Replication Document

This document can be used to replicate the quantitative lithic analyses presented in the paper "Multiple hominin dispersals into Southwest Asia over the last 400,000 years" by Groucutt et al. 2020. All analyses were conducted in R.

Load libraries

First, we will load the psych library for convenient PCA-related tests, and the ggplot2 and ggpubr libraries for plotting results.

```
library(psych)
library(ggplot2)
library(ggpubr)
```

Data

Next, we load the lithic data as follows:

```
LP <- read.csv(file="./LP.csv")
MIS67 <- read.csv(file="./MIS67.csv")
```

These data sets come with several variables (columns). The Lower Palaeolithic data look like this:

```
head(LP)
```

```
##   Assemblage  ID N..scars Flaking.Length Width.at.Midpoint Proximal.Width
## 1  KAM-4 A.E  14        3          31.33             21.50          21.29
## 2  KAM-4 A.E  40        5          38.03             34.17          31.96
## 3  KAM-4 A.E  58        4          45.94             35.65          32.32
## 4  KAM-4 A.E  59        5          57.07             34.42          34.79
## 5  KAM-4 A.E  61        3          38.96             25.45          28.52
## 6  KAM-4 A.E 108        4          45.97             30.32          31.99
##   Distal.Width Thickness.at.midpoint Platform.Width Platform.Thickness
## 1        12.76                  5.76          18.05               4.87
## 2        23.17                  7.95          26.32               4.47
## 3        26.41                 12.43          29.02              12.18
## 4         6.41                  9.36          36.18               7.72
## 5         2.18                  5.83          29.80               5.26
## 6         3.44                  7.84          29.78               8.89
```

The data from the transition between Marine Isotope Stage 6 and 7 (MIS67) look like this:

```
head(MIS67)
```

```
##   Assemblage   ID N..scars Flaking.Length Width.at.Midpoint Proximal.Width
## 1    KAM-4-C   61        6          46.27             41.22          25.09
## 2    KAM-4-C 5031        4          52.39             39.74          22.64
## 3    KAM-4-C   77        8          34.08             43.92          23.57
```

```
## 4    KAM-4-C 1427        4          41.68             24.16           24.01
## 5    KAM-4-C 1431        5          23.78             24.05           16.64
## 6    KAM-4-C 1455        3          30.16             32.35           28.43
##   Distal.Width Thickness.at.midpoint Platform.Width Platform.Thickness
## 1        38.77                  7.95          25.18               3.34
## 2        27.48                  7.99          19.28               7.45
## 3        20.64                  9.74          24.30              10.16
## 4        18.35                  7.18          21.56               5.87
## 5        14.62                  6.81          19.06               3.27
## 6        26.15                  7.04          26.55               4.88
```

A summary of sample sizes for the two data sets and each individual site are as follows.

Samples sizes by data set (time-period):

```
sample_size_period <- cbind(c("LP","MIS67"),c(dim(LP)[1],dim(MIS67)[1]))
sample_size_period
```

```
##      [,1]    [,2]
## [1,] "LP"    "404"
## [2,] "MIS67" "92"
```

Sample sizes by assemblage:

```
assemblage_n <- table(LP$Assemblage)
sample_size_assemblage <- data.frame(Assemblage = names(assemblage_n),
                               n = as.numeric(assemblage_n),
                               period = "LP")
assemblage_n <- table(MIS67$Assemblage)
sample_size_assemblage <- rbind(sample_size_assemblage,
                            data.frame(Assemblage = names(assemblage_n),
                                       n = as.numeric(assemblage_n),
                                       period = "MIS67"))
sample_size_assemblage
```

```
##    Assemblage  n period
## 1       ANW-3 50     LP
## 2         BNS 32     LP
## 3       JSM-1 36     LP
## 4   KAM-4 A.D 39     LP
## 5   KAM-4 A.E 14     LP
## 6    Kebara X 50     LP
## 7      MDF-61 50     LP
## 8  Qafzeh XIX 50     LP
## 9   Tor Faraj 50     LP
## 10      Wusta 33     LP
## 11        AHS 45  MIS67
## 12    KAM-4-C 21  MIS67
## 13    Misliya 26  MIS67
```

PrePCA tests

Before running the analysis, we used a couple of simple preliminary tests to determine whether the variation in the data was sufficiently greater in at least one or more dimensions that sensible principle components could be extracted. One test involved the “Kaiser, Meyer, Olkin Measure of Sampling Adequacy”:

```r
KMO(MIS67[,c(3:10)])
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = MIS67[, c(3:10)])
## Overall MSA =  0.73
## MSA for each item =
##              N..scars        Flaking.Length     Width.at.Midpoint
##                  0.39                  0.56                  0.76
##        Proximal.Width          Distal.Width Thickness.at.midpoint
##                  0.76                  0.54                  0.84
##        Platform.Width    Platform.Thickness
##                  0.76                  0.87
```

```r
KMO(LP[,c(3:10)])
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = LP[, c(3:10)])
## Overall MSA =  0.78
## MSA for each item =
##              N..scars        Flaking.Length     Width.at.Midpoint
##                  0.81                  0.76                  0.85
##        Proximal.Width          Distal.Width Thickness.at.midpoint
##                  0.74                  0.72                  0.87
##        Platform.Width    Platform.Thickness
##                  0.67                  0.83
```

The other involved "Bartlett's Test for Sphericity",

```r
cortest.bartlett(MIS67[,c(3:10)])
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 396.6637
##
## $p.value
## [1] 9.26127e-67
##
## $df
## [1] 28
```

```r
cortest.bartlett(LP[,c(3:10)])
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 1673.095
##
## $p.value
## [1] 0
##
## $df
## [1] 28
```

PCA

Then, we can perform the simple PCA on the relevant lithic variables,

```
pca_MIS67 <- prcomp(
                x = MIS67[,c(3:10)],
                retx = T,
                center = T,
                scale = T)

pca_LP <- prcomp(
            x = LP[,c(3:10)],
            retx = T,
            center = T,
            scale = T)
```

We can then look at the loadings tables to see how the variables each correlate with the extracted components:

```
pca_MIS67
```

```
## Standard deviations (1, .., p=8):
## [1] 2.0011960 1.1348036 1.0049351 0.7980139 0.6561803 0.5683053 0.4131356
## [8] 0.3694453
##
## Rotation (n x k) = (8 x 8):
##                             PC1         PC2         PC3           PC4
## N..scars             -0.1058251 -0.39746054 -0.83309126  0.0384716968
## Flaking.Length       -0.2716468  0.59598817 -0.36794310  0.0842223507
## Width.at.Midpoint    -0.4406385  0.03718445 -0.11629804  0.1711988731
## Proximal.Width       -0.4203641 -0.06395559  0.08578626  0.5259180120
## Distal.Width         -0.2461758 -0.67761951  0.16768726 -0.0008845973
## Thickness.at.midpoint -0.3901399  0.10314382 -0.07125931 -0.5906260781
## Platform.Width       -0.4203846  0.09235265  0.30097262  0.2629855791
## Platform.Thickness   -0.3931995 -0.05490312  0.16096859 -0.5172294977
##                              PC5          PC6         PC7         PC8
## N..scars             -0.32384544  0.038105596  0.09084923  0.14395680
## Flaking.Length        0.32673335  0.229091300  0.42398898 -0.29920965
## Width.at.Midpoint     0.50918830 -0.023858599 -0.48401857  0.51720870
## Proximal.Width       -0.29613546 -0.019801464 -0.36564977 -0.55988086
## Distal.Width          0.48948292  0.002214711  0.39833865 -0.23202878
## Thickness.at.midpoint -0.06436812 -0.663712162 -0.06261620 -0.18616461
## Platform.Width       -0.38311201 -0.147983442  0.51602374  0.46909211
## Platform.Thickness   -0.22558571  0.694756052 -0.12300839  0.01538124
```

```
pca_LP
```

```
## Standard deviations (1, .., p=8):
## [1] 1.9996583 1.1776678 0.8742895 0.7786343 0.7233013 0.5773857 0.5039168
## [8] 0.3651587
##
## Rotation (n x k) = (8 x 8):
##                           PC1         PC2         PC3         PC4         PC5
## N..scars            0.1590189  0.61961998 -0.29015781  0.67459897 -0.22150927
## Flaking.Length      0.3029657 -0.28228962 -0.76499848 -0.08123662  0.19863697
## Width.at.Midpoint   0.4275462  0.09095345 -0.19931229 -0.25273546 -0.07998487
## Proximal.Width      0.4208977 -0.29799462  0.07430390  0.02886084 -0.39638989
```

```
## Distal.Width          0.3027166  0.47334148  0.20144533 -0.53300155 -0.33431170
## Thickness.at.midpoint 0.4061573  0.15930225  0.04942317 -0.09537377  0.42950637
## Platform.Width        0.3526375 -0.43305342  0.27457724  0.35384812 -0.33255125
## Platform.Thickness    0.3768552  0.04843245  0.40856307  0.23497529  0.58405831
##                               PC6         PC7         PC8
## N..scars               0.02229839 -0.02255268 -0.03790208
## Flaking.Length        -0.12335312 -0.41861468  0.08240988
## Width.at.Midpoint     -0.34692688  0.75028123  0.12511328
## Proximal.Width         0.16461594 -0.01750473 -0.73699426
## Distal.Width          -0.10273428 -0.46387180  0.14880830
## Thickness.at.midpoint  0.76776004  0.11326691  0.10658168
## Platform.Width         0.04208609 -0.07370184  0.60804346
## Platform.Thickness    -0.48481860 -0.16609147 -0.17244640
```

Next, we can extract PC scores for the original observations (project the data onto the component axis):

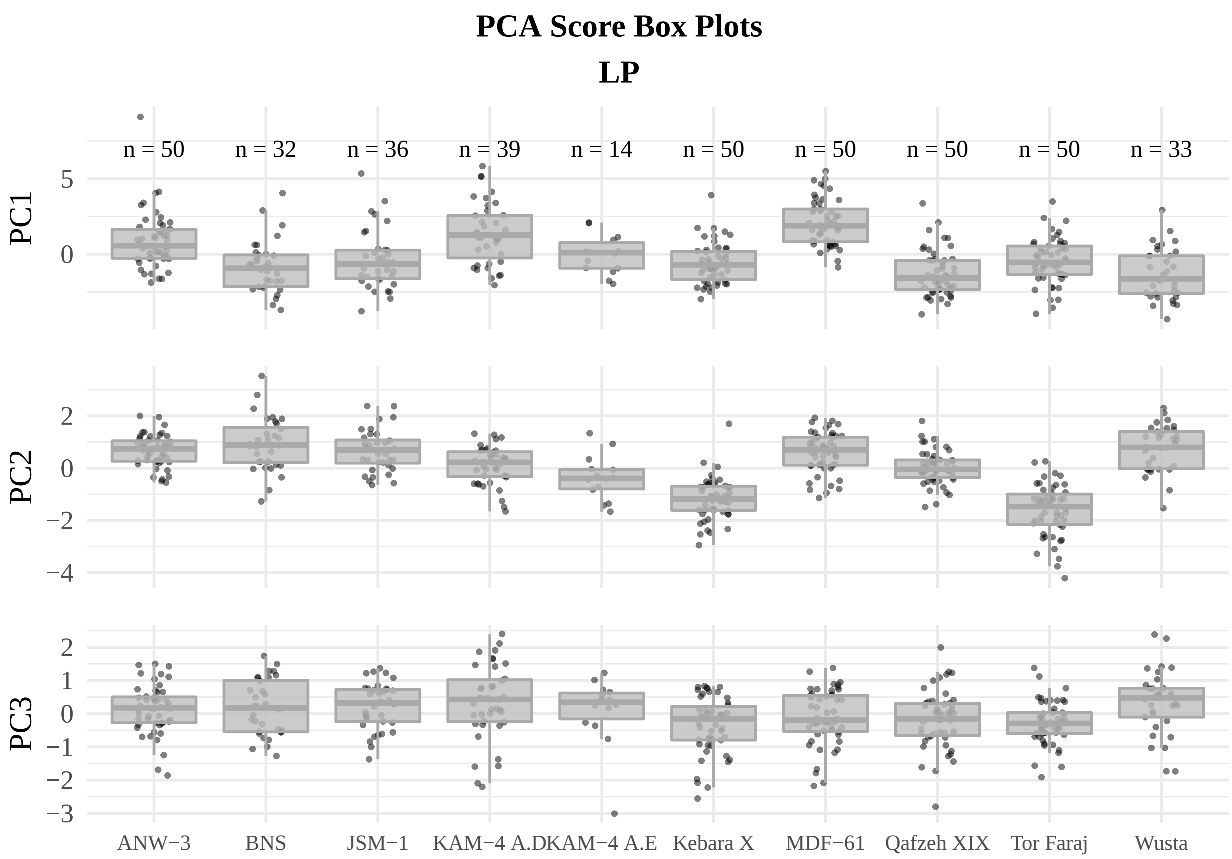
```
LP_scores <- cbind(
                LP[,c(1:2)],
                pca_LP$x)

MIS67_scores <- cbind(
                    MIS67[,c(1:2)],
                    pca_MIS67$x)
```

Ploting

Lastly, we plot the results using ggplot2 as follows. The first plot will contain the results for the analysis of the LP data,

```
sample_name <- "LP"

sample_sizes_LP <- subset(sample_size_assemblage, period == "LP")
sample_sizes_LP$label <- paste("n = ",sample_sizes_LP$n,sep="")

p1 <- ggplot(
        data = get(paste(sample_name,"_scores",sep="")),
        mapping = aes(Assemblage, PC1, group = Assemblage)) +
        geom_jitter(width = 0.15,
                     height = 0,
                     alpha = 0.5,
                     size = 0.5) +
        geom_boxplot(colour = "darkgrey",
                    fill = "grey",
                    alpha = 0.8,
                    outlier.shape = NA) +
        geom_text(data = sample_sizes_LP,
            mapping = aes(x = 1:10, y = 7, label = label),
            size = 3,
            family = "Times") +
        theme_minimal() +
        theme(text = element_text(family="Times", size=12),
            plot.title = element_text(face="bold",hjust=0.5,size=15),
            axis.text.x = element_blank(),
            axis.title.x = element_blank())
```

```
p2 <- ggplot(
        data = get(paste(sample_name,"_scores",sep="")),
        mapping = aes(Assemblage, PC2, group = Assemblage)) +
        geom_jitter(width = 0.15,
                      height = 0,
                      alpha = 0.5,
                      size = 0.5) +
        geom_boxplot(colour = "darkgrey",
                      fill = "grey",
                      alpha = 0.8,
                      outlier.shape = NA) +
        theme_minimal() +
        theme(text = element_text(family="Times", size=12),
            plot.title = element_text(face="bold",hjust=0.5,size=15),
            axis.text.x = element_blank(),
            axis.title.x = element_blank())

p3 <- ggplot(
        data = get(paste(sample_name,"_scores",sep="")),
        mapping = aes(Assemblage, PC3, group = Assemblage)) +
        geom_jitter(width = 0.15,
                      height = 0,
                      alpha = 0.5,
                      size = 0.5) +
        geom_boxplot(colour = "darkgrey",
                      fill = "grey",
                      alpha = 0.8,
                      outlier.shape = NA) +
        theme_minimal() +
        theme(text = element_text(family="Times", size=12),
            plot.title = element_text(face="bold",hjust=0.5,size=15),
            axis.text.x = element_text(size=8),
            axis.title.x = element_blank())

fig <- ggarrange(p1,p2,p3,
            ncol=1,
            nrow=3,
            align="v")

annotate_figure(fig,
               top = text_grob("PCA Score Box Plots\nLP",
                                 family="Times",
                                 face="bold"),
               fig.lab.pos = "top")
```



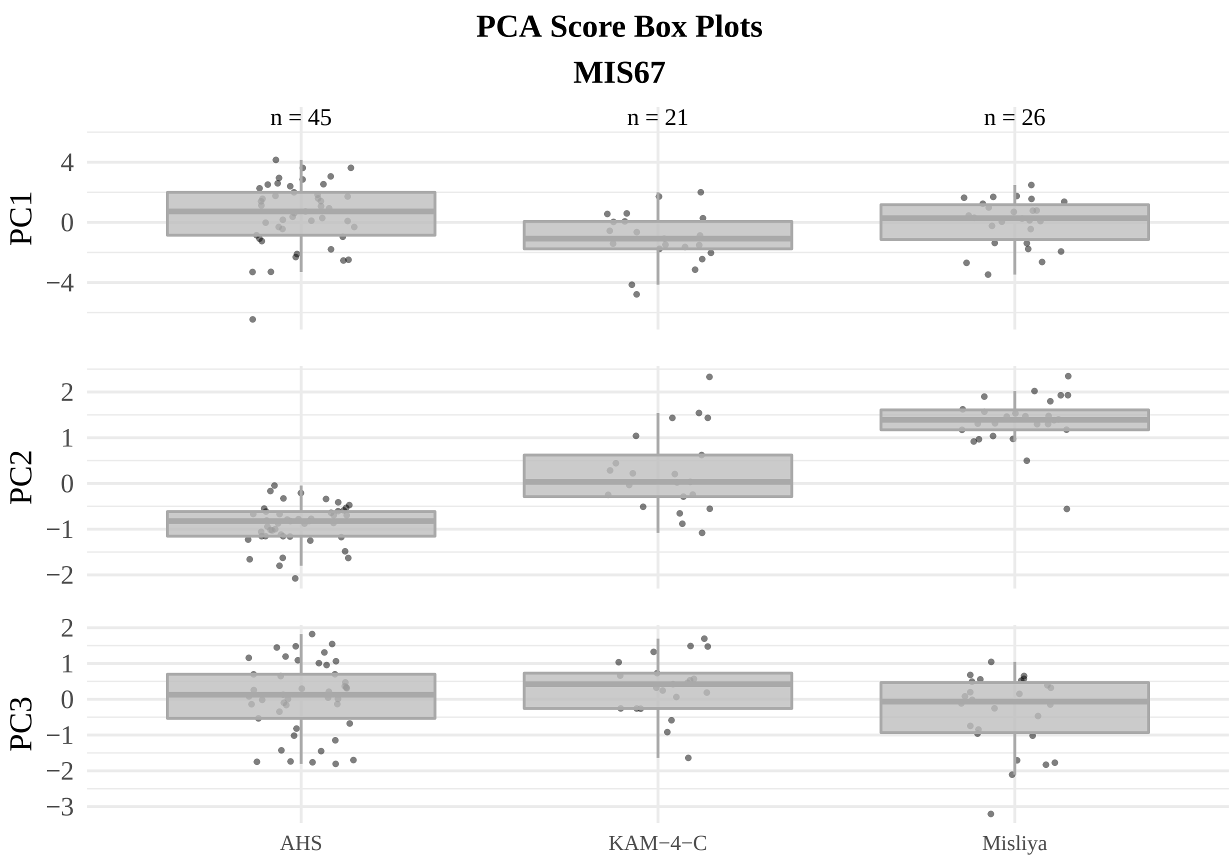

```r
ggsave(filename="./pca_LP_box.pdf",device="pdf")
```

```
## Saving 6.5 x 4.5 in image
```

The second plot contains the results pertaining to the MIS67 data,

```r
sample_name <- "MIS67"

sample_sizes_MIS67 <- subset(sample_size_assemblage, period == "MIS67")
sample_sizes_MIS67$label <- paste("n = ",sample_sizes_MIS67$n,sep="")

p1 <- ggplot(
        data = get(paste(sample_name,"_scores",sep="")),
        mapping = aes(Assemblage,PC1,group = Assemblage)) +
        geom_jitter(width = 0.15,
                    height = 0,
                    alpha = 0.5,
                    size = 0.5) +
        geom_boxplot(colour = "darkgrey",
                   fill = "grey",
                   alpha = 0.8,
                   outlier.shape = NA) +
        geom_text(data = sample_sizes_MIS67,
                mapping = aes(x = 1:3, y = 7, label = label),
                size = 3,
                family = "Times") +
        theme_minimal() +
```

```
        theme(text = element_text(family="Times", size=12),
            plot.title = element_text(face="bold",hjust=0.5,size=15),
            axis.text.x = element_blank(),
            axis.title.x = element_blank())

p2 <- ggplot(
        data = get(paste(sample_name,"_scores",sep="")),
        mapping = aes(Assemblage,PC2,group = Assemblage)) +
        geom_jitter(width = 0.15,
                      height = 0,
                      alpha = 0.5,
                      size = 0.5) +
        geom_boxplot(colour = "darkgrey",
                    fill = "grey",
                    alpha = 0.8,
                    outlier.shape = NA) +
        theme_minimal() +
        theme(text = element_text(family="Times", size=12),
            plot.title = element_text(face="bold",hjust=0.5,size=15),
            axis.text.x = element_blank(),
            axis.title.x = element_blank())

p3 <- ggplot(
        data = get(paste(sample_name,"_scores",sep="")),
        mapping = aes(Assemblage,PC3,group = Assemblage)) +
        geom_jitter(width = 0.15,
                      height = 0,
                      alpha = 0.5,
                      size = 0.5) +
        geom_boxplot(colour = "darkgrey",
                    fill = "grey",
                    alpha = 0.8,
                    outlier.shape = NA) +
        theme_minimal() +
        theme(text = element_text(family="Times", size=12),
            plot.title = element_text(face="bold",hjust=0.5,size=15),
            axis.text.x = element_text(size=8),
            axis.title.x = element_blank())

fig <- ggarrange(p1,p2,p3,
            ncol=1,
            nrow=3,
            align="v")

annotate_figure(fig,
               top = text_grob("PCA Score Box Plots\nMIS67",
                                 family="Times",
                                 face="bold"),
               fig.lab.pos = "top")
```




```r
ggsave(filename="./pca_MIS67_box.pdf",device="pdf")
```

```
## Saving 6.5 x 4.5 in image
```