STA 135: Discussion 1

Chris Conley

Jan. 11, 2017

Basic Descriptive Statistics

Suppose that we have an $n \times p$ data array $\mathbf{X}$. That is, we have n measurements on p variables. We can compute statistics that describe basic properties of center, scale, and linear association between variables. For the ith and kth variables, we have

- Sample means $\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \qquad k = 1, \ldots, p$
- Sample variances $s_{kk} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2 \qquad k = 1, \ldots, p$
- Sample covariance
  $s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \qquad i = 1, \ldots, p \;\; k = 1, \ldots, p$
- Sample correlations (Pearson's product moment correlation coefficient)

$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}s_{kk}}} \qquad i = 1, \ldots, p \;\; k = 1, \ldots, p$

Exercise 1.3 Computing Array of Statistics

Store 5 × 3 data array.

```r
X <- matrix(c(9,2,6,5,8,
              12,8,6,4,10,
               3,4,0,2,1),5,3)
colnames(X) <- paste0("x", 1:3)
rownames(X) <- paste0("n", 1:5)
X
```

```
##    x1 x2 x3
## n1  9 12  3
## n2  2  8  4
## n3  6  6  0
## n4  5  4  2
## n5  8 10  1
```

On an exam, you may be required to compute these descriptive statistics by hand. In chapter 3, we will learn some matrix representations of how to compute the array of descriptive statistics.

Exercise 1.3 Computing Mean Vector

Using the apply function, we can compute the basic statistics over each variable (i.e. column) in the data array $\mathbf{X}$ by specifying the 2nd margin and telling what function to apply. First we compute the mean vector $\bar{\mathbf{x}}$.

```r
xbar <- apply(X = X, MARGIN = 2, FUN = mean)
xbar
```

```
## x1 x2 x3
##  6  8  2
```

Exercise 1.3 Computing Sample Variances

Now compute the sample variances, s_{kk}'s. Note that R's `var` function will use the $n - 1$ divisor instead of n. So multiply the output of `var` by $\frac{n-1}{n}$

```r
n <- nrow(X)
skk <- ( (n - 1) / n ) * apply(X = X, MARGIN = 2,
                                FUN = var)
skk
```

```
## x1 x2 x3
##  6  8  2
```

Exercise 1.3 Computing Covariance, $\mathbf{S_n}$

R has a built in function to compute the sample covariance of the data array.

```
Sn <- ( (n - 1) / n ) * cov(X)
Sn
```

```
##       x1  x2   x3
## x1  6.0 4.0 -1.4
## x2  4.0 8.0  1.2
## x3 -1.4 1.2  2.0
```

Let's confirm the s_{12} entry by hand .

Exercise 1.3 Computing Correlation **R**

```
cov2cor(Sn)
```

```
##             x1        x2         x3
## x1  1.0000000 0.5773503 -0.4041452
## x2  0.5773503 1.0000000  0.3000000
## x3 -0.4041452 0.3000000  1.0000000
```

Let's confirm the r_{12} entry by hand.

Exercise 1.6 Read in the data

Read in the 42 measurements on air-pollution variables recorded at 12:00 noon in Los Angelos on different days.

```r
file <- "~/../Box Sync/sta135-winter17/discussion/air-pollution.
ap <- read.table(file, header = T)
X <- as.matrix(ap)
head(X)
```

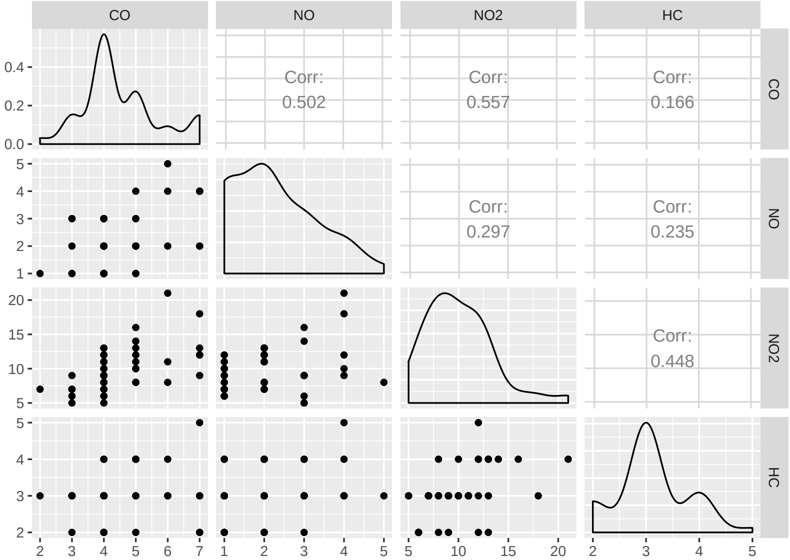
```
##      Wind Radiation CO NO NO2 O3 HC
## [1,]    8        98  7  2  12  8  2
## [2,]    7       107  4  3   9  5  3
## [3,]    7       103  4  3   5  6  3
## [4,]   10        88  5  2   8 15  4
## [5,]    6        91  4  2   8 10  3
## [6,]    8        90  5  2  12 12  4
```

Exercise 1.6 Scatter plot matrix

Pairwise relationships can quickly be visualized in R. The following command generates all pairwise scatter plots for the pollutants CO, NO, NO2, HC. Also it shows the marginal distribution of each variable along the diagonal.

```r
library(GGally)
ggpairs(ap[,c("CO", "NO", "NO2", "HC")])
```

Exercise 1.6 Scatter plot matrix



Statistical distance

- How might it differ from regular euclidean distance in p dimensions?

Statistical distance

- How might it differ from regular euclidean distance in p dimensions?

Differences in scale of the variables matter.

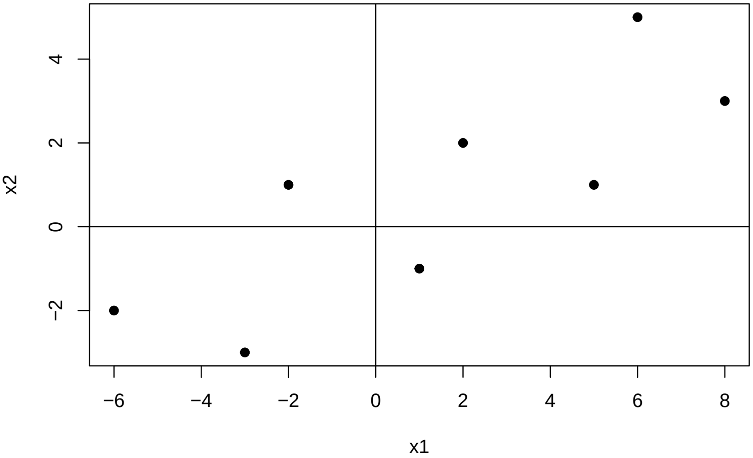
Correlation among variables has meaningful impact on distance. Need to rotate our coordinate system to adjust for correlation structure.

Exercise 1.9

```
X <- matrix(c(-6, -3, -2, 1, 2, 5, 6, 8,
              -2, -3, 1, -1, 2, 1, 5, 3), 8, 2)
X
```

```
##      [,1] [,2]
## [1,]   -6   -2
## [2,]   -3   -3
## [3,]   -2    1
## [4,]    1   -1
## [5,]    2    2
## [6,]    5    1
## [7,]    6    5
## [8,]    8    3
```

Exercise 1.9 (a) Plot the data



Exercise 1.9 (a)

Obtain s_{11}, s_{22}, s_{12}

```
Sn <- cov(X)
Sn
```

```
##           [,1]      [,2]
## [1,] 23.41071 10.392857
## [2,] 10.39286  7.071429
```

Exercise 1.9 (b)

Rotate the data vectors (x_1, x_2) about the new axes to obtain $(\tilde{x}_1, \tilde{x}_2)$, where

$$\tilde{x}_1 = x_1 cos(\theta) + x_2 sin(\theta)$$

$$\tilde{x}_1 = -x_1 sin(\theta) + x_2 cos(\theta)$$

Under an angle of $\theta = 26$.

```r
theta <- 26
#rotate x1 axis
rotx1 <- function(theta, x1, x2) x1*cos(theta) + x2*sin(theta)
rx1 <- rotx1(26, X[,1], X[,2])
#rotate x2 axis
rotx2 <- function(theta, x1, x2) -x1*sin(theta) + x2*cos(theta)
rx2 <- rotx2(26, X[,1], X[,2])
```

Exercise 1.9 (c)

Compute $\tilde{s}_{11}, \tilde{s}_{22}$.

```
rs11 <- var(rx1)
rs11
```

```
## [1] 24.16337
```

```
rs22 <- var(rx2)
rs22
```

```
## [1] 6.318768
```

## Exercise 1.9 (d)

Compute the distance of the new point $(x_1, x_2) = (4, -2)$ by transforming into $(\tilde{x}_1, \tilde{x}_2)$ and evaluating.

$$d(O, P) = \sqrt{\frac{\tilde{x_1^2}}{\tilde{s_{11}}} + \frac{\tilde{x_2^2}}{\tilde{s_{22}}}}$$

```
rx1_new <- rotx1(26, 4, -2)
rx2_new  <- rotx2(26, 4, -2)
d <- sqrt(rx1_new^2/rs11 + rx2_new^2/rs22)
d
```

```
## [1] 1.741614
```