Giảm chiều dữ liệu sử dụng PCA

Le Nhat Tung

Contents

1 Giới thiệu về PCA

1.1 Khái niệm

Principal Component Analysis (PCA) là một kỹ thuật giảm chiều được sử dụng rộng rãi trong phân tích dữ liệu và học máy. PCA chuyển đổi một tập dữ liệu có nhiều biến (nhiều chiều) thành một tập dữ liệu với ít biến hơn (ít chiều hơn) nhưng vẫn giữ được thông tin quan trọng nhất.

1.2 Mục tiêu của PCA

PCA có các mục tiêu chính sau:

- **Giảm số lượng biến:** Chuyển đổi dữ liệu sang không gian mới với ít chiều hơn
- **Giữ lại thông tin quan trọng:** Các thành phần chính (principal components) mới sẽ giữ lại phần lớn sự biến thiên của dữ liệu gốc
- **Loại bỏ đa cộng tuyến:** Các thành phần chính không tương quan với nhau
- **Trực quan hóa dữ liệu:** Có thể hiển thị dữ liệu nhiều chiều trên không gian 2D hoặc 3D

1.3 Quy trình thực hiện PCA

1. Chuẩn hóa dữ liệu (nếu cần)
2. Tính ma trận hiệp phương sai (covariance matrix) hoặc ma trận tương quan (correlation matrix)
3. Tính các giá trị riêng (eigenvalues) và vector riêng (eigenvectors) của ma trận
4. Sắp xếp các vector riêng theo thứ tự giảm dần của giá trị riêng tương ứng
5. Chọn k vector riêng đầu tiên để tạo ma trận chiếu
6. Chiếu dữ liệu gốc lên không gian mới k chiều

2 Ứng dụng PCA với bộ dữ liệu mtcars

2.1 Tải thư viện cần thiết

```
library(tidyverse)   # Bộ thư viện chứa nhiều công cụ xử lý dữ liệu như dplyr, tidyr, ggplot2,... để th
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(ggplot2)     # Thư viện tạo đồ thị với cú pháp ngữ pháp đồ họa (grammar of graphics) giúp tạo b
#install.packages(GGally)
library(GGally)      # Mở rộng của ggplot2, cung cấp các hàm để tạo ma trận tương quan, biểu đồ cặp (pa
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(factoextra)  # Thư viện cho phân tích đa chiều, hỗ trợ phân tích thành phần chính (PCA) và phân
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

2.2 Hiểu về bộ dữ liệu mtcars

```r
# Xem cấu trúc bộ dữ liệu mtcars
str(mtcars)
```

```
## '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 ...
```

```r
# Hiển thị một số dòng đầu tiên
head(mtcars)
```

```
##                   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
```

```r
# Tóm tắt thống kê
summary(mtcars)
```

```
##       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
```

Bộ dữ liệu mtcars chứa thông tin về 32 mẫu xe với 11 biến mô tả các đặc điểm kỹ thuật:

- mpg: Miles per gallon (hiệu suất tiêu thụ nhiên liệu)
- cyl: Số xi-lanh
- disp: Dung tích xi-lanh
- hp: Mã lực
- drat: Tỷ số truyền động sau
- wt: Trọng lượng (1000 lbs)
- qsec: Thời gian chạy 1/4 dặm
- vs: Kiểu động cơ (0 = chữ V, 1 = thẳng hàng)
- am: Kiểu hộp số (0 = tự động, 1 = số sàn)
- gear: Số lượng số
- carb: Số lượng bộ chế hòa khí

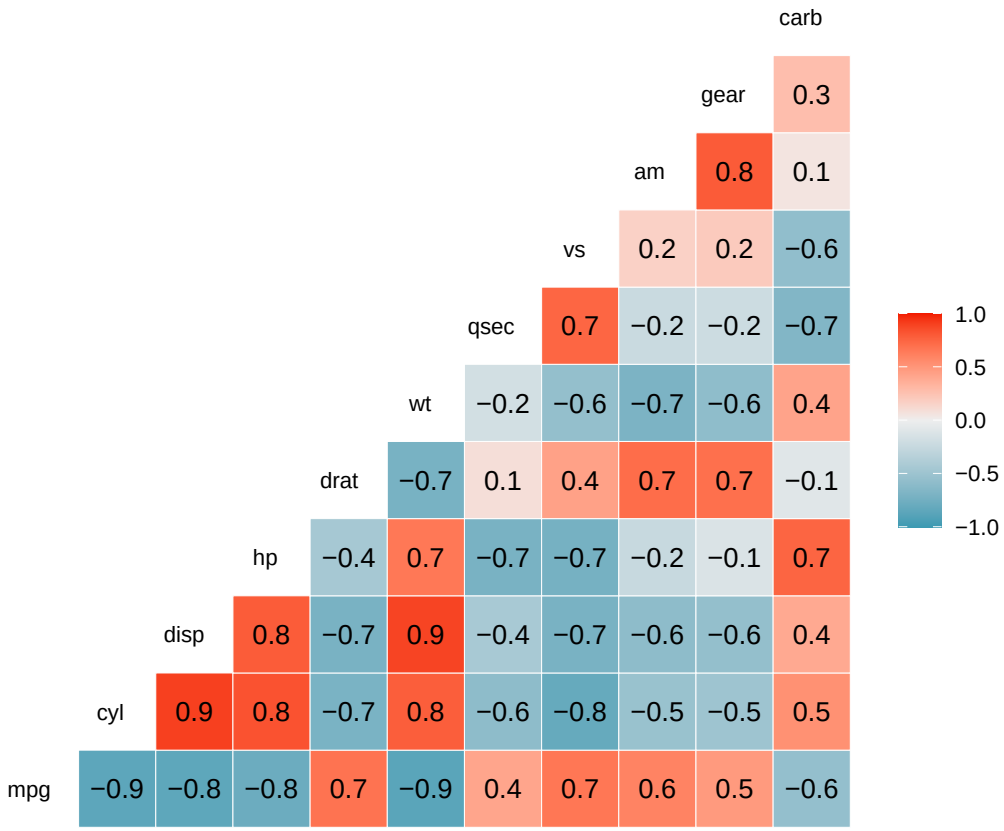
2.3 Phân tích tương quan

Trước khi thực hiện PCA, chúng ta hãy xem các biến có tương quan như thế nào:

```r
# Tạo ma trận tương quan
cor_matrix <- cor(mtcars)
round(cor_matrix, 2)
```

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
## vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
## am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

```r
# Visualize correlation matrix
ggcorr(mtcars,
       method = c("everything", "pearson"),
       label = TRUE,
       hjust = 0.75,
       size = 3,
       layout.exp = 2)
```



Qua ma trận tương quan, chúng ta thấy nhiều biến có tương quan mạnh với nhau. Ví dụ:

- cyl, disp, hp và wt có tương quan dương mạnh với nhau
- Các biến trên có tương quan âm mạnh với mpg

Điều này chỉ ra rằng dữ liệu có thể có đa cộng tuyến và phù hợp để áp dụng PCA.

2.4 Thực hiện PCA

```r
# Chuẩn hóa dữ liệu (center và scale)
mtcars_scaled <- scale(mtcars)

# Thực hiện PCA
```

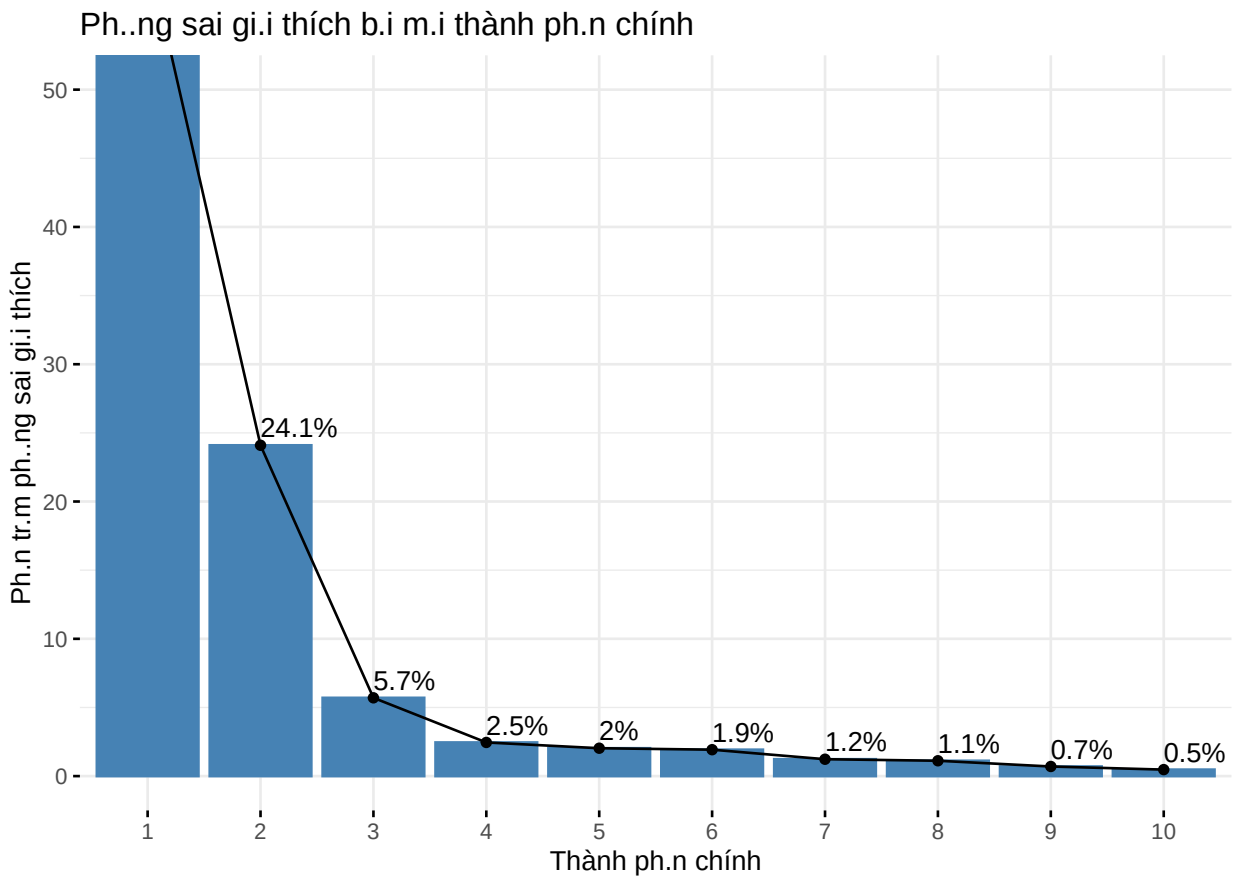
```r
?prcomp
pca_result <- prcomp(mtcars_scaled, center = TRUE, scale.=TRUE)

# Xem kết quả
summary(pca_result)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6    PC7
## Standard deviation     2.5707 1.6280 0.79196 0.51923 0.47271 0.46000 0.3678
## Proportion of Variance 0.6008 0.2409 0.05702 0.02451 0.02031 0.01924 0.0123
## Cumulative Proportion  0.6008 0.8417 0.89873 0.92324 0.94356 0.96279 0.9751
##                           PC8    PC9    PC10   PC11
## Standard deviation     0.35057 0.2776 0.22811 0.1485
## Proportion of Variance 0.01117 0.0070 0.00473 0.0020
## Cumulative Proportion  0.98626 0.9933 0.99800 1.0000
```

2.5 Phân tích các thành phần chính

Phương sai giải thích bởi mỗi thành phần

```r
fviz_eig(pca_result,
         addlabels = TRUE,
         ylim = c(0, 50),
         main = "Phương sai giải thích bởi mỗi thành phần chính",
         xlab = "Thành phần chính",
         ylab = "Phần trăm phương sai giải thích")
```



Ý nghĩa thực tế của biểu đồ này:

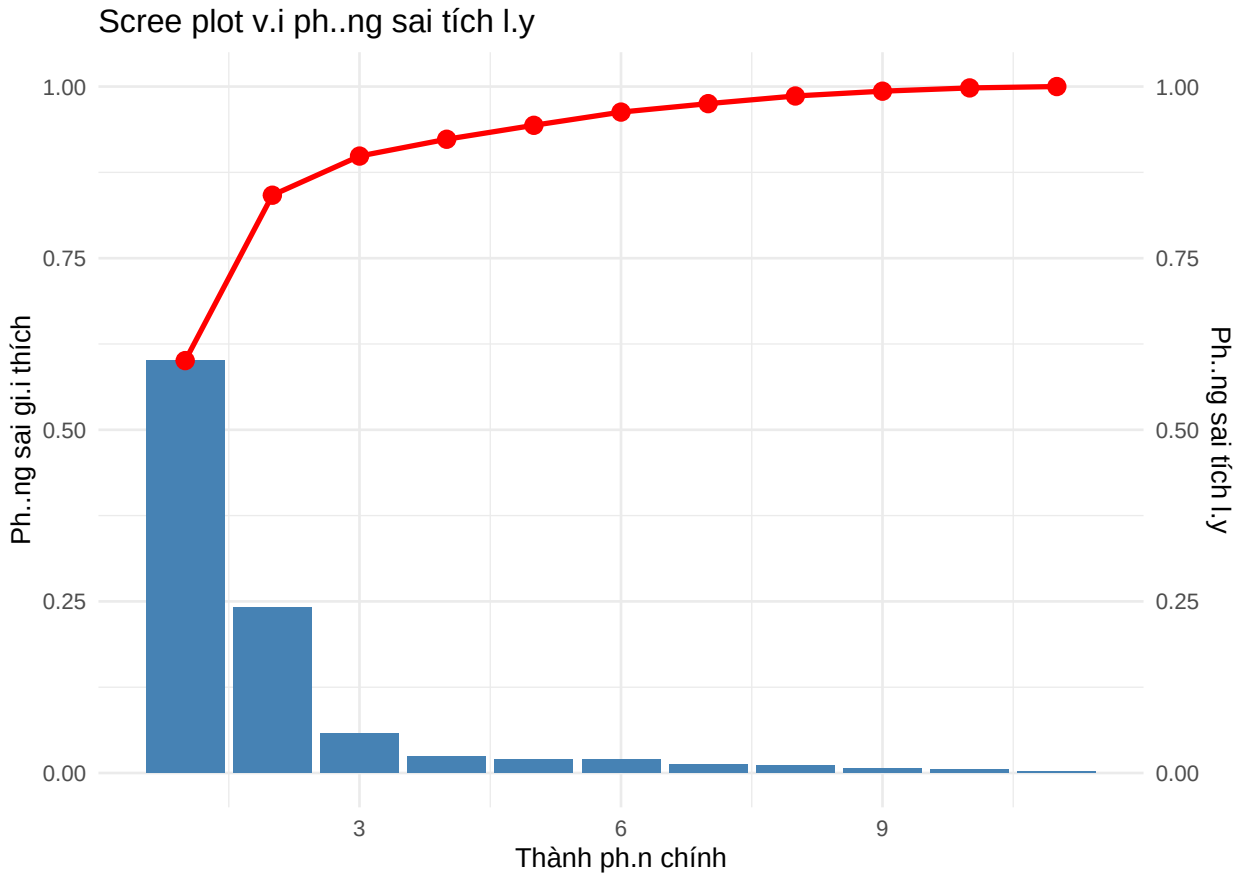
- Hai thành phần chính đầu tiên (PC1 và PC2) đã giải thích khoảng 84% phương sai (60% + 24.1%)
- Điều này có nghĩa là bạn có thể giảm số chiều dữ liệu từ 10 xuống còn 2 mà vẫn giữ được phần lớn thông tin
- Theo quy tắc "elbow" (điểm gấp khúc), bạn có thể chọn giữ lại 2-3 thành phần chính đầu tiên vì sau PC3, lượng phương sai giải thích giảm đáng kể

Scree plot kết hợp với phương sai tích lũy

```
# Tính phương sai giải thích
var_explained <- pca_result$sdev^2 / sum(pca_result$sdev^2)
cum_var_explained <- cumsum(var_explained)

# Tạo dataframe cho biểu đồ
var_df <- data.frame(
  PC = 1:length(var_explained),
  Variance = var_explained,
  Cumulative = cum_var_explained
)

# Vẽ biểu đồ
ggplot(var_df, aes(x = PC)) +
  geom_col(aes(y = Variance), fill = "steelblue") +
  geom_line(aes(y = Cumulative), color = "red", size = 1) +
  geom_point(aes(y = Cumulative), color = "red", size = 3) +
```

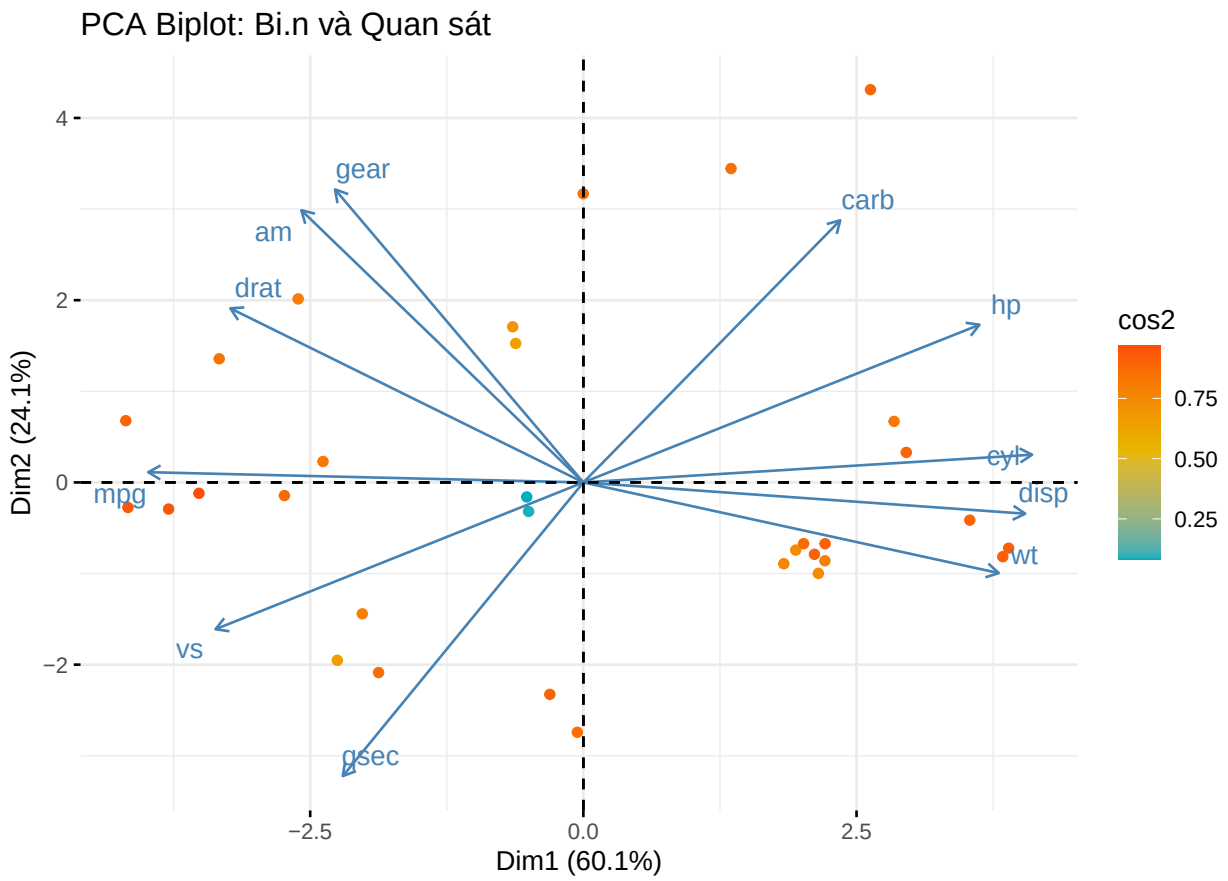
```
  scale_y_continuous(name = "Phương sai giải thích",
                     sec.axis = sec_axis(~., name = "Phương sai tích lũy")) +
  labs(title = "Scree plot với phương sai tích lũy",
       x = "Thành phần chính") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



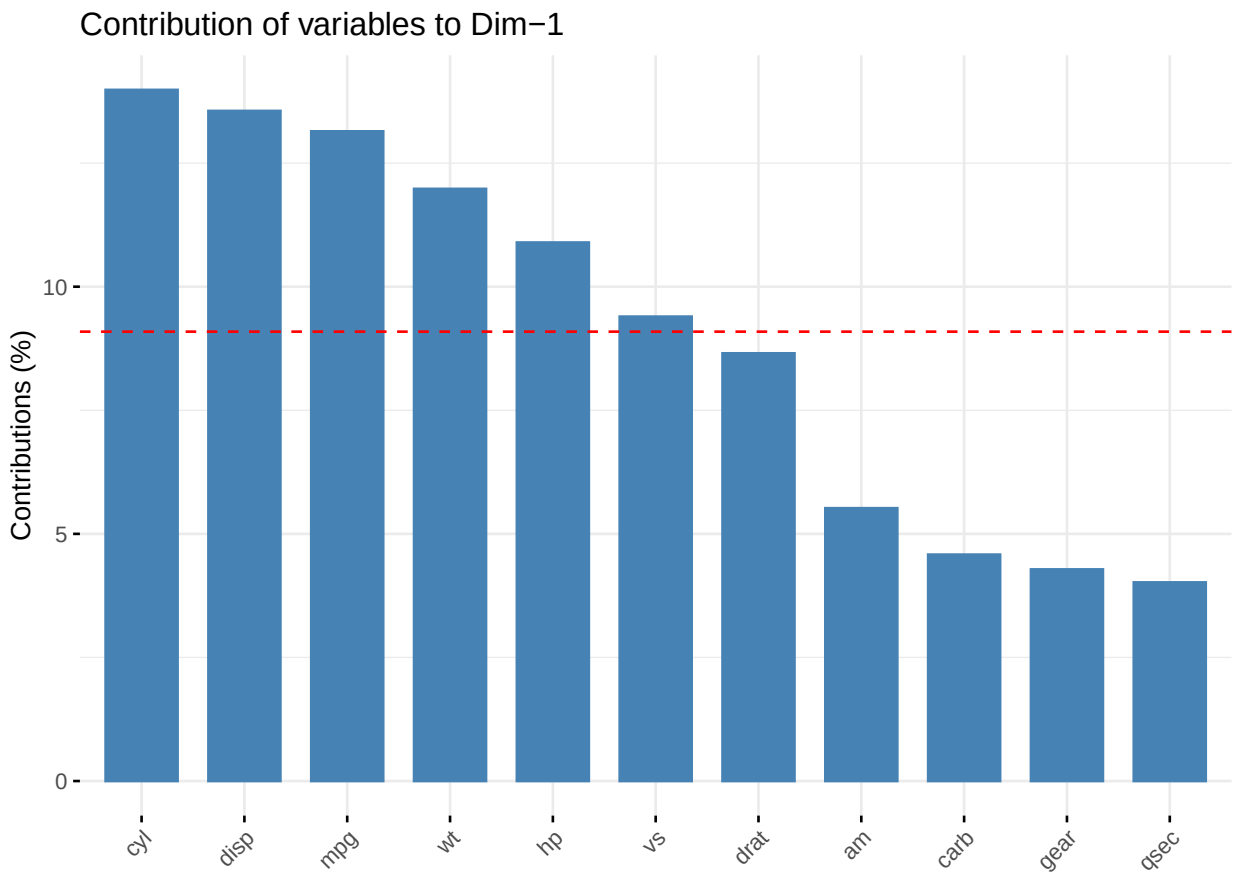
Đường màu đỏ thể hiện phương sai tích lũy - tổng phương sai được giải thích khi kết hợp các thành phần chính phía trước.

```
# Biplot
fviz_pca_biplot(pca_result,
                label = "var",
                col.ind = "cos2",
                gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                repel = TRUE,
                title = "PCA Biplot: Biến và Quan sát")
```

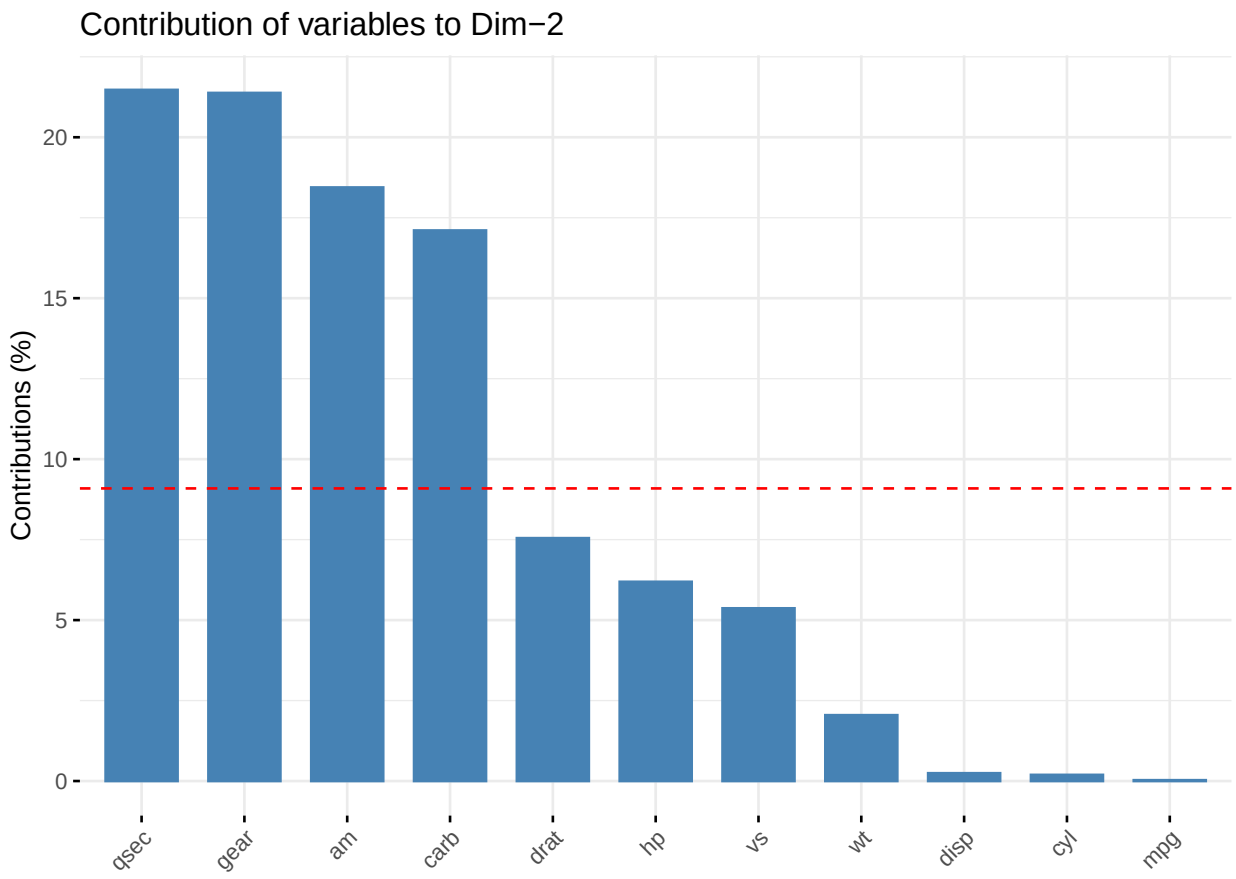


Phân tích đóng góp của các biến vào thành phần chính

```
# Contribution of variables to PC1
fviz_contrib(pca_result, choice = "var", axes = 1, top = 11)
```

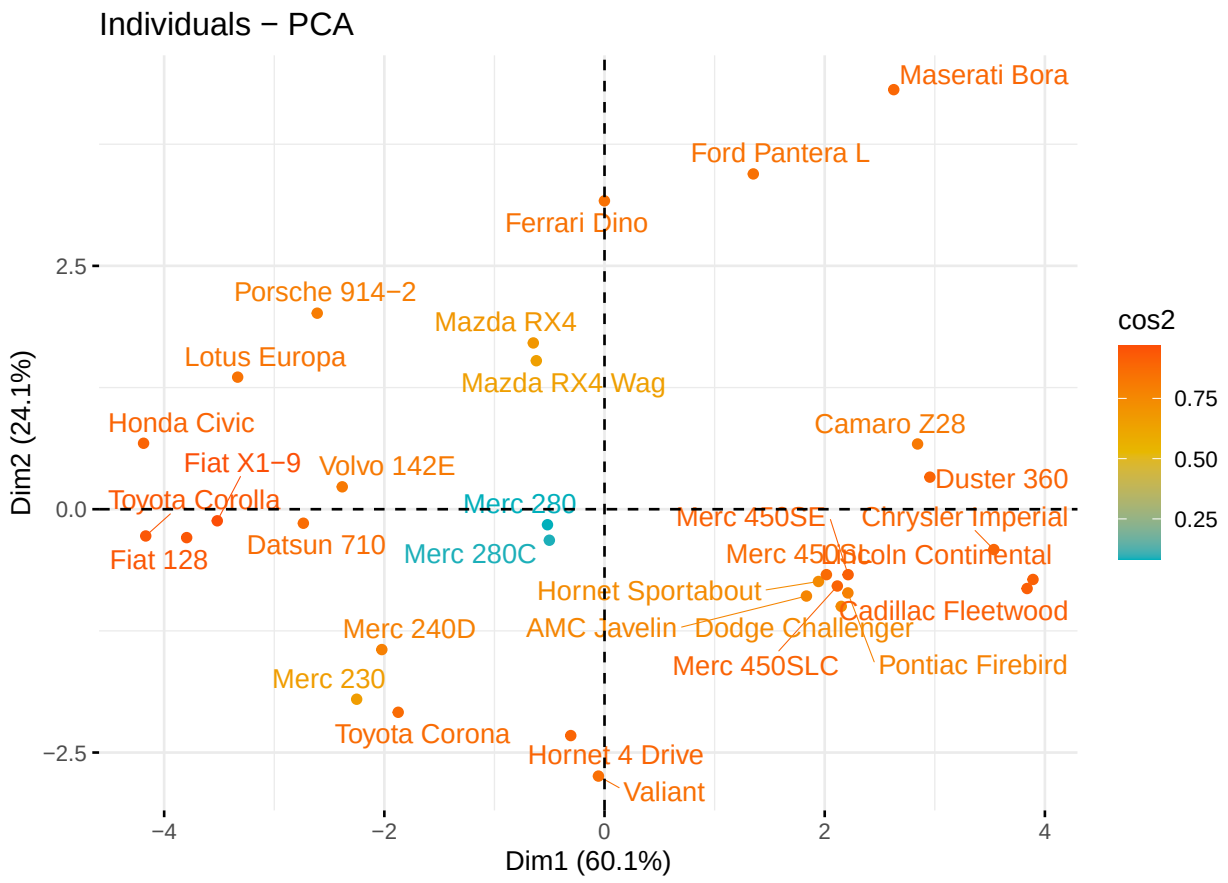



```r
# Contribution of variables to PC2
fviz_contrib(pca_result, choice = "var", axes = 2, top = 11)
```



Trực quan hóa quan sát trên không gian mới

```
# Vẽ các quan sát trên không gian PC1 và PC2
# Thêm tên của từng xe
fviz_pca_ind(pca_result,
             col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```

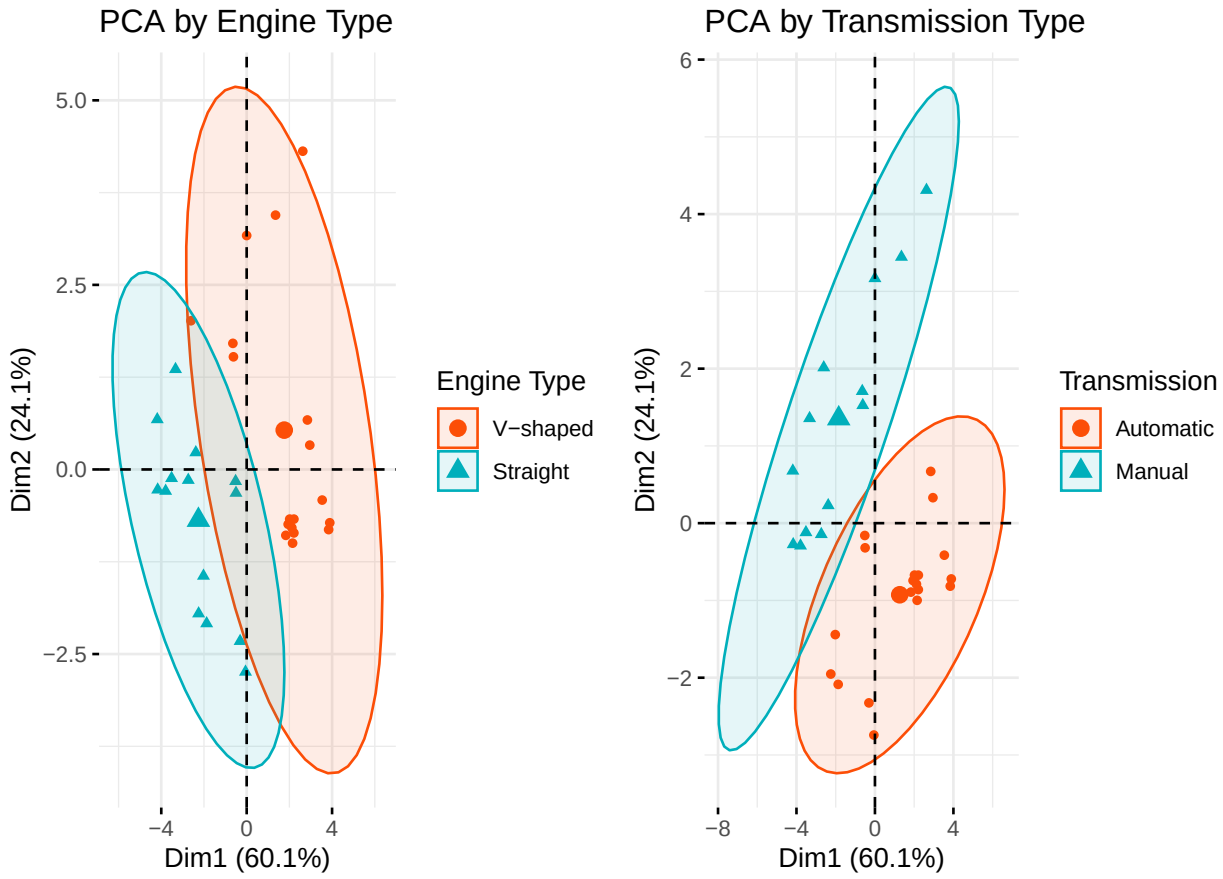


2.6 Sử dụng PCA cho phân loại

Chúng ta sẽ thêm thông tin về loại động cơ (vs) và kiểu hộp số (am) để xem liệu PCA có thể phân tách các nhóm xe này hay không:

```
# Tạo dataframe với thông tin về loại động cơ và hộp số
mtcars_info <- mtcars %>%
  mutate(vs_factor = factor(vs, labels = c("V-shaped", "Straight")),
         am_factor = factor(am, labels = c("Automatic", "Manual")))

# Trực quan hóa theo kiểu động cơ
p1 <- fviz_pca_ind(pca_result,
             geom.ind = "point",
             col.ind = mtcars_info$vs_factor,
             palette = c("#FC4E07", "#00AFBB"),
             addEllipses = TRUE,
             legend.title = "Engine Type",
             title = "PCA by Engine Type")

# Trực quan hóa theo kiểu hộp số
p2 <- fviz_pca_ind(pca_result,
             geom.ind = "point",
             col.ind = mtcars_info$am_factor,
             palette = c("#FC4E07", "#00AFBB"),
```

```
            addEllipses = TRUE,
            legend.title = "Transmission",
            title = "PCA by Transmission Type")

# Hiển thị cả hai biểu đồ
gridExtra::grid.arrange(p1, p2, ncol = 2)
```



3 Giảm chiều dữ liệu

3.1 Lựa chọn số lượng thành phần

Có một số tiêu chí:

- Kaiser's rule: Giữ các thành phần có eigenvalue > 1
- Elbow method: Chọn điểm gãy trên scree plot
- Phương sai tích lũy: Giữ đủ số thành phần để giải thích ít nhất 80-90% phương sai

3.2 Tạo dữ liệu đã giảm chiều

4 Lấy 2 thành phần chính đầu tiên

```
pca_data_2d <- as.data.frame(pca_result$x[, 1:2])
head(pca_data_2d)
```

```
##                           PC1        PC2
## Mazda RX4         -0.64686274  1.7081142
## Mazda RX4 Wag     -0.61948315  1.5256219
## Datsun 710        -2.73562427 -0.1441501
## Hornet 4 Drive    -0.30686063 -2.3258038
## Hornet Sportabout  1.94339268 -0.7425211
## Valiant           -0.05525342 -2.7421229
```

5 Lấy 4 thành phần chính đầu tiên

```
pca_data_4d <- as.data.frame(pca_result$x[, 1:4])
head(pca_data_4d)
```

```
##                           PC1        PC2        PC3         PC4
## Mazda RX4         -0.64686274  1.7081142 -0.5917309  0.11370221
## Mazda RX4 Wag     -0.61948315  1.5256219 -0.3763013  0.19912121
## Datsun 710        -2.73562427 -0.1441501 -0.2374391 -0.24521545
## Hornet 4 Drive    -0.30686063 -2.3258038 -0.1336213 -0.50380035
## Hornet Sportabout  1.94339268 -0.7425211 -1.1165366  0.07446196
## Valiant           -0.05525342 -2.7421229  0.1612456 -0.97516743
```

5.1 Kết hợp với mô hình khác

Thêm mpg vào dữ liệu đã giảm chiều

```
pca_data_with_mpg <- cbind(pca_data_4d, mpg = mtcars$mpg)


# Xây dựng mô hình hồi quy
lm_model <- lm(mpg ~ ., data = pca_data_with_mpg)
summary(lm_model)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = pca_data_with_mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6102 -1.1754 -0.1933  1.0695  4.1308
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.09062    0.35862  56.022  < 2e-16 ***
## PC1         -2.18495    0.14174 -15.416 6.63e-15 ***
## PC2          0.09718    0.22381   0.434  0.66758
## PC3         -1.36055    0.46008  -2.957  0.00638 **
## PC4         -0.13585    0.70174  -0.194  0.84795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.029 on 27 degrees of freedom
## Multiple R-squared:  0.9013, Adjusted R-squared:  0.8867
## F-statistic: 61.65 on 4 and 27 DF,  p-value: 3.48e-13
```