R practical – Principal Components Analysis

1. Arrests

   We will use the USArrests data set available within R. The data set contains statistics, in arrests per 100,000 residents, for assault, murder, and rape in each of the 50 US states.

   Use `?USArrests` to read more details on the data set and `View(USArrests)` to view the data set.

   Plot all the pairwise correlations between the variables using

   `corrplot(cor(USArrests), method="ellipse").`

   Note that first you might need to install and load the *corrplot* package using

   ```
   install.packages("corrplot")
   ```

   ```
   library(corrplot)
   ```

   What do you see? Answer

   Now perform principal component analysis on this data set using the covariance matrix

   ```
   arr_pca_cov <- prcomp(USArrests)
   ```

   Look at the results of this analysis, and especially the loadings of the first principal component, by running `arr_pca_cov`. What do you see? Answer

   Now perform your analysis using the correlation matrix

   `arr_pca_cor <- prcomp(apply(USArrests, 2, scale))` and look at the results.

   ```
   # view results
   summary(arr_pca_cor)
   # these are the sqrt of the eigenvalues
   arr_pca_cor$sdev
   # these are the eigenvectors
   arr_pca_cor$rotation
   # these are the scores
   arr_pca_cor$x
   # screeplot
   screeplot(arr_pca_cor)
   ```

   What do you conclude in terms of the number of PC you need to keep? Answer

Suppose that we decide to keep the first two principal components. Produce a biplot using

```
biplot(arr_pca_cor, cex=0.5, col=c("black","red"),
  xlabs = row.names(USArrests))
```

What does this show? Answer

2. Pollution

We will use the USairpollution data set available in package HSAUR2.

Use `data("USairpollution", package = "HSAUR2")` to load the data, `?USairpollution` to read more details on the data set and `View(USairpollution)` to view the data set.

Plot all the pairwise correlations between the variables and comment on the result. Is there a variable that you could transform to potentially make interpretation easier?

Answer

Now perform principal component analysis on this data set using the correlation matrix, since the variables are measured on completely different scales. What do you conclude in terms of the number of PC you need to keep? Answer

Produce biplots for all pairs of components that you have decided to keep, for example for plotting component 1 against component 3 you can use

```
biplot(pol_pca_cor, choices=c(1,3), cex=0.7,
 col=c("black","red"), scale = 0, xlabs = row.names(USairpollution))
```

Summarise your findings. Answer