MthStat 768

February 19, 2024

Chapter 7: Principal Component Analysis

```
food <- read.csv(file = '../Data_csv/food.csv') # two points for knitting, one point for running it in
food_names <- food$food_type
food <- food[, -1]
food2 <- sweep(x = food, MARGIN = 1, STATS = food$weight_grams, FUN = "/")
food2 <- food2[, -6]
```

```
apply(X = food2, MARGIN = 2, FUN = sd)
```

```
##            fat_grams food_energy_calories  carbohydrates_grams
##           0.20446639           1.93566372           0.24959716
##        protein_grams       cholesterol_mg  saturated_fat_grams
##           0.08993624           0.67548156           0.06610351
```

```
food3 <- scale(food2)
```

```
S <- cov(food3)
sum(diag(S))
```

```
## [1] 6
```

```
eig <- eigen(S)
eig$values
```

```
## [1] 2.66367512 1.33532725 1.03360142 0.68131153 0.27812812 0.00795656
```

```
cumsum(eig$values)
```

```
## [1] 2.663675 3.999002 5.032604 5.713915 5.992043 6.000000
```

```
cumsum(eig$values) / sum(eig$values)
```

```
## [1] 0.4439459 0.6665004 0.8387673 0.9523192 0.9986739 1.0000000
```

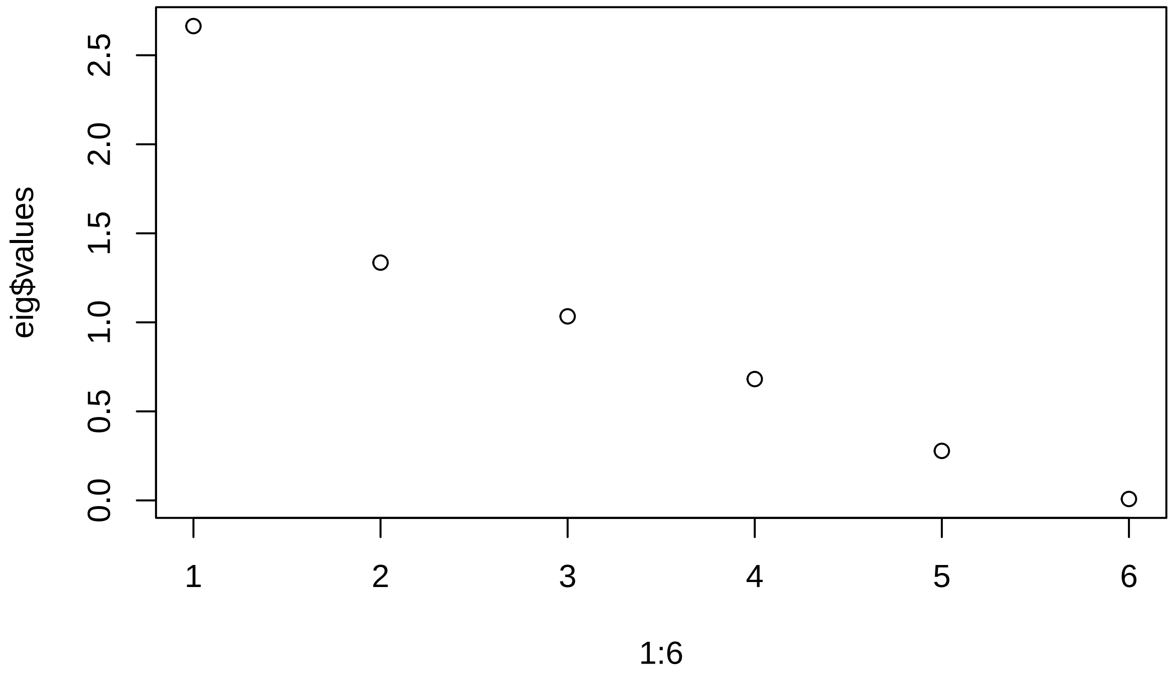
First 3 account for 84% of the total variability.

Then $\mathcal{H}$ is the space spanned by the first 3 eigenvectors:

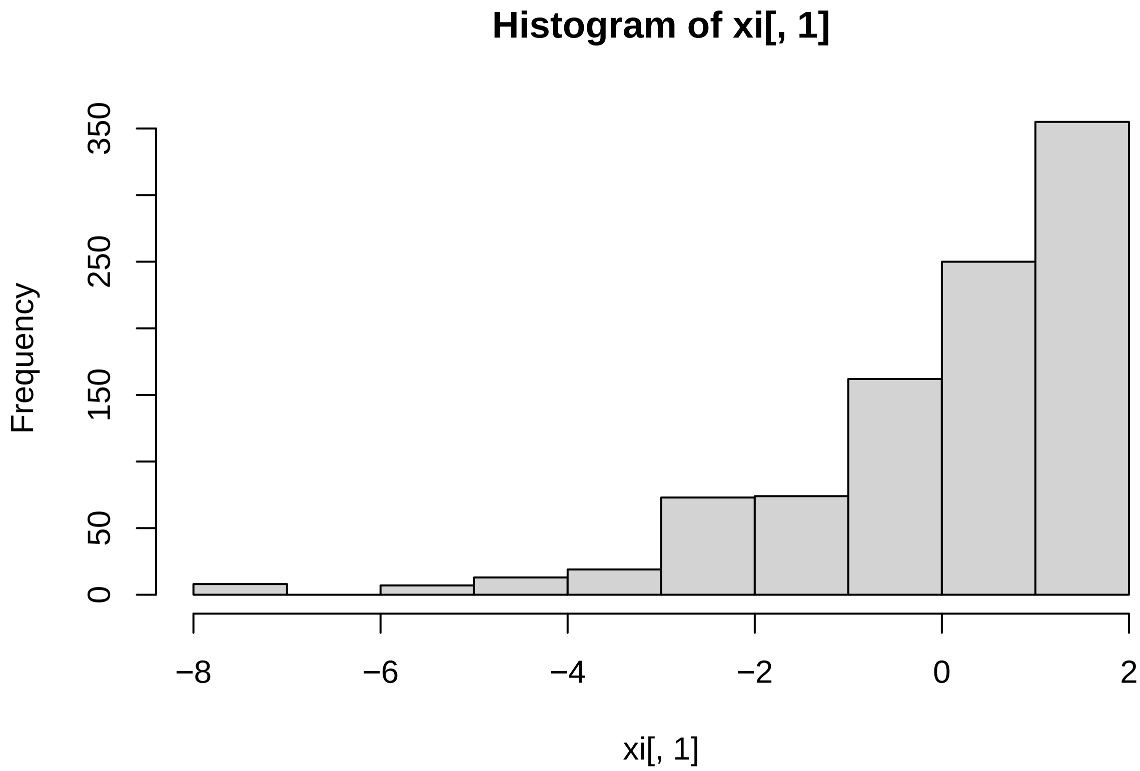
```r
eig$vectors[, 1:3]
```

```
##             [,1]        [,2]       [,3]
## [1,] -0.56114323  0.11208627  0.2793506
## [2,] -0.54369431  0.34362668 -0.1505871
## [3,]  0.02634175  0.65345964 -0.5942314
## [4,] -0.22434167 -0.39322136 -0.6230013
## [5,] -0.24222227 -0.53569538 -0.3179456
## [6,] -0.52898022 -0.02748262  0.2386539
```

```r
plot(1:6, eig$values)
```



```r
xi <- food3 %*% eig$vectors[, 1:3]
hist(xi[,1])
```



```
food_names[which(xi[,1] <= -6)]
```

```
## [1] "BUTTER, SALTED"   "BUTTER, SALTED"   "BUTTER, SALTED"   "BUTTER, UNSALTED"
## [5] "BUTTER, UNSALTED" "BUTTER, UNSALTED" "LARD"             "LARD"
```