HW 3

임효진

April 10, 2021

1

(a)

```
cov=matrix(c(5, 2, 2, 2), nrow = 2)
eigen(cov)
```

```
## eigen() decomposition
## $values
## [1] 6 1
##
## $vectors
##            [,1]       [,2]
## [1,] -0.8944272  0.4472136
## [2,] -0.4472136 -0.8944272
```

```
a1=eigen(cov)$vectors[, 1]
a2=eigen(cov)$vectors[, 2]

# The principle component of y1 is t(a1)%*%x and y2 is t(a2)%*%x.
```

(b)

```
6/(1+6)
```

```
## [1] 0.8571429
```

(c)

```
cor=cov2cor(cov)
eigen(cor)
```

```
## eigen() decomposition
## $values
## [1] 1.6324555 0.3675445
##
## $vectors
```

```
##            [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
```

```
b1=eigen(cor)$vectors[, 1]
b2=eigen(cor)$vectors[, 2]

# The principle component of y1 is t(b1)%*%x and y2 is t(b2)%*%x.

1.6324555/(1.6324555+0.3675445)
```

```
## [1] 0.8162278
```

(d)

The components obtained from the covariance matrix are different from that obtained from the correlation matrix.

2

(a)

```
s=matrix(0, 6, 6)
diag(s)=c(3266, 722, 179, 475, 10, 21)
s[lower.tri(s)]=c(1344, 732, 1176, 163, 238,
                      324, 537, 80, 118,
                      281, 39, 57,
                      64, 95,
                      14)
s[upper.tri(s, diag=F)]=t(s)[upper.tri(s, diag = F)]

eigen(s)
```

```
## eigen() decomposition
## $values
## [1] 4478.8520839  152.1779705   32.2516481    7.9138797    1.3761054
## [6]    0.4283124
##
## $vectors
##            [,1]        [,2]         [,3]        [,4]         [,5]         [,6]
## [1,] 0.84931097  0.46948805  0.228424561 -0.07769519  0.005584319  0.002517378
## [2,] 0.36853352 -0.84784176  0.364562832 -0.01522645  0.110171038 -0.008475517
## [3,] 0.19417675 -0.05461497 -0.293282177  0.93353621  0.031838195  0.028189257
## [4,] 0.31474469 -0.21276944 -0.852728689 -0.34749729  0.080728307 -0.035317209
## [5,] 0.04397973 -0.05892352 -0.008486377  0.02953198 -0.627965936 -0.774150134
## [6,] 0.06442733 -0.09497212 -0.041030351 -0.02480435 -0.765479726  0.631325161
```

```r
eigen(s)$values[1]/sum(diag(s))
```

```
## [1] 0.9584533
```

The data can be effectively summarized by one principal component.

(b)

```r
r=cov2cor(s)

eigen(r)
```

```
## eigen() decomposition
## $values
## [1] 5.65597128 0.17659709 0.05764131 0.05026065 0.03642490 0.02310479
##
## $vectors
##            [,1]       [,2]        [,3]        [,4]       [,5]        [,6]
## [1,] -0.4032700 -0.5609237 -0.16701057  0.60076955 -0.3657813 -0.01119623
## [2,] -0.4037062  0.5273396  0.54612023  0.45467380  0.1395127 -0.18572803
## [3,] -0.4095549 -0.3924816  0.11592987 -0.15182547  0.7418507  0.30230737
## [4,] -0.4118106 -0.2170011  0.31780324 -0.59342128 -0.3183269 -0.47837520
## [5,] -0.4085635  0.3312894 -0.74505982 -0.05770698  0.1986505 -0.35414259
## [6,] -0.4124891  0.3104732 -0.06563076 -0.23198664 -0.3944664  0.72092315
```

```r
eigen(r)$values[1]/sum(diag(r))
```

```
## [1] 0.9426619
```

The data can be effectively summarized by one principal component.

(c)

From both analysis we can conclude that the data can be summarized by the first principal component. Analyzing with the covariance matrix though, the principal component depend upon the arbitrary choice of the units while the correlation matrix assumes that the variables are equally important.

3

(a)

```r
stock=read.table("C:/Users/limj/Downloads/stock.dat")
```

The correlation matrix will be selected for this analysis to treat all variables equally.

(b)

```r
pca=prcomp(stock, center = T, scale = T)

pca$rotation
```

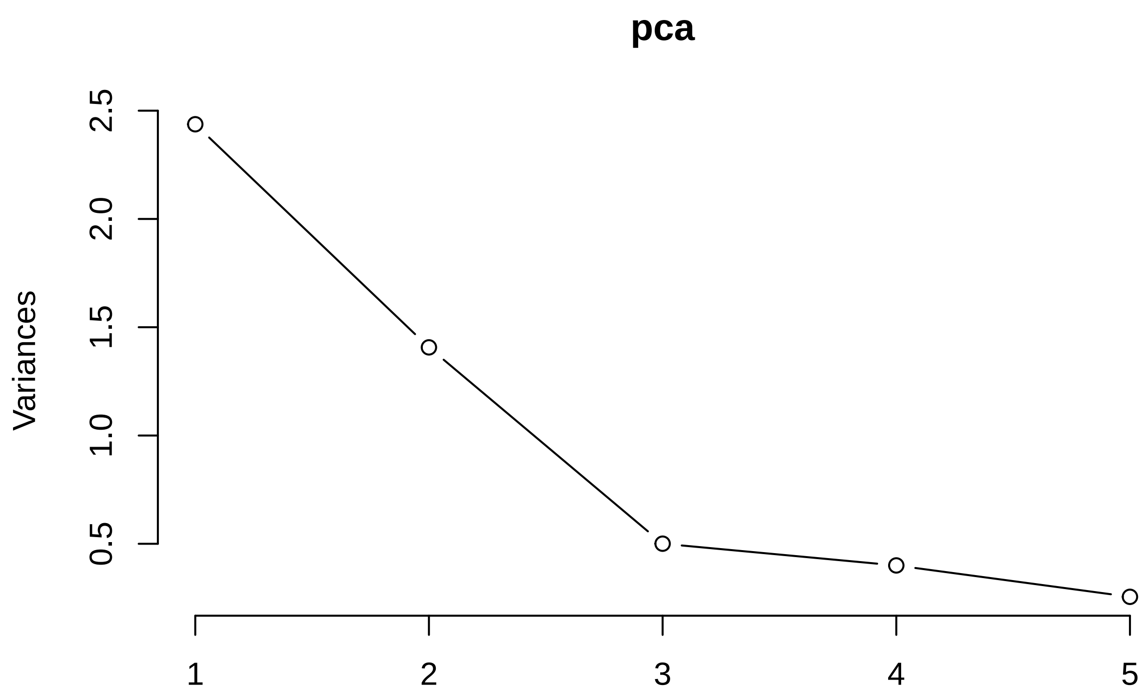
```
##           PC1        PC2         PC3        PC4         PC5
## V1 -0.4690832  0.3680070 -0.60431522  0.3630228  0.38412160
## V2 -0.5324055  0.2364624 -0.13610618 -0.6292079 -0.49618794
## V3 -0.4651633  0.3151795  0.77182810  0.2889658  0.07116948
## V4 -0.3873459 -0.5850373  0.09336192 -0.3812515  0.59466408
## V5 -0.3606821 -0.6058463 -0.10882629  0.4934145 -0.49755167
```

(c)

```r
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.5612 1.1862 0.7075 0.63248 0.50514
## Proportion of Variance 0.4874 0.2814 0.1001 0.08001 0.05103
## Cumulative Proportion  0.4874 0.7689 0.8690 0.94897 1.00000
```

```r
screeplot(pca, type="l")
```



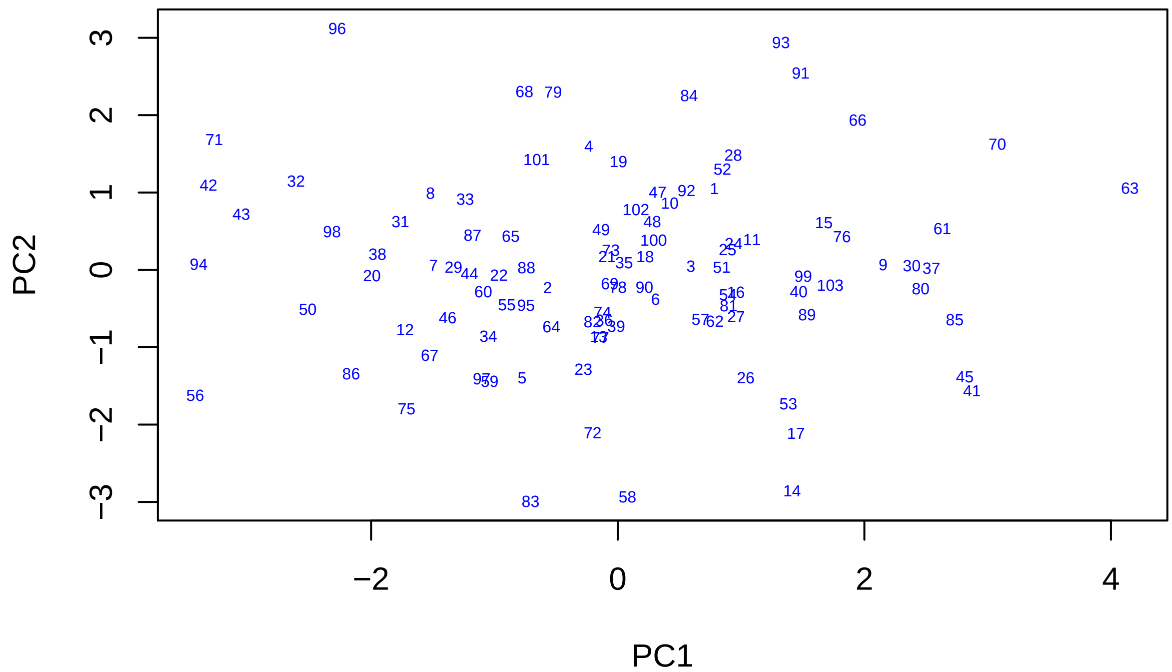
The proportion explained by the first two principal components is equal to 0.7689 which implies that the first two components convey about 77% of the infomation.

(d)

Yes.

(e)

```
plot(pca$x[,1], pca$x[,2], xlab="PC1",
     ylab="PC2", type="n", lwd=2)
text(pca$x[,1], pca$x[,2], labels=row.names(stock),
     col="blue",cex=.5,lwd=2)
```



4

(a)

```r
air=read.table("C:/Users/limj/Downloads/airpollution.dat", header = T)
```

Since the variables have different scales and values, it is efficient to use the correlation matrix to apply PCA.

(b)

```r
pca=prcomp(air, center = T, scale = T)

pca$rotation
```

```
##           PC1          PC2        PC3          PC4         PC5          PC6
## X1 -0.2368211  0.278445138 -0.6434744  0.172719491 -0.56053441  0.223579220
## X2  0.2055665 -0.526613869 -0.2244690  0.778136601  0.15613432  0.005700851
## X3  0.5510839 -0.006819502  0.1136089  0.005301798 -0.57342221  0.109538907
## X4  0.3776151  0.434674253  0.4070978  0.290503052  0.05669070  0.450234781
## X5  0.4980161  0.199767367 -0.1965567 -0.042428178 -0.05021430 -0.744968707
## X6  0.3245506 -0.566973655 -0.1598465 -0.507915905 -0.08024349  0.330583071
```

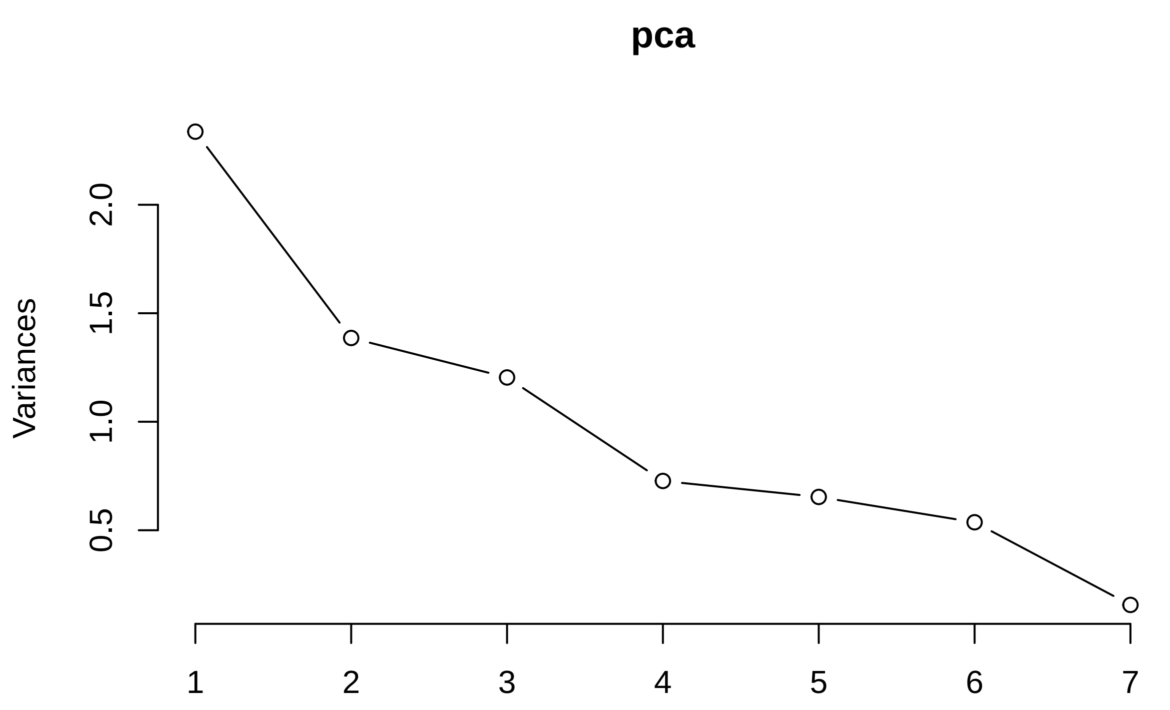
```
## X7  0.3194032  0.307882771 -0.5410484 -0.143082348  0.56607057  0.266469812
##             PC7
## X1  0.24146701
## X2  0.01126548
## X3 -0.58524622
## X4  0.46088973
## X5  0.33784371
## X6  0.41707805
## X7 -0.31391372
```

(c)

```r
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.5287 1.1773 1.0973 0.8527 0.80838 0.73259 0.39484
## Proportion of Variance 0.3338 0.1980 0.1720 0.1039 0.09335 0.07667 0.02227
## Cumulative Proportion  0.3338 0.5318 0.7038 0.8077 0.90106 0.97773 1.00000
```

```r
screeplot(pca, type="l")
```



Choosing four principal components is appropriate to summarize this data.

(d)

The first four principal components conveys 80% of the data.

(e)

Yes.