MACT 4233: Applied Multivariate Analysis - Dr. Ali Hadi

Project 4

The Application of Principal Component Analysis on the Birds Dataset

Katia Gabriel - 900202272

Marina Guindy- 900191946

Farida Simaika - 900201753

Joyce Wassef - 900191248

Introduction

Birds are fascinating creatures that have captured the imagination of humans for centuries. From their beautiful plumage to their unique songs, birds have long been a subject of study and admiration. However, perhaps one of the most intriguing and understudied aspects of birds is their skeletal structure. The Birds' Bone dataset provides an opportunity to explore this aspect of birds in detail. This dataset contains information on 420 birds, representing a wide variety of species, each with 11 distinct measurements of their bone structure.

PCA is a statistical method that is used to reduce the number of variables in a large dataset while retaining most of the information. This technique is useful in analyzing large datasets with many variables, as it helps to identify and eliminate any redundant variables. The importance of PCA lies in its ability to simplify complex datasets and to identify the underlying patterns in the data.

Implementing PCA on the birds' bone dataset can be useful for several reasons. The birds' bone dataset usually contains a large number of variables that describe different features of the bones, such as length, width, and curvature. Using PCA, we can reduce the number of variables, while still preserving most of the information in the dataset. This can help us to identify the most important features that contribute to the variability in the dataset.

Data Description

The Birds' Bones dataset encompasses 420 observations and 12 variables.

Variables	Explanation	Units of Measurements
huml	Length of Humerus	mm
humw	Diameter of Humerus	mm
ulnal	Length of Ulna	mm
ulnaw	Diameter of Ulna	mm
feml	Length of femur	mm
femw	Diameter of femur	mm
tibl	Length of Tibiotarsus	mm
tibw	Diameter of Tibiotarsus	mm
tarl	Length of Tarsometatarsus	mm
tarw	Diameter of Tarsometatarsus	mm
type	Ecological Group	-
id	Sequential ID	-

Data Wrangling and Analysis

After describing each variable and understanding their meaning, the dataset was checked for missing values. Accordingly, it was evident that the dataset had exactly 15 missing values. However, this accounts only for around 4% of the dataset so they were dropped. Moreover, the dataset has two categorical variables namely the "index" variable and the "type" variable. In order to perform the principal component analysis both variables were dropped to ensure that the data consists only of numerical variables. Moreover, the dataset was scaled to eliminate any difference between the variables due to different variances. This is of huge importance because the PCA

depends on maximizing the variance which would have been extremely affected if the data was not scaled.

Moreover, the application of PCA assumes that variables in the dataset are highly correlated because otherwise the PCA would not be of any benefit since it does dimensionality reduction to remove dependent and redundant variables from the dataset. In order to check this assumption, the following correlation matrix was plotted:

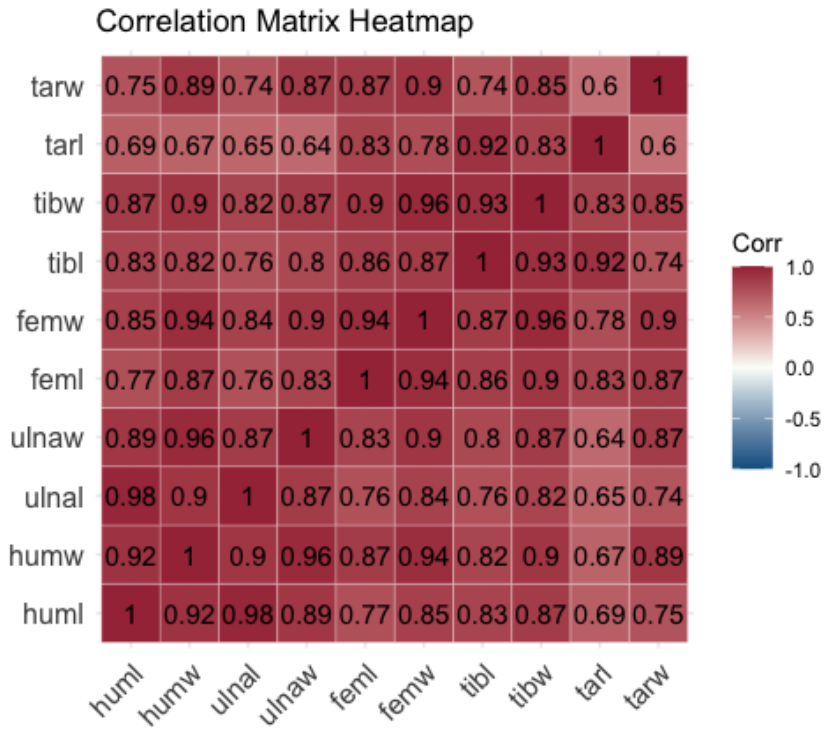


Figure 1 - Correlation Matrix Heatmap

It is evident from the above plot that the variables are highly correlated. An example of two highly correlated variables are “tibw” and “femw” which have a correlation value of 0.96 and have

therefore a strong positive linear relationship. This underscores the need for the implementation of the PCA on this dataset.

Non-Robust PCA

The PCA is a type of variable transformation that is implemented to eliminate the dependence between the various variables in the dataset and therefore solve the multicollinearity problem. As a first step the PCA will be implemented on the original dataset without checking the existence of any outliers. This implementation will be referred to as the non-robust PCA. The target of PCA is to find a set of variables C = $(C_1, C_2, C_3, \ldots, C_m)$ so that the variables in C are linear functions of the X variables and that the variables in C are independent. The PCA was implemented on the birds dataset and yielded the following results:

```
> pc=princomp(x, cor=T); summary(pc,loadings=T)
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4      Comp.5
Standard deviation     2.9228043 0.80949520 0.65251123 0.35234470 0.300755078
Proportion of Variance 0.8542785 0.06552825 0.04257709 0.01241468 0.009045362
Cumulative Proportion  0.8542785 0.91980673 0.96238382 0.97479850 0.983843865
                           Comp.6      Comp.7      Comp.8     Comp.9
Standard deviation     0.265592826 0.190599495 0.168596166 0.13784374
Proportion of Variance 0.007053955 0.003632817 0.002842467 0.00190009
Cumulative Proportion  0.990897820 0.994530637 0.997373104 0.99927319
                           Comp.10
Standard deviation     0.0852529498
Proportion of Variance 0.0007268065
Cumulative Proportion  1.0000000000
```

Figure 2 - Non-Robust Principal Component Analysis Output Summary

```
Loadings:
      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
huml   0.317  0.228  0.476         0.205  0.145  0.193  0.151  0.321  0.628
humw   0.329  0.264                -0.205 -0.170  0.451 -0.687  0.186 -0.189
ulnal  0.308  0.308  0.479  0.405         0.152 -0.213        -0.316 -0.491
ulnaw  0.320  0.303        -0.414 -0.623 -0.163 -0.340  0.315
feml   0.320 -0.193 -0.323  0.561 -0.207 -0.179  0.380  0.451  0.124
femw   0.333        -0.219  0.150  0.236 -0.464 -0.288 -0.233 -0.488  0.416
tibl   0.315 -0.391  0.156 -0.467         0.160  0.433  0.148 -0.515
tibw   0.331 -0.129        -0.289  0.561 -0.298 -0.175  0.151  0.432 -0.377
tarl   0.281 -0.663  0.183  0.143 -0.280  0.221 -0.374 -0.314  0.238
tarw   0.305  0.212 -0.571         0.172  0.700 -0.112
```

Figure 3 - Non-Robust Principal Component Analysis Eigenvectors

We have at our disposal three methods for comparing and selecting the most appropriate number of components for this dataset. The first method involves examining the scree plot and identifying its elbow point. A scree plot is a graphical representation of the eigenvalues (variances) of the principal components. It displays the eigenvalues of each component in descending order on the y-axis and the corresponding number of components on the x-axis. The second method involves selecting m number of components such that their corresponding λ values are greater than either the average of variances of C or $\underline{\lambda}$.

$$\underline{\lambda} = trace(\lambda)/p$$

We are using the correlation matrix so $\underline{\lambda}$ is equal to 1.

The third and final method consists of choosing m lambdas that cumulatively satisfy a desired proportion of variability.

1. Scree Plot Approach

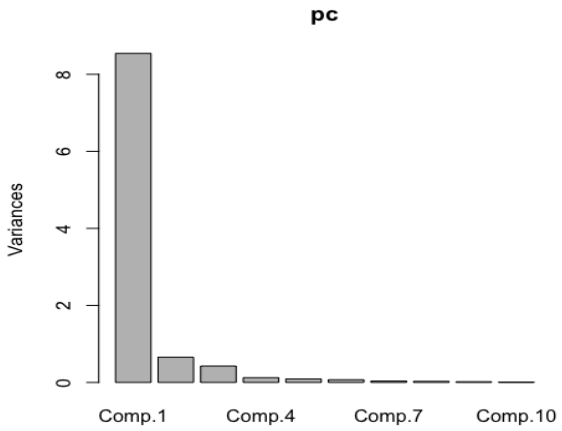


Figure 4 - Non-Robust PCA Scree Plot

The scree plot helps to determine the optimal number of components to retain for further analysis by identifying the elbow point. The elbow point represents the point at which additional components do not contribute significantly to the explained variance of the data. As seen above, the elbow is at Component 2. According to the scree plot, the optimal number of components is two. From the R output, the first two components of this dataset account for nearly 92% of its total variability. These two components are linear functions that combine almost all the variables in the dataset (except for the variable *femw (*diameter of femur) in Component 2), implying that almost all variables significantly contribute (with varying weights) to these principal components. The plot of the first two components can be used to show that unlike the original variables, the two components are independent of each other.

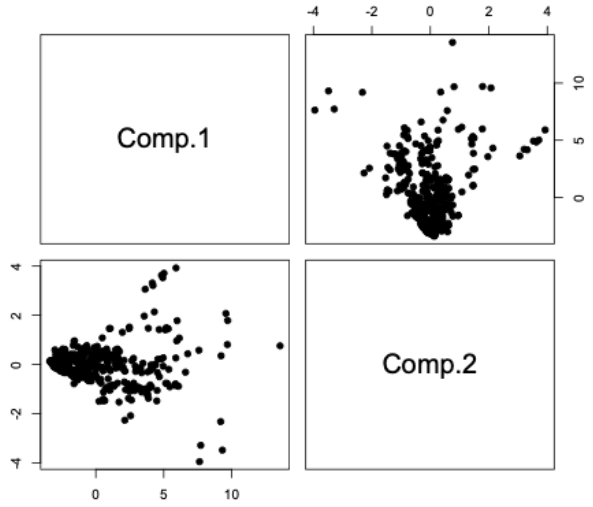


Figure 5 - Robust PCA Components Pairwise Scatter Plots

2. Average Lambda Approach:

The values of the lambdas were analyzed. The lambdas represent the variances of each principal component. We select components with lambda values greater than 1. Based on the above R output, we can conclude that component 1 is the most significant. Component 1 has a lambda value (standard deviation squared) that is greater than 1 (almost at 8.53).

It can be observed that the standard deviation of Component 2 is relatively high as well. Upon squaring the standard deviation of Component 2 its variance was found to be around 0.65, which is not close enough to 1. Therefore, we can conclude that the only significant component is component 1, as it is the only one with a variance greater than 1. This contradicts the conclusion drawn from the scree plot (optimal number of components according to the scree plot is two).

3. Cumulative proportion of variability

The third method for selecting the number of components involves choosing m number of components such that the sum of their corresponding λ's accounts for 95% of the total variance.

From the R output, it is obvious that the optimal number of components is found to be 3. The three components cumulatively account for 96.3% of the total variance. Components 1 and 2 are linear functions that combine almost all the variables in the dataset (except for the variable *femw (*diameter of femur) in Component 2), implying that almost all variables significantly contribute (with varying weights) to these principal components. On the other hand, Component 3 is a linear combination of all variables in this dataset except for the following features: *humw* (diameter of humerus), *ulnaw* (diameter of ulna) and *tibw* (diameter of tibiotarsus). These three variables do not contribute to the third component.

Non-Robust PCA Overview:

Each of the three methods yielded to a different number of optimal components. Both the scree plot and cumulative proportion of variability methods suggested that an appropriate number of components is 2. On the other hand, the average λ approach indicated that an adequate number of components for this dataset would be 1. All methods have significantly contributed to reducing the dimensions of this dataset, identifying redundant variables while being able to retain most of the information.

Robust PCA

Since outliers can heavily influence the output of the principal component analysis technique, therefore we decided to compute the Robust PCA. In this method we are going to perform BACON to detect outliers and then apply the principle component analysis technique in the basic subset.

From the BACON output we can observe that the algorithm started with a basic subset of size m = 40 (which implies that c = 4 which is the default, multiplied by p=10 which is the number of variables). After 15 iterations, the algorithm ended with a basic subset of a size 247 which implies

that 40.19% of the observations were detected by the algorithm as outliers. This can also be seen in the scatter plot in the Appendix.

```
> b=mvBACON(x)
rank(x.ord[1:m,] >= p  ==> chosen m =  40
MV-BACON (subset no. 1): 40 of 413 (9.69 %)
MV-BACON (subset no. 2): 184 of 413 (44.55 %)
MV-BACON (subset no. 3): 200 of 413 (48.43 %)
MV-BACON (subset no. 4): 210 of 413 (50.85 %)
MV-BACON (subset no. 5): 213 of 413 (51.57 %)
MV-BACON (subset no. 6): 216 of 413 (52.3 %)
MV-BACON (subset no. 7): 219 of 413 (53.03 %)
MV-BACON (subset no. 8): 221 of 413 (53.51 %)
MV-BACON (subset no. 9): 224 of 413 (54.24 %)
MV-BACON (subset no. 10): 232 of 413 (56.17 %)
MV-BACON (subset no. 11): 237 of 413 (57.38 %)
MV-BACON (subset no. 12): 242 of 413 (58.6 %)
MV-BACON (subset no. 13): 246 of 413 (59.56 %)
MV-BACON (subset no. 14): 247 of 413 (59.81 %)
MV-BACON (subset no. 15): 247 of 413 (59.81 %)
```

Figure 6 - BACON Output

After removing the outliers subset from the data the PCA was then applied on the basic subset, and we obtained the following output.

```
> pcr=princomp(x,covmat=R); summary(pcr,loadings=T)
Warning message:
In princomp.default(x, covmat = R) :
  both 'x' and 'covmat' were supplied: 'x' will be ignored
Importance of components:
                          Comp.1    Comp.2     Comp.3      Comp.4      Comp.5      Comp.6     Comp.7      Comp.8      Comp.9      Comp.10
Standard deviation     2.9207284 0.92646088 0.59250302 0.295798256 0.231066374 0.207433152 0.17158466 0.149933493 0.125884773 0.0909452465
Proportion of Variance 0.8530654 0.08583298 0.03510598 0.008749661 0.005339167 0.004302851 0.00294413 0.002248005 0.001584698 0.0008271038
Cumulative Proportion  0.8530654 0.93889840 0.97400438 0.982754046 0.988093212 0.992396064 0.99534019 0.997588199 0.999172896 1.0000000000

Loadings:
      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
huml   0.307  0.448  0.153  0.250  0.256         0.158  0.162  0.163  0.689
humw   0.327  0.237 -0.226 -0.340               -0.311        -0.738  0.123
ulnal  0.302  0.471  0.187                0.578        -0.123        -0.537
ulnaw  0.322  0.308 -0.145 -0.286 -0.173 -0.490 -0.280 -0.280  0.508 -0.117
feml   0.310 -0.401 -0.186         0.696  0.188 -0.221 -0.363
femw   0.320 -0.299 -0.237 -0.341 -0.214  0.349         0.582  0.349
tibl   0.329 -0.121  0.348  0.132  0.319 -0.482         0.483 -0.108 -0.396
tibw   0.333 -0.152 -0.138 -0.127 -0.137 -0.162  0.807 -0.328 -0.165
tarl   0.290 -0.340  0.704        -0.371        -0.226 -0.244         0.207
tarw   0.320 -0.150 -0.379  0.763 -0.333        -0.182
```

Figure 7 - Robust Principal Component Analysis Output Summary

We can now compare the three methods for choosing the most adequate number of components for this dataset. The first method is choosing the elbow point in the scree plot that plots the

variances of C versus their index. The second method involves selecting m number of components such that their corresponding λ values are greater than either the average of variances of C or $\underline{\lambda}$. $\underline{\lambda} = trace(\lambda)/p$, which, in case of using the correlation matrix, is equal to 1. The third method is choosing m lambdas that cumulatively satisfy a desired proportion of variability.

1. **Scree plot**

According to the scree plot, the elbow can be observed at C2, this suggests that the adequate number of components is two components. The first two components help achieve almost 93.9% of the total variability of the data. This component is a linear function of all the variables of the dataset, which suggests that these variables are all significant (with different weights), based on this principal component.

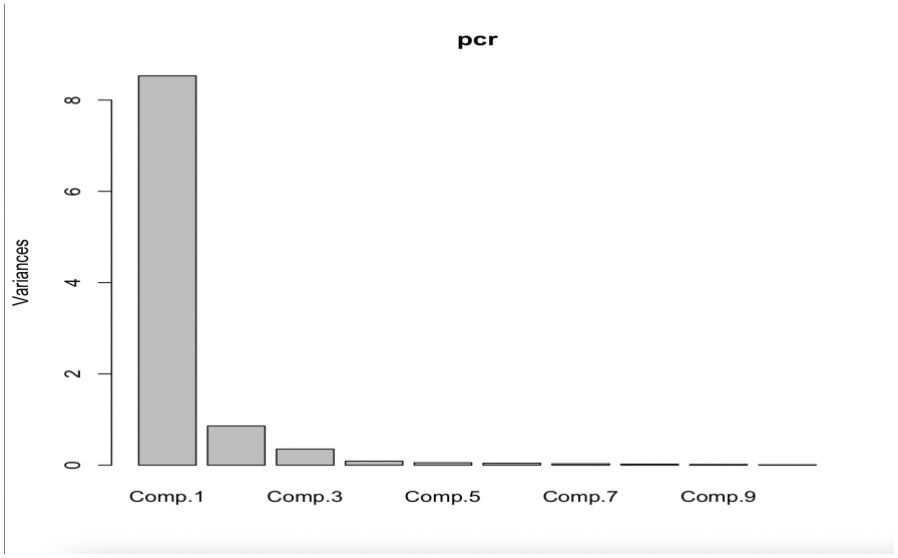


Figure 8 - Robust Principal Component Analysis Scree Plot

2. **Average Lambda Approach**

For this method, we observe the values of the λ's which are the variances, and choose the values that are greater than 1. From the output obtained, we can see that it suggests that component 1 is the most significant given that the value of its λ (which is the square of the standard deviation) is

greater than 1, and almost equal to 8.5. Additionally, we initially thought that we can consider component 2 to be significant as well given that the value of its standard deviation is relatively high, however, when we squared the standard deviation we found that the variance is equal to 0.85 which is not that close to 1, therefore we would conclude that the only significant component is component 1 given that it's the only one that has a variance greater than 1. The conclusion of this method contradicts with the conclusion obtained from the scree plot.

3. Cumulative proportion of variability

The third method for choosing the number of components is choosing m such that m number of λ's achieve a 95% desired proportion of the total variance. Accordingly, the most suitable number of components is m = 3. The first 3 components help achieve 97.4% of the total variance. Each component is a linear function of all the variables, with the coefficients represented by the loadings obtained by R, which are, in fact, the eigenvectors of the correlation matrix, which indicates their significance for all of the chosen components.

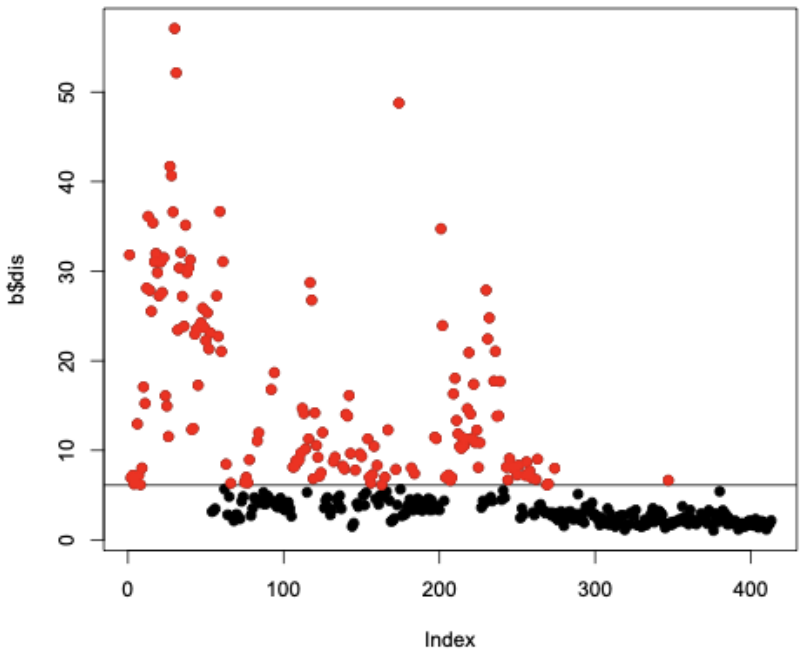
Overview of Robust PCA

Each of the three methods resulted in a different number of components: 2 for the scree plot, 1 for the average variability and 3 for the 95% cumulative variability. Nevertheless, for the cumulative proportion of variability methods, it can be noted that 94% of the variance can be achieved by choosing m = 2, which is close enough to 95%. We can therefore choose m = 2 components.

Conclusion

By implementing PCA on the birds' bones dataset we were able to successfully reduce the number of variables, while still preserving most of the information in the dataset. PCA allowed us to identify the most important features that contribute to the variability in the dataset. It also allowed us to identify and eliminate any redundant variables. We were able to simplify this complex dataset. Each of the three methods used to select an optimal number of components yielded different results. Both the scree plot and cumulative proportion of variability methods suggested that an appropriate number of components is 2. On the other hand, the average λ approach indicated that an adequate number of components for this dataset would be 1. However, all methods have significantly contributed to reducing the dimensions of this dataset.

Appendix



BACON distances scatter plot