Week 6

Bùi Khánh Duy

2023-04-12

```
print(utils::getSrcFilename(function(){}, full.names = TRUE))
```

```
## [1] "<text>"
```

Đọc dữ liệu:

```
dl = read.csv("table1.csv")
x = dl[, c("X1", "X2", "X3")]
s = cov(x)

s
```

```
##          X1        X2       X3
## X1 6.631579  6.368421 3.000000
## X2 6.368421 12.526316 3.578947
## X3 3.000000  3.578947 5.944737
```

```
eicov = eigen(s)
eicov
```

```
## eigen() decomposition
## $values
## [1] 18.331135  4.385884  2.385613
##
## $vectors
##            [,1]        [,2]       [,3]
## [1,] -0.5155034  0.06535209  0.8543918
## [2,] -0.7818109 -0.44401310 -0.4377489
## [3,] -0.3507533  0.89363386 -0.2799833
```

```
eicov$values
```

```
## [1] 18.331135  4.385884  2.385613
```

```
# lambda_1 = 18.331135, lambda_2 = 4.385884, lambda_3 = 2.385613

eicov$vectors # vecto rieng
```

```
##            [,1]        [,2]       [,3]
## [1,] -0.5155034  0.06535209  0.8543918
## [2,] -0.7818109 -0.44401310 -0.4377489
## [3,] -0.3507533  0.89363386 -0.2799833
```

=> $e_1 = (-0.5155034, -0.7818109, -0.3507533)^T$ Biểu diễn các thành phần chính theo biến ban đầu X1, X2, X3

Thành phần chính thứ nhất PC_1

$$PC_1 = e_1^T * X$$

=>

$$PC_1 = -0.5155034 * X1 - 0.7818109 * X2 - 0.3507533 * X3$$

=> PC_1 chứa nhiều thông tin X nhất

=> PC_2 chứa nhiều thông tin của X = (X1, X2, X3)^T mà PC_1 chưa thể hiện

$$\lambda_1 + \lambda_2 + \lambda_3 = tr(S) = \sigma_{11} + \sigma_{22} + \sigma_{33}$$

(tr(S) = vết của S)

Tỷ lệ $\frac{\lambda_i}{tr(S)}$ là phần thông tin của X được chứa trong PC_i và được gọi là ***tỷ lệ biến sai tổng cộng***

Thực tế, khi tỉ lệ $\frac{\lambda_m}{tr(S)}$ đủ nhỏ thì không cần đến các thành phần chính $PC_m, PC_{m+1}, ..., PC_k$ để biểu diễn X

Hỏi: Để thu được 95% thông tin về tập dữ liệu ban đầu thì cần bao nhiêu thành phần chính?

Tìm m nguyên dương nhỏ nhất sao cho $\frac{(\lambda_1 + ... + \lambda_m)}{tr(S)} > 0.95$

```
head(x)
```

```
##   X1 X2 X3
## 1 34 30 32
## 2 37 32 37
## 3 38 30 36
## 4 36 33 39
## 5 38 29 33
## 6 43 32 38
```

```
pca = princomp(x)
summary(pca)
```

```
## Importance of components:
##                            Comp.1    Comp.2     Comp.3
## Standard deviation     4.1730777 2.0412227 1.50543412
## Proportion of Variance 0.7302475 0.1747181 0.09503436
## Cumulative Proportion  0.7302475 0.9049656 1.00000000
```

Tỉ lệ biến sai tổng cộng của thành phần chính thứ nhất, thứ 2 và thứ 3 là 0.7302475, 0.1747181, 0.09503436

Hỏi: Để thu được 90% thông tin về tập dữ liệu ban đầu thì cần bao nhiêu thành phần chính?

Ta có: $\frac{(\lambda_1 + \lambda_2)}{tr(S)} = 0.9049656 > 90\%$ nên chỉ cần hai thành phần chính, ta sẽ thu được 90% thông tin về tập dữ liệu ban đầu.

```r
summary(pca, loadings=T)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3
## Standard deviation     4.1730777 2.0412227 1.50543412
## Proportion of Variance 0.7302475 0.1747181 0.09503436
## Cumulative Proportion  0.7302475 0.9049656 1.00000000
##
## Loadings:
##    Comp.1 Comp.2 Comp.3
## X1  0.516         0.854
## X2  0.782 -0.444 -0.438
## X3  0.351  0.894 -0.280
```

$$PC_1 = e_1^T * X$$

$$PC_1 = 0.516 * X1 + 0.782 * X2 + 0.351 * X3$$

```r
pcacov = princomp(covmat=s)
summary(pcacov) # Tỉ lệ biến sai tổng cộng
```

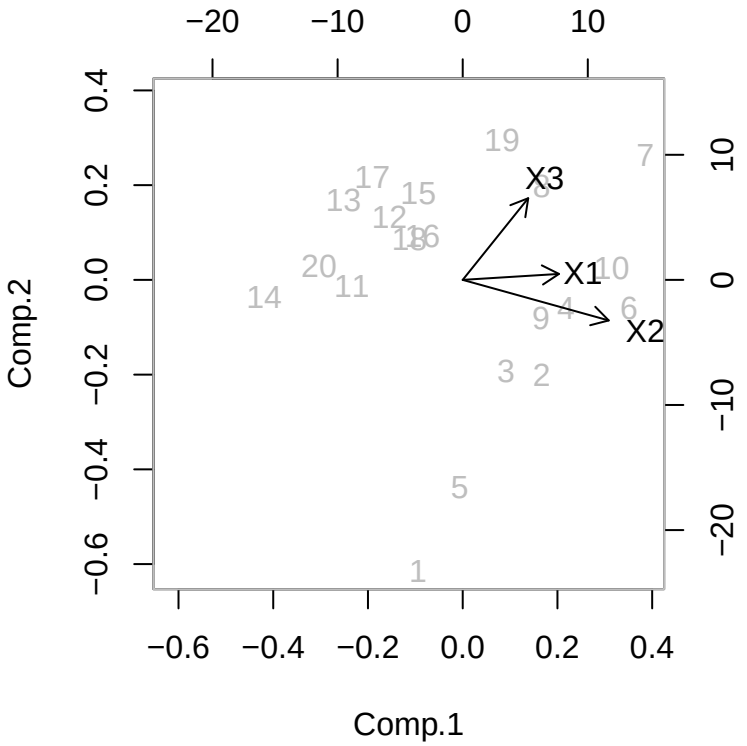
```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3
## Standard deviation     4.2814874 2.0942503 1.54454282
## Proportion of Variance 0.7302475 0.1747181 0.09503436
## Cumulative Proportion  0.7302475 0.9049656 1.00000000
```

```r
summary(pcacov, loadings=T)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3
## Standard deviation     4.2814874 2.0942503 1.54454282
## Proportion of Variance 0.7302475 0.1747181 0.09503436
## Cumulative Proportion  0.7302475 0.9049656 1.00000000
##
## Loadings:
##    Comp.1 Comp.2 Comp.3
## X1  0.516         0.854
## X2  0.782 -0.444 -0.438
## X3  0.351  0.894 -0.280
```

Vẽ đồ thị nào

```r
# install.packages("stats")
library(stats)
biplot(pca, col = c("grey", "black"))
```



Góc giữa các vectơ thể hiện độ tương quan giữa các biến, góc càng nhỏ thì hệ số tương quan càng lớn.