MthStat 768

February 14, 2024

Chapter 7: Principal Component Analysis

Example: Food Data

Import csv files: `read.csv(..)`. This creates a dataframe with variable names read from the file. There are $n = 961$ observations, and $r = 8$ variables for each observation. We have a sample $\overrightarrow{x_1}, \ldots, \overrightarrow{x_n} \in \mathbb{R}^r$.

`View(..)` shows the dataframe.

```
food <- read.csv(file = '../Data_csv/food.csv') # two points for knitting, one point for running it in
#View(food)
```

```
food_type <- food$food_type
food <- food[, -1] # delete first column
```

We want to divide each column of the dataframe by the column `weight_grams`. In R we use `sweep(..)`.

```
food2 <- sweep(x = food, MARGIN = 1, STATS = food$weight_grams, FUN = "/")
#View(food2)
```

```
food2 <- food2[, -6]
#View(food2)
```

We are left with $r = 6$ numerical variables.

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} \overrightarrow{x_i}$$

```
xbar <- apply(X = food2, MARGIN = 2, FUN = mean) # apply the mean on the columns
```

Population covariance matrix:

$$\Sigma = \mathbb{E}\{(\overrightarrow{X} - \overrightarrow{\mu}) - (\overrightarrow{X} - \overrightarrow{\mu})^T\}$$

Sample covariance matrix:

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (\overrightarrow{X_i} - \overline{X}) - (\overrightarrow{X_i} - \overline{X})^T$$

```
S <- cov(food2)
print(S)
```

```
##                          fat_grams food_energy_calories carbohydrates_grams
## fat_grams              0.041806506           0.33962407        -0.007912943
## food_energy_calories   0.339624070           3.74679403         0.156691797
## carbohydrates_grams   -0.007912943           0.15669180         0.062298744
## protein_grams          0.002435462           0.04537026        -0.001946321
## cholesterol_mg         0.021486317           0.17752106        -0.027413224
## saturated_fat_grams    0.010106102           0.08228867        -0.002317922
##                        protein_grams cholesterol_mg saturated_fat_grams
## fat_grams              0.0024354616     0.02148632        0.0101061015
## food_energy_calories   0.0453702565     0.17752106        0.0822886688
## carbohydrates_grams   -0.0019463206    -0.02741322       -0.0023179221
## protein_grams          0.0080885274     0.01992273        0.0008484866
## cholesterol_mg         0.0199227307     0.45627533        0.0138441855
## saturated_fat_grams    0.0008484866     0.01384419        0.0043696735
```

$$\rho_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)\text{Var}(X_j)}} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

Let $\Delta = \text{diag}(S)$. Then the sample correlation matrix is

$$R = \Delta^{-1/2} S \Delta^{-1/2}$$

In R we just use `cor(..)`.

```
R <- cor(food2)
print(R)
```

```
##                          fat_grams food_energy_calories carbohydrates_grams
## fat_grams              1.0000000           0.8581172        -0.15505167
## food_energy_calories   0.8581172           1.0000000         0.32432222
## carbohydrates_grams   -0.1550517           0.3243222         1.00000000
## protein_grams          0.1324417           0.2606193        -0.08670418
## cholesterol_mg         0.1555702           0.1357708        -0.16259492
## saturated_fat_grams    0.7477170           0.6431106        -0.14048653
##                        protein_grams cholesterol_mg saturated_fat_grams
## fat_grams                 0.13244166      0.1555702           0.7477170
## food_energy_calories      0.26061931      0.1357708           0.6431106
## carbohydrates_grams      -0.08670418     -0.1625949          -0.1404865
## protein_grams             1.00000000      0.3279447           0.1427203
## cholesterol_mg            0.32794472      1.0000000           0.3100483
## saturated_fat_grams       0.14272030      0.3100483           1.0000000
```

Dimension Reduction

Our sample $\{\overrightarrow{X_1}, \ldots, \overrightarrow{X_n}\}$ is a cloud of points in $\mathbb{R}^r$. We are going to try to find a subspace $\mathcal{H}$ of $\dim(\mathcal{H}) = q < r$ that approximates the data. The dispersion in the orthogonal directions to $\mathcal{H}$ are as small as possible.

In mathematical terms: If P is a projection matrix onto some subspace $\mathcal{H}$, we want to find P that minimizes

$$D = \sum_{i=1}^{n} \left\| \underbrace{(\overrightarrow{X_i} - \overline{X})}_{=\tilde{X}_i} - P(\overrightarrow{X_i} - \bar{X}) \right\|^2$$

Using the norm, we can rewrite D in terms of the trace.