BEUTH HOCHSCHULE
FÜR TECHNIK
BERLIN
University of Applied Sciences

**Prof. Tim Downie**

# Workshop 2 – 18 April 2018
# Higher Dimensional K-Means, Principal Components Analysis

### Exercise 1

As a similar exercise to Exercise 3 in Workshop 1, we will simulate a three dimensional matrix with three clusters.

Start Rstudio as usual, clear your workspace and save a new script file with a sensible name. You will use the function `plot3d` which is in the `rgl` library. Try the command

```
> library(rgl)
```

If an error occurs install the package

```
> install.packages("rgl")
```

and then try the `library` command again. Copy and paste the following code into the script file.

```
set.seed=(10)
x=matrix(rnorm(75*3),ncol=3)
x[1:25,1]=x[1:25,1]+5
x[51:75,2]=x[1:25,2]-6
truth<-rep(1:3,c(25,25,25))
pairs(x,col=truth+1)
plot3d(x)
```

(a) Make sure you understand the given code.

(b) Rotate and play about with the 3D plot. Try to find a viewing angle that differentiates the clusters well. Notice that the clusters are well enough separated that with the correct viewing angle the colours are not required.

(c) Run the `kmeans` function on this data with 3 clusters.

(d) Compare the output clusters with the known clusters.

(e) Colour the clusters and add the cluster means.

```
plot3d(x,col=km.out$cluster+1)
plot3d(km.out$centers,add=TRUE,col=2:4,type="s")
```

(f) Repeat the above using just two clusters. How can you describe the two clusters in one sentence?

(g) Generate a new matrix called `y` with 10 columns instead of 3, define the clusters in exactly the same way as above. Run `kmeans` with 3 centres on `x` and on `y`. Use the R function `table` to compare how many rows have been correctly and incorrectly assigned.

**Exercise 2 USA arrests data: PCA and clustering**

In James et al. you have read about the `USArrests` data frame, see `help(USArrests)`. Notice that these data were collected over 40 ago. The first part of this exercise closely follows *10.4 Lab 1* on page 401 on PCA. Then you will use K-Means clustering on the original data and on the PCA data. The aims of this exercise are:

- to learn about the principal components method,
- to see how many clusters are appropriate,
- to investigate the characteristics of each cluster (if possible)
- and to see if it makes a difference between clustering the original data or clustering the principal components.

Note that there is no right answer here! we are primarily trying to learn more about the data as opposed to proposing a good model. Work through the following using *10.4 Lab 1* in James et al. to help you.

(a) Obtain names of the four variables in `USArrests`.

(b) For each variable obtain the mean and the standard deviation.

(c) Use `prcomp` to get the principal components for this dataset.

(d) What is the difference between `pr.out$scale` and `pr.out$sd`?

(e) Create the *scree plot* and *cumulative variance plot* plots as in James. The first two Principal components explain 87% of the variance, so we will use just the first two.

(f) Obtain the `biplot` using

```
> biplot(pr.out,xlabs=state.abb)
```

`state.abb` contains the standard US state abbreviations in the correct order, giving a neater presentation, than that given in James et al.

(g) Run K-means on the PCA data with 2 clusters and plot the results. Note that the function `biplot` does not allow the points to be coloured. Make sure you understand the two plotting commands below.

```
> km.out<-kmeans(pr.out$x,centers=2,nstart=20)
> plot(pr.out$x[,1:2],type="n")
> text(pr.out$x[,1],pr.out$x[,2],labels=state.abb,col=km.out$cluster)
```

How many states are in each cluster.

(h) Adapt the code blow to obtain the within sum of squares using 1 to 10 clusters plotting the results. At which value of $k$ is there an „Elbow”?

```
wss.vec<-rep(NA,??)
for(k in 1:??){
  km.out<-??(??,centers=?,nstart=20)
  wss.vec[?]<-km.out$???
}
plot(???,type="b")
```

(i) Using the “best” number plot the principal components coloured by cluster. With just 4 variables a `pairs` plot is also worthwhile.

(j) Repeat parts g, h and i clustering on the original `USArrests` data frame, and compare the results.

### Exercise 3 Hamburg decathlon data, PCA and clustering

You have already met these data in Visualising Data. The decathlon data[1] contain the results of each discipline along with other information such as age number of points gained for each competitor. The data file is provided in Moodle. You should download and save this file in the usual way. The data needs some preprocessing. In particular quite a few competitors did not complete all the events and so have missing values. Because of this we will restrict the data only to those who completed the entire decathlon. The code to do this (which includes some code from Prof. Grömping) is given on the next page.

[1] https://www.10-kampf.de/ergebnisse/

```
decathlon <- read.csv2("Data/Zehnkampf2017Hamburg.csv",
                       stringsAsFactors = FALSE)[-c(1:3,seq(10,28,2))]
## transform times to seconds
hilf <- decathlon$Zeit.400m
decathlon$Zeit.400m <- 60*as.numeric(substr(hilf,1,2)) +
  as.numeric(gsub(",",".",substr(hilf,4,nchar(hilf))))
hilf <- decathlon$Zeit.1500m
decathlon$Zeit.1500m <- 60*as.numeric(substr(hilf,1,2)) +
  as.numeric(gsub(",",".",substr(hilf,4,nchar(hilf))))
colnames(decathlon) <- c("YearOfBirth", "Class", "Points",
            "Day1", "Day2", "Time.100m", "LongJump", "ShotPut",
            "HighJump", "Time.400m", "Hurdles", "Discus", "PoleVault",
            "JavelinThrow", "Time.1500m")

#remove any competitors without complete data, and select just the event results
temp <- apply(decathlon,1,function(x) sum(is.na(x)))
clustermat<-decathlon[temp==0,6:15]
pairs(clustermat)
```

(a) Run the code above. The pairs plot is not pretty, but we can nevertheless see some strong correlations. Which combination of disciplines have a large positive or large negative correlation, and is this what you would expect?

(b) Use `prcomp` to get the principal components for this dataset and plot the first two using `biplot`. What characteristic is clear to see in the first principal component ($x$-axis)? Can you suggest a characteristic to summarise the second PC? Note also that there are 4 clear outliers.

(c) `< plot3d(pr.out$x[,1:3])` creates an interactive plot of the first three PCs. It is not so easy to learn anything more than the biplot tells us.

(d) Create a *cumulative variance plot* for the principal components. 3 PCs is probably a good choice.

(e) Use what you have learnt in Exercise 2 to carry out a $k$-means cluster analysis on these data. In particular, show that the clusters obtained using `clustermat` are heavily dependent on the 1500m times. Why is this? This would suggest that clustering the PC data matrix is a better approach.