Prelims Data Analysis TT 2020
Sheet 7

At the end of this exercise sheet there are Practical Exercises in `R` and `Matlab`. Students should ask their college tutor whether to use `R` or `Matlab`. The course website has an Introduction to `R`, which students should work through before starting the `R` exercises.

1. Let $\mathbf{X}$ be a mean centered $n \times p$ data matrix and let $\mathbf{Z}$ be the corresponding $n \times p$ scores matrix. Show that the sample covariance of the scores matrix is diagonal. What is the interpretation of this result?

2. Let $\mathbf{X}$ be a mean centered $n \times p$ data matrix.
   (i) Define the sample covariance matrix $\mathbf{S}$ in terms of $\mathbf{X}$.
   (ii) If $\mathbf{W}$ be a diagonal matrix with entries $\mathbf{S}_{ii}$ for $i \in 1, \ldots, p$ and $\mathbf{R}$ is the sample correlation matrix, how can $\mathbf{R}$ be written in terms of $\mathbf{S}$ and $\mathbf{W}$.
   (iii) Show that the PCA components derived from using the sample covariance matrix $\mathbf{S}$ will be equivalent to those derived using the sample correlation matrix $\mathbf{R}$ when the variances of the p variables are all equal.

3. Suppose $\mathbf{X}$ is a mean centered data matrix, and let $\widetilde{\mathbf{X}} = z_1 w_1^T$ be the best rank-1 approximation to $\mathbf{X}$, where z_1 is an n-column vector and w_1 is a p-column vector with $w_1^T w_1 = 1$.
   (i) How is w_1 related to the eigendecomposition of $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$?
   (ii) Show that $z_1 = \mathbf{X}w_1$ and that $\widetilde{\mathbf{X}} = \mathbf{X}w_1 w_1^T$
   (iii) Consider the sum of the squared differences between $\mathbf{X}$ and $\widetilde{\mathbf{X}}$ defined as

$$d = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{p}(\mathbf{X}_{ij} - \widetilde{\mathbf{X}}_{ij})^2$$

   Show that d is equal to each of the following expressions
   (a) $\frac{1}{n-1}tr\Big((\mathbf{X} - \widetilde{\mathbf{X}})(\mathbf{X} - \widetilde{\mathbf{X}})^T\Big)$
   (b) $\frac{1}{n-1}tr(\mathbf{X}^T\mathbf{X}) - \lambda_1$ where λ_1 is the largest eigenvalue of $\mathbf{S}$.
   (c) $\sum_{i=2}^{p}\lambda_i$ where λ_i $i = 1, \ldots, p$ are the eigenvalues of $\mathbf{S}$.

Practical Exercises using R

Students should carry out these practical exercises and produce a report summarizing the results of their analysis i.e. produce a document that contains the plots produced and hand this in to your tutor.

NOTE To run these exercises in R you will need to install a few packages called MASS, stats and car. To do this in RStudio click

Tools $->$ Install Packages

and then type in the names of the packages and install them. Make sure to click the box that says "install dependencies"

1. The Crabs dataset is in the MASS library which can be loaded using

   ```
   library(MASS)
   ```

   Create a new dataset with the 5 main variables as follows

   ```
   varnames = c('FL','RW','CL','CW','BD')
   Crabs = crabs[,varnames]
   ```

   Load stats library using

   ```
   library(stats)
   ```

   Run PCA on the scaled dataset using

   ```
   f1 = prcomp(Crabs, scale = TRUE, retx = TRUE)
   ```

   Produce a pairs plot of the PCs using

   ```
   pairs(f1$x)
   ```

   Plot of the 2nd and 3rd PCs using

   ```
   plot(f1$x[,2:3])
   ```

   Produce the scree plot using

   ```
   barplot(f1$sdev)
   ```

   Look at the loadings matrix using

   ```
   f1$rotation
   ```

2. Download the EU indicators dataset from

   www.stats.ox.ac.uk/~sejdinov/teaching/data/eu.csv

   Load stats library using

   ```
   library(stats)
   ```

   Load the dataset into R using

   ```
   eu = read.csv("eu.csv",sep="", hea = T, row.names = 1)
   ```

   Note you will need to change the command so that file includes the path to its location on your computer.

   Look at the dataset using

   ```
   eu
   ```

   Run PCA on the scaled dataset using

   ```
   f2 = prcomp(eu[,-1], scale = TRUE, retx = TRUE)
   ```

   Plot of the 1st and 2nd PCs and label the points using

   ```
   plot(f2$x[,1:2])
   text(f2$x[,1:2], labels = rownames(eu), pos = 4, offset = 1)
   ```

   make a biplot using

   ```
   biplot(f2)
   ```

   What happens when you don't scale the dataset?

3. Download the Single Cell Genomics dataset from

   www.stats.ox.ac.uk/~sejdinov/teaching/data/single_cell.data

   Load the dataset into `R` using

   ```
   load("single_cell.data")
   ```

   This creates a data matrix object called `X`.

   Note you will need to change the command so that file includes the path to it's location on your computer.

   Run the PCA using

   ```
   f3=prcomp(X, scale = TRUE, retx = TRUE)
   ```

   Plot the 1st and 2nd PCs using

   ```
   plot(f3$x[,1:2], xlab="PC1", ylab="PC2", col="blue", pch=16)
   ```

   Load `car` library using

   ```
   library(car)
   ```

   Create an interactive 3D plot using

   ```
   scatter3d(x = f3$x[,1], y = f3$x[,2], z = f3$x[,3], point.col = "blue", pch = 16, surface = FALSE,
   xlab = "PC1", ylab = "PC2", zlab = "PC3")
   ```

   Plot the scree plot. How much variance is contained in the first 10 PCs?

Practical Exercises using Matlab

Students should carry out these practical exercises and produce a report summarizing the results of their analysis i.e. produce a document that contains the plots produced and hand this in to your tutor.

NOTE If you get the error: Undefined function or variable 'princomp'." when you try to run the PCA commands, then you might not have the "Statistics and Machine Learning Toolbox" installed with your copy of matlab. If you rerun the matlab installer that you used for your previous matlab course, you should be able to add the "Statistics and Machine Learning Toolbox" to your installation by following the instructions here:
http://uk.mathworks.com/help/install/ug/install-mathworks-software.html
You might have to download the matlab installer again in case you've deleted it since your previous matlab course. You can download it again using this link: https://www.maths.ox.ac.uk/members/it/software-personal-machines/matlab

1. Download the Crabs dataset from

   www.stats.ox.ac.uk/~sejdinov/teaching/data/crabs.txt

   Load the dataset into Matlab using

   ```
   crabs = readtable('crabs.txt', 'Delimiter', 'space');
   ```

   Create a new dataset with the 5 main variables as follows

   ```
   varnames =  'FL' 'RW' 'CL' 'CW' 'BD' ;
   Crabs = crabs(:, varnames);
   ```

   Run PCA on the scaled dataset using

   ```
   X1 = table2array(Crabs);
   for d = 1:size(X1, 2)
   X1(:, d) = X1(:, d) - mean(X1(:, d));
   X1(:, d) = X1(:, d)/std(X1(:, d), 1);
   end
   [coeff1, score1, latent1] = pca(X1);
   ```

   Produce a pairs plot of the PCs using

   ```
   corrplot(score1);
   ```

   Plot of the 2nd and 3rd PCs using

   ```
   plot(score1(:, 2), score1(:, 3), 'o');
   ```

   Produce the scree plot using

   ```
   sdev = std(score1);
   bar(sdev);
   ```

   Look at the loadings matrix

   ```
   coeff1
   ```

2. Download the EU indicators dataset from

   www.stats.ox.ac.uk/~sejdinov/teaching/data/eu.csv

   Load the dataset using

   ```
   eu = readtable(' /Desktop/eu.csv', 'Delimiter', 'space');
   ```

   Look at the dataset

   ```
   eu
   ```

   Run PCA on the scaled dataset

   ```
   X2 = table2array(eu(:, 3:end));
   for d = 1:size(X2, 2)
   X2(:, d) = X2(:, d) - mean(X2(:, d));
   X2(:, d) = X2(:, d)/std(X2(:, d), 1);
   end
   ```

```
[coeff2, score2, latent2] = pca(X2);
```

Plot the 1st and 2nd PCs using

```
plot(score2(:, 2), score2(:, 3), 'o');
text(score2(:, 2), score2(:, 3), eu.Countries);
```

Make a biplot

```
vbls = eu.Properties.VariableNames(1, 3:end);
biplot(coeff2(:,1:2), 'scores', score2(:,1:2), 'varlabels',vbls);
```

3. Download the Single Cell Genomics dataset from

   www.stats.ox.ac.uk/~sejdinov/teaching/data/single_cell.X.csv

   Load the dataset

   ```
   cells = readtable(' /Desktop/single_cell.X.csv', 'Delimiter', 'space');
   ```

   Note The "warning" on this readtable command is raised because some of the column names of "single_cell" contain periods or dashes. MATLAB will convert these both to underscores.

   Run the PCA

   ```
   X3 = table2array(cells(:, 2:end));
   for d = 1:size(X3, 2)
   X3(:, d) = X3(:, d) - mean(X3(:, d));
   X3(:, d) = X3(:, d)/std(X3(:, d), 1);
   end
   [coeff3, score3, latent3] = pca(X3);
   ```

   Plot the 1st and 2nd PCs

   ```
   plot(score3(:, 1), score3(:, 2), 'o');
   xlabel('PC1');
   ylabel('PC2');
   ```

   Create an interactive 3D plot

   ```
   scatter3(score3(:, 1), score3(:, 2), score3(:, 3));
   xlabel('PC1');
   ylabel('PC2');
   zlabel('PC3');
   ```