# Tutorial 3 Principal Component Analysis

## Lecture

### Principal Component Analysis

Last week, lectures covers the section on Principal Component Analysis (PCA). PCA will probably be the best 'trick' you will learn in this subject. It is a very popular data mining method. The idea of PCA is to condense information spread across from dozens of variables into a much smaller set of variables called *Principal Components*. The Principal Components are *orthogonal* too. Orthogonal variables can be handy as you don't have to worry about numerical problems that are created from highly correlated variables.

**Textbook Reading**

## Prescribed reading

| Author | Relevant chapters |
|---|---|
| James et al | 12.1 -12.2 & 12.5.1 |
| Hastie et al | 14.5 |

In a group of two, discuss the following topics

- What is principal component analysis?
- What is the key objective pf principal component analysis?
- What are the main characteristics of the components in PCA?
- When does PCA fail to achieve its objective?

### Part 1: PCA theory

During the lecture this week, I skipped over the theory behind PCA. It is however crucial to grasp how PCA works. Since students in MA3405 have varying levels of mathematical expertise, we will take a step-by-step approach to cover the theory of PCA thoroughly.

We will use the `NCI60` data in `ISLR2` package for demonstration. We will start by loading the dataset on the working directory,

```
library(ISLR2)
nci.labs <- NCI60$labs
table(nci.labs)

nci.data <- NCI60$data
dim(nci.data)
```

`NCI60` data contain gene expression level of 6830 genes from 64 cancer cell lines. The format is a list containing two elements: `data` and `labs`. `data` is a 64 by 6830 matrix of the expression values while `labs` is a vector listing the cancer types for the 64 cell lines.

Principal components are orthogonal vectors which explains variation of the data. The goal of a PCA is to find a new set of variables (i.e. smaller than original number of variables) that best describe the variation in

the dataset. This is achieved via eigendecomposition of the covariance matrix. Therefore, the first step is to find the covariance matrix of centred **nci.data** with `cov` funciton. TO save computation time, we will only use the expression of first 500 genes,

```
nci.sub<- nci.data[, 1:500]
#Centering
nci.sub.scale<-scale(nci.sub, scale = TRUE)
var.cov<-cov(nci.sub)
```

The next step is to eigenden composition of the variance-covariance matrix. To do this in R, we need to use `eigen` function in `matlib` library,

```
library(matlib)
e_decom<-eigen(var.cov)
```

The function returns two elements, the first element is the eigenvalue of var-cov matris, while the second element is the eigenvector. We can see amount of variation explained by dividing each eigenvalue by the numebr of sample -1,

```
e_var<-e_decom$values/(nrow(nci.sub.scale)-1)
#We can convert this into percentage
e_var_per<-e_var/sum(e_var)
```

We can use `cumsum` function to see cumulative variance explained by each eigenvector,

```
cumsum(e_var_per)
```

The first 35 eigenvectors explained around 90% variation, and the 62 eigenvectors explained 99% of vairation. Let's produce a scree plot,

```
library(ggplot2)
pp<-data.frame("PC"=c(1:12), PER=e_var_per[1:12])
pp

ggplot(pp, aes(x = PC, y = PER)) +
  geom_col(width = 0.5, color = "black") +
  xlab("Principal component") +
  ylab("Percentage of variation (%)") +
  theme_classic()
```

Let's use `princomp` function in R to run PCA,

```
pca<-prcomp(nci.sub, center = TRUE)
summary(pca)
```

`summary` shows amount of variance explained by each component, these are the same as the values we got using eigen decomposition (`pp`). There is however some variation in the factor loading, this is because `princomp` uses Single value decomposition instead of eigen decomposition.

## Independendent Learning

### Labs

12.5.1 Principal Components Analysis
12.5.4 NCI60 Data Example

**Question** Exercise 8, Chapter 12

**Question** Download the sparrow2.csv data from LearnJCU. This dataset consists of seven morphological variables taken from 1026 sparrows. The seven variables were

- wingcrd = wingcord
- flating = flattened wing
- tarsus = leg
- head = bill tip to back of skull
- culmen = beak length
- nalopsi = bill tip to nostril
- weight

Conduct exploratory data analysis to check if there is correlation between variables.

Perform PCA and answer the following questions:

1. How much variation is explained in the (i) first (ii) and (iii) principal component analysis?
2. How many principal components do you recommend using? Why?
3. Can you describe the first two principal components?
4. Interpret any interesting features in the biplot.