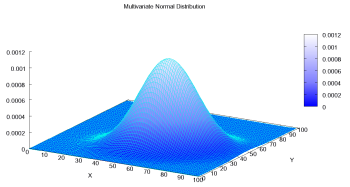
Estadística Aplicada III

Distribución normal multivariada

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México



Semana 3

Distribución normal multivariada

Distribución normal multivariada I

- Este tema se basará en los capítulos 4 a 6 de Johnson y Wichern, y en los capítulos 3 y 5 de Mardia, Kent y Bibby. Un enfoque más teórico se puede encontrar en los capítulos 2 a 10 del libro de Anderson.
- El tema de normalidad multivariada es fundamental en varios contextos de la estadística. Para no ir muy lejos, baste considerar:
  - Diseño de experimentos
  - Análisis de regresión
  - Muchos de los métodos multivariados que veremos requieren normalidad para poder llevar a cabo inferencia.
  - Teoría clásica de portafolios basados en la teoría de Harry Markowitz
  - Muchos modelos en Economía y Finanzas se basan en la distribución normal
- En lo que sigue, se seguirá el camino análogo al que se sigue en el caso del análisis univariado de la distribución normal.

Distribución normal multivariada II

Def (Distribución normal multivariada)

Un vector aleatorio $\mathbf{x}$ tiene una *distribución multinormal* o *normal multivariada* con media $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$ que se denota por $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, si su densidad es:

$$f(\mathbf{x}) = \det(2\pi\boldsymbol{\Sigma})^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

donde $\boldsymbol{\Sigma}$ es una matriz definida positiva ($\boldsymbol{\Sigma} > \mathbf{0}$).

Notas:

- El determinante también se puede escribir como: $|2\pi\boldsymbol{\Sigma}|^{-1/2} = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}$. Entonces depende explícitamente de la varianza generalizada.
- $E(\mathbf{X}) = \boldsymbol{\mu}$, y $\text{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$.

Distribución normal multivariada III

- En particular, en el caso de la densidad bivariada, se puede escribir explícitamente en términos de correlación:

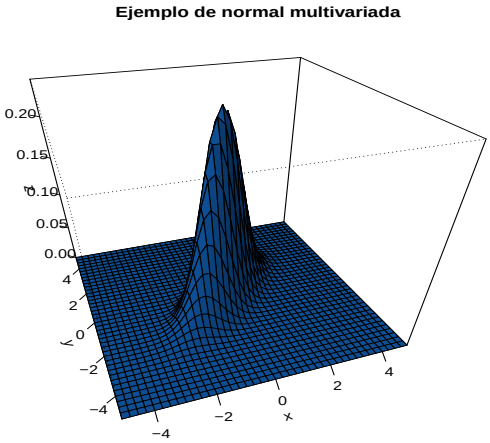
$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{X_1-\mu_1}{\sigma_1}\frac{X_2-\mu_2}{\sigma_2} + \left(\frac{X_2-\mu_2}{\sigma_2}\right)^2\right\}\right]$$

donde $\boldsymbol{\Sigma}$ se puede escribir como $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

- Entonces la normal bivariada depende de 5 parámetros: $\mu_1, \mu_2, \sigma_1, \sigma_2$ y ρ.

```
library(mnormt) # funciones de normal multivariada, hasta dimensión 20
x <- y <- seq(-5,5,0.25) # partición del dominio
    mu <- c(0,0)
sigma <- matrix(c(1, 0.75, 0.75, 1), nrow = 2) # correlación =0.75, mismas varianzas unitarias
f <- function(x, y) dmnorm(cbind(x,y),mu,sigma)
z <- outer(x, y, f)
mycolor <- rgb(colorr[1],colorr[2],colorr[3], maxColorValue = 255) # Asigna un color (igual a la presentación)
persp(x, y, z, theta = -20, phi = 35, ticktype = "detailed", col = mycolor, shade = 0.1, main = "Ejemplo de normal multivariada")
```

Distribución normal multivariada IV



Distribución normal multivariada estándar

Al igual que en el caso univariado, un vector normal se puede estandarizar.

Teorema

Si $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces $\mathbf{y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$

Demostración.

Noten que $\mathbf{y}'\mathbf{y} = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ Si consideramos el jacobiano correspondiente de la transformación $T(\mathbf{y}) = \boldsymbol{\Sigma}^{1/2}\mathbf{y} + \boldsymbol{\mu} = \mathbf{x}$ se tiene $|J| = |\boldsymbol{\Sigma}|^{\frac{1}{2}}$. Por lo tanto, la densidad queda:

$$g(\mathbf{y}) = f(T(\mathbf{y})) \cdot |J| = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{\sum_{i=1}^{p} y_i^2}{2}\right) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2}\mathbf{y}'\mathbf{y}\right)$$

que corresponde a la densidad de la normal estándar. □

Geometría de la distribución multinormal I

Las *curvas de nivel*, es decir, los valores de $\mathbf{x}$ de la función de densidad multinormal donde se mantiene constante corresponden al lugar geométrico definido por

$$\mathcal{C}_k = \left\{\mathbf{x} \in \mathbb{R}^p | (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = k^2\right\}$$

con k constante. Estos contornos $\mathcal{C}_k$ corresponden a elipsoides de igual concentración.

- Si $\boldsymbol{\mu} = \mathbf{0}$ y $\boldsymbol{\Sigma} = \mathbf{I}$, los contornos corresponden a hiperesferas alrededor del origen.
- Si consideramos la descomposición espectral de $\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$ con $\mathbf{P} = [\mathbf{e}_1|\cdots|\mathbf{e}_p]$, la *rotación a componentes principales* está dada por $\mathbf{y} = \mathbf{P}'(\mathbf{x}-\boldsymbol{\mu})$. Bajo esta transformación, los contornos se pueden expresar:

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})'\mathbf{P}\boldsymbol{\Lambda}^{-1}\mathbf{P}'(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{y}'\boldsymbol{\Lambda}^{-1}\mathbf{y} = \sum_{i=1}^{p}\frac{y_i^2}{\lambda_i} = k^2$$

  En el espacio correspondiente a esta transformación, las componentes principales de $\mathbf{y}$ son los ejes del elipsoide.

Ejemplo I

Considerando los siguientes valores: $\boldsymbol{\Sigma} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ y $\boldsymbol{\mu} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$, los contornos se muestran a continuación junto con las direcciones dadas por las direcciones correspondientes a la transformación de las componentes principales.

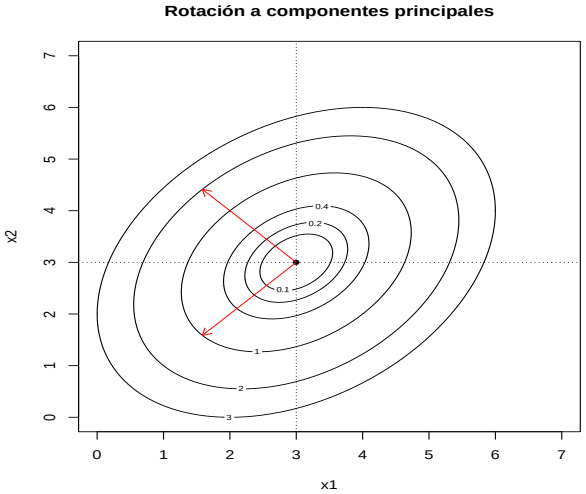
```
Sinv <- solve(matrix(c(3,1,1,3),ncol=2)) #inversa de la matriz de covarianzas
mu <- c(3,3) #vector media
DE <- eigen(Sinv) #descomposición espectral de Sinv

#función del lugar geométrico y version para vectorizar
elipse <- function(x, y, mu, Sigma){as.numeric((c(x,y) - mu) %*% Sigma %*% (c(x,y) - mu))}
elipse2 <- function(x, y)apply(cbind(x, y), 1, function(x)elipse(x[1], x[2], mu = mu, Sigma = Sinv))

x <- seq(-1, 10, 0.01)
par(pty = "s")  #Para hacer una gráfica cuadrada (no se distorsionen las escalas)
contour(x, x, outer(x, x, elipse2), levels = c(0.1, 0.2, 0.4, 1, 2, 3), xlim = c(0,7), ylim = c(0,7),
        main = "Rotación a componentes principales", xlab = "x1", ylab = "x2")
points(mu[1], mu[2], pch = 16, cex = 0.9)

#Ejes en la dirección de las componentes principales (dibuja las flechas):
s <- 2 #factor de escala
arrows(x0 = mu[1], y0 = mu[2],
       x1 = mu[1] + s*DE$vectors[1,1], y1 = mu[2] + s*DE$vectors[2,1], length = 0.1, col = "red")
arrows(x0 = mu[1], y0 = mu[2],
       x1 = mu[1] + s*DE$vectors[1,2], y1 = mu[2] + s*DE$vectors[2,2], length = 0.1, col = "red")
abline(h = 3, v = 3, lty = 3)
```

Ejemplo II


Rotación a componentes principales
x2
x1
0
1
2
3
4
5
6
7
0.4
0.2
0.1
1
2
3

Distribución de la forma cuadrática

Con ayuda de la transformación de componentes principales, podemos calcular la distribución de la forma cuadrática $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$:

Teorema

Si $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces $\mathbf{u} = (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi^2_{(p)}$

Demostración.

Sea $\mathbf{y} = \mathbf{A}'(\mathbf{x}-\mu)$, donde $\mathbf{A}' = \mathbf{P}\boldsymbol{\Lambda}^{-1/2}$. Como $\mathbf{y}$ es una transformación lineal de $\mathbf{x}$, $\mathbf{y}$ tiene distribución $\mathcal{N}_p(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.
Por lo tanto $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \boldsymbol{\Lambda}^{-1/2}\mathbf{P}'\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'\mathbf{P}\boldsymbol{\Lambda}^{-1/2} = \mathbf{I}$.
Entonces $\mathbf{u} = \mathbf{y}'\mathbf{y} = \sum_{i=1}^{p} y_i^2$ con $y_i \sim \mathcal{N}(0,1)$ entonces $y_i^2 \sim \chi^2_{(1)}$ y como las p variables son independientes, $\sum_{i=1}^{p} y_i^2 \sim \chi^2_{(p)}$. □

Función característica de la multinormal I

En algunos resultados utilizaremos la función característica del vector **x**.

Teorema

Si $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, su función característica está dada por:

$$\phi_{\mathbf{x}}(\mathbf{t}) = \exp\{i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\}$$

donde $\mathbf{t} \in \mathbb{R}^p$ y $i^2 = -1$.

Demostración.

Función característica de la multinormal II

Si $\mathbf{y} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ entonces $\mathbf{x} = \boldsymbol{\Sigma}^{1/2}\mathbf{y} + \boldsymbol{\mu}$. Entonces por definición:

$$\begin{aligned}
\phi_{\mathbf{x}}(\mathbf{t}) &= E[e^{i\mathbf{t}'\mathbf{x}}] = E[e^{i\mathbf{t}'(\boldsymbol{\Sigma}^{1/2}\mathbf{y}+\boldsymbol{\mu})}] \\
&= e^{i\mathbf{t}'\boldsymbol{\mu}} E[e^{i\mathbf{t}'\boldsymbol{\Sigma}^{1/2}\mathbf{y}}] \\
&= e^{i\mathbf{t}'\boldsymbol{\mu}} E[e^{i\mathbf{u}'\mathbf{y}}], \text{ donde } \mathbf{u} = \boldsymbol{\Sigma}^{1/2}\mathbf{t}
\end{aligned}$$

Como $y_j \sim \mathcal{N}(0,1)$ para $j = 1, 2, \ldots, p$ y estas variables son mutuamente independientes,

$$E[e^{i\mathbf{u}'\mathbf{y}}] = \prod_{j=1}^{p} \phi_{y_j}(u_j) = e^{-\sum_{j=1}^{p} u_j^2/2} = e^{-\mathbf{u}'\mathbf{u}/2}$$

ya que la función característica de la normal estándar está dada por $\phi_y(t) = e^{-t^2/2}$. Entonces:

$$\phi_{\mathbf{x}}(\mathbf{t}) = e^{i\mathbf{t}'\boldsymbol{\mu}} e^{-\mathbf{u}'\mathbf{u}/2} = e^{i\mathbf{t}'\boldsymbol{\mu}' - (\mathbf{t}'\boldsymbol{\Sigma}^{1/2})(\boldsymbol{\Sigma}^{1/2}\mathbf{t})/2} = e^{i\mathbf{t}'\boldsymbol{\mu}' - \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2}$$

□

Caracterización de una distribución multivariada a través de combinaciones lineales I

- Trabajar directamente con densidades multivariadas puede ser excesivamente demandante en términos de notación y herramientas analíticas.
- Para simplificar el trabajo sin tener que escribir directamente las densidades, el siguiente resultado, basado en la función característica, nos permite trabajar exclusivamente en términos de combinaciones lineales.
- El siguiente teorema no sólo aplica a la distribución normal, sino a otras distribuciones como la Cauchy, t multivariada, etc., siempre y cuando las distribuciones multivariadas existan.

Caracterización de una distribución multivariada a través de combinaciones lineales II

Teorema (Cramér-Wold)

La distribución de un vector aleatorio $\mathbf{x}$ está completamente determinada por el conjunto de todas las distribuciones unidimensionales de combinaciones lineales $\mathbf{t}'\mathbf{x}$, para $\mathbf{t} \in \mathbb{R}^p$. Entonces la distribución normal multivariada queda completamente definida especificando la distribución de todas sus combinaciones lineales.

Demostración.

Sea $y = \mathbf{t}'\mathbf{x}$ y consideremos su función característica: $\phi_y(s) = E[e^{isy}] = E[e^{is\mathbf{t}'\mathbf{x}}]$. Para $s = 1$, podemos obtener la función característica de $\mathbf{x}$:

$$\phi_y(1) = E[e^{i\mathbf{t}'\mathbf{x}}] = \phi_{\mathbf{x}}(\mathbf{t})$$

Por lo tanto la función característica de $\mathbf{x}$ está completamente determinada a partir de la combinación lineal $y = \mathbf{t}'\mathbf{x}$.

Caracterización de una distribución multivariada a través de combinaciones lineales III

- A partir del teorema de Cramér-Wold se puede redefinir un vector normal sin hacer referencia directa a la función de densidad o de distribución, y nos permitirá agilizar algunas demostraciones.
- Podemos redefinir un vector normal del siguiente modo:

vector normal multivariado

Un vector $\mathbf{x}$ sigue una distribución normal multivariada $\iff \mathbf{a}'\mathbf{x} = \sum_{i=1}^{p} a_i x_i$ es una variable aleatoria normal univariada $\forall \mathbf{a} \in \mathbb{R}^p$.

- Una interpretación geométrica a partir de la definición anterior es que $\mathbf{a}'\mathbf{x}$ son proyecciones en un subespacio unidimensional que son normales univariadas con media $\mathbf{a}'\boldsymbol{\mu}$ y varianza $\mathbf{a}'\Sigma\mathbf{a}$.

Caracterización de una distribución multivariada a través de combinaciones lineales IV

- En general podemos tomar q combinaciones lineales simultáneamente. En muchos de las situaciones que se verán más adelante, el paso importante será encontrar qué combinaciones lineales son relevantes para el modelo o resultado en consideración.
- A continuación veremos algunos teoremas que se pueden extender a matrices.

Teorema

Si $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ entonces $\underset{q\times 1}{\mathbf{y}} = \underset{q\times p}{\mathbf{A}}\,\mathbf{x} + \underset{q\times 1}{\mathbf{c}} \sim \mathcal{N}_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.

Caracterización de una distribución multivariada a través de combinaciones lineales V

Demostración.

Si $\mathbf{b} \in \mathbb{R}^q$, entonces $\mathbf{b}'\mathbf{y} = \mathbf{b}'\mathbf{A}\mathbf{x} + \mathbf{b}'\mathbf{c}$. Tomando $\mathbf{a} = \mathbf{A}\mathbf{b}$ y $\mathbf{d} = \mathbf{b}'\mathbf{c}$, tenemos que $\mathbf{b}'\mathbf{y} = \mathbf{a}'\mathbf{x} + \mathbf{d}$. Como $\mathbf{a}'\mathbf{x}$ es normal univariada, $\mathbf{b}'\mathbf{y}$ también es normal univariada y por la definición en términos de combinaciones lineales, $\mathbf{y}$ sigue una distribución normal multivariada. □

Resultados relativos a componentes y particiones de vectores multinormales I

Los siguientes resultados serán utilizados con frecuencia. Las demostraciones están basadas en tomar las combinaciones lineales correctas.

Teorema

Todos los subconjuntos de un vector normal $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ son normales: si se considera el vector particionado $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$, respectivamente $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$ y $\boldsymbol{\Sigma} = \begin{pmatrix} \underset{(p_1 \times p_1)}{\boldsymbol{\Sigma}_{11}} & \underset{(p_1 \times p_2)}{\boldsymbol{\Sigma}_{12}} \\ \underset{(p_2 \times p_1)}{\boldsymbol{\Sigma}_{21}} & \underset{(p_2 \times p_2)}{\boldsymbol{\Sigma}_{22}} \end{pmatrix}$, entonces $\mathbf{x}_i \sim \mathcal{N}_{p_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii})$, $i = 1, 2$, donde $p_1 + p_2 = p$.

Resultados relativos a componentes y particiones de vectores multinormales II

Teorema

1. Si $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2$ son vectores aleatorios de dimensiones p_1 y p_2 respectivamente, entonces $\text{cov}(\mathbf{x}_1, \mathbf{x}_2) = \underset{p_1 \times p_2}{\mathbf{0}}$
2. Si $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim \mathcal{N}_{p_1+p_2}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$ entonces $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2$ si y sólo si $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21} = \mathbf{0}$.
3. Si $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2$ y $\mathbf{x}_1 \sim \mathcal{N}_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ y $\mathbf{x}_2 \sim \mathcal{N}_{p_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, entonces el vector agregado $\mathbf{x}_1 = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim \mathcal{N}_{p_1+p_2}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$.

Distribuciones normales condicionales I

La distribución condicional de un vector normal con respecto a un subconjunto de componentes de ese vector juega un papel muy importante en los modelos lineales en general. El siguiente resultado es fundamental para el análisis de regresión multivariado.

Teorema

Si $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim \mathcal{N}_p \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$ donde $|\boldsymbol{\Sigma}_{22}| > 0$, entonces la distribución condicional $\mathbf{x}_1 | \mathbf{x}_2$ es normal con media $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ y varianza $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Demostración.

Noten que para cualquier matriz **B**,

Distribuciones normales condicionales II

$$\begin{bmatrix} \mathbf{I} & -\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{B}' & \mathbf{I} \end{bmatrix} =$$

$$\begin{bmatrix} \Sigma_{11} - \Sigma_{12}\mathbf{B}' - \mathbf{B}\Sigma_{21} + \mathbf{B}\Sigma_{22}\mathbf{B}' & \Sigma_{12} - \mathbf{B}\Sigma_{22} \\ \Sigma_{21} - \Sigma_{22}\mathbf{B}' & \Sigma_{22} \end{bmatrix}$$

En particular, si tomamos $\mathbf{B} = \Sigma_{12}\Sigma_{22}^{-1}$, entonces obtenemos:

$$\begin{pmatrix} \mathbf{x}_1 - \mathbf{B}\mathbf{x}_2 \\ \mathbf{x}_2 \end{pmatrix} \sim \mathcal{N}_p \left(\begin{pmatrix} \boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{bmatrix} \right)$$

Como $\mathbf{x}_1 - \mathbf{B}\mathbf{x}_2 \perp\!\!\!\perp \mathbf{x}_2$ entonces en distribución, $\mathbf{x}_1 - \mathbf{B}\mathbf{x}_2 | \mathbf{x}_2 = \mathbf{x}_1 - \mathbf{B}\mathbf{x}_2$, por lo que

$$\mathbf{x}_1 - \mathbf{B}\mathbf{x}_2 | \mathbf{x}_2 \sim \mathcal{N}_{p_1} \left(\boldsymbol{\mu}_1 - \mathbf{B}\boldsymbol{\mu}_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right)$$

Distribuciones normales condicionales III

y condicional a $\mathbf{x}_2$, podemos pasar la constante al otro lado para obtener que:

$$\mathbf{x}_1|\mathbf{x}_2 \sim \mathcal{N}_{p_1}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

□

Estimación

Verosimilitud y suficiencia I

Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una muestra aleatoria de una distribución multivariada con densidad $f(\mathbf{x}, \boldsymbol{\theta})$ donde $\boldsymbol{\theta}$ es un vector de parámetros.

Def

La función de verosimilitud es la densidad conjunta de la muestra aleatoria como función de $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i, \boldsymbol{\theta})$$

y la función de log-verosimilitud esta dada por:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\{f(\mathbf{x}_i, \boldsymbol{\theta})\}$$

Verosimilitud y suficiencia II

En el caso normal, con $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, las funciones de verosimilitud y log-verosimilitud son, respectivamente:

$$
\begin{aligned}
L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= |2\pi\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right\} \\
l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{n}{2} \log |2\pi\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})
\end{aligned}
$$

Otra representación de la (log-) verosimilitud I

La parte más importante de la (log-) verosimilitud es la forma cuadrática que aparece en ella, que es donde aparece la muestra. Se puede reescribir de otra manera:

$$\begin{aligned}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) &= (\mathbf{x}_i-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\boldsymbol{\mu}) \\ &= (\mathbf{x}_i-\bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})+(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \\ &\quad -2(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})\end{aligned}$$

el último término de la suma se anula:

$$\sum_{i=1}^{n} 2(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}) = 2(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\left[\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})\right] = 2(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\mathbf{0} = 0$$

entonces (la parte amarilla se justificará adelante):

Otra representación de la (log-) verosimilitud II

$$
\begin{aligned}
\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) &= \sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})+\sum_{i=1}^{n}(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \\
&= \sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})+n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \\
&= tr\left[\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'\right]+n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \\
&= tr\left[\boldsymbol{\Sigma}^{-1}n\mathbf{S}_n\right]+n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})
\end{aligned}
$$

Así que finalmente:

$$
l(\boldsymbol{\mu},\boldsymbol{\Sigma})=-\frac{n}{2}\log|2\pi\boldsymbol{\Sigma}|-\frac{n}{2}tr\left[\boldsymbol{\Sigma}^{-1}n\mathbf{S}_n\right]-\frac{n}{2}(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})
$$

Para justificar la parte marcada de amarillo, tenemos el siguiente resultado:

Otra representación de la (log-) verosimilitud III

Teorema

Si $\mathbf{A}_{k\times k}$ es una matriz simétrica y $\mathbf{x} \in \mathbb{R}^k$, entonces:

1. $\mathbf{x}'\mathbf{A}\mathbf{x} = tr\{\mathbf{x}'\mathbf{A}\mathbf{x}\} = tr\{\mathbf{A}\mathbf{x}\mathbf{x}'\}$
2. $tr\{\mathbf{A}\} = \sum_{i=1}^{k} \lambda_i$, donde $\lambda_i \in eigen(\mathbf{A})$.

Demostración.

Para la parte 1, hay que recordar que para cualesquiera matrices $\mathbf{B}_{k\times l}$ y $\mathbf{C}_{l\times k}$, el elemento j de la diagonal se obtiene multiplicando el renglón j de $\mathbf{B}$ por la columna j de $\mathbf{C}$. Así que $tr(\mathbf{BC}) = \sum_{j=1}^{l} \sum_{i=1}^{k} b_{ji}c_{ij}$. El mismo argumento se obtiene intercambiando los sumandos para obtener el elemento j de $\mathbf{CB}$. □

Función score e información de Fisher I

Def (Información de Fisher)

La derivada de la función de log-verosimilitud $l(\boldsymbol{\theta})$ se conoce como la *función score*:

$$s(\mathbf{x}, \boldsymbol{\theta}) = s(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{L(\boldsymbol{\theta})} \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

En particular, la función score, como función de la muestra, se puede ver como una variable aleatoria. A la matriz de covarianzas de $s(\boldsymbol{\theta})$ se le llama *matriz de información de Fisher* y se denota por $\mathbf{F} = \text{Var}(s(\mathbf{x}, \boldsymbol{\theta}))$.

Observaciones:

Función score e información de Fisher II

i. $\int f(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} = 1$ Por lo tanto, bajo condiciones de regularidad $\frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} = 0 = \int \frac{\partial f(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\, d\mathbf{x}$. Ahora bien, como $\frac{\partial \log(f(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{1}{f(\mathbf{x}|\boldsymbol{\theta})} \frac{\partial f(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, entonces:

$$\begin{aligned} \int \frac{\partial f(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\, dx &= \int \frac{\partial \log(f(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} f(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} \\ &= E[s(\mathbf{x}, \boldsymbol{\theta})] \end{aligned}$$

para θ fijo.

ii. Lo anterior implica que $\text{Var}(s(\mathbf{x}, \boldsymbol{\theta})) = E[s^2(\mathbf{x}, \boldsymbol{\theta})] = E\left[-\frac{\partial^2}{\partial \boldsymbol{\theta}^2} f(\mathbf{x}|\boldsymbol{\theta})\right]$.

Más adelante veremos varias propiedades de la información de Fisher que son relevantes en la estimación y que nos darán información relevante sobre los estimadores de los parámetros.

Estimadores máximo verosímiles de μ y Σ I

Para obtener los estimadores máximo verosímiles en el caso multivariado, se requiere el siguiente lema:

Lema

Dada una matriz simétrica $p \times p$ definida positiva **B** y un escalar $b > 0$, se cumple la siguiente desigualdad:

$$\frac{1}{|\mathbf{\Sigma}|^b} \exp\left(-tr(\mathbf{\Sigma}^{-1}\mathbf{B})/2\right) \leq \frac{1}{|\mathbf{B}|^b}(2b)^{pb} \exp(-bp)$$

para todas las matrices definidas positivas $\mathbf{\Sigma}_{p\times p}$, y la igualdad se alcanza solo para $\mathbf{\Sigma} = (1/2b)\mathbf{B}$

Demostración.

Estimadores máximo verosímiles de μ y Σ II

Como $\mathbf{B}$ es cuadrada y simétrica definida positiva, $\exists \quad \mathbf{B}^{1/2}$. Entonces

$$tr(\mathbf{\Sigma}^{-1}\mathbf{B}) = tr((\mathbf{\Sigma}^{-1}\mathbf{B}^{1/2})\mathbf{B}^{1/2}) = tr(\mathbf{B}^{1/2}(\mathbf{\Sigma}^{-1}\mathbf{B}^{1/2}))$$

Como esta matriz es definida positiva, ya que

$$\mathbf{y}'\mathbf{B}^{1/2}\mathbf{\Sigma}^{-1}\mathbf{B}^{1/2}\mathbf{y} = (\mathbf{B}^{1/2}\mathbf{y})'\mathbf{\Sigma}^{-1}(\mathbf{B}^{1/2}\mathbf{y}) > 0 \text{ si } \mathbf{y} \neq \mathbf{0},$$

sus eigenvalores η_i son positivos, y $tr(\mathbf{\Sigma}^{-1}\mathbf{B}) = tr(\mathbf{B}^{1/2}\mathbf{\Sigma}^{-1}\mathbf{B}^{1/2}) = \sum_{i=1}^{p}\eta_i$ y $|\mathbf{B}^{1/2}\mathbf{\Sigma}^{-1}\mathbf{B}^{1/2}| = \prod_{i=1}^{p}\eta_i$.
Ahora bien, por las propiedades de determinantes:

$$\begin{aligned} |\mathbf{B}^{1/2}\mathbf{\Sigma}^{-1}\mathbf{B}^{1/2}| &= |\mathbf{B}^{1/2}||\mathbf{\Sigma}^{-1}||\mathbf{B}^{1/2}| = |\mathbf{\Sigma}^{-1}||\mathbf{B}^{1/2}||\mathbf{B}^{1/2}| \\ &= |\mathbf{\Sigma}^{-1}||\mathbf{B}| \\ &= \frac{1}{|\mathbf{\Sigma}|}|\mathbf{B}| \end{aligned}$$

Estimadores máximo verosímiles de μ y Σ III

Es decir: $\frac{1}{|\boldsymbol{\Sigma}|} = \frac{\prod_{i=1}^{p} \eta_i}{|\mathbf{B}|}$ Multiplicando ambos lados de la ecuación por $e^{-tr(\boldsymbol{\Sigma}^{-1}\mathbf{B})/2} = e^{-\sum_{i=1}^{p} \eta_i/2}$ y elevando a la b, se obtiene:

$$\frac{1}{|\boldsymbol{\Sigma}|^b} e^{-tr(\boldsymbol{\Sigma}^{-1}\mathbf{B})/2} = \frac{\left(\prod_{i=1}^{p} \eta_i\right)^b}{|\mathbf{B}|^b} e^{-\sum_{i=1}^{p} \eta_i/2} = \frac{1}{|\mathbf{B}|^b} \prod_{i=1}^{p} \eta_i^b e^{-\eta_i/2}$$

La función $h(\eta) = \eta^b e^{-\eta/2}$ tiene un máximo en $\eta = 2b$ de $h(2b) = (2b)^b e^{-b}$. Por lo tanto, en este máximo

$$\frac{1}{|\boldsymbol{\Sigma}|^b} e^{-tr(\boldsymbol{\Sigma}^{-1}\mathbf{B})/2} \leq \frac{1}{|\mathbf{B}|^b} (2b)^{pb} e^{-bp}.$$

Por último, noten que para la elección $\boldsymbol{\Sigma} = (1/2b)\mathbf{B}$,

$$\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2} = \mathbf{B}^{1/2}(2b)\mathbf{B}^{-1}\mathbf{B}^{1/2} = (2b)\mathbf{I}$$

y $tr(\boldsymbol{\Sigma}^{-1}\mathbf{B}) = tr(\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}) = tr(2b\mathbf{I}) = 2bp$ y del mismo modo $\frac{1}{|\boldsymbol{\Sigma}|} = \frac{(2b)^p}{|\mathbf{B}|}$. Sustituyendo estas expresiones en la desigualdad se obtiene la igualdad.

Estimadores máximo verosímiles de μ y Σ IV

Los estimadores máximo verosímiles de μ y Σ son los valores $\hat{\mu}$ y $\hat{\Sigma}$ que maximizan la verosimilitud.

Teorema

Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una muestra aleatoria de una normal multivariada con media μ y covarianza Σ.
Entonces $\hat{\mu} = \bar{\mathbf{x}}$ y $\hat{\Sigma} = \frac{n-1}{n}\mathbf{S}$ son los estimadores máximo verosímiles de μ y Σ respectivamente.

Demostración.

El exponente en la función de verosimilitud es:

$$-\frac{n}{2}\left(\log|2\pi\Sigma| + tr\left[\Sigma^{-1}n\mathbf{S}\right] + (\bar{\mathbf{x}}-\mu)'\Sigma^{-1}(\bar{\mathbf{x}}-\mu)\right)$$

Estimadores máximo verosímiles de μ y Σ V

Como $\Sigma^{-1} > \mathbf{0}$, entonces la forma cuadrática $(\bar{\mathbf{x}} - \boldsymbol{\mu})'\Sigma^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) > 0$ a menos que $\boldsymbol{\mu} = \bar{\mathbf{x}}$. Por lo tanto la verosimilitud se maximiza con respecto a $\boldsymbol{\mu}$ en $\bar{\mathbf{x}}$.
Entonces nos queda la expresión $l(\bar{\mathbf{x}}, \boldsymbol{\Sigma}) = -\frac{n}{2}\left(\log|2\pi\boldsymbol{\Sigma}| + tr\left[\boldsymbol{\Sigma}^{-1}n\mathbf{S}\right]\right)$, que si tomamos exponenciales de ambos lados, obtenemos la forma de la expresión del lema:

$$L(\bar{\mathbf{x}}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-n/2}\exp[-\frac{n}{2}tr(\boldsymbol{\Sigma}^{-1}n\mathbf{S})]$$

Para ver el máximo con respecto a Σ, por el lema anterior con $b = n/2$ y $\mathbf{B} = \mathbf{S}$, el máximo se alcanza en $\hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\mathbf{S}$.

□

- Recordar que los estimadores máximo verosímiles (EMV) poseen la propiedad de *invarianza*: si θ es un parámetro y $\hat{\theta}$ su EMV, entonces el EMV de una función del parámetro $h(\theta)$ es $\widehat{h(\theta)} = h(\hat{\theta})$.
- En particular, la expresión $\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$ tiene EMV: $\bar{\mathbf{x}}\mathbf{S}^{-1}\bar{\mathbf{x}}$.

Estimadores máximo verosímiles de μ y Σ VI

- También $\bar{\mathbf{x}}$ y $\mathbf{S}$ son *estadísticas suficientes*, ya que la densidad conjunta depende de la muestra sólo a través de $\bar{\mathbf{x}}$ y $\mathbf{S}$. Esto significa que toda la información sobre $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ está contenida en esos dos estimadores, independientemente del tamaño n de la muestra.
- El punto anterior no es cierto para poblaciones que no son normales en general. Por eso es importante verificar el supuesto de normalidad multivariada antes de extraer conclusiones sólo de los parámetros $\bar{\mathbf{x}}$ y $\mathbf{S}$.