Certamen 1 - Análisis Multivariado

Fabián Ramírez

Problemas

1. Problema *Demostración:*

Notar que:

$$
\begin{aligned}
\overline{x}_n &= \frac{1}{u_n}\sum_{i=1}^{n} u_i x_i \\
&= \frac{1}{U_n}\left(\sum_{i=1}^{n-1} u_i x_i + u_n x_n\right) \\
&= \frac{1}{U_n}\left[U_{n-1}\overline{x}_{n-1} + u_n x_n\right] \\
&= \frac{1}{U_n}\left[\left(U_n - u_n\right)\overline{x}_{n-1} + u_n x_n\right] \\
&= \overline{x}_{n-1} - \frac{u_n}{U_n}\overline{x}_{n-1} + \frac{u_n}{U_n}x_n \\
&= \overline{x}_{n-1} + \frac{u_n}{U_n}\left(-\overline{x}_{n-1} + x_n\right) \\
&= \overline{x}_{n-1} + \frac{u_n}{U_n}d_n
\end{aligned}
$$

Luego note que:

$$
\begin{aligned}
\boldsymbol{Q}_n &= \sum_{i=1}^{n} u_i\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n\right)^\top \\
&= \sum_{i=1}^{n-1} u_i\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n\right)^\top + u_n(x_n - \overline{x}_n)(x_n - \overline{x}_n)^\top \\
&= \sum_{i=1}^{n-1} u_i\left(\boldsymbol{x}_i - \overline{x}_{n-1} - \frac{u_n}{U_n}d_n\right)\left(\boldsymbol{x}_i - \overline{x}_{n-1} - \frac{u_n}{U_n}d_n\right)^\top + u_n\left(x_n - \overline{x}_{n-1} - \frac{u_n}{U_n}d_n\right)\left(x_n - \overline{x}_{n-1} - \frac{u_n}{U_n}d_n\right)^\top \\
&= \sum_{i=1}^{n-1} u_i\left(x_i - \overline{x}_{n-1}\right)\left(x_i - \overline{x}_{n-1}\right)^\top + \cancelto{0}{\sum_{i=1}^{n-1} -u_i\left(x_i - \overline{x}_{n-1}\right)\left(\frac{u_n}{U_n}d_n\right)^\top} - \cancelto{0}{\sum_{i=1}^{n-1} u_i\left(\frac{u_n}{U_n}d_n\right)\left(x_i - \overline{x}_{n-1}\right)^\top} + \\
&+ \sum_{i=1}^{n-1} u_i\left(\frac{u_n}{U_n}d_n\right)\left(\frac{u_n}{U_n}d_n\right)^\top + u_n\left(x_n - \overline{x}_{n-1} - \frac{u_n}{U_n}d_n\right)\left(x_n - \overline{x}_{n-1} - \frac{u_n}{U_n}d_n\right)^\top \\
&= Q_{n-1} + \frac{u_n^2}{U_n^2}(U_n - u_n)d_n d_n^\top + u_n(x_n - \overline{x}_{n-1})(x_n - \overline{x}_{n-1})^\top - u_n(x_n - \overline{x}_{n-1})\left(\frac{u_n}{U_n}\right)d_n^\top - u_n\left(\frac{u_n}{U_n}\right)d_n(x_n - \overline{x}_{n-1})^\top \\
&+ \frac{u_n^3}{U_n^2}d_n d_n^\top \\
&= Q_{n-1} + \frac{u_n^2}{U_n}d_n d_n^\top + u_n d_n d_n^\top - \frac{u_n^2}{U_n}d_n d_n^\top - \frac{u_n^2}{U_n}d_n d_n^\top \\
&= Q_{n-1} + \left(u_n - \frac{u_n^2}{U_n}\right)d_n d_n^\top
\end{aligned}
$$

2. Problema (Ver anexo para ver las cuentas)

Solución

Notemos que la log verosimilitud viene dada por:

$$\ell(p) = -n\log(2\pi) - \frac{n}{2}\log\left[\left(1-p^2\right)\right] - \frac{1}{2}\sum_{l=1}^{n}\frac{1}{1-p^2}\cdot\left[x_i^2 + y_i^2 - 2px_iy_i\right]$$

Notemos que al resolver $\ell'(p) = 0$ se obtiene la siguiente ecuación de estimación:

$$-p^3 - \frac{1}{n}s_{xy}\cdot p^2 + \left(1 - \frac{1}{n}s_{xx} - \frac{1}{n}s_{yy}\right)p + \frac{1}{n}s_{xy} = 0$$

El cual es un polinomio cubico del cual se le pueden obtener 3 raíces explícitamente (es demasiado largo el desarrollo para obtener las raíces explícitamente). Ahora bien nota que si aumentamos la cantidad de datos se tiene que esta ecuación de estimación converge a:

$$-p^3 + p = 0$$

Cuyas raíces son $0, 1, -1$, el problema es que tanto en $1, -1$ la densidad no existe (ver el determinante de la matriz). Por tanto consideraremos que el estimador máximo verosímil de p es aquella raíz de la ecuación de estimación que converge a 0 cuando n converge a infinito y cuya convergencia asintótica es a 0. Por tanto la distribución a la que converge la variable aleatoria que genera los datos es a una normal bi-variada con media 0 y varianza identidad.

3. Problema (Ver anexo para ver las cuentas)

Demostración:

$$\ell(\Sigma) = \log\left(\frac{1}{2^{mn}\Gamma(n/2)}\right) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}A\right) + \frac{(n-m-1)}{2}\log|A|$$

Note que:

$$\mathrm{d}\ell(\Sigma) = -\frac{n}{2}\cdot\operatorname{tr}\left(\Sigma^{-1}\mathrm{d}\Sigma\right) - \frac{1}{2}\operatorname{tr}\left(-\Sigma^{-1}\mathrm{d}\Sigma\Sigma^{-1}A\right)$$

Resolviendo $\mathrm{d}\ell(\Sigma) = 0$ se tiene que una ecuación de estimación viene dada por:

$$\operatorname{tr}\left(\mathrm{d}\Sigma\left(-n\mathrm{Id} + \Sigma^{-1}A\right)\Sigma^{-1}\right) = 0$$

Note que $\mathrm{d}\Sigma\left(-n\mathrm{Id} + \Sigma^{-1}A\right)\Sigma^{-1} \in \operatorname{Ker}(\operatorname{tr})$ por tanto en particular puedo tomar Σ tal que $\mathrm{d}\Sigma\left(-n\mathrm{Id} + \Sigma^{-1}A\right)\Sigma^{-1} = 0$ el cual existe por propiedad de una transformación lineal,(de otra forma puedo encontrar otro estimador). De esta forma despejando se concluye que:

$$\widehat{\Sigma} = \frac{1}{n}A$$

4. Problema. *Demostración:*

1. Sea $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ y $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$. Notemos que $\boldsymbol{z}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, luego es fácil chequear que $\overline{\boldsymbol{x}} \sim \mathcal{N}_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$. Se sabe que $\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{z}}$ y también que $\widehat{\boldsymbol{\mu}}_1 = \overline{\boldsymbol{x}}$, $\widehat{\boldsymbol{\mu}}_2 = \overline{\boldsymbol{y}}$. Obteniendo así una relación entre los estimadores $\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}} - \overline{\boldsymbol{y}}$. De la misma manera sabemos que $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S}$, pero por independencia de las muestras se puede obtener la separación de la varianza de la suma en la suma de las varianzas por lo tanto, $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_1 + \widehat{\boldsymbol{\Sigma}}_2 = \boldsymbol{S}_1 + \boldsymbol{S}_2$. Finalmente basándonos en un resultado visto en clases tenemos que la estadística $T^2 = n\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)^\top \widehat{\boldsymbol{\Sigma}}^{-1}\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) \sim \mathrm{T}^2(p, n-1)$, en nuestro caso esta estadística toma la siguiente forma $T^2 = n\left(\overline{\boldsymbol{x}} - \overline{\boldsymbol{y}} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right)^\top (\boldsymbol{S}_1 + \boldsymbol{S}_2)^{-1}\left(\overline{\boldsymbol{x}} - \overline{\boldsymbol{y}} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right) \sim \mathrm{T}^2(p, n-1)$ la cual es una variable aleatoria que depende del parámetro de interés ($\boldsymbol{\mu}$) pero su distribución no depende del parámetro de interés. Finalmente dada la relación entre la distribución T^2 y la F tenemos que $F = \frac{T^2}{n-1}\frac{n-p}{p} \sim F(p, n-p)$. Y con esta estadística podemos realizar el test $H_0 : \boldsymbol{\mu} = 0$ lo cual es equivalente a decir que $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ por lo tanto el estadístico de prueba toma la forma $F_0 = \frac{n\left(\overline{\boldsymbol{x}} - \overline{\boldsymbol{y}}\right)^\top (\boldsymbol{S}_1 + \boldsymbol{S}_2)^{-1}\left(\overline{\boldsymbol{x}} - \overline{\boldsymbol{y}}\right)}{n-1}\frac{n-p}{p}$, y tendremos el siguiente criterio de decisión que es rechazar H_0 cuando $F_0 \geq F_{1-\alpha}(p, n-p)$ a la significancia α.

2. Siguiendo la misma idea del punto anterior lo que necesitamos determinar es como es la nueva estimación de $\boldsymbol{\Sigma}$ dada la información de que ahora $\boldsymbol{\Sigma} = k\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_2 = (k+1)\boldsymbol{\Sigma}_2$, notemos que Σ es una función parámetro de $\boldsymbol{\Sigma}_2$ y dado que k es conocido tenemos que $\widehat{\boldsymbol{\Sigma}} = (k+1)\widehat{\boldsymbol{\Sigma}}_2 = (k+1)\boldsymbol{S}_2$, luego $\widehat{\boldsymbol{\Sigma}}^{-1} = \frac{1}{k+1}(\boldsymbol{S}_2)^{-1}$. Entonces utilizando el mismo resultado anterior tenemos que $T^2 = n\left(\overline{\boldsymbol{x}} - \overline{\boldsymbol{y}} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right)^\top \frac{1}{k+1}(\boldsymbol{S}_2)^{-1}\left(\overline{\boldsymbol{x}} - \overline{\boldsymbol{y}} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right) \sim \mathrm{T}^2(p, n-1)$ es cantidad pivotal para $\boldsymbol{\mu}$, luego tenemos que dada la relación con la Fisher tenemos que $F = \frac{T^2}{n-1}\frac{n-p}{p} \sim F(p, n-p)$. Y con esta estadística podemos realizar el test $H_0 : \boldsymbol{\mu} = 0$ lo cual es equivalente a decir que $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ por lo tanto el estadístico de prueba toma la forma $F_0 = \frac{n\left(\overline{\boldsymbol{x}} - \overline{\boldsymbol{y}}\right)^\top \frac{1}{k+1}(\boldsymbol{S}_2)^{-1}\left(\overline{\boldsymbol{x}} - \overline{\boldsymbol{y}}\right)}{n-1}\frac{n-p}{p}$, y tendremos el siguiente criterio de decisión que es rechazar H_0 cuando $F_0 \geq F_{1-\alpha}(p, n-p)$ a la significancia α.

5. Problema. *Demostración:*

Nota que $\overline{\boldsymbol{x}} \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$ y $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, como son provenientes de una muestra aleatoria entonces tenemos independencia y por tanto se tiene que $\boldsymbol{x} - \overline{\boldsymbol{x}} \sim \mathcal{N}\left(\boldsymbol{\mu} - \boldsymbol{\mu}, \boldsymbol{\Sigma} + \frac{1}{n}\boldsymbol{\Sigma}\right)$, si llamamos $\boldsymbol{z} = \boldsymbol{x} - \overline{\boldsymbol{x}} \sim \mathcal{N}\left(\boldsymbol{0}, \left(1 + \frac{1}{n}\right)\boldsymbol{\Sigma}\right)$. Usando la misma idea del Problema 4 sabemos que $T^2 = n\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)^\top \widehat{\boldsymbol{\Sigma}^*}^{-1}\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) \sim \mathrm{T}^2(p, n-1)$. En nuestro caso $\boldsymbol{\mu} = \boldsymbol{0}$, $\Sigma^* = \frac{n+1}{n}\boldsymbol{\Sigma}$, $\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{z}} = \boldsymbol{x} - \overline{\boldsymbol{x}}$ y $\widehat{\boldsymbol{\Sigma}^*} = \frac{n+1}{n}\widehat{\boldsymbol{\Sigma}}$, por lo tanto tenemos que $\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) = \boldsymbol{x} - \overline{\boldsymbol{x}}$ y $\widehat{\boldsymbol{\Sigma}^*}^{-1} = \frac{n}{n+1}\widehat{\boldsymbol{\Sigma}}^{-1} =$, luego $T^2 = n\left(\boldsymbol{x} - \overline{\boldsymbol{x}}\right)^\top \frac{n}{n+1}\widehat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{x} - \overline{\boldsymbol{x}}\right) = \left(\boldsymbol{x} - \overline{\boldsymbol{x}}\right)^\top \frac{n}{n+1}\left[\frac{1}{n}\widehat{\boldsymbol{\Sigma}}\right]^{-1}\left(\boldsymbol{x} - \overline{\boldsymbol{x}}\right)$, luego $\left[\frac{1}{n}\widehat{\boldsymbol{\Sigma}}\right]$ es la estimación de la covarianza del promedio por tanto definimos S como la estimación de la varianza del promedio mediante n datos, de esa forma se tiene por construcción que:

$$T^2 = \left(\boldsymbol{x} - \overline{\boldsymbol{x}}\right)^\top \frac{n}{n+1}[S]^{-1}\left(\boldsymbol{x} - \overline{\boldsymbol{x}}\right) \sim \mathrm{T}^2(p, n-1)$$