

Principal Component Analysis

Dian Ramadhani

08/01/2020

Principal Component Analysis

Analisis Komponen Utama (Principal Component Analysis) adalah analisis multivariate yang mentransformasi variabel-variabel asal yang saling berkorelasi menjadi variabel-variabel baru yang tidak saling berkorelasi dengan mereduksi sejumlah variabel sehingga mempunyai dimensi yang lebih kecil namun dapat menerangkan sebagian besar keragaman variabel aslinya.

Di praktek kali ini, kita akan mencoba PCA menggunakan dataset yang sederhana dan mudah dipahami yaitu dataset “mtcars” yang telah disediakan oleh R. Dataset ini terdiri dari data 32 model mobil, yang diambil dari American motoring magazine (1974 Motor Trend magazine). Setiap mobil memiliki 11 fitur yang dinyatakan dalam berbagai unit.

Install Packages

```
# Menginstall package(s)
install.packages("factoextra") # visualisasi plot pca
```

Import Library

```
# Mengaktifkan package(s)
library(factoextra)
```

Import Data

Dataset kali ini telah disediakan oleh R. Data berjudul “mtcars” dapat digunakan dengan perintah berikut.

```
# Mengimport data
data(mtcars)
```

data “mtcars” diimpor menjadi tabel bernama “mtcars”

Eksplorasi Data

Data yang telah diimpor selanjutnya dieksplorasi untuk mengetahui strukturnya.

```
# Melihat attribute dan struktur data
names(mtcars) # menampilkan nama kolom
```

```
## [1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs" "am" "gear"
## [11] "carb"
```

```
dim(mtcars) # menampilkan dimensi tabel
```

```
## [1] 32 11
```

```
head(mtcars) # menampilkan beberapa data teratas
```

```
##           mpg cyl disp  hp drat   wt  qsec vs am gear carb
## Mazda RX4    21.0   6  160 110 3.90 2.620 16.46 0  1    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02 0  1    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61 1  1    4    1
```

```
## Hornet 4 Drive      21.4    6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout  18.7    8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant            18.1    6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
str(mtcars) # menampilkan struktur data
```

```
## 'data.frame':    32 obs. of  11 variables:
## $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
## $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
## $ disp: num  160 160 108 258 360 ...
## $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
## $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
## $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
## $ qsec: num  16.5 17 18.6 19.4 17 ...
## $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
## $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
## $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
## $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
summary(mtcars) # menampilkan rangkuman data
```

```
##      mpg          cyl          disp          hp
##  Min.   :10.40    Min.   :4.000    Min.   : 71.1    Min.   : 52.0
## 1st Qu.:15.43    1st Qu.:4.000    1st Qu.:120.8    1st Qu.: 96.5
## Median :19.20    Median :6.000    Median :196.3    Median :123.0
## Mean   :20.09    Mean   :6.188    Mean   :230.7    Mean   :146.7
## 3rd Qu.:22.80    3rd Qu.:8.000    3rd Qu.:326.0    3rd Qu.:180.0
## Max.   :33.90    Max.   :8.000    Max.   :472.0    Max.   :335.0
##      drat          wt          qsec          vs
##  Min.   :2.760    Min.   :1.513    Min.   :14.50    Min.   :0.0000
## 1st Qu.:3.080    1st Qu.:2.581    1st Qu.:16.89    1st Qu.:0.0000
## Median :3.695    Median :3.325    Median :17.71    Median :0.0000
## Mean   :3.597    Mean   :3.217    Mean   :17.85    Mean   :0.4375
## 3rd Qu.:3.920    3rd Qu.:3.610    3rd Qu.:18.90    3rd Qu.:1.0000
## Max.   :4.930    Max.   :5.424    Max.   :22.90    Max.   :1.0000
##      am          gear          carb
##  Min.   :0.0000    Min.   :3.000    Min.   :1.000
## 1st Qu.:0.0000    1st Qu.:3.000    1st Qu.:2.000
## Median :0.0000    Median :4.000    Median :2.000
## Mean   :0.4062    Mean   :3.688    Mean   :2.812
## 3rd Qu.:1.0000    3rd Qu.:4.000    3rd Qu.:4.000
## Max.   :1.0000    Max.   :5.000    Max.   :8.000
```

```
# Mengetahui jumlah data kosong
```

```
sum(is.na(mtcars))
```

```
## [1] 0
```

Kita akan mengecualikan dua variabel kategorik dalam bentuk biner (0 dan 1) pada tabel “mtcars” yaitu variabel “vs” dan “am”. Data baru tanpa kedua variabel kategorik tersebut kemudian dijadikan tabel dengan nama “df.mtcars”.

```
# Memilih data
```

```
df.mtcars <- subset(mtcars, select = -c(vs, am))
```

```
View(df.mtcars)
```

Membuat Model PCA

Model PCA diterapkan pada tabel “df.mtcars” dan model diberi nama “pca”.

```
# Membuat model PCA
pca <- prcomp(df.mtcars, scale = TRUE)

# Menampilkan model PCA
pca

## Standard deviations (1, ..., p=9):
## [1] 2.3782219 1.4429485 0.7100809 0.5148082 0.4279704 0.3518426 0.3241326
## [8] 0.2418962 0.1489644
##
## Rotation (n x k) = (9 x 9):
##          PC1      PC2      PC3      PC4      PC5
## mpg  -0.3931477  0.02753861 -0.22119309 -0.006126378 -0.3207620
## cyl   0.4025537  0.01570975 -0.25231615  0.040700251  0.1171397
## disp  0.3973528 -0.08888469 -0.07825139  0.339493732 -0.4867849
## hp    0.3670814  0.26941371 -0.01721159  0.068300993 -0.2947317
## drat -0.3118165  0.34165268  0.14995507  0.845658485  0.1619259
## wt    0.3734771 -0.17194306  0.45373418  0.191260029 -0.1874822
## qsec -0.2243508 -0.48404435  0.62812782 -0.030329127 -0.1482495
## gear -0.2094749  0.55078264  0.20658376 -0.282381831 -0.5624860
## carb  0.2445807  0.48431310  0.46412069 -0.214492216  0.3997820
##          PC6      PC7      PC8      PC9
## mpg   0.72015586 -0.38138068 -0.12465987  0.11492862
## cyl   0.22432550 -0.15893251  0.81032177  0.16266295
## disp -0.01967516 -0.18233095 -0.06416707 -0.66190812
## hp    0.35394225  0.69620751 -0.16573993  0.25177306
## drat -0.01536794  0.04767957  0.13505066  0.03809096
## wt   -0.08377237 -0.42777608 -0.19839375  0.56918844
## qsec  0.25752940  0.27622581  0.35613350 -0.16873731
## gear -0.32298239 -0.08555707  0.31636479  0.04719694
## carb  0.35706914 -0.20604210 -0.10832772 -0.32045892
```

Output dari pemodelan “PCA” yaitu standar deviasi dan korelasi antara kesembilan komponen (atau variabel) di dalam tabel “df.mtcars”.

```
# Menampilkan summary model PCA
summary(pca)

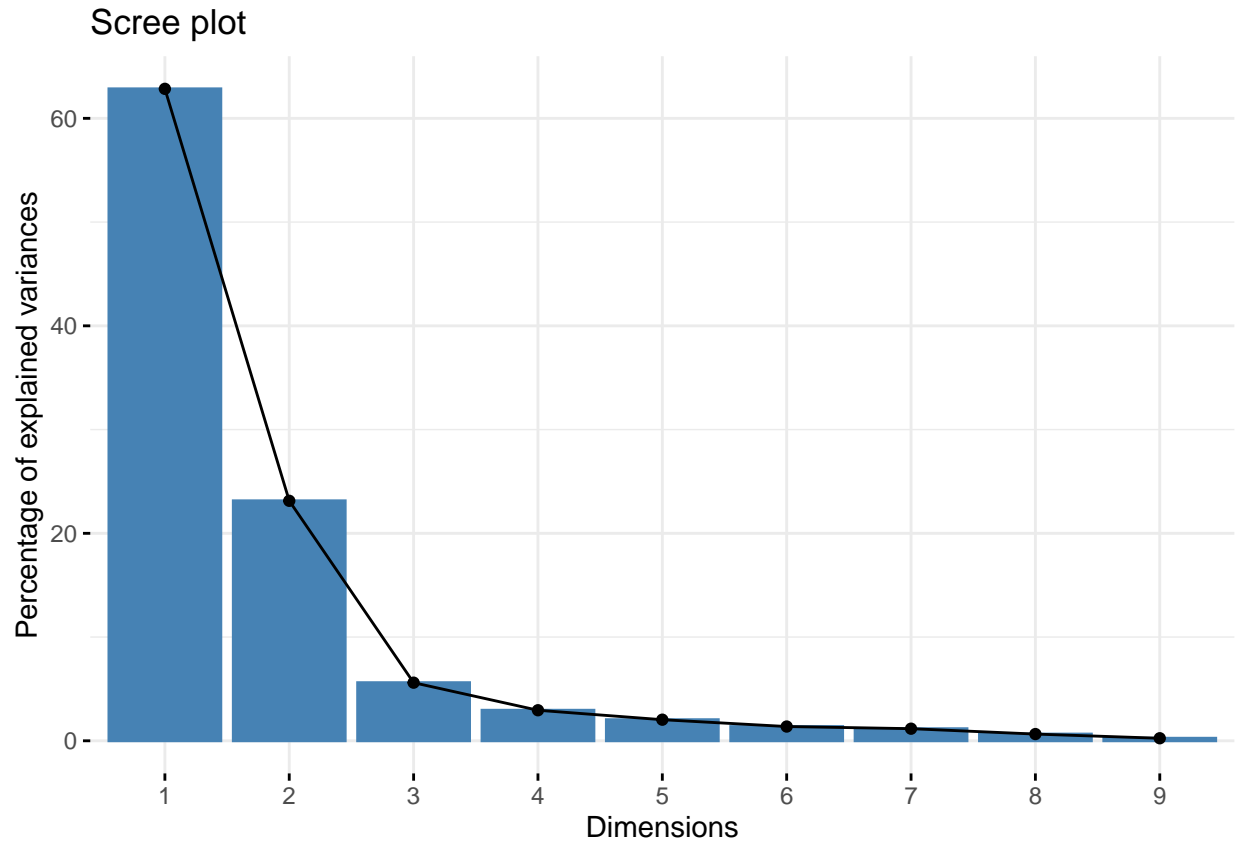
## Importance of components:
##          PC1      PC2      PC3      PC4      PC5      PC6
## Standard deviation      2.3782 1.4429 0.71008 0.51481 0.42797 0.35184
## Proportion of Variance  0.6284 0.2313 0.05602 0.02945 0.02035 0.01375
## Cumulative Proportion  0.6284 0.8598 0.91581 0.94525 0.96560 0.97936
##          PC7      PC8      PC9
## Standard deviation      0.32413 0.2419 0.14896
## Proportion of Variance  0.01167 0.0065 0.00247
## Cumulative Proportion  0.99103 0.9975 1.00000
```

Rangkuman pemodelan PCA berisi informasi tentang standar deviasi, proporsi varians, dan proporsi kumulatif dari kesembilan komponen (atau variabel) di dalam tabel “df.mtcars”.

Melalui informasi proporsi varians, dapat diketahui bahwasannya principal component 1 mampu mewakili 62.8% dari total varians sedangkan principal component 2 mampu mewakili sebesar 23.1% dari total varians.

Melalui informasi proporsi kumulatif, dapat diketahui bahwasannya dengan menggunakan 1 principal component, kita sudah dapat mewakili sebesar 62.8% dari total varians, sedangkan ketika menggunakan principal component 1 dan 2, kita sudah dapat mewakili sebesar 85.9% dari total varians.

```
# Visualisasi scree plot PCA  
fviz_eig(pca)
```



Plot ini menampilkan informasi yang sama dengan informasi proporsi varians.