Homework Assignment 2

DS4043, Spring 2022

Due on March 13, 2022 at 11:59 pm

1. Consider the multivariate normal distribution vector $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3)^{\mathrm{T}}$ having mean vector $\boldsymbol{\mu} = (0, 1, 2,)^{\mathrm{T}}$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.5 & 0.5 \\ -0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix}$$

a) Generate 100 random observations from the multivariate normal distribution given above with *set.seed(12)*. (Hint: see *?mvrnorm*) You may need to use the package MASS.

Solution:

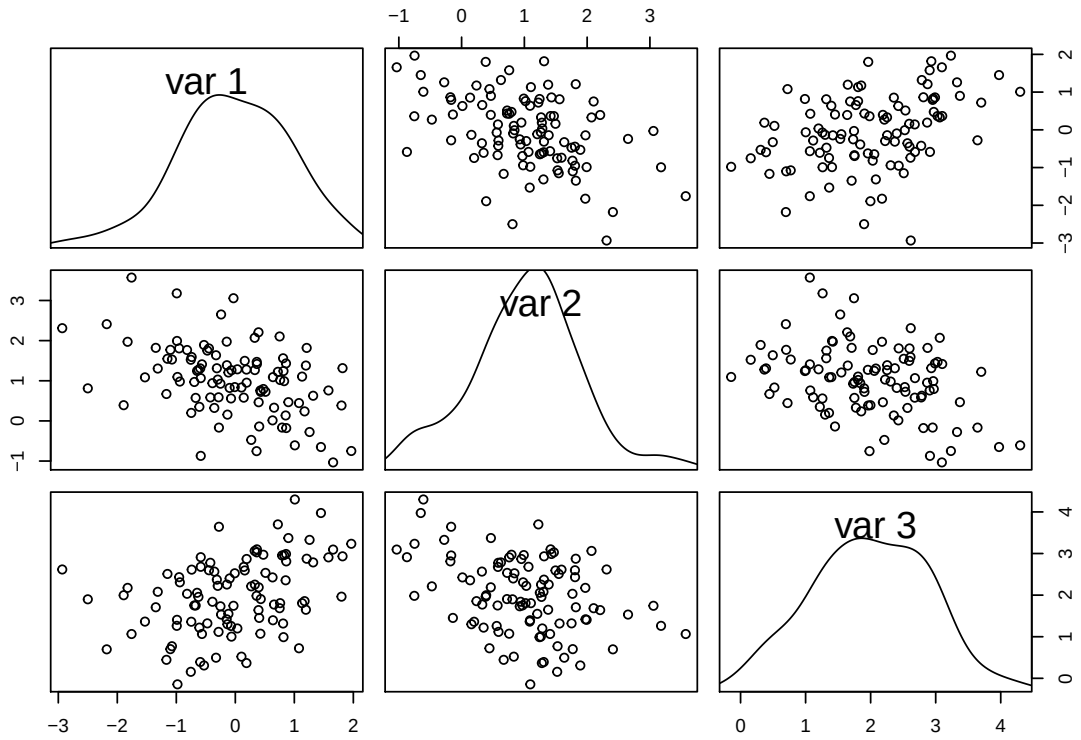
```
library(MASS)
set.seed(12)
mu <- c(0,1,2)
sgv <- c(1,-0.5,0.5,-0.5,1,-0.5,0.5,-0.5,1)
sgm<- matrix(sgv,3,3)
xm <- mvrnorm(100,mu,sgm)
sgm
```

```
##      [,1] [,2] [,3]
## [1,]  1.0 -0.5  0.5
## [2,] -0.5  1.0 -0.5
## [3,]  0.5 -0.5  1.0
```

b) Construct a scatterplot matrix for $\mathbf{X}$ and add a fitted smooth density curve on the diagonal panels for each $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3$ to verify that the location and correlation for each plot agrees with the parameters of the corresponding bivariate distributions.

Solution:

```
panel.d <- function(x, ...) {
  usr <- par("usr"); on.exit(par(usr));
  par(usr = c(usr[1:2], 0, .5))
  lines(density(x))
}
pairs(xm,diag.panel = panel.d)
```



Look at each picture horizontally or vertically. Their data points gather around 0, 1, and 2 which match their corresponding μ values. Also they match their corresponding correlation. For example, picture at row 1 column 2, has correlation -0.5; picture at row 1 column 3, has correlation 0.5; picture at row 2 column 3, has correlation -0.5.

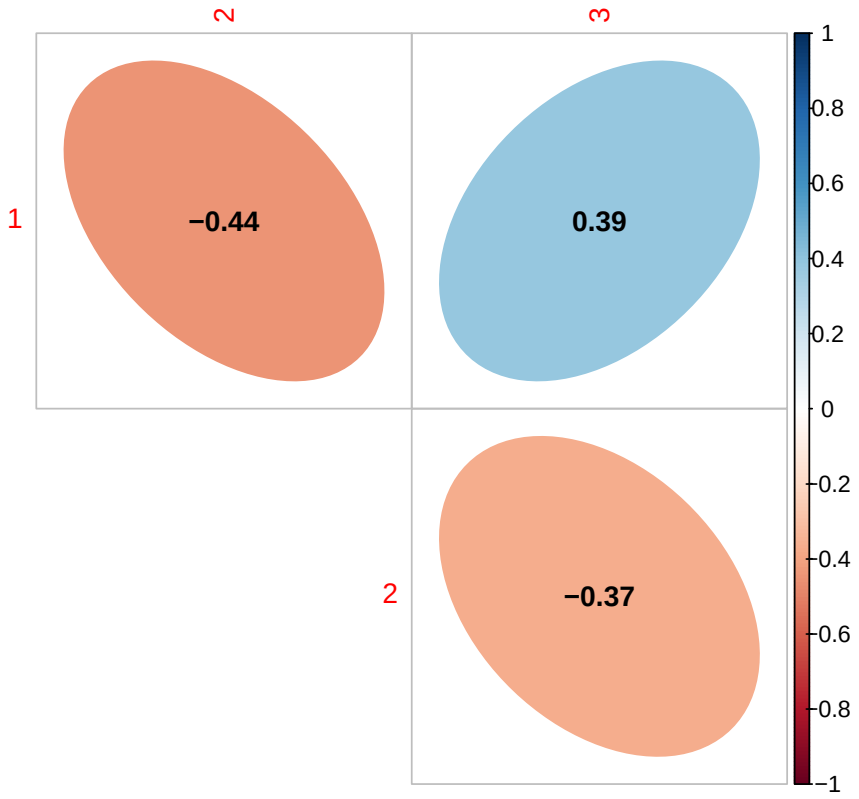
c) Obtain the correlation plot for the generated sample $\mathbf{X}$, where coefficients are added to the plot whose magnitude are presented by different colors. Let the visualization method of correlation matrix to be ellipse.

Solution:

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corMat <- cor(xm);
corrplot(corMat, type = "upper",method = "ellipse", addCoef.col = "black", diag=FALSE)
```



d) Given the covariance matrix Σ, find σ_{x_1}, σ_{x_2} and $\rho_{x_1x_2}$. Consider the joint PDF of bivariate normal distribution

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right\},$$

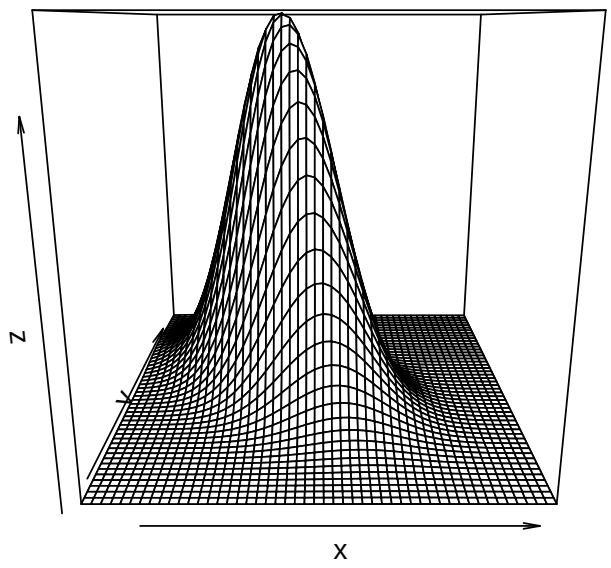
sketch a surface plot for X_1 and X_2, based on their bivariate probability density function. (Hint: if you want to use *curve3d*, please install and use the package *emdbook*)

Solution: From the covariance matrix Σ, we can easily find $\sigma_{x_1} = 1$, $\sigma_{x_2} = 1$ and $\rho_{x_1x_2} = -0.5$.

```
library(emdbook)
```

```
## Warning: package 'emdbook' was built under R version 4.0.5
```

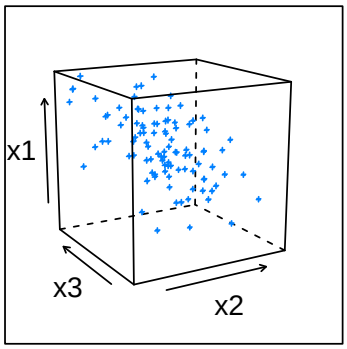
```
sg1 <- 1; sg2 <- 1; ro <- -0.5; mu1 <- 0; mu2<- 1;
f <- function(x,y){
  z <- (1/(2*pi*sg1*sg2*sqrt(1-ro^2))) * exp( -1/(2*(1-ro^2)) *
(((x-mu1)/sg1)^2 + ((y-mu2)/sg2)^2 - 2*ro*(x-mu1)*(y-mu2)/(sg1*sg2)))
}
r <-range(xm)
y <- x <- seq(r[1], r[2], length= 50)
z <- outer(x, y, f);
#curve3d(f, from=c(r[1],r[1]), to=c(r[2],r[2]), sys3d="persp", ticktype="detailed");
# students can also use this one
persp(x, y, z);
```

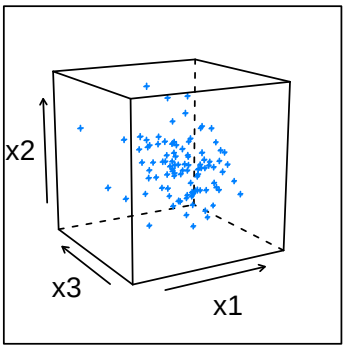


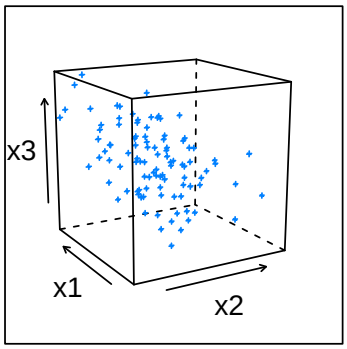
e) Sketch 3-D scatter plots for each of X_1, X_2 and X_3 as a z axis and rest two variables as x and y axes. Put these 3 plots in one picture.

Solution:

```
library(lattice)
x1<-xm[,1];x2<-xm[,2];x3<-xm[,3]
print(cloud(x1~x2+x3,screen = list(z = 30, x = -75, y = 0)), split = c(1, 1, 2, 2), more = TRUE)
print(cloud(x2~x1+x3,screen = list(z = 30, x = -75, y = 0)),split = c(2, 1, 2, 2), more = TRUE)
print(cloud(x3~x2+x1,screen = list(z = 30, x = -75, y = 0)), split = c(1, 2, 2, 2), more = TRUE)
```







2. A continuous random variable X has the probability density function

$$f_X(t) = \begin{cases} at + bt^2 & 0 < t < 1 \\ 0 & \text{otherwise} \end{cases}.$$

If $E[X] = 1/2$, find (a) a and b; (b) $P(X < 1/2)$; (c) $\text{Var}(X)$; (d) Generate the density plot of X

Solution:

(a) we'll first need to find the values of a and b. To do so, we can use the information that

$$\int_{-\infty}^{\infty} f_X(t)\, dt = 1$$

(since X is a cont. R.V.), and information about the expected value of X. So,

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(t)\, dt &= \int_0^1 (at + bt^2)\, dt \\ &= \left[\frac{at^2}{2} + \frac{bt^3}{3}\right]_0^1 \\ &= \frac{a}{2} + \frac{b}{3} \end{aligned}$$

And we find that

$$\frac{a}{2} + \frac{b}{3} = 1.$$

From the problem, we are told $E[X] = 1/2$. In other words,

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} t f_X(t)\,dt = \int_0^1 t(at + bt^2)\,dt \\ &= \int_0^1 (at^2 + bt^3)\,dt = \left[\frac{at^3}{3} + \frac{bt^4}{4}\right]_0^1 \\ &= \frac{a}{3} + \frac{b}{4} \end{aligned}$$

Then,

$$\frac{a}{3} + \frac{b}{4} = 1/2.$$

Solving these two equations:

$$\begin{aligned} \begin{cases} \frac{a}{2} + \frac{b}{3} = 1 \\ \frac{a}{3} + \frac{b}{4} = \frac{1}{2} \end{cases} &\implies \begin{cases} 3a + 2b = 6 \\ 4a + 3b = 6 \end{cases} \\ &\implies \begin{cases} 12a + 8b = 24 \\ 12a + 9b = 18 \end{cases} \\ &\implies -b = 6 \\ &\implies a = 6, b = -6. \end{aligned}$$

(b) To find the $P(X < 1/2)$, We input these values into our pdf for X to get:

$$f_X(t) = \begin{cases} 6t - 6t^2 & 0 < t < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} P(X < 1/2) &= \int_{-\infty}^{1/2} f_X(t)\,dt = \int_0^{1/2} (6t - 6t^2)\,dt \\ &= \left[3t^2 - 2t^3\right]_0^{1/2} = 3(1/2)^2 - 2(1/2)^3 - 0 \\ &= \frac{3}{4} - \frac{2}{8} = 1/2 \end{aligned}$$

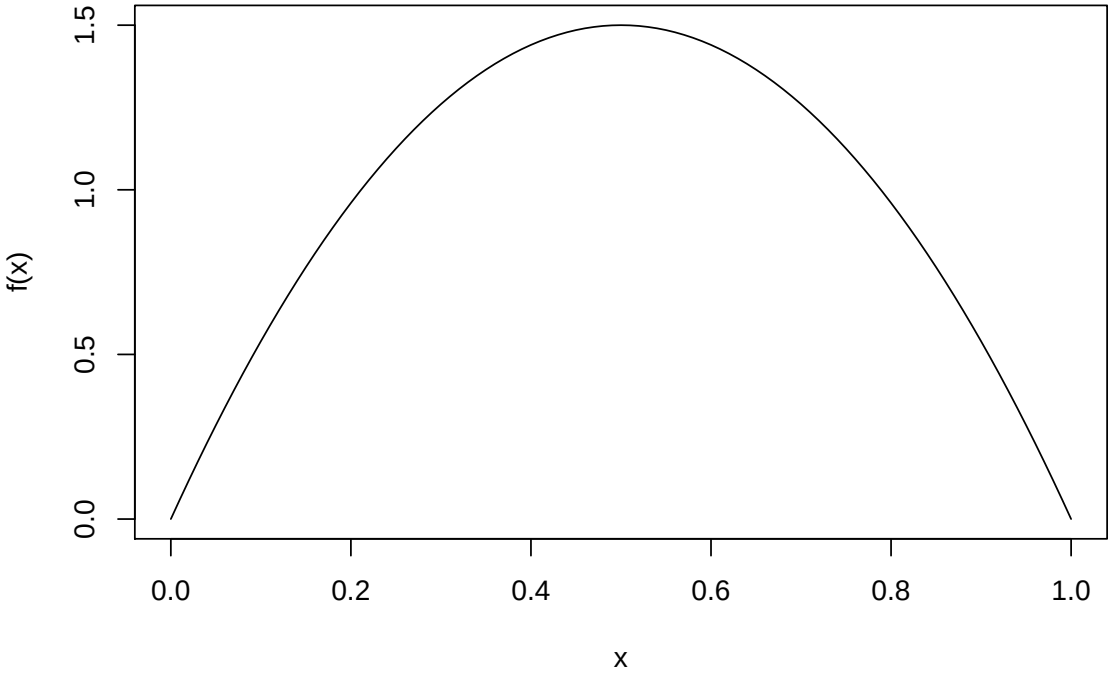
(c) Recall that $\text{Var}(X) = E[X^2] - E[X]^2$. We know $E[X] = 1/2$ from the problem; thus, we know $E[X]^2 = 1/4$. Now, we'll find $E[X^2]$.

$$\begin{aligned} E\left[X^2\right] &= \int_{-\infty}^{\infty} t^2 f_X(t)\,dt = \int_0^1 t^2(6t - 6t^2)\,dt \\ &= \int_0^1 (6t^3 - 6t^4)\,dt = \left[\frac{6t^4}{4} - \frac{6t^5}{5}\right]_0^1 \\ &= \frac{6}{4} - \frac{6}{5} = \frac{3}{10} \end{aligned}$$

Then,

$$\text{Var}(X) = E[X^2] - E[X]^2 = \frac{3}{10} - \frac{1}{4} = \frac{1}{20}.$$

(d)
```
f <- function(x){6*x-6*x^2}
curve(f(x),0,1)  ## note that the density plot should be from 0 to 1.
```



3. Consider a nonparametric regression model

$$y_i = g(x_i) + \epsilon_i, \quad 1 \le i \le n,$$

where y_i's are observations, g is an unknown function, and ϵ_i's are independent and identically distributed random errors with zero mean and variance σ^2. n is the number of observations. Usually one fits the mean function g first and then estimates the variance σ^2 from residual sum of squares $\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\epsilon}_i^2/(n-1)$ where $\hat{\epsilon}_i = y_i - \hat{g}(x_i)$. However this method requires an estimate of the unknown function g. Then some researchers proposed some difference-based estimators which does not require the estimation of g. Assume that x is univariate and $0 \le x_1 \le \cdots \le x_n \le 1$. Rice (1984) proposed the first order difference-based estimator

$$\hat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2$$

Gasser, Sroka and Jennen-Steinmetz (1986) proposed the second order difference based estimator and for equidistant design points (i.e. x_i and x_{i+1} have the same distance for all $i = 1, 2, \ldots, n$), $\hat{\sigma}_{GSJ}^2$ reduces to

$$\hat{\sigma}_{GSJ}^2 = \frac{2}{3(n-2)} \sum_{i=2}^{n-1} \left(\frac{1}{2} y_{i-1} - y_i + \frac{1}{2} y_{i+1} \right)^2.$$

Consider the temperature anomaly dataset. Temperature anomalies in degrees Celsius are based on the new version HadCRUT4 land-sea dataset (Morice et al., 2012). We focus on the global median annual temperature anomalies from 1850 to 2019 relative to the 1961-1990 average. We try to build up the model between time and global median temperature y_i and year x_i.

(a) Use *read.csv* to read the temperature anomaly dataset. Let x be the vector of years from 1850-2019, y be the vector of corresponding global median annual temperature anomalies, and n be the number of observations

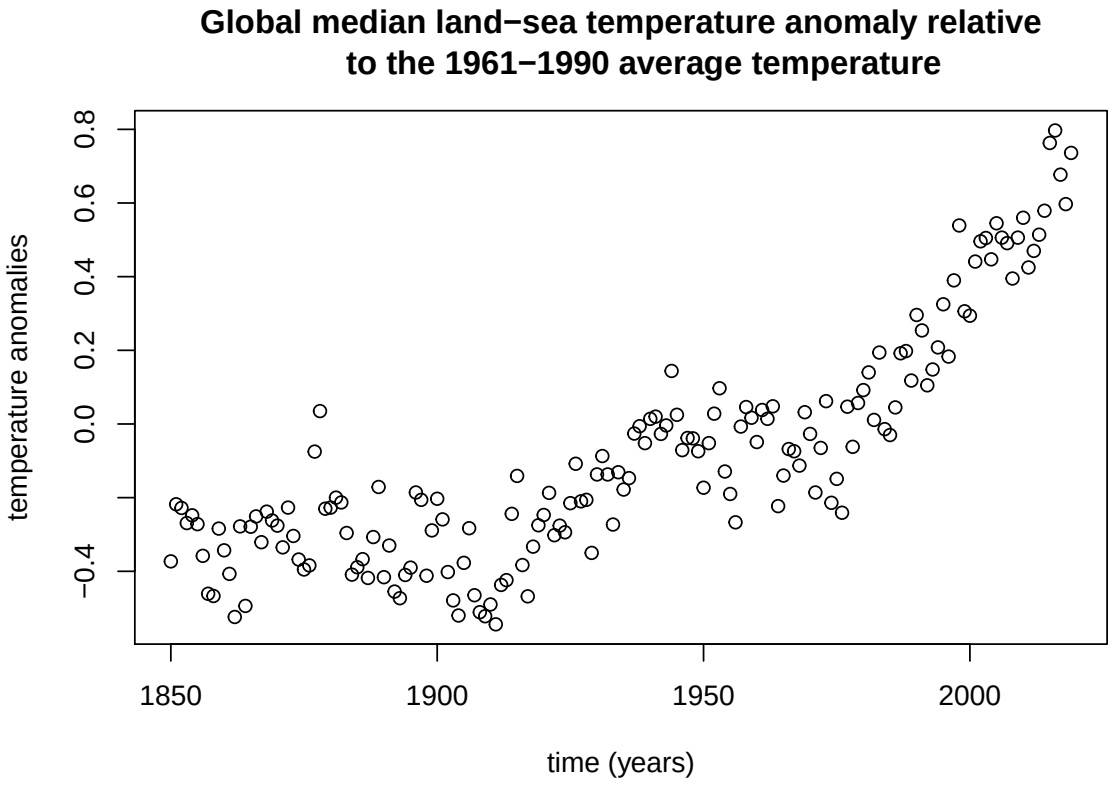
Solution:

```
temp<- read.csv("temperature-anomaly.csv",header=T)
x=temp$Year[1:170]
y=temp$Median[1:170]
n=length(x);
```

(b) Display a scatter plot between global median annual temperature anomalies and years with caption "Global median land-sea temperature anomaly relative to the 1961-1990 average temperature", $x-$label years and $y-$label temperature anomalies.

Solution:

```
plot(x,y,main="Global median land-sea temperature anomaly relative
     to the 1961-1990 average temperature",
     xlab="time (years)", ylab = "temperature anomalies")
```



(c) Change the years x to a new vector x such that $x_i = i/n$. Compute the first order difference-based estimator. (Note: the change of x or not will not affect the computation of the estimator)

Solution:

```
x <- (x-min(x)+1)/n;
sg1 <- sum((y[2:n] - y[1:(n-1)])^2)/(2*(n-1))
sg1
```

```
## [1] 0.006658
```

(d) Compute the second order difference-based estimator.

Solution:

```r
sg2 <- 2*sum((1/2*y[1:(n-2)] - y[2:(n-1)]+1/2* y[3:n])^2)/(3*(n-2))
sg2
```

```
## [1] 0.005406
```