Tutorial 4: Principal Component Analysis

Exercise 1: Sparrow Data

In 1898, H.C. Bumpus collected data from 49 female sparrows, which he picked up after a severe storm. Birds 1 to 21 survived; birds 22 to 49 died. The variables measured were X_1 total length, X_2 alar extent X_3 length of beak and head, X_4 length of humerus, X_5 length of keel of sternum, X_6 returns a 1 if the bird survived and a 0 otherwise. This data set is found in `sparrow.dat` on the course home page.

```
www<-"https://www.mimuw.edu.pl/~noble/courses/MultivariateStatistics/data/spar
row.dat"
sparrow <- read.table(www,header=T,quote="\"")
View(sparrow)
```

A principal component analysis can be carried out quite simply using the command `prcomp`.

```
pca <- prcomp(sparrow[,-6],scale=TRUE)
print(pca)
```

Interpret the output.
The correlation matrix may be obtained in the following way:

```
> cormat <- cor(sparrow[,-6])
> cormat
```

The eigenvalues and eigenvectors from the correlation matrix can be obtained in the following way:

```
> ev <- eigen(cor(sparrow[,-6]))
> ev
```

Note that $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 = 5$, the sum of the trace.

The first component accounts for $\frac{3.5475}{5} \times 100 = 70.95\%$ of the total variance. The other principal components account for 10.95%, 7.70%, 6.63% and 3.77% respectively of the total variance.

Another way of looking at it is as follows: after standardisation, all the original variables have variance 1. Therefore, the first principal component has a variance 3.616 times as much as one of the original variables, while the second only accounts for half as much as any of the original variables. The first principal component is clearly by far the most important.

The first principal component, in terms of the standardised variables, is

$$Y_1 = 0.4529Z_1 + 0.4481Z_2 + 0.4559Z_3 + 0.4749Z_4 + 0.4008Z_5.$$

The coefficients are all nearly equal, so Y_1 is an index of the size of the sparrows. Therefore, about 72.3% of the variation in the data is due to differences in the size of the sparrows.

The second principal component is

$$Y_2 = -0.0819Z_1 + 0.4020Z_2 + 0.2694Z_3 + 0.1627Z_4 - 0.8560Z_5.$$

This contrasts Z_2, Z_3 and Z_4 on the one hand, with the length of the keel of the sternum Z_5 on the other. Here Y_2 represents a shape difference between the sparrows.
Now let us make a scatter plot of the first two principal components. A nice package for plots and visualisation is **ggplot2**. Firstly, we add the first two principal components to the 'sparrow' data frame as follows:

```
spa <- sparrow
spa$pc1 <- pca$x[,1]
spa$pc2 <- pca$x[,2]
```

Using the plotting command from **ggplot2**:

```
qplot(pc2,pc1,colour=SURVIVE,data=spa)
```

gives a scatterplot, which illustrates that the population of birds that did not survive has more 'extreme' values than the population that did survive for the first two PCs.

Make a scree plot indicating the amount of the total variation explained by each PC.

Bootstrap for Confidence Intervals

We can use bootstrap methods to estimate confidence intervals. Consider the Fisher Iris data (found in the data `iris`, which comes with R). Suppose we want a 95% confidence interval for the loading of the first principal component with respect to Sepal length.

```
library(boot)

getPrcStat <- function (samdf,vname,pcnum){
  prcs <- prcomp(samdf[1:4]) # returns matrix
  return(prcs$rotation[ vname,pcnum ])   # pick out the thing we need
}

bootEst <- function(df,d){
   sampledDf <- df[ d, ]  # resample dataframe
   return(getPrcStat(sampledDf,"Sepal.Length",1))
```

```
}

bootOut <- boot(iris,bootEst,R=10000)
boot.ci(bootOut,type=c("basic"))
```

Interpret the output.

Exercise 2: Employment Country Profile Data

This example considers the percentages of people employed in nine industry sectos in various European countries in the years from 1989 to 1995. 30 countries are considered and 9 different industry sectors (X_1 = AGR: agriculture forestry and fishing, X_2 = MIN: mining and quarrying, X_3 = MAN: manufacturing, X_4 = PS: power and water supplies, X_5 = CON: construction, X_6 = SER: services, X_7 = FIN: finance, X_8 = SPS: social and personal services and X_9 = TC: transport and communications. In addition, the countries were classified as to whether they belonged to the EU, or EFTA, or were Eastern European, or 'other'. The data is from 1995 and uses the classifications that were appropriate then. For 'USSR' read 'former USSR', for 'Yugoslavia' read 'former Yugoslavia', etc ... The data set is found on the course home page under `employment.csv`.

```
www<-"https://www.mimuw.edu.pl/~noble/courses/MultivariateStatistics/data/empl
oyment.csv"
employment <- read.csv(www)
```

Firstly, the sample correlation matrix for the standardised variables for the nine industries may be obtained:

```
> cormat2 <- cor(employment[,3:11])
> cormat2
```

The eigenvalues and eigenvectors may be obtained by:

```
> ev2 <- eigen(cormat2)
> ev2
```

The eigenvalues, with the percentages of the total of nine in parentheses, are
3.112(34.6%), 1.809(20.1%), 1.496(16.6%),
1.063(11.8%), 0.710(7.9%), 0.311(3.5%), 0.293(3.3%), 0.204(2.3%), 0.000(0.0%).
The last value is necessarily zero, because the data is *percentages* of the workforce, which necessarily adds up to 100. Therefore, although there are 9 variables, there are only 8 *free* variables. This information may be obtained as follows:

```
> pca2 <- prcomp(employment[,3:11],scale = TRUE)
> summary(pca2)
```

It is a matter of judgement whether or not to use 4 or 5 components. The first 4 components account for 83% of the variation; the first 5 account for over 90% of the variation. From the eigenvectors,

$$Z_1 = 0.51(AGR) + 0.37(MIN) \\ -0.25(MAN) - 0.31(PS) - 0.22(CON) - 0.38(SER) - 0.13(FIN) - 0.42(SPS) - 0.21(TC),$$

$$Z_2 = -0.02(AGR) + 0.00(MIN) \\ +0.43(MAN) + 0.11(PS) - 0.24(CON) - 0.41(SER) - 0.55(FIN) + 0.05(SPS) + 0.52(TC).$$

and the others similarly.

The first component contrasts (AGR) and (MIN) on the one hand with (MAN), (PS), (CON), (SER), (FIN), (SPS) and (TC) on the other hand.
The second component gives little or no weight to (AGR), (MIN), (SPS) and constrasts (MAN), (PS), (TC) with (CON), (SER), (FIN).

Interpretations for the other components may be derived similarly.

While PC1 and PC2 are uncorrelated when taken with respect to the *whole* data set, ignoring the classifications, it is interesting to plot PC1 against PC2, using different symbols for the four categories (Western EU, EFTA, Eastern European, Other). When `prcomp` is used and the results stored in `pca`, the principal component values are stored under `pca$x`. They may be added to the data frame in the following way:

```
> emp <- employment
> emp$pc1 <- pca2$x[,1]
> emp$pc2 <- pca2$x[,2]

> library("ggplot2")
> qplot(pc2,pc1,colour=group,data=emp)
```

The package **GPArotation** can make a varimax rotation.

```
> fit <- principal(emp[,3:11], nfactors=4, rotate="varimax")
> fit
```

Verify the results and the interpretation found in the lecture.

Exercise 3

Generate a random sample of size 100 from a three-dimensional Gaussian distribution, where one variable has a very large variance compared with the other two. Carry out a PCA on the data using the covariance matrix and then using the correlation matrix. In each case, find the eigenvalues and eigenvectors, draw the scree plot, compute the PC scores and plot all pairwise PC scores in a matrix plot. Compare results.

The **GGally** package enables the pairwise plots of principal components; the `ggpairs`. You'll get an idea of how it works here:

```
library(GGally)
library(ggplot2)
data(flea)
ggpairs(flea, columns = 2:4, ggplot2::aes(colour=species))
```

Exercise 4

Carry out a PCA on Fisher's `iris` data. The data consists of 50 observations on each of three species of iris: *Iris setosa*, *Iris versicolor* and *Iris virginica*. The four measured variables are sepal length, sepal width, petal length and petal width. Ignore the species labels. Compute the PC scores and plot all pairwise sets of PC scores in a matrix plot. Explain your results, taking into consideration the species labels.

Canonical Correlation

Exercise 5: Butterfly Data Work through the butterfly example in the lecture note (where the two sets of variables are geographical and genetic). The R code is in the script accompanying the tutorial. The data in the butterfly file can be used to illustrate the procedure. There are 16 colonies of the butterfly *Euphydryas editha* in California and Oregon. These vary with respect to four environmental variables (altitude, annual precipitation, annual maximum temperature and annual minimum temperature) and six genetic variables (percentages of six phosphoglucose-isomerase (Pgi) genes as determined by electrophoresis). The data may be found in `butterfly.dat` on the course home page. Significant relationships are of interest, because they may indicate that the butterfly has adapted to local environments.

Since there are fewer environment variables, the environmental variables have to be treated as X and the gene variables as Y. But all the gene frequencies *cannot* be used, since they are percentages and add up to 100.

Therefore, the 1.30 gene frequency may be omitted. It also seemed a good idea to combine the 0.40 and 0.60 gene fequencies. Thus the X variables considered are X_1 altitude, X_2 annual precipitation, X_3 annual maximum temperature, X_4 annual minimum temperature and the Y variables considered are Y_1 frequency of 0.4 and 0.6 gene, Y_2 frequency of 0.8 gene, Y_3 frequency of 1.0 gene, Y_4 frequency of the 1.16 gene.

FIRST these columns are *standardised*, because it is only *correlations* that are of interest here, and the rest of the analysis is performed using the standardised variables.

NEXT the correlation matrix C for the standardised variables is obtained (which is the same as the correlation matrix for the raw variables) and is partitioned into C_{11}, C_{22}, C_{12} as described earlier.

NEXT find the eigenvalues $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and eigenvectors $\underline{b}_1$, $\underline{b}_2$, $\underline{b}_3$, $\underline{b}_4$ which solve the eigenvalue problem

$$(C_{22}^{-1/2} C_{21} C_{11}^{-1} C_{12} C_{22}^{-1/2} - \lambda_j I_4)\underline{b}_j = 0$$

where I_4 denotes the 4×4 identity matrix. Then

$$C_{22}^{-1} Q^t = L_2^t = (\underline{b}_1, \underline{b}_2, \underline{b}_3, \underline{b}_4).$$

Here $\lambda_j = \rho_j^2$.
The canonical correlations are given by:

```
> canon$corr
```

The canonical variables $\underline{U} = HC_{11}^{-1/2}\underline{X}$ and $\underline{V} = QC_{22}^{-1/2}\underline{Y}$ are found by:

```
> canon
```

The correlations are between observed variables and canonical variables are known as the *canonical loadings*.

To perform tests of the significance of Canonical Correlation, try:

```
> canon$chisq
```

To 'look up' the chi squared table,

```
> pchisq(canon$chisq,canon$df,ncp=0)
```

Although the canonical correlations seem quite large, Bartlett's test does not reject the null hypothesis (that they are insignificant) because the sample size is rather small. It is found that $X^2 = 18.41$ with $16df$. The probability of obtaining a value greater than this is 0.30. This is not sufficiently small to reject the null hypothesis.

Nevertheless, the canonical correlation provides useful pointers at a descriptive level. U_1 is mainly a contrst between the maximum and minimum temperatures on the one hand and precipitation on the other. V_1 has moderate to large coefficients for Y_1, Y_2 and Y_4, with a small negative coefficient for Y_3. It appears that the $0.4, 0.6, 0.8$ and 1.16 genes tend to be frequent in colonies with high temperatures and low precipitation.

Exercise 6: Penguin Data Start by loading `tidyverse`. We'll use this for plots. We'll illustrate canonical correlation using the `cancor()` function in the base R stat package. The data is found in `palmer_penguins.csv` in the course data directory.

From this data set, prepare two datasets for CCA. Omit species/island (we'll see later how the canonical variables can be used for classification). Prepare two data sets for Canonical Correlation Analysis (CCA). Let X (the first data set) be `bill_depth_mm` and `bill_length_mm` and Y (the second data set) be `flipper_length_mm` and `body_mass_g`.

Perform a CCA. What are the coefficients (loadings) for the CCA? What are the correlations between the canonical variables in each pair? Remember to use scaled variables for the CCA. The relevant library is CCA and the function for canonical correlation is `cancor()`.

Now separate the pairs of canonical variates for each species. Verify that the correlation between each pair is the same when we take it species by species as when we take it overall.

The `mutate` command keeps the current columns of a data frame the same and adds new columns. Hence the `mutate` command used in the script adds columns we have created to the data frame.

Make a scatter plot of the first pair of canonical covariates.

Plot the first canonical variate of X against species.

Plot the first canonical variate of Y against species.

Make a scatter plot of the first pair of canonical variates, coloured by species.

Make a scatter plot of the second pair of canonical variates, coloured by species.

What are the correlations between the pairs of canonical variables?

Exercise 7: Psychological Traits and Academic Features The data in `mmreg.csv` on the course page contains 8 variables. The first three measure psychological features, while the remaining 5 relate to academic ability. We want to see how these two sets are related.

Perform a canonical correiation analysis.

Now test to see how many of the canonical correlations are significant. This does not seem to be available in any package, hence the test has to be programmed by hands. A script is available in the script to go with the tutorial. Perform and report the significance tests.

Finally, standardize the canonical coefficients for `psych` and `acad`.