STAT 636, Fall 2015 - Assignment 1

SOLUTIONS

1. The data in Table 6.12 of the textbook contain $p = 4$ oxygen volume measurements for 25 males and 25 females. The variables are X_1: oxygen volume (L/min.) while resting, X_2: oxygen volume (mL/kg/min.) while resting, X_3: oxygen volume (L/min.) during strenuous exercise, and X_4: oxygen volume (mL/kg/min.) during strenuous exercise.

   (a) Report a table showing the sample averages and standard deviations for each variable, by gender. Comment.

   SEE CODE AT THE END OF THIS DOCUMENT.

	FEMALE		MALE	
VARIABLE	MEAN	S.D.	MEAN	S.D.
X_1	0.3136	0.0987	0.3972	0.0844
X_2	5.1788	1.6675	5.3296	1.0697
X_3	2.3152	0.3471	3.6876	0.6752
X_4	38.1548	4.8229	49.4204	7.4332

   A FEW THINGS STAND OUT:

   - THERE IS CLEARLY A BIG DIFFERENCE IN OXYGEN VOLUME BETWEEN RESTING AND EXERCISE CONDITIONS. THE MAGNITUDE OF THIS DIFFERENCE IS SIMILAR FOR THE TWO GENDERS.
   - ALL OF THE FEMALE MEANS ARE LESS THAN THEIR MALE COUNTERPARTS.
   - THERE ARE SOME DIFFERENCES BETWEEN GENDERS IN THE VARIABLE STANDARD DEVIATIONS. FOR EXAMPLE, THE S.D. FOR X_2 AMONG FEMALES IS ABOUT 1.5 TIMES THAT AMONG MALES.

   (b) Make a pairs plot like we did for the pottery data. Comment on any relationships you see. Which individual would you say is an outlier?

   SEE FIGURE 1. THE TWO VOLUME MEASURES TAKEN WHILE RESTING ARE HIGHLY POSITIVELY CORRELATED WITH EACH OTHER, AS WOULD BE EXPECTED. SIMILARLY FOR THE TWO VOLUME MEASURES TAKEN UNDER EXERTION. THERE IS MUCH LESS CORRELATION BETWEEN THE RESTING MEASUREMENTS AND THE EXERTION MEASUREMENTS. AN OUTLIER IS APPARENT IN THE SCATTERPLOTS INVOLVING X_1 AND X_2. THIS CORRESPONDS TO SUBJECT 48, THE 23RD FEMALE.

   (c) Make a coplot, like we did for the pottery data, to compare X_1 to X_3 by gender. Does there appear to be a difference for this pair of variables between genders?

   SEE FIGURE 2. THERE DOES NOT APPEAR TO BE MUCH OF A RELATIONSHIP BETWEEN THESE TWO VARIABLES IN FEMALES. ALTHOUGH, THIS MAY BE DUE TO THE OUTLIER. IF THE OUTLIER WERE REMOVED, WE MIGHT SEE MORE EVIDENCE OF POSITIVE CORRELATION LIKE THAT FOR MALES. THERE IS CLEARLY A DIFFERENCE BETWEEN PAIRS (X_1, X_3) FOR FEMALES AND PAIRS FOR MALES, WITH MALES TENDING TO HAVE LARGER VALUES FOR BOTH VARIABLES.

2. The multivariate normal distribution is defined by its probability density function (pdf)

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}$$

for $-\infty < x_i < \infty$, $i = 1, 2, \ldots, p$. Volumes underneath this surface equal probabilities. The mean *vector* of this distribution is $\boldsymbol{\mu}$, and the covariance *matrix* is $\boldsymbol{\Sigma}$. In the bivariate setting ($p = 2$), we have

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

Thus, draws from this distribution are *pairs* (vectors of length $p = 2$). The averages of the two components among all pairs in the population equal μ_1 and μ_2, respectively. Similarly, the variances of the two components among all pairs in the population equal σ_{11} and σ_{22}, respectively. Finally, the *covariance* between the two components equals σ_{12}, which means that the correlation equals $\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$.

For the bivariate normal distribution:

(a) Recall that the distance between the point $P = (x_1, x_2)$ and $Q = (\mu_1, \mu_2)$ can be written as

$$d(P, Q) = \sqrt{a_{11}(x_1 - \mu_1)^2 + 2a_{12}(x_1 - \mu_1)(x_2 - \mu_2) + a_{22}(x_2 - \mu_2)^2}$$

We will see that the statistical distance between the two *vectors* $\mathbf{x}$ and $\boldsymbol{\mu}$ can be written as

$$d(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

And it turns out that $d(P, Q) = d(\mathbf{x}, \boldsymbol{\mu})$. Use this result to derive the values of a_{11}, a_{12}, and a_{22}.

WE HAVE

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$$

SO THAT

$$\begin{aligned}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) &= \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \Big(\sigma_{22}(x_1 - \mu_1)^2 - 2\sigma_{12}(x_1 - \mu_1)(x_2 - \mu_2) \\ &\qquad + \sigma_{11}(x_2 - \mu_2)^2 \Big) \\ &= a_{11}(x_1 - \mu_1)^2 + 2a_{12}(x_1 - \mu_1)(x_2 - \mu_2) + a_{22}(x_2 - \mu_2)^2\end{aligned}$$

FOR

$$a_{11} = \frac{\sigma_{22}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}, \quad a_{12} = -\frac{\sigma_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}, \quad a_{22} = \frac{\sigma_{11}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}$$

(b) Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & -1.6 \\ -1.6 & 4.0 \end{pmatrix}$$

i. Use the `persp` function to graph the pdf. You can do this by evaluating the pdf over a grid of `x` values. Code the pdf manually; i.e., do not use the `dmvnorm` function (or any other predefined function). Use the `ticktype = ''detailed''` option to include axis tick marks and labels.

Here are some `R` functions and operations that you will find useful: `sqrt` computes the square root; `det` computes the determinant of a matrix; `t(v)` computes the transpose of the vector / matrix `v`; `u %*% v` computes the vector / matrix product of the vectors / matrices `u` and `v`; `exp(a)` equals e^a; `solve` inverts a matrix. If you need a refresher on the vector / matrix operations, see Supplement 2A in the textbook. We will revisit them in more detail in Topic 2.

See Figure 3. The surface is peaked and angled in a negative direction because of the negative correlation between the variables. Also, there is more variation in the second component than in the first.

ii. Let O be the origin $(0, 0)$, P be the point $(0, -2)$, and Q be the point $(\mu_1, \mu_2) = (1, -1)$. Which of O or P is "closer" to $\boldsymbol{\mu}$, based on statistical distance? Which of O or P is closer to $\boldsymbol{\mu}$, based on straight-line distance?

First, note that $a_{11} = 2.7778$, $a_{12} = 1.1111$, and $a_{22} = 0.6944$. The statistical distances are then

$$d(O, Q) = \sqrt{2.7778(0-1)^2 + 2(1.1111)(0-1)(0+1) + 0.6944(0+1)^2} = 1.1180$$
$$d(P, Q) = \sqrt{2.7778(0-1)^2 + 2(1.1111)(0-1)(-2+1) + 0.6944(-2+1)^2} = 2.3863$$

We can verify that these equal the distance calculations involving matrix operations. We therefore have that O is closer to $\boldsymbol{\mu}$ than P, in terms of statistical distance. Now we compute the straight-line distances:

$$d(O, Q) = \sqrt{(0-1)^2 + (0+1)^2} = \sqrt{2}$$
$$d(P, Q) = \sqrt{(0-1)^2 + (-2+1)^2} = \sqrt{2}$$

Thus, while O and P are equal straight-line distance from $\boldsymbol{\mu}$, O is located in a more "likely" region than is P.

iii. Consider all of the pairs (x_1, x_2) located inside a small square centered at O. That is, let R_O be the square containing all pairs (x_1, x_2) such that $-\epsilon \leq x_1 \leq \epsilon$ and $-\epsilon \leq x_2 \leq \epsilon$ for some small value of ϵ (e.g., $\epsilon = 0.01$). Similarly, let R_P consist of all pairs located inside an equally-small square centered at P, for which $-\epsilon \leq x_1 \leq \epsilon$ and $-2 - \epsilon \leq x_2 \leq -2 + \epsilon$. Let $P(\mathbf{x} \in R_O)$ be the probability that a randomly-drawn pair from this bivariate normal distribution falls within R_O. Similarly, let $P(\mathbf{x} \in R_P)$ be the probability that a randomly-drawn pair falls within R_P. Is $P(\mathbf{x} \in R_O) < P(\mathbf{x} \in R_P)$, $P(\mathbf{x} \in R_O) = P(\mathbf{x} \in R_P)$, or $P(\mathbf{x} \in R_O) > P(\mathbf{x} \in R_P)$? Why? No calculations are required to answer this.

We know that probabilities equal volumes underneath the pdf. Thus, for example, the double integral of the pdf over the region defined by R_O equals $P(\mathbf{x} \in R_O)$. Based on inspection of the bivariate normal pdf, we see the negative squared statistical distance inside the exponential term. The value of the pdf therefore gets smaller (closer to zero) the further away from $\boldsymbol{\mu}$ you get (defining distance as statistical distance). Since the statistical distance from P to $\boldsymbol{\mu}$ is greater than that from O to $\boldsymbol{\mu}$, we have $P(\mathbf{x} \in R_O) > P(\mathbf{x} \in R_P)$. In other words, we are more likely to see pairs near the origin than we are to see pairs near the point $(0, 2)$.

```
####
#### (1)
####

##
## Input data.
##

dta <- read.delim("T6-12.DAT", header = FALSE, sep = "")
colnames(dta) <- c("X_1", "X_2", "X_3", "X_4", "Gender")
attach(dta)

n_1 <- n_2 <- 25
p <- 4

## Summary statistics.
x_bar_M <- colMeans(dta[1:n_1, 1:p])
x_bar_F <- colMeans(dta[n_1 + (1:n_2), 1:p])
sg_M <- apply(dta[1:n_1, 1:p], 2, sd)
sg_F <- apply(dta[n_1 + (1:n_2), 1:p], 2, sd)

##
## Pictures.
##

## Pairs plot. The two volume measures taken while resting are highly positively
## correlated with each other, as would be expected. Similarly for the two volume
## measures taken under exertion. There is much less correlation between the resting
## measurements and the exertion measurements. Subject 48, the 23rd female, appears to be
## an outlier, based on her resting oxygen levels. Interestingly, her oxygen volumes
## under exertion are consistent those for the other females.
pairs(dta)

dta[which.max(dta$X_1), ]

## Coplot.
coplot(X_3 ~ X_1 | Gender, data = dta)

####
#### (2)
####

mu <- c(1, -1)
rho <- -0.8
sg <- c(1, 2)
```

```
Sigma <- matrix(c(sg[1] ^ 2, rho * prod(sg), rho * prod(sg), sg[2] ^ 2), nrow = 2)

##
## A perspective plot.
##

## Function to evaluate the pdf. Make it take the determinant and inverse of Sigma as
## arguments, so we only have to compute them once.
f <- function(x, mu, Sigma_det, Sigma_inv) {
  p <- length(x)

  f_x <- (1 / ((2 * pi) ^ (p / 2) * sqrt(Sigma_det))) *
    exp(-0.5 * t(x - mu) %*% Sigma_inv %*% (x - mu))
  return(f_x)
}

Sigma_det <- det(Sigma)
Sigma_inv <- solve(Sigma)

x <- seq(-2, 4, length = 50)
y <- seq(-7, 6, length = 50)
z <- outer(x, y, function(x, y) { apply(cbind(x, y), 1, f, mu, Sigma_det, Sigma_inv) })

persp(x, y, z, xlab = "x_1", ylab = "x_2", zlab = "f", ticktype = "detailed")

##
## Statistical distance.
##

O <- c(0, 0)
P <- c(0, -2)

a_11 <- Sigma[2, 2] / (Sigma[1, 1] * Sigma[2, 2] - Sigma[1, 2] ^ 2)
a_12 <- -Sigma[1, 2] / (Sigma[1, 1] * Sigma[2, 2] - Sigma[1, 2] ^ 2)
a_22 <- Sigma[1, 1] / (Sigma[1, 1] * Sigma[2, 2] - Sigma[1, 2] ^ 2)

## Statistical distance using the constants a_11, a_12, and a_22.
sqrt(a_11 * (O[1] - mu[1]) ^ 2 + 2 * a_12 * (O[1] - mu[1]) * (O[2] - mu[2]) +
  a_22 * (O[2] - mu[2]) ^ 2)
sqrt(a_11 * (P[1] - mu[1]) ^ 2 + 2 * a_12 * (P[1] - mu[1]) * (P[2] - mu[2]) +
  a_22 * (P[2] - mu[2]) ^ 2)

## Equivalently, using the matrix operation.
sqrt(t(O - mu) %*% Sigma_inv %*% (O - mu))
sqrt(t(P - mu) %*% Sigma_inv %*% (P - mu))
```

```
## Straight-line distances.
sqrt((O[1] - mu[1]) ^ 2 + (O[2] - mu[2]) ^ 2)
sqrt((P[1] - mu[1]) ^ 2 + (P[2] - mu[2]) ^ 2)
```

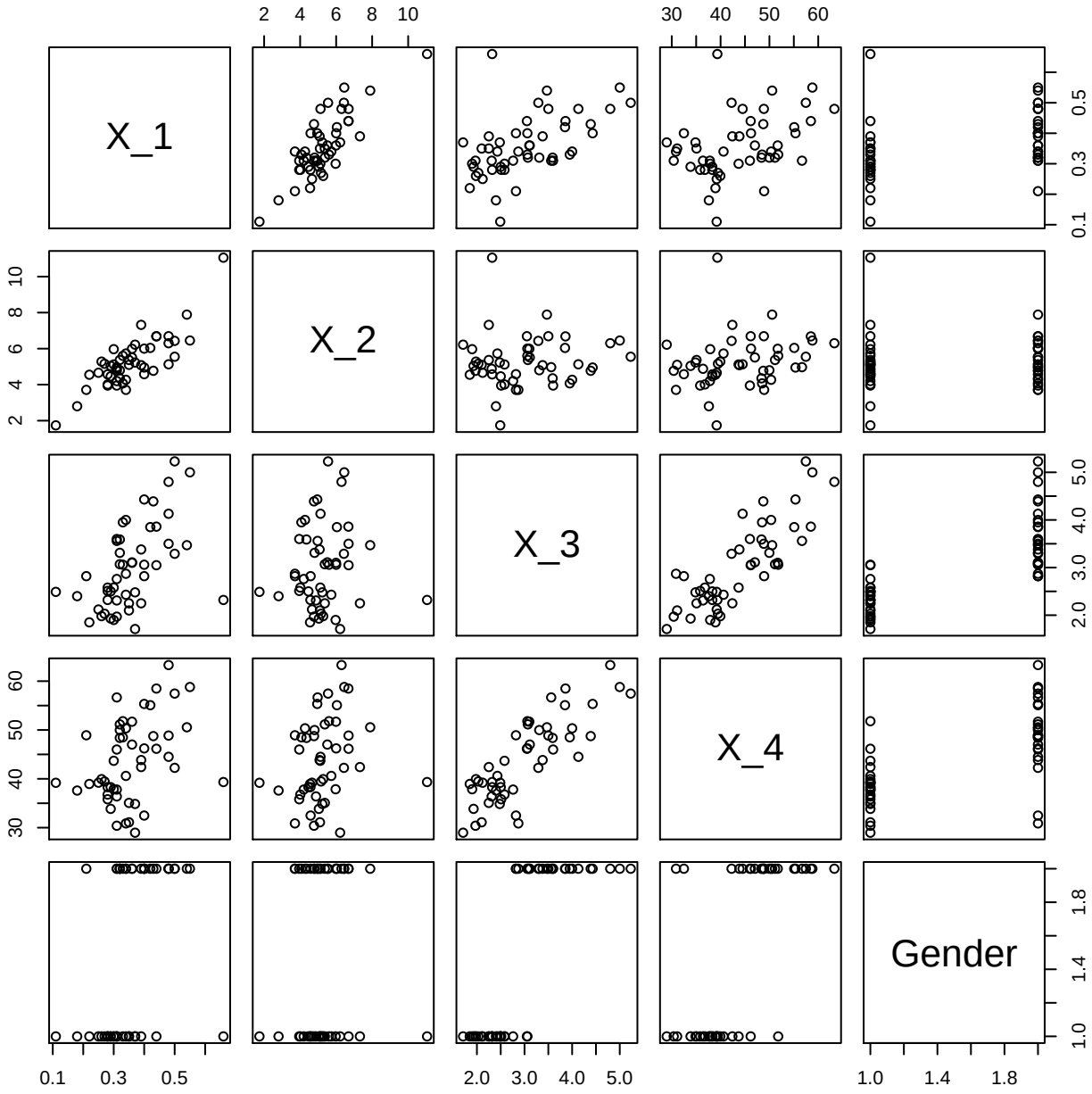

Figure 1: Pairs plot for the oxygen volume data.

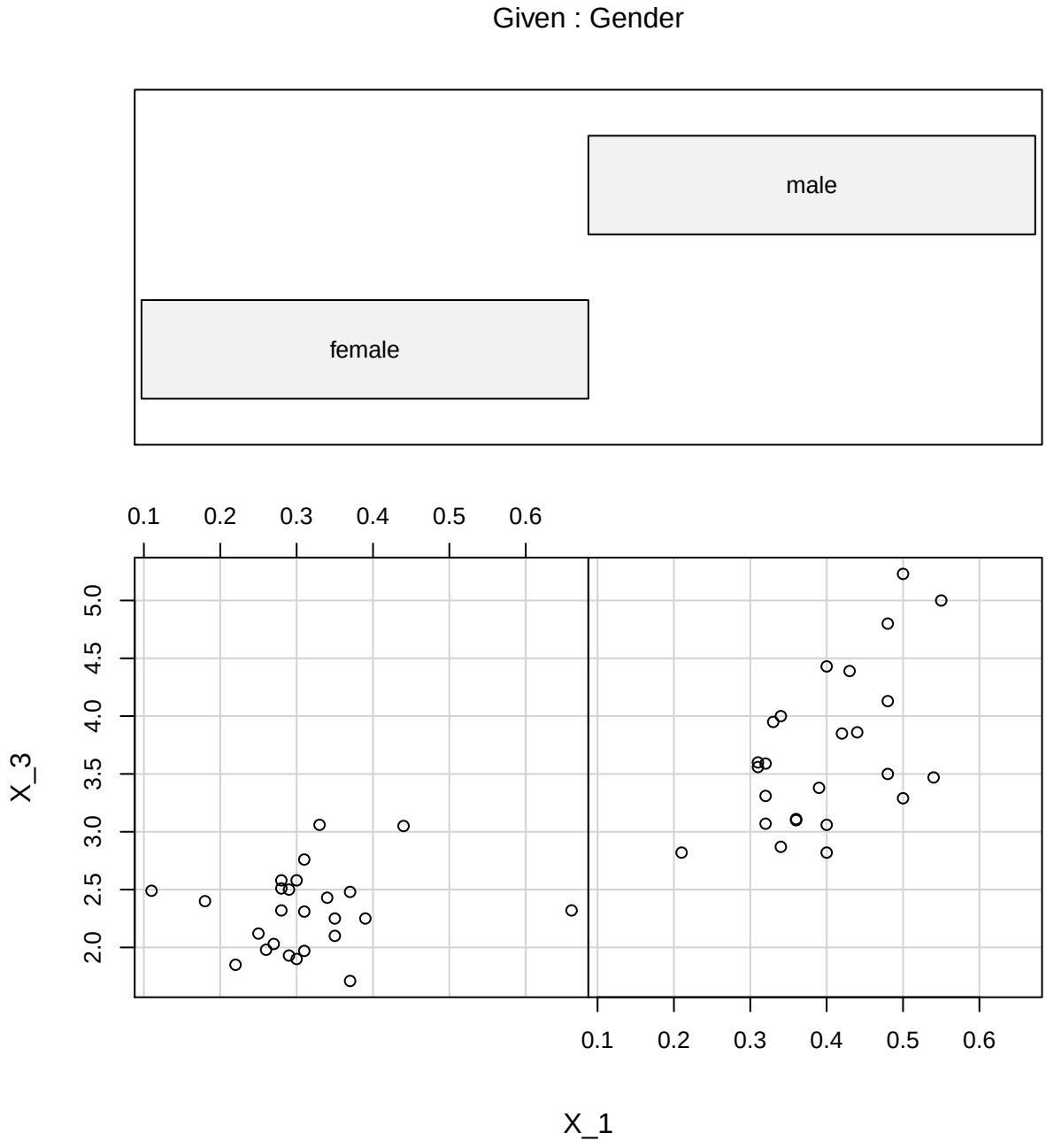

Figure 2: Coplot for comparing X_1 to X_3 by gender in the oxygen volume data.

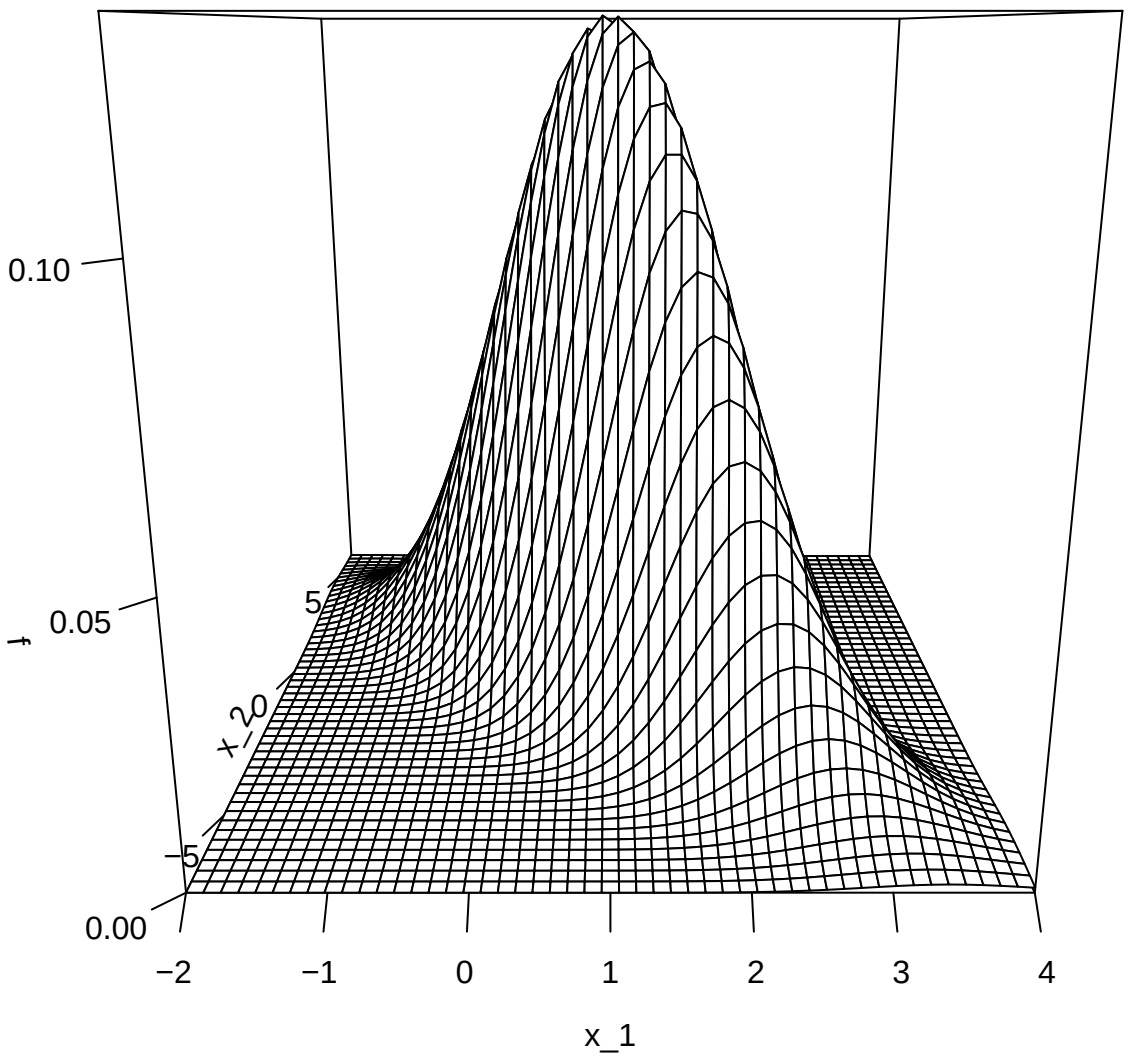

Figure 3: The pdf for the bivariate normal distribution from number 2.