MA5810: Introduction to Data Mining

Week 5; Collaborate Session 1: Principal Components Analysis

Martha Cooper, PhD

JCU Masters of Data Science

2019-21-9 (updated: 2020-11-25)

Housekeeping

- Collaborate 1 = Wednesdays 6-7pm (Martha)
- Collaborate 2 = Thursdays 7-8pm (Hongbin)

For my Collaborate Sessions, you can get the slides & R code for each week here:

https://github.com/MarthaCooper/MA5810

Assessments

Next week's collaborate session 1 will focus on:

- Common mistakes made in Assessment 1
- Clarification for the Capstone project (Assessment 3)

Subject: MA5810 Intro to Data Mining

MA5810 Learning Outcomes

1. Overview of Data Mining and Examples
2. Unsupervised data mining methods e.g. clustering and outlier detection;
3. Unsupervised and supervised techniques for dimensionality reduction (Today = PCA);
4. Supervised data mining methods for pattern classification;
5. Apply these concepts to real data sets using R (Today).

Today's Goals

- Understand the background behind Principal Components Analysis (PCA)
- Understand the pros and cons of PCA
- Apply PCA to real datasets using R

Unsupervised Learning

A set of statistical tools to understand a set of features, $X_1, \ldots Xp$, without having an associated response variable, Y, to predict.

Data visualization, identification of subgroups & dimensionality reduction

Principal Components Analysis (PCA)

A technique for summarizing a large set of variables into a smaller number of representative variables that collectively explain most of the variation in the original set.

PCA looks to find a low-dimensional representation of the observations that explain a good fraction of the variance

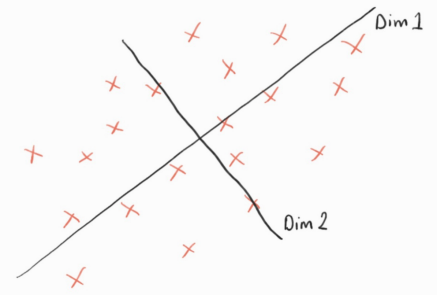
Why reduce dimensions?

Problems:

- Correlated variables
- A large number of variables

Solutions:

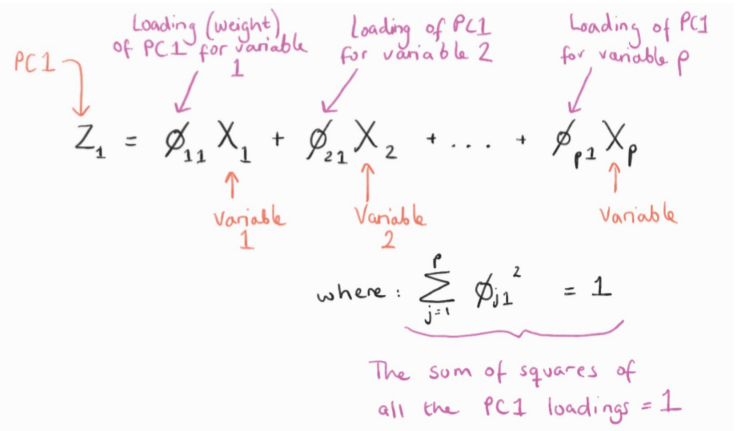
- A new set of variables that are *uncorrelated* and *explain as much variance as possible*
- *The best combination* of all the variables that *explains the original data set with less variables*



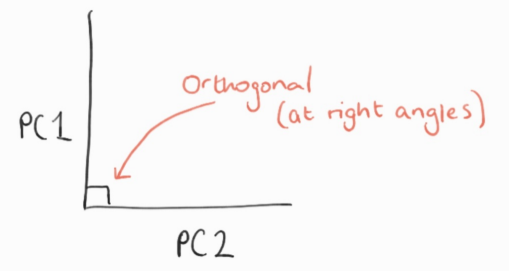
Finding the Prinicpal Components

- Transform the data to a small number of interesting dimensions
- Interesting = Highest Variance
- These dimensions (Principal Components) are:

1. (Normalised) Linear combinations of the original variables

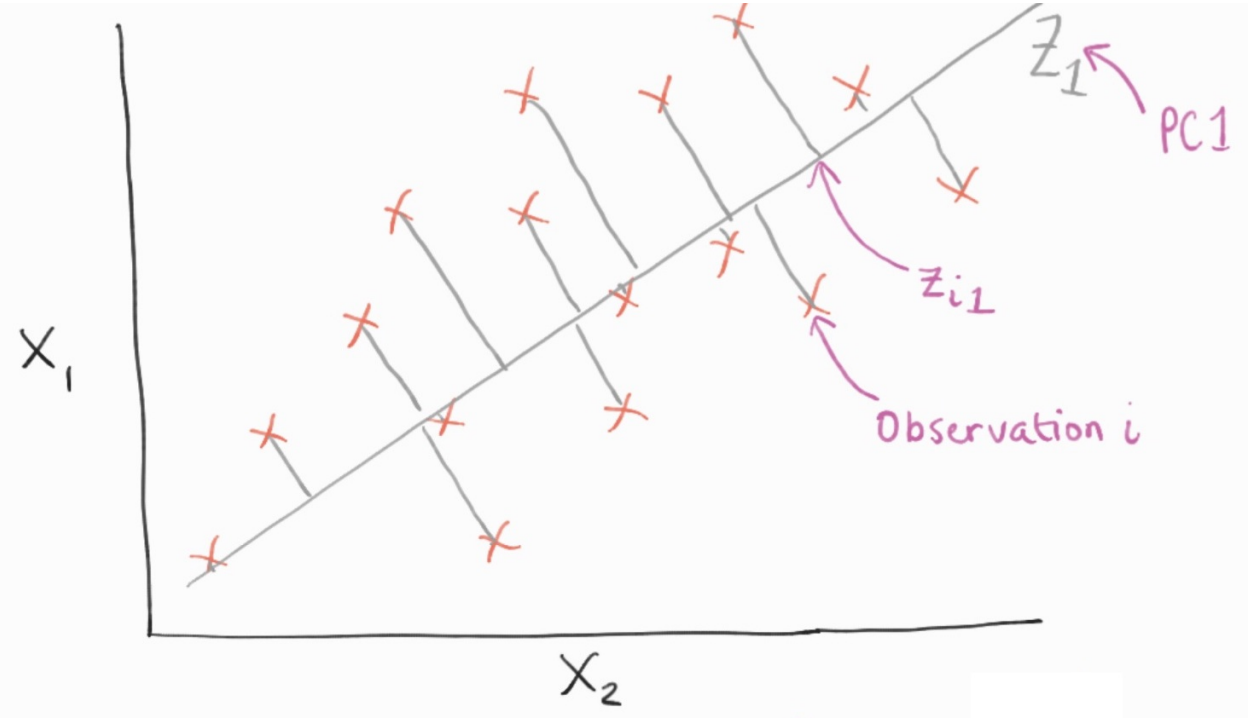


2. Uncorrelated



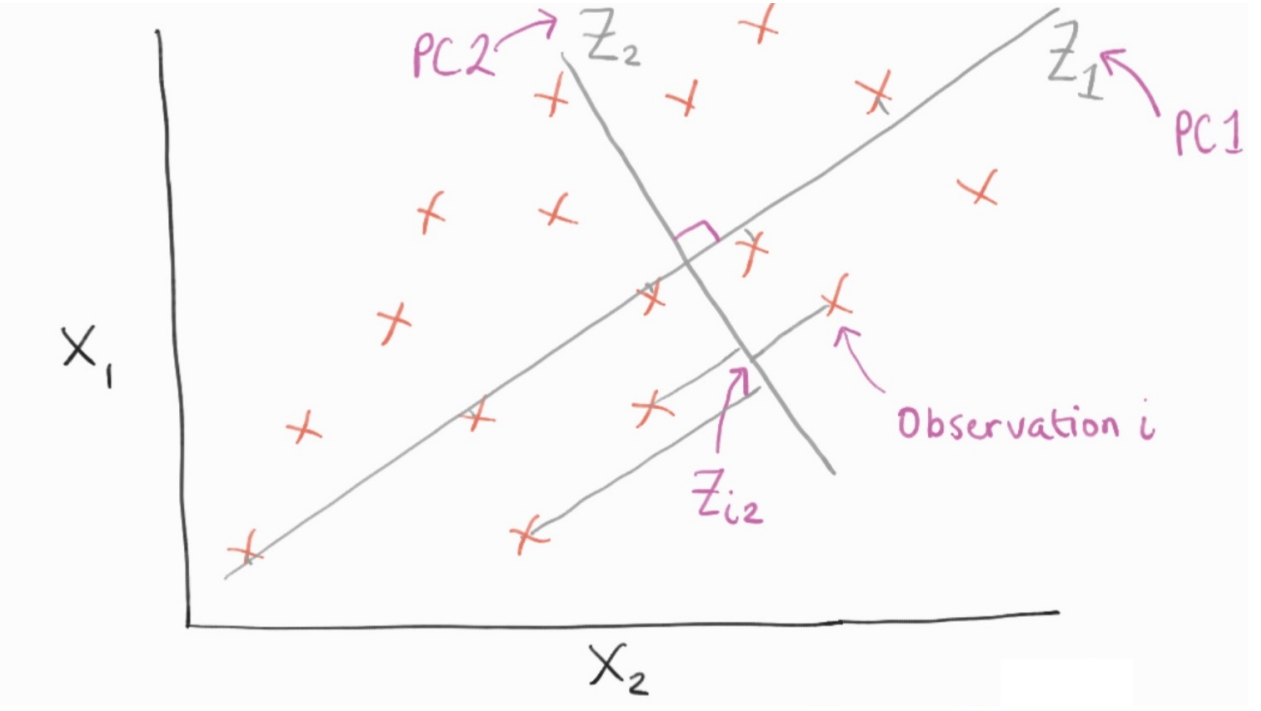
Finding the Prinicipal Components

- How do we choose the loadings that cause PC1 to explain the most variance in the data?



Finding the Prinicipal Components

- How do we choose PC2 - the second biggest source of variation & uncorrelated?



Pros and Cons

Pros

- Reduce number of predictors
- Reduce number of correlated predictors
- Identify subgroups in our dataset
- Identify outliers

Cons

- Subjective
- Exploratory data analysis
- Difficult to assess results

PCA in R

```
head(iris) #data
dat <- iris[ ,1:4] # remove Species column

pc <- prcomp(dat, center = T, scale = T) #why center? why scale?

pc$rotation #loadings - matrix of variable loadings
pc$x #scores - the coordinates of the observations on each PC
```

Visualising PCA in R

```r
library(factoextra)

fviz_screeplot(pc) #scree plot - proportion of variance explained
# How might we choose how many PCs to keep?

fviz_pca_ind(pc) #PC1 vs PC2

fviz_pca_biplot(pc) #biplot
```

Interpretting PCA with domain knowledge

```
fviz_pca_ind(pc,
             habillage = iris$Species,
             addEllipses = TRUE)
```

Extra reading

- Chapter 10.2 of James *et al.*, ISLR

References

- James *et al.*, ISLR

Slides

- xaringhan, xaringanthemer, remark.js, knitr, R Markdown