Maestría en Computo Estadístico
Estadística Multivariada
Tarea 1
16 de febrero de 2021
Enrique Santibáñez Cortés
Repositorio de Git: Tarea 1, EM.

Ejercicio 1. Demuestre que la matriz de centrado $\mathbf{P} = \mathbf{I} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{1}\mathbf{1}'$ cumple las siguientes propiedades:

- Tiene rango $(n-1)$, es decir, tiene $n-1$ columnas o renglones linealmente indepenties.

RESPUESTA
Lo haremos por inducción.

Paso 1. Demostrar para algún n. Sea n=2, tenemos que la matriz de centrado es

$$\mathbf{P} = \mathbf{I} - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{1}\mathbf{1}' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{\mathbf{1}}{\mathbf{2}}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Ahora ocupemos eliminación gaussiana para llevar a la matriz $\mathbf{P}$ a su forma escalonada.

$$\begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \overset{R_2 \to R_2 + R_1}{\longrightarrow} \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ 0 & 0 \end{pmatrix}.$$

Por lo tanto, observando la forma escalonada de $\mathbf{P}$ podemos ver que esta tiene solo un pivote diferente de cero por lo que podemos decir, que el rango de $\mathbf{P}$ es 1 $(n-1=2-1=1)$.

Paso 2. Suponemos que se cumple para $n-1$. Es decir, la matriz

$$\mathbf{P} = \mathbf{I} - \frac{\mathbf{1}}{\mathbf{n-1}}\mathbf{1}\mathbf{1}' = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}}_{\mathbf{n-1}} - \frac{\mathbf{1}}{\mathbf{n-1}} \underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}}_{\mathbf{n-1}}$$

tiene rango $n-2$. **Paso 3.** Demostremos que se cumple para n. Tenemos que la matrz de centrado es

$$\mathbf{P} = \mathbf{I} - \frac{\mathbf{1}}{\mathbf{n-1}}\mathbf{1}\mathbf{1}' = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}}_{\mathbf{n}} - \frac{\mathbf{1}}{\mathbf{n}} \underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & \cdots & 1 & 1 \end{pmatrix}}_{\mathbf{n}} = \underbrace{\begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} & 1-\frac{1}{n} \end{pmatrix}}_{\mathbf{n}}.$$

Ahora, llevemos a la matriz $\mathbf{P}$ a su forma escalonada

$$\underbrace{\begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} & 1-\frac{1}{n} \end{pmatrix}}_{n} \overset{R_i \to R_1 - R_i, i=1,\ldots,n}{\longrightarrow} \underbrace{\begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} & -\frac{1}{n} \\ 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \cdots & -1 & 0 \\ 1 & 0 & \cdots & 0 & -1 \end{pmatrix}}_{n} \overset{R_n \leftrightarrow R_2}{\longrightarrow}$$

$$\underbrace{\begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} & -\frac{1}{n} \\ 1 & 0 & \cdots & 0 & -1 \\ 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \cdots & -1 & 0 \end{pmatrix}}_{n} \overset{R_2 \to R_i - R_2, i=3,\ldots,n}{\longrightarrow} \underbrace{\begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} & -\frac{1}{n} \\ 1 & 0 & \cdots & 0 & -1 \\ 0 & -1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}}_{n} \overset{R_1 \to R_1 - R_i/n, i=3,\ldots,n}{\longrightarrow}$$

$$\underbrace{\begin{pmatrix} 1-\frac{1}{n} & 0 & \cdots & 0 & -\frac{n-1}{n} \\ 1 & 0 & \cdots & 0 & -1 \\ 0 & -1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}}_{n} \longrightarrow \underbrace{\begin{pmatrix} 1-\frac{1}{n} & 0 & \cdots & 0 & -1+\frac{1}{n} \\ 1 & 0 & \cdots & 0 & -1 \\ 0 & -1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}}_{n} \overset{R_2 \to R_1 - R2(1-\frac{1}{n})}{\longrightarrow}$$

$$\underbrace{\begin{pmatrix} 1-\frac{1}{n} & 0 & \cdots & 0 & -1+\frac{1}{n} \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}}_{n} \overset{R_2 \leftrightarrow R_n}{\longrightarrow} \underbrace{\begin{pmatrix} 1-\frac{1}{n} & 0 & \cdots & 0 & -1+\frac{1}{n} \\ 0 & -1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}}_{n}.$$

Observando la forma escalonada, tenemos que la matriz tiene $n-1$ pivotes **lo que implica que la rango de la matriz P sea** $n-1$ ■.

- Sus valores propios son 1 o 0.

RESPUESTA
Para este problema ocupemos el siguiente teorema,

Teorema: 1 *Sea A una matriz idempotente si y solo si todos los valores propios son 0 o 1.*

Demostración. Si A es idempotente, para cualquier λ valor propio y V un vector propio correspondiente (distinto de $\mathbf{0}$) entonces

$$\lambda v = Av = AAv = \lambda Av = \lambda^2 v.$$

Entonces, como $v \neq \mathbf{0}$ entonces

$$\lambda - \lambda^2 = \lambda(1-\lambda) = 0.$$

Esto implica que los valores propios sean $\lambda = 0$ o $\lambda = 1$.

Ahora, por lo visto en clase sabemos que la matriz de centradoes simétrica e idempotente (ver diapositiva 23). Entonces podemos **concluir que por ser idempotente esto implica que los valores propios sean 1 o 0** ■.

Ejercicio 2. Dado los siguientes datos:

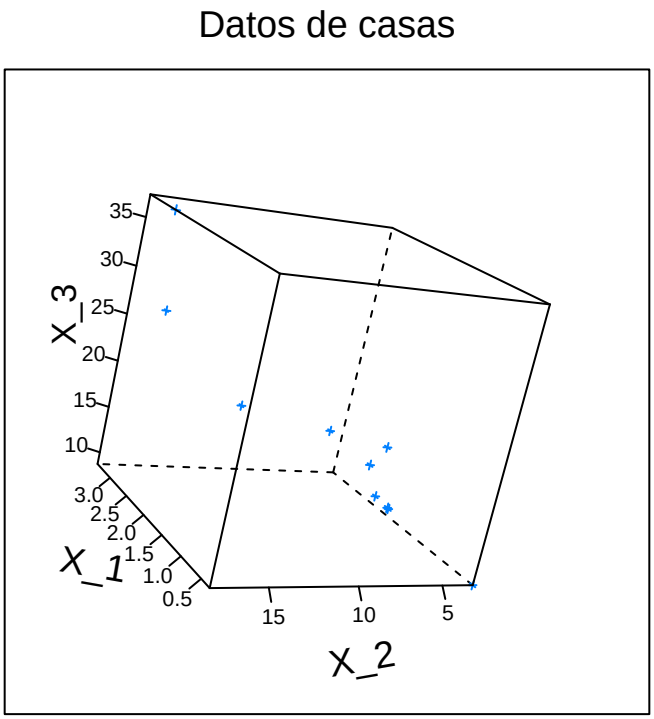
Promotora	X_1 =Duración media hipoteca (años)	X_2 =Precio medio (millones euros)	X_3 =Superficie media (m^2) de cocina
1	8.7	0.3	3.1
2	14.3	0.9	7.4
3	18.9	1.8	9.0
4	19.0	0.8	9.4
5	20.5	0.9	8.3
6	14.7	1.1	7.6
7	18.8	2.5	12.6
8	37.3	2.7	18.1
9	12.6	1.3	5.9
10	25.7	3.4	15.9

a) Dibújese al diagrama de dispersión múltiple y coméntese el aspecto del gráfico.

RESPUESTA

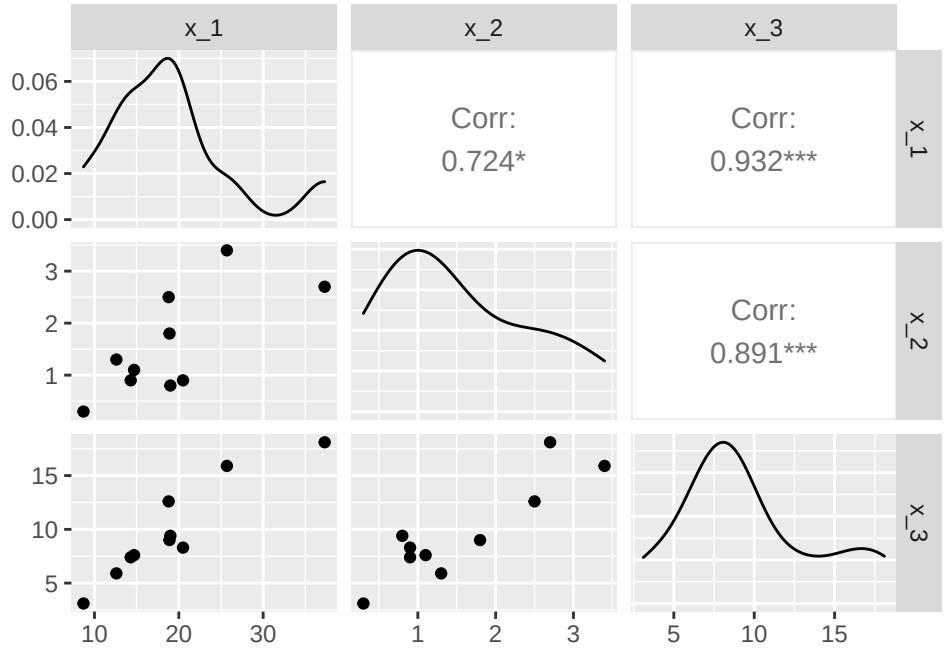
Con ayuda de software R, y la librería 'lattice' gráficamos las tres variables.

```
# cargamos los datos
library(tidyverse)
library(lattice)
x_1 <- c(8.7, 14.3, 18.9, 19.0, 20.5, 14.7, 18.8, 37.3, 12.6, 25.7)
x_2 <- c(0.3, 0.9, 1.8, 0.8, 0.9, 1.1, 2.5, 2.7, 1.3, 3.4)
x_3 <- c(3.1, 7.4, 9.0, 9.4, 8.3, 7.6, 12.6, 18.1, 5.9, 15.9)

# graficamos todas las variables.
cloud(x_1~x_2*x_3, ticktype="detailed", main=expression(paste("Datos de casas")),
    screen=list(z=80,x=-70, y=40), scales=list(arrows=FALSE,col="black",distance=1,cex=.7),
  xlab=list(expression(paste("X_1")),rot=-10,cex=1.2), ylab=list("X_2",rot=10,cex=1.2),
    zlab=list("X_3", rot=90,cex=1.1))
```



Del gráfico anterior podemos observar una relación entre las variables, podría afirmar que están altamente correlacionadas. El gráfico con las tres variables no es tan claro determinar algo, es por eso que graficaremos las variables dos a dos para tener un enfoque más claro.

```r
library(GGally) # cargamos un la libraria para graficar
datos_2 <- data.frame(x_1, x_2, x_3) # creamos un data frame.

# graficamos todas las variables
datos_2 %>% ggpairs(.)
```



El lás gráficas anteriores ya se observa claramente la correlación lineal entre las variables, la variable X_1 y X_2 son las que están más altamente correlacionadas. En términos de los datos, podemos decir que, la duración media de la hipoteca (supongo que de una casa) esta correlacionada (positivamente) con la superficie media de la cocina.

b) Para X_1 y X_2 calcúlanse, respectivamente, las medias muestrales $\bar{x}_1$ y $\bar{x}_2$, las varianzas muestrales s_{11} y s_{22}, la covarianza entre X_1 y X_2, s_{12}, y la correlación entre ambas r_{12}. Interprétrese el valor obtenido de r_{12}.

RESPUESTA

Tenemos que

$$\bar{x}_j = \frac{1}{n}\sum_{j=1}^{n} x_{ij}, \quad s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_j - \bar{x_j})^2, \quad s_{jk} = \frac{1}{n}\sum_{i=1}^{2}(x_{ij} - \bar{x}_j)(x_{jj} - \bar{x}_k) \quad \& \quad r_{jk} = \frac{s_{jk}}{s_j s_k}.$$

Entonces con ayuda de R calculamos los solicitado.

```r
# medias muestrales
bar_x_1 <- mean(x_1)
bar_x_2 <- mean(x_2)

# varianzas muestrales
s_x_1 <- sum((x_1-bar_x_1)^2)/length(x_1)
s_x_2 <- sum((x_2-bar_x_2)^2)/length(x_2)

# covarianza
s_x_1_x_2 <- sum((x_1-bar_x_1)*(x_2-bar_x_2))/length(x_1)
```

```r
# correlacion
r_x_1_x_2 <- s_x_1_x_2/sqrt(s_x_1*s_x_2)
```

Por lo tanto, **las medias muestrales $\bar{x}_1$ =19.05 y $\bar{x}_2$ =1.57, las varianzas muestrales s_{11} =56.9685 y s_{22} =0.8941, la covarianza entre X_1 y X_2, s_{12} =5.1705, y la correlación entre ambas r_{12} =0.7244728. Observamos que el coeficiente de correlación es positivo y a mi parecer es un valor alto. En términos del los datos, podemos decir que la duración media hipoteca esta relacionada al precio medio.**

c) Utilizando la matriz de datos $\mathbf{X}$ y la de centrado $\mathbf{P}$, calcúlense el vector de medias muestrales $\bar{\mathbf{x}}$ y la matriz de covarianzas muestrales $\mathbf{S}$. A partir de ésta obténgase la matriz de correlaciones $\mathbf{R}$.

RESPUESTA

Tenemos que las siguientes igualdades vistas en clases,

$$\bar{\mathbf{x}} = \frac{1}{n}X'1 \quad S = \frac{1}{n}X'PX, \quad \text{donde } P = I - \frac{1}{n}11'$$
$$R = D^{-1/2}SD^{-1/2}, \quad D \text{ es la matriz diagonal con las varianzas.}$$

Entonces ocupando lo anterior, calculemoslo

```r
# datos del problema
X <- matrix(c(x_1,x_2,x_3), nrow = 10) # matriz de los datos
n <- length(x_1) # numero de observaciones
P <- diag(1,10)-matrix(1,nrow=10, ncol=10)/n # matriz P

# calculamos la matriz de medias
x_bar <- t(X)%*%matrix(1,nrow=10, ncol=1)/n
x_bar
```

```
##       [,1]
## [1,] 19.05
## [2,]  1.57
## [3,]  9.73
```

```r
# calculamos la matriz de covarianzas
S <- (t(X)%*%P%*%X)/n
S
```

```
##          [,1]   [,2]    [,3]
## [1,] 56.9685 5.1705 30.4775
## [2,]  5.1705 0.8941  3.6479
## [3,] 30.4775 3.6479 18.7641
```

```r
# calculamos la matriz de correlaciones
R <- solve(diag(sqrt(diag(S))))%*%S%*%solve(diag(sqrt(diag(S))))
R
```

```
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.7244728 0.9321763
## [2,] 0.7244728 1.0000000 0.8906068
## [3,] 0.9321763 0.8906068 1.0000000
```

Ejercicio 3. Considérese la muestra $\mathbf{x_1}, \cdots, \mathbf{x_n}$ de vectores de . Pruébese que la matriz de covarianzas $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})'$, se puede expresar como $\frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i}\mathbf{x_i'} - \bar{\mathbf{x}}\bar{\mathbf{x}}'$.

RESPUESTA
Utilizando solo propiedades vectoriales, tenemos

$$\begin{aligned}\mathbf{S} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{x_i}-\bar{\mathbf{x}})' &= \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{x_i}-\bar{\mathbf{x}})\mathbf{x_i'} - \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=n}}^{\mathbf{n}}(\mathbf{x_i}-\bar{\mathbf{x}})\bar{\mathbf{x}}' \\ &= \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}\mathbf{x_i}\mathbf{x_i'} - \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}\bar{\mathbf{x}}\mathbf{x_i'} + \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}\mathbf{x_i}\bar{\mathbf{x}}' - \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}\bar{\mathbf{x}}\bar{\mathbf{x}}' \\ &= \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}\mathbf{x_i}\mathbf{x_i'} - \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}\bar{\mathbf{x}}\bar{\mathbf{x}}' \\ &= \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}\mathbf{x_i}\mathbf{x_i'} - \bar{\mathbf{x}}\bar{\mathbf{x}}'. \quad \blacksquare\end{aligned}$$

Ejercicio 4. Considere una población normal bivariada con $\mu_1 = 0, \mu_2 = 2, \sigma_{11} = 2, \sigma_{22} = 1$, y $\rho_{12} = 0{,}5$.

a) Escriba la densidad normal bivariada explicitamente.

RESPUESTA
Tenemos que la función de densidad normal p−dimensional es (Johnson and Wichern 2007):

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

donde $-\infty \leq x_i \leq \infty, \quad i = 1, \cdots, p$. Entonces del problema tenemos que

$$\mu = \begin{pmatrix}0\\2\end{pmatrix}, \quad \sigma_{\mathbf{12}} = \sigma_{\mathbf{21}} = \rho_{\mathbf{12}}\sqrt{\sigma_{\mathbf{11}}}\sqrt{\sigma_{\mathbf{22}}} = \frac{\sqrt{\mathbf{2}}}{\mathbf{2}} \Rightarrow \mathbf{\Sigma} = \begin{pmatrix}2 & \frac{\sqrt{2}}{2}\\ \frac{\sqrt{2}}{2} & 1\end{pmatrix}.$$

Ahora calculemos el determinante y la inversa de la matriz de covarianza $\mathbf{\Sigma}$,

$$\begin{aligned}|\mathbf{\Sigma}| &= 2 - \frac{2}{4} = \frac{3}{2} \Rightarrow |\Sigma|^{1/2} = \sqrt{\frac{3}{2}}. \\ \mathbf{\Sigma^{-1}} &= \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{21}^2}\begin{pmatrix}\sigma_{22} & -\sigma_{12}\\ -\sigma_{21} & \sigma_{11}\end{pmatrix} = \frac{2}{3}\begin{pmatrix}1 & -\frac{\sqrt{2}}{2}\\ -\frac{\sqrt{2}}{2} & 2\end{pmatrix} = \begin{pmatrix}\frac{2}{3} & -\frac{\sqrt{2}}{3}\\ -\frac{\sqrt{2}}{3} & \frac{4}{3}\end{pmatrix}.\end{aligned}$$

Con lo anterior podemos concluir **que la función de densidad normal bivariada es**

$$\begin{aligned}f(x) &= \frac{\sqrt{2}}{2\pi\sqrt{3}} e^{-\begin{pmatrix}x_1 & x_2-2\end{pmatrix}\begin{pmatrix}\frac{2}{3} & -\frac{\sqrt{2}}{3}\\ -\frac{\sqrt{2}}{3} & \frac{4}{3}\end{pmatrix}\begin{pmatrix}x_1\\ x_2-2\end{pmatrix}/2} \\ &= \frac{1}{\pi\sqrt{6}} e^{-\begin{pmatrix}x_1 & x_2-2\end{pmatrix}\begin{pmatrix}\frac{2}{3}x_1 - \frac{\sqrt{2}}{3}(x_2-2)\\ -\frac{\sqrt{2}}{3}x_1 + \frac{4}{3}(x_2-2)\end{pmatrix}/2} \\ &= \frac{1}{\pi\sqrt{6}} e^{-\begin{pmatrix}x_1 & x_2-2\end{pmatrix}\begin{pmatrix}\frac{1}{3}x_1 - \frac{1}{3\sqrt{2}}x_2 + \frac{1}{6\sqrt{2}}\\ -\frac{1}{3\sqrt{2}}x_1 + \frac{2}{3}x_2 - \frac{4}{3})\end{pmatrix}} \\ &= \frac{1}{\pi\sqrt{6}} e^{-\left(x_1\left(\frac{1}{3}x_1 - \frac{1}{3\sqrt{2}}x_2 + \frac{1}{6\sqrt{2}}\right) + (x_2-2)\left(-\frac{1}{3\sqrt{2}}x_1 + \frac{2}{3}x_2 - \frac{4}{3}\right)\right)}.\end{aligned}$$

De forma explícita nosé si se tenía que desarrollar más.

b) Escriba la expresión de distancia cuadrada generalizada $(\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)$ como función de x_1 y x_2.

RESPUESTA

Sabemos que la distancia cuadrada generalizada se puede escribir como (Johnson and Wichern 2007)

$$(\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu) = \frac{1}{1-\rho_{12}^2}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]$$

Entonces para este problema la distancia cuadrada generalizada se escribe en función de x_1 y x_2 de la sigueinte manera

$$\begin{aligned}(\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu) &= \frac{1}{1-\rho_{12}^2}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right] \\ &= \frac{1}{1-0{,}5^2}\left[\left(\frac{x_1}{\sqrt{2}}\right)^2 + (x_2-2)^2 - \left(\frac{x_1}{\sqrt{2}}\right)(x_2-2)\right] \\ &= \frac{4}{3}\left[\frac{x_1^2}{2} + (x_2-2)^2 - \frac{x_1(x_2-2)}{\sqrt{2}}\right].\end{aligned}$$

De igual manera, podemos reescribir la distribución calculada en el inciso a) de la siguiente forma

$$\begin{aligned}f(x) = \cdots &= \frac{1}{\pi\sqrt{6}}e^{-\left(x_1\left(\frac{1}{3}x_1-\frac{1}{3\sqrt{2}}x_2+\frac{1}{6\sqrt{2}}\right)+(x_2-2)\left(-\frac{1}{3\sqrt{2}}x_1+\frac{2}{3}x_2-\frac{4}{3}\right)\right)} \\ &= \frac{1}{\pi\sqrt{6}}e^{-\frac{2}{3}\left[\frac{x_1^2}{2}+(x_2-2)^2-\frac{x_1(x_2-2)}{\sqrt{2}}\right]}.\end{aligned}$$

c) Determine y grafique el contorno de densidad constante que contiene el 50 % de la probabilidad.

RESPUESTA

Sabemos que la distancia generalizada tiene una distribución conocida, en concreto, $(\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu) \sim \chi^{\mathbf{2}}_{\mathbf{p}}$ con p grados de libertad. Entonces el elipsoide sólido de valores de x que satisface

$$(\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu) \leq \chi^{\mathbf{2}}_{\mathbf{p}}(\alpha)$$

tiene una probabilidad $1-\alpha$ donde $\chi^2_p(\alpha)$ denota el percentil superior $(100\alpha)\,\%$ de la distribución χ^2_p. Ahora, si asumimos $\alpha = 0{,}5$ se espera que el 50 % de los datos estén contenidos dentro del contorno estimado del 50 %. Tenemos que $\chi^2_p(0{,}5) = 1{,}3863$ (ocupando R 'qchisq(0.5,2)'), entonces buscamos el elipsoide que cumple

$$(\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu) \leq \chi^{\mathbf{2}}_{\mathbf{p}}(\mathbf{0{,}5}) = \mathbf{1{,}3863} = (\mathbf{1{,}774})^{\mathbf{2}}$$

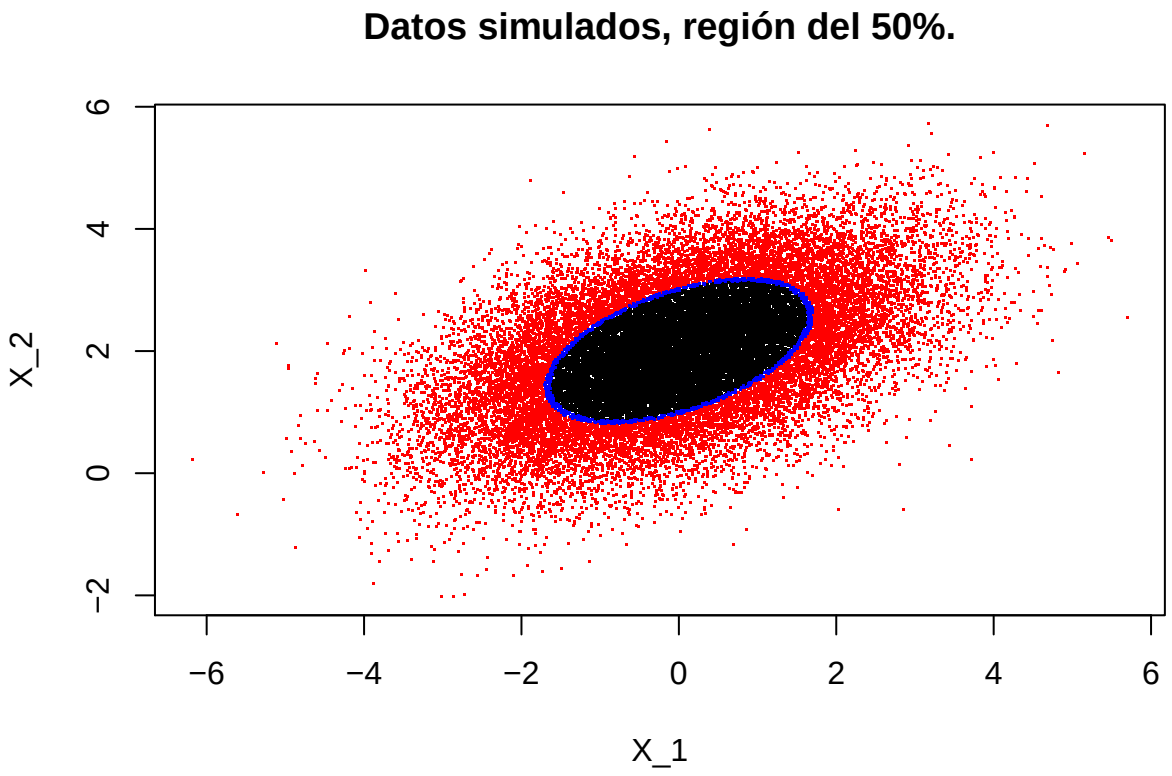
Entonces, para graficar el contorno del elipsoide ahora calculemos los valores propios de la matriz de covarianzas y vectores propios correspondientes a los valores propios

```
sigma_4 <- matrix(c(2,1/sqrt(2), 1/sqrt(2), 1), 2) #matriz de covarianzas
eigen(sigma_4) # valores y vectores propios
```

```
## eigen() decomposition
## $values
## [1] 2.3660254 0.6339746
##
## $vectors
##            [,1]       [,2]
## [1,] -0.8880738  0.4597008
## [2,] -0.4597008 -0.8880738
```

Entonces los vectores propios son $(\lambda_1, \lambda_2) = 2{,}36602540{,}6339746$ y sus vectores propios $\begin{pmatrix} e_1 & e_2 \end{pmatrix} = \begin{pmatrix} -0{,}8880738 & 0{,}4597008 \\ -0{,}4597008 & -0{,}8880738 \end{pmatrix}$. Por lo tanto, los ejes del elipsoide que se espera cumpla que el 50 % de los datos estén obtenidos en este son $c\sqrt{(\lambda_1)} = 1{,}1774 * \sqrt{2{,}36602} = 1{,}81106$ y $c\sqrt{(\lambda_1)} = 1{,}1774 * \sqrt{0{,}6339} = 0{,}9375$. Entonces simulemos puntos de esta distribución bivariada y posteriormente graficamos la elipsoide encontrada, se ocupara la libreria ´MASS´.

```r
# cargamos las librerias y los datos del problema.
library(MASS)
n <- 30000
mu <- c(0, 2)
sigma <- sigma_4

# simulamos datos.
datos_4 <- mvrnorm(n, mu, sigma)
delta <- 0.06 # delta.
# calculamos las distancias
distancias <- mahalanobis(datos_4, colMeans(datos_4), cov(datos_4))

# graficamos los puntos y la región de rechazo.
plot(datos_4, pch=".", xlab="X_1", ylab="X_2", main="Datos simulados, región del 50%.")
points(datos_4[distancias>1.3863,], pch='.', col='red')
points(datos_4[(1.3863-delta)<distancias & distancias<(1.3863+delta),], pch='.', col='blue')
```



Los puntos rojos son los puntos que se encuentran fuera de la región del 50 %, la linea azul es el límite de la región que contiene al 50 % de los datos (se ocupo un $\delta = 0{,}6$ para resaltar el límite) y llos puntos negros son los que se encuentran dentro de la región. Otra manera de gráficar la región es utilizando los párametros (centro, eje menor y mayor) de la elipsoide, pero con simulación fue más sencillo de hacer la

gráfica.

d) Especifique la distribución condicional de X_1 dado que $X_2 = x_2$.

RESPUESTA
Ocupemos la siguiente propiedad de las distribuciones normales multivariadas (Johnson and Wichern 2007).

Teorema: 2 *(Ver página 181) Sea* $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p(\mu, \mathbf{\Sigma})$, *con* $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \mathbf{\Sigma} = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right), |\mathbf{\Sigma_{22}}| > \mathbf{0}$. *Entonces la distribución condidicional de* X_1 *dado* $X_2 = x_2$, $f(X_1|X_2)$ *es normal multivariada con*

$$\begin{aligned} \textit{Media} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \textit{Covarianza} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

Entonces para este problema tenemos $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p(\mu, \mathbf{\Sigma})$, con $\mu = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \mathbf{\Sigma} = \left(\begin{array}{c|c} 2 & \frac{1}{\sqrt{2}} \\ \hline \frac{1}{\sqrt{2}} & 1 \end{array}\right), |\mathbf{\Sigma_{22}}| = \mathbf{1} > \mathbf{0}$. Entonces la distribución condidicional de X_1 dado $X_2 = x_2$, $f(X_1|X_2)$ es normal multivariada con

$$\begin{aligned} \text{Media} &= \frac{1}{\sqrt{2}}(x_2 - 2) = \frac{x_2 - 2}{\sqrt{2}} \\ \text{Covarianza} &= 2 - \frac{1}{\sqrt{2}}\frac{1}{\sqrt{2})} = \frac{3}{2}. \end{aligned}$$

Es decir, $\mathbf{X_1}|\mathbf{X_2} = \mathbf{x_2} \sim \mathbf{N}\left(\frac{\mathbf{x_2 - 2}}{\sqrt{\mathbf{2}}}, \frac{\mathbf{3}}{\mathbf{2}}\right)$ ■.

Ejercicio 5. Sea X un vector aleatorio de distribución normal con media $\mu = \begin{pmatrix} -1 & 1 & 0 \end{pmatrix}'$ y matriz de covarianza

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 2 \end{pmatrix} \tag{1}$$

a) Hállese la distribución de $X_1 + 2X_2 - 3X_3$.

RESPUESTA
Por lo visto en clase (Johnson and Wichern 2007), tenemos la siguiente propiedad

Teorema: 3 *(Ver página 177) Si* $\mathbf{x} \sim \mathbf{N_p}(\mu, \mathbf{\Sigma})$, *entonces cualquier combinación lineal*

$$\mathbf{a'x} = \sum_{\mathbf{i=1}}^{\mathbf{p}} \mathbf{a_i x_i} \sim \mathbf{N}(\mathbf{a'}\mu, \mathbf{a'\Sigma a}).$$

Además, si $\mathbf{a'x} = \sum_{\mathbf{i=1}}^{\mathbf{p}} \mathbf{a_i x_i} \sim \mathbf{N}(\mathbf{a'}\mu, \mathbf{a'\Sigma a})$ *para cada* $\mathbf{a}$, *entonces* $\mathbf{x} \sim \mathbf{N_p}(\mu, \mathbf{\Sigma})$.

Entonces ocupando lo anterior, estamos buscando la combinación lineal

$$Y = X_1 + 2X_2 - 3X_3 = \begin{pmatrix} 1 & 2 & -3 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix},$$

entonces tenemos que $\mathbf{a} = \begin{pmatrix} 1 & 2 & -3 \end{pmatrix}'$. Ahora calculemos los parametros de la distribución

$$\mathbf{a}'\mu = \begin{pmatrix} 1 & 2 & -3 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = -1 + 2 = 1$$

$$\mathbf{a}'\mathbf{\Sigma}\mathbf{a} = \begin{pmatrix} 1 & 2 & -3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & -3 \end{pmatrix} \begin{pmatrix} -2 \\ 3 \\ -3 \end{pmatrix} = 13.$$

Por lo tanto, $\mathbf{X_1 + 2X_2 - 3X_3 \sim N(-1, 13)}$.

b) Hállese un vector $\mathbf{a}_{(\mathbf{2}\times\mathbf{1})}$ tal que las variables X_1 y $\mathbf{X_1} - \mathbf{a}' \begin{pmatrix} X_2 \\ X_3 \end{pmatrix}$ sean independientes.

RESPUESTA
Ocupemos la siguiente propiedad de las distribuciones multivariadas normales (Johnson and Wichern 2007).

Teorema: 4 *(Ver página 180)*

a) Si $X_1 \sim N_{p_1}(\mu_1, \sigma_{11})$ *y* $X_2 \sim N_{p_2}(\mu_2, \Sigma_{22})$ *son independientes, entonces* $Cov(X_1, X_2) = \Sigma_{12} = 0$.

b) Si

$$\begin{pmatrix} X_1 \\ \hline X_2 \end{pmatrix} \sim N_{p_1+p_2}\left(\begin{pmatrix} \mu_1 \\ \hline \mu_2 \end{pmatrix}, \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array} \right) \right)$$

entonces X_1 *y* X_2 *son independientes si y solo si* $\Sigma_{12} = 0$.

Tenemos la variable X_1 y sea $Y = \mathbf{X_1} - \mathbf{a}' \begin{pmatrix} X_2 \\ X_3 \end{pmatrix} = \mathbf{X_1} - \mathbf{a_1 X_2} - \mathbf{a_3 X_3}$, entonces tenemos la matrix $A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & -a_1 & -a_2 \end{pmatrix}$. Por lo que ocupando el teorema 3 tenemos

$$AX = \begin{pmatrix} 1 & 0 & 0 \\ 1 & -a_1 & -a_2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_1 - a_1 X_2 - a_2 X_3 \end{pmatrix} \sim N_2\left(A\mu, A\Sigma A^T\right).$$

Ahora, calculemos explicitamente la matriz de covarianza

$$\begin{aligned} A\Sigma A^T &= \begin{pmatrix} 1 & 0 & 0 \\ 1 & -a_1 & -a_2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & -a_1 \\ 0 & -a_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & -a_1 & -a_2 \end{pmatrix} \begin{pmatrix} 1 & 1 - a_2 \\ 0 & -3a_1 - a_2 \\ 1 & 1 - a_1 - 2a_2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 - a_2 \\ 1 - a_2 & 1 - a_2 - a_1(-3a_1 - a_2) - a_2(1 - a_1 - 2a_2) \end{pmatrix}. \end{aligned}$$

Por lo tanto, **ocupando el teorema 4 podemos decir que** X_1 **y** $Y = \mathbf{X_1} - \mathbf{a}' \begin{pmatrix} X_2 \\ X_3 \end{pmatrix} = \mathbf{X_1} - \mathbf{a_1 X_2} - \mathbf{a_3 X_3}$ **son independientes si solo si** $a' = \begin{pmatrix} a_1 & 1 \end{pmatrix}, \quad \forall a_1 \in \mathbb{R}$.

c) Calcúlese la distribución de X_3 condicionada a $X_1 = x_1$ y $X_2 = x_2$.

RESPUESTA

Tenemos $\mu_1 = 0, \mu_2 = \begin{pmatrix} -1 & 1 \end{pmatrix}', \Sigma_{11} = 2, \Sigma_{22} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}, \Sigma_{12} = \begin{pmatrix} 1 & 1 \end{pmatrix}'$. Entonces ocupando el teorema 2 sabemos que la distribución de X_3 condicionada a $X_1 = x_1,\ X_2 = x_2$ es normal con

$$\begin{aligned}
\text{Media } &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) = \begin{pmatrix} -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}^{-1} \begin{pmatrix} x_1 + 1 \\ x_2 - 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} x_1 + 1 \\ x_2 - 1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} x_1 + 1 \\ x_2 - 1 \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} x_1 + 1 \\ x_2 - 1 \end{pmatrix} = x_1 + 1 + \frac{x_2 - 1}{3} = x_1 + \frac{x_2}{3} + \frac{2}{3}. \\
\text{Covarianza } &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 2 - \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2 - \begin{pmatrix} 1 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&= 2 - \left(1 + \frac{1}{3} \right) = \frac{2}{3}.
\end{aligned}$$

Es decir, **podemos concluir que** $X_3 | X_1 = x_1, X_2 = x_2 \sim N\left(x_1 + \frac{x_2}{3} + \frac{2}{3}, \frac{2}{3} \right)$ ■.

Nota: En algunos ejercicios no se desarrollo la función de distribución por que no se especificaba (aunque no se si era correcto), solo se encontraba la distribución y los parámetros de esta.

Bibliografía

Johnson, Richard Arnold, and Dean W. Wichern. 2007. *Applied Multivariate Statistical Analysis*. 6. ed. Upper Saddle River, NJ: Prentice Hall. http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+330798693&sourceid=fbw_bibsonomy.