MA554 Applied Multivariate Analysis HW1

Aukkawut Ammartayakun

2023-09-19

Problem 1

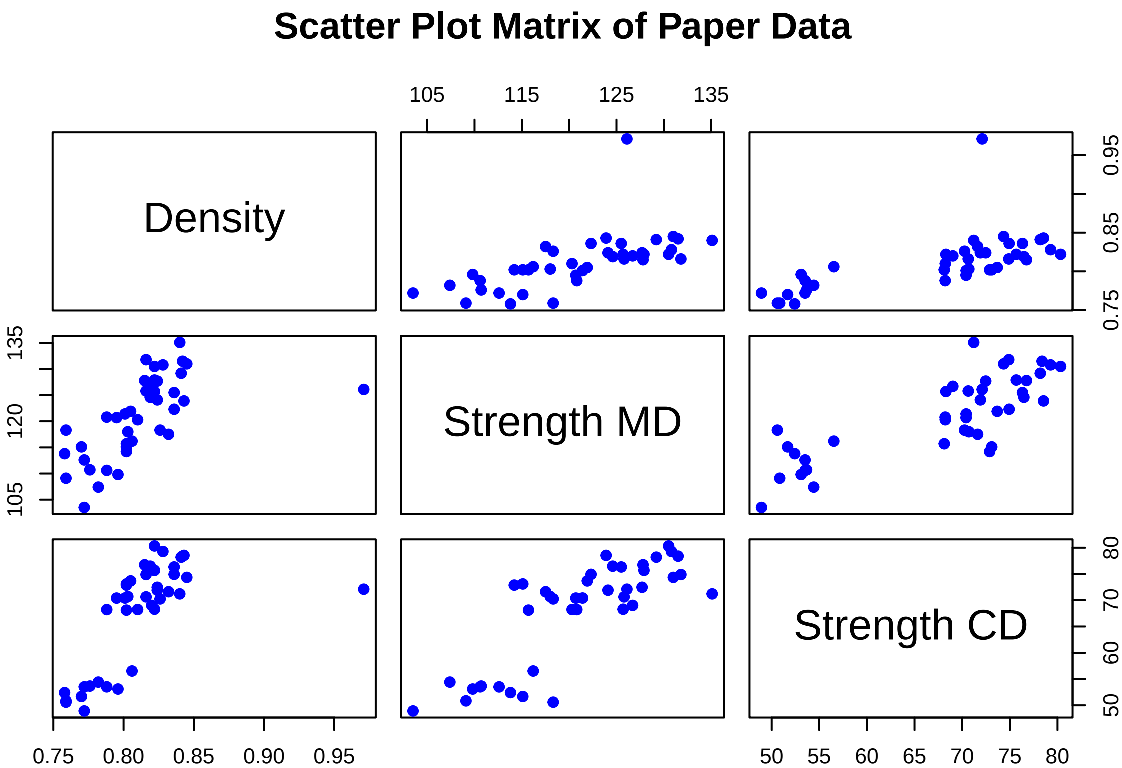
Paper is manufactured in continuous sheets several feet wide. Because of the orientation of fibers within the paper, it has a different strength when measured in the direction produced by the machine than when measured across, or at right angles to, the machine direction. The dataset in `T1-2.txt` contains 41 measurements of X_1 = density (grams/cubic centimeter), X_2 = strength (pounds) in the machine direction, and X_3 = strength (pounds) in the cross direction. Use any statistical package to answer the following.

```r
var_summary <- function(data) {
    s <- summary(data)
    var_data <- sapply(data, var, na.rm = TRUE)
    r <- rbind(s, Variance = var_data)
    return(r)
}
var_summary(data)
```

a) Report summary statistics (means and variances) of the three variables, and present a scatter plot matrix.

```
##                               Density              Strength MD
##          "Mode:logical  " "Min.   :0.7580  "    "Min.   :103.5  "
##          "NA's:41       " "1st Qu.:0.7950  "    "1st Qu.:115.1  "
##          NA               "Median :0.8150  "    "Median :121.4  "
##          NA               "Mean   :0.8119  "    "Mean   :121.0  "
##          NA               "3rd Qu.:0.8260  "    "3rd Qu.:126.7  "
##          NA               "Max.   :0.9710  "    "Max.   :135.1  "
## Variance NA               "0.00126457804878049" "59.3211480487805"
##           Strength CD
##          "Min.   :48.93  "
##          "1st Qu.:56.53  "
##          "Median :70.70  "
##          "Mean   :67.72  "
##          "3rd Qu.:74.89  "
##          "Max.   :80.33  "
## Variance "95.8566671951219"
```

From this we can see that $(\bar{X}_1, S_1^2) = (0.8119, 0.0013)$, $(\bar{X}_2, S_2^2) = (121.0, 59.321)$, $(\bar{X}_3, S_3^2) = (70.70, 95.857)$



b) Identify an outlier Visually, we can see that the point where density is greater than 0.96. That is the maximum of the density.

```
outlier <- data[data$Density == max(data$Density), ]
outlier
```

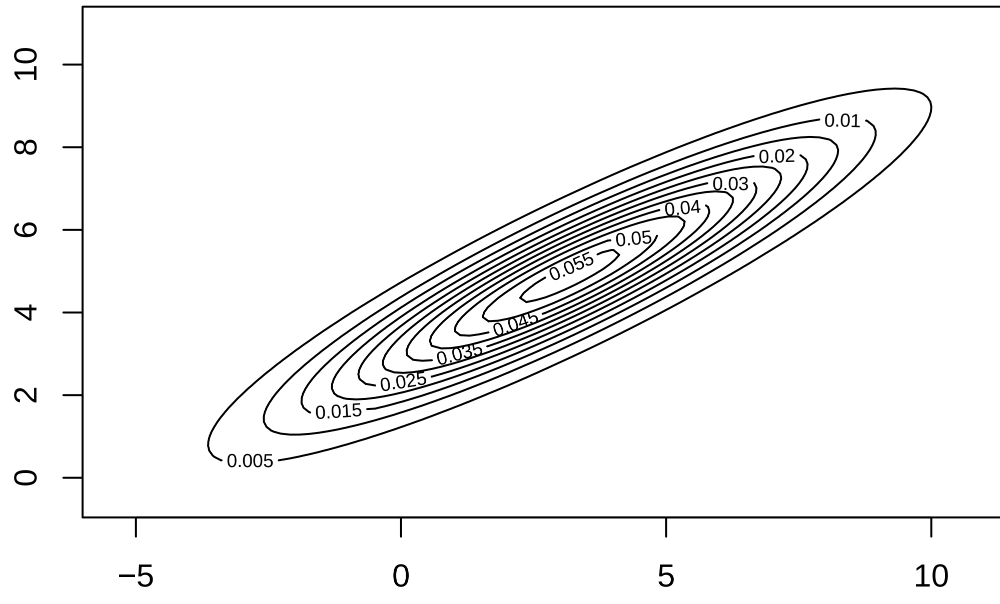
```
## # A tibble: 1 x 4
##   ``    Density `Strength MD` `Strength CD`
##   <lgl>   <dbl>         <dbl>         <dbl>
## 1 NA      0.971          126.          72.1
```

Problem 2

Consider a bivariate normal distribution for $\mathbf{X} = (X_1, X_2)^\top$, with $\mathbb{E}[X_1] = 3, \mathbb{E}[X_2] = 5, \mathrm{Var}[X_1] = 10, \mathrm{Var}[X_2] = 4, \mathrm{Corr}(X_1, X_2) = 0.9$

```
library(mvtnorm)
library(MASS)
mu <- c(3, 5)
sigma1 <- sqrt(10)
sigma2 <- sqrt(4)
rho <- 0.9
cov_12 <- rho * sigma1 * sigma2
Sigma <- matrix(c(10, cov_12, cov_12, 4), ncol=2)
samples <- mvrnorm(n = 100, mu = mu, Sigma = Sigma)
x <- seq(min(samples[,1]) - 1, max(samples[,1]) + 1, length=100)
y <- seq(min(samples[,2]) - 1, max(samples[,2]) + 1, length=100)
z <- matrix(0, ncol=length(x), nrow=length(y))
for (i in 1:length(x)) {
```

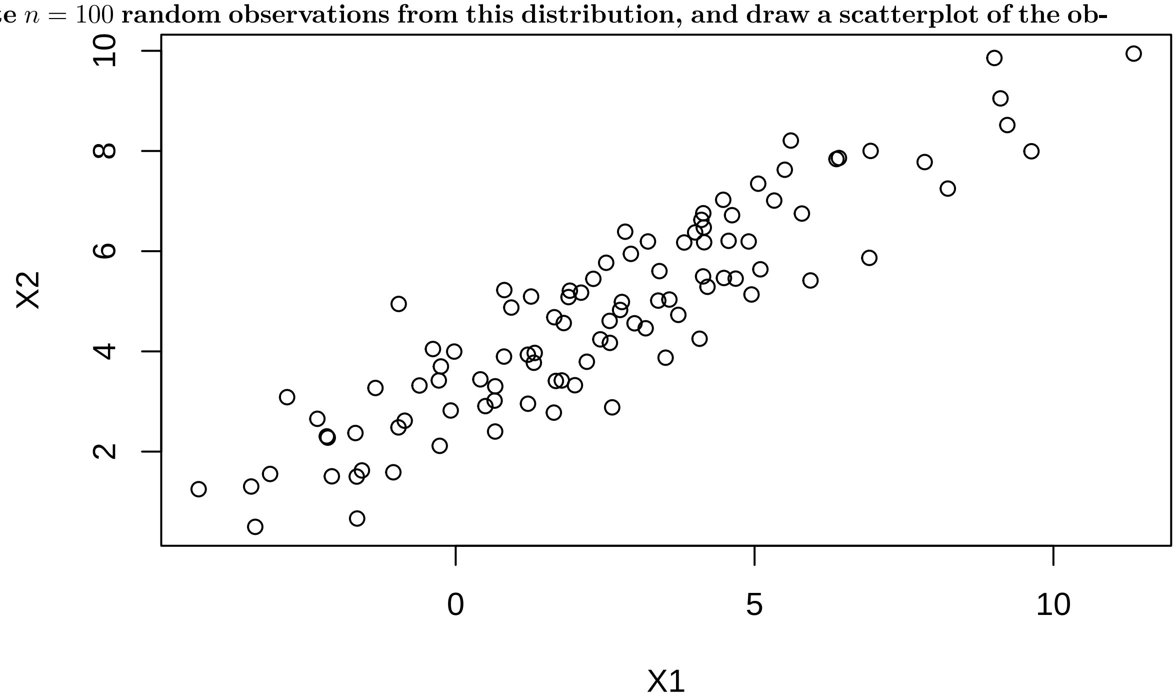
```r
  for (j in 1:length(y)) {
    z[j,i] <- dmvnorm(c(x[i], y[j]), mean=mu, sigma=Sigma)
  }
}

contour(x, y, z)
```



a) Sketch a contour of the density

```r
plot(samples, xlab="X1", ylab="X2")
```

b) Generate $n = 100$ random observations from this distribution, and draw a scatterplot of the observation.



```
summary(samples)
```

c) Report the sample mean, and compare it with the population mean.

```
##        V1               V2
##  Min.   :-4.3006   Min.   :0.4984
##  1st Qu.: 0.3034   1st Qu.:3.2953
##  Median : 2.4687   Median :4.7785
##  Mean   : 2.4329   Mean   :4.7368
##  3rd Qu.: 4.2780   3rd Qu.:6.1821
##  Max.   :11.3438   Max.   :9.9436
```

From the result, we can see that the sample mean is $\bar{X} = (2.8398, 4.8294)^\top$ while the population mean is $\mathbb{E}[X] = (3, 5)^\top$

d) Report the sample covariance matrix S, and sketch the equidistance set given by the sample Mahalanobis distance, i.e., $E_c = \{\mathbf{y} \in \mathbb{R}^2 | (\mathbf{y} - \bar{\mathbf{x}})^\top S^{-1} (\mathbf{y} - \bar{\mathbf{x}}) = c^2\}$ for $c > 0$

Problem 3

By expressing a correlation matrix $\mathbf{R}_{n\times n}$ with equal correlation ρ as $\mathbf{R} = (1-\rho)\mathbb{I} + \rho\mathbb{J}$, where $\mathbb{J}$ is an $n \times n$ matrix of ones, find $\det(\mathbf{R})$ and $\mathbf{R}^{-1}$.

Hint: Consider the fact that $\det(\mathbb{I}_p + AB) = \det(\mathbb{I}_q + BA)$ for $A_{p\times q}$ and $B_{q\times p}$ and the Woodbury formula

---

bla

Problem 4

Prove Corollary 12 of Lecture 1.

> **Corollary 12** Let $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$ and $\mathbf{A}$ be a $p \times p$ symmetric matrix. Then
>
> $$Y = \mathbf{X}^\top \mathbf{A} \mathbf{X} \sim \chi^2(m)$$
>
> if and only if either i) m of the eigenvalues of $\mathbf{A}\Sigma$ are 1 and the rest are 0 or ii) m of the eigenvalues of $\Sigma\mathbf{A}$ are 1 and the rest are 0.

Hint: Similar to the proof of Corollary 13.

Let $\mathbf{B} = \Sigma^{\frac{1}{2}} \mathbf{A} \Sigma^{\frac{1}{2}}$, we can see that from theorem 11,

Problem 5

Prove Theorem 15 of Lecture 1.

> **Theorem 15** Let $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$. Suppose $\mathbf{A}$ and $\mathbf{B}$ are $p \times p$ symmetric matrices. Then if $\mathbf{B}\Sigma\mathbf{A} = 0$, $\mathbf{X}^\top\mathbf{A}\mathbf{X}$ and $\mathbf{X}^\top\mathbf{B}\mathbf{X}$ are independent.

Hint: Similar to the proof of Theorem 14.

---

bla

Problem 6

Let $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \sigma^2\mathbb{I})$ and $\mathbf{A}$ be a $p \times p$ symmetric matrix. Then $Y = \frac{1}{\sigma^2}\mathbf{X}^\top \mathbf{A}\mathbf{X} \sim \chi^2(m, \delta)$ with $\delta = \frac{\boldsymbol{\mu}^\top \mathbf{A}\boldsymbol{\mu}}{\sigma^2}$, if and only if $\mathbf{A}$ is an idempotent of rank $m \leq p$.

Hint: Similar to the proof of Theorem 11. Note: Let $X_1, \ldots, X_n$ be independent random variables and $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, n$. The distribution of the random variable $Y = \frac{(X_1^2 + \cdots + X_n^2)}{\sigma^2}$ is called the noncentral chi-square distribution with degrees of freedom n and the noncentrality parameter $\delta = \frac{(\mu_1^2 + \cdots + \mu_n^2)}{\sigma^2}$, denoted by $\chi^2(n, \delta)$

---

bla