Análisis Multivariado I - Práctica 3 - Parte 2

Los ejercicios marcados en rojo no son para elegir para exponer, aunque deben hacerse.

Inferencia para matrices de covarianza

1. Consideremos el modelo de regresión lineal

$$y_i = \theta_0 + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta} + \varepsilon_i$$

donde $\mathbf{x}_i$ m.a. $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ independientes de ε_i m.a. $N(0, \sigma^2)$.

   (a) ¿Qué distribución tiene $\mathbf{z}_i = (\mathbf{x}_i^{\mathrm{T}}, y_i)^{\mathrm{T}}$?

   (b) ¿Qué expresión tiene (en términos de $\mathbf{x}_i^{\mathrm{T}}$ e y_i) el estadístico del test del cociente de verosimilitud para la independencia de bloques aplicado a las $\mathbf{z}_i$?

Comentario: el test de F usual para $H_0 : \boldsymbol{\theta} = \mathbf{0}$ es equivalente al test del cociente de verosimilitud anterior.

2. Dada un muestra $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ mostrar que el estadístico del cociente de verosimilitud para testear $H_0 : \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ cumple

$$\ell^{2/(nd)} = \frac{|\mathbf{Q}|^{1/d}}{\operatorname{tr}(\mathbf{Q})/d}.$$

Escribirlo en función de los autovalores e interpretar. ¿Cuál es la distribución asintótica de $-2\ln \ell$ bajo H_0?

3. Consideremos una muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Hallar el estadístico del test del cociente de verosimilitud para las hipótesis

$$H_0 : \boldsymbol{\Sigma} = \sigma^2 (1 - \rho) \mathbf{I} + \sigma^2 \rho \mathbf{1}_d \mathbf{1}_d^{\mathrm{T}}$$

versus

$$H_1 : \nexists\, \sigma^2 \text{ ó } \rho \in (0, 1) \text{ tal que } \boldsymbol{\Sigma} = \sigma^2 (1 - \rho) \mathbf{I} + \sigma^2 \rho \mathbf{1}_d \mathbf{1}_d^{\mathrm{T}}.$$

4. Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una muestra de una $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Encontrar el test basado en el principio de unión–intersección para testear $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$. Explicar cómo se podrían calcular los valores críticos del test.

5. La Tabla 1 y el archivo `P3-2-ej5-2019.txt` contiene información fisonómica de 25 hermanos. Las variables son

   x_1 = longitud de la cabeza del primer hijo,

   x_2 = ancho de la cabeza del primer hijo,

$x_3 =$ longitud de la cabeza del segundo hijo,

$x_4 =$ ancho de la cabeza del segundo hijo;

referidas a 25 familias distintas. Consideremos el siguiente vector aleatorio $\mathbf{y} = (y_1, y_2, y_3, y_4)^{\mathrm{T}}$ donde

$$y_1 = x_1 + x_3\,, \quad y_2 = x_2 + x_4\,, \quad y_3 = x_1 - x_3\,, \quad y_4 = x_2 - x_4\,.$$

Testear si las primeras dos coordenadas de $\mathbf{y}$ son independientes de las segundas dos, a nivel 0.05 con el test de máxima verosimilitud y con el test basado en el principio de unión-intersección. Comparar las conclusiones.

6. Para los datos del ejercicio anterior, se desea estudiar si el primer y segundo hijo difieren en la dimensión media de su cabeza.

   (a) Grafique los datos e indique las medias del primer hijo y del segundo hijo en distinto color.

   (b) Como plantearía esta hipótesis? Qué supuestos necesita?

   (c) Qué decisión toma al 5%? y al 1%? Qué opina?

   (d) Supongamos que decidimos aceptar la hipótesis nula. Encuentre el estimador de máxima verosimilitud de (μ_1, μ_2) cuando la hipótesis nula es cierta.

   Ayuda: *Defina* $\mathbf{y}$ *como en el ejercicio 5 y use la invariancia de los estimadores de máxima verosimilitud.*

7. Sea $\mathbf{x} = \begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{pmatrix}$ donde $\mathbf{x}^{(1)} \in \mathbb{R}^p$, $\mathbf{x}^{(2)} \in \mathbb{R}^p$. Supongamos que $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ donde

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix} \quad \text{y} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

con $\boldsymbol{\mu}^{(1)} \in \mathbb{R}^p$, $\boldsymbol{\mu}^{(2)} \in \mathbb{R}^p$, $\boldsymbol{\Sigma}_{11} \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Sigma}_{22} \in \mathbb{R}^{p\times p}$.

Se desea testear que $\mathbf{x}^{(1)}$ y $\mathbf{x}^{(2)}$ son intercambiables o sea, que $\mathbf{x}$ tiene la misma distribución que $\begin{pmatrix} \mathbf{x}^{(2)} \\ \mathbf{x}^{(1)} \end{pmatrix}$ basado en una muestra aleatorio $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

   (a) Teniendo en cuenta que $\mathbf{x}$ es normal, muestre que esta hipótesis puede escribirse como

$$H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)} \qquad \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22} \qquad \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}\,,$$

   es decir,

$$H_0 : \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\theta} \end{pmatrix} \text{ para algún } \boldsymbol{\theta} \in \mathbb{R}^p \text{ y } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Psi} & \boldsymbol{\Upsilon} \\ \boldsymbol{\Upsilon} & \boldsymbol{\Psi} \end{pmatrix} \text{para matrices simétricas } \boldsymbol{\Psi}, \boldsymbol{\Upsilon} \in \mathbb{R}^{p\times p}$$

   (b) Sea $\mathbf{y} = \begin{pmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{pmatrix} = \begin{pmatrix} \mathbf{x}^{(1)} + \mathbf{x}^{(2)} \\ \mathbf{x}^{(1)} - \mathbf{x}^{(2)} \end{pmatrix}$.

i) Calcule la distribución de $\mathbf{y}$. Muestre que H_0 es equivalente a testear

$$H_0^\star : \boldsymbol{\nu}^{(2)} = \mathbf{0} \quad \boldsymbol{\Gamma}_{12} = \boldsymbol{\Gamma}_{21} = \mathbf{0}$$

donde $\boldsymbol{\nu} = \begin{pmatrix} \boldsymbol{\nu}^{(1)} \\ \boldsymbol{\nu}^{(2)} \end{pmatrix} = \mathbb{E}\mathbf{y}$ y $\boldsymbol{\Gamma} = \mathrm{Cov}(\mathbf{y})$.

ii) Mostrar que el test del cociente de verosimilitud para H_0 basado en una muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n$ es equivalente a rechazar cuando

$$\frac{\det(\widehat{\boldsymbol{\Gamma}})}{\det(\widehat{\boldsymbol{\Gamma}}_{11})\det(\widetilde{\boldsymbol{\Gamma}}_{22})}$$

es menor que una constante (que depende del nivel del test), donde

$$\begin{aligned} \widehat{\boldsymbol{\Gamma}} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^{\mathrm{T}} \\ \widehat{\boldsymbol{\Gamma}}_{11} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i^{(1)} - \overline{\mathbf{y}}^{(1)})(\mathbf{y}_i^{(1)} - \overline{\mathbf{y}}^{(1)})^{\mathrm{T}} \\ \widetilde{\boldsymbol{\Gamma}}_{22} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i^{(2)}(\mathbf{y}_i^{(2)})^{\mathrm{T}} \end{aligned}$$

Cuál es la distribución asintótica del test?

(c) Para los datos del tamaño de la cabeza del primer y segundo hijo, dados en el ejercicio 6, llamemos $\mathbf{x}_i^{(1)} = (x_{i1}, x_{i2})^{\mathrm{T}}$ y sea $\mathbf{x}_i^{(2)} = (x_{i3}, x_{i4})^{\mathrm{T}}$. Luego, $\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_i^{(1)} \\ \mathbf{x}_i^{(2)} \end{pmatrix}$. Sea $\mathbf{y}$ definido como en el ejercicio 6, es decir,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{pmatrix} = \begin{pmatrix} \mathbf{x}^{(1)} + \mathbf{x}^{(2)} \\ \mathbf{x}^{(1)} - \mathbf{x}^{(2)} \end{pmatrix}$$

y definamos

$$\mathbf{y}_i = \begin{pmatrix} \mathbf{y}_i^{(1)} \\ \mathbf{y}_i^{(2)} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_i^{(1)} + \mathbf{x}_i^{(2)} \\ \mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)} \end{pmatrix}.$$

Se desea testear que el tamaño de la cabeza de ambos hijos tiene la misma distribución.

i) Usando lo obtenido en b) especifique el estadístico del test y su distribución asintótica.

ii) A qué conclusión llegaría?

8. Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una m.a. $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Se desea testear que las d variables son independientes, es decir

$$H_0 : \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{dd}).$$

Consideremos la matriz de correlación muestral $\mathbf{R} = (R_{j\ell})_{1\le j,\ell\le d}$ cuyos elementos son

$$R_{jk} = \frac{\widehat{\sigma}_{jk}}{(\widehat{\sigma}_{jj}\ \widehat{\sigma}_{kk})^{1/2}} = \frac{Q_{jk}}{(Q_{jj}\ Q_{kk})^{1/2}},$$

donde

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{Q} = (\widehat{\sigma}_{j\ell})_{1\le j,\ell\le d} \qquad \mathbf{Q} = (Q_{j\ell})_{1\le j,\ell\le d}$$

(a) Probar que el estadístico del test del cociente de verosimilitud es $\Lambda = |\mathbf{R}|^{n/2}$.

(b) ¿Cuál es la distribución asintótica de $-2\ln\Lambda$ bajo H_0?

Test para igualdad de covarianza entre dos poblaciones

9. En el ejercicio 1 de la Parte 1 de la Práctica 3 se estudiaba el costo de transporte de la leche desde las granjas hasta las lecherías para $n_1 = 36$ camiones nafteros y $n_2 = 23$ camiones a diesel.

   (a) Para testear si había diferencias entre los vectores de costos esperados, se supuso que las dos poblaciones tenían igual matriz de covarianza. ¿Es este supuesto razonable? Tomar $\alpha = 0.01$.

   (b) Si la hipótesis de igualdad de matrices de covarianzas es rechazada en la parte (a), ¿cómo testearía la igualdad de vectores esperados?

10. En el ejercicio 4 de la Parte 1 de la Práctica 3 se estudiaban los resultados de tomar un test de habilidad sicolingual a dos grupos de 27 chicos de edades 8-9 años. Nos interesaba estudiar las siguientes hipótesis:

    H_{01}: los dos perfiles son similares

    H_{02}: los dos perfiles están al mismo nivel

    H_{03}: no hay diferencias entre las medias de los tests

    Sean

    - x_1 = recepción auditiva
    - x_2 = recepción visual
    - x_3 = memoria visual
    - x_4 = asociación auditiva
    - x_5 = memoria auditiva
    - x_6 = asociación visual
    - x_7 = oclusión visual
    - x_8 = expresión oral
    - x_9 = oclusión gramatical
    - x_{10} = destreza manual

    (a) Para cada una de las hipótesis anteriores, ¿es necesario suponer que las matrices de covarianza del vector $\mathbf{x} = (x_1, \ldots, x_{10})^{\mathrm{T}}$ en las dos poblaciones sea la misma o puede hacer un supuesto más débil? Exprese matemáticamente la hipótesis sobre la igualdad de covarianzas que sea necesaria en cada caso.

    (b) Para cada una de las situaciones de interés estudie si el supuesto de igualdad de covarianzas es razonable. Tomar $\alpha = 0.01$.

x_1	x_2	x_3	x_4
191	155	179	145
195	149	201	152
181	148	185	149
183	153	188	149
176	144	171	142
208	157	192	152
189	150	190	149
197	159	189	152
188	152	197	159
192	150	187	151
186	161	179	158
179	147	183	147
195	153	174	150
202	160	190	159
194	154	188	151
163	137	161	130
195	155	183	158
186	153	173	148
181	145	182	146
175	140	165	137
192	154	185	152
174	143	178	147
176	139	176	143
197	167	200	158
190	153	187	150

Table 1: Datos fisonómicos de 25 hermanos. Corresponden a la Tabla 6.2.1 de Flury (1997)

11. En el ejercicio 5 de la Parte 1 de la Práctica 3 se estudiaba las medidas de cinco variables biométricas sobre gorriones hembra, recogidos casi moribundos después de una tormenta. 21 de ellos sobrevivieron mientras que los 28 restantes no lo consiguieron.

    (a) Para testear si había diferencias entre los vectores de medidas esperadas, se supuso que las dos poblaciones tenían igual matriz de covarianza. ¿Es este supuesto razonable? Tomar $\alpha = 0.01$.

    (b) Si la hipótesis de igualdad de matrices de covarianzas es rechazada en la parte (a), ¿cómo testearía la igualdad de los vectores esperados?