1 Análisis Multivariado I - Práctica 1

1.1 Distribución Normal Multivariada

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

1. Sean $\mathbf{y}_i \sim N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ independientes $(1 \leq i \leq n)$ y $\mathbf{a} = (a_1, \ldots, a_n)^{\mathrm{T}} \in \mathbb{R}^n$. Probar que $\sum_{i=1}^n a_i \mathbf{y}_i \sim N_d\left(\sum_{i=1}^n a_i \boldsymbol{\mu}_i, \sum_{i=1}^n a_i^2 \boldsymbol{\Sigma}_i\right)$.

   Sugerencia: usar la distribución normal univariada.

2. Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una m.a. $N_d(\mathbf{0}, \boldsymbol{\Sigma})$. Llamemos $\mathbf{X}^{\mathrm{T}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$.

   (a) Si $\mathbf{a}$ de $n \times 1$ es un vector no aleatorio, entonces $\mathbf{X}^{\mathrm{T}}\mathbf{a} \sim N_d\left(\mathbf{0}, \|\mathbf{a}\|^2 \boldsymbol{\Sigma}\right)$.

   (b) Si $\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$ es un conjunto de vectores ortogonales no aleatorios, entonces los vectores aleatorios $\mathbf{u}_i = \mathbf{X}^{\mathrm{T}}\mathbf{a}_i$, $1 \leq i \leq r$, son independientes.

   (c) Si $\mathbf{b}$ de $d \times 1$ es un vector no aleatorio, entonces $\mathbf{Xb} \sim N_n(\mathbf{0}, (\mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{b})\mathbf{I}_n)$. En particular, $\mathbf{x}^{(j)} \sim N_n(\mathbf{0}, \sigma_{jj}\mathbf{I}_n)$, con $\boldsymbol{\Sigma} = (\sigma_{ij})$.

Definición: Si las variables aleatorias $X_1, X_2, ..., X_n$ son i.i.d. $N_1(\mu_i, \sigma^2)$, entonces

$$U = \sum_{i=1}^n \frac{X_i^2}{\sigma^2} \sim \chi_n^2(\delta)$$

es decir que la distribución de la variable aleatoria U se denomina χ^2 *no central* con parámetro de centralidad $\delta = \sum_{i=1}^n (\mu_i^2/\sigma^2)$.

3. Consideremos $\mathbf{x} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

   (a) Probar que $\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} \sim \chi_d^2(\delta)$ con $\delta = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.

   (b) Si $\mathbf{B}$ es simétrica de rango k y $\mathbf{B}\boldsymbol{\Sigma}$ es idempotente, probar que $\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x} \sim \chi_k^2(\delta)$ con $\delta = \boldsymbol{\mu}^{\mathrm{T}}\mathbf{B}\boldsymbol{\mu}$.

4. (a) Sea $\mathbf{x} \sim N_2(\mathbf{0}, \boldsymbol{\Sigma})$.

      i. Usando que existe una matriz triangular (descomposición de Cholesky)

$$\mathbf{L} = \begin{pmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{pmatrix}$$

      tal que $\mathbf{L}\mathbf{L}^{\mathrm{T}} = \boldsymbol{\Sigma}$, encontrar los parámetros ℓ_{11}, ℓ_{21} y ℓ_{22} como función de los coeficientes de la matriz de covarianzas $\boldsymbol{\Sigma}$.

      ii. Interpretar los resultados del ítem anterior en base a las distribuciones marginales y condicionales de las variables.

(b) Aplicar la descomposición de Cholesky del ítem anterior a la matriz de covarianzas

$$\mathbf{\Sigma} = \begin{pmatrix} 9 & 3 \\ 3 & 4 \end{pmatrix}$$

5. Sea $\mathbf{x} \in \mathbb{R}^p$ un vector aleatorio y $\mathbf{\Sigma}$ su matriz de covarianzas. Paticionemos a $\mathbf{x}$ y $\mathbf{\Sigma}$ como

$$\mathbf{x} = (\mathbf{x}_1^{\mathrm{T}}, \mathbf{x}_2^{\mathrm{T}})^{\mathrm{T}} \qquad \text{y} \qquad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$$

donde $\mathbf{x}_1 \in\in \mathbb{R}^{p_1}$, $\mathbf{x}_2 \in\in \mathbb{R}^{p_2}$. Indiquemos por $\mathbf{L}_{11}$, $\mathbf{L}_{12}$ y $\mathbf{L}_{22}$ a las matrices correspondientes a la descomposición de Cholesky de $\mathbf{\Sigma}$, es decir, $\mathbf{L}\mathbf{L}^{\mathrm{T}} = \mathbf{\Sigma}$ con

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_{11} & \mathbf{0} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{pmatrix}.$$

Probar que $\mathbf{L}_{11}\mathbf{L}_{11}^{\mathrm{T}} = \mathbf{\Sigma}_{11}$, $\mathbf{L}_{21} = \mathbf{\Sigma}_{21}(\mathbf{L}_{11}^{\mathrm{T}})^{-1}$ y $\mathbf{L}_{22}\mathbf{L}_{22}^{\mathrm{T}} = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}$.

6. (a) Generar muestras de una distribución normal bivariada $\mathbf{x} \sim N_2(\boldsymbol{\mu}, \mathbf{\Sigma})$ donde

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

por el método siguiente:

(1) Generar una observación de la distribución marginal de la primera variable, es decir, genere una observación $x_{1,0}$ con distribución $N(\mu_1, \sigma_{11})$.

(2) Genere una observacion $x_{2,0}$ de la distribución de x_2 dada que $x_1 = x_{1,0}$.

Aplicarlo para generar valores al azar de un vector aleatorio normal con media $\boldsymbol{\mu} = (0, 5)^{\mathrm{T}}$ y matriz de covarianza $\mathbf{\Sigma} = \begin{pmatrix} 4 & 3 \\ 3 & 9 \end{pmatrix}$.

(b) Demostrar que el método anterior es equivalente a generar dos variables aleatorias normales estándar, $\mathbf{z} = (z_1, z_2)^{\mathrm{T}}$, y obtener los valores las variables x_1 y x_2 mediante la transformación $\mathbf{x} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z}$, donde $\mathbf{L}$ es la matriz triangular de la descomposición de Cholesky.

7. (a) Obtener las distribuciones condicionales de la normal bivariada con media cero y matriz de covarianzas $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

(b) La distribución de los gastos en dos productos (x, y) de un grupo de consumidores sigue una distribución normal bivariada con medias 2 y 3 euros respectivamente y matriz de covarianzas $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 2 \end{pmatrix}$.

Calcular la distribución condicional de los datos en el producto y para los consumidores que gastan 4 euros en el producto x.

¿Cuál es la probabilidad de que el gasto en el producto y sea mayor a 2 euros cuando el gasto en el producto x fue de 4 euros?

8. Cuando aumenta la dimensión del vector de datos, la *maldición de la dimensión* se manifiesta en que cada vez hay menos densidad de datos en una región fija del espacio. Para ilustrar este problema, considere vectores normales estándar $\mathbf{x} \sim N_p(\mathbf{0}, \mathbf{I}_p)$ y calcule usando la distribución de χ^2 la probabilidad de encontrar un valor en la bola unidad definida como $\mathcal{B}_1 = \{\mathbf{u} \in \mathbb{R}^p : \mathbf{u}^{\mathrm{T}}\mathbf{u} \leq 1\}$, cuando la dimensión p varía entre 2 y 16. Graficar las probabilidades en función de la dimensión. ¿Qué observa?

9. En lo que sigue realizaremos un pequeño estudio de simulación para observar el comportamiento del Test de Shapiro Wilks Multivariado propuesto en Villasenor Alva y González Estrada (2009) cuando, efectivamente, los datos provienen de una muestra de vectores normales multivariados.

   (a) Fijar la semilla.

   (b) Generar NITER $= 500$ muestras de tamaños $n = 20, 50$ y 100 de vectores $\mathbf{x}_i$ con la misma distribución que $\mathbf{x}$ tal que

      i. $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ con x_1 y x_2 independientes y con distribución $N(0, 1)$

      ii. $\mathbf{x} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $\boldsymbol{\mu} = (0, 5)^{\mathrm{T}}$ y $\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 3 \\ 3 & 9 \end{pmatrix}$, utilizando la función `mvrnorm` de la librería `MASS` de `R`

      iii. $\mathbf{x} \sim N_3(\mathbf{0}, \mathbf{I}_3)$

   (c) Indicar, para cada n, la proporción de muestras para las que el test no rechaza la hipótesis de normalidad multivariada a nivel 0.10. ¿A qué número debería parecerse dicha proporción?

   Para realizar el Test de Shapiro Wilk Multivariado utilizar la función `mvShapiro.Test` de la librería `mvShapiroTest` de `R`.

   Referencia

Villasenor Alva, J. y González Estrada, E. (2009). A Generalization of ShapiroWilk's Test for Multivariate Normality. *Communications in Statistics: Theory and Methods*, **38**, 1870-1883.

10. Sea $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_p)$ donde $p = 10$.

    (a) Sea D la distancia de $\mathbf{x}$ al centro de la distribución, en este caso $\mathbf{0}$. Calcule $\mathbb{E}(D^2)$

    (b) Sean $\mathbf{x}_1, \ldots, \mathbf{x}_n$, una muestra aleatoria de $\mathbf{x}$ y sea $\mathbf{x}_0 \sim \mathbf{x}$ independiente de $\mathbf{x}_i$, $1 \leq i \leq n$.

       i. Qué distribución tiene $\mathbf{x}_0^{\mathrm{T}}\mathbf{x}_i / \|\mathbf{x}_0\|$?

       ii. Calcule la distancia al cuadrado esperada entre el centro de los datos y $\mathbf{x}_0$ dado $\mathbf{x}_0$, es decir,

       $$\mathbb{E}\left(\|\overline{\mathbf{x}} - \mathbf{x}_0\|^2 \middle| \mathbf{x}_0\right)$$

       iii. Deduzca el valor de $\mathbb{E}\left(\|\overline{\mathbf{x}} - \mathbf{x}_0\|^2\right)$. Qué observa?

11. Sea $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $\boldsymbol{\mu} = (-1, 1, 0)^{\mathrm{T}}$ y

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

(a) Hallar la distribución $x_1 + 2\,x_2 - 3\,x_3$.

(b) Hallar $\mathbf{a} \in \mathbb{R}^2$ tal que las variables x_1 y

$$x_1 - \mathbf{a}^{\mathrm{T}} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix}$$

sean independientes.

(c) Obtener la distribución condicional de x_3 dado $(x_1, x_2) = (x_{10}, x_{20})$.