STAA 574: Homework 4

Spring 2020

Jon Dollard

Use R-Markdown to organize your work.

You will need to carry out a principal analysis on two data sets. You should type up approximately a 1-2 page summary. At a minimum, this will include:

- Give the covariance and correlation matrices used for the analysis.
- Make a decision on using the covariance or correlation matrix and justify your decision.
- Describe the eigenvalues and fractional contributions to the variance.
- Describe the criteria you used for how many PC's you decide to keep.
- Give the eigenvectors for the principal components you retain.
- Make appropriate plots of pairs of principal components. Make observations about the plots.
- Use any additional plots or analysis to make conclusions if appropriate.

The data sets that are to be analyzed are on Canvas and are:

1. Flea Beetle Data
2. Mali Farm Data

The flea beetle dataset has two groups/species. Run a principal component analysis on each group individually for each dataset. Compare your results from the two groups.

You will only need to run one PCA for the Mali farm dataset.

For problem 1 we will complete principal component analysis (PCA) for a flea beetle data set that contains data for 2 species of flea beetle. The first PCA will be for Haltica oleracea and the second for H. carduourum. For Haltica oleracea we begin by reading in the data set to R to assist us with the necessary calculations and visualization. Next, we plot the pairs of variables to visually check the data for correlation, multicollinearity, and determine if PCA is appropriate for the data set.

```
#read in the flea beetle data set and manipulate it to a usable form
Flea_Beetle_Data <- read_excel("fleabeetledata.xlsx")
Haltica_Oleracea_Data <- Flea_Beetle_Data[1:19,2:5]
Haltica_Oleracea_Data <- data.frame(Haltica_Oleracea_Data)
colnames(Haltica_Oleracea_Data) <- c("TG", "Elytra", "Second Antenna","Third Antenna")
head(Haltica_Oleracea_Data)
```

```
##    TG Elytra Second Antenna Third Antenna
## 1 189    245            137           163
## 2 192    260            132           217
## 3 217    276            141           192
## 4 221    299            142           213
## 5 171    239            128           158
## 6 192    262            147           173
```

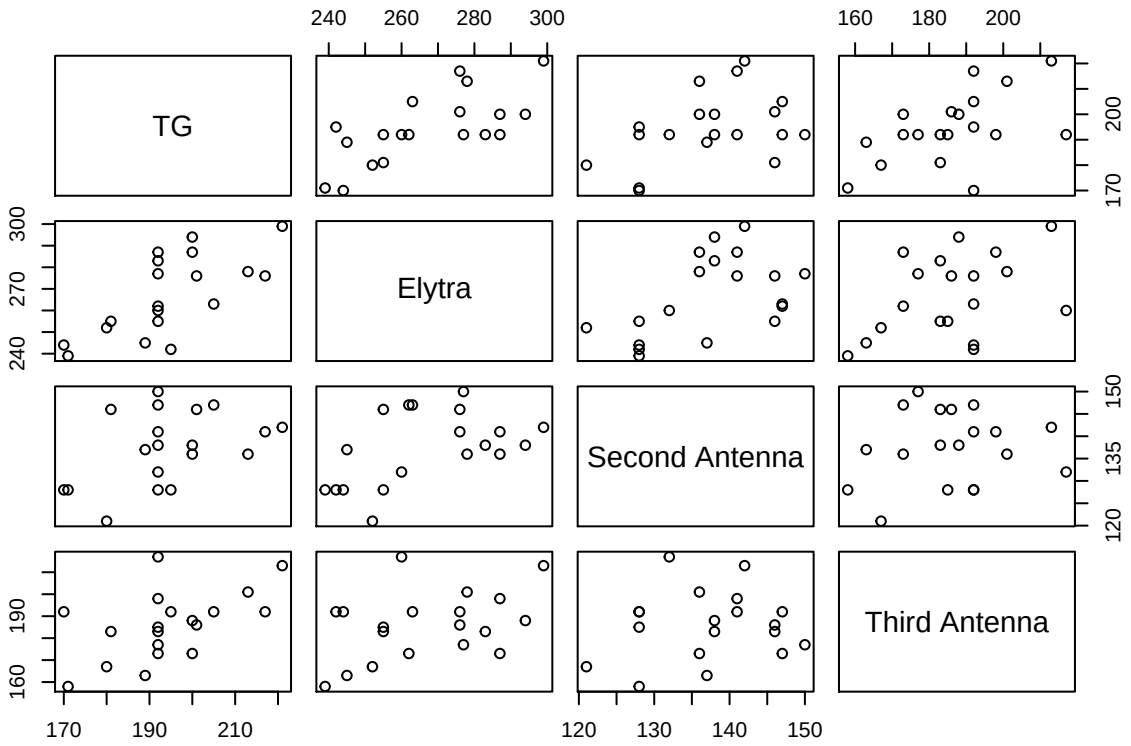

Figure 1: Pairs Plot of Haltica oleracea variables TG, Elytra, Second Antenna, and Third Antenna observations.

From the pairs plot it is clear that the variables in the Haltica oleracea data set are correlated and that multicollinearity exists in the data. Therefore, PCA is an appropriate analysis technique for the data. We can also inspect the covariance and correlation matrices to further quantify correlation and multicollinearity. We will use R to calculate the covariance and correlation matrices.

```r
#calculate the covariance matrix for the Haltica Oleracea data set
S_HO <- var(Haltica_Oleracea_Data)
round(S_HO, 1)
```

```
##                    TG Elytra Second Antenna Third Antenna
## TG              187.6  176.9           48.4          113.6
## Elytra          176.9  345.4           76.0          118.8
## Second Antenna   48.4   76.0           66.4           16.2
## Third Antenna   113.6  118.8           16.2          239.9
```

```r
#kable(S_HO, digits = 1, align = 'c')
```

$$\mathbf{S}_{H.oleracea} = \begin{bmatrix} 187.6 & 176.9 & 48.4 & 113.6 \\ 176.9 & 345.4 & 76.0 & 118.8 \\ 48.4 & 76.0 & 66.4 & 16.2 \\ 113.6 & 118.8 & 16.2 & 239.9 \end{bmatrix}$$

```r
#calculate the correlation matrix for the Haltica Oleracea data set
Corr_HO <- cor(Haltica_Oleracea_Data)
round(Corr_HO, 2)
```

```
##                  TG Elytra Second Antenna Third Antenna
## TG             1.00   0.69           0.43          0.54
## Elytra         0.69   1.00           0.50          0.41
## Second Antenna 0.43   0.50           1.00          0.13
## Third Antenna  0.54   0.41           0.13          1.00
```

$$\mathbf{R}_{H.oleracea} = \begin{bmatrix} 1.00 & 0.69 & 0.43 & 0.54 \\ 0.69 & 1.00 & 0.50 & 0.41 \\ 0.43 & 0.50 & 1.00 & 0.13 \\ 0.54 & 0.41 & 0.13 & 1.00 \end{bmatrix}$$

From the covariance matrix we can see that the first 2 variables, TG and Elytra, both account for a significant portion of the variance. From the correlation matrix it is now very clear that the variables are correlated and multicollinearity between variables exists. Since the data for Haltica oleracea includes measurements of different units (TG, Second Antenna, and Third Antenna are measured in microns whereas Elytra is in 0.01mm) we will proceed with PCA using standardized data and the correlation matrix.

Using R and the built in function, princomp, we can very easily generate the eigenvalues and eigenvectors for the principal components.

```
HO_PC_S <- princomp(Haltica_Oleracea_Data, cor=TRUE)
sum_HO_PC_S <- summary(HO_PC_S, loadings = TRUE)
sum_HO_PC_S
```

```
## Importance of components:
##                            Comp.1    Comp.2    Comp.3     Comp.4
## Standard deviation     1.5476548 0.9417681 0.6571892 0.53473353
## Proportion of Variance 0.5988088 0.2217318 0.1079744 0.07148499
## Cumulative Proportion  0.5988088 0.8205406 0.9285150 1.00000000
##
## Loadings:
##                 Comp.1 Comp.2 Comp.3 Comp.4
## TG               0.573  0.112  0.310  0.750
## Elytra           0.562 -0.117  0.525 -0.629
## Second Antenna   0.419 -0.700 -0.578
## Third Antenna    0.425  0.695 -0.543 -0.203
```

Observing the output from the princomp function we can assess the eigenvalues fractional contributions to the variance of the data set. Principal component 1 (PC1) accounts for about 60% of the total variance. PC2 accounts for about 22%, PC3 about 11%, and PC4 about 7%. The first 2 PC's account for over 80% of the total variance in the Haltica oleracea data. We can make a scree plot to help us decide how many PC's we will keep in this PCA.

```
screeplot(HO_PC_S, main = "Scree Plot of Principal Components")
```

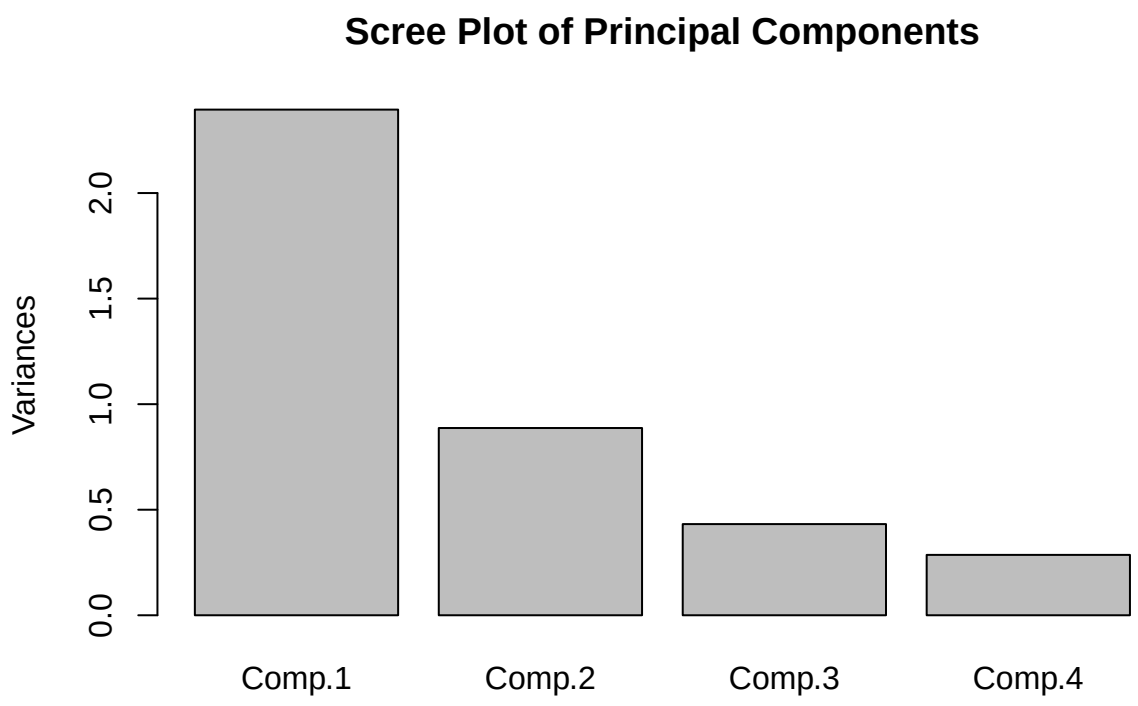

Figure 2: Scree plot of the principal components for the Haltica oleracea data set.

From the scree plot it would be reasonable to choose to keep the first 2 PCs. The elbow of this plot appears to happen at PC2. Also, we know from our previous discussion of the fractional contribution that the first two PCs account for over 80% of the total variance in the data. Therefore, we will keep the first 2 PCs for this PCA.

The eigenvectors for PC1 and PC2 are shown below:

```
loadings_matrix <- cbind(c(0.573, 0.562, 0.419, 0.424), c(0.112, -0.117, -0.700, 0.695))
rownames(loadings_matrix) <- c("TG", "Elytra", "Second Antenna", "Third Antenna")
colnames(loadings_matrix) <- c("PC1", "PC2")
kable(loadings_matrix, align = 'c')
```

	PC1	PC2
TG	0.573	0.112
Elytra	0.562	-0.117
Second Antenna	0.419	-0.700
Third Antenna	0.424	0.695

Since I chose to perform the PCA using standardized data the interpretation becomes a bit more difficult than using non-standardized data. However, we can draw some conclusions about the PC's.

Each variable in PC1 appears to contribute a similar proportion to the total. Therefore, it seems as though the first PC is describing the overall size of the beetle and that each variable contributes a similar proportion to the size. If a beetle has large TG, Elytra, Second and Third Antenna measurements then this would tend to be a larger than average beetle.

The second PC, PC2, appears to represent a contrast between the second and third antenna measurements. The contrast between TG and Elytra appear small and potentially neglible for this analysis. Therefore, I think we can conclude that a contrast between the second and third antenna measurements exist for the Haltica oleracea beetle. For example, if the beetle has a small second antenna measurement we would expect that it would have a large third antenna measurement and vice versa.

In order to visualize our PCA we can graph PC1 and PC2 and show the vectors associated with each.

```
ggbiplot(HO_PC_S, labels = rownames(Haltica_Oleracea_Data)) +
  ggtitle("Principal Component 1 vs. Principal Component 2") +
  theme(plot.title = element_text(hjust = 0.5))
```

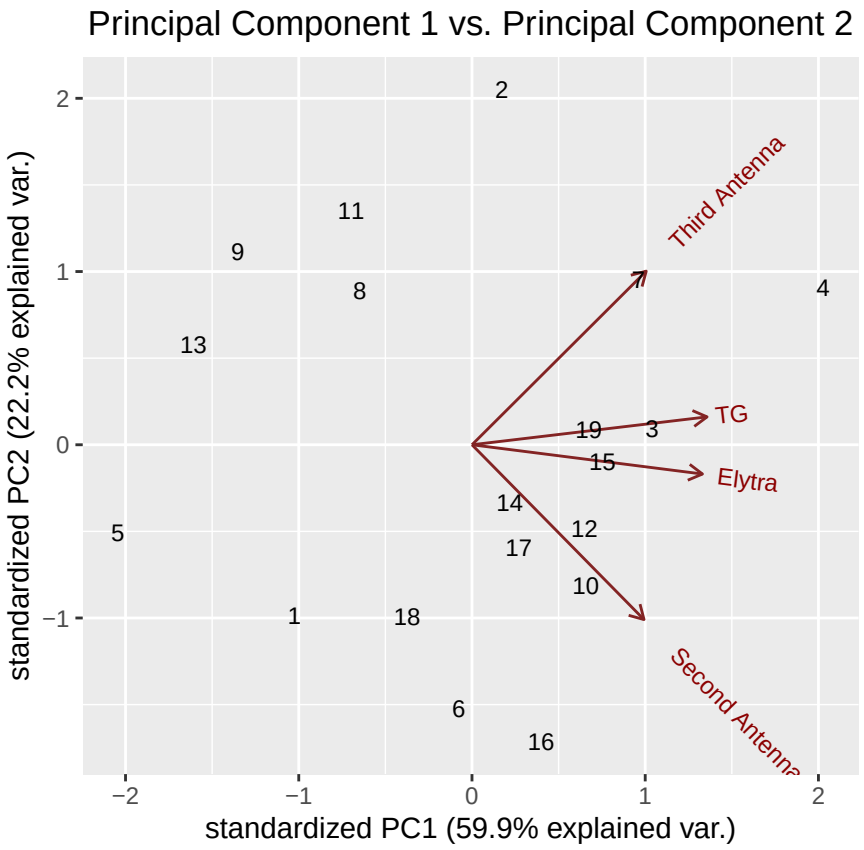

Figure 3: PC biplot of PC1 and PC2 for the Haltica oleracea data set.

Now that we have completed the PCA and chosen the PC's to keep in the analysis, we want to look at some particular observations to determine if our conclusions about the PC's are consistent with what we see in the data set. First, we will calculate the means for each variable.

```
HO_means <- colMeans(Haltica_Oleracea_Data)
HO_means <- as.matrix(HO_means)
colnames(HO_means) <- c("Mean")
kable(HO_means, digits = 2, align = 'c')
```

	Mean
TG	194.47
Elytra	267.05
Second Antenna	137.37
Third Antenna	185.95

One data point we can examine would be observation number 4. We would expect from our PCA and PC plot above that observation 4 would represent a large beetle or in other words a beetle that has larger than average measurements for all 4 variables. The values for TG, Elytra, Second Antenna, and the Third Antenna for observation 4 are 221, 229, 142 and 213 respectively. Clearly, all values are larger than the average as we expected.

Another data point of interest is observation number 2. From it's position on the PC biplot we would expect this beetle to have larger than average third antenna and a correspondingly smaller than average second antenna measurement. Recall that PC2 represents a contrast between those measurements so a very large measurement for one would have a correspondingly small measurement in the other. The TG and Elytra measurements we would expect to be about average. The values for TG, Elytra, Second Antenna, and the Third Antenna for observation 2 are 192, 260, 132 and 217 respectively. These values for observation 2 confirm our expectation.

In contrast to observation 2 we can investigate observation 6. Observation 6 roughly falls on the opposite end of the PC biplot from observation 2. Since PC2 is a contrast then in this case we would expect the second antenna measurement for Observation 6 to be larger than average and the corresponding measurement for the third antenna to be smaller than average. Similar to observation 2 we would expect the measurements for TG and Elytra to be about average. The values for TG, Elytra, Second Antenna, and the Third Antenna for observation 6 are 192, 262, 147 and 173 respectively. The values we observe for observation 6 are consistent with what we expect from our PCA.

Another point of interest is observation 5. In contrast to observation 4 this appears to be a beetle with smaller than average measurements for each variable. This might indicate a smaller than average beetle. The values for TG, Elytra, Second Antenna, and the Third Antenna for observation 5 are 171, 239, 128 and 158 respectively. Clearly, all values are smaller than the average as we expected.

Finally, let's investigate observation 14. This beetle appears to be very close to the center of PC biplot. This would be an indication that this beetle has essentially average measurements for all 4 variables. The values for TG, Elytra, Second Antenna, and the Third Antenna for observation 5 are 192, 283, 138 and 183 respectively. With the exception of the Elytra measurement all of the values are very near average. For the Elytra measurement we may be seeing a bit of the contrast of PC2 being picked up here.

PCA is a descriptive technique for data analysis and we can see from our evaluation of a few observations that our PCA for Haltica oleracea, using PC1 and PC2, is descriptive of the data. Our expectation of what the measurements for a given beetle should be based on the PC biplot or the PC eigenvectors is consistent with what we find when we investigate that individual observation. Since, this is the case I am satisfied with the result of this PCA.

The second PCA for problem 1 will be for Haltica carduourum.

For Haltica carduourum we begin by reading in the data set to R to assist us with the necessary calculation and visualization. We begin the PCA by plotting the pairs of variables to visually check the data for correlation, multicollinearity, and determine if PCA is appropriate for the data set.

```
Flea_Beetle_Data <- read_excel("fleabeetledata.xlsx")
Haltica_carduourum_Data <- Flea_Beetle_Data[1:20,6:9]
Haltica_carduourum_Data <- data.frame(Haltica_carduourum_Data)
colnames(Haltica_carduourum_Data) <- c("TG", "Elytra", "Second Antenna","Third Antenna")
head(Haltica_carduourum_Data)
```

```
##    TG Elytra Second Antenna Third Antenna
## 1 181    305             184            209
## 2 158    237             133            188
## 3 184    300             166            231
## 4 171    273             162            213
## 5 181    297             163            224
## 6 181    308             160            223
```

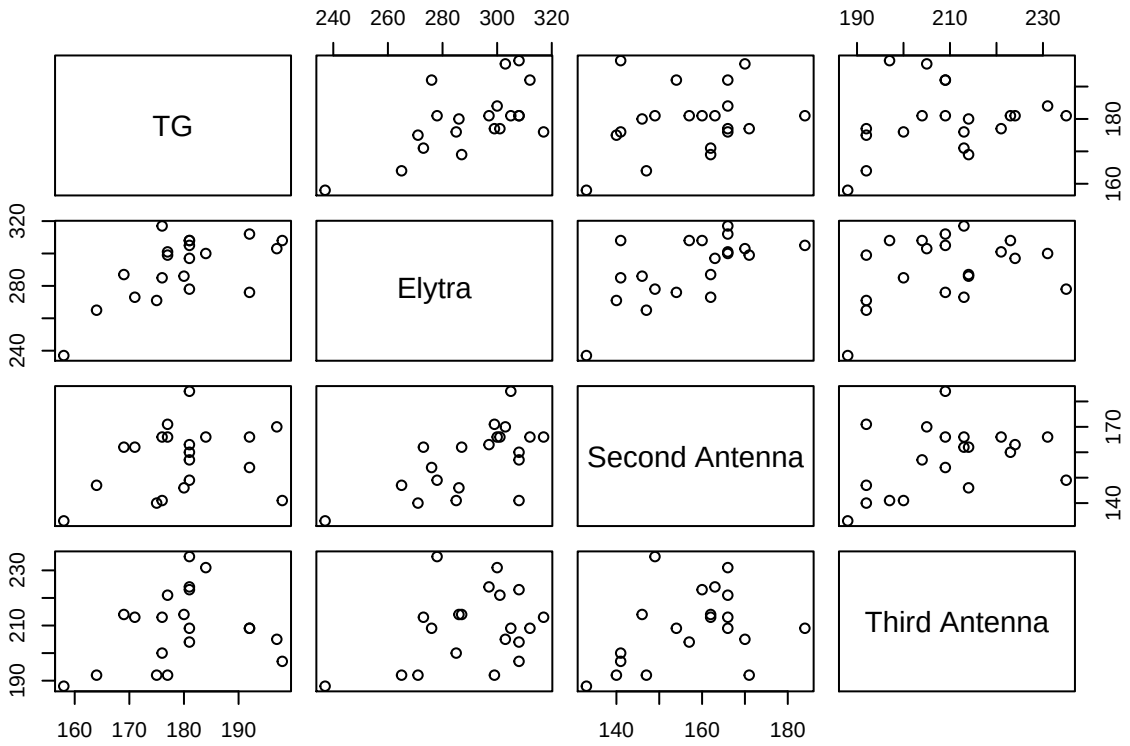

Figure 4: Pairs Plot of Haltica oleracea variables TG, Elytra, Second Antenna, and Third Antenna observations.

From the pairs plot it is clear that the variables in the Haltica carduourum data set are correlated and that multicollinearity exists in the data. Therefore, PCA is an appropriate analysis for the data. We can also inspect the covariance and correlation matrices to quantify correlation and multicollinearity.

We will use R to calculate the covariance and correlation matrices.

```r
#calculate the covariance matrix for the Haltica carduourum data set
S_Hc <- var(Haltica_carduourum_Data)
round(S_Hc,1)
```

```
##                    TG Elytra Second Antenna Third Antenna
## TG              101.8  128.1           37.0          32.6
## Elytra          128.1  389.0          165.4          94.4
## Second Antenna   37.0  165.4          167.5          66.5
## Third Antenna    32.6   94.4           66.5         177.9
```

$$\mathbf{S}_{H.corduourum} = \begin{bmatrix} 101.8 & 128.1 & 37.0 & 32.6 \\ 128.1 & 389.0 & 165.4 & 94.4 \\ 37.0 & 165.4 & 167.5 & 66.5 \\ 32.6 & 94.4 & 66.5 & 177.9 \end{bmatrix}$$

```r
#calculate the correlation matrix for the Haltica carduourum data set
Corr_Hc <- cor(Haltica_carduourum_Data)
round(Corr_Hc, 2)
```

```
##                    TG Elytra Second Antenna Third Antenna
## TG              1.00   0.64           0.28          0.24
## Elytra          0.64   1.00           0.65          0.36
## Second Antenna 0.28   0.65           1.00          0.39
## Third Antenna  0.24   0.36           0.39          1.00
```

$$\mathbf{R}_{H.oleracea} = \begin{bmatrix} 1.00 & 0.64 & 0.28 & 0.24 \\ 0.64 & 1.00 & 0.65 & 0.36 \\ 0.28 & 0.65 & 1.00 & 0.39 \\ 0.24 & 0.36 & 0.39 & 1.00 \end{bmatrix}$$

From the covariance matrix we can see that the Elytra dominates the amount of variance accounted for in the data. It is important to note here that the Elytra has different units of measurement than the other 3 variables. From the correlation matrix we see that the variables are correlated and multicollinearity between variables exists. Since the data for Haltica carduourum includes measurements of different units (TG, Second Antenna, and Third Antenna are measured in microns whereas Elytra is in 0.01mm) we will proceed with PCA using standardized data and the correlation matrix.

Using R and built in function, princomp, we can very easily generate the eigenvalues and eigenvectors for the principal components.

```
Hc_PC_S <- princomp(Haltica_carduourum_Data, cor=TRUE)
summary(Hc_PC_S, loadings = TRUE)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4
## Standard deviation     1.5212946 0.9099492 0.8002720 0.46606860
## Proportion of Variance 0.5785843 0.2070019 0.1601088 0.05430499
## Cumulative Proportion  0.5785843 0.7855862 0.9456950 1.00000000
##
## Loadings:
##                 Comp.1 Comp.2 Comp.3 Comp.4
## TG               0.475  0.615  0.433  0.457
## Elytra           0.594  0.224 -0.173 -0.753
## Second Antenna   0.512 -0.281 -0.660  0.472
## Third Antenna    0.400 -0.702  0.588
```

Observing the output from the princomp function we can assess the eigenvalues fractional contributions to the variance of the data set. Principal component 1 (PC1) accounts for about 58% of the total variance. PC2 accounts for about 21%, PC3 about 16%, and PC4 about 5%. The first 2 PCs account for almost 80% of the total variance in the Haltica carduourum data. PC3 does appear to be a bit more significant in this PCA than for Haltica oleracea. We can make a scree plot to help us decide how many PC's we will keep in this PCA.

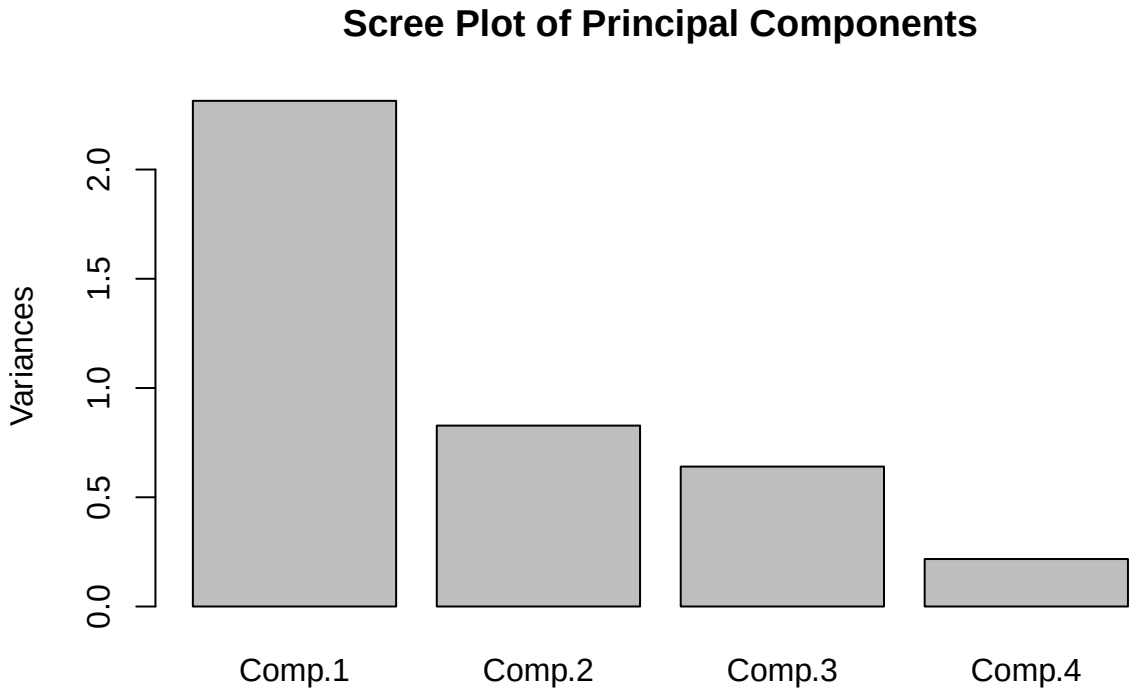

Figure 5: Scree plot of the principal components for the Haltica corduourum data set

From the scree plot it would be reasonable to choose to keep the first 2 PCs. The elbow of this plot appears

to be at PC2. Also, we know from our previous discussion of the fractional contribution that the first two PCs account for about 78% of the total variance in the data. Investigating PC3 a little further we see that we seem to get similar information from PC3 as we do from PC2, but in a slightly different way (different contrasts, but similar conclusions about the data). Since it doesn't appear that much information about the data set is lost by excluding PC3 we will keep the first 2 PCs for this PCA. If during our review of points of interest (at the conclusion of this report) we find that the PCA with only PC1 and PC2 is not adequately descriptive we will need to revist this part of the analysis and include PC3.

The eigenvectors for PC1 and PC2 are:

	PC1	PC2
TG	0.475	0.615
Elytra	0.594	0.224
Second Antenna	0.512	-0.281
Third Antenna	0.400	-0.702

Since I chose to perform the PCA using standardized data the interpretation becomes a bit more difficult than using non-standardized data. However, we can draw some conclusions about the PC's.

Each variable in PC1 appears to contribute a similar proportion to the total with the Elytra being the most dominant. Therefore, we can conclude that the first PC is describing the overall size of the beetle and that each variable contributes a similar proportion to the size. If a beetle had large TG, Elytra, Second and Third Antenna measurements then this would tend to be a larger beetle. This is not surprising as this was the same conclusion we reached for PC1 in the Haltica oleracea PCA.

The second PC, PC2, appears to represent a contrast between the TG and third antenna measurements and to a lesser extent a contrast between the Elytra and second antenna measurements. The contrast between TG and Elytra is significantly smaller and potentially neglible. Therefore, our conclusion is that a contrast between TG and the third antenna measurements exist for the Haltica carduourum beetle. For example, if the beetle has a small third antenna measurement we would expect that it would have a large TG measurement and vice versa.

In order to visualize our PCA we can graph PC1 and PC2 and show the vectors associated with each.

```
ggbiplot(Hc_PC_S, labels = rownames(Haltica_carduourum_Data)) +
  ggtitle("Principal Component 1 vs. Principal Component 2") +
  theme(plot.title = element_text(hjust = 0.5))
```

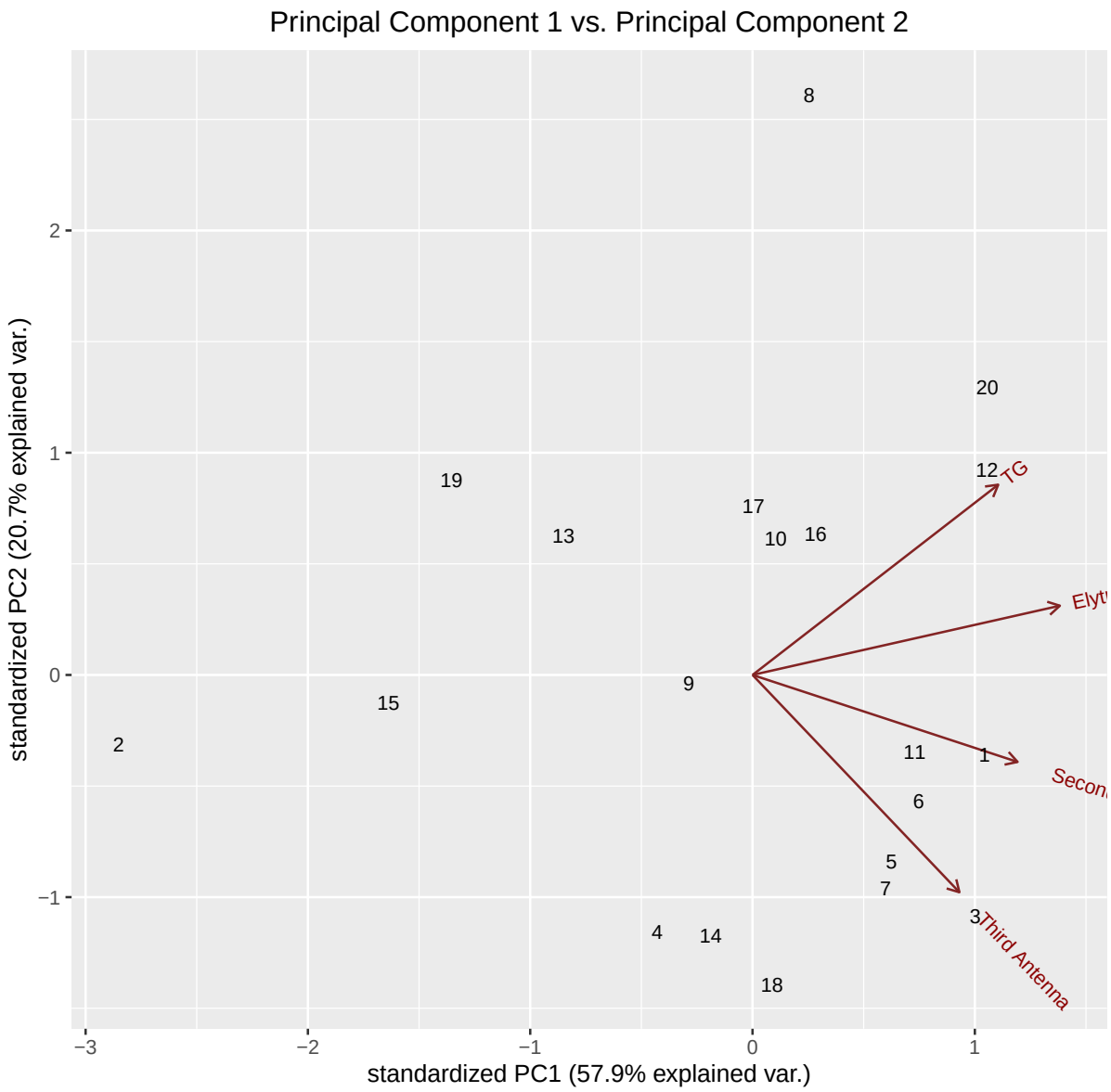

Figure 6: PC biplot for PC1 and PC2

Now that we have completed the PCA and chosen the PC's to keep in the analysis, we want to look at some particular points to determine if our conclusions about PC's are consistent with what we see in the data set. First, we will calculate the means for each variable.

```
Hc_means <- colMeans(Haltica_carduourum_Data)
Hc_means <- as.matrix(Hc_means)
colnames(Hc_means) <- c("Mean")
kable(Hc_means, digits = 2, align = 'c')
```

	Mean
TG	179.55
Elytra	290.80
Second Antenna	157.20
Third Antenna	209.25

One data point of interest is observation number 8. We would expect from our PCA and PC's plot above that observation 8 would have a larger than average TG measurement and a correspondingly smaller average third antenna measurement due to the contrast of these variables. We would also anticipate to see some of the effect of the other contrast that this beetle will have a larger than average Elytra and smaller than average second antenna measurement. The values for TG, Elytra, Second Antenna, and the Third Antenna for observation 8 are 198, 308, 141 and 197 respectively. With regard to observation 8 we see what we would expect.

Another data point of interest is observation number 2. From it's position on the PC biplot we would expect this beetle to have smaller than average measurements for all variables. This indicates that this is a smaller than average beetle. The values for TG, Elytra, Second Antenna, and the Third Antenna for observation 2 are 158, 237, 133 and 188 respectively. These values for observation 2 confirm our expectation.

In contrast to observation 8 we can investigate observation 14. Observation 14 roughly falls on the opposite end of the PC biplot from observation 8. Since PC2 is a contrast, then in this case we would expect the third antenna measurement for Observation 14 to be larger than average and the corresponding measurement for TG to be smaller than average. We also expect to see some of the contrast between the Elytra and second antenna. However, for this observation the second antenna measurement we expect to be larger than average and the Elytra to be smaller than average. The values for TG, Elytra, Second Antenna, and the Third Antenna for observation 14 are 169, 287, 162 and 214 respectively. The values we observe for observation 14 are consistent with what we expect from our PCA.

Finally, for this PCA let's investigate observation 9 further. Observation 9 falls very close to zero for both PCs. This is an indication to us that this beetle is of average size and we would expect to see measurements for this beetle close to the sample averages for each variable. The values for TG, Elytra, Second Antenna, and the Third Antenna for observation 9 are 180, 286, 146 and 214 respectively. These values fall close to the sample average values as expected.

After investigating a few observations of interest I will conclude that the Haltica carduourum PCA, using PC1 and PC2, is descriptive of the data. Our expectation of what the measurements for a given beetle should be based on the PC biplot and the PC eigenvectors is consistent with what we find when we investigate that individual observation. Since, this is the case I am satisfied with the result of this PCA.

Comparison of the both beetle species:

1) It was interesting, but not surprising, that for both beetle species the first principal component was a description of the overall size of the beetle.

2) The contrasts of the second principal components were different. The second principal component for the Haltica oleracea beetle showed a contrast between the second and third antenna measurements. The second principal component for the Haltica corduourum beetle was a contrast between the TG and

third antenna measurements and to a lesser degree the Elytra and second antenna measurements. I doesn't seem unreasonable that between species we would find the PCA descriptions of the data varies as we add more principal components.

3) Overall I thought that keeping two principal component did a very good job describing the data. It was nice to be able to reduce the data dimensions in half (from 4 variables to 2) and still retain the ability to describe the data set with reasonable accuracy.

Problem 2 will be a PCA for the Mali Farm data set.

First, we will read in the Mali farm data using R. We begin the PCA by plotting the pairs of variables to visually check the data for correlation, multicollinearity, and determine if PCA is appropriate for the data set.

```
Mali_Farm_Data <- read_excel("malifarmdata.xlsx")
Mali_Farm_Data <- Mali_Farm_Data[1:76,2:10]
Mali_Farm_Data <- data.frame(Mali_Farm_Data)
colnames(Mali_Farm_Data) <- c("Family", "DistRD", "Cotton","Maize", "Sorg",
                              "Millet", "Bull", "Cattle", "Goats")
head(Mali_Farm_Data)
```

```
##   Family DistRD Cotton Maize Sorg Millet Bull Cattle Goats
## 1     12     80    1.5   1.0    3   0.25    2      0     1
## 2     54      8    6.0   4.0    0   1.00    6     32     5
## 3     11     13    0.5   1.0    0   0.00    0      0     0
## 4     21     13    2.0   2.5    1   0.00    1      0     5
## 5     61     30    3.0   5.0    0   0.00    4     21     0
## 6     20     70    0.0   2.0    3   0.00    2      0     3
```

```
pairs(Mali_Farm_Data)
```

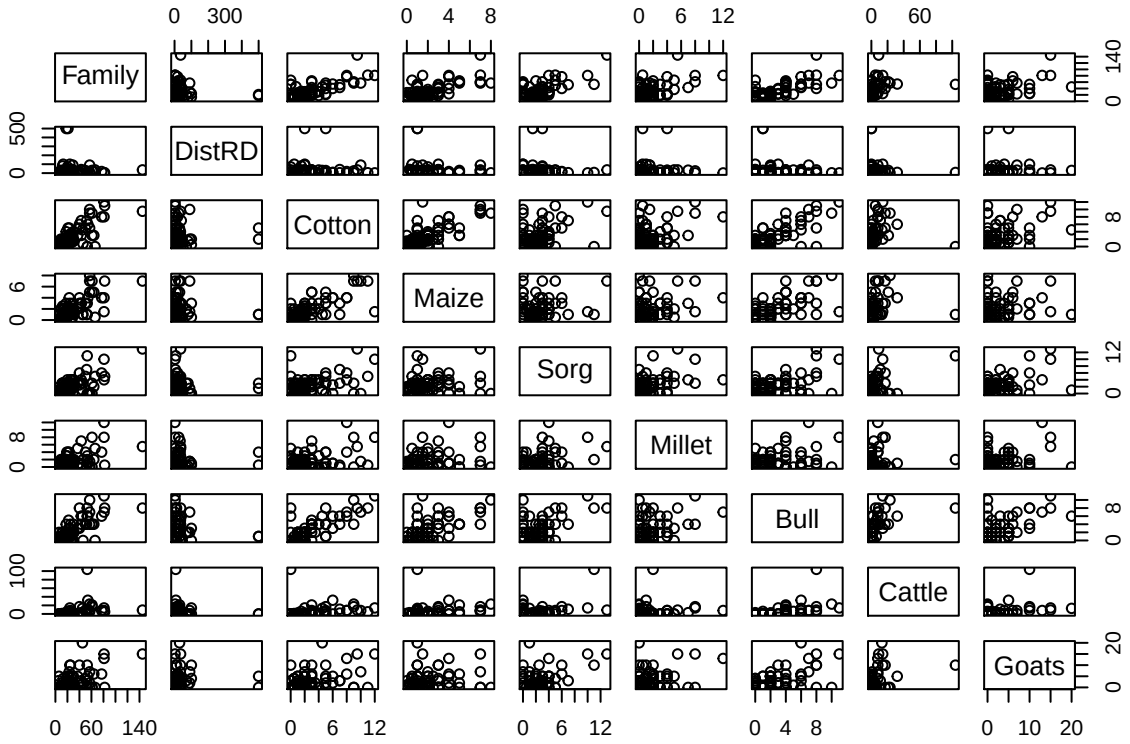

Figure 7: Pairs Plot of the Mali Farm data set.

From the pairs plot we can see that there is correlated data as well as multicollinearity. A few of the variables clearly contain, at first glance, the presence of outliers in the data. For example DistRd and Cattle both appear to have outliers in the data. We will use R to calculate the covariance and correlation matrices and further assess correlation and multicollinearity.

```r
#calculate the covariance matrix for the Mali Farm data set
S_Mf <- var(Mali_Farm_Data)
round(S_Mf,2)
```

```
##         Family  DistRD Cotton Maize   Sorg Millet   Bull Cattle Goats
## Family  550.88 -158.77  48.12 29.54  31.84  26.39  45.46 103.75 46.81
## DistRD -158.77 6533.75   6.44 -8.11 -13.69   3.94 -19.02 -67.35 10.36
## Cotton   48.12    6.44   8.01  3.83   2.58   2.45   5.76   6.50  4.65
## Maize    29.54   -8.11   3.83  3.43   0.48   0.89   3.07   4.81  1.04
## Sorg     31.84  -13.69   2.58  0.48   5.70   2.03   2.82  12.70  4.17
## Millet   26.39    3.94   2.45  0.89   2.03   4.94   2.09   2.37  2.80
## Bull     45.46  -19.02   5.76  3.07   2.82   2.09   7.09  18.21  6.15
## Cattle  103.75  -67.35   6.50  4.81  12.70   2.37  18.21 173.08 19.36
## Goats    46.81   10.36   4.65  1.04   4.17   2.80   6.15  19.36 17.01
```

```r
#calculate the correlation matrix for the Mali farm data set
Corr_Mf <- cor(Mali_Farm_Data)
round(Corr_Mf, 2)
```

```
##        Family DistRD Cotton Maize  Sorg Millet  Bull Cattle Goats
## Family   1.00  -0.08   0.72  0.68  0.57   0.51  0.73   0.34  0.48
## DistRD  -0.08   1.00   0.03 -0.05 -0.07   0.02 -0.09  -0.06  0.03
## Cotton   0.72   0.03   1.00  0.73  0.38   0.39  0.76   0.17  0.40
## Maize    0.68  -0.05   0.73  1.00  0.11   0.22  0.62   0.20  0.14
## Sorg     0.57  -0.07   0.38  0.11  1.00   0.38  0.44   0.40  0.42
## Millet   0.51   0.02   0.39  0.22  0.38   1.00  0.35   0.08  0.31
## Bull     0.73  -0.09   0.76  0.62  0.44   0.35  1.00   0.52  0.56
## Cattle   0.34  -0.06   0.17  0.20  0.40   0.08  0.52   1.00  0.36
## Goats    0.48   0.03   0.40  0.14  0.42   0.31  0.56   0.36  1.00
```

From the covariance matrix we can see that the family size (Family) and the distance from the nearest passable road (DistRD) dominates the amount of variance accounted for in the data. From the correlation matrix we see that the variables are correlated and multicollinearity between variables exists. Since the data for the Mali Farm includes measurements of different units we will proceed with PCA using standardized data and the correlation matrix. Another reason to standardize our data is that we have data that is related to family size and geography as well as data that is based on the farm's agricultural output and size. If we didn't standardize the date we may not be able to fairly detect the influence of the farm's agricultural output and size since Family and DistRD dominate the variance of the data set.

Using R and built in function, princomp, we can very easily generate the eigenvalues and eigenvectors for the principal components.

```
Mf_PC_S <- princomp(Mali_Farm_Data, cor=TRUE)
summary(Mf_PC_S, loadings = TRUE)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     2.0357507 1.1119255 1.0286536 0.9594127 0.77833471
## Proportion of Variance 0.4604757 0.1373754 0.1175698 0.1022747 0.06731166
## Cumulative Proportion  0.4604757 0.5978510 0.7154209 0.8176956 0.88500726
##                            Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     0.71023781 0.52156176 0.38344629 0.33382508
## Proportion of Variance 0.05604864 0.03022518 0.01633678 0.01238213
## Cumulative Proportion  0.94105590 0.97128108 0.98761787 1.00000000
##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## Family  0.444                0.123         0.127  0.579  0.454  0.461
## DistRD                0.831 -0.502  0.194
## Cotton  0.411  0.342               -0.100  0.216 -0.509 -0.372  0.504
## Maize   0.337  0.554 -0.170 -0.164  0.134         0.352 -0.360 -0.499
## Sorg    0.311 -0.452         0.229  0.361  0.632        -0.139 -0.300
## Millet  0.269         0.385  0.606  0.182 -0.594
## Bull    0.440        -0.122 -0.197 -0.129 -0.110 -0.458  0.621 -0.357
## Cattle  0.247 -0.458 -0.278 -0.486  0.392 -0.407        -0.215  0.225
## Goats   0.309 -0.379  0.173 -0.100 -0.770         0.242 -0.242
```

Investigating the output from the princomp function we can assess the eigenvalues fractional contributions to the variance of the data set. Principal component 1 (PC1) accounts for about 46% of the total variance. PC2 accounts for about 14%, PC3 about 12%, and PC4 about 10%. The remaining 4 PCs account for the remaining 18% in decreasing order of importance. The first 4 PCs account for about 82% of the total variance in the Mali farm data. We can make a scree plot to help us decide how many PC's we will keep in this PCA using the "elbow" method.

```
screeplot(Mf_PC_S, main = "Scree Plot of Principal Components")
```

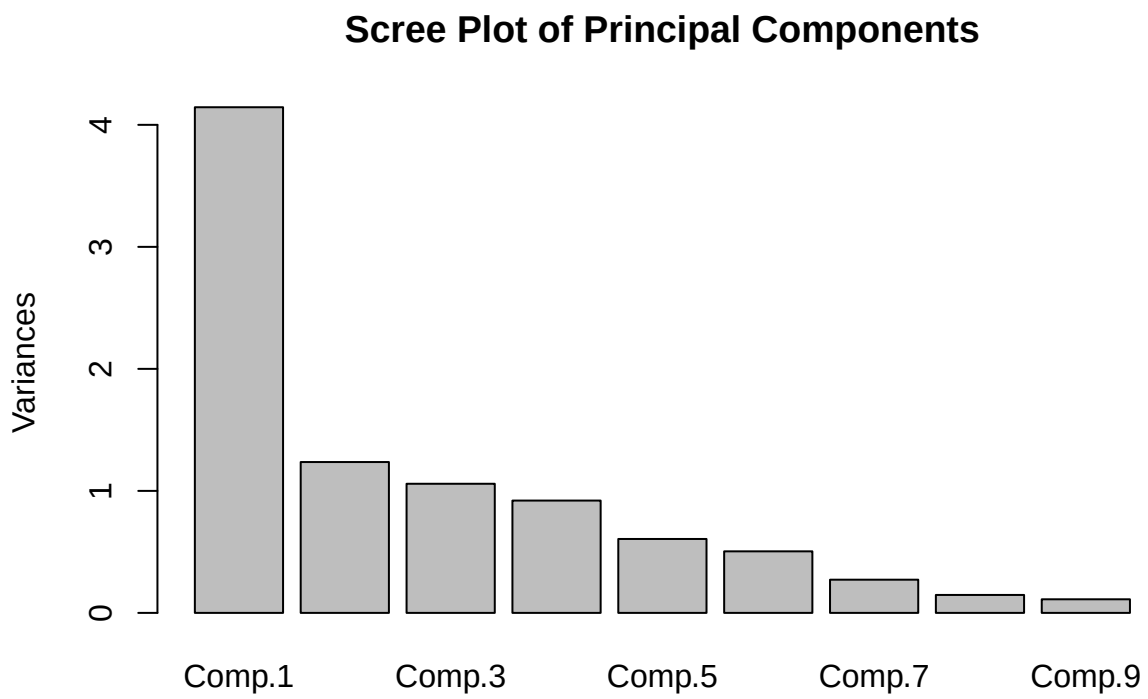

Figure 8: Scree plot of the principal components for the Mali farm data set

From the scree plot it would be reasonable to choose to keep the first 2 PCs. The elbow of this plot appears to be at PC2. One of the downsides to keeping only 2 PCs in this case is the potential loss in information since PC1 and PC2 combined only account for 60% of the variance in the data. Retaining PC 3 or 4 could improve the amount of data retained, but it could also make the result of the PCA much more difficult to interpret. Since I have chosen to standardize the data, for reasons previously stated, we already know that interpretation will be more difficult. Since dimension reduction and interpretablity are my main goals for this PCA I want to retain only the first 2, PC1 and PC2. I know that some information will be lost with this approach, but I am prioritizing the interpretability and dimensional reduction over the loss in information. If the loss in information appears to be too great then we may need to re-evaluate this PCA and consider incorporating additional PCs. This is not difficult to accomplish.

The eigenvectors for PC1 and PC2 are:

```
loadings_matrix <- cbind(c(0.444, 0.000 , 0.411, 0.337, 0.311, 0.269, 0.440, 0.247, 0.309),
                         c( 0.000,  0.000, 0.342, 0.554, -0.452, 0.000, 0.000, -0.458, -0.379))
rownames(loadings_matrix) <- c("Family", "DistRD", "Cotton", "Maize",
                               "Sorg", "Millet", "Bull", "Cattle", "Goats")
colnames(loadings_matrix) <- c("PC1", "PC2")
kable(loadings_matrix, align = 'c')
```

	PC1	PC2
Family	0.444	0.000
DistRD	0.000	0.000
Cotton	0.411	0.342
Maize	0.337	0.554
Sorg	0.311	-0.452
Millet	0.269	0.000
Bull	0.440	0.000
Cattle	0.247	-0.458
Goats	0.309	-0.379

Since I chose to perform the PCA using standardized data the interpretation becomes more difficult than using non-standardized data. In order to improve the interpretability of the data I chose to retain only the first 2 PCs.

Each variable in PC1 appears to contribute a somewhat similar proportion to the total with the Family, Cotton, and Bull being the most dominant factors. Therefore, we could conclude that the first PC is describing the overall size of the farm. If a farm has large values for this PC then we can conclude that this is a larger than average farm. It is interesting to note that the distance of the farm to the nearest passable road is not a component of PC1. It looks as though this PC is telling us that farm size is not influenced by it's distance from the nearest road.

The second PC, PC2, appears to represent a contrast between Sorg, Cattle, and Goats all being negatively loaded and Cotton and Maize with positive loadings. I think that this contrast can be viewed a couple of ways. One way is to view it how the farm chooses to farm it's available land. If we have large herds of cattle and goats this would indicate that we plant fewer acres for row crops such as cotton and maize. This seems to make sense because we need grazing land to support livestock and a given farm will only have a fixed amount of land to utilize. Another reasonable way to view this contrast is between the row crops sorghum, cotton, and maize. What we see is that if we plant sorghum in large quantities then we have lower quantites of cotton and maize. It is very likely that sorghum, cotton, and maize are planted in a rotation due to the demand on the soils. In a year we plant sorghum on our farm then we won't plant as much cotton or maize. It would be interesting to see data year over year to test this idea in the contrast we see. Already we can see how interpretability is becoming difficult.

In order to visualize our PCA we can graph PC1 and PC2 and show the vectors associated with each.

```
ggbiplot(Mf_PC_S, labels = rownames(Mali_Farm_Data)) +
  ggtitle("Principal Component 1 vs. Principal Component 2") +
  theme(plot.title = element_text(hjust = 0.5))
```

Now that we have completed the PCA and chosen the PC's to keep in the analysis, we want to look at some particular points to determine if our conclusions about PC's are consistent with what we see in the data set. First, we will calculate the means for each variable.

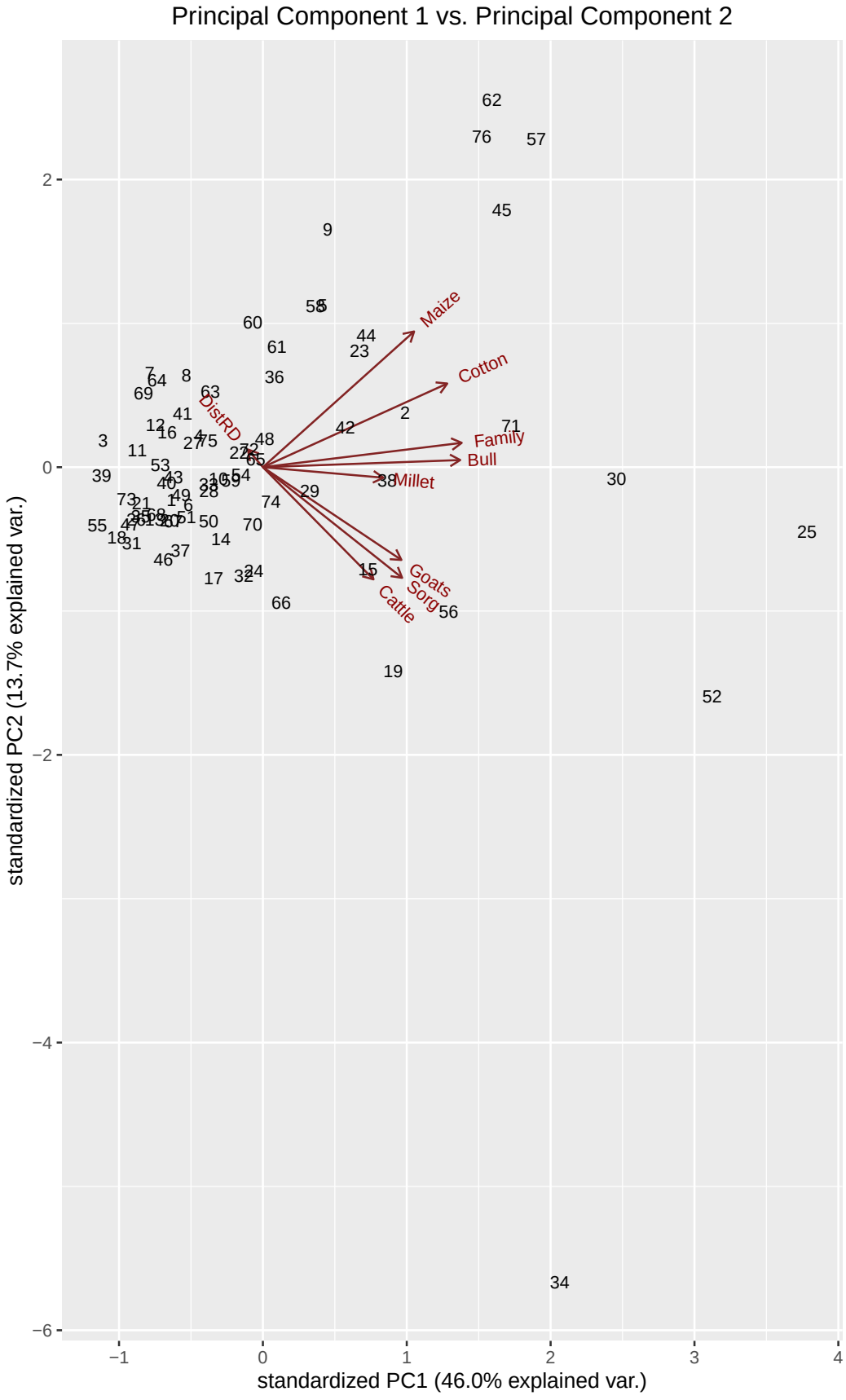

Figure 9: PC biplot of PC1 and PC2 for the Mali farm.

```r
Mf_means <- colMeans(Mali_Farm_Data)
Mf_means <- as.matrix(Mf_means)
colnames(Mf_means) <- c("Mean")
kable(Mf_means, digits = 2, align = 'c')
```

	Mean
Family	32.37
DistRD	32.86
Cotton	2.99
Maize	2.08
Sorg	2.65
Millet	1.67
Bull	2.63
Cattle	4.66
Goats	2.97

There are quite a few interesting farms to investigate. One that stands out in the biplot is number 25. Given our interpretation of PC1 we would expect that farm 25 is a large farm with larger than average values for each variable. Indeed, this is what we see and helps give confidence to our interpretation of PC1.

Another data point of interest is farm 60. Farm 60 shows an observable contrast. But which one? Is is the cattle and row crops contrast or is it with the crop rotation of row crops (planting maize and cotton as opposed to sorghum). This is where interpretability of PC2 becomes difficult. We have to look at farm 60 to see that they have no cattle and larger than average planting for cotton, maize, and sorghum. So in this case it appears the contrast is with row crop versus cattle farming. Farm 60 is a nice farm to observe because it doesn't appear to contain any outliers.

A couple other farms to investigate are are farms 3 and 39. From PC1 we would expect that these are just smaller than average farms. when we look at them in the data set this is what we find.

The Mali farm data set is challenging in many ways. With PCA we want to reduce the dimensionality of the data, without losing information, but also retaining interpretability of the result. In this PCA I felt that retaining more than 2 PCs, while preserving the information, led to a result that was difficult and confusing to interpret. So while information could be retained the interpretability of that information was lost. It was already becoming difficult to interpret with two PCs. As I mentioned with farm 60 the contrast isn't immediately clear. There are numerous examples of this interpretability issue in this data set. We know that the first few PCs are very sensitive to outliers that tend to inflate variance. We know that Cattle likely has outliers and is included as part of PC1 and PC2. It would be good to further evaluate this data set and understand the outliers. It seems reasonable that doing PCA on the Mali farm data set after the outliers have been evaluated and dealt with might produce a result that is more descriptive and easier to interpret.