Principal Component Analysis

```
#Data and Plot
library(mvtnorm)
library(ggplot2)
set.seed(9)
sig <- matrix(c(9, 5, 5, 4), nrow = 2)
x <- rmvnorm(200, sigma = sig, mean = c(20,40))
head(x)
```

```
##          [,1]     [,2]
## [1,] 16.94723 37.78654
## [2,] 19.29332 39.38189
## [3,] 19.87350 38.53251
## [4,] 23.29282 41.31503
## [5,] 18.90085 39.12075
## [6,] 23.02213 40.66738
```

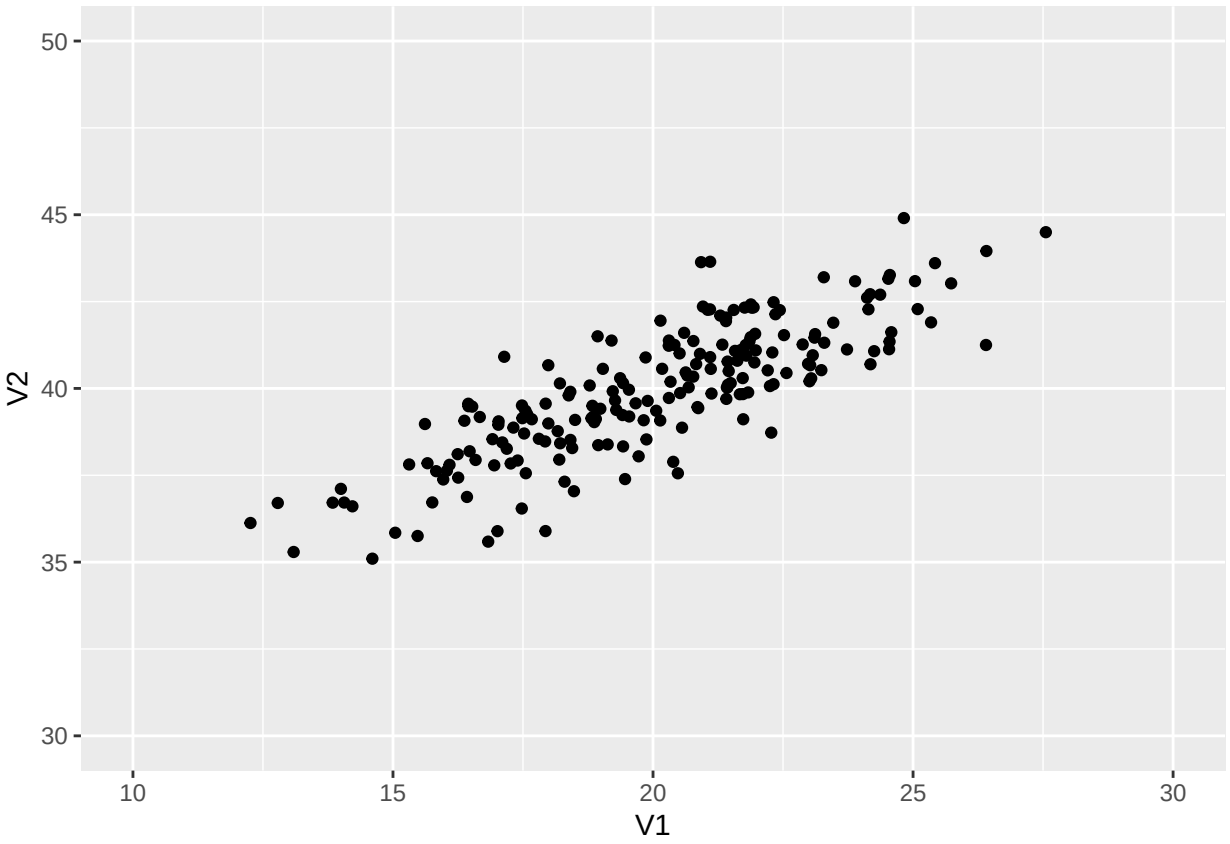
```
cor(x)
```

```
##           [,1]      [,2]
## [1,] 1.0000000 0.8238307
## [2,] 0.8238307 1.0000000
```

```
cov(x)
```

```
##          [,1]     [,2]
## [1,] 8.486099 4.613268
## [2,] 4.613268 3.695160
```

```
p <- ggplot(as.data.frame(x), aes(x = V1, y = V2)) + geom_point() +
    xlim(10, 30) + ylim(30,50)
print(p)
```




```
#Centering data & Finding Column Means
means = colMeans(x)
xc <- x - cbind(rep(1, 200))%*%means
colMeans(xc)
```

```
## [1] 4.263256e-16 7.815970e-16
```

```
#Checking the result
print(1/(200-1)*t(xc)%*%xc)
```

```
##          [,1]     [,2]
## [1,] 8.486099 4.613268
## [2,] 4.613268 3.695160
```

```
print(cov(x))
```

```
##          [,1]     [,2]
## [1,] 8.486099 4.613268
## [2,] 4.613268 3.695160
```

```
#Eigenvectors and Eigenvalues
e <- eigen(t(xc)%*%xc/(200-1))
Q <- e$vectors
print(e$values)
```

```
## [1] 11.2887555  0.8925038
```

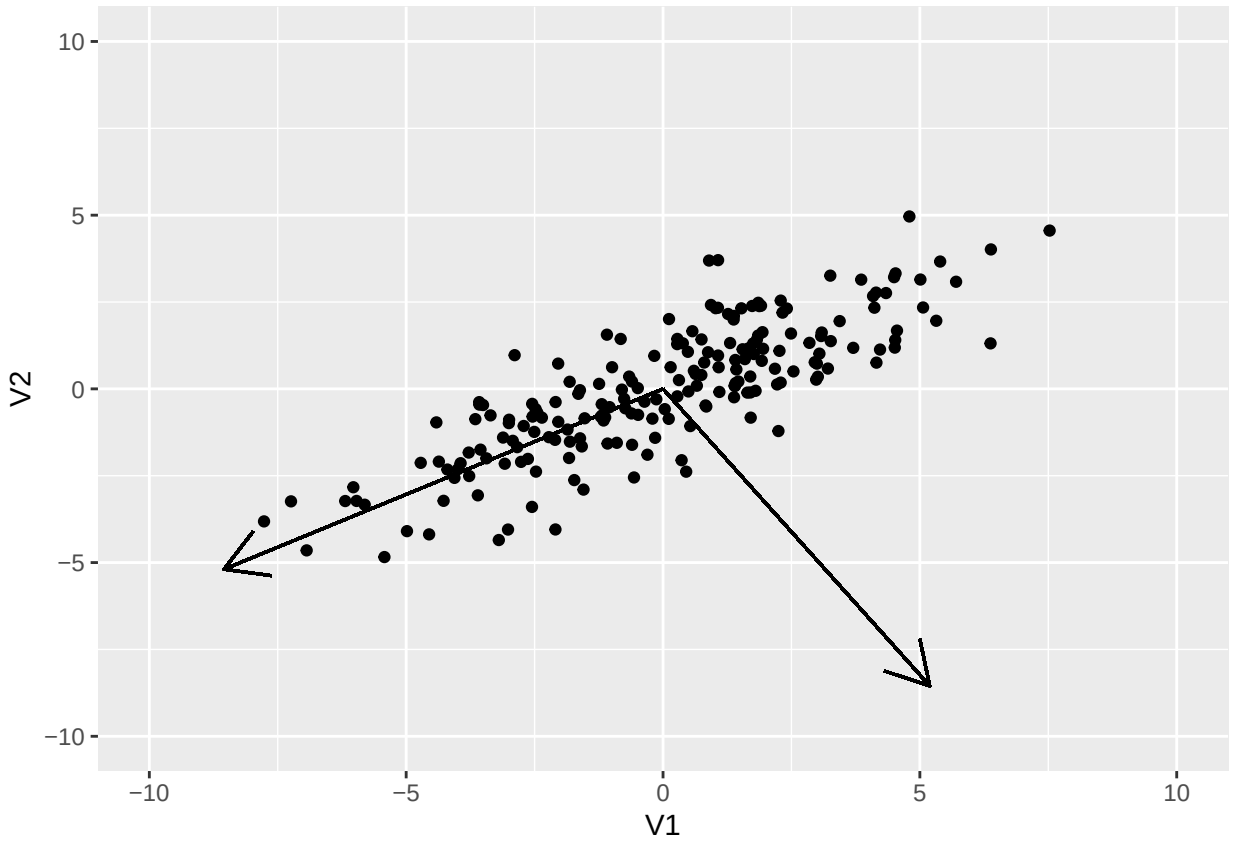
```
print(Q)
```

```
##            [,1]       [,2]
## [1,] -0.8546442  0.5192142
## [2,] -0.5192142 -0.8546442
```

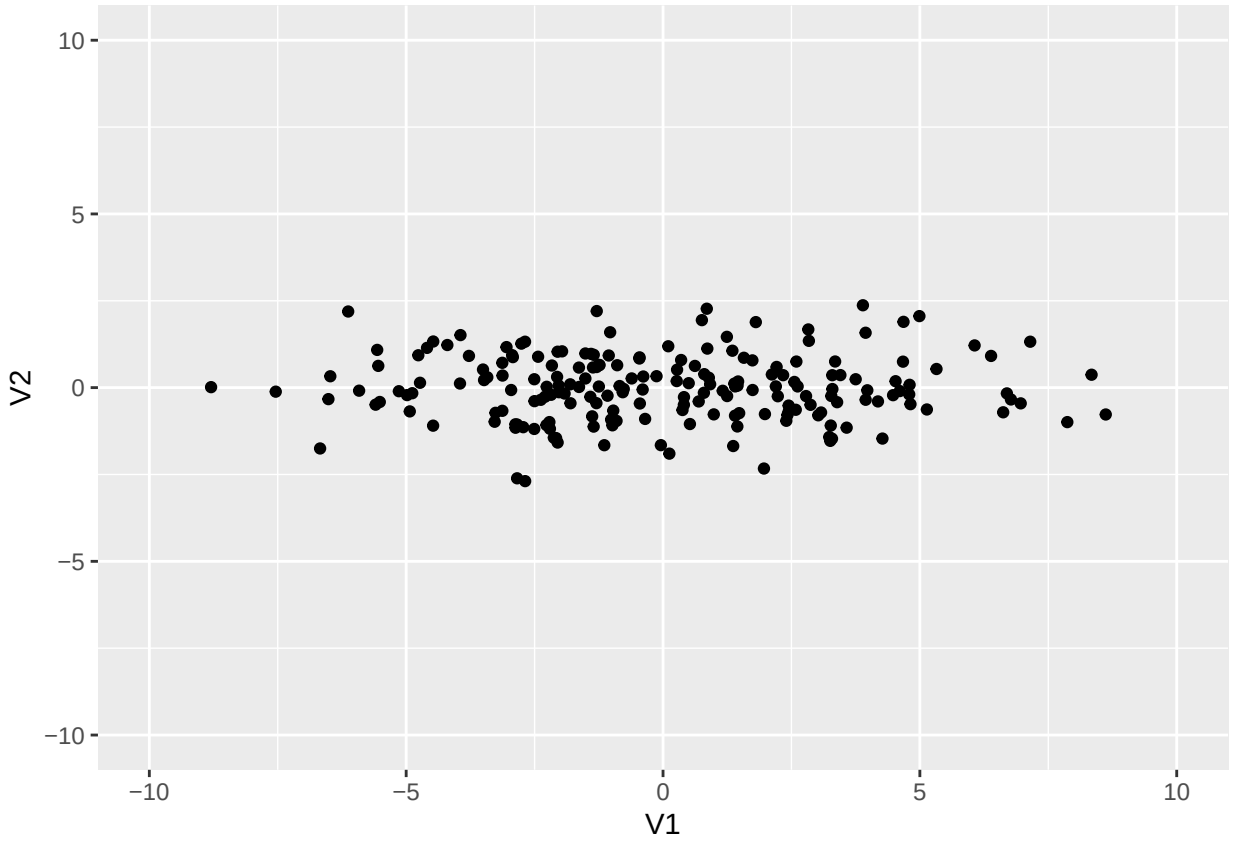
```
print(Q%*%diag(e$values)%*%t(Q))
```

```
##          [,1]     [,2]
## [1,] 8.486099 4.613268
## [2,] 4.613268 3.695160
```

```
#Principal Direction
p <- ggplot(as.data.frame(xc), aes(x = V1, y = V2)) + geom_point() +
    xlim(-10, 10) + ylim(-10,10) +
    geom_segment(x = 0, y = 0, xend = e$vector[1,1]*10,
                 yend = e$vector[2,1]*10, arrow = arrow()) +
    geom_segment(x = 0, y = 0, xend = e$vector[1,2]*10,
                 yend = e$vector[2,2]*10, arrow = arrow())
print(p)
```



```r
#Origial data multiplied by eigenvectors
pc <- xc %*% Q
p <- ggplot(as.data.frame(pc), aes(x = V1, y = V2)) + geom_point() +
    xlim(-10, 10) + ylim(-10,10)
print(p)
```




```r
#Covariance & Correlation of Principal Component
print(cov(pc))
```

```
##               [,1]          [,2]
## [1,]  1.128876e+01 -4.530886e-16
## [2,] -4.530886e-16  8.925038e-01
```

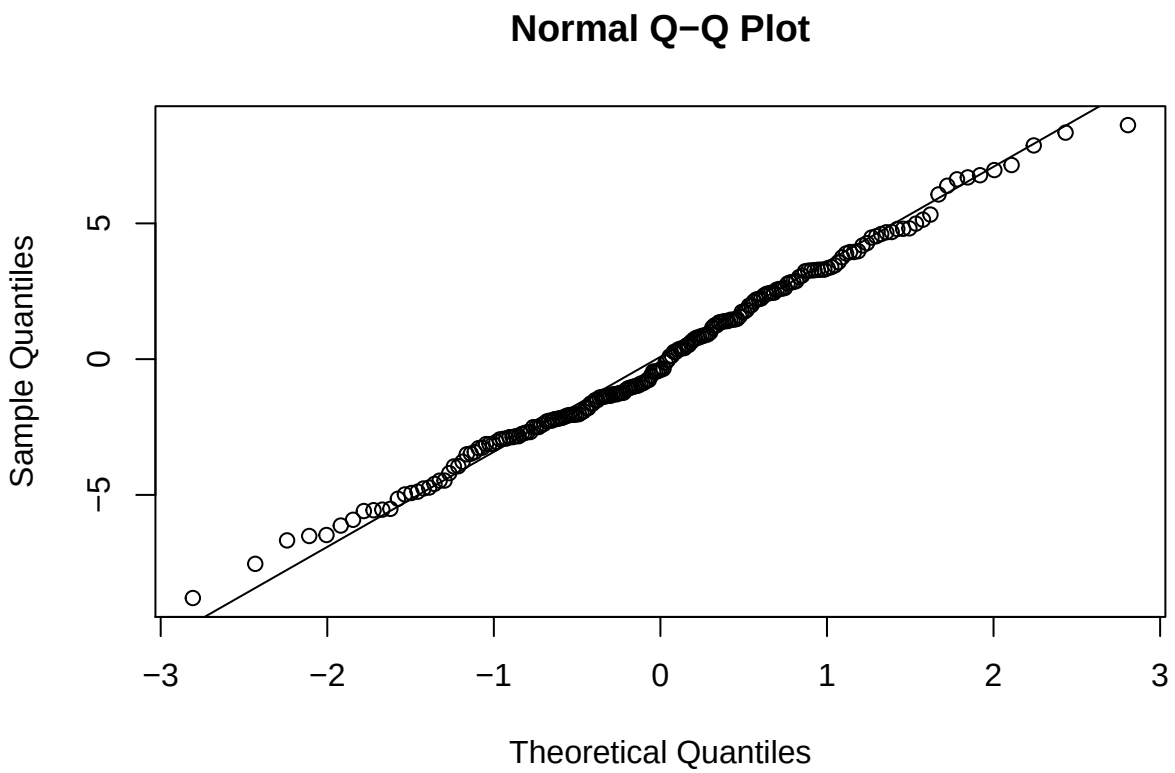
```r
print(cor(pc))
```

```
##               [,1]          [,2]
## [1,]  1.000000e+00 -1.427431e-16
## [2,] -1.427431e-16  1.000000e+00
```

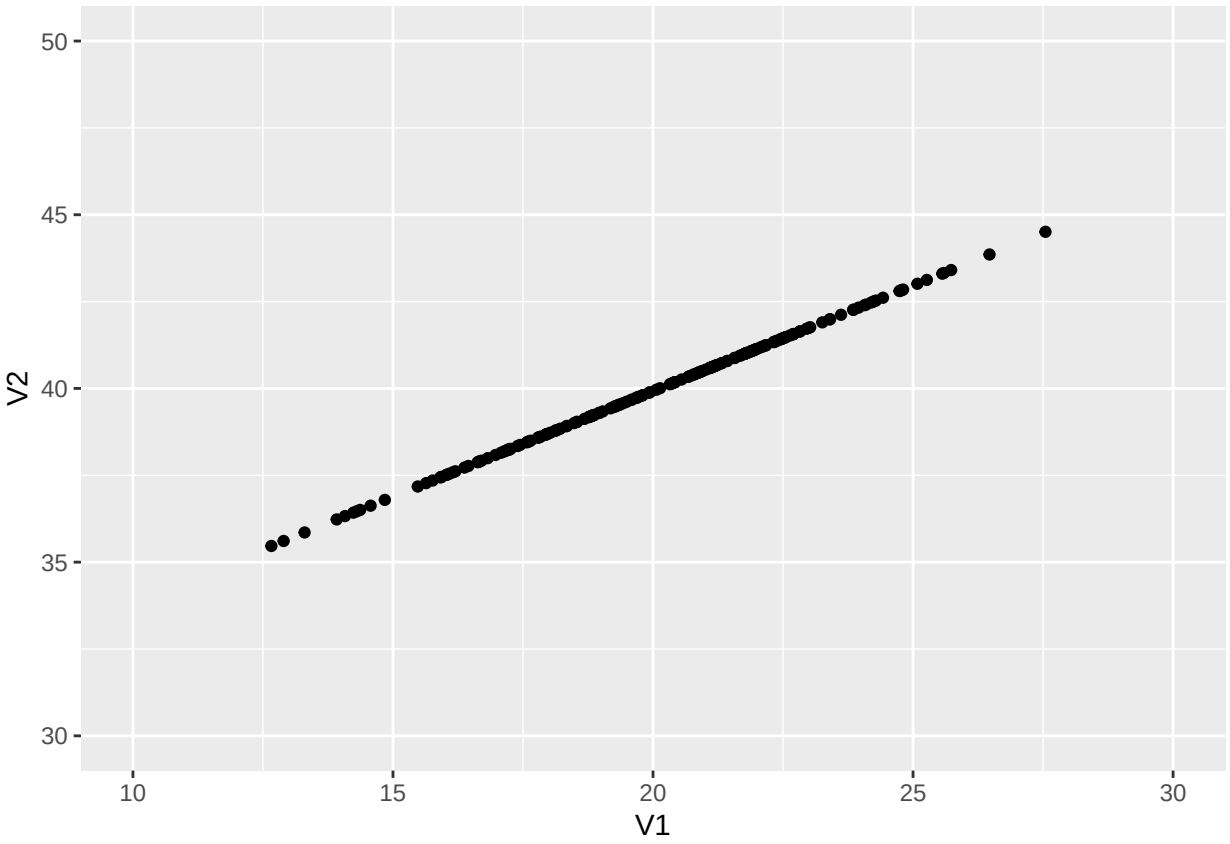
```r
#First Principal Component
pc1 <- pc[,1]
c(mean(pc1), sd(pc1))
```

```
## [1] -7.778717e-16  3.359874e+00
```

```r
qqnorm(pc1); qqline(pc1)
```




```r
#Projection of data onto the first principal component direction
xrec <- pc1 %*% t(Q[,1])
xrec <- xrec + cbind(rep(1,200))%*%means
p <- ggplot(as.data.frame(xrec), aes(x = V1, y = V2)) + geom_point()  + xlim(10, 30) + ylim(30,50)
print(p)
```


50
45
40
35
30
V2
10
15
20
25
30
V1