PRINCIPAL COMPONENT ANALYSIS OF ANIMAL SLEEPING AND HEART DISEASE DATASET

COURSE: CSIC 5011

NAMES AND STUDENT IDs:

MARADESA ADELEKE	20724523
OGEDENGBE IKEOLUWA IREOLUWA	20724157

1. Introduction

In this era of Big data, the volume of data collected and archived has increased significantly. For quantitative planning and evidence-based decision making, data analysis is a key factor. Given the need to extract information quickly, the dimension of data to be analyzed to this end thus becomes a point of interest to the researcher. When the number of features increases, there should be corresponding increase in the number of observations (sample size). It has been shown that many machine learning algorithms can handle big data efficiently, in contrast, their predictive efficiency decreases with increasing dimensionality [1]. In fact, a model built with respect to many features relies on data with low predictive efficiency and places wide bound on the error of estimation (unreliable credibility region). To arrive at a model with reliable estimates, it is necessary to determine the number of components which explain most of the variability in the data, and the method of principal component is a good candidate in this aspect [1, 2].

Also, in regression analysis, a situation may arise where two or more regressors are correlated (serial correlation), the problem of multicollinearity can be resolved by using principal component analysis (PCA) in which the number of features will be reduced to the number of uncorrelated features. The core idea behind PCA is to reserve more variability and obtaining new variables which are not correlated with one another and have optimal variance. To finds new variables with maximized variance, we resort to finding adequate solution to the associated eigenproblem [3] – [4]. In many healthcare systems, there exist many variables that are of great interest to researcher. In this research, we are interested in investigating the sleeping hour of some selected animal species and carry out rigorous explorative analysis by using principal component analysis. Furthermore, a heart disease dataset is also analyzed with the following objectives: to investigate the number of principal components that explain larger variability within the dataset, to determine the contribution of each variables to the principal component, that is, the relative importance of each variables and investigation of significant influential variables using correlation plot. Also, Horn parallel analysis is performed to determine the number of principal components to be retained, if the Horn's condition is satisfied.

to make reasonable conclusion, based on quantitative analysis and exploratory studies, for evidence-based decision making.

2. Theory

Principal component analysis is a form of exploratory data analysis which involves data reduction. We let $Y \in R^{n\,x\,p}$,with p-variates and n-dimensional vectors Y = $\left(y_1, y_2, \dots, y_p\right)$, has jth variable and column. We construct a linear combination of X matrix with optimum variance which is given by:

$$\sum_{j=1}^{p} \alpha_j y_j = Y\alpha \qquad (1)$$

Where α is a vector of constant values, given by $\alpha = \left(\alpha_1, \alpha_2, \dots, \alpha_p\right)$. We also obtain the associated covariance as $Var(Y\alpha) = \alpha'\Psi\alpha$, where Ψ is a sample covariance. We can then obtain the vector α (p-dimensional) which maximizes $\alpha'\Psi\alpha$. To achieve this, we restrict that $\alpha'\alpha = I$, where I is an identity matrix. We can define a problem that maximizes:

$$\alpha'\Psi\alpha - \lambda(\alpha'\alpha - 1) \qquad (2)$$

Where λ is a Lagrange multiplier; we solve (2) and express it solution as $\Psi\alpha - \lambda\alpha = 0$, where λ and α are the eigen value and vector respectively and $Var(Y\alpha) = \alpha'\Psi\alpha = \lambda\alpha'\alpha = \lambda$. We define the symmetric matrix $(p\ x\ p)$ in such a way that the matrix Ψ has λ_k as its eigenvalues, where $k = 1, \dots, p$, and their eigenvector is constructed in a way to form orthonormal vectors $\alpha_k'\alpha_k = 1$. The method of Lagrange multiplier is used to obtain the solution and show that the set of eigenvectors of Ψ provide solution to problem constructing up to p linear combinations $\sum_{j=1}^{p} \alpha_{jk} y_j = Y\alpha_k$ and this minimizes the variance with respect to non-autocorrelation conditions. That is, there is not autocorrelation with previous linear combinations. When this happens, $Y\alpha_k$ is the principal component of the data under consideration [2, 4].

3. Principal Component Analysis

3.1 PCA for Animal sleeping data

Here, we analyze Animal sleeping data using the method of PCA

Table 1: The proportion of explained variance

Principal Component	Eigenvalue	Variance percent	Cumulative variance percent
PC1	4.80166302	48.0166302	48.01663
PC2	2.13718237	21.3718237	69.38845
PC3	1.25553611	12.5553611	81.94381
PC4	0.74797894	7.4797894	89.42360
PC5	0.36122655	3.6122655	93.03587
PC6	0.23888329	2.3888329	95.42470
PC7	0.20599556	2.0599556	97.48466
PC8	0.17375571	1.7375571	99.22222
PC9	0.05014598	0.5014598	99.72368
PC10	0.02763247	0.2763247	100.00000

From table 1, for the PC1, the eigenvalue is 4.80166302 and the proportion of variance explained is 0.480166302 and this is an indication that PC1 account for 48.16% of the total variation within the dataset. From PC1 to PC3, the cumulative variance percentage is about 81.94381%, we deduce that first three principal component account for 81.94381% of the total variation within the dataset

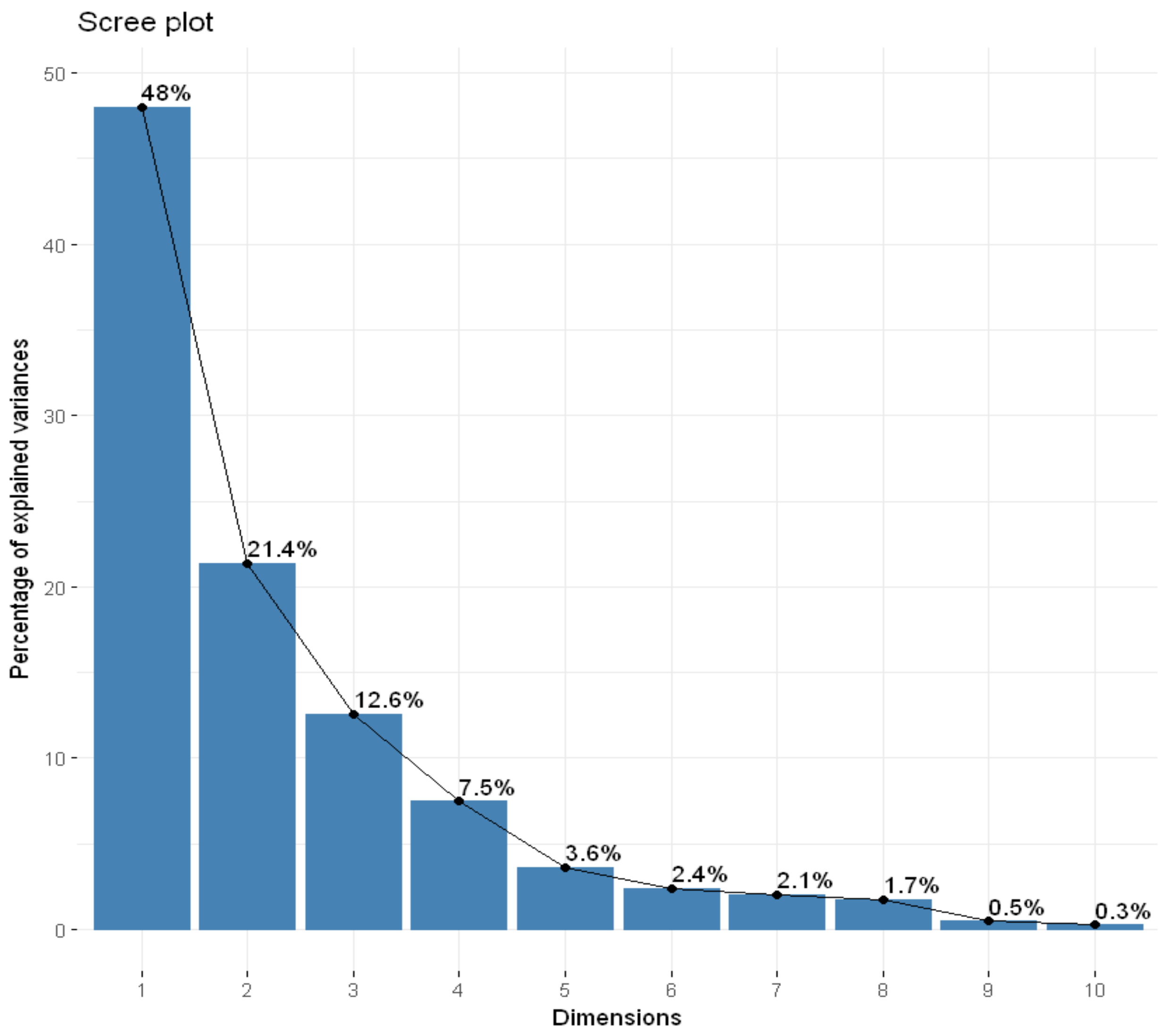

Fig 1: The Scree plot shows the percentage of explained variance

From figure 1, the scree plot shows the percentage of variance for each principal component. We deduced that the variance explained with respect to different principal component are about 48%, 21.4% and 12.6% of explained variances are observed with respect to PC1, PC2 and PC3 respectively. Since about 81.94381% of the total variation areas explained using only PC1, PC2 and PC3, we can say that PC1, PC2 and PC3 provide ample relevance, hence they are to be retained

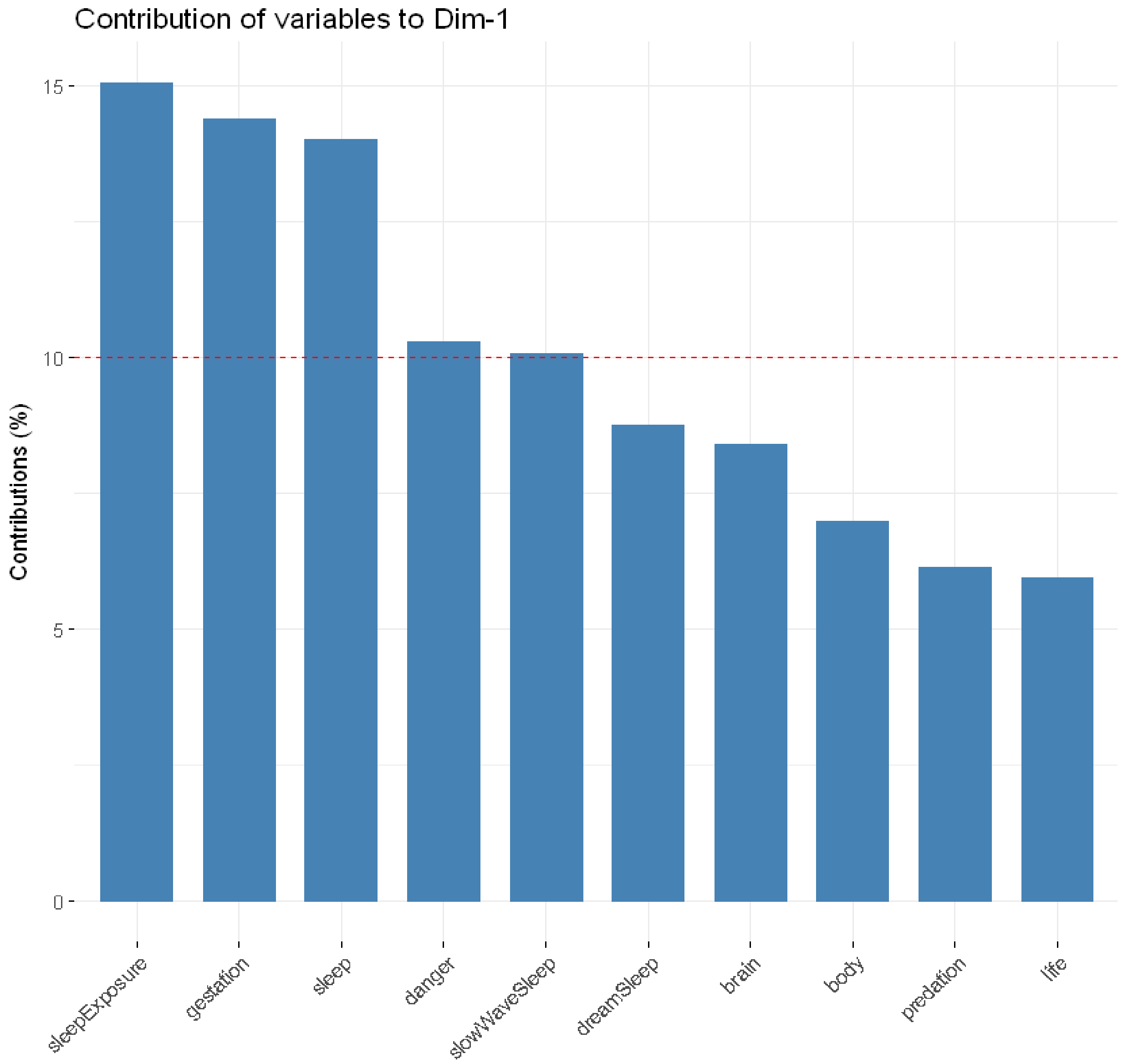

Fig 2: the contribution of variable to PC1

It has been revealed that PC1 explained about 48% of the total variation within our dataset. There is a need to investigate the variables that contribute to this performance of PC1. We used visualization method to determine the percentage contribution of each variable to PC1 and we discovered that sleepExposure, gestation and sleep contribute the most to the principal component.

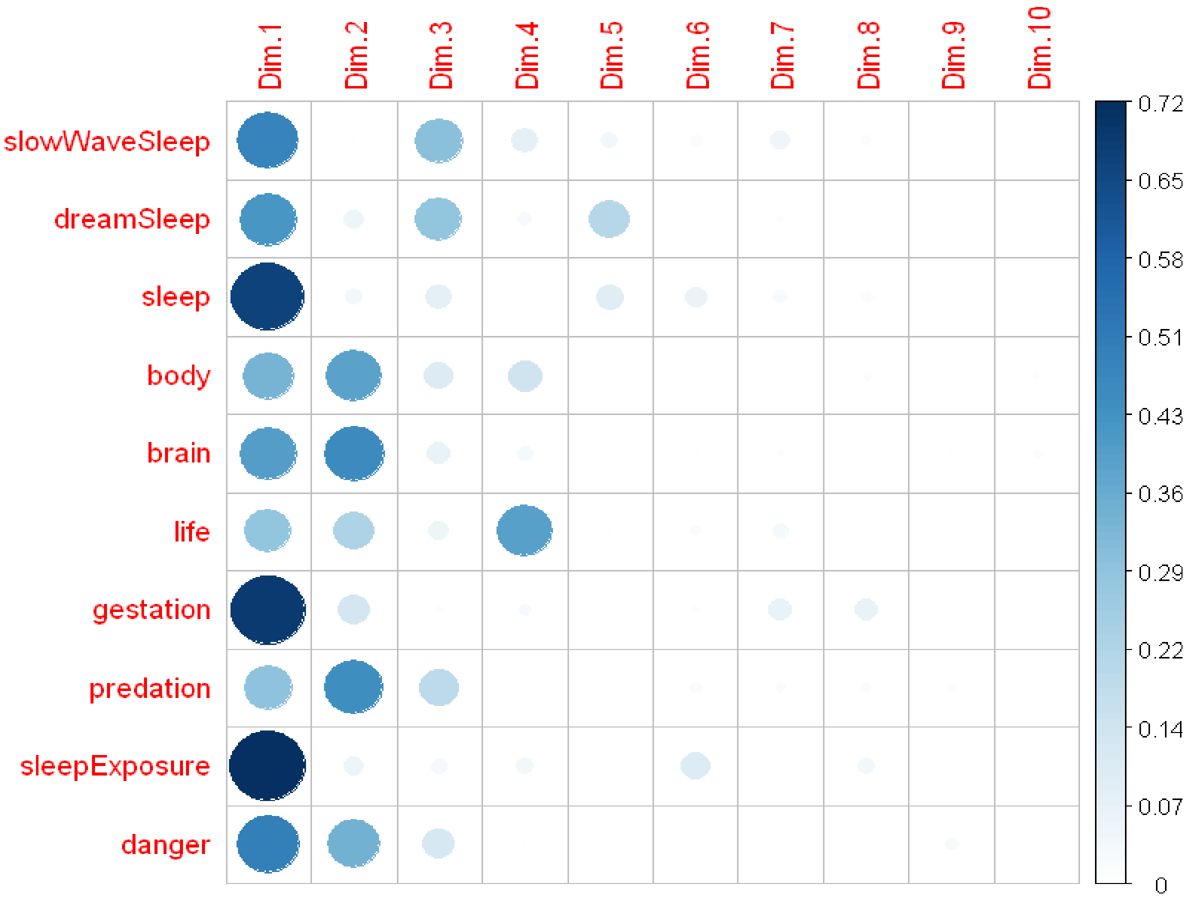

Fig 3: The correlation plot

From fig 3, the PC1 shows good representation of all the variables of interest and their respective contributions. The sleepExposurte, sleep and gestation showed significant contribution to PC1 as compared to other variables. In PC2, brain, predation body, and danger contribute to PC2, and the contribution of SlowWaveSleep and DreamSleeep to PC3 is insignificant. We can say that PC1 contributes the most to total variation.

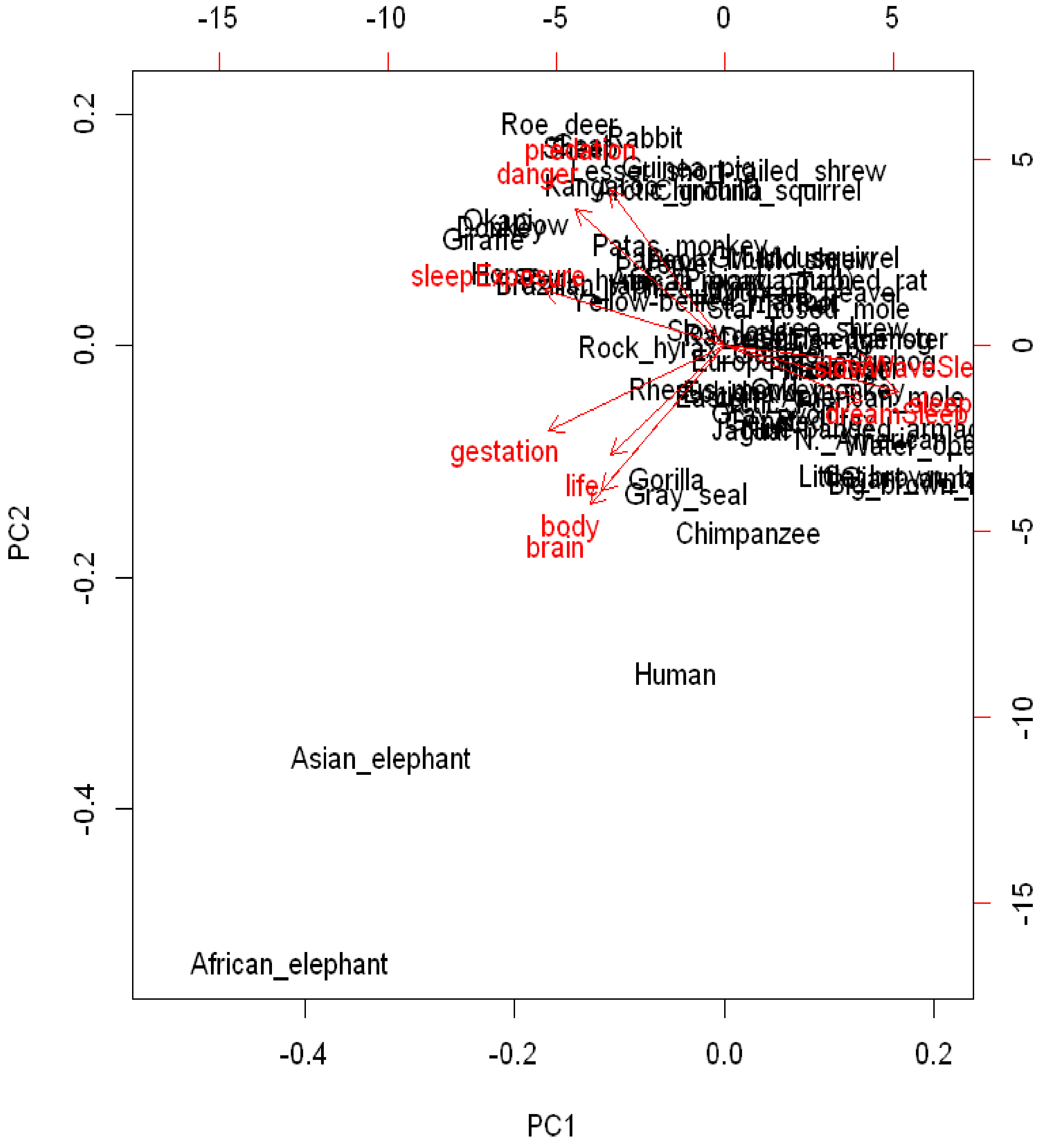

Fig 4: Principal Component Scatter Plot

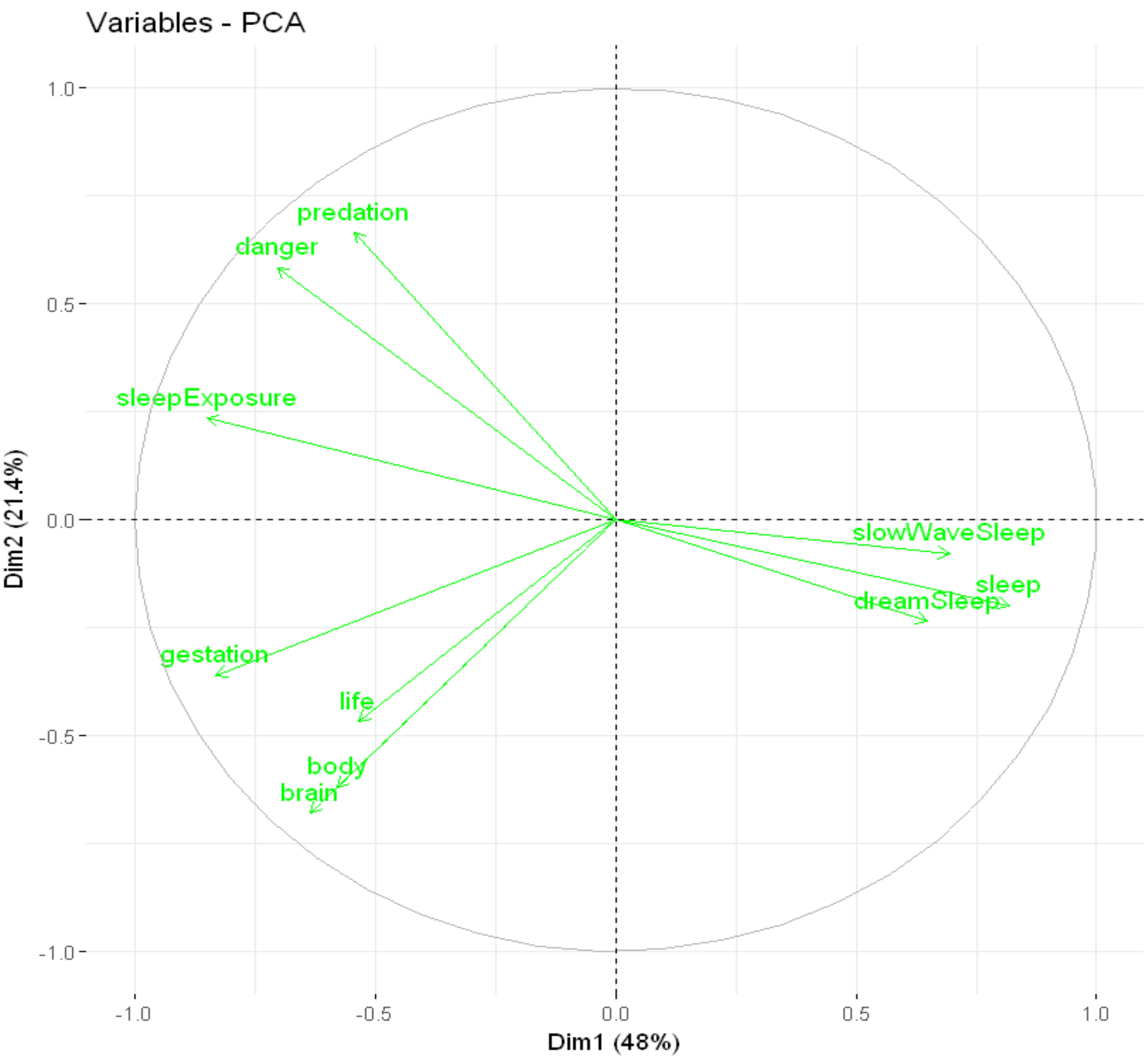

Fig 5: Principal Component Polar Plot

From fig 4, we can say that PC1 explains 48% of total variation in the data set. The African and Asian elephant are outliers in the dataset.

Table 2: Horn Parallel Analysis

Component	Adjusted eigenvalue	Umadjusted Eigenvalue	Estimated Biase
PC1	4.112125	4.801663	0.689537
PC2	1.678054	2.137182	0.459127

In Table 2, the Horn parallel analysis automatically retained only two principal component *i.e* PC1 and PC2. The first and second principal component are retained because their adjusted eigenvalues are greater than one ($\lambda_k > 1$) and this has validated the number of PC to be retained.

3.2 PCA for Heart Attack data

3.2.1 Data Description

The nature of the data used and the corresponding parameters are given below:

Data source: *https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset*

Data parameters: (excluding the target)

Age: Age of the patient

Sex: Sex of the patient

exang: exercise induced angina (1 = yes; 0 = no)

ca: number of major vessels (0-3)

cp: Chest Pain type chest pain type

Value 1: typical angina

Value 2: atypical angina

Value 3: non-anginal pain

Value 4: asymptomatic

trtbps: resting blood pressure (in mm Hg)

chol: cholestoral in mg/dl fetched via BMI sensor

fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

rest_ecg: resting electrocardiographic results

Value 0: normal

Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

thalach : maximum heart rate achieved

3.2.2 Results

Table 3: The Proportion of Variance Explained

Principal Component	Eigenvalue	Variance percent	Cumulative variance percent
PC1	2.7630269	21.254053	21.25405
PC2	1.5366920	11.820708	33.07476
PC3	1.2228343	9.406418	42.48118
PC4	1.1811455	9.085735	51.56691
PC5	1.0219665	7.861281	59.42819
PC6	0.9700159	7.461661	66.88985
PC7	0.8627699	6.636692	73.52655
PC8	0.7759454	5.968811	79.49536
PC9	0.7189255	5.530196	85.02555
PC10	0.6215702	4.781309	89.80686
PC11	0.5301048	4.077729	93.88459
PC12	0.4231424	3.254941	97.13953
PC13	0.3718607	2.860467	100.0000

From table 1, for the PC1, the eigenvalue is 2.7630269 and the proportion of variance explained is 0.21254053 and this is an indication that PC1 account for 21.25% of the total variation within the dataset. From PC1 to PC4, the cumulative variance percentage is about 51.56691%, we deduce that first four principal components account for about 52% of the total variation within the dataset

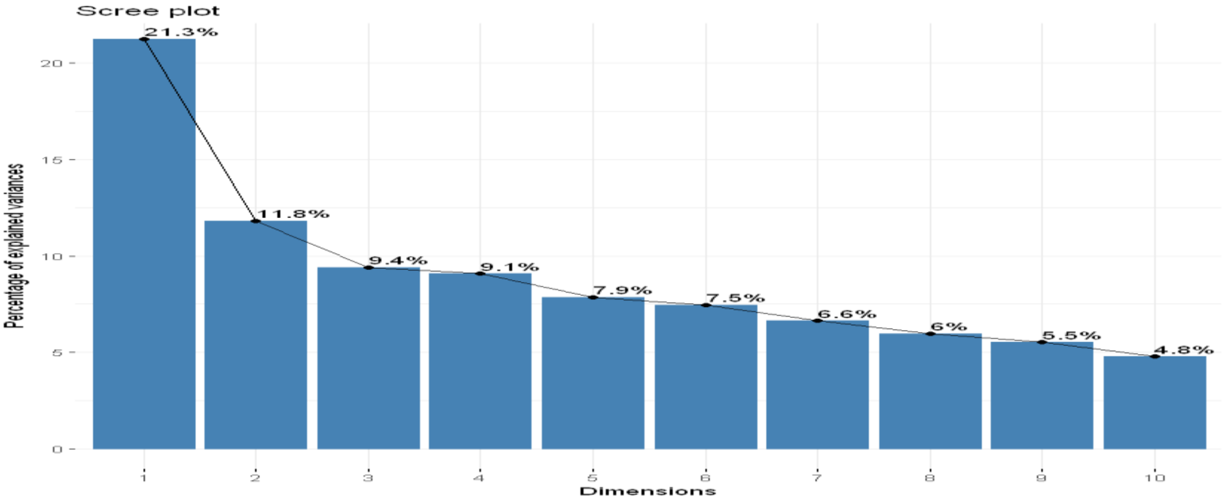

Fig 6: The Scree plot for Heart disease datasets

From fig 5, the scree plot shows the percentage of variance for each principal component. We deduced that the variance explained with respect to different principal component are about 21.3%, 11.8% 9.4% and 9.1% of explained variances are observed with respect to PC1,PC2, PC3 and PC4 respectively. Since about 52% of the total variation are explained using only PC1-PC4, they are thereby retained.

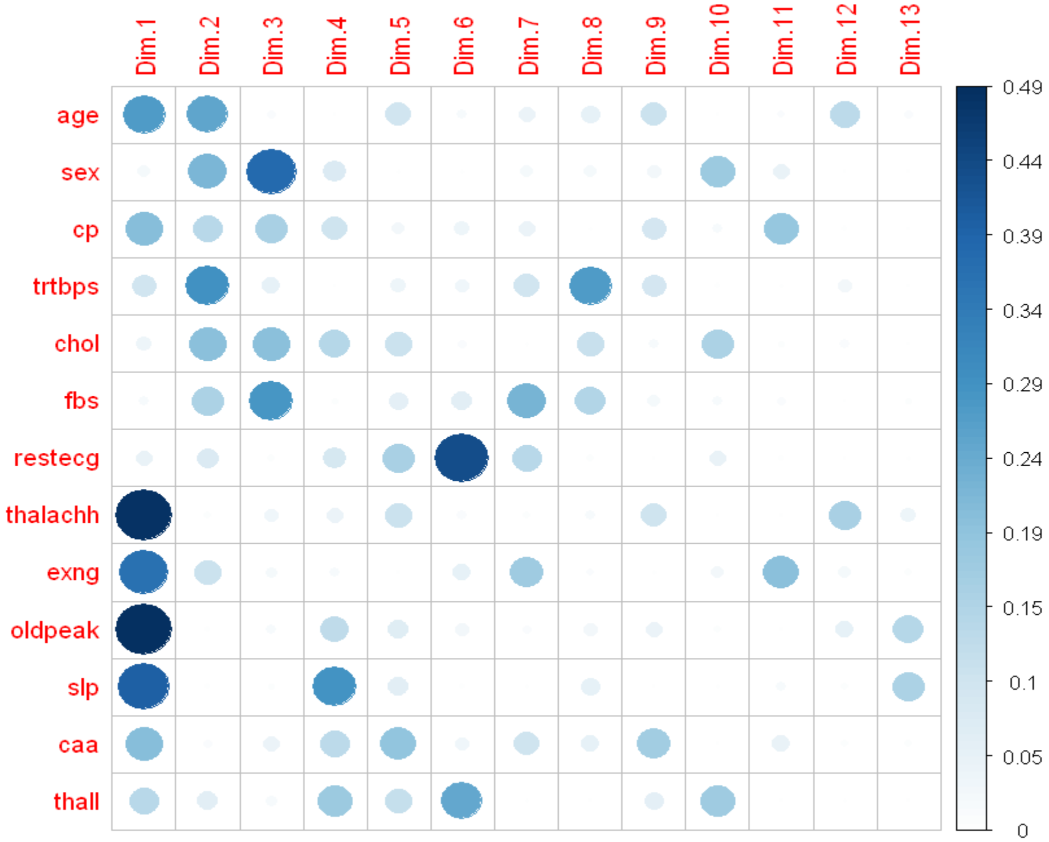

Fig 7: The correlation plot for heart disease dataset

The fig 6 reveals the relative importance of each of the principal components with respect to the contributions of each variable under consideration. In PC1, thalach (maximum heart rate achieved), oldpeak, exng (exercise induced angina) and slip show significant contribution as compared to other variables. Age and trtbps (resting blood pressure) are the major players in PC2. Sex and fbs (fasting blood sugar > 120 mg/dl) show good contribution to PC3 and slp also contributed to PC4.

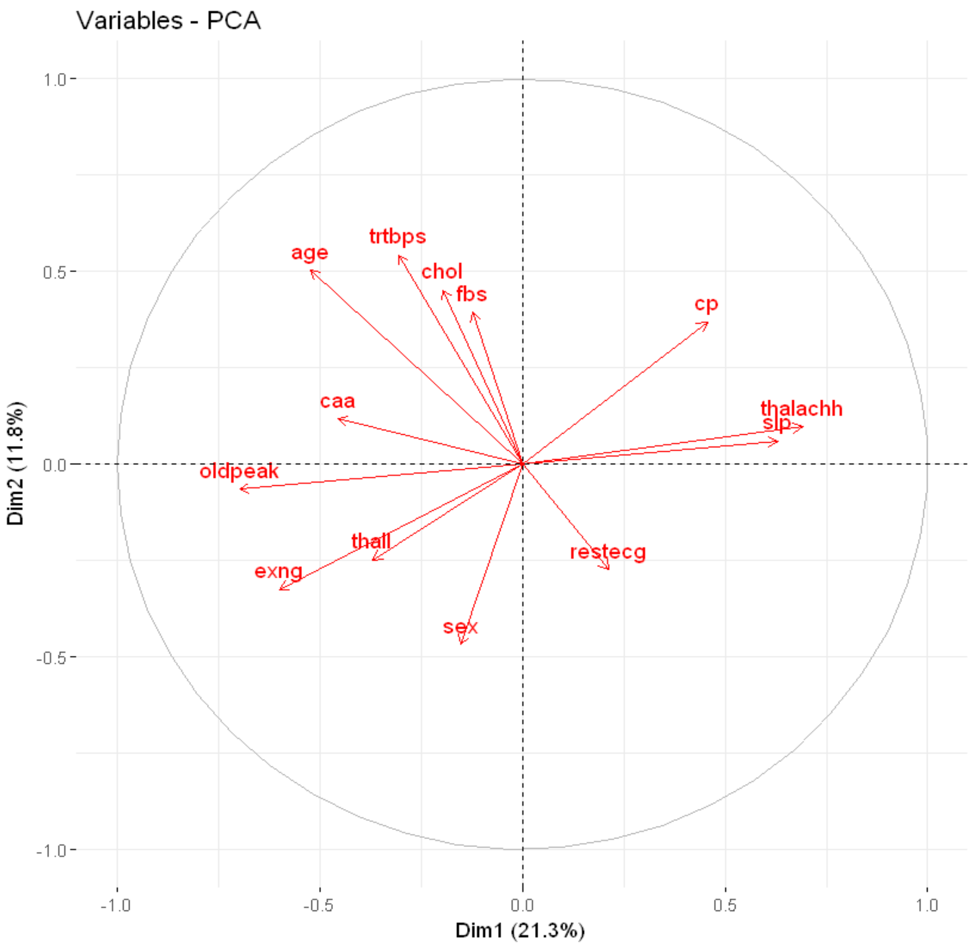

Fig 8: The Principal Component Polar plot (Heart disease data)

The fig 7 shows the percentage variation explained by principal components (PC1 and PC2). PC1 and PC2 explained about 21.3% and 11.8% of the total variation within the dataset

Table 4: Horn Parallel Analysis for Heart disease data

Component	Adjusted eigenvalue	Umadjusted Eigenvalue	Estimated Biase
PC1	2.406281	2.763026	0.356745
PC2	1.269155	1.536692	0.267536
PC3	1.023294	1.222834	0.199540
PC4	1.040195	1.181145	0.140950

From the result of Horn parallel analysis (table 4), we retained only four principal components (PC1-PC4) because their adjusted eigenvalues are all greater than one (λ_k> 1).

4. Conclusion

In this work, we have applied the method of principal component analysis to analyzed two different datasets to determining the number of principal components required to explain most variability in the datasets. We conclude that the first two principal components (PC1-PC2) and the first four principal components (PC1-PC4) explained more variability in the Animal sleeping and Heart disease datasets, respectively.

Reference

[1] C. L. Sabharwal and B. Anjum, “Principal Component Analysis as an Integral Part of Data Mining in Health Informatics,” p. 7.

[2] E. F. Jackson, A. Siddiqui, H. Gutierrez, A. M. Kanté, J. Austin, and J. F. Phillips, “Estimation of indices of health service readiness with a principal component analysis of the Tanzania Service Provision Assessment Survey,” *BMC Health Serv. Res.*, vol. 15, no. 1, p. 536, Jun. 2015, doi: 10.1186/s12913-015-1203-7.

[3] F. Anowar, S. Sadaoui, and B. Selim, “Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE),” *Comput. Sci. Rev.*, vol. 40, p. 100378, May 2021, doi: 10.1016/j.cosrev.2021.100378.

[4] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.