Estadística Aplicada III

Normalidad multivariada

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México

Semana 4: 5/7 de septiembre de 2018

Distribución normal multivariada

Distribución normal multivariada

Def (Distribución normal multivariada)

Un vector aleatorio $\mathbf{X}$ tiene una *distribución multinormal* o *normal multivariada* con media $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$ si su densidad es:

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

donde $\boldsymbol{\Sigma}$ es definida positiva ($\boldsymbol{\Sigma} > \mathbf{0}$) y se denota $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Notas:

- El determinante también se puede escribir como: $|2\pi\boldsymbol{\Sigma}|^{-1/2} = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}$
- $E(\mathbf{X}) = \boldsymbol{\mu}$, y $\text{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$
- En particular, la densidad bivariada se puede escribir explícitamente en términos de correlación como:

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

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

Distribución normal multivariada estándar

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:

$$g(\mathbf{y}) = f(T(\mathbf{y}))|J| = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{\sum_{i=1}^{p} y_i^2}{2}\right)$$

□

Geometría de la distribución multinormal I

Los valores de $\mathbf{x}$ de la función de densidad multinormal donde la multinormal es constante corresponden al lugar geométrico definido por

$$\{\mathbf{x}|(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=k^2\}$$

con k constante. Estos contornos 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 *transformación de componentes principales* está dada por $\mathbf{y}=\mathbf{P}'(\mathbf{x}-\boldsymbol{\mu})$. Bajo esta transformación, los contornos se pueden expresar como:

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

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

Ejemplo I

Si $\boldsymbol{\Sigma} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ 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.

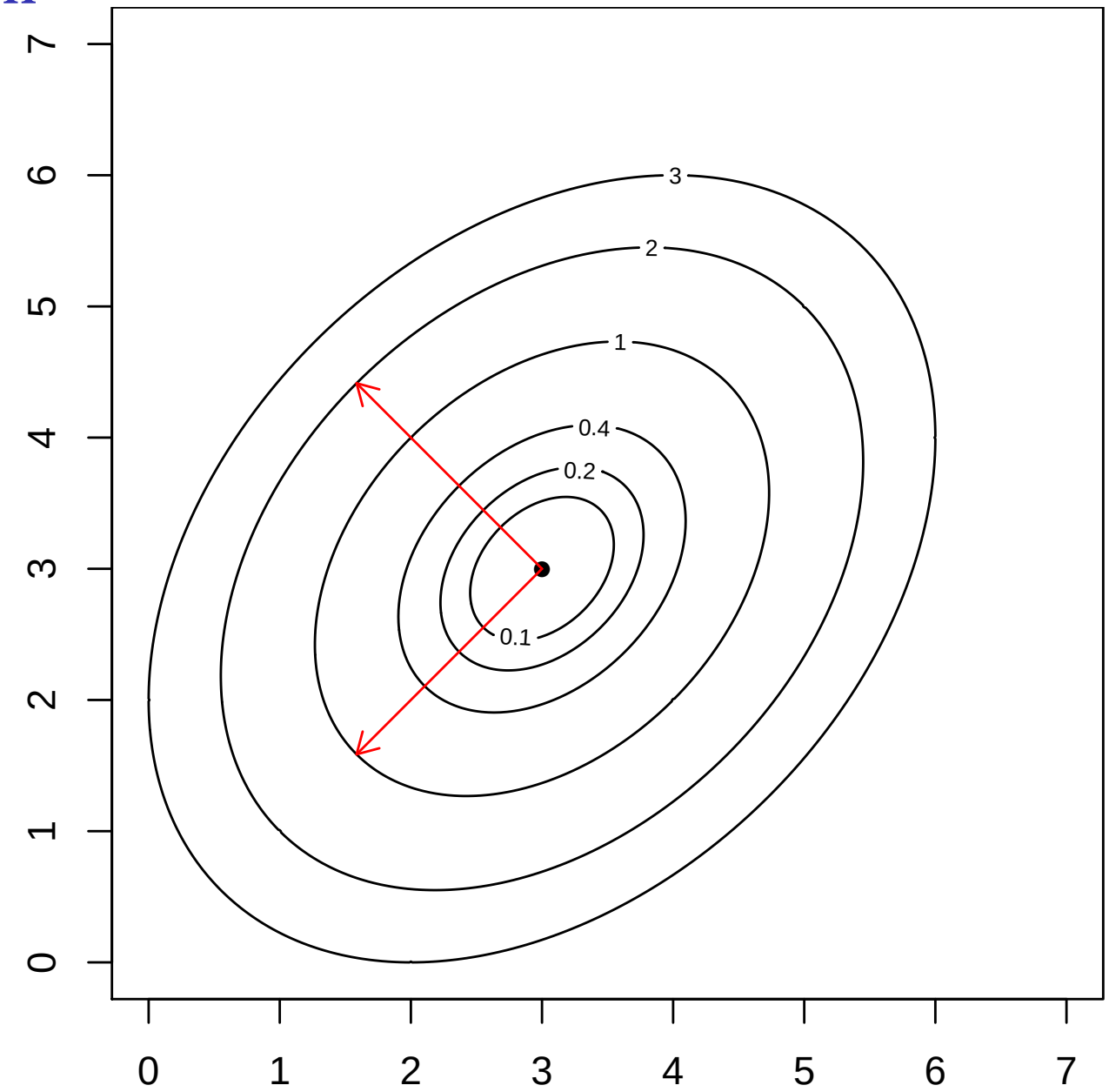
```
Sinv <- solve(matrix(c(3,1,1,3),ncol=2)) #inversa de la matriz de covarianzas
mu <- c(3,3) #vector media
DE <- eigen(Sinv) #descomposición espectral de Sinv

#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")
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))
points(mu[1],mu[2], pch = 16, cex = 0.9)

#Ejes en la dirección de las componentes principales:
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")
```

Ejemplo II



Distribución de la forma cuadrática

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

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.

Como $\mathbf{y}'\mathbf{y} = \sum_{i=1}^{p} y_i^2$ 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 $\mathbf{x}$.

Teorema

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

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

Demostración.

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:

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

Como $y_i \sim \mathcal{N}(0,1)$ para $i = 1, 2, \ldots, p$ y son mutuamente independientes,

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

Función característica de la multinormal II

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 multivariadas como la Cauchy, siempre y cuando las distribuciones multivariadas existan.

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 las combinaciones lineales.

Caracterización de una distribución multivariada a través de combinaciones lineales II

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 $s = 1$, tenemos que para $s = 1$ podemos obtener la función carcaterística 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 $y = \mathbf{t}'\mathbf{x}$.

□

- 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: un vector $\mathbf{x}$ es multinormal si y sólo si $\mathbf{a}'x$ es 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 III

- 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.

Teorema

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

Demostración.

Si $\mathbf{b} \in \mathbb{R}^q$, entonces $\mathbf{b}'\mathbf{y} = \mathbf{b}'\mathbf{A}\mathbf{x} + \mathbf{b}'\mathbf{c} = \mathbf{a}'\mathbf{x} + \mathbf{b}'\mathbf{c}$. 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, $\mathbf{y}$ es multinormal. □

Resultados relativos a componentes y particiones

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 $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$ y $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$ entonces $\mathbf{x}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$

Teorema

1. Si $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2$ de dimensiones q_1 y q_2 respectivamente, entonces $\text{cov}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{0}_{q_1 \times q_2}$
2. Si $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim \mathcal{N}_{q_1+q_2}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right)$ 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 $\mathbf{x}_1 \sim \mathcal{N}_{q_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ y $\mathbf{x}_2 \sim \mathcal{N}_{q_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, entonces el vector agregado $\mathbf{x}_1 = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim \mathcal{N}_{q_1+q_2}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right)$.

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 $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim \mathcal{N}_p \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right)$ 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 $\mathbf{B}$,

Distribuciones normales condicionales II

$$\begin{pmatrix} \mathbf{I} & -\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{B}' & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\mathbf{B}' - \mathbf{B}\boldsymbol{\Sigma}_{21} + \mathbf{B}\boldsymbol{\Sigma}_{22}\mathbf{B}' & \boldsymbol{\Sigma}_{12} - \mathbf{B}\boldsymbol{\Sigma}_{22} \\ \boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{22}\mathbf{B}' & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

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

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

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

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

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

Sea $\mathbf{x}_1, \ldots \mathbf{x}_n$ una muestra aleatoria de una distribución 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 log-verosimilitud esta dada por:

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

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 rescribir de otra manera:

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

y como el último término es

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

entonces (la parte amarilla se justificará adelante)

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

Otra representación de la (log-)verosimilitud II

Asi que finalmente:

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

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

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 $\mathbf{CB}$. □

Información de Fisher

Def (Información de Fisher)

a la derivada de la función de log-verosimilitud $l(\boldsymbol{\theta})$ se le conoce como *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 se puede ver como una variable aleatoria como función de la muestra. A la matriz de covarianzas de s se le llama *matriz de información de Fisher* y se denota por $\mathbf{F} = cov(s(\mathbf{x}, \boldsymbol{\theta}))$

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 $\mathbf{B}$ y un escalar $b > 0$, se cumple la siguiente desigualdad:

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

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

Demostración.

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

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

Estimadores máximo verosímiles de μ y Σ II

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:

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

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

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

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

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

Estimadores máximo verosímiles de μ y Σ III

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

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

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

□

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

Teorema

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

Demostración.

El exponente en la función de verosimilitud es, como vimos en la otra representación de la verosimilitud:

$$tr\left[\mathbf{\Sigma}^{-1}n\mathbf{S}\right] + n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})$$

Estimadores máximo verosímiles de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ IV

Como $\boldsymbol{\Sigma}^{-1} > \mathbf{0}$, entonces la forma cuadrática $(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) > 0$ a menos que $\boldsymbol{\mu} = \bar{\mathbf{x}}$. Por lo tanto la verosimilitud se maximiza con respecto a $\boldsymbol{\mu}$ en $\bar{\mathbf{x}}$. Para ver el máximo con respecto a $\boldsymbol{\Sigma}$, por el lema anterior con $b = n/2$ y $\mathbf{B} = \mathbf{S}$, el máximo se alcanza en $\hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\mathbf{S}$

□