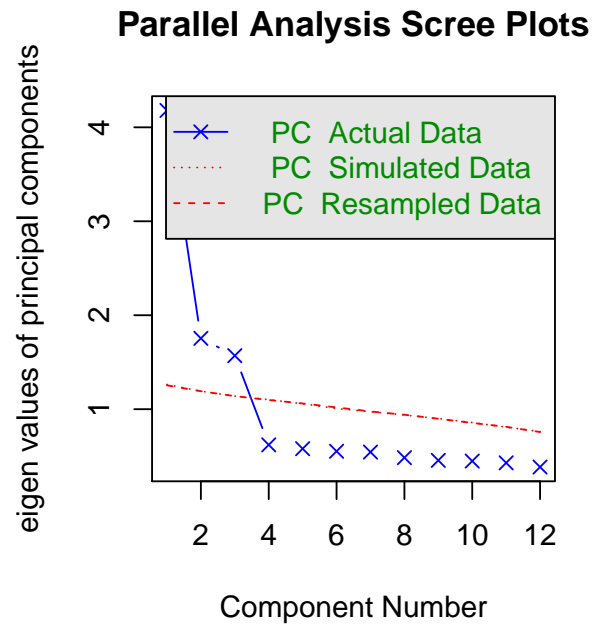
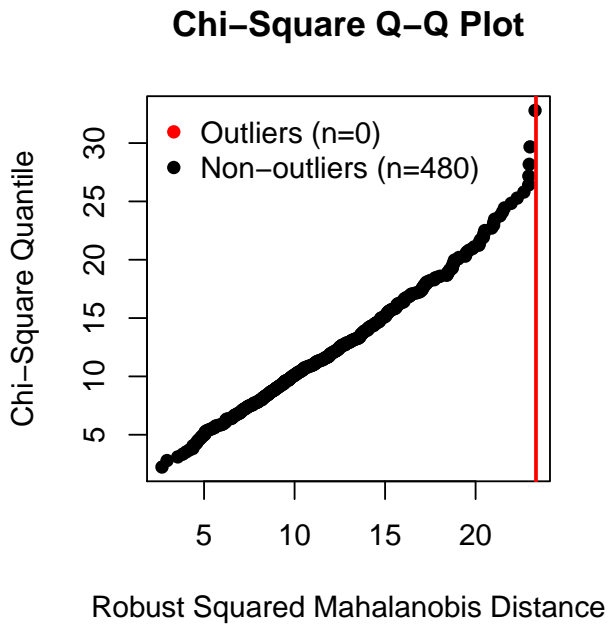
Robust Squared Mahalanobis Distance

**Parallel Analysis Scree Plots**



Component Number

```
## [1] "Case 107 removed. 3 factors remain. This is the 20 round"
```



```
## [1] "Case NA removed. 3 factors remain. This is the 21 round"

pca_final <- principal(
  tmp
  , nfactors = scree$ncomp
  , rotate = "none"
  , residuals = T
  , scores = TRUE)
```

In total, using the mardia test, it took 21 rounds to remove the following outliers: cases .

Our conclusions do not change. We would still extract 3 components. However, the percentage of variance changes (see below):

```
pca_final$Vaccounted %>% data.frame %>% mutate(m = rownames(.)) %>%
  mutate_at(vars(PC1:PC3), funs(round(.,2))) %>%
  select(m, everything()) %>%
  kable(., "latex", booktabs = T, escape = F)
```

m	PC1	PC2	PC3
SS loadings	4.18	1.75	1.57
Proportion Var	0.35	0.15	0.13
Cumulative Var	0.35	0.49	0.63
Proportion Explained	0.56	0.23	0.21
Cumulative Proportion	0.56	0.79	1.00