KJ_Pipho_PS_10

Krista Pipho

November 28, 2018

Question 1. Dimensionality and Principal Componant of nci60 Data Set

```
# a) Load the data set
load(file="NCI60")
# b) Confirm the set
# Here we make two objects, one to hold the label of NCI60 and one to hold the data
nci.lab <- NCI60$labs
# See the number of unique cancer lines
unique(nci.lab)
```

```
##  [1] "CNS"         "RENAL"       "BREAST"      "NSCLC"       "UNKNOWN"
##  [6] "OVARIAN"     "MELANOMA"    "PROSTATE"    "LEUKEMIA"    "K562B-repro"
## [11] "K562A-repro" "COLON"       "MCF7A-repro" "MCF7D-repro"
```

```
nci.data <- NCI60$data
# Now we read the number of dimensions and data type
dim(nci.data)
```

```
## [1]   64 6830
```

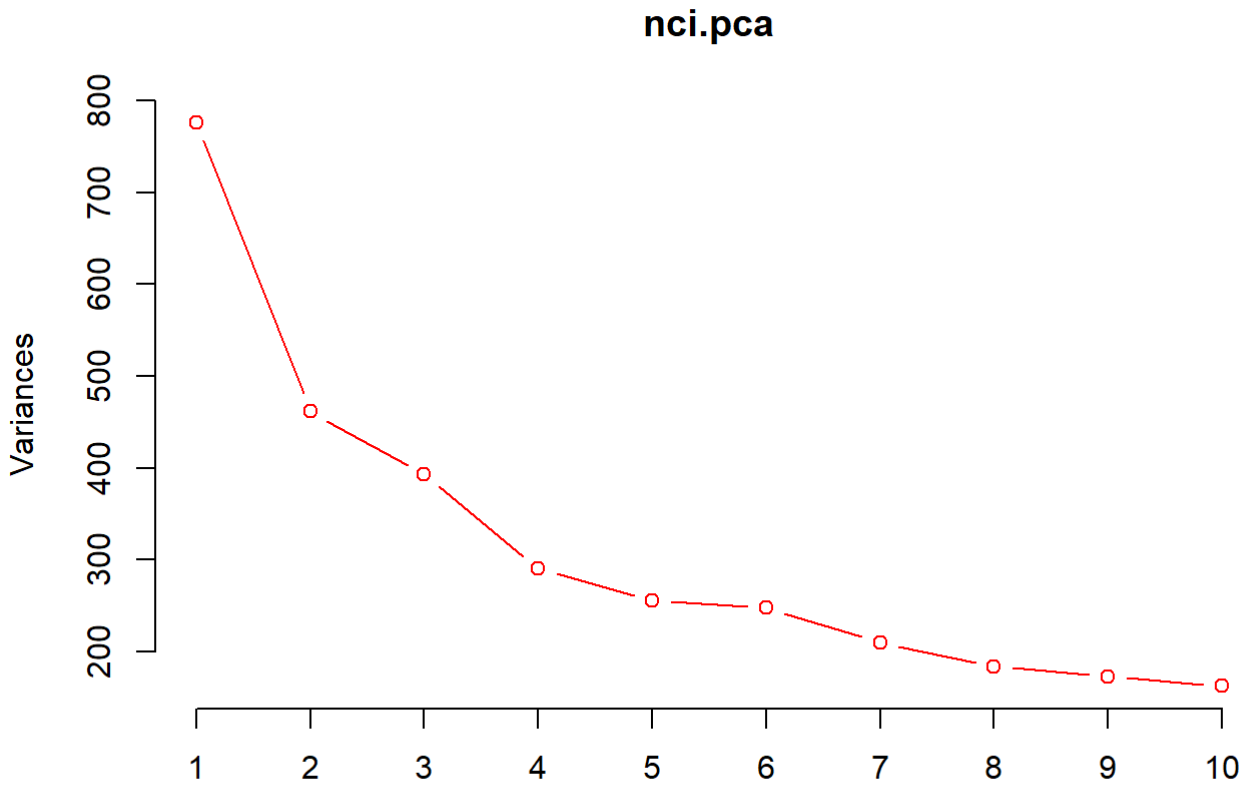
```
class(nci.data)
```

```
## [1] "matrix"
```

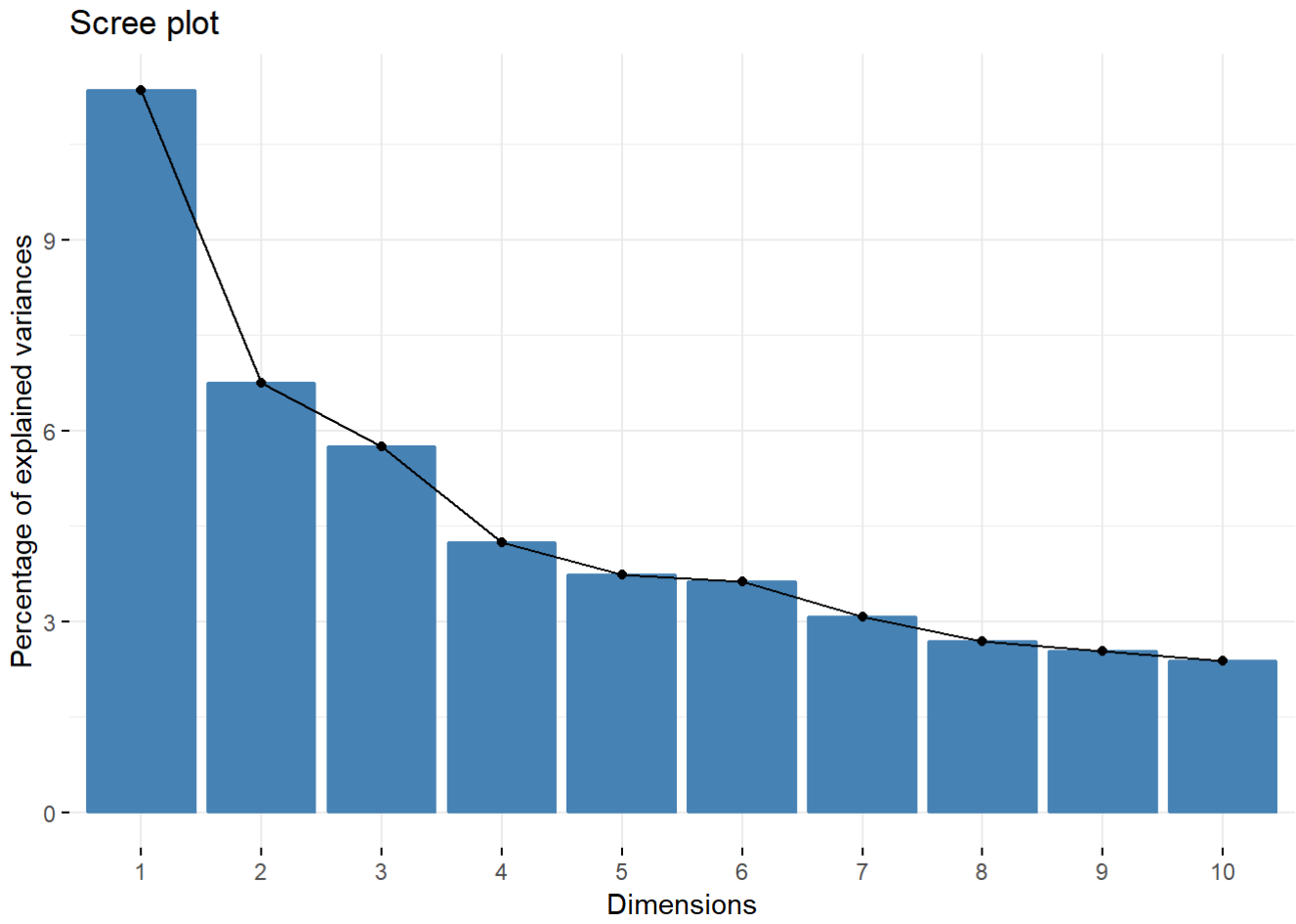
```
# c) Onwards to PCA
# Next we look at a summary of the data, and look under cumulative probabuility for capture o
f variance
#summary(nci.data) too long to show
# Next we run principal componants analysis on the data and save the result as nci.pca
nci.pca <- prcomp(nci.data, scale=TRUE)
#nci.pca too long to show
# Now we create a skew plot pf the data
plot(nci.pca, type="lines", col="red")
# d) Final visualization
# And finally, using factoextra we will make a pretty visualization of our PCA analysis
# Here we install factoextra
#install.packages("factoextra") - this is already installed
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFC
Z
```




```
# Pretty graph wow
fviz_eig(nci.pca)
```



You need 32 of the columns to find 80% of the variance, and 63 of the columns to understand 100% of it.

Question 2. Are the species differences seen in cricket's stridulating pulse rates driven by body temperature?

To answer this question

```
# First we load in the file and examine it
Cricketunes <- read.csv("C:\\Users\\xenon\\Desktop\\R Studio 2018\\crickets.csv")
names(Cricketunes)
```

```
## [1] "Species" "Temp"    "Pulse"
```

```
head(Cricketunes)
```

```
##   Species Temp Pulse
## 1      ex 20.8  67.9
## 2      ex 20.8  65.1
## 3      ex 24.0  77.3
## 4      ex 24.0  78.7
## 5      ex 24.0  79.4
## 6      ex 24.0  80.4
```

```
tail(Cricketunes)
```

```
##    Species Temp Pulse
## 26     niv 24.2  70.9
## 27     niv 25.9  76.2
## 28     niv 26.5  76.1
## 29     niv 26.5  77.0
## 30     niv 26.5  77.7
## 31     niv 28.6  84.7
```

```
# Next we generate two linear models,one which tests Temp and Species individually for signif
icance against Pulse, and one which tests their cumulative interaction.
CricketunesNoInteraction <- lm(Cricketunes$Pulse ~ Cricketunes$Temp + Cricketunes$Species)
CricketunesInteraction <- lm(Cricketunes$Pulse ~ Cricketunes$Temp* Cricketunes$Species)

#Now we generate a linear model prediction for the interaction model
CricketunesPrediction <- predict(CricketunesInteraction)

# First we check ANOVA on the model that assumes interacton
anova(CricketunesInteraction)
```

```
## Analysis of Variance Table
##
## Response: Cricketunes$Pulse
##                                    Df Sum Sq Mean Sq  F value    Pr(>F)
## Cricketunes$Temp                    1 7894.8  7894.8 2505.583 < 2.2e-16
## Cricketunes$Species                 1  598.0   598.0  189.789 9.907e-14
## Cricketunes$Temp:Cricketunes$Species  1    4.3     4.3    1.357    0.2542
## Residuals                          27   85.1     3.2
##
## Cricketunes$Temp                    ***
## Cricketunes$Species                 ***
## Cricketunes$Temp:Cricketunes$Species
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the simple anova table, we can see that Temperature and species as an overall interaction do not generate a statistically distinct group in relation to pulse.

```
# Now we test the model that does not assume interaction
anova(CricketunesNoInteraction)
```

```
## Analysis of Variance Table
##
## Response: Cricketunes$Pulse
##                     Df Sum Sq Mean Sq F value    Pr(>F)
## Cricketunes$Temp     1 7894.8  7894.8  2474.0 < 2.2e-16 ***
## Cricketunes$Species  1  598.0   598.0   187.4 6.272e-14 ***
## Residuals           28   89.3     3.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using a no interaction model, it seems like body temperature and species both have highly significant effects on pulse. However, given the results of the interaction model, it seems likely that these are not distinct effects. When accounting for temperature the rates are probably not distinct. The climates inhabited during mating

season seem to give rise to an artefactual difference in pulse rate between species.