

08 PCA

Dien Giau (Richard) Bui

4/20/2022

Load required packages

```
library(tidyverse)
```

```
## -- Attaching packages -----
```

```
## v ggplot2 3.3.3      v purrr    0.3.4
```

```
## v tibble  3.1.1      v dplyr    1.0.5
```

```
## v tidyr   1.1.3      v stringr  1.4.0
```

```
## v readr   1.4.0      v forcats  0.5.1
```

```
# if you're using macOS, you can run: library(dplyr)
```

```
library(skimr)
```

```
library(broom)
```

```
library(modelr)
```

Introduction

- ▶ A data table is a space of information
 - ▶ Each column/variable is a feature/dimension that adds more information to our understanding about a problem
- ▶ When data is bigger and bigger, the information set gets bigger and bigger
- ▶ For example, data can have more than 1,000 variables
 - ▶ We need to analyze statistics of each variable and do some tests like t-test to understand the data
 - ▶ But very exhausted

Dimension reduction

- ▶ That's why in data science, people think out a way to reduce the dimension of data
- ▶ Say, how to combine 1,000 variables into a few variables that can capture most of information in the data
- ▶ It is similar to news summary by the end of the day:
 - ▶ Instead of reading all 1,000 news articles, we just need to read 2-3 news summary by the end and understand most of things happen in today
- ▶ Today's lecture will introduce to you that skill: Principal component analysis or PCA

An overview of PCA

- ▶ Combine features into a smaller set of principal component:
 - ▶ Each component has score: which is linear combination of features
- ▶ Each component aims to explain the highest variance/variation in the data
 - ▶ The first component explains most of variance
 - ▶ The second component continues to explain a bit
 - ▶ and so on

An example

- ▶ We have a data of crime in 50 US states with 4 variables:
 - ▶ Assault: the number of assault arrests/100,000 residents
 - ▶ Murder: the number of murder arrests/100,000 residents
 - ▶ Rape: the number of rape arrests/100,000 residents
 - ▶ UrbanPop: the percent of population in each state living in urban areas

Import data

```
Crime = USArrests  
head(Crime)
```

##	Murder	Assault	UrbanPop	Rape
## Alabama	13.2	236	58	21.2
## Alaska	10.0	263	48	44.5
## Arizona	8.1	294	80	31.0
## Arkansas	8.8	190	50	19.5
## California	9.0	276	91	40.6
## Colorado	7.9	204	78	38.7

Skim statistics

```
skim_without_charts(Crime) %>%  
  as.data.frame() %>%  
  select(var=skim_variable, mean=numeric.mean,  
         sd=numeric.sd)
```

##	var	mean	sd
## 1	Murder	7.788	4.355510
## 2	Assault	170.760	83.337661
## 3	UrbanPop	65.540	14.474763
## 4	Rape	21.232	9.366385

Discussion

- ▶ Assault has the highest variance and mean
- ▶ UrbanPop has different units with other variables

To run PCA in R

- ▶ Pretty simple:

```
pca_out = prcomp(Crime)
names(pca_out)
```

```
## [1] "sdev"      "rotation" "center"    "scale"     "x"
```

Output of PCA

- ▶ `center` and `scale` show means and standard deviations of the variables that were used for scaling prior to implementing PCA
- ▶ `rotation`: provides the principal component loadings, which is how each variables contribute to a principal component
- ▶ `x`: principal component scores, which is linear combination of all variables
- ▶ `sdev`: standard deviation of each principal component

rotation: principal component loadings

```
pca_out$rotation
```

	PC1	PC2	PC3	PC4
## Murder	0.04170432	-0.04482166	0.07989066	-0.99492173
## Assault	0.99522128	-0.05876003	-0.06756974	0.03893830
## UrbanPop	0.04633575	0.97685748	-0.20054629	-0.05816914
## Rape	0.07515550	0.20071807	0.97408059	0.07232502

- ▶ Assault contributes mostly to PC1
- ▶ It just simply ignore the other variables' information
- ▶ It may raise concern that we lose so much money
 - ▶ Units and each variable variance is very important and can affect our PCA analysis
 - ▶ It is better to scale data before running PCA, let's do again

Scaled PCA

```
pca_out = prcomp(Crime, scale. = TRUE)
pca_out$rotation
```

##		PC1	PC2	PC3	PC4
##	Murder	-0.5358995	0.4181809	-0.3412327	0.64922780
##	Assault	-0.5831836	0.1879856	-0.2681484	-0.74340748
##	UrbanPop	-0.2781909	-0.8728062	-0.3780158	0.13387773
##	Rape	-0.5434321	-0.1673186	0.8177779	0.08902432

- Now PC1 contains information from each variable

How good is a PCA?

- ▶ How much variance of dataset can be explained by PC?
 - ▶ If PC can explain much of variance of dataset, it means that we do capture much of information in the data
- ▶ In R, we need to calculate the variance explained by each principal component and draw the scree plot

Variance explained by each component

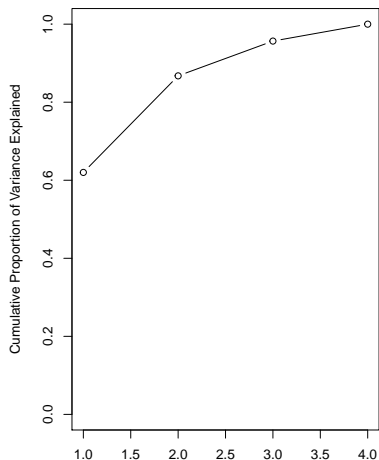
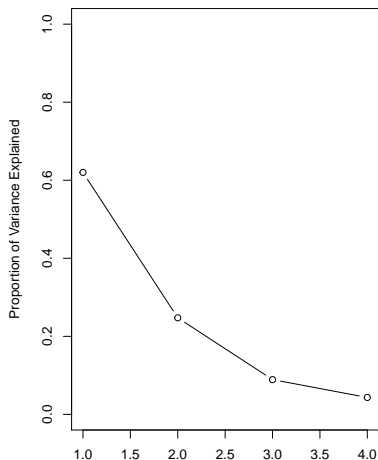
```
pca_var = pca_out$sdev^2  
pve = pca_var/sum(pca_var)  
pve
```

```
## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

- ▶ So PC1 explains 62% of data variance
- ▶ PC2 explains 24.7% of data variance
- ▶ So only two PC we can explain around 87% of data variance

Scree plot

```
par(mfrow = c(1, 2))  
plot(pve, xlab = "Principal Component", ylab = "Proportion",  
plot(cumsum(pve), xlab = "Principal Component", ylab = "Cum
```



Practice time

- ▶ In class, we will practice more