Tarea 7: Introducción a Ciencia de Datos.

Rojas Gutiérrez Rodolfo Emmanuel

31 de octubre de 2021

1. Introducción

A lo largo de esta tarea, $M_{n\times m}(\mathbb{R})$ denota el espacio de todas las matrices de dimensión $n \times m$ con coeficientes en los reales, adicionalmente si $A \in M_{n\times m}(\mathbb{R})$ entonces A' denota a la matriz transpuesta de A.

2. Ejercicios

Para comenzar con la sección de ejercicios, se dará una Observación concerniente únicamente al Ejercicio 1.

Observación 1. *Sean Y y Z vectores aleatorios que toman valores en $\mathbb{R}^n$, entonces, es posible que (Y, Z) no sea un vector aleatorio ni discreto ni continuo, en cuyo caso no tiene sentido hablar de la función de densidad o función masa de probabilidad conjunta de (Y, Z), no obstante, si se supone que Y es un vector aleatorio absolutamente continuo y Z es un vector aleatorio discreto, existe una noción similar a la idea de densidad la cual se denomina densidad generalizada conjunta de Z e Y. Esto se comenta, pues el término será empleado a lo largo de este ejercicio.*[1] *Adicionalmente, dado que bajo el enfoque frecuentista de la estadística, los parámetros de los modelos no son considerados variables aleatorias, la notación $\mathbb{E}[Y|Z;\theta]$ ó $\mathbb{P}[Y \in A|Z;\theta]$ sera usada para indicar de manera respectiva lo siguiente: La esperanza, dependiendo de un vector de parámetros θ, de Y dado el vector aleatorio Z y la probabilidad dependiendo en algún vector de parámetros θ, de que el vector aleatorio Y pertenezca a algún boreliano A dado el vector Z.*[2] *Se resalta este último aspecto, pues en McLachlan y Krishnan (2008) suelen emplear esta notación, o en su defecto una notación con el vector de parámetros como subíndice, para hacer referencia a esta misma idea.* △

Ejercicio 1:

Solución. El siguiente ejemplo fue extraído de McLachlan y Krishnan (2008). Considere que se han mezclado dos modelos de series de tiempo del tipo[3] $AR(1)$ y, que se sabe que los parámetros de este

[1] Para más información sobre densidades generalizadas ver Martín González (2021), pp 15-17.

[2] Al decir dependiendo en un vector de parámetros θ, se debe asumir que las distribuciones implicadas en los cálculos de la esperanza y probabilidad anteriores, dependen de este vector de parámetros.

[3] $AR(1)$ es la abreviación para modelo auto-regresivo de primer orden.

modelo mezclado, se rigen por una cadena de Markov de dos estados $\{S_1, S_2\}$. En otras palabras, considere una serie de tiempo $\{w_j : j \in \mathbb{N}\}$ de la forma:

$$w_j - \mu_{s_j} = \beta_{s_j}(w_{j-1} - \mu_{s_{j-1}}) + \varepsilon_j, \text{ con } j \in \mathbb{N}, \tag{1}$$

donde, para $(j,i) \in \mathbb{N} \times \{1,2\}$, se tiene que $s_j = i$ si y solo si ω_j se encuentra en el estado S_i de la cadena de Markov oculta y, donde $\{\varepsilon_j : j \in \mathbb{N}\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas (i.i.d), con distribución común: Normal de parámetros $(0, \sigma^2)$ para alguna $\sigma > 0$. Luego, a modo de simplificación del modelo planteado en (1) se supondrá que $\mu_1 = \mu_2 = 0$, con lo que, es posible escribir el mismo como:

$$w_j = \beta_{s_j} w_{j-1} + \varepsilon_j, \text{ con } j \in \mathbb{N}, \tag{2}$$

Ahora, para cada $(j,i) \in \mathbb{N} \times \{1,2\}$, defina las variables indicadoras $Z_{ij} = \mathbb{1}_{\{s_j = i\}}$ y para cada $n \in \mathbb{N}$ establezca $\overline{Z}_n = \left(Z_1', \ldots, Z_n'\right)'$, donde, para $j \in \mathbb{N}$ se tiene que $Z_j = \begin{pmatrix} Z_{1j} & Z_{2j} \end{pmatrix}$. De este modo, observe que la dependencia entre las indicadoras de estado Z_j, queda especificada mediante la Cadena de Markov oculta que determina el estado en el que se encuentra la serie de tiempo, cuya matriz de probabilidades de transición se denotará por A y sus entradas $(A)_{ij}$, con $i,j \in \{1,2\}$, se denotaran por a_{ij}. Esto pues, las variables Z_j forman una cadena de Markov con espacio de estados $E = \{(0,1),(1,0)\}$ y mismas probabilidades de transición que la cadena original,[4] ya que para $j \in \mathbb{N}$ y $x, y \in E$ con $x = (x_1, x_2)$ e $y = (y_1, y_2)$, se cumple que

$$\begin{aligned}
\mathbb{P}[Z_{j+1} = x | Z_j = y] &= \mathbb{P}[Z_{j+1} = (1,0) | Z_j = (1,0)] \mathbb{1}_{\{x_1=1, y_1=1\}} + \mathbb{P}[Z_{j+1} = (1,0) | Z_j = (0,1)] \mathbb{1}_{\{x_1=1, y_2=1\}} \\
&+ \mathbb{P}[Z_{j+1} = (0,1) | Z_j = (1,0)] \mathbb{1}_{\{x_2=1, y_1=1\}} + \mathbb{P}[Z_{j+1} = (0,1) | Z_j = (0,1)] \mathbb{1}_{\{x_2=1, y_2=1\}} \\
&= \mathbb{P}[s_{j+1} = 1 | s_j = 1] \mathbb{1}_{\{x_1=1, y_1=1\}} + \mathbb{P}[s_{j+1} = 1 | s_j = 2] \mathbb{1}_{\{x_1=1, y_2=1\}} \\
&+ \mathbb{P}[s_{j+1} = 2 | s_j = 1] \mathbb{1}_{\{x_2=1, y_1=1\}} + \mathbb{P}[s_{j+1} = 2 | s_j = 2] \mathbb{1}_{\{x_2=1, y_2=1\}} \\
&= \mathbb{P}[w_{j+1} \in S_1 | w_j \in S_1] \mathbb{1}_{\{x_1=1, y_1=1\}} + \mathbb{P}[w_{j+1} \in S_1 | w_j \in S_2] \mathbb{1}_{\{x_1=1, y_2=1\}} \\
&+ \mathbb{P}[w_{j+1} \in S_2 | w_j \in S_1] \mathbb{1}_{\{x_2=1, y_1=1\}} + \mathbb{P}[w_{j+1} \in S_2 | w_j \in S_2] \mathbb{1}_{\{x_2=1, y_2=1\}} \\
&= a_{11} \mathbb{1}_{\{x_1=1, y_1=1\}} + a_{21} \mathbb{1}_{\{x_1=1, y_2=1\}} + a_{12} \mathbb{1}_{\{x_2=1, y_1=1\}} + a_{22} \mathbb{1}_{\{x_2=1, y_2=1\}},
\end{aligned} \tag{3}$$

(4)

donde, la segunda igualdad se debe a la forma en la que se definieron las indicadoras $\{Z_j : j \in \mathbb{N}\}$, mientras que, la penúltima igualdad se debe a la definición de $\{s_j : j \in \mathbb{N}\}$, finalmente, para deducir la última igualdad, se uso que el estado en el que se encuentra la serie de tiempo, forma una cadena de Markov con matriz de transición A.[5] Ahora, como puede infererir del cálculo en (3), es posible calcular para $i,h \in \{1,2\}$ el valor de a_{hi}, en términos de que componentes de Z_{j+1} y Z_j deben ser iguales a uno, al proceder de la siguiente manera:

$$a_{hi} = \mathbb{P}[Z_{i,j+1} = 1 | Z_{hj} = 1], \text{ con } i,h \in \{1,2\}. \tag{5}$$

Por lo que, es posible expresar a las probabilidades de transición de la cadena $\{Z_j : n \in \mathbb{N}\}$, en términos de las componentes de cada variable que forma la cadena.

Ahora, tome $n \in \mathbb{N}$ y para $j \in \{1, \ldots, n\}$ denote por

$$\overline{y}_j = (w_1, w_2, \ldots, w_j)'.$$

[4] Que describe el cambio de estado en la serie de tiempo.

[5] La propiedad de Markov para el proceso $\{Z_j : j \in \mathbb{N}\}$, se puede probar haciendo uso de la propiedad de Markov que describe el estado de la serie de tiempo, al realizar exactamente lo mismo que se hizo en (3), iniciando con $\mathbb{P}[Z_{j+1} = x | Z_j = y, Z_{j-1} = y_{j-1}, \ldots, Z_1 = y_1]$ con $x, y, y_{j-1}, \ldots, y_1 \in E$.

Luego, suponga que se ha observado el vector $\overline{y}_n$, entonces, bajo el contexto dado en la introducción de este problema, puede probarse que para $j \in \{1, \ldots, n\}$, se tiene que:[6]

$$w_j|\overline{y}_{j-1}, s_j \sim N(\beta_{s_j} w_{j-1}, \sigma^2). \tag{6}$$

donde, para $j = 1$ la notación $w_1|\overline{y}_0, s_1$, debe interpretarse simplemente como $w_1|s_1$. Luego, sea $\theta = (\beta_1, \beta_2, \sigma^2)'$, entonces este modelo posee un vector de parámetros desconocidos Φ, el cual consta de los cuatro elementos en la matriz A, más los elementos en el vector θ. De este modo, si se desea estimar el vector de parámetros Φ el algoritmo EM es claramente el algoritmo a considerar, pues, inicialmente contamos con un vector de datos observados $\overline{y}_n$, no obstante, dichos datos no representan toda la información existente sobre los parámetros a estimar, pues, por el contexto que se ha dado se desconoce el valor de los indicadores $\{Z_j : j \in \{1, \ldots, n\}\}$, por lo que, estos datos representan un conjunto de datos no observados.

Ahora, para llevar a cabo el algoritmo EM debe calcularse primeramente la verosimilitud de la muestra completa. Para ello, note que si $f(\overline{y}_n, Z|\phi)$ representa a la densidad generalizada de $(\overline{y}_n, Z)$, donde $Z = \overline{Z}_n = (Z_1, \ldots, Z_n)$, entonces

$$\begin{aligned} f(\overline{y}_n, Z; \Phi) &= f(w_1, \ldots, w_n|Z; \Phi) f(Z; \Phi) \\ &= f(w_n|w_{n-1}, \ldots, w_1, Z; \Phi) \cdots f(w_2|w_1, Z; \Phi) f(w_1|Z; \Phi) f(Z; \Phi) \\ &= f(w_n|\overline{y}_{n-1}, \ldots, w_1, Z; \Phi) \cdots f(w_2|\overline{y}_1, Z; \Phi) f(w_1|Z; \Phi) f(Z; \Phi) \\ &= \left[\prod_{j=1}^{n} f(w_j|\overline{y}_{j-1}, Z; \Phi) \right] f(Z; \Phi), \end{aligned} \tag{7}$$

donde, en la primer igualdad se empleado la Ley de probabilidad total, mientras que, para la segunda igualdad se ha usado la Ley de condicionamiento Iterado. [7] Ahora, es posible probar que:[8]

$$f(w_j|\overline{y}_{j-1}, Z; \Phi) = f(w_j|\overline{y}_{j-1}, s_j; \theta). \tag{8}$$

Más aún, en virtud de que para $j \in \{1, \ldots, n\}$, se cumple que:

$$Z_j = \begin{pmatrix} Z_{1j} & Z_{2j} \end{pmatrix} = \begin{cases} (1,0) & \text{si } s_j = 1, \\ (0,1) & \text{si } s_j = 2. \end{cases}$$

Entonces, el lado derecho de (8) puede escribirse como

$$f(w_j|\overline{y}_{j-1}, s_j; \theta) = \prod_{i=1}^{2} f^{Z_{ij}}(w_j|\overline{y}_{j-1}, s_j = i; \theta). \tag{9}$$

[6] De manera intuitiva, observe que dado y_{j-1} y s_j, se conoce el valor de w_{j-1} y el estado de la cadena en que se encuentra w_j, por ende, también se la constante β_{s_j}, así, es posible aplicar de manera ingenua el Lema 3.2 aunado al hecho de que $\varepsilon_j \sim N(0, \sigma^2)$, para obtener el resultado citado.

[7] Ver anexo. Adicionalmente, cabe destacar que se esta suponiendo que $(\overline{y}_n, Z)$ tiene densidad conjunta generalizada, es decir, se esta suponiendo que $\overline{y}_n$ es absolutamente continua.

[8] Intuitivamente, dado $j \in \{1, \ldots, n\}$ si se conoce el vector al completo de etiquetas Z, entonces, se conoce el valor en particular de Z_j y, dado que para $i \in \{1, 2\}$ se cumple que $s_j = i$ si y solo si $Z_{ij} = 1$, entonces, también se conoce el valor de s_j. De manera un poco menos intuitiva, la propiedad de Markov de Z_j es suficiente para probar esta propiedad.

A modo de ejemplo, del porque se da la igualdad anterior, note que en el evento $s_j = 1$ se tiene que $f(w_j|\overline{y}_{j-1}, s_j; \theta)$ es igual a $f(w_j|\overline{y}_{j-1}, s_j = 1; \theta)$, mientras que, por la definición de Z_{1j} y Z_{2j}, se tiene que en el evento $s_j = 1$ se cumple que $Z_{1j} = 1$ y $Z_{2j} = 0$, entonces:

$$\begin{aligned}\prod_{i=1}^{2} f^{Z_{ij}}(w_j|\overline{y}_{j-1}, s_j = i; \theta) &= f^{Z_{1j}}(w_j|\overline{y}_{j-1}, s_j = 1; \theta) f^{Z_{2j}}(w_j|\overline{y}_{j-1}, s_j = 2; \theta) \\ &= f^{1}(w_j|\overline{y}_{j-1}, s_j = 1; \theta) f^{0}(w_j|\overline{y}_{j-1}, s_j = 2; \theta) \\ &= f(w_j|\overline{y}_{j-1}, s_j = 1; \theta),\end{aligned}$$

lo que corrobora en este caso la igualdad en (9), mientras que, un razonamiento similar sirve al analizar el evento en el que $s_j = 2$. De este modo, por la igualdad en (8) y la igualdad en (9), es posible escribir la ecuación en (7) de manera equivalente, como

$$f(y_n, Z; \Phi) = \left[\prod_{j=1}^{n}\prod_{i=1}^{2} f^{Z_{ij}}(w_j|\overline{y}_{j-1}, s_j = i; \theta)\right] f(Z; \Phi)$$

Así, la logverosimilitud de los datos completos esta dada por:

$$L_c(\Phi) = \log(f(y_n, Z; \Phi)) = \sum_{j=1}^{n}\sum_{i=1}^{2} Z_{ij} \log f(w_j|\overline{y}_{j-1}, s_j = i; \theta) + \log f(Z; \Phi).$$

Ahora, para $j \in \{1, \ldots, n\}$ e $i \in \{1, 2\}$ defina

$$f_i(w_j|\overline{y}_{j-1}; \theta) = f(w_j|\overline{y}_{j-1}, s_j = i; \theta)$$

entonces, por lo comentado en (6), se tiene que $f_i(w_j|\overline{y}_{j-1}; \theta)$ corresponde a una densidad Normal con media $\beta_i w_{j-1}$ y varianza σ^2 evaluada en w_j, más aún, se sigue que

$$L_c(\Phi) = \sum_{j=1}^{n}\sum_{i=1}^{2} Z_{ij} \log f_i(w_j|\overline{y}_{j-1}; \theta) + \log f(Z; \Phi). \tag{10}$$

Ahora, se calculara

$$Q(\Phi|\overline{y}_n, \Phi^*) = \mathbb{E}[\log(L_c(\Phi))|\overline{y}_n, \Phi^*].$$

Para ello, note que al tomar esperanzas condicionales a ambos lados de la igualdad en (10), se obtiene que

$$\begin{aligned}Q(\Phi|\overline{y}_n; \Phi^*) &= \mathbb{E}[\log(L_c(\Phi))|\overline{y}_n; \Phi^*] \\ &= \sum_{j=1}^{n}\sum_{i=1}^{2} \mathbb{E}\left[Z_{ij} \log f_i(w_j|\overline{y}_{j-1}; \theta)|\overline{y}_n; \Phi^*\right] + \mathbb{E}\left[\log f(Z; \Phi)|\overline{y}_n; \Phi^*\right] \\ &= \sum_{j=1}^{n}\sum_{i=1}^{2} \left[\mathbb{E}\left[Z_{ij}|\overline{y}_n; \Phi^*\right] \log f_i(w_j|\overline{y}_{j-1}; \theta)\right] + \mathbb{E}\left[\log f(Z; \Phi)|\overline{y}_n; \Phi^*\right].\end{aligned}$$

donde, en la primer igualdad se ha usado la linealidad de la esperanza condicional, mientras que, en la segunda igualdad se ha empleado el hecho de que

$$\log f_i(w_j|\overline{y}_{j-1};\theta),$$

es una cantidad conocida dado $\overline{y}_n$, pues, para $j \in \{1,\dots,n\}$ se tiene que $\overline{y}_{j-1}$ es un subvector de $\overline{y}_n$ y w_j es la componente j-ésima de $\overline{y}_n$, por lo que, este término sale de la esperanza condicional.[9] Como comentario adicional, en el texto de McLachlan y Krishnan (2008) al sacar este termino de la esperanza, se cambia el parámetro θ por Φ, no obstante esto no es necesario, pues al recordar que Φ es el vector de parámetros $\Phi = (\theta, a_{11}, a_{12}, a_{21}, a_{22})$, entonces, lo que indica la notación $\log f_i(w_j|\overline{y}_{j-1};\theta)$ es que dicho término depende de Φ únicamente a través de θ. De esta manera, se tiene que

$$Q(\Phi|\overline{y}_n;\Phi^*) = \sum_{j=1}^{n}\sum_{i=1}^{2}\left[\mathbb{E}\left[Z_{ij}|\overline{y}_n;\Phi^*\right]\log f_i(w_j|\overline{y}_{j-1};\theta)\right] + \mathbb{E}\left[\log f(Z;\Phi)|\overline{y}_n;\Phi^*\right].$$

Finalmente, denote por $P_{\Phi^*}(Z) = \mathbb{E}\left[\log f(Z;\Phi)|\overline{y}_n;\Phi^*\right]$,[10] entonces, la igualdad anterior puede expresarse de forma equivalente como:

$$Q(\Phi|\overline{y}_n;\Phi^*) = \sum_{j=1}^{n}\sum_{i=1}^{2}\left[\mathbb{E}\left[Z_{ij}|\overline{y}_n;\Phi^*\right]\log f_i(w_j|\overline{y}_{j-1};\theta)\right] + P_{\Phi^*}(Z).$$

De este modo, después de p-iteraciones del algoritmo EM, el paso E del mismo consiste en calcular $Q(\Phi|\overline{y}_n;\Phi^{(P)})$, donde $\Phi^{(p)}$ es la estimación para el vector de parámetros Φ obtenida en la p-ésima iteración del algoritmo, así, se debe de calcular

$$Q(\Phi|\overline{y}_n;\Phi^{(p)}) = \sum_{j=1}^{n}\sum_{i=1}^{2}\left[\mathbb{E}\left[Z_{ij}|\overline{y}_n;\Phi^{(p)}\right]\log f_i(w_j|\overline{y}_{j-1};\theta)\right] + P_{\Phi^{(p)}}(Z). \tag{11}$$

De este modo, si para $i \in \{1,2\}$ y $j \in \{1,\dots,n\}$ se define $\tau_{ij}^{(p)}$ como

$$\tau_{ij}^{(p)} = \mathbb{E}\left[Z_{ij}|\overline{y}_n;\Phi^{(p)}\right] = \mathbb{P}\left[Z_{ij}=1|\overline{y}_n;\Phi^{(p)}\right].$$

donde, la última igualdad se debe a la definición de esperanza y al hecho de que Z_{ij} solo toma valores en $\{0,1\}$. Entonces, es posible expresar la ecuación en (11), como

$$Q(\Phi|\overline{y}_n;\Phi^{(p)}) = \sum_{j=1}^{n}\sum_{i=1}^{2}\left[\tau_{ij}^{(p)}\log f_i(w_j|\overline{y}_{j-1};\theta)\right] + P_{\Phi^{(p)}}(Z). \tag{12}$$

Así, el paso E consiste en calcular explícita o computacionalmente la función $Q(\Phi|\overline{y}_n;\Phi^{(p)})$, al calcular tanto las constantes $\tau_{ij}^{(p)}$ como $P_{\Phi^{(p)}}(Z)$.[11] Por otro lado, para explicar en que consiste el

[9] Formalmente, lo que se tiene es que $\overline{y}_{j-1}$ y w_j son medibles respecto a la σ-álgebra generada por $\overline{y}_n$, por lo que, $\log f_i(w_j|\overline{y}_{j-1};\theta)$ puede salir de la esperanza condicional, pues, es una función de variables medibles respecto a la σ-álgebra generada por $\overline{y}_n$.

[10] En McLachlan y Krishnan (2008) se hace uso de una notación parecida sin especificar su significado, por lo cual, en clase con el Dr. Quiroga se discutió el mismo y se llego a la conclusión anterior.

[11] Pues, las funciones $f_i(w_j|\overline{y}_{j-1};\theta)$ son simplemente densidades normales de parámetros desconocidos.

paso M, será necesario recordar que para $j \in \{1, \ldots, n\}$ e $i \in \{1, 2\}$, se tiene que $f_i(w_j|\overline{y}_{j-1}; \theta)$ es una densidad Normal de media $\beta_j w_{j-1}$ y varianza σ^2 evaluada en w_j,[12] es decir

$$f_i(w_j|\overline{y}_{j-1}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(w_j - \beta_i w_{j-1})^2\right].$$

pues, de este modo se obtiene que

$$\log f_i(w_j|\overline{y}_{j-1}; \theta) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(w_j - \beta_i w_{j-1})^2}{2\sigma^2}.$$

Así, es posible escribir la igualdad en (12) como

$$Q(\Phi|\overline{y}_n; \Phi^{(p)}) = \sum_{j=1}^{n}\sum_{i=1}^{2}\left[\tau_{ij}^{(p)}\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(w_j - \beta_i w_{j-1})^2}{2\sigma^2}\right)\right] + P_{\Phi^{(p)}}(Z).$$

Ahora, para el paso M se debe obtener la actualización para el vector de parámetros

$$\Phi = (\theta, a_{11}, a_{12}, a_{21}, a_{22}).$$

Para ello, se procederá a máximizar respecto a Φ a la función Q y se establecerá por $\Phi^{(p+1)} = \text{argmax}_\Phi\, Q(\Phi|\overline{y}_n; \Phi^{(p)})$. Con esto en mente, note que para $l \in \{1, 2\}$, se sigue que

$$\begin{aligned}
\frac{\partial}{\partial\beta_l}\left[Q(\Phi|\overline{y}_n; \Phi^{(p)})\right] &= \sum_{j=1}^{n}\sum_{i=1}^{2}\left[\tau_{ij}^{(p)}\frac{\partial}{\partial\beta_l}\left[\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(w_j - \beta_i w_{j-1})^2}{2\sigma^2}\right)\right]\right] \\
&= \sum_{j=1}^{n}\left[\tau_{ij}^{(p)}\frac{\partial}{\partial\beta_l}\left[\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(w_j - \beta_l w_{j-1})^2}{2\sigma^2}\right)\right]\right] \\
&= \sum_{j=1}^{n}\left[\tau_{ij}^{(p)}\frac{2w_{j-1}(w_j - \beta_l w_{j-1})}{2\sigma^2}\right] = \sum_{j=1}^{n}\left[\tau_{ij}^{(p)}\frac{w_{j-1}(w_j - \beta_l w_{j-1})}{\sigma^2}\right],
\end{aligned}$$

donde, la primer igualdad se debe a que $\tau_{ij}^{(p)}$ y $P_{\Phi^{(p)}}$ son constantes respecto a β_l. Así, al igualar a cero las parciales anteriores, se deduce que la actualización para el coeficiente β_l, dada después de p iteraciones del método EM, en el paso M y denotada por $\beta_l^{(p+1)}$, es la solución a la ecuación:

$$\sum_{j=1}^{n}\left[\tau_{ij}^{(p)}\frac{w_{j-1}(w_j - \beta_l^{(p+1)} w_{j-1})}{\sigma^2}\right] = 0, \text{ con } l \in \{1, 2\}.$$

O equivalentemente, es la solución a la ecuación

$$\sum_{j=1}^{n}\left[\tau_{ij}^{(p)} w_{j-1}(w_j - \beta_l^{(p+1)} w_{j-1})\right] = 0, \text{ con } l \in \{1, 2\}.$$

Así, despejando $\beta_l^{(p+1)}$ de la ecuación anterior, se obtiene que

$$\beta_l^{(p+1)} = \frac{\sum_{j=1}^{n}\tau_{ij}^{(p)} w_{j-1} w_j}{\sum_{j=1}^{n}\tau_{ij}^{(p)} w_j^2} = \frac{\sum_{j=1}^{n} w_j^{(p)}\overline{w}_j^{(p)}}{\sum_{j=1}^{n}(w_j^{(p)})^2}, \text{ con } l \in \{1, 2\}. \tag{13}$$

[12] Por (6) ya que $f_i(w_j|\overline{y}_{j-1}; \theta) = f(w_j|\overline{y}_{j-1}, s_j = i; \theta)$

donde, para $j \in \{1, \ldots, n\}$ se tiene que

$$w_j^{(k)} = \sqrt{\tau_{ij}^{(p)}} w_j, \quad \overline{w}_j^{(p)} = \sqrt{\tau_{ij}^{(p)}} w_{j-1}.$$

Por otro lado, note que

$$\begin{aligned}\frac{\partial}{\partial \sigma}\left[Q(\Phi|\overline{y}_n; \Phi^{(p)})\right] &= \sum_{j=1}^{n}\sum_{i=1}^{2}\left[\tau_{ij}^{(p)}\frac{\partial}{\partial \sigma}\left[\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(w_j - \beta_i w_{j-1})^2}{2\sigma^2}\right)\right]\right] \\ &= \sum_{j=1}^{n}\sum_{i=1}^{2}\left[\tau_{ij}^{(p)}\left[-\frac{1}{\sigma} + \frac{w_j - \beta_i w_{j-1}}{\sigma^3}\right]\right],\end{aligned}$$

donde, la primer igualdad se debe a que $\tau_{ij}^{(p)}$ y $P_{\Phi^{(p)}}$ son constantes respecto a σ. De este modo, al igualar a cero la parcial anterior, se sigue que la actualización para el coeficiente σ^2 dada en el paso M y denotada por $(\sigma^{(p+1)})^2$, es la solución a la ecuación:

$$\sum_{j=1}^{n}\sum_{i=1}^{2}\left[\tau_{ij}^{(p)}\left[-\frac{1}{\sigma^{(p+1)}} + \frac{w_j - \beta_i^{(p+1)} w_{j-1}}{(\sigma^{(p+1)})^3}\right]\right] = 0.$$

Así, al despejar $(\sigma^{(p+1)})^2$ de la ecuación anterior, se obtiene

$$(\sigma^{(p+1)})^2 = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{2}\left[\tau_{ij}^{(p)}(w_j - \beta_i^{(p+1)} w_{j-1})^2\right]. \tag{14}$$

Finalmente, Hamilton (1900) probó que si $\hat{\Phi}$ denota al estimador de máxima verosimilitud de Φ, entonces, para $h, l \in \{1, 2\}$ el estimador de máxima verosimilitud $\hat{a}_{hl}$ de a_{hl}, satisface la ecuación

$$\hat{a}_{hl} = \frac{\sum_{j=1}^{n-1}\mathbb{P}[Z_{hj} = 1, Z_{i,j+1} = 1|\overline{y}_n; \hat{\Phi}]}{\sum_{j=1}^{n-1}\mathbb{P}[Z_{hj} = 1|\overline{y}_n; \hat{\Phi}]}$$

Así, las estimaciones de las probabilidades de transición pueden actualizarse haciendo uso de la igualdad anterior, como la solución a la ecuación

$$\begin{aligned}a_{hl}^{(p+1)} &= \frac{\sum_{j=1}^{n-1}\mathbb{P}[Z_{hj} = 1, Z_{i,j+1} = 1|\overline{y}_n; \Phi^{(p+1)}]}{\sum_{j=1}^{n-1}\mathbb{P}[Z_{hj} = 1|\overline{y}_n; \Phi^{(p+1)}]} \\ &= \frac{\sum_{j=1}^{n-1}\mathbb{P}[Z_{hj} = 1, Z_{i,j+1} = 1|\overline{y}_n; (\theta^{(p+1)}, a_{11}^{(p+1)}, \ldots, a_{22}^{(p+1)})]}{\sum_{j=1}^{n-1}\mathbb{P}[Z_{hj} = 1|\overline{y}_n; \theta^{(p+1)}, a_{11}^{(p+1)}, \ldots, a_{22}^{(p+1)}]},\end{aligned}$$

donde, $\theta^{(p+1)} = (\beta_1^{(p+1)}, \beta_2^{(p+1)}, (\sigma^{(p+1)})^2)$ con $(\sigma^{(p+1)})^2$ como en (14) y con $\left(\beta_1^{(p+1)}, \beta_2^{(p+1)}\right)$ como en (13). ■

Ejercicio 2:

Solución. El siguiente ejemplo fue se basa en gran medida en la pagina 545 de Hastie y col. (2009). Primeramente, se simulo en Julia una muestra aleatoria de 450 variables aleatorias uniformes en $[0, 2\pi]$ y, otra muestra aleatoria de 900 variables normales de media 0 y desviación estándar 0.25, así, denote por $\{\Theta_n : n \in \{1, \ldots, 450\}\}$ a la muestra aleatoria de las uniformes y por $\{Z_n : n \in \{1, \ldots, 900\}\}$ a la muestra Normal. A partir de las muestras anteriores se simularon tres conjuntos de datos, cada uno con tamaño 150, cuyos puntos se distribuyeran de manera uniforma en los círculos de radios 1, 2.8 y 5 respectivamente y, a cada coordenada de cada punto de estos conjuntos se le agrego un ruido gausiano con desviasión estándar 0.25, de la siguiente manera:

Datos para el circulo de radio 1 + perturbación Gaussiana.

$$X_i^{(11)} = \cos(\Theta_i) + Z_{2i-1},$$
$$X_i^{(12)} = \sin(\Theta_i) + Z_{2i}, \text{ con } i \in \{1, \ldots, 150\}.$$

Datos para el circulo de radio 2.8 + perturbación Gaussiana.

$$X_i^{(21)} = 2.8\cos(\Theta_i) + Z_{2i-1},$$
$$X_i^{(22)} = 2.8\sin(\Theta_i) + Z_{2i}, \text{ con } i \in \{151, \ldots, 300\}.$$

Datos para el circulo de radio 5 + perturbación Gaussiana.

$$X_i^{(31)} = 5\cos(\Theta_i) + Z_{2i-1},$$
$$X_i^{(32)} = 5\sin(\Theta_i) + Z_{2i}, \text{ con } i \in \{301, \ldots, 450\}.$$

De este modo, se tienen tres conjuntos de pares de puntos, los cuales se denotaran por:

$$C_j = \left\{(X_i^{(j1)}, X_i^{(j2)}) : i \in \{(j-1)\cdot 150 + 1, \ldots, j \cdot 150\}\right\} \text{ con } j \in \{1, 2, 3\}.$$

Luego, haciendo uso de estos conjuntos de datos simulados, se construyo la gráfica en la esquina superior izquierda de la Figura 1, en la cual se puede observar en color naranja los 150 puntos del conjunto C_1, es decir, aquellos que fueron simulados con un radio 1, en azul los 150 puntos del conjunto C_2, es decir, aquellos que fueron simulados con un radio 2.8 y en verde los 150 puntos del conjunto C_3, es decir, aquellos que fueron simulados con un radio de 5. Ahora, para llevar a cabo el algoritmo de Clustering Espectral se cálculo una matriz de similaridades W y el Laplaciano L para estos datos, de la siguiente forma:

1. Se calculó la distancia euclideana entre todos los puntos de los conjuntos $\{C_j : j \in \{1, 2, 3\}\}$ y, se calculo una medida de similaridad para estos puntos, de la siguiente manera:

   Para $j, k \in \{1, 2, 3\}$, $i \in \{150 \cdot (j-1) + 1, \ldots, 150 \cdot j\}$ y $l \in \{150 \cdot (k-1) + 1, \ldots, 150 \cdot k\}$, la distancia euclidiana entre los puntos $(X_i^{(j1)}, X_i^{(j2)}) \in C_j$ y $(X_l^{(k1)}, X_l^{(k2)}) \in C_k$ se denotará por $d_{i,l}$ y, esta dada por:

$$d_{i,l} = \sqrt{\left(X_i^{(j1)} - X_l^{(k1)}\right)^2 + \left(X_i^{(j2)} - X_l^{(k2)}\right)^2}.$$

Haciendo uso de esta distancia, se definió $s_{i,l} = \exp\left\{-\beta d_{i,l}^2\right\}$ donde β debe ser una constante positiva, que para fines de reproducción del ejercicio se tomo igual a $1/2$. A modo de observación, $s_{i,l}$ es la medida de similaridad mencionada a la cual llamaremos distancia radial entre $(X_i^{(j1)}, X_i^{(j2)})$ y $(X_l^{(k1)}, X_l^{(k2)})$. Así, sea S la matriz en $M_{450\times 450}(\mathbb{R})$ tal que

$$S_{ij} = s_{i,j}.$$

Entonces, S es simétrica pues las distancias radiales lo son, ya que están basadas en la métrica euclidiana y los primeros 150 índices de S corresponden a puntos en el conjunto C_1, los siguientes 150 indices a puntos en el conjunto C_2 y los últimos 150 índices de S corresponden a puntos en el conjunto C_3. Lo anterior, en el sentido de que si se eligen $j, k \in \{1, 2, 3\}$, $i \in \{150 \cdot (j-1) + 1, \ldots, 150 \cdot j\}$ y $l \in \{150 \cdot (k-1) + 1, \ldots, 150 \cdot k\}$, entonces, S_{il} contiene la distancia radial del punto $(X_i^{(j1)}, X_i^{(j2)})$ en C_j, al punto $(X_l^{(k1)}, X_l^{(k2)})$ en C_k.

2. Ahora, para cada punto se eligieron los 10 vecinos con mayor distancia radial a ellos, i.e, con menor distancia euclídea a ellos, de la siguiente manera:

   Para $j \in \{1, 2, 3\}$ e $i \in \{150 \cdot (j-1) + 1, \ldots, 150 \cdot j\}$ sea

   $$ind_i = \{S_{il} : l \in \{1, \ldots, 450\}\}.$$

   Entonces, los 10 vecinos con mayor distancia radial a i, son los 10 índices j en ind_i para los que se alcanzan los 10 mayores valores de S_{ij}.[13] Así, para $j \in \{1, 2, 3\}$ e $i \in \{150 \cdot (j-1) + 1, \ldots, 150 \cdot j\}$, se denotará al conjunto de 10 vecinos con mayor distancia radial a i como $V(i)_{10}$. Bajo esta definición, note que si $j \in V(i)_{10}$ entonces: Si j esta entre los primeros 150 índices se sigue que i tiene un vecino cercano en el conjunto C_1, si j esta entre los índices 151 hasta 350, entonces, i tiene un vecino en el conjunto C_2 y algo similar para C_3, esto debido a la manera en la que están asignados los índices de la matriz S, a los conjuntos de puntos C_1, C_2 y C_3.

3. Finalmente, se obtuvo la matriz de similaridades $W \in M_{450\times 450}(\mathbb{R})$ y el Laplaciano $L \in M_{450\times 450}(\mathbb{R})$ para este conjunto de puntos, de la siguiente manera:

   Para $j, k \in \{1, 2, 3\}$, $i \in \{150 \cdot (j-1) + 1, \ldots, 150 \cdot j\}$ y $l \in \{150 \cdot (k-1) + 1, \ldots, 150 \cdot k\}$, defina la entrada il de W como:

   $$W_{il} = \begin{cases} S_{il} & \text{si } i \in V(l)_{10} \text{ o } l \in V(i)_{10} \\ 0 & \text{e.o.c} \end{cases}$$

   Finalmente, defina para cada $i \in \{1, \ldots, 450\}$ la cantidad d_i como $d_i = \sum_{j=1}^{450} W_{ij}$ y, defina $D = Diag(d_1, \ldots, d_{450})$. Entonces, el Laplaciano de este problema esta dado por

   $$L = D - W.$$

   Como comentario extra, al igual que los índices en las matrices anteriores, los primeros 150 índices de L y W se pueden asociar a puntos en C_1, los siguientes 150 a puntos en C_2 y los últimos 150 a puntos en C_3.

[13] Sin contar al índice i.

Cabe destacar que, todos los cálculos anteriores se llevaron a cabo en Julia y pueden consultarse en el Script adjunto a este trabajo. Adicionalmente, una vez obtenido el Laplaciano L se calcularon sus 15 valores propios más pequeños, con lo que, se construyo la gráfica mostrada en la esquina superior derecha de la Figura 1. Finalmente, de acuerdo con Hastie y col. (2009) una vez teniendo el Laplaciano L, el método de Clustering Spectral consiste en encontrar los m vectores propios asociados a los m valores propios más pequeños de L, ignorando el vector propio asociado al valor propio trivial 0, los cuales se pueden acomodar como columnas de una matriz $Z \in M_{450\times m}(\mathbb{R})$, para luego proceder a aplicar algún algoritmo como K-medias para formar clusters de las filas de Z, con lo que, es posible inducir una partición del conjunto de datos original. Por ejemplo, en este caso se calcularon los vectores asociados con el segundo y tercer valor propio más pequeños de L, con lo que, se obtuvo la siguiente matriz $Z \in M_{450,2}(\mathbb{R})$:

$$Z = \begin{pmatrix} -0.0426967 & 0.0530574 \\ -0.042046 & 0.0500462 \\ -0.042656 & 0.0527221 \\ \vdots & \vdots \\ 0.0756577 & 0.0230397 \\ 0.0692118 & 0.0137433 \\ 0.0740207 & 0.0207076 \end{pmatrix}$$

La matriz anterior, contiene en su primer columna al vector propio normalizado de L asociado al segundo valor propio más pequeño de L, mientras que, su tercer columna esta constituida por el vector propio normalizado de L asociado al tercer valor propio más pequeño de L. Así, dado que los primeros 150 índices de L se pueden asociar a puntos en C_1, los segundos 150 índices de L se asocian a puntos en C_2 y los últimos 150 índices de L pueden asociarse a puntos en C_3, entonces, los vectores propios anteriores se vinculan de la misma forma a los puntos en estos conjuntos.[14] Bajo esta idea, se construyo la gráfica en la esquina inferior izquierda de la Figura 1, en dicha gráfica se puede apreciar una gráfica de las columnas de Z contra su índice. Cabe destacar que los puntos en ambas gráficas, se han coloreado de acuerdo al criterio de que los primeros 150 índices de ambas columnas corresponden al grupo de puntos C_1, los siguientes 150 índices al grupo C_2 y los últimos 150 índices corresponden al conjunto de puntos C_3, usando de manera respectiva los colores naranja, azul y verde. Como puede puede notar, la relación que se mencionó tienen estos vectores propios con los conjuntos C_1, C_2 y C_3, se vuelve notoria al comparar esta gráfica con la gráfica en el panel en la esquina superior izquierda de la Figura 1, pues, recuerde que en esta última gráfica los conjuntos C_1, C_2 y C_3 se han coloreado igualmente con los colores naranja, azul y verde de forma respectiva. Finalmente, note que las filas de Z inducen pares de puntos en $\mathbb{R}^2$, los cuales pueden verse en la gráfica en la esquina inferior derecha de la Figura 1, los colores se seleccionaron nuevamente como naranja para las primeras 150 filas de Z, azul para las siguientes 150 filas de Z y verde para las últimas 150 filas, como puede ver, en esta gráfica se forman tres conjuntos de puntos que corresponden, por color naranja, azul y verde, a los grupos C_1, C_2 y C_3 y, como se menciono con anterioridad, a este conjunto de puntos es al que se le puede aplicar un algoritmo como K-medias, para inducir una agrupación para los puntos originales. Este ultimo paso no fue solicitado, no obstante si no se conocieran los grupos, uno puede notar de manera clara que en la gráfica en la esquina inferior derecha de la Figura 1, existen tres grupos distintos de puntos, así, en la Figura 2 se muestra la misma gráfica coloreada de acuerdo a una clasificación hecha mediante el

[14] Dado que, en particular son vectores propios de L.

algoritmo de K-medias, evidentemente, se aprovecho que ya se conoce la pertenencia de los diversos puntos en esta gráfica a los conjuntos originales, para colorear los clusters obtenidos por K-medias de la forma en la que se colorearon, pues, al hacer esto se puede notar que el algoritmo de K-medias separa perfectamente los tres conjuntos de puntos existentes. Así, bajo esta idea de los índices, se puede obtener la partición deseada del conjunto original.

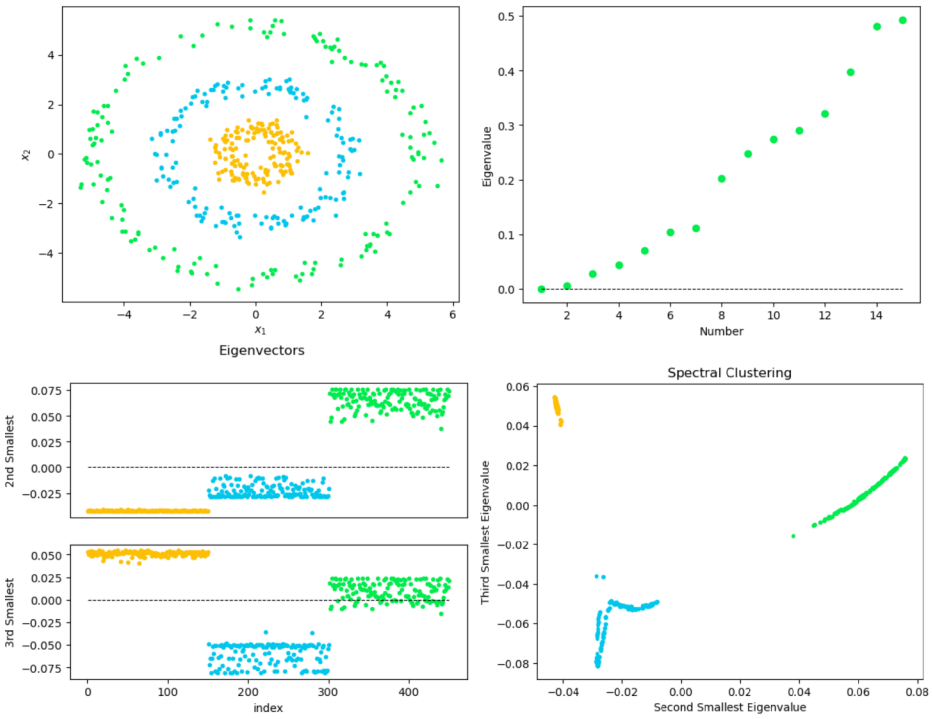

Figura 1: Replica de la gráfica 14.29 en Hastie y col. (2009). En la esquina superior izquierda se aprecia una gráfica de los datos simulados, separados por los colores verde, azul y rojo, de acuerdo al diámetro 5, 2.8 y 1 con el que fueron simulados. En la esquina superior derecha puede encontrar una gráfica con los 15 valores propios más pequeños del Laplaciano, L, de este problema de clasificación espectral. Por otra parte, la gráfica en la esquina inferior izquierda corresponde a las gráficas del segundo y tercer vector propio más pequeños de L. Finalmente, la última gráfica corresponde a una gráfica de dispersión del segundo vector propio más pequeño de L, contra el tercer vector propio más pequeño de L.

■

3. Anexo.

Se enunciara de manera rápida y poco formal versiones de la Ley de Condicionamiento Iterado y la Ley de Probabilidad Total, que fueron de utilidad a lo largo del ejercicio 1.:

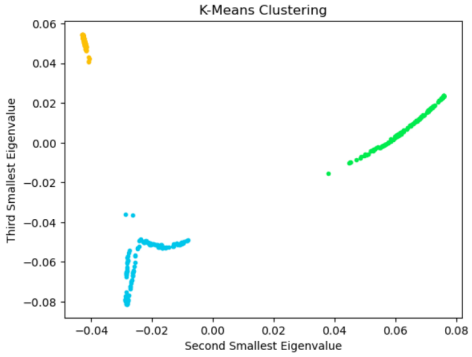

Figura 2: Agrupación usando K-means clustering.

Lema 3.1. *Sean $\overline{Y}_n = (Y_1, \ldots, Y_n)$ un vector aleatorio absolutamente continuo, $Z = (Z_1, \ldots, Z_n)$ un vector aleatorio discreto y $f(\overline{Y}_n, Z)$ la densidad generalizada de $(\overline{Y}_n, Z)$. Entonces*

1. *(Ley de Probabilidad Total.) Se cumple que*

$$f(\overline{Y}_n, Z) = f(\overline{Y}_n|Z)f(Z).$$

donde, $f(\overline{Y}_n|Z)$ es la densidad de Y dado Z y $f(Z)$ es la función masa de probabilidades de Z.

2. *(Ley de Condicionamiento Iterado.) Sea $f(Y|Z)$ como en el inciso anterior. Si para $j \in \{2, \ldots, n\}$ se define $\overline{Y}_{j-1} = (Y_1, \ldots, Y_{j-1})$, entonces*

$$f(Y_n|\overline{Y}_{n-1}, Z) \cdots f(Y_2|\overline{Y}_1, Z)f(Y_1|\overline{Y}_0, Z).$$

donde, para $j \in \{2, \ldots, n-1\}$ se tiene que $f(Y_j|\overline{Y}_{j-1}, Z)$ denota la densidad condicional de Y_j dado $\overline{Y}_{j-1}, Z$, mientras que, $f(Y_1|\overline{Y}_0, Z) = f(Y_1|Z)$ denota a la densidad condicional de Y_1 dado Z.

Adicionalmente, se enunciará un Lema sobre distribuciones normales que fue empleado a lo largo del Ejercicio.

Lema 3.2. *Sean Y una variable aleatoria normal, de media μ y varianza σ^2, y a y b dos números reales. Entonces, $aY + b$ posee distribución normal de media $a\mu + b$ y varianza $a^2\sigma^2$.*

Referencias

Hastie, T., Tibshirani, R. & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed). Springer.

Martín González, E. M. (2021). *Probabilidad* (Notas de clase).

McLachlan, G. J. & Krishnan, T. (2008). *The EM algorithm and extensions* (2nd ed) [OCLC: ocn137325058]. Wiley-Interscience.