UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

Tesis de Licenciatura

Selección de modelos: una aplicación a datos biológicos

Yamila Mercedes Barrera

Directora: Dra. Mariela Sued
Codirectora: Dra. Daniela Rodriguez

Diciembre de 2014

Agradecimientos

A mi mamá y mi papá, por su amor y apoyo incondicional.
A mis hermanos Araceli y Francisco, por llenar de alegría mi vida y estar en todo momento.
A Dani Cuesta, por su amistad infinita.
A mi comunidad del Movi, porque caminar juntos hace todo más lindo y sencillo.
Al grupo de jóvenes viejitos de Cristo Rey.
A todos mis compañeros matemáticos con quienes compartimos travesias en el 28, charlas, tardes de estudio, miles de mates, viajes a la UMA. Gracias por llenar de sonrisas estos años y por el empujón que me dieron en este último tramo.
A Cecilia Lorusso, por enseñar matemática con una pasión que contagia.
A mis alumnos, compañeros y directivos del Colegio Emaús por todo el apoyo en este camino.
A mis alumnos y compañeros del Apoyo Escolar.
A los Zapp, porque son un ejemplo de vida y su historia merece ser compartida.
A las investigadoras del laboratorio de Dinámica y Transporte Intracelular, por ayudarme a entender el problema desde el punto de vista biológico.
A Julieta Molina, porque sus consejos me guiaron durante estos años de estudio y sus palabras me dieron aliento en todo momento.
A Daniela y Mariela porque trabajar con ellas es un lujo. Gracias por guiarme en mis primeros pasos en la estadística, por la generosidad, por la paciencia, por estar siempre predispuestas a escuchar mis dudas, por la dedicación, por la buena onda. Es un placer estar al lado de dos personas siempre dispuestas a enseñar.

Índice general

Capítulo 1

Introducción

En varias ramas de las Ciencias Naturales es común tener, luego de realizar repetidas veces algún experiemento, un conjunto de datos de los cuales se desconoce la estructura probabilística que los gobierna. El deseo del científico es descubrir qué distribución probabilística está detrás de esos datos. En este escenario la pregunta que motiva esta tesis es: *¿Qué herramientas podemos emplear para descubrirla?*. Una opción es suponer que los datos responden a cierta distribución absolutamente conocida salvo por finitos parámetros. Este camino asume que es razonable conjeturar que los datos son observaciones independientes de una variable aleatoria con cierta distribución paramétrica (una normal (μ, σ^2) por ejemplo) y que lo único que resta es estimar esos parámetros (μ y σ^2).

Uno de los criterios más naturales para estimar parámetros es el de máxima verosimilitud. En algunos casos no es sencillo (y en otros es directamente imposible) hallar fórmulas cerradas para los estimadores de máxima verosimilitud. Así se vuelve necesario recurrir a técnicas de optimización numérica. En este sentido, el algoritmo Expectation Maximization es una herramienta que facilita la tarea de hallar numéricamente los estimadores de máxima verosimilitud. Dedicamos el **Capítulo 2** a contar cómo funciona este algoritmo, a exhibir varios ejemplos de aplicación y a probar algunos resultados de convergencia.

El criterio de máxima verosimilitud depende de suponer cierta estructura paramétrica. Muchas veces se puede asumir cierto modelo paramétrico por la naturaleza de los datos en cuestión. Sin embargo, en muchos casos, la realidad del científico experimental es la de poseer observaciones independientes de cierta variable y querer inferir los mecanismos que la regulan a partir de las observaciones, sin tener ideas previas sobre la distribución de la variable. En estos casos una posibilidad es proponer varias familias paramétricas y elegir la que en algún sentido sea la que mejor ajusta los datos. Aquí es necesaria alguna medida de bondad de ajuste entendida como una medida que resume la discrepancia entre los valores observados y los valores esperados bajo el modelo propuesto. En este sentido los test de bondad de ajuste nos permiten decidir, con cierto nivel de confianza, si determinada curva proporciona un ajuste razonable de los datos. Sin embargo, podría ser que varios de los modelos propuestos proporcionen un buen ajuste, entonces ¿qué criterio usar para seleccionar alguno? Modelos más complejos, es decir, con mayor cantidad de parámetros independientes, son más flexibles para adaptarse a las particularidades de los datos y por lo tanto proveen un mejor ajuste que modelos simples. El problema cuando buscamos seleccionar un modelo es encontrar el compromiso o balance apropiado entre bondad de ajuste y dimensión. En el

Capítulo 3 nos dedicamos a estudiar el test de bondad de ajuste chi-cuadrado, introducido por Pearson [16] en el año 1900. Presentamos el estadístico del test y contamos la propuesta de Mann y Wald [13] para elegir la cantidad y ancho de los intervalos que intervienen en dicho test. En el **Capítulo 4** exponemos el Criterio de Información de Akaike (AIC) que brinda una respuesta al problema de la selección de un modelo.

Por último, en el **Capítulo 5** presentamos un conjunto de datos obtenido por las investigadoras del grupo de Dinámica y Transporte Intracelular del Departamento de Química Biológica de esta facultad. Proponemos ciertas familias paramétricas como posibles modelos para explicar la aleatoriedad de los datos y utilizamos las técnicas descriptas en los primeros capítulos para encontrar aquella familia que mejor describe al conjunto de datos.

Capítulo 2

Algoritmo Expectation-Maximization

Supongamos que disponemos de n observaciones independientes $y_1, y_2, ..., y_n$ de cierta variable aleatoria con distribución $F \in \mathcal{F} = \{F(\cdot\,, \theta) : \theta \in \Theta\}$ discreta o continua. Suponemos también que dicha distribución tiene función de probabilidad puntual o de densidad f que depende de r parámetros desconocidos $\underline{\theta} = (\theta_1, ..., \theta_r) \in \Theta \subset \mathbb{R}^r$. Aquí el espacio de parámetros es un subconjunto de $\mathbb{R}^r$. Por ejemplo, f podría ser la función de densidad de una variable normal y $\underline{\theta}$ podría ser su media y varianza: $\underline{\theta} = (\mu, \sigma^2)$. A partir de las observaciones buscamos estimar los parámetros desconocidos de la función f.

Como las observaciones son independientes, la densidad conjunta resulta ser

$$f(y_1, y_2, ..., y_n|\underline{\theta}) = \prod_{i=1}^{n} f(y_i|\underline{\theta}).$$

Pensamos que las observaciones $y_1, y_2, ..., y_n$ están fijas, entonces $f(y_1, y_2, ..., y_n|\underline{\theta})$ es una función de $\underline{\theta}$, que se conoce como la función de verosimilitud y se nota $\mathcal{L}$.

$$\mathcal{L}(\underline{\theta}|\underline{y}) = f(y_1, y_2, ..., y_n|\underline{\theta})$$

donde $\underline{y} = (y_1, y_2, ..., y_n)$. El criterio de máxima verosimilitud para estimar los parámetros nos dice que seleccionemos el $\underline{\theta}$ que maximiza $\mathcal{L}$, es decir

$$\hat{\underline{\theta}} = \arg\max_{\underline{\theta} \in \Theta} \mathcal{L}(\underline{\theta}|\underline{y}).$$

La idea es que $\hat{\underline{\theta}}(y_1, y_2, ..., y_n)$ es el valor del parámetro $\underline{\theta}$ que hace "más probable" haber observado $y_1, y_2, ..., y_n$. Bajo ciertas condiciones de regularidad sobre f podremos tomar logartimo de la función de verosimilitud y maximizar $log\ \mathcal{L}(\underline{\theta}|\underline{y})$ en vez de $\mathcal{L}(\underline{\theta}|\underline{y})$.

En algunos casos podemos hallar una fórmula cerrada para las estimaciones de los parámetros derivando con respecto a cada componente de $\underline{\theta}$ e igualando a cero. Sin embargo, es sabido que encontrar máximos de funciones de varias variables es, en la mayoría de los casos, algo complicado y no siempre el sistema de ecuaciones resultante es sencillo de resolver analíticamente.

Ejemplo A *Mezcla de distribuciones*

Supongamos que la observación y_i se produce de la siguiente manera: se elige al azar, con probabilidad p_j, una de m funciones de densidad posibles. Luego, de acuerdo con los parámetros de la función elegida se produce la medición y_i. Si las funciones posibles son $f_1, f_2, ..., f_m$ que dependen de los parámetros $\underline{\theta}_1, \underline{\theta}_2, ..., \underline{\theta}_m$ respectivamente, la función de densidad de esta mezcla de distribuciones es

$$f(\cdot\,|\underline{\theta}) = \sum_{j=1}^{m} p_j \; f_j(\cdot\,|\underline{\theta}_j)$$

en donde

- $f_j(\cdot\,|\underline{\theta}_j)$ es la función de densidad de la j-ésima componente cuando los parámetros de dicha densidad vienen dados por $\underline{\theta}_j$
- p_j es la probabilidad de que se elija la función f_j y $\sum_{j=1}^{m} p_j = 1$
- $\underline{\theta} = (p_1, p_2, ..., p_m, \underline{\theta}_1, \underline{\theta}_2, ..., \underline{\theta}_m)$

Si desconocemos los parámetros $\underline{\theta}$ pero disponemos de una muestra aleatoria $\underline{y} = (y_1, y_2, ..., y_n)$ podríamos intentar estimar esos parámetros con el criterio de máxima verosimilitud. Esto supone encontrar el $\underline{\theta}$ que maximice el logaritmo de la función de verosimilitud.

$$\begin{aligned} log\ \mathcal{L}(\underline{\theta}|\underline{y}) &= log\ \prod_{i=1}^{n} f(y_i|\underline{\theta}) \\ &= \sum_{i=1}^{n} log\ f(y_i|\underline{\theta}) \\ &= \sum_{i=1}^{n} log \left(\sum_{j=1}^{m} p_j \; f_j(y_i|\underline{\theta}_j) \right) \end{aligned}$$

Esta expresión es difícil de maximizar analíticamente en $\underline{\theta}$ pues el logaritmo abarca la sumatoria.

2.1. Idea general del algoritmo

En la literatura encontramos varios algoritmos que, bajo ciertas condiciones, convergen a valores extremos de una función. La particularidad del algoritmo Expectation Maximization (EM en adelante) es que utiliza la estructura probabilística del problema. Este algoritmo fue introducido por Dempster, Laird y Robin en 1977 [3]. Para dar una idea general podemos decir que es un método para encontrar estimadores de máxima verosimilitud cuando hay datos faltantes. El algoritmo EM se usa esencialmente en dos circunstancias:

1. cuando las observaciones tienen datos faltantes debido a problemas o limitaciones en la medición

2. cuando es analíticamente intratable la maximización de la función de verosimilitud (como en el ejemplo de la mezcla de distribuciones) pero se puede simplificar asumiendo la existencia de ciertas variables que no son observadas (variables "ocultas"). Estas variables son una construcción artificial.

Estrictamente hablando, el algoritmo EM no proporciona directamente el valor de $\hat{\underline{\theta}}$, el estimador de máxima verosimilitud, sino que a partir de cierta estimación inicial $\underline{\theta}_0$ construye una sucesión $\left\{\underline{\theta}^{(k)}\right\}_{k\in\mathbb{N}_0}$ que bajo ciertas condiciones converge a $\hat{\underline{\theta}}$. A continuación presentamos un ejemplo, tomado de [23], en donde el algoritmo EM es útil, ya no porque la maximización es complicada desde el punto de vista analítico (como en el caso de mezcla de distribuciones) sino porque hay datos que no se registraron.

Ejemplo B

En varias regiones de un país se han detectado personas infectadas con un virus muy peligroso, que puede llegar a causar la muerte. En esas n regiones se ha tomado registro de la cantidad de personas infectadas y de la densidad poblacional (miles de habitantes por kilómetro cuadrado). Para $1 \leqslant i \leqslant n$ sean :

$$\begin{aligned} Y_i = &\text{ cantidad de personas infectadas en la región } i \\ Z_i = &\text{ densidad poblacional en la región } i \end{aligned}$$

Es decir, se observó $y_1, y_2, ..., y_n$ y $z_1, z_2, ..., z_n$. Supongamos que en cada región la densidad poblacional es independiente de la cantidad de personas infectadas y a su vez estas dos variables son independientes con las de las otras regiones. Se propone un modelo en donde la cantidad de personas infectadas en la región i tiene una distribución Poisson de parámetro $\beta\sigma_i$ y la densidad poblacional en la región i tiene una distribución Poisson de parámetro σ_i en donde σ_i es un factor que influye en la densidad poblacional y β es un factor que refleja la incidencia de la enfermedad. Los parámetros $\sigma_1, \sigma_2, ..., \sigma_n$ y β son desconocidos y buscamos estimarlos vía el criterio de máxima verosimilitud. Por problemas en los registros se desconoce el valor de z_1. Es decir, hay un dato que es desconocido. Los datos son incompletos por la propia naturaleza del problema (en este caso por una falla en la toma de registros) y estamos en una situación completamente distinta a la de la mezcla de distribuciones salvo porque en ambos casos es una buena idea usar el algoritmo Expectation-Maximization para estimar los parámetros desconocidos.

En lo que sigue detallaremos cómo procede el algoritmo para construir la estimación $\underline{\theta}^{(k+1)}$ si suponemos conocida $\underline{\theta}^{(k)}$. En la sección siguiente hay varios ejemplos que ayudarán a la comprensión del algoritmo.

Sean $\underline{y} = (y_1, y_2, ..., y_n)$ realizaciones independientes de una variable aleatoria Y con función de densidad $f(\cdot\,|\underline{\theta})$ que depende de ciertos parámetros $\underline{\theta} = (\theta_1, ..., \theta_r) \in \Theta$ desconocidos, de los que buscamos los estimadores de máxima verosimilitud. Llamemos $X_1, X_2, ..., X_m$ a las variables ocultas y $\underline{x} = (x_1, x_2, ..., x_m)$ a los datos no observados. Si estamos en un caso como el del virus, entonces las variables ocultas son aquellas que por problemas en la medición no fueron observadas. Las variables ocultas pueden ser reales o bien una construcción artificial cuya conceptualización permite introducir el procedimiento Expectation-Maximization. A estas alturas puede resultar sorprendente el hecho de que introducir variables que no son observables ayude al cálculo de los estimadores de

máxima verosimilitud, pero esa es justamente la clave del algoritmo EM. En los ejemplos de la sección siguiente mostramos qué variables ocultas conviene elegir en el caso de mezclas de distribuciones. Llamamos:

datos incompletos	$\underline{y} = (y_1, y_2, ..., y_n)$
datos completos	$(\underline{x}, \underline{y}) = (x_1, x_2, ..., x_m, y_1, y_2, ..., y_n)$
función de verosimilitud de los datos completos	$\mathcal{L}(\underline{\theta}\|\underline{x}, \underline{y}) = f(\underline{x}, \underline{y}\|\underline{\theta})$
log-verosimilitud de los datos completos	$l(\underline{\theta}\|\underline{x}, \underline{y}) = log\ f(\underline{x}, \underline{y}\|\underline{\theta})$

El algoritmo EM primero encuentra una estimación de $l(\underline{\theta}|\underline{y})$ y luego maximiza esa expresión. Para hallar esta estimación es importante notar que $x_1, x_2, ..., x_n$ son realizaciones de las variables ocultas. Sin embargo, como estas nos son observables debemos considerar cada x_i como una variable aleatoria X_i. Luego, la log-verosimilitud de los datos completos es una variable aleatoria. Su esperanza, dados los datos observados y una estimación de los parámetros $\underline{\theta}^{(k)}$ es

$$Q(\underline{\theta}, \underline{\theta}^{(k)}) = E[l(\underline{\theta}|\underline{X}, \underline{y})|\underline{y}, \underline{\theta}^{(k)}] \tag{2.1}$$

Equivalentemente podemos notar (2.1) así

$$E_{\underline{X}|\underline{y},\underline{\theta}^{(k)}}[l(\underline{\theta}|\underline{X}, \underline{y})] \tag{2.2}$$

en donde enfatizamos que tomamos la esperanza con respecto a la distribución condicional de $\underline{X}$ dado $\underline{y}$ bajo la k-ésima estimación de los parámetros $\underline{\theta}^{(k)}$. Así (2.1) es una estimación de $l(\underline{\theta}|\underline{y})$. El método propuesto halla el $\underline{\theta}$ que maximiza esta esperanza, es decir, construye la estimación $k+1$ así:

$$\underline{\theta}^{(k+1)} = \arg\max_{\underline{\theta}\in\Theta} E_{\underline{X}|\underline{y},\underline{\theta}^{(k)}}[l(\underline{\theta}|\underline{X}, \underline{y})]$$

En lineas generales, podemos resumir el algoritmo Expectation-Maximization así:

1. Obtener el logaritmo de la función de verosimilitud de los datos completos. Proponer una estimación inicial de los parámetros $\underline{\theta}^{(0)}$.
2. Tomar esperanza usando la distribución condicional de las variables ocultas dados los datos incompletos y la estimación actual de los parámetros.
3. Maximizar
4. Repetir (2) y (3) hasta que la sucesión $\left\{\underline{\theta}^{(k)}\right\}_{k\in\mathbb{N}_0}$ converja.

Con el algoritmo construimos una sucesión $\{\theta^{(k)}\}$ que bajo ciertas condiciones converge al $\underline{\theta} \in \Theta$ que maximiza $l(\underline{\theta}|\underline{y})$. Notemos que $l(\underline{\theta}|\underline{y})$ admite la siguiente escritura cualquiera sea $\underline{\theta}^{(k)}$:

$$\begin{aligned} l(\underline{\theta}|\underline{y}) &= E[l(\underline{\theta}|\underline{X}, \underline{y})|\underline{y}, \underline{\theta}^{(k)}] - \left(E[l(\underline{\theta}|\underline{X}, \underline{y})|\underline{y}, \underline{\theta}^{(k)}] - l(\underline{\theta}|\underline{y})\right) \\ &= Q(\underline{\theta}, \underline{\theta}^{(k)}) - H(\underline{\theta}, \underline{\theta}^{(k)}) \end{aligned} \tag{2.3}$$

en donde Q fue definido en (2.1) y $H(\underline{\theta}, \underline{\theta}^{(k)}) = E[l(\underline{\theta}|\underline{X}, \underline{y})|\underline{y}, \underline{\theta}^{(k)}] - l(\underline{\theta}|\underline{y})$. El algoritmo, dada una estimación $\underline{\theta}^{(k)}$ de los parámetros, maximiza Q pero no tiene en cuenta el término que involucra

a H. En la sección de convergencia veremos condiciones bajo las cuales tenemos una cota para H. Notemos que en el procedimiento que propone Expectation Maximization también hay una maximización de una función de múltiples variables. En varios casos, aunque no en todos, es posible hallar una fórmula cerrada para ese máximo.

Una de las primeras críticas que recibió el paper de 1977 [3] es que más que un algoritmo Expectation-Maximization es un método general para hallar estimadores de máxima verosimilitud. Los críticos objetaron que la descripción es demasiado general y que en cada ejemplo de aplicación varía sustancialmente. En la sección siguiente veremos algunos ejemplos que ayudarán a entender cómo proceder en distintos escenarios.

2.2. Ejemplos de aplicación

Ejemplo 1. ***Mezcla de dos distribuciones***

Empezaremos con un ejemplo sencillo para concentrarnos en entender cómo funciona el algoritmo. Sean $f_0(\cdot\,)$ y $f_1(\cdot\,)$ dos funciones de densidad cuyos parámetros son conocidos y sean $y_1, y_2, ..., y_n$ observaciones independientes de una variable aleatoria con densidad:

$$f(\cdot\,) = (1-p)\ f_0(\cdot\,) + p\ f_1(\cdot\,)$$

donde p es un parámetro desconocido con $0 \leqslant p \leqslant 1$.

Según el esquema del algoritmo EM, $\underline{y} = (y_1, y_2, ..., y_n)$ son los datos incompletos y la log-verosimilitud de los datos incompletos es

$$\begin{aligned} l(p|\underline{y}) &= log\left(\prod_{i=1}^{n} f(y_i|p)\right) \\ &= \sum_{i=1}^{n} log\left[(1-p)\ f_0(y_i) + p\ f_1(y_i)\right] \end{aligned}$$

Para hallar el estimador de máxima verosimilitud de p via el algoritmo EM necesitamos definir las variables ocultas. Para $i = 1, 2, ..., n$ sean:

$$X_i = \begin{cases} 1 & \text{si la } i\text{-ésima observación proviene de la densidad } f_1 \\ 0 & \text{si la } i\text{-ésima observación proviene de la densidad } f_0 \end{cases}$$

Las variables $X_1, X_2, ..., X_n$ son independientes y tienen distribución Bernoulli de parámetro p. Esta elección de variables ocultas puede resultar mágica. La intuición detrás es que si esa información estuviera disponible sería fácil estimar p. Concretamente en este caso si conocieramos de que distribución proviene cada observación y_i, es decir, si conocieramos $x_1, x_2, ..., x_n$ entonces el estimador de máxima verosimilitud de p sería $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$.

Los datos $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ son los datos completos. Además, $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ son variables independientes con "densidad" común:

$$f(x, y|p) = [(1-p)\ f_0(y)]^{1-x}\ [p\ f_1(y)]^x$$

Ahora sigamos los pasos que nos propone el algoritmo Expectation-Maximization.

Paso 1 Calculamos la log-verosimilitud de los datos completos

$$\begin{aligned} l(p|\underline{x},\underline{y}) &= log\left(\prod_{i=1}^{n} f(x_i,y_i|p)\right) \\ &= \sum_{i=1}^{n} log\ f(x_i,y_i|p) \\ &= \sum_{i=1}^{n} (1-x_i)\ log[(1-p)\ f_0(y_i)] + x_i\ log[p\ f_1(y_i)] \\ &= \sum_{i=1}^{n} log[(1-p)\ f_0(y_i)] - x_i\ log[(1-p)\ f_0(y_i)] + x_i\ log[p\ f_1(y_i)] \end{aligned}$$

Paso 2: Tomamos esperanza. Buscamos una expresión para $E[l(p|\underline{X},\underline{y})|\underline{y},p^{(k)}]$.

Observemos que $E[X_i|\underline{y},p^{(k)}] = E[X_i|y_i,p^{(k)}]$ por la independecia de X_i e Y_j para $j \neq i$.

$$\begin{aligned} E[X_i|y_i,p^{(k)}] &= \frac{f(X_i=1,y_i|p^{(k)})}{f(y_i|p^{(k)})} \\ &= \frac{p^{(k)}\ f_1(y_i)}{(1-p^{(k)})\ f_0(y_i) + p^{(k)}\ f_1(y_i)} =: \tilde{p_i}^{(k)} \end{aligned}$$

Entonces

$$\begin{aligned} E[l(p|\underline{X},\underline{y})|\underline{y},p^{(k)}] &= \sum_{i=1}^{n} log[(1-p)\ f_0(y_i)] - \tilde{p_i}^{(k)}\ log[(1-p)\ f_0(y_i)] + \tilde{p_i}^{(k)} log[p\ f_1(y_i)] \\ &= \sum_{i=1}^{n} log(1-p) + log(f_0(y_i)) - \tilde{p_i}^{(k)}\ log(1-p) - \tilde{p_i}^{(k)}\ log(f_0(y_i)) \\ &\qquad + \tilde{p_i}^{(k)}\ log(p) + \tilde{p_i}^{(k)}\ log(f_1(y_i)) \end{aligned}$$

En la última expresión el único valor desconocido es el parámetro p. Además observemos que $\tilde{p_i}^{(k)}$ no depende de p sino de la k-ésima estimación de p: $p^{(k)}$. Recordemos que $p^{(k+1)}$ será aquel que maximice $E[l(p|\underline{X},\underline{y})|\underline{y},p^{(k)}]$ entre todos los posibles p.

Paso 3: Maximizar

En este caso podemos hallar una expresión cerrada para $p^{(k+1)} = \underset{p\in[0,1]}{\arg\max}\ E_{\underline{X}|\underline{y},p^{(k)}}[l(p|\underline{X},\underline{y})]$.

En efecto,

$$\begin{aligned}
&\frac{\partial}{\partial p} E[l(p|\underline{X},\underline{y})|\underline{y},p^{(k)}]\Big|_{p=p^{(k+1)}} = 0 \\
\Leftrightarrow \sum_{i=1}^{n} &\frac{-1}{1-p^{(k+1)}} + \frac{\tilde{p_i}^{(k)}}{1-p^{(k+1)}} + \frac{\tilde{p_i}^{(k)}}{p^{(k+1)}} = 0 \\
\Leftrightarrow &\frac{-n}{1-p^{(k+1)}} + \frac{\sum_{i=1}^{n} \tilde{p_i}^{(k)}}{p^{(k+1)}\ (1-p^{(k+1)})} = 0 \\
\Leftrightarrow &\ p^{(k+1)} = \frac{\sum_{i=1}^{n} \tilde{p_i}^{(k)}}{n}
\end{aligned}$$

Entonces

$$p^{(k+1)} = \frac{\sum_{i=1}^{n} \tilde{p_i}^{(k)}}{n} = \frac{1}{n}\sum_{i=1}^{n} \frac{p^{(k)}\ f_1(y_i)}{(1-p^{(k)})\ f_0(y_i) + p^{(k)} f_1(y_i)}.$$

Es fácil verificar que $0 \leq p^{(k+1)} \leq 1$:

$$\sum_{i=1}^{n} \tilde{p_i}^{(k)} = \sum_{i=1}^{n} E[X_i|\underline{y},p^{(k)}] = E\left[\sum_{i=1}^{n} X_i|\underline{y},p^{(k)}\right] \leq E[n|\underline{y},p^{(k)}] = n.$$

<u>*Conclusión:*</u> En este caso, el algoritmo Expectation-Maximization construye la sucesión

$$\begin{cases} p^{(0)} = \frac{1}{2} \\ p^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{p^{(k)} f_1(y_i)}{(1-p^{(k)})\ f_0(y_i) + p^{(k)}\ f_1(y_i)} \end{cases}$$

donde $p^{(0)} \in [0,1]$ es una estimación arbitraria, para fijar ideas, propuse $p^{(0)} = \frac{1}{2}$. Bajo ciertas condiciones, que analizaremos en la sección siguiente, la sucesión $\left\{p^{(k)}\right\}_{k\in\mathbb{N}_0}$ converge al estimador de máxima verosimilitud $\hat{p}$.

Ejemplo 2. ***Mezcla de*** m ***distribuciones***

Este ejemplo es una generalización del anterior. Sean $f_1(\cdot), f_2(\cdot), ..., f_m(\cdot)$ m distribuciones cuyos parámetros son conocidos y sean $y_1, y_2, ..., y_n$ observaciones independientes de una variable aleatoria con densidad

$$f(\cdot) = \sum_{j=1}^{m} p_j\ f_j(\cdot)$$

con $0 \leq p_j \leq 1$ y $\sum_{j=1}^{m} p_j = 1$. Aquí Θ es el m-simplex estandard

$$\Theta = \{(p_1,\ldots,p_m) \in \mathbb{R}^m\ :\ \sum_{j=1}^{m} p_j = 0\ \text{ y }\ p_j \geqslant 0\ \ \forall 1 \leqslant j \leqslant m\}.$$

Los pesos p_j son desconocidos. Buscamos estimarlos con el criterio de máxima verosimilitud via el algoritmo EM. Ya mencionamos que maximizar esta expresión (derivando e igualando a cero) es un camino complicado desde el punto de vista analítico.

Para utilizar el esquema del algoritmo EM definamos las variables ocultas

$$\begin{aligned} \underline{X}_1 &= (X_{11}, X_{12}, ..., X_{1m}) \\ \underline{X}_2 &= (X_{21}, X_{22}, ..., X_{2m}) \\ &\vdots \\ \underline{X}_n &= (X_{n1}, X_{n2}, ..., X_{nm}) \end{aligned} \tag{2.4}$$

donde

$$X_{ij} = \begin{cases} 1 & \text{si la } i\text{-ésima observación proviene de la densidad } f_j \\ 0 & \text{si no.} \end{cases}$$

Las $\underline{X}_i$ son variables aleatorias independientes y cada una tiene distribución $multinomial(1, p_1, p_2, ..., p_m)$. Vamos a notar $\underline{X}$ a $(\underline{X}_1, \underline{X}_2, ..., \underline{X}_n)$.

$(\underline{X}_1, Y_1), (\underline{X}_2, Y_2), ..., (\underline{X}_n, Y_n)$ son independientes con densidad común

$$f(\underline{x}_i, y_i|p) = \prod_{j=1}^{m} [p_j \ f_j(y_i)]^{x_{ij}}.$$

Recordemos que la idea detrás de la elección de las variables ocultas es que, en caso de conocerlas, sería sencillo calcular los estimadores de máxima verosimilitud. En efecto, si conocieramos $\underline{x}_1, \underline{x}_2, ...\underline{x}_n$ conoceríamos de que distribución provieve cada observación y nuestro problema se reduciría a hallar los estimadores de máxima verosimilitud de una distribución $multinomial(1, p_1, ..., p_m)$. Supongamos que efectivamente conocemos las variables ocultas y hallemos los estimadores de máxima verosimilitud para la multinomial. Para ello notaremos $\underline{p}$ a $(p_1, ..., p_m)$. La función de probabilidad puntual viene dada por:

$$p_{\underline{X}_i}(\underline{x}_i) = \frac{1!}{x_{i1}! \dots x_{im}!} \ p_1^{x_{i1}} \ p_2^{x_{i2}} \dots p_m^{x_{im}}.$$

Buscamos $\underline{p}$ que maximice

$$\begin{aligned} l(\underline{p}|\underline{X}) &= log\left(\prod_{i=1}^{n} p_{\underline{X}_i}(\underline{x}_i)\right) \\ &= \sum_{i=1}^{n} log\left(\frac{1!}{x_{i1}! \dots x_{im}!}\right) + \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} \ log(p_j). \end{aligned} \tag{2.5}$$

Recurrimos al Teorema de los Multiplicadores de Lagrange, ya que los pesos están ligados a que su suma sea uno. Equivalentemente podríamos reemplazar p_m por $1 - \sum_{j=1}^{m-1} p_j$ y maximizar sin ligaduras.

Teorema 2.1 (Teorema de los Multiplicadores de Lagrange)**.** *Sean $f, g : \mathbb{R}^s \to \mathbb{R}$ de clase C^1 y sea $\underline{p}_0 \in S = \{\underline{p} \in \mathbb{R}^s \ : \ g(\underline{p}) = 0\}$ tal que $f(\underline{p}) \leqslant f(\underline{p}_0)$ para todo $\underline{p} \in S$. Entonces*

$$\nabla g(\underline{p}_0) \neq \vec{0} \Rightarrow \exists \lambda \in \mathbb{R} \text{ tal que } \nabla f(\underline{p}_0) = \lambda \nabla g(\underline{p}_0)$$

En nuestro problema $g(\underline{p}) = \left(\sum_{j=1}^{m} p_j\right) - 1$ y $f(\underline{p}) = \sum_{i=1}^{n} log\left(\frac{1!}{x_{i1}!...x_{im}!}\right) + \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}\ log(p_j)$.
Entonces existe $\lambda \in \mathbb{R}$ tal que para $1 \leq j \leq m$:

$$\begin{aligned} \frac{\partial f}{\partial p_j} &= \lambda \frac{\partial g}{\partial p_j} \\ \sum_{i=1}^{n} \frac{x_{ij}}{p_j} &= \lambda \ ,1 \\ \frac{1}{\lambda}\sum_{i=1}^{n} x_{ij} &= p_j \end{aligned} \tag{2.6}$$

Sumando en j nos queda $\frac{1}{\lambda}\sum_{j=1}^{m}\sum_{i=1}^{n} x_i j = \sum_{j=1}^{m} p_j = 1$ y por lo tanto $\lambda = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} = n$.

Reemplazando en (2.6), nos queda que $p_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$ es el estimador de máxima verosimilitud para p_j basado en la muestra $\underline{x}_1 \ldots \underline{x}_n$ ¿Y que pasá si no conocemos de que componente provino cada observación? Recurrimos al algoritmo EM.

<u>Paso 1</u> *Calculamos la log-verosimilitud de los datos completos*

$$\begin{aligned} l(\underline{p}|\underline{y}, \underline{x}_1, \underline{x}_2, ..., \underline{x}_n) &= log\left(\prod_{i=1}^{n} f(\underline{x}_i, y_i|p)\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}\ [log\ p_j + log\ f_j(y_i)]. \end{aligned}$$

<u>Paso 2</u> *Tomamos esperanza*

Queremos calcular $E_{\underline{X}|\underline{y},\underline{p}^{(k)}}[l(\underline{p}|\underline{y}, \underline{X}_1, \underline{X}_2, ..., \underline{X}_n)]$. Primero, algunas cuentas auxiliares.

$$\begin{aligned} E[X_{ij}|\underline{y}, \underline{p}^{(k)}] &= E[X_{ij}|y_i, \underline{p}^{(k)}] \\ &= \frac{f(X_{ij} = 1, y_i|\underline{p}^{(k)})}{f(y_i|\underline{p}^{(k)})} \\ &= \frac{p_j^{(k)}\ f_j(y_i)}{\sum_{s=1}^{m} p_s^{(k)}\ f_s(y_i)} =: \tilde{p_{ij}}^{(k)} \end{aligned}$$

Luego

$$E_{\underline{X},y|\underline{p}^{(k)}}[l(\underline{p}|\underline{y}, \underline{X}_1, \underline{X}_2, ..., \underline{X}_n)] = \sum_{i=1}^{n}\sum_{j=1}^{m} \tilde{p_{ij}}^{(k)}[log\ p_j + log\ f_j(y_i)] \tag{2.7}$$

<u>Paso 3</u> *Maximizamos*

Observemos que la función que queremos maximizar en $\underline{p}$ (2.7) es como (2.5) en donde la única diferencia es que en (2.7) en lugar de tener x_{ij} tenemos $\tilde{p_{ij}}^{(k)}$ (que vendría a ser una estimación de

x_{ij}). Para hallar $\underline{p} = (p_1, p_2, ..., p_m)$ que maximice (2.7) nuevamente recurrimos al Teorema de los Multiplicadores de Lagrange. Aquí $g(\underline{p}) = \left(\sum_{j=1}^{m} p_j\right) - 1$ y $f(\underline{p}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \tilde{p_{ij}}^{(k)}[log\ p_j + log\ f_j(y_i)]$. Procediendo como para el caso de la multinomial nos queda que

$$p_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tilde{p_{ij}}^{(k)}}{n} \quad 1 \leqslant j \leqslant m$$

es el que maximiza $E_{\underline{X}|\underline{y},\underline{p}^{(k)}}[l(\underline{p}|\underline{y}, \underline{X}_1, \underline{X}_2, ..., \underline{X}_n)]$.

Conclusión: El algoritmo Expectation-Maximization construye la sucesión

$$\begin{cases} p^{(0)} \in \Theta \text{ arbitrario} \\ p_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tilde{p_{ij}}^{(k)}}{n} \text{ para } 1 \leqslant j \leqslant m. \end{cases}$$

Ejemplo 3. ***Mezcla de dos distribuciones normales con parámetros desconocidos***

Retomemos el ejemplo 1 pero ahora con

$$f_0(\cdot) = \frac{1}{\sqrt{2\pi}\sigma_0}\ e^{\frac{-(\cdot-\mu_0)^2}{2\sigma_0^2}}$$
$$f_1(\cdot) = \frac{1}{\sqrt{2\pi}\sigma_1}\ e^{\frac{-(\cdot-\mu_1)^2}{2\sigma_1^2}}$$

donde (μ_0, σ_0^2) y (μ_1, σ_1^2) son parámetros desconocidos.

Sea $f(\cdot)$ la densida mezcla

$$f(\cdot) = (1-p)\ f_0(\cdot) + p\ f_1(\cdot).$$

A partir de $\underline{y} = (y_1, y_2, ..., y_n)$ una muestra aleatoria buscamos hallar los estimadores de máxima verosimilitud de $p, \mu_0, \sigma_0^2, \mu_1$ y σ_1^2 con el algoritmo EM. Este ejemplo se asemeja más a los problemas de las ciencias naturales: no solo se desconoce la proporción de las densidades involucradas sino también los parámetros que las gobiernan.

Notación

$$\underline{\theta} = (p, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$$
$$\underline{\theta}_0 = (\mu_0, \sigma_0^2)$$
$$\underline{\theta}_1 = (\mu_1, \sigma_1^2)$$

En este ejemplo utilizaremos las mismas variables ocultas que en el ejemplo 1:

$$X_i = \begin{cases} 1 & \text{si la } i\text{-ésima observación proviene de la densidad } f_1 \\ 0 & \text{si la } i\text{-ésima observación proviene de la densidad } f_0 \end{cases}$$

con $1 \leq i \leq n$.

Entonces

datos incompletos	$\underline{y} = (y_1, y_2, ..., y_n)$
datos completos	$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$
densidad de (X_1, Y_1)	$f(x_i, y_i\|\underline{\theta}) = [(1-p)\ f_0(y_i\|\underline{\theta}_0)]^{1-x_i}\ [p\ f_1(y_i\|\underline{\theta}_1)]^{x_i}$

Paso 1 *Calculamos el logaritmo de la función de verosimilitud de los datos completos*

$$l(\underline{\theta}|\underline{x}, \underline{y}) = \sum_{i=1}^{n}(1-x_i)[log(1-p) + log\ f_0(y_i|\underline{\theta}_0)] + x_i\ [log(p) + log(f_1(y_i|\underline{\theta}_1))]$$

Paso 2 *Tomamos esperanza*

$$\begin{aligned} E[X_i|y_i, \underline{\theta}^{(k)}] &= \frac{f(X_i = 1, y_i|\underline{\theta}^{(k)})}{f(y_i|\underline{\theta}^{(k)})} \\ &= \frac{p^{(k)}\ f_1(y_i|\underline{\theta}_1^{(k)})}{(1-p^{(k)})\ f_0(y_i|\underline{\theta}_0^{(k)}) + p^{(k)}\ f_1(y_i|\underline{\theta}_1^{(k)})} =: \tilde{p_i}^{(k)} \end{aligned}$$

Entonces

$$\begin{aligned} E[l(\underline{\theta}|\underline{X}, \underline{y})|\underline{y}, \underline{\theta}^{(k)}] &= E[l(\underline{\theta}|\underline{X}, \underline{y})|y_i, \underline{\theta}^{(k)}] \\ &= \sum_{i=1}^{n}(1-\tilde{p_i}^{(k)})\ log(1-p) + \tilde{p_i}^{(k)}\ log(p) \\ &\quad + \sum_{i=1}^{n}(1-\tilde{p_i}^{(k)})\ log\ f_0(y_i|\underline{\theta}_0) \\ &\quad + \sum_{i=1}^{n}\tilde{p_i}^{(k)}\ log\ f_1(y_i|\underline{\theta}_1) \end{aligned}$$

Paso 3 *Maximizamos*

Maximización con respecto a p

$$\begin{aligned} 0 &= \frac{\partial}{\partial p} E[l(\underline{\theta}|\underline{X}, \underline{y})|\underline{y}, \underline{\theta}^{(k)}]\Big|_{\underline{\theta}=(p^{(k+1)}, \underline{\theta}_0^{(k+1)}, \underline{\theta}_1^{(k+1)})} \\ 0 &= \sum_{i=1}^{n}\frac{-1+\tilde{p_i}^{(k)}}{1-p^{(k+1)}} + \frac{\tilde{p_i}^{(k)}}{p^{(k+1)}} \\ 0 &= \sum_{i=1}^{n}\frac{-p^{(k+1)} + \tilde{p_i}^{(k)}}{p^{(k+1)}\ (1-p^{(k+1)})} \\ p^{(k+1)} &= \frac{1}{n}\sum_{i=1}^{n}\tilde{p_i}^{(k)} \end{aligned}$$

Maximización con respecto a μ_0

$$0 = \frac{\partial}{\partial \mu_0} E[l(\underline{\theta}|\underline{X},\underline{y})|\underline{y},\underline{\theta}^{(k)}]\Big|_{\underline{\theta}=(p^{(k+1)},\underline{\theta}_0^{(k+1)},\underline{\theta}_1^{(k+1)})}$$

$$0 = \frac{\partial}{\partial \mu_0}\left[\sum_{i=1}^{n}(1-\tilde{p_i}^{(k)})\left(-log(\sqrt{2\pi}) - log(\sigma_0) - \frac{(y_i-\mu_0)^2}{2\sigma_0^2}\right)\right]\Big|_{\underline{\theta}=(p^{(k+1)},\underline{\theta}_0^{(k+1)},\underline{\theta}_1^{(k+1)})}$$

$$0 = \sum_{i=1}^{n}(1-\tilde{p_i}^{(k)})\frac{(y_i-\mu_0^{(k+1)})^2}{(\sigma_0^2)^{(k+1)}}$$

$$\mu_0{}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}(1-\tilde{p_i}^{(k)})y_i$$

Maximización con respecto a σ_0

$$0 = \frac{\partial}{\partial \sigma_0} E[l(\underline{\theta}|\underline{X},\underline{y})|\underline{y},\underline{\theta}^{(k)}]\Big|_{\underline{\theta}=(p^{(k+1)},\underline{\theta}_0^{(k+1)},\underline{\theta}_1^{(k+1)})}$$

$$0 = \sum_{i=1}^{n}(1-\tilde{p_i}^{(k)})\left(\frac{-1}{\sigma_0^{(k+1)}} + \frac{(y_i-\mu_0^{(k+1)})^2}{(\sigma_0^{(k+1)})^3}\right)$$

$$(\sigma_0^2)^{(k+1)} = \frac{\sum_{i=1}^{n}(1-\tilde{p_i}^{(k)})\ (y_i-\mu_0^{(k+1)})^2}{\sum_{i=1}^{n}(1-\tilde{p_i}^{(k)})}$$

Procediendo de forma análoga podemos hallar $\mu_1^{(k+1)}$ y $(\sigma_1^2)^{(k+1)}$.

Ejemplo 4. ***Mezcla de*** m ***distribuciones normales con parámetros desconocidos***

Generalizamos el ejemplo 3 y ahora consideramos $f_1, f_2, ..., f_m$ funciones de densidad normales con parámetros desconocidos $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), ..., (\mu_m, \sigma_m^2)$ respectivamente.

Sea $f(\cdot\,)$ la densidad de la mezcla de las m normales

$$f(\cdot\,) = \sum_{j=1}^{n} p_j\ f_j(\cdot\,|\underline{\theta}_j).$$

A partir de una muestra aleatoria $y_1, y_2, ..., y_n$ de f buscamos los estimadores de máxima verosimilitud de $\underline{\theta} = (p_1, ..., p_m, \mu_1, ..., \mu_m, \sigma_1^2, ..., \sigma_m^2)$. Siguiendo con la notación del ejemplo 2, llamamos $\underline{\theta}_j = (\mu_j, \sigma_j^2)$ a los parámetros de f_j. En ese caso, $\underline{\theta} = (\underline{p}, \underline{\theta}_1, ..., \underline{\theta}_m)$. Naturalmente, tomamos las mismas variables ocultas que en el ejemplo 2 de la página 10:

$$\underline{X}_1 = (X_{11}, X_{12}, ..., X_{1m})$$
$$\underline{X}_2 = (X_{21}, X_{22}, ..., X_{2m})$$
$$\vdots$$
$$\underline{X}_n = (X_{n1}, X_{n2}, ..., X_{nm}).$$

Entonces

datos incompletos	$\underline{y} = (y_1, y_2, ..., y_n)$
datos completos	$(\underline{x}_1, y_1), ..., (x_n, y_n)$
densidad conjunta de $(\underline{X}_i, Y_i)$	$f(\underline{x}_i, y_i) = \prod_{j=1}^{m} (p_j \ f_j(y_i\|\underline{\theta}_j))^{x_{ij}}$

Paso 1 Calculamos la log-verosimilitud de los datos completos:

$$l(\underline{\theta}|\underline{x}_1, ..., \underline{x}_n, \underline{y}) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}(log\ p_j + log\ f_j(y_i))$$

Paso 2 Tomamos esperanza. Primero observemos que

$$E[X_{ij}|y_i, \underline{\theta}^{(k)}] = \frac{f(x_{ij}=1, y_i|\underline{\theta}^{(k)})}{f(y_i|\underline{\theta}^{(k)})} = \frac{p_j^{(k)}\ f_j(y_i|\mu_j^{(k)}, (\sigma_j^2)^{(k)})}{\sum_{s=1}^{m} p_s^{(k)}\ f_s(y_i|\mu_s^{(k)}, (\sigma_s^2)^{(k)})} =: \tilde{p_{ij}}^{(k)}$$

Y por lo tanto

$$E[l(\underline{\theta}|\underline{X}_1, ..., \underline{X}_n, \underline{y})|\underline{y}, \underline{\theta}^{(k)}] = \sum_{i=1}^{n}\sum_{j=1}^{m} \tilde{p_{ij}}^{(k)}[log\ p_j + log\ f_j(y_i|\underline{\theta}_j)]$$

Paso 3 Maximizamos: Tenemos que recurrir al Teorema de los Multiplicadores de Lagrange (enunciado en la página 10) ya que los pesos están ligados a que su suma sea 1.

Maximización con respecto a $\underline{p}$. La cuenta es análoga al caso del ejemplo 2.

$$p_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tilde{p_{ij}}^{(k)}}{n}$$

Observemos que hasta este momento no hicimos uso de que las densidades son normales. En el ejemplo siguiente, en donde analizamos el caso de combinación convexa de gammas, retomaremos en este punto. (2.8)

Maximización con respecto a μ_j

$$\begin{aligned} 0 &= \frac{\partial}{\partial \mu_j} E[l(\underline{\theta}|\underline{X}_1, ..., \underline{X}_n, \underline{y})|\underline{y}, \underline{\theta}^{(k)}]\Big|_{\underline{\theta}=\underline{\theta}^{(k+1)}} \\ 0 &= \sum_{i=1}^{n} \tilde{p_{ij}}^{(k)} \frac{\partial}{\partial \mu_j} log\ f_j(y_i|\underline{\theta}_j)\Big|_{\underline{\theta}=\underline{\theta}^{(k+1)}} \\ 0 &= \sum_{i=1}^{n} \tilde{p_{ij}}^{(k)} \frac{(y_i - \mu_j^{(k+1)})}{(\sigma_j^2)^{(k+1)}} \end{aligned} \tag{2.9}$$

Entonces

$$\mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tilde{p_{ij}}^{(k)}\ y_i}{\sum_{i=1}^{n} \tilde{p_{ij}}^{(k)}}$$

Maximización con respecto a σ_j

$$0 = \frac{\partial}{\partial \sigma_j} E[l(\underline{\theta}|\underline{X}_1, ..., \underline{X}_n, \underline{y})|\underline{y}, \underline{\theta}^{(k)}]\Big|_{\underline{\theta}=\underline{\theta}^{(k+1)}}$$
$$0 = \sum_{i=1}^{n} \tilde{p_{ij}}^{(k)} \frac{\partial}{\partial \sigma_j} log\ f_j(y_i|\underline{\theta}_j)\Big|_{\underline{\theta}=\underline{\theta}^{(k+1)}}$$
$$0 = \sum_{i=1}^{n} \tilde{p_{ij}}^{(k)} \frac{-1}{\sigma_j^{(k+1)}} + \frac{(y_i - \mu_j{}^{(k+1)})^2}{\left(\sigma_j^{(k+1)}\right)^3}$$

Entonces

$$\left(\sigma_j^2\right)^{(k+1)} = \frac{\sum_{i=1}^{n} \tilde{p_{ij}}^{(k)}\ (y_i - \mu_j^{(k+1)})^2}{\sum_{i=1}^{n} \tilde{p_{ij}}^{(k)}}$$

Ejemplo 5. ***Mezcla de*** m ***distribuciones exponenciales con parámetros desconocidos***

Si $f_1, f_2, ..., f_m$ son densidades de variables aleatorias exponenciales de parámetros $\lambda_1, \lambda_2, ..., \lambda_m$ respectivamente el análisis es similar al del ejemplo anterior. En efecto, consideramos las mismas variables ocultas y la misma notación salvo que ahora $\underline{\theta} = (p_1, ..., p_m, \lambda_1, ..., \lambda_m)$

Paso 1 Calculamos el logaritmo de la función de verosimilitud de los datos completos:

$$l(\underline{\theta}|\underline{x}_1, ..., \underline{x}_n, \underline{y}) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}(log\ p_j + log\ f_j(y_i|\lambda_j))$$

donde $f_j(t) = \lambda_j\ e^{-\lambda_j\ t}$.

Paso 2 Tomamos esperanza. Primero calculamos

$$E[X_{ij}|y_i, \underline{\theta}^{(k)}] = \frac{f(x_{ij} = 1, y_i|\underline{\theta}^{(k)})}{f(y_i|\underline{\theta}^{(k)})} = \frac{p_j^{(k)}\ f_j(y_i|\lambda_j^{(k)})}{\sum_{s=1}^{m} p_s^{(k)}\ f_s(y_i|\lambda_s^{(k)})} =: \tilde{p_{ij}}^{(k)}.$$

Y por lo tanto $E[l(\underline{\theta}|\underline{X}_1, ..., \underline{X}_n, \underline{y})|\underline{y}, \underline{\theta}^{(k)}] = \sum_{i=1}^{n} \sum_{j=1}^{m} \tilde{p_{ij}}^{(k)}[log\ p_j + log\ f_j(y_i|\lambda_j)]$.

Paso 3 Maximizamos

Maximización con respecto a $\underline{p}$: Es análogo al caso del ejemplo 2 en donde los parámetros eran conocidos. Esta maximización no depende de las estimaciones de λ_j y resuelve el mismo problema que máxima verosimilitud para los parámetros de una $multinomial(1, p_1 \dots p_m)$ con la diferencia que cambian las observaciones x_{ij} por sus valores esperados. Luego para $1 \leqslant j \leqslant m$:

$$p_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tilde{p_{ij}}^{(k)}}{n}$$

Maximización con respecto a λ_j

Observemos que $log\ f_j(y_i) = log(\lambda_j) - \lambda_j y_i$. Entonces

$$\begin{aligned} 0 &= \frac{\partial}{\partial \lambda_j} E[l(\underline{\theta}|\underline{X}_1, ..., \underline{X}_n, \underline{y})|\underline{y}, \underline{\theta}^{(k)}]\Big|_{\underline{\theta}=\underline{\theta}^{(k+1)}} \\ 0 &= \sum_{i=1}^{n} \tilde{p_{ij}}^{(k)} \frac{\partial}{\partial \lambda_j} log\ f_j(y_i|\lambda_j)\Big|_{\lambda_j=\lambda_j^{(k+1)}} \\ 0 &= \sum_{i=1}^{n} \tilde{p_{ij}}^{(k)} \left(\frac{1}{\lambda_j^{(k+1)}} - y_i\right) \end{aligned} \tag{2.10}$$

Y por lo tanto

$$\lambda_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tilde{p_{ij}}^{(k)}}{\sum_{i=1}^{n} \tilde{p_{ij}}^{(k)}\ y_i}$$

Luego de estos ejemplos puede quedar la idea de que, en el caso de tener una combinación convexa de algún tipo de distribución simpre será posible despejar una fórmula cerrada para el argumento que maximiza $E[l(\underline{\theta}|\underline{X}_1, ..., \underline{X}_n, \underline{y})|\underline{y}, \underline{\theta}^{(k)}]$. Esto no es así en varios casos, como por ejemplo cuando hay una mezcla de gammas. En efecto, si $\underline{\theta}_j = (\alpha_j, \lambda_j)\ :\ \alpha_j > 0, \lambda_j > 0$ para $1 \leqslant j \leqslant m$ y

$$f_j(y_i|\underline{\theta}_j) = \frac{\lambda_j^{\alpha_j}}{\Gamma(\alpha_j)} y_i^{\alpha_j - 1} e^{-\lambda_j y_i} 1_{(0,+\infty)}(y_i)$$

donde

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha} e^{-x} dx \quad \text{para } \alpha > 0.$$

Retomando desde (2.8) tenemos :

$$E[l(\underline{\theta}|\underline{X}_1, ..., \underline{X}_n, \underline{y})|\underline{y}, \underline{\theta}^{(k)}] = \sum_{i=1}^{n} \sum_{j=1}^{m} \tilde{p_{ij}}^{(k)} [log\ p_j + log\ f_j(y_i|\underline{\theta}_j)]$$

Nuevamente recurrimos al Teorema de los Multiplicadores de Lagrange para hallar el máximo de esta función. La maximización con respecto a los pesos es idéntica a los ejemplos anteriores.

Maximización con respecto a λ_j

$$\begin{aligned} 0 &= \frac{\partial}{\partial \lambda_j} E[l(\underline{\theta}|\underline{X}_1, ..., \underline{X}_n, \underline{y})|\underline{y}, \underline{\theta}^{(k)}]\Big|_{\underline{\theta}=\underline{\theta}^{(k+1)}} \\ 0 &= \sum_{i=1}^{n} \tilde{p_{ij}}^{(k)} \frac{\partial}{\partial \lambda_j} log\ f_j(y_i|\underline{\theta}_j)\Big|_{\underline{\theta}_j=\underline{\theta}_j^{(k+1)}} \\ 0 &= \sum_{i=1}^{n} \tilde{p_{ij}}^{(k)} \left(\frac{\alpha_j^{(k+1)}}{\lambda_j^{(k+1)}} - y_i\right) \\ \frac{\alpha_j^{(k+1)}}{\lambda_j^{(k+1)}} &= \frac{\sum_{i=1}^{n} \tilde{p_{ij}}^{(k)} y_i}{\sum_{i=1}^{n} \tilde{p_{ij}}^{(k)}} \end{aligned} \tag{2.11}$$

Maximización con respecto a α_j

$$\begin{aligned} 0 &= \frac{\partial}{\partial \alpha_j} E[l(\underline{\theta}|\underline{X}_1, ..., \underline{X}_n, \underline{y})|\underline{y}, \underline{\theta}^{(k)}]\Big|_{\underline{\theta}=\underline{\theta}^{(k+1)}} \\ 0 &= \sum_{i=1}^{n} \tilde{p_{ij}}^{(k)} \frac{\partial}{\partial \alpha_j} log\ f_j(y_i|\underline{\theta}_j)\Big|_{\underline{\theta}_j=\underline{\theta}_j^{(k+1)}} \\ 0 &= \sum_{i=1}^{n} \tilde{p_{ij}}^{(k)} \left(log\ \lambda_j^{(k+1)} - \frac{1}{\Gamma(\alpha_j^{(k+1)})} \frac{\partial}{\partial \alpha_j} \Gamma(\alpha_j)\Big|_{\alpha_j=\alpha_j^{(k+1)}} + log\ y_i \right) \end{aligned} \tag{2.12}$$

De (2.11) y (2.12) no es posibe despejar $\alpha_j^{(k+1)}$ ni $\lambda_j^{(k+1)}$. En este caso serán necesarios métodos de optimización numéricos para hallarlos. Este hecho no debería ser sorprendente ya que es sabido que no hay una fórmula cerrada para los estimadores de máxima verosimilitud de una gamma (y eso es justamente lo que estamos tratando de hallar cuando maximizamos con respecto a λ_j y α_j).

Ejemplo 6. ***Muestras con datos faltantes***

Los ejemplos anteriores son aplicaciones del esquema que propone el algoritmo Expectation-Maximization cuando es analíticamente intratable la maximización de la función de verosimilitud. Veamos ahora un ejemplo de aplicación cuando las observaciones tienen datos faltantes debido a problemas en la medición. Retomamos el ejemplo del virus de la sección 1.

$\underline{y} = (y_1, y_2, ..., y_n) \quad \underline{z}_{(-1)} = (z_2, ..., z_n)$	datos incompletos
$\underline{y} = (y_1, y_2, ..., y_n) \quad \underline{z} = (z_1, z_2, ..., z_n)$	datos completos
$\underline{\theta} = (\beta, \sigma_1, ..., \sigma_n)$	parámetros desconocidos

Paso 1 *Calculamos el logaritmo de la función de verosimilitud de los datos completos*

Tenemos que

$$P(\underline{y}, \underline{z}|\beta, \sigma_1, ..., \sigma_n) = \prod_{i=1}^{n} \frac{e^{-\beta\sigma_i}(\beta\sigma_i)^{y_i}}{y_i!} \cdot \frac{e^{-\sigma_i}\sigma_i^{z_i}}{z_i!}$$

Y por lo tanto

$$l((\beta, \sigma_1, ..., \sigma_n)|\underline{y}, \underline{z}) = \sum_{i=1}^{n} -\beta\sigma_i + y_i\ log(\beta\sigma_i) - log(y_i!) - \sigma_i + z_i\ \log(\sigma_i) - log(z_i!) \tag{2.13}$$

Supongamos por un momento que conocemos los datos completos (es decir, que no se perdió z_1. En ese caso, podemos hallar los estimadores de máxima verosimilitud analíticamente de la siguiente manera:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta} l((\beta, \sigma_1, ..., \sigma_n)|\underline{y}, \underline{z})\Big|_{\hat{\beta}, \hat{\sigma_2}, ..., \hat{\sigma_n}} \\ 0 &= \sum_{i=1}^{n} -\hat{\sigma}_i + \frac{y_i}{\hat{\beta}} \\ \hat{\beta} &= \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \hat{\sigma}_i}. \end{aligned} \tag{2.14}$$

Para $i \in \{1, 2, ..., n\}$

$$\begin{aligned} 0 &= \frac{\partial}{\partial \sigma_i} l((\beta, \sigma_1, ..., \sigma_n)|\underline{y}, \underline{z}) \\ 0 &= -\hat{\beta} + \frac{y_i}{\hat{\sigma}_i} - 1 + \frac{z_i}{\hat{\sigma}_i} \\ \hat{\sigma}_i &= \frac{y_i + z_i}{\hat{\beta} + 1}. \end{aligned}$$

Luego, $\sum_{i=1}^{n} \hat{\sigma}_i = \frac{\sum_{i=1}^{n} y_i + z_i}{\hat{\beta}+1}$ Reemplazando en (2.14) nos queda:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\frac{\sum_{i=1}^{n} y_i + z_i}{\hat{\beta}+1}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} z_i}$$

Y por lo tanto,

$$\begin{cases} \hat{\sigma}_i &= \frac{y_i + z_i}{\hat{\beta}+1} \qquad 1 \leq i \leq n \\ \hat{\beta} &= \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} z_i}. \end{cases}$$

Estos son los estimadores de máxima verosimilitud en el caso de que no haya datos faltantes. Ahora, si nuestra situación es como la que planteamos inicialmente donde no conocemos a z_1 no resulta sencillo hallar una expresión analítica para dichos estimadores. En efecto, la función de verosimilitud para los datos incompletos es

$$f(\underline{y}, \underline{z}_{(-1)}|\beta, \sigma_1, ..., \sigma_n) = \prod_{i=1}^{n} \frac{e^{-\beta\sigma_i}(\beta\sigma_i)^{y_i}}{y_i!} \times \prod_{i=2}^{n} \frac{e^{-\sigma_i}\sigma_i^{z_i}}{z_i!}$$

y por lo tanto

$$l((\beta, \sigma_1, ..., \sigma_n)|\underline{y}, \underline{z}_{(-1)}) = \sum_{i=1}^{n} -\beta\sigma_i + y_i \ log(\beta\sigma_i) - log(y_i!) + \sum_{i=2}^{n} -\sigma_i + z_i \ \log(\sigma_i) - log(z_i!)$$

Derivando e igualando a cero hallamos las ecuaciones que deben verificar los estimadores de máxima verosimilitud.

$0 = \frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} -\hat{\sigma}_i + \frac{y_i}{\hat{\beta}} \Leftrightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \hat{\sigma}_i}$

$0 = \frac{\partial l}{\partial \sigma_1} = -\hat{\beta} + \frac{y_1}{\hat{\sigma}_1} \Leftrightarrow \hat{\sigma}_1 = \frac{y_1}{\hat{\beta}}$

$0 = \frac{\partial l}{\partial \sigma_i} = -\beta + \frac{y_i}{\sigma_i} - 1 + \frac{z_i}{\sigma_i} \Leftrightarrow \hat{\sigma}_i = \frac{y_i + z_i}{\hat{\beta}+1} \qquad 2 \leq i \leq n$

Nos queda entonces que los estimadores de máxima verosimilitud $\hat{\beta}, \hat{\sigma}_1, ..., \hat{\sigma}_n$ verifican las ecuaciones.

$$\begin{cases} \hat{\beta} &= \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \hat{\sigma}_i} \\ \hat{\sigma}_1 &= \frac{y_1}{\hat{\beta}} \\ \hat{\sigma}_i &= \frac{y_i + z_i}{\hat{\beta}+1} \qquad 2 \leq i \leq n. \end{cases}$$

Este sistema de ecuaciones no es sencillo de resolver. Veamos entonces cómo utilizar el esquema del algoritmo EM para encontrar una sucesión que converja al estimador de máxima verosimilitud. Aquí, $\underline{\theta} = (\beta, \sigma_1, ..., \sigma_n)$.

Paso 1

Ya calculamos la log-verosimilitud de los datos completos en (2.13):

$$l((\beta, \sigma_1, ..., \sigma_n)|\underline{y}, \underline{z}) = \sum_{i=1}^{n} -\beta\sigma_i + y_i \ log(\beta\sigma_i) - log(y_i!) - \sigma_i + z_i \ \log(\sigma_i) - log(z_i!)$$

Aquí $y_1, y_2, ..., y_n, z_2, ..., z_n$ son los datos observados y z_1 es un símbolo que representa la variable aleatoria Z_1.

Paso 2 Tomamos esperanza

Observemos que $E[z_1|\underline{y}, \underline{z}_{(-1)}, \underbrace{\beta^{(k)}, \sigma_1^{(k)}, ..., \sigma_n^{(k)}}_{\underline{\theta}^{(k)}}] = \sigma_1^{(k)}$. Entonces

$$\begin{aligned} E[l(\underline{\theta}|\underline{y}, \underline{z})|\underline{y}, \underline{z}_{(-1)}, \underline{\theta}^{(k)}] = &\sum_{i=1}^{n} -\beta\sigma_i + y_i \ log(\beta\sigma_i) - log(y_i!) \\ &- \sigma_1 + \sigma_1^{(k)} \ log(\sigma_1) + E[log(z_1!)|\underline{y}, \underline{z}_{(-1)}, \underline{\theta}^{(k)}] \\ &\sum_{i=2}^{n} -\sigma_i + z_i \ log(\sigma_i) - log(z_i!) \\ = &-\sum_{i=1}^{n}(1+\beta)\sigma_i + y_i \ log(\beta\sigma_i) + \sum_{i=2}^{n} z_i \ log(\sigma_i) + \sigma_1^{(k)} \ log(\sigma_1) \\ &+ \sum_{i=1}^{n} log(y_i!) + \sum_{i=2}^{n} log(z_i!) \end{aligned}$$

Paso 3 Maximizamos

Buscamos $\underline{\theta}^{(k+1)} = (\beta^{(k+1)}, \sigma_1^{(k+1)}, ..., \sigma_n^{(k+1)})$ que maximicen esta expresión.

$$0 = \frac{\partial}{\partial\beta} E[l(\underline{\theta}|\underline{y}, \underline{z})|\underline{y}, \underline{z}_{(-1)}, \underline{\theta}^{(k)}]\Big|_{\underline{\theta}^{(k+1)}} = -\sum_{i=1}^{n} {\sigma_i}^{(k+1)} + \frac{y_i}{\beta^{(k+1)}}$$

$$\beta^{(k+1)} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} {\sigma_i}^{(k+1)}} \tag{2.15}$$

$$0 = \frac{\partial}{\partial\sigma_1} E[l(\underline{\theta}|\underline{y}, \underline{z})|\underline{y}, \underline{z}_{(-1)}, \underline{\theta}^{(k)}]\Big|_{\underline{\theta}^{(k+1)}} = -(1+\beta^{(k+1)}) + \frac{y_1}{\sigma_1^{(k+1)}} + \frac{\sigma_1^{(k)}}{\sigma_1^{(k+1)}}$$

$$\sigma_1^{(k+1)} = \frac{y_1 + \sigma_1^{(k)}}{1+\beta^{(k+1)}}$$

Para $2 \leq i \leq n$

$$0 = \frac{\partial}{\partial \sigma_1} E[l(\underline{\theta}|\underline{y}, \underline{z})|\underline{y}, \underline{z}_{(-1)}, \underline{\theta}^{(k)}]\Big|_{\underline{\theta}^{(k+1)}} = -(1+\beta^{(k+1)}) + \frac{y_i}{\sigma_i^{(k+1)}} + \frac{z_i}{\sigma_i^{(k+1)}}$$

$$\sigma_i^{(k+1)} = \frac{y_i + z_i}{1+\beta^{(k+1)}}.$$

Entonces

$$\sum_{i=1}^{n} \sigma_i^{(k+1)} = \frac{\sum_{i=1}^n y_i + \sum_{i=2}^n z_i + \sigma_1^{(k)}}{1+\beta^{(k+1)}}.$$

Reemplazando en (2.15)

$$\beta^{(k+1)} = \frac{\sum_{i=1}^n y_i (1+\beta^{(k+1)})}{\sum_{i=1}^n y_i + \sum_{i=2}^n z_i + \sigma_1^{(k)}}$$
$$\beta^{(k+1)} = \frac{\sum_{i=1}^n y_i}{\sum_{i=2}^n z_i + \sigma_1^{(k)}}.$$

En resumen, ado $\underline{\theta}^{(k)} = (\beta^{(k)}, \sigma_1^{(k)}, ..., \sigma_m^{(k)})$ una estimación de los parámetros tenemos fórmulas cerradas, y fáciles de computar, para $\underline{\theta}^{(k+1)} = (\beta^{(k+1)}, \sigma_1^{(k+1)}, ..., \sigma_m^{(k+1)})$:

$$\begin{cases} \beta^{(k+1)} & = \frac{\sum_{i=1}^n y_i}{\sum_{i=2}^n z_i + \sigma_1^{(k)}} \\ \sigma_1^{(k+1)} & = \frac{y_1 + \sigma_1^{(k)}}{1+\beta^{(k+1)}} \\ \sigma_i^{(k+1)} & = \frac{y_i + z_i}{1+\beta^{(k+1)}} \qquad 2 \leq i \leq n. \end{cases}$$

2.3. Convergencia

El propósito del método Expectation-Maximization es proporcionar un algoritmo iterativo para el estimador de máxima verosimilitud. ¿Bajo qué condiciones esa sucesión converge?, ¿el algoritmo encuentra un máximo local o un valor estacionario de la función de verosimilitud? Para responder estos interrogantes probaremos algunos resultados que fueron expuestos por Jeff Wu en [22]. Lo interesante de este enfoque es que mira Expectation-Maximization como un algoritmo de optimización y utiliza los resultados existentes en ese área.

Notación En busca de mayor claridad simplificaremos la notación utilizada hasta el momento:

	notación anterior	nueva notación
parámetros	$\underline{\theta}$	θ
$log\ f(\underline{y}\|\underline{\theta})$	$l(\underline{\theta}\|\underline{y})$	$l(\theta)$
$\frac{f(\underline{x},\underline{y}\|\underline{\theta})}{f(\underline{y},\underline{\theta})}$	ninguna	$k(\underline{x}, \underline{y}\|\underline{y}, \theta)$

Aquí k es la densidad condicional de los datos completos $(\underline{x}, \underline{y})$ dado $\underline{y}$ cuando el parámetro es θ.

Como vimos en las secciones anteriores, el algoritmo EM contruye, dada θ_0 una estimación inicial de los parámetros, una sucesión $\left\{\theta^{(k)}\right\}_{k\in\mathbb{N}_0}$ así

$$\begin{cases} \theta_0 \in \Theta & \text{arbitrario} \\ \theta^{(k+1)} & = \arg\max_{\theta\in\Theta} E[l(\theta|\underline{X}, \underline{y})|\underline{y}, \theta^{(k)}] \end{cases}$$

Recordamos la notación introducida en (2.3):

$$l(\theta) = log\ f(\underline{y}|\theta) = Q(\theta, \theta^{(k)}) - H(\theta, \theta^{(k)})$$

donde

$$\begin{aligned} Q(\theta, \theta^{(k)}) &= E[log\ f(\underline{X}, \underline{y}|\theta)|\underline{y}, \theta^{(k)}] \\ H(\theta, \theta^{(k)}) &= E[log\ f(\underline{X}, \underline{y}|\theta)|\underline{y}, \theta^{(k)}] - l(\theta) \ = \ E[log\ k(\underline{X}, \underline{y}|\underline{y}, \theta)|\underline{y}, \theta^{(k)}]. \end{aligned}$$

Esta escritura del logaritmo de la función de verosimilitud será muy útil en este sección. Así, una iteración del algoritmo Expectation-Maximization $\theta^{(k)} \to \theta^{(k+1)} \ \in M(\theta^{(k)})$ viene dada por:

$$\begin{aligned} &\textbf{Paso E} \ \text{Determinar } Q(\theta, \theta^{(k)}) \\ &\textbf{Paso M} \ \text{Elegir } \theta^{(k+1)} \text{cualquier valor } \theta \in \Theta \text{ que maximice } Q(\theta, \theta^{(k)}) \\ &M \text{ es una aplicación que a cada } \theta \text{ le asigna el conjunto de parámetros} \\ &\text{que maximizan } Q(\theta', \theta) \text{ sobre los } \theta' \in \Theta \end{aligned} \tag{2.16}$$

En esta sección analizaremos esencialmente dos aspectos sobre la convergencia del algorimo EM:

- ¿el algoritmo encuentra un máximo local o un valor estacionario de la log-verosimilitud?
- ¿la sucesión $\{\theta_k\}_{k\in\mathbb{N}}$ converge?

2.3.1. Convergencia de $\{l(\theta^{(k)})\}_{k\in\mathbb{N}}$

Nuestro primer objetivo es probar la siguiente propiedad:

Propiedad 2.1. *Propiedad ascendente del algoritmo Expectation-Maximization*

$$l(\theta^{(k+1)}) \geq l(\theta^{(k)})$$

Para la demostración recordamos algunos resultados.

Proposición 2.1. *Desigualdad de Jensen*

Sea X una variable aleatoria, $g : \mathbb{R} \longrightarrow \mathbb{R}$ una función convexa. Entonces

$$E[g(X)] \geq g(E[X])$$

Nos será útil considerar la función convexa $-log(x)$. En ese caso

$$-E[log(X)] \geq -log\ E[X]. \tag{2.17}$$

Proposición 2.2. *Desigualdad de entropía*

Si f_1 y f_2 son dos funciones de densidad, entonces

$$E_{f_1}\left[log\left(\frac{f_1(x)}{f_2(x)}\right)\right] \geq 0$$

Demostración. $E_{f_1}\left[log\left(\frac{f_1(x)}{f_2(x)}\right)\right] = -E_{f_1}\left[log\left(\frac{f_2(x)}{f_1(x)}\right)\right] \geq -log\ E_{f_1}\left[\frac{f_2(x)}{f_1(x)}\right] = 0$ donde la desigualdad vale por (2.17). □

Lema 2.1. *Para todo $k \in \mathbb{N}$ vale que*

$$H(\theta^{(k)}, \theta^{(k)}) \geqslant H(\theta, \theta^{(k)}) \textit{para todo } \theta \in \Theta$$

Demostración.

$$\begin{aligned}
H(\theta, \theta^{(k)}) &= E_{\underline{X}|\underline{y},\theta^{(k)}}[log\ f(\underline{X}, \underline{y}|\theta)] - l(\theta) = E_{\underline{X}|\underline{y},\theta^{(k)}}[log\ f(\underline{X}, \underline{y}|\theta) - log\ f(\underline{y}|\theta)] \\
&= E_{\underline{X}|\underline{y},\theta^{(k)}}\left[log\ \frac{f(\underline{X}, \underline{y}|\theta)}{f(\underline{y}|\theta)}\right] = E_{\underline{X}|\underline{y},\theta^{(k)}}[log\ f(\underline{X}|\underline{y}, \theta)] \\
&\leq E_{\underline{X}|\underline{y},\theta^{(k)}}[log\ f(\underline{X}|\underline{y}, \theta^{(k)})] \\
&= E_{\underline{X}|\underline{y},\theta^{(k)}}\left[log\ \frac{f(\underline{X}, \underline{y}|\theta^{(k)})}{f(\underline{y}|\theta^{(k)})}\right] \\
&= E_{\underline{X}|\underline{y},\theta^{(k)}}[l(\theta^{(k)}|\underline{x}, \underline{y})] - l(\theta^{(k)}) \\
&= H(\theta^{(k)}, \theta^{(k)}) \qquad \square
\end{aligned}$$

donde la desigualdad vale por la proposición (2.2).

Ahora sí contamos con las herramientas para demostrar la propiedad (2.1)

Demostración.

$$\begin{aligned}
l(\theta^{(k+1)}) &= Q(\theta^{(k+1)}, \theta^{(k)}) - H(\theta^{(k+1)}, \theta^{(k)}) \\
&\overset{\geqslant}{\text{lema}}\ Q(\theta^{(k+1)}, \theta^{(k)}) - H(\theta^{(k)}, \theta^{(k)}) \\
&\overset{\geqslant}{\text{EM}}\ Q(\theta^{(k)}, \theta^{(k)}) - H(\theta^{(k)}, \theta^{(k)}) \\
&= l(\theta^{(k)}) \qquad \square
\end{aligned}$$

Si $\{l(\theta^{(k)})\}_{k\in\mathbb{N}_0}$ fuera acotada superiormente, como es creciente, convergería en forma monótona a un l^*. Es de interés saber si l^* es un máximo global de $l(\theta)$ sobre Θ. Si no, ¿es un máximo local o un punto estacionario?. Para lo que sigue haremos las siguientes suposiciones:

1. Θ es un subconjunto de $\mathbb{R}^r$
2. $\Theta_{\theta_0} = \{\theta \in \Theta : l(\theta) \geqslant l(\theta_0)\}$ es compacto para cualquier θ_0 tal que $l(\theta_0) > -\infty$
3. l es continua en Θ y diferenciable en el interior de Θ.

Como consecuencia de estas tres hipótesis tenemos que

$$\{l(\theta^{(k)})\}_{k\in\mathbb{N}_0} \text{ está acotada superiormente para cualquier } \theta_0 \in \Theta \tag{2.18}$$

Como vamos a querer derivar l, H y Q asumimos que $\theta^{(k)}$ está en el interior de Θ. Esto se cumple si, por ejemplo

$$\Theta_{\theta_0} \text{ está en el interior de } \Theta \text{ para } \theta_0 \in \Theta \tag{2.19}$$

Bajo los supuestos (1),(2) y (3), $\{l(\theta^{(k)})\}_{k\in\mathbb{N}_0}$ converge, pero no hay garantías que lo haga a un máximo de $l(\theta)$ sobre Θ. En general, si l tiene varios máximos (locales o globales) y puntos estacionarios la convergencia a cada uno dependerá de la estimación inicial θ_0. En [14] Murray exhibió un ejemplo en donde $\{l(\theta^{(k)})\}_{k\in\mathbb{N}_0}$ converge a un punto estacionario (y no a un máximo local) para cierto valor de θ_0. De todas formas, no podemos esperar convergencia a un máximo global (ningún algoritmo de optimización suficientemente general garantiza convergencia a extremos globales). El asunto será decidir en qué casos la sucesión generada por Expectation-Maximization converge a máximos (ya sean locales o globales) o a puntos estacionarios. Como mencionamos nuestro enfoque será utilizar los resultados existentes sobre algoritmos de optimización.

Definición Diremos que A es una aplicación punto-a-conjunto definida sobre X si a cada elemento de X le asigna un subconjunto de X.

Definición Una aplicación A definida sobre un conjunto X al conjunto de partes de X la llamaremos una aplicación cerrada en x si $x_k \to x, x_k \in X$ e $y_k \to y,\ y_k \in A(x_k)$, implica que $y \in A(x)$.

Teorema 2.2. *Teorema de Convergencia Global*

Sea la sucesión $\{x_k\}_{k\in\mathbb{N}_0}$ generada por $x_{k+1} \in A(x_k)$ donde A es una aplicación punto-a-conjunto en X. Sea $\Gamma \subset X$ un conjunto solución dado y supongamos que

(I) todos los punto x_k están contenidos en un conjunto compacto $S \subset X$

(II) A es cerrada sobre el complemento de Γ

(III) hay una función continua α en x tal que :

a) *Si $x \notin \Gamma \Rightarrow \alpha(y) > \alpha(x)\ \forall y \in A(x)$*

b) *Si $x \in \Gamma \Rightarrow \alpha(y) \geqslant \alpha(x)\ \forall y \in A(x)$*

Entonces todos los puntos límites de $\{x_k\}_{k\in\mathbb{N}_0}$ *están el el conjunto solución* Γ *y* $\{\alpha(x_k)\}_{k\in\mathbb{N}_0}$ *converge en forma monótona a* $\alpha(x)$ *para algún* $x \in \Gamma$.

La demostración puede consultarse en la página 91 de [24].

En el caso del algoritmo Expectation-Maximization :

- $\alpha(x)$ es la log-verosimilitud de los datos incompletos l
- M es la aplicación definida en 2.16 en la página 22
- el conjunto solución Γ será el conjunto de máximos locales en el interior de Ω (que llamaremos $\mathcal{M}$) o bien es el conjunto de puntos estacionarios en el interior de Ω (que llamaremos $\mathcal{S}$).

La condicion $(iii)(b)$ se cumple por la propopiedad ascendente (2.1). Por otro lado, (i) se deduce del supuesto (2) y la contención es estricta por (2.19). Entonces, tenemos el siguiente resultado

Teorema 2.3. *Sea* $\{\theta^{(k)}\}_{k\in\mathbb{N}_0}$ *una sucesión generada por el algoritmo EM en donde* $\theta^{(k+1)} \in M(\theta^{(k)})$ *y supongamos que*

*(*I*)* M *es una aplicación punto-a-conjunto cerrada sobre el complemente de* $\mathcal{S}$ *(respectivamente* $\mathcal{M}$*)*

*(*II*)* $l(\theta^{(k+1)}) > l(\theta^{(k)}) \ \ \forall\ \theta^{(k)} \notin \mathcal{S}$ *(respectivamente* $\mathcal{M}$*)*

Entonces todos los puntos límite de $\{\theta^{(k)}\}_{k\in\mathbb{N}_0}$ *son puntos estacionarios (respec. máximos locales) de* l *y* $l(\theta^{(k)})$ *converge en forma monótona a* $l^* = l(\theta^*)$ *para algún* $\theta^* \in \mathcal{S}$ *(respec.* $\mathcal{M}$*).*

Una condición suficiente para que M sea cerrada es que $Q(\psi, \phi)$ sea continua en ψ y ϕ. Esta condición se verifica en varias situaciones como por ejemplo si el log-verosimilitud de los datos completos pertenece a una familia exponencial [3], [22] .

Propiedad 2.2. *Si* $Q(\psi, \phi)$ *es continua en* ψ *y* ϕ *entonces* M *es cerrada en* Ω.

Demostración. Sea $x \in \Omega$ y $x_k \to x$. Sean $y_k \in M(x_k)$ tal que $y_k \to y$. Queremos ver que $y \in M(x)$. Como $y_k \in M(x_k)$

$$Q(y_k, x_k) \geqslant Q(z, x_k) \ \ \forall z \in \Omega$$

Haciendo $k \to \infty$ tenemos

$$Q(y, x) \geqslant Q(z, x) \ \ \forall z \in \Omega$$

y por lo tanto $y \in M(x)$.

Notación

$$D^{ij}F(\psi, \phi) = \frac{\partial^{i+j}F(\psi, \phi)}{\partial\psi^i \partial\phi^j}$$

Teorema 2.4. *Si* Q *es continua en ambas variables entonces todos los puntos de límites de la sucesión* $\{\theta^{(k)}\}_{k\in\mathbb{N}_0}$ *generada por el algoritmo EM son puntos estacionarios de* l *y* $l(\theta^{(k)})$ *converge en forma monótona a* $l(\theta^*)$ *para algún punto estacionario* θ^*.

Demostración. Basta verificar (ii) del teorema (2.3) pues (i) se deduce a partir de la continuidad de Q y la propiedad (2.2). Sea $\theta^{(k)} \notin \mathcal{S}$. Por (2.19), $\theta^{(k)}$ está en el interior de Ω. En (2.1) probamos que $H(\theta^{(k)}, \theta^{(k)}) \geqslant H(\theta, \theta^{(k)}) \ \ \forall \theta \in \ \Omega$ y por lo tanto $\theta^{(k)}$ es un máximo de $H(\theta, \theta^{(k)})$ sobre los $\theta \in \Omega$. En consecuencia

$$DH^{10}(\theta^{(k)}, \theta^{(k)}) = 0$$

y

$$Dl(\theta^{(k)}) = D^{10}Q(\theta^{(k)}, \theta^{(k)}) - D^{10}H(\theta^{(k)}, \theta^{(k)}) = D^{10}Q(\theta^{(k)}, \theta^{(k)})$$

y $Dl(\theta^{(k)}) \neq 0$ pues $\theta^{(k)}$ está en el complemento de $\mathcal{S}$. Por lo tanto $D^{10}Q(\theta^{(k)}, \theta^{(k)}) \neq 0$. En consecuencia $\theta^{(k)}$ no es máximo de $Q(\theta, \theta^{(k)})$ sobre los $\theta \in \Omega$. Esto justifica la desigualdad estricta en $Q(\theta^{(k+1)}, \theta^{(k)}) > Q(\theta^{(k)}, \theta^{(k)})$. Nuevamente por (2.1) $H(\theta^{(k)}, \theta^{(k)}) \geqslant H(\theta^{(k+1)}, \theta^{(k)})$ entonces

$$l(\theta^{(k+1)}) = Q(\theta^{(k+1)}, \theta^{(k)}) - H(\theta^{(k+1)}, \theta^{(k)}) > Q(\theta^{(k)}, \theta^{(k)}) - H(\theta^{(k)}, \theta^{(k)}) = l(\theta^{(k)}). \quad \square$$

No podemos usar el mismo argumento para probar que los puntos límites son máximos locales. Es decir, no es cierto que valga (ii) cuando $\theta^{(k)} \notin \mathcal{M}$. En efecto, si $\theta^{(k)} \in \mathcal{S} \setminus \mathcal{M}$ ($\theta^{(k)}$ es un punto estacionario pero no un máximo de l), $DL(\theta^{(k)}) = D^{10}Q(\theta^{(k)}, \theta^{(k)}) = 0$ y por lo tanto $\theta^{(k)}$ podría ser un máximo de $Q(\theta, \theta^{(k)})$ sobre los $\theta \in \Omega$. En ese caso, $\theta^{(k+1)} = \theta^{(k)}$ y no vale que $l(\theta^{(k+1)}) > l(\theta^{(k)})$. Por lo tanto para garantizar convergencia a un máximo local necesitamos una condición extra: (2.20).

Teorema 2.5. *Supongamos que Q es continua y*

$$\sup_{\theta' \in \Omega} Q(\theta', \theta) > Q(\theta, \theta) \text{ para todo } \theta \in \mathcal{S} \setminus \mathcal{M} \tag{2.20}$$

Entonces todos los puntos límite de la sucesión generada por el EM $\{\theta^{(k)}\}_{k \in \mathbb{N}_0}$ son máximos locales de l y $\{l(\theta^{(k)})\}_{k \in \mathbb{N}_0}$ converge en forma monótona a $l^* = l(\theta^*)$ para algún máximo local θ^*.

En general 2.20 es una condición difícil de verificar y eso limita bastante la utilidad del teorema.

2.3.2. Convergencia de $\{\theta^{(k)}\}_{k \in \mathbb{N}}$

Ahora nos ocuparemos de la convergencia de la sucesión $\{\theta^k\}_{k \in \mathbb{N}}$. Serán necesarias condiciones más fuertes que las de la sección anterior para que el algoritmo converja en este sentido. Definimos:

$$\mathcal{S}(\alpha) = \{\theta \in \mathcal{S} : l(\theta) = \alpha\}$$
$$\mathcal{M}(\alpha) = \{\theta \in \mathcal{M} : l(\theta) = \alpha\}$$

Bajo las condiciones del teorema (2.3), $l(\theta^k) \to l*$ y todos los puntos límite de $\{\theta^{(k)}\}$ están en $\mathcal{S}(l*)$ (resp. $\mathcal{M}(l*)$). Si fuera que $\mathcal{S}(l*)$ (resp. $\mathcal{M}(l*)$) tiene un único elemento $\theta*$, es decir, no hay dos puntos estacionarios (resp. máximos locales) distintos con el mismo $L*$, entonces $\theta^{(k)} \to \theta*$. Así tenemos el siguiente resultado.

Teorema 2.6. *Sea $\{\theta^{(k)}\}$ una sucesión generada por el algoritmo EM que satisface las condiciones (i) y (ii) del teorema 2.3. Si $\mathcal{S}$ (resp. $\mathcal{M}(l*)$) $= \{\theta*\}$ donde $l*$ es el límite de $\{l(\theta^{(k)})\}$ en el teorema 2.3 entonces $\theta^{(k)} \to \theta*$.*

La condición de que $\mathcal{S}(l*) = \{\theta*\}$ puede ser relajada si asumimos que $\|\theta^{(k+1)} - \theta^{(k)}\| \to 0$ cuando $k \to \infty$.

Resumen de los resultados de convergencia

1. Cualquier sucesión $\{\theta^{(k)}\}$ generada por el algoritmo EM incrementa la log-verosimilitud y $\{l(\theta^{(k)})\}_{k\in\mathbb{N}}$, si está acotada superiormente, converge a $l*$.

2. Si $Q(\psi, \phi)$ es continua en ψ y ϕ, $l*$ es un valor estacionario de l.

3. Si, además de (2), Q no se estanca en ningún θ_0 que sea un punto estacionario pero no un máximo local, es decir, $\sup_{\theta\in\Theta} Q(\theta, \theta_0) > Q(\theta_0, \theta_0)$ entonces $l*$ es un máximo local de l. Esta condición es, en general, difícil de verificar. Como la convergencia a un punto estacionario o máximo local o máximo global depende de la estimación incial de θ es recomendable correr el algoritmo EM con diferentes estimaciones iniciales θ_0 que sean representativas del espacio de parámetros Θ.

4. Si, además de (2), no existen dos puntos estacionarios (máximos locales) distintos con el mismo valor de l, entonces $\{\theta^{(k)}\}$ converge a un punto estacionario (máximo local).

Capítulo 3

Bondad de ajuste

Sean $x_1, x_2, ..., x_n$ observaciones independientes de una variable aleatoria con función de distribución acumulada F, desconocida. Por algún motivo, sospechamos que F_0 es la función de distribución acumulada que gobierna esos datos y queremos saber si las observaciones $x_1, x_2, ..., x_n$ no contradicen esa suposición. Es decir, buscamos testear la validez de un modelo probabilístico.

Una manera informal de hacer esto es confeccionar el histograma de los datos y juzgar "a ojo" la cercanía del histograma y la función de densidad asociada a la función F_0 propuesta. Este método, al depender de la mirada del observador, es muy subjetivo. Fisher, en su libro Statistical Methods for Research Workers comenta sobre esta manera de evaluar la bondad de ajuste:

> No eye observation of such diagrams, however experienced, is really capable of discriminating whether or not the observations differ from the expectation by more than could be expected from circumstances of random sampling.

(R. A. Fisher, 1925, p. 36). Es decir, la mera obseración no es capaz de distinguir si los datos difieren de lo que se espera de ellos bajo F_0 más de lo esperable por el azar que intervino al tomar la muestra. En el año 1900 Pearson introdujo el test que denominó χ^2 (chi-cuadrado) [16]. Según sus propias palabras [17], el objetivo del test es permitir al científico decidir si determinada curva es un ajuste razonale de las observaciones. El test de Pearson fue el primero de los llamados test de bondad de ajuste.

3.1. Generalidades

Las observaciones $x_1, x_2, ..., x_n$ son realizaciones independientes de una variable aleatoria con distribución F desconocida. Sea F_0 una función de distribución acumulada perteneciente a una familia de distribuciones $\mathcal{F}_0 = \{F_0(\cdot, \theta) \ : \ \theta \in \Theta \subseteq \mathbb{R}^s\}$. Deseamos testear si es razonable pensar que la muestra aleatoria fue generada por una variable con distribución en $\mathcal{F}_0$. Es decir :

$$H_0 : F \in \mathcal{F}_0 \qquad versus \qquad H_1 : F \notin \mathcal{F}_0$$

La hipótesis H_0 puede ser simple o compuesta. Si la hipótesis H_0 es simple, la familia $\mathcal{F}_0$ de

la hipótesis nula tiene una única distribución F_0, que es conocida y completamente especificada. Podemos pensar que los parámetros de la distribución se conocen por experiencias previas o argumentos teóricos. Si H_0 es una hipótesis compuesta, sólo el tipo de distribución es especificado y por lo tanto la familia $\mathcal{F}_0$ contiene más a más de una distribución. En este caso, los parámetros serán estimados a partir de la muestra. Por ejemplo, una hipótesis simple es $H_0 : \ F(x) = \underbrace{\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt}_{F_0(x)}$ en donde la hipótesis que queremos testear es si los datos fueron generados por una distribución normal estándar. En cambio, una hipótesis compuesta es suponer que los datos son generados por una variable con distribución normal (sin espeficificar los parámetros).

El espíritu de los test de bondad de ajuste es evaluar si se encontró una curva de ajuste razonable, no afirmar que H_0 sea cierta o falsa. Claro que H_0 es o bien cierta o bien falsa pero el test no pretende ser un criterio de verdad de H_0, sino una medida para testear qué tan adecuado es un modelo probabilístico para ajustar los datos observados.

Dado un test, podemos cometer dos tipos de errores: rechazar H_0 cuando es verdadera o no rechazar H_0 cuando es falsa. Usualmente se llama *error de tipo I* al primero y *error de tipo II* al segundo. El *nivel de significación de un test* es la probabilidad de que dicho test nos conduzca a rechazar H_0 cuando en realidad es cierta, es decir, es la probabilidad de cometer un error de tipo I.

3.2. Test chi-cuadrado

3.2.1. Descripción del test

Primer caso: Hipótesis simples

Vamos a analizar primero el caso en donde la familia $\mathcal{F}_0$ contiene una única distribución: $\mathcal{F}_0 = \{F_0\}$ y F_0 es una función de densidad completamente especificada (es decir, conocemos sus parámetros).
Consideremos Π una partición finita de $\mathbb{R}$, $\Pi : I_1, I_2, ..., I_k$ donde $I_1, I_2, ..., I_k$ son intervalos.
Intuitivamente, si fuera que la cantidad de observaciones en cada intervalo no es muy distinta de la cantidad de observaciones esperadas bajo F_0 entonces no es descabellado pensar que F_0 es un modelo aceptable para la distribución que gobierna esos datos.
Formalmente, para $1 \leq j \leq k$ llamamos O_j a la cantidad de datos observados en el intervalo I_j y f_j a la probabilidad del intervalo I_j bajo H_0:

$$O_j = \sum_{i=1}^{n} I_{\{x_i \in I_j\}} \qquad\qquad f_j = P_{F_0}(I_j)$$

donde n es el tamaño de la muestra. Así, la frecuencia esperada bajo H_0 en el intervalo I_j es $n\ f_j$.
Observemos que si H_0 es válida, $(O_1, O_2, ..., O_k)$ tiene distribución multinomial con parámetros n, $(f_1, f_2, ..., f_k)$.

Una forma de medir la diferencia entre los datos observados y lo que se espera de ellos bajo

H_0 es

$$T_n = \sum_{j=1}^{k} \frac{(nf_j - O_j)^2}{nf_j} \qquad = \quad \sum \frac{(esperados - observados)^2}{esperados}$$

en donde se suma el cuadrado de la diferencia en cada intervalo, ponderada por el inverso de la frecuencia esperada. Pearson probó que bajo la hipótesis nula T_n converge en distribución a una variable chi-cuadrado con $k-1$ grados de libertad.

Teorema 3.1. *Si* $(O_1, O_2, ..., O_k)$ *tiene distribución multinomial con parámetros* n , $a = (a_1, a_2, ..., a_k) > 0$ *entonces la sucesión*

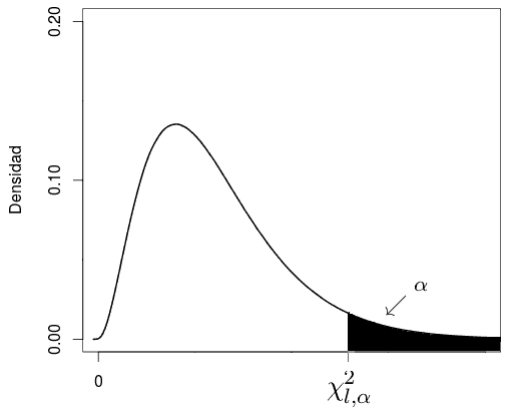
$$C_n = \sum_{j=1}^{k} \frac{(na_j - O_j)^2}{na_j}$$

converge en distribución a una variable χ^2_{k-1}.

La demostración puede consultarse en [18].

Como observamos anteriormente $(O_1, O_2, ..., O_k)$ tiene distribución multinomial con parámetros n, $(f_1, f_2, ..., f_k)$. Luego, si n es grande, podemos asumir que T_n tiene distribución χ^2_{k-1}. La calidad de la aproximación depende las probabilidades f_j. Como se menciona en [18] si 1001 intervalos tienen probabilidades $10^{-23}, 10^{-23}, ..., 10^{-23}, 1 - 10^{20}$ para valores moderados de n todos salvo un intervalo estarán vacios y será necesario un n muy grande para que la aproximación χ^2_{1000} funcione. Como una regla práctica, se suele sugerir que la elección de la partición sea tal que la cantidad de observaciones esperadas por intervalo sea al menos 5, i.e., $nf_j \geqslant 5$ para $1 \leqslant j \leqslant k$.

Notación: Llamamos $\chi^2_{l,\alpha}$ al percentil α de la distribución chi-cuadrado con l grados de libertad. Si T es una variable aleatoria con distribución χ^2_l entonces $P(T > \chi^2_{l,\alpha}) = \alpha$



Test chi-cuadrado con nivel de significación asintótico $\alpha \in [0,1]$ para hipótesis simples

$$H_0 : F = F_0 \qquad H_1 : F \neq F_0$$

A partir de las observaciones $x_1, x_2, ..., x_n$, calcular $T_n = \sum_{j=1}^{k} \frac{(nf_j - O_j)^2}{nf_j}$

- Si $T_n \geqslant \chi^2_{k-1,\alpha}$, rechazar H_0
- Si $T_n < \chi^2_{k-1,\alpha}$, no rechazar H_0

Observemos que efectivamente el nivel asintótico del test es α :

$$P(rechazar\ H_0 | H_0\ es\ verdadera) = P(T_n \geqslant \chi^2_{k-1,\alpha} | H_0) = P_{F_0}(T_n \geqslant \chi^2_{k-1,\alpha}) \underrightarrow{n \to \infty}\ \alpha$$

Para un conjunto de observaciones $x_1, x_2, ..., x_n$ fijo, llamamos p-valor al menor nivel de significación para el que rechazamos la hipótesis H_0. Otra forma de interpretar el p-valor es la siguiente: es la probabilidad de observar un valor del estadístico tan o más extremo que el observado bajo la hipótesis nula H_0. Supongamos que para un conjunto de datos dado se evalúa el estadístico del test y se obtiene un p-valor de 0.002. Para interpretarlo pensemos que la hipótesis nula es cierta e imaginemos a otros investigadores repitiendo la experiencia en idénticas condiciones. El valor 0.002 nos dice que sólo dos investigadores de cada 1000 puede obtener un valor del estadístico tan o más extremo que el obtenido. Por lo tanto, la diferencia entre los datos y los que se espera de ellos bajo H_0 no puede atribuirse meramente a variación aleatoria. Cuanto menor sea el p-valor más evidencia en contra de H_0 tenemos. Por el contrario, p-valores altos indican que la probabilidad de observar, bajo la hipótesis nula, un valor del estadístico tan o más extremo que el observado es alta y consecuentemente los datos no contradicen la suposición hecha en H_0. En las aplicaciones, utilizamos el test chi-cuadrado para verificar que determinada curva F_0 es un modelo probabilistico razonable de las observaciones. Por lo tanto, si obtenemos un p-valor lo suficientemente alto no rechazamos H_0 como modelo. Usualmente se considera que los datos no contradicen la suposición hecha en H_0 cuando el p-valor es mayor a 0.2. Vale la pena notar que nos interesaría poder controlar el error de tipo II pero, aún en el caso de una alternativa fija, en muchos casos no es sencillo hallar la distribución del estadístico bajo la alternativa.

Segundo caso: Hipótesis compuestas

Si buscamos testear si nuestros datos tienen distribución normal, exponencial, gamma, o cualquier distribución paramétrica pero desconocemos los parámetros de la distribución, podemos estimarlos a partir de la muestra. En ese caso, las frecuencias esperadas en cada intervalo $n\hat{f}_1 = n\ P_{\hat{F}_0}(I_1)$, $n\hat{f}_2 = n\ P_{\hat{F}_0}(I_2)$, ... , $n\hat{f}_k = n\ P_{\hat{F}_0}(I_k)$ serán aproximaciones basadas en una estimación $\hat{F}_0$ de F_0. Con lo cual, se ve afectado el estadístico del test

$$T_n = \sum_{j=1}^{k} \frac{(n\hat{f}_j - O_j)^2}{n\hat{f}_j}.$$

La distribución asintótica de este estadístico no es necesariamente chi-cuadrado, depende del estimador de los parámetros utilizado. Si el estimador de los parámetros es asintóticamente eficiente

(por ejemplo, el estimador de máxima verosimilitud) y está basado en el vector $O_1, O_2, ..., O_k$ (no en las observaciones $x_1, x_2, ..., x_n$ sino en los datos ya agrupados), resulta que T_n también tiene distribución asintótica chi-cuadrado, pero ahora con $k-1-s$ grados de libertad donde s es la cantidad de parámetros estimados a partir de la muestra.

Supongamos que no disponemos de las observaciones originales $x_1, x_2, ..., x_n$, pero sí se conoce $O_1, O_2, ..., O_k$. En ese caso resulta natural basar la estimación en ese vector. Sin embargo, en las aplicaciones es común disponer de las observaciones $x_1, x_2, ..., x_n$, con lo cual puede parecer artificial basar la estimación en los datos agrupados. ¿Qué pasaría si basamos la estimación en $x_1, x_2, ..., x_n$? Responderemos a esta pregunta en breve.

Test chi-cuadrado con nivel de significación asintótico $\alpha \in [0,1]$ para hipótesis compuestas

$$H_0 : F \in \mathcal{F}_0 = \{F_0(\cdot\,, \theta) \ : \ \theta \in \Theta \subseteq \mathbb{R}^s\} \qquad H_1 : F \notin \mathcal{F}_0$$

A partir de las observaciones $x_1, x_2, ..., x_n$, calcular $O_1, O_2, ..., O_k$ y estimar los parámetros $\theta \subseteq \mathbb{R}^s$ con el estimador de máxima verosimilitud. Calcular $T_n = \sum_{j=1}^{k} \frac{(n\hat{f}_j - O_j)^2}{n\hat{f}_j}$

- Si $T_n \geqslant \chi^2_{k-1-s,\alpha}$, rechazar H_0
- Si $T_n < \chi^2_{k-1-s,\alpha}$, no rechazar H_0

Pearson introdujó el test chi-cuadrado en una publicación de 1900 [16] y la modificación con parámetros estimados fue considera por Fisher [9], quien corrigió la creencia de que la estimación de los parámetros no modificaba la distribución límite del estadístico.

Posteriormente Chernoff y Lehmann en [7] mostraron que si para estimar los parámetros se usaban estimadores de máxima verosimilitud basados en las observaciones originales, la distribución asintótica del estadístico deja de ser chi-cuadrado. Sin embargo, en la práctica se suelen usar las observaciones $x_1, x_2, ..., x_n$ y no los datos agrupados. Analizando el resultado de Chernoff y Lehmann intentaremos justificar esta práctica.

Si estimamos los parámetros a partir de las observaciones $x_1, x_2, ..., x_n$, llamemos $\tilde{F}_o$ a distribución estimada. Luego, la frecuencia esperada en el intervalo I_j será $n\ \tilde{f}_j = n\ P_{\tilde{F}_0}(I_j)$.

Sea

$$\tilde{T}_n = \sum_{j=1}^{k} \frac{(n\tilde{f}_j - O_j)^2}{n\tilde{f}_j}.$$

Observar que $\tilde{T}_n$ es como el estadístico del test T_n donde la única diferencia está en cómo fueron estimados los parámetros.

Teorema 3.2. *Bajo H_0, si $F = F_0(\cdot\,, \theta)$ con $\theta \in \Theta \subseteq \mathbb{R}^s$, la distribución asintótica de $\tilde{T}_n$ coincide con la de*

$$\sum_{i=1}^{k-s-1} y_i^2 + \sum_{i=k-s}^{k-1} \lambda_i\ y_i^2$$

donde las y_i son variables independientes normalmente distribuidas con media cero y varianza unitaria y los λ_i están entre 0 y 1 y pueden depender de los s parámetros.

Notemos que la primer sumatoria corresponde a una variable χ^2_{k-1-s} y la segunda es una variable positiva. Un detalle importante es que, a diferencia de T_n, la distribución límite depende de los parámetros. Quienes son exactamente los λ_i y una demostración del teorema pueden consultarse en [7].

Supongamos que basamos la estimación en las observaciones originales $x_1, x_2, ..., x_n$ y erroneamente suponemos que $\tilde{T}_n$ tiene distribución asintótica χ^2_{k-1-s}. Si queremos testear con nivel de significación asintótico α, rechazaremos la hipótesis nula si $\tilde{T}_n \geqslant \chi^2_{k-1-s,\ \alpha}$ (esto bajo la creencia de que $\tilde{T}_n$ tiene distribución asintótica χ^2_{k-1-s}). El resultado de Lehmann nos demuestra que esto no es cierto, con lo cual el nivel del test que suponemos es asintóticamente α en verdad no lo es,

$$P(rechazar\ H_0 | H_0\ es\ cierta) = P_{F_0}(\tilde{T}_n > \chi^2_{k-1-s}) \quad \underrightarrow{n \to \infty} \quad \alpha* > \alpha$$

Es decir, el nivel del test es mayor de lo que suponiamos: la probabilidad de rechazar la hipótesis nula cuando es válida pensabamos que era asintóticamente α cuando en realidad es superior.

En [7], Chernoff y Lehmann nos proporcionan un ejemplo. Queremos testear si observaciones $x_1, x_2, ..., x_n$ independientes son realizaciones de una variable aleatoria normal. Para la construcción del test utilizamos los intervalos $(-\infty, -1), (-1, 0), (0, 1), (1, +\infty)$. Supongamos que la muestra efectivamente proviene de una población normal con media $\mu = 0$ y varianza $\sigma^2 = 1$. En ese caso, $\lambda_1 = 0{,}8$ y $\lambda_2 = 0{,}2$ (λ_1 y λ_2 son los autovalores de cierta matriz relacionada con el test).

La probabilidad $\alpha*$ viene dada por

$$\alpha* = P(U + \lambda_1 V + \lambda_2 W \geqslant \chi^2_{k-1-s,\alpha})$$

donde U, V y W son variables chi-cuadrado con 1 grado de libertad. Consideremos el caso $\alpha = 0{,}05$. Como una cota inferior de $\alpha*$ se computó $P(U + 0{,}8V \geqslant \chi^2_{k-1-s,0,05}) = 0{,}12$. Esto nos indica que si utilizamos las observaciones originales $x_1, x_2, ..., x_n$ para estimar los parámetros, el error de tipo I asintótico no es 0.05 como se pensaba sino que es mayor a 0.12. Esto parece desalentador, pero recordemos que en general, como nos interesa no rechazar H_0, buscamos obtener p-valores altos.

Para una muestra fija $\underline{x} = (x_1, x_2, ..., x_n)$ sea $\tilde{p}$ el p-valor obtenido bajo la suposición errónea de que $\tilde{T}_n$ tiene distribución asintótica χ^2_{k-1-s} y sea p el p-valor verdadero. Así, si U tiene distribución χ^2_{k-1-s} y R tiene la distribución asintótica de $\tilde{T}_n$ tenemos que

$$\tilde{p} = P(U \geqslant T_n(\underline{x})) \ \leqslant \ P(R \geqslant T_n(\underline{x})) = p.$$

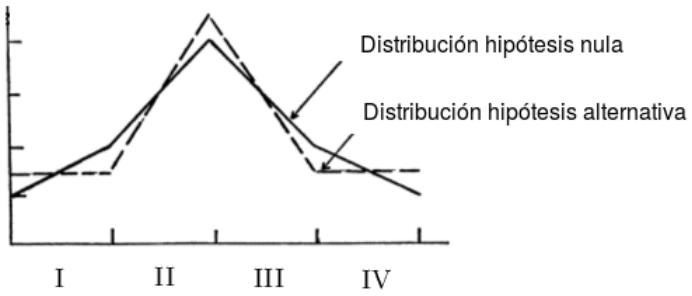
Con lo cual $\tilde{p}$ nos da una cota inferior para el p-valor del test. Si buscamos validar H_0, buscamos obtener p-valores altos y por lo tanto tener una cota inferior para el p-valor es útil.

3.2.2. Limitaciones

Si bien este test es usado frecuentemente presenta algunos inconvenientes y limitaciones a tener en cuenta. Los ejemplos que se presentan a continuación fueron tomados de [21].

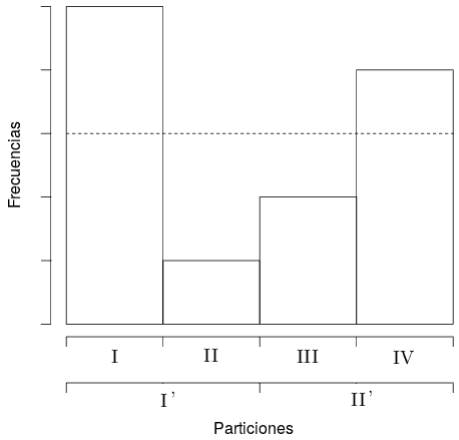
- En general hay varias distribuciones que tienen la misma frecuencia esperada por clase que la hipótesis nula. Es posible que no rechacemos la hipótesis de que determinada muestra fue extraída de una población normal, cuando en realidad fue extraída de otra población con las

mismas frecuencias esperadas por clase. Si la distribución de esa otra población fuera puesta como hipótesis nula, esta nueva H_0 no sería rechazada pues el valor del estadístico del test de Pearson sería el mismo. Un ejemplo de esto se aprecia en la siguiente figura.



Están graficadas las funciones de densidad de dos distribuciones: f_0 y f_1. Para las clases que se marcan en la figura, la cantidad de observaciones esperadas bajo f_0 y f_1 coinciden (esto lo podemos chequear graficamente: el área bajo f_0 y f_1 en cada intervalo es la misma). Es necesario remarcar que el no rechazo de la hipótesis nula no es una afirmación de que los datos tiene esa distribución, sino que nos da la pauta de que H_0 es un modelo que no entra en contradicción con los datos y que, en consecuencia, podría ser adoptado. Esto no dice que no haya otros modelos probabilísticos que también ajusten los datos de manera tan precisa como H_0, o incluso, sean más precisos.

- La cantidad y ancho de los intervalos que particionan la recta real puede ser elegida de muchas maneras distintas ya que el test no impone restricciones sobre la partición. Pero no todas esas elecciones conducen a los mismos resultados. Por ejemplo, en la siguientes figuras la linea punteada representa la frecuencia esperada bajo H_0 y los bloques las frecuencias observadas según las distintas particiones. Si usamos las clases I,II, III y IV rechazamos H_0. En cambio, si utilizamos I' y II' , no rechazamos.



3.2.3. ¿Cómo elegir la cantidad y ancho de los intervalos?

Como mencionamos, la elección de la partición es un asunto importante. Distintas elecciones pueden derivar en distintos resultados. Supongamos que efectuamos dos particiones distintas y obtenemos un p-valor de 0.001 para una (lo cual nos conduciría a rechazar H_0) y un p-valor de 0.6 para la otra (lo cual nos lleva a no rechazar H_0). Entonces, ¿F_0 es o no es un buen modelo para ajustar las observaciones? Esta pregunta es independiente de las particiones. Si testeamos la bondad de ajuste de F_0 mediante un test chi-cuadrado con nivel α, sabemos que la probabilidad de rechazar H_0 cuando es verdadera es asintóticamente α. Pero nada sabemos de la probabilidad de que el test nos conduzca a rechazar H_0 cuando vale la hipótesis alternativa H_1. Entonces, es razonable pensar que la partición debe ser elegida de manera de maximizar la potencia del test, es decir, maximizar la probabilidad de que el test nos conduzca a rechazar H_0 cuando efectivamente H_0 no es válida. En general, es dificil analizar la potencia de un test para hipótesis compuestas, ya que no se conoce una expresión analítica para la distribución del estadístico T_n bajo la hipótesis alternativa.

Propuesta de Mann & Wald

En el caso de que los parámetros de la hipótesis nula estén especificados (es decir, no son estimados a partir de la muestra) y que la distribución de la hipótesis nula sea continua Mann & Wald propusieron una solución, que fue publicada en [13]. Ellos proponen utilizar cierta partición, que en un sentido que detallaremos más adelante, es óptima Para elegir las particiones del test chi-cuadrado con nivel de significación asintótico α la propuesta es:

- Elegir el número de clases k así:

$$k = \left[4\sqrt[5]{\frac{2(n-1)^2}{c^2}} \right]$$

  donde n es el tamaño de la muestra y c es el percentil α de una distribución $N(0,1)$:

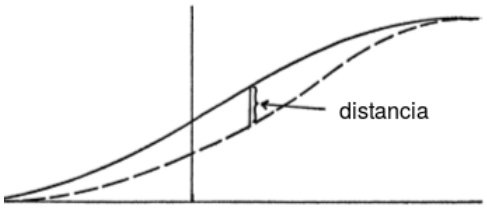
$$\alpha = \int_c^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} dy$$

- Elegir el ancho de los intervalos de manera de que la frecuencia esperada bajo H_0 sea igual para todos: $\frac{n}{k}$

Por ejemplo, si efectuamos un test con nivel de significación α para una muestra de tamaño 1000, la sugerencia es elegir $k = \left[4\sqrt[5]{\frac{2(1000-1)^2}{1{,}64^2}} \right] = 59$.

Ventajas y desventajas

Primero, consideremos $F_1(t)$ y $F_2(t)$ dos funciones de distribución acumulada y definimos la distancia entre ellas como $d(F_1, F_2) = \sup_{t\in\mathbb{R}} |F_1(t) - F_2(t)|$.



Ventajas

Las pruebas de las siguientes afirmaciones se pueden encontrar en el artículo de Mann y Wald.

1. Se minimiza la distancia máxima entre las funciones de distribución acumulada que tienen las mismas frecuencias esperadas por clase que la hipótesis nula. Es decir, si bien no se elimina la primer restricción, se minimiza la distancia entre esas alternativas.

2. La potencia del test para aquellas funciones de distribución acumulada cuya distancia a la función de distribución acumulada de la hipótesis nula es mayor o igual que $\Delta = \frac{5}{k} - \frac{4}{k^2}$ es mayor o igual a 0,5. Es decir, la probabilidad de rechazar la hipótesis nula cuando los datos son generados por una variable cuya función de distribución acumulada está a distancia mayor o igual que Δ de la hipótesis nula es mayor o igual a $\frac{1}{2}$

3. Si se elige una cantidad de clases distinta a la que propone la fórmula, existe al menos una función de distribución acumulada cuya distancia a la función de distribución acumulada de la hipótesis nula es mayor o igual que Δ y la potencia del test para esa función es menor a 0,5

4. Cuando la hipótesis nula es verdadera, hay más probabilidades de no rechazo que cuando es falsa.

Desventajas

1. La teoría es asintótica. Fue probada rigurosamente para muestras de tamaño mayor o igual 450 y nivel de significación 0.05 y para muestras de tamaño mayor o igual a 300 y nivel de significación 0.01.

2. Elegir los intervalos de manera de que bajo H_0 resulten equiprobables lleva un tiempo considerable.

3. Se desconoce la potencia del test para aquellas distribuciones cuya distancia a la hipótesis nula es mayor que Δ. Mucho más serio es el cuestionamiento de si la distancia definida es un criterio útil. Por ahí resultaría más interesante hablar de la potencia del test para distribuciones que son similares a la hipótesis nula en algún otro sentido.

4. Se asume los parámetros de la distribución nula son conocidos.

5. La distribución de la hipótesis nula tiene que ser continua.

Según Williams en [21], la propuesta de Mann y Wald es demasiado conservadora y en la mayoría de los casos reducir la cantidad de intervalos a la mitad no varía significativamente la potencia del test.

Capítulo 4

Criterio de Información de Akaike

En la elección de un modelo hay esencialmente dos pasos. Primero, elegir una o varias familias de curvas (la ˝forma˝ que podría tener la curva de ajuste). Segundo, encontrar dentro de esas familias la curva que ajusta los datos de forma más precisa. Este segundo paso requiere algún criterio para medir la bondad de ajuste. El criterio de información de Akaike conocido como AIC por sus siglas en inglés (Akaike Information Criterion) fue introducido por Hirotugu Akaike en la publicación *'Information theory and an extension of maximum likelihood principle'* en 1973 [1].

El paradigma tradicional de la estimación por máxima verosimilitud provee un mecanismo para estimar los parámetros desconocidos de un modelo con una dimensión y estructura especificados. Akaike extendió este paradigma considerando un marco de trabajo en donde la dimensión del modelo es también desconocida y por lo tanto debe ser determinada a partir de las observaciones.

Para un modelo paramétrico, la función de verosimilitud refleja la conformidad del modelo con los datos observados. Si consideramos otro modelo parámetrico más complejo, el modelo se vuelve más flexible para adaptarse a las características de los datos. Luego, si buscamos el modelo que maximize la función de verosimilitud indefectiblemene elegiremos, entre los modelos posibles, el más complejo.

Sin embargo, es común que el científico tenga el deseo de describir el mecanismo que genera las observaciones de una forma simple. ¿Qué características hacen que una curva de ajuste sea más simple que otra?, ¿por qué deberían ser preferibles las curvas menos complejas?. Si un modelo pertenece a alguna familia paramétrica es razonable asociar la complejidad del modelo con la cantidad de parámetros independientes. La segunda pregunta abre un largo e interesante camino de discusión filosófica que puede seguirse en [11].

Estos dos requisitos entran inevitablemente en conflicto. Maximizar la simplicidad del modelo usualmente requiere sacrificar bondad de ajuste (un modelo simple es menos flexible para adaptarse a las fluctuaciones de los datos). Por otro lado, un modelo que ajuste ˝perfectamente˝ los datos generalmente es muy complejo. ¿Cómo seleccionar un modelo que tenga un balance entre estos dos requisitos? Dedicaremos este capítulo a estudiar la propuesta de Akaike.

4.1. Generalidades

Disponemos de un conjunto de observaciones. Estas observaciones son los valores de alguna variable aleatoria cuya distribución es desconocida (o tenemos un conocimiento parcial de ella). De la información provista por los datos queremos hacer inferencias respecto de los aspectos desconocidos de la distribución subyacente. Formalmente sea $\underline{y} = (y_1, y_2, ..., y_n)$ un conjunto de observaciones que son realizaciones independientes de una variable aleatoria continua con función de densidad $g(\cdot)$ desconocida. El objetivo es hallar, entre un conjunto de densidades paramétricas propuestas, la que en algún sentido sea la que mejor aproxime la densidad verdadera $g(\cdot)$. En lo que sigue expresaremos un modelo en terminos de una función de densidad. El vector aleatorio $\underline{y} = (y_1, y_2, ..., y_n)$ tiene función de densidad conjunta $\prod_{i=1}^{n} g(y_i)$ que también notaremos $g(\cdot)$. Los modelos serán los de la familia $\mathcal{F} = \{\mathcal{F}(k_1), \mathcal{F}(k_2), ..., \mathcal{F}(k_l)\}$ en donde

$$\mathcal{F}(k_i) = \{f(\cdot\,|\underline{\theta}_{k_i}) : \underline{\theta}_{k_i} = (\theta^1, \theta^2, ..., \theta^{k_i}) \in \Theta(k_i) \subset \mathbb{R}^{k_i}\}$$

es una familia de densidades en donde el espacio de parámetros tiene dimensión k_i. Por simplicidad, la notación asume que cada moldelo se distingue por la dimensión k_i.

Definición Para $1 \leqslant i \leqslant l$ sea $\hat{\underline{\theta}}_{k_i}(\underline{y})$ el estimador de máxima verosimilitud basado en la muestra $\underline{y}$ bajo el modelo $\mathcal{F}(k_i)$:

$$\hat{\underline{\theta}}_{k_i}(\underline{y}) = \underset{\underline{\theta}_{k_i} \in \Theta(k_i)}{\arg\max}\ log\ f(\underline{y}|\underline{\theta}_{k_i}).$$

Luego, $f(\cdot\,|\hat{\underline{\theta}}_{k_i}(\underline{y}))$ es la densidad de la familia $\mathcal{F}(k_i)$ que mejor aproxima a $g(\cdot)$ en el sentido de máxima verosimilitud.

Nuestro problema es ajustar las observaciones $\underline{y} = (y_1, y_2, ..., y_n)$ con algún modelo de la familia $\mathcal{F} = \{\mathcal{F}(k_1), \mathcal{F}(k_2), ..., \mathcal{F}(k_l)\}$. Para cada modelo (es decir, para cada $1 \leqslant i \leqslant l$), el Criterio de Información de Akaike(AIC) nos propone calcular $\hat{\underline{\theta}}_{k_i}(\underline{y})$ el estimador de máxima verosimilitud bajo el modelo $\mathcal{F}(k_i)$, luego calcular

$$AIC_i = 2k_i - 2\ log\ f(\underline{y}|\hat{\underline{\theta}}_{k_i}(\underline{y})) \tag{4.1}$$

y seleccionar el modelo cuyo valor AIC_i sea minímo. En (4.1) y a lo largo de todo el capítulo *log* refiere al logaritmo natural.

Podemos interpretar el valor de AIC_i como una log-verosimilitud penalizada. El término de penalización está relacionado con la cantidad de parámetros independientes del modelo: $2k_i$. Informalmente podemos decir que al seleccionar el modelo k_i que minimiza AIC_i estamos eligiendo aquel que presenta el mejor balance entre conformidad con las observaciones y dimensión del modelo. Como mencionamos en la introducción, los modelos más complejos son capaces de adaptarse a las características de los datos y por lo tanto el término de menos la log-verosimilitud será menor que en el caso de un modelo con menor cantidad de parámetros independientes. Es decir, en la elección del modelo hay un compromiso entre dimensión y bondad de ajuste. De alguna forma AIC describe cuánto tiene que ser la mejora en el ajuste para preferir una curva más compleja. Una leve mejora no será suficiente para compensar el término que penaliza la complejidad.

En la práctica, los valores de AIC para los distintos modelos son fáciles de calcular (o al menos de igual dificultad que calcular los estimadores de máxima verosimilitud). Esto le aporta al

criterio mucha practicidad. Akaike meciona en [2] que una ventaja de este critero es que no necesita elecciones subjetivas como es la del nivel de significación en un procedimiento de test de hipótesis.

A medida que el tamaño de la muestra n aumenta el segundo término de AIC $-2\ log\ f(\underline{y}|\hat{\underline{\theta}}_{k_i}(\underline{y})) = -2\ \sum_{j=1}^{n} log\ f(y_i|\hat{\underline{\theta}}_{k_i}(\underline{y}))$ disminuye pero el término de penalización $2k_i$ permanece constante. Eso significa que el término de penalización tiene poco efecto si el tamaño de muestra n es grande. De todas formas es importante notar que cualquier problema real tiene un tamaño muestral n finito y ciertamente AIC provee una respuesta a la pregunta de cuánto tiene que mejorar el ajuste la incorporación de un parámetro independiente antes de ser incluido en el modelo y en qué escala debe ser medida esa mejora en el ajuste.

La interpretación de AIC como la log-verosimilitud penalizada por la cantidad de parámetros es bastante general y no supone ninguna condición sobre las familias $\mathcal{F}(k_1), \mathcal{F}(k_2), ..., \mathcal{F}(k_l)$ de modelos candidatos considerados. Por ejemplo, puede ser usadado para comparar modelos no anidados basados en distintas distribuciones de probabilidad. Asimismo esta interpretación deja el interrogante de por qué dos veces la cantidad de parámetros es una medida de penalización adecuada. El Criterio de Información de Akaike tiene otra motivación, más relacionada con su surgimiento histórico, basada en la siguiente idea: si tuvieramos una medida de distancia entre un modelo y la distribución verdadera, un procedimiento natural sería buscar un modelo que minimice esta distancia. Yendo en ese sentido la divergencia de Kullback-Leibler es una herramienta clave.

4.2. La divergencia de Kullback-Leibler como medida de bondad de ajuste

Cómo calcularla

Consideremos $f(x)$ y $g(x)$ dos funciones de densidad tales que el soporte no depende del parámetro. La divergencia de Kullback-Leibler entre $g(x)$ y $f(x)$, que notamos $I(g, f)$ se calcula como:

$$I(g,f) = \begin{cases} \int\limits_{\{x:g(x)>0\}} g(x)\ log\left(\frac{g(x)}{f(x)}\right) dx = E_g\left[log\left(\frac{g(X)}{f(X)}\right)\right] & \text{si } (\forall x) f(x) = 0 \text{ implica } g(x) = 0 \\ \infty & \text{en caso contrario .} \end{cases}$$

La expresión para la divergencia de Kullback-Leibler en el caso discreto es

$$I(p,\pi) = \begin{cases} \sum\limits_{i=1}^{k} p_i\ log\left(\frac{p_i}{\pi_i}\right) & \text{si } (\forall i)\ p_i = 0 \text{ implica } \pi_i = 0 \\ \infty & \text{en caso contrario} \end{cases}$$

donde las variables aleatorias subyacentes pueden tomar k valores distintos, la probabilidad bajo p del i-ésimo valor es p_i, la probabilidad bajo π del i-ésimo valor es π_i con $0 < p_i \leqslant 1$, $0 < \pi_i \leqslant 1$ y $\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} \pi_i = 1$.

Propiedades

En [12] Kullback analiza en detalle las propiedades de $I(f,g)$. Mencionaremos solamente algunas que nos serán de utilidad:

1. $I(g,f) > 0$ si g y f son densidades distintas en casi todo punto.
2. $I(g,f) = 0$ si y sólo si g y f son iguales en casi todo punto.

Sin embargo, $I(f,g)$ no es una distancia formal pues no es simétrica ni verifica la desigualdad triangular. La propiedad 1 también es conocida como la Desigualdad de Entropía y fue probada en el capítulo 2, página 23.

Interpretación

La divergencia de Kullback-Leibler puede ser pensada como una medida de discrepancia entre g y f en el sentido que es cero si y sólo si g y f son iguales y es siempre positiva. En [5] se sugiere interpretar $I(g,f)$ como una cuantificación de la ¨información¨ perdida cuando f es usada para aproximar g. Cuando buscamos un modelo de ajuste, buscamos que este ¨pierda¨ la mínima información posible, esto es equivalente a buscar -entre un conjunto de candidatos- el modelo f que minimice $I(g,f)$. Minimizar $I(g,f)$ es encontrar el modelo ¨más cercano¨ (en el sentido de la divergencia de Kullback-Leibler) a la verdad.

Relación con el estimador de máxima verosimilitud

Analizemos el caso en donde los modelos candidatos pertenecen a una única familia paramétrica $\mathcal{F} = \{f(\cdot\,|\underline{\theta}) : \underline{\theta} \in \Theta\}$. La pregunta ahora es ¿Cuál es el $\underline{\theta}$ que hace que $f(\cdot\,|\underline{\theta})$ sea el modelo que mejor se ajusta a los datos de entre los de $\mathcal{F}$?

Bajo el paradigma de máxima verosimilitud la respuesta sería $f(\cdot\,|\hat{\underline{\theta}})$ donde $\hat{\underline{\theta}}$ al estimador de máxima verosimilitud de $\underline{\theta}$ bajo el modelo $f(\cdot\,|\underline{\theta})$ basado en las observaciones $\underline{y} = (y_1, y_2, ..., y_n)$. Por otra parte, si utilizamos como criterio para la bondad de ajuste la divergencia de Kullback-Leibler , el mejor modelo de la clase $\mathcal{F}$ es aquel que minimiza $I(g, f(\cdot\,|\underline{\theta}))$ entre los $\underline{\theta}$ en Θ. Llamamos θ^* a ese mínimo. Es decir

$$\min_{\underline{\theta}\in\Theta} I(g, f(\cdot\,|\underline{\theta})) = \int g(x)\ log\left(\frac{g(x)}{f(x|\theta^*)}\right) d\underline{x}.$$

Entonces $f(\cdot\,|\theta^*)$ es el mejor modelo entre los de $\mathcal{F}$ bajo el criterio de la divergencia de Kullback-Leibler .

Observemos que no asumimos que la verdadera distribución $g(\cdot\,)$ subyacente a las observaciones pertenece a la familia paramétrica $\mathcal{F} = \{f(\cdot\,|\underline{\theta}) : \underline{\theta} \in \Theta\}$ que define al estimador de máxima verosimilitud. Un resultado muy interesante en este sentido es la consistencia del estimador de máxima verosimilitud: bajo ciertas condiciones de regularidad $\hat{\underline{\theta}}(\underline{y})$ converge casi seguramente a θ^* [10],[20] .

4.3. Deducción de AIC a partir de la divergencia de Kullback-Leibler

La divergencia de Kullback-Leibler interpretada como una medida de discrepancia entre dos distribuciones nos proporciona un criterio para seleccionar un modelo que ajuste las observaciones entre los de la familia $\mathcal{F} = \{\mathcal{F}(k_1), \mathcal{F}(k_2), ..., \mathcal{F}(k_l)\}$: elegir la densidad $f(\cdot\,)$ que minimice la divergencia de Kullback-Leibler entre $g(\cdot\,)$ y $f(\cdot\,)$. Es decir, seleccionaremos el modelo ¨más cercano¨ a $g(\cdot\,)$.

Según lo expuesto en la sección 4.1 el Criterio de Información de Akaike puede ser resumido así:

Buscamos ajustar las observaciones $\underline{y} = (y_1, y_2, ..., y_n)$ con algún modelo de la familia $\mathcal{F} = \{\mathcal{F}(k_1), \mathcal{F}(k_2), ..., \mathcal{F}(k_l)\}$ donde $\mathcal{F}(k_i) = \{f(\cdot\,|\underline{\theta}_{k_i}) : \underline{\theta}_{k_i} \in \Theta_{k_i}\}$ es una familia paramétrica donde el parámetro $\underline{\theta}_{k_i}$ tiene k_i componentes independientes. Para cada modelo (es decir, para cada $1 \leqslant i \leqslant l$) calcular $\hat{\underline{\theta}}_{k_i}(\underline{y})$ el estimador de máxima verosimilitud basado en $\underline{y}$ bajo la suposición de que los datos son observaciones de una variable aleatoria con función de densidad en la familia $\mathcal{F}(k_i)$. Para cada modelo k_i, calcular $AIC_i = -2log\ f(\underline{y}|\hat{\underline{\theta}}_{k_i}(\underline{y})) + 2k_i$ y seleccionar el modelo cuyo valor AIC_i sea minímo.

En esta sección nos dedicaremos a interpretar el Criterio de Información de Akaike vía la divergencia de Kullback-Leibler . Para ello hay que restringir los modelos candidatos y suponer que la ¨verdadera¨ distribución subyacente a las observaciones $\underline{y} = (y_1, y_2, ..., y_n)$ puede expresarse como :

$$f(\cdot\,|\theta_K) \quad : \qquad \theta_K = (\theta_1, \ldots, \theta_K) \in \mathbb{R}^K.$$

Notaremos θ_K^* al valor ¨verdadero ¨ del vector de parámetros θ_K. Es decir,

$$g(\cdot\,) = f(\cdot\,|\theta_K^*).$$

Los modelos que consideraremos serán, en términos paramétricos, restricciones del vector θ_K:

$$\mathcal{F}(k) = \{f(\cdot\,|\theta_k) \quad : \quad \theta_k = (\theta_1, \ldots, \theta_k, 0, \ldots, 0)\}.$$

Es decir, bajo el modelo k, el vector de parámetros está restringido al espacio Θ_k con

$$\Theta_k = \{(\theta_1, ..., \theta_k, \theta_{k+1}, ..., \theta_K) \in \mathbb{R}^K \quad : \quad \theta_{k+1} = \theta_{k+2} = \ldots = \theta_K = 0\}.$$

Naturalmente, asociamos k con la dimensión del modelo. Extendiendo la notación de la sección anterior llamamos θ_k^* al argumento que minimiza la distancia de Kullback-Leibler entre $f(\cdot\,|\theta_K^*)$ y las densidades de $\mathcal{F}(k)$:

$$\theta_k^* = \underset{\theta_k \in \Theta_k}{\operatorname{argmin}}\, I(f(\cdot\,|\theta_K^*), f(\cdot\,|\theta_k)).$$

Asimismo notaremos $\hat{\theta}_k$ al estimador de máxima verosimilitud bajo el modelo k:

$$\hat{\theta}_k(\underline{y}) = \underset{\theta_k \in \Theta_K}{\arg\max}\, log f(\underline{y}|\theta_k).$$

Por último, simplificaremos la notación para la divergencia de Kullback-Leibler así:

$$I(\alpha,\beta) := I(f(\cdot\,|\alpha), f(\cdot\,|\beta)) = \int f(x|\alpha)\ log\left(\frac{f(x|\alpha)}{f(x|\beta)}\right) dx.$$

Como disponemos de una muestra aleatoria $\underline{y} = (y_1, y_2, ..., y_n)$, podemos calcular $\hat{\theta}_k(\underline{y})$ y estimar la divergencia de Kullback-Leibler entre $g(\cdot\,) = f(\cdot\,|\theta_K^*)$ y $f(\cdot\,|\theta_k^*)$, $I(\theta_K^*, \theta_k^*)$, como

$$I(\theta_K^*, \hat{\theta}_k(\underline{y})) = \int f(x|\theta_K^*)\ log\left(\frac{f(x|\theta_K^*)}{f(x|\hat{\theta}_k(\underline{y}))}\right) dx.$$

La consistencia del estimador de máxima verosimilitud fundamenta esta estimación.

Si pudieramos calcular θ_k^* y por lo tanto $I(\theta_K^*, \theta_k^*)$ podríamos juzgar la bondad de ajuste en comparación con el modelo perfecto $I(\theta_K^*, \theta_K^*) = 0$. Pero la realidad es que no conocemos θ_k^*, solo la estimación $\hat{\theta}_k(\underline{y})$. Con probabilidad 1, $\hat{\theta}_k$ no es igual a θ_k^* y entonces

$$I(\theta_K^*, \hat{\theta}_k(\underline{y})) > I(\theta_K^*, \theta_k^*).$$

Por otro lado, el estimador $\hat{\theta}_k(\underline{y})$ varía según la muestra $\underline{y} = (y_1, y_2, ..., y_n)$ observada. Por lo tanto, tiene sentido minimizar el valor esperado $E_{\underline{y}}[I(\theta_K^*, \hat{\theta}_k(\underline{y}))]$ más que el valor verdadero (pero desconocido) $I(\theta_K^*, \theta_k^*)$. Intuitivamente

$$E_{\underline{y}}[I(\theta_K^*, \hat{\theta}_k(\underline{y}))] = E_{\underline{y}}\left[E_{\theta_K^*}[-2log\ f(x|\underline{\theta})]\Big|_{\underline{\theta}=\hat{\theta}_k(\underline{y})}\right]$$

representa la separación esperada entre la verdadera densidad $f(\cdot\,|\theta_K^*)$ y los modelos con estructura $f(\cdot\,|\hat{\theta}_k)$. Todas las esperanzas son tomadas con respecto a la verdadera distribución $f(\cdot\,|\theta_K^*)$, más allá de la notación utilizada para la variable aleatoria involucrada ($\underline{y}$ y x en este caso).

Por todo lo expuesto tiene sentido el siguiente criterio:

$$\textit{Seleccionar el modelo } f(\cdot\,|\hat{\underline{\theta}}_k) \in \mathcal{F}(k) \textit{ que minimice } E_{\underline{y}}[I(\theta_K^*, \hat{\theta}_k(\underline{y}))]. \tag{4.2}$$

Si interpretamos la divergencia de Kullback-Leibler como una medida de discrepancia entre modelos, aquel que minimiza (4.2) es aquel que minimiza la separación esperada entre $f(\cdot\,|\theta_K^*)$ y los modelos considerados. De todas formas, no es posible calcular $E_{\underline{y}}[I(\theta_K^*, \hat{\theta}_k)]$ ya que depende de la verdadera distribución $f(\cdot\,|\theta_K^*)$. En su trabajo, Akaike propone AIC_k como un estimador de $E_{\underline{y}}[I(\theta_K^*, \hat{\theta}_k)]$. Se presenta a continuación la justificación de esta última afirmación tal como se expone en varias publicaciones (por ejemplo mirar [1], [2], [4] y [8]). Entiendo que esta es la manera en que lo concibió Akaike y, si bien algunas aproximaciones están dudosamente justificadas, sirve como una motivación para entender el Criterio de Información desde la perspectiva de la divergencia de Kullback-Leibler .

En la página 26 de [12] Kullback presenta la siguiente aproximación:

$$I(\alpha,\beta) \cong \frac{1}{2}||\alpha - \beta||_J^2. \tag{4.3}$$

donde J es la matriz

$$(J)_{ij} = E_\beta \left[\left(\frac{\partial}{\partial \theta_i} log f(x|\theta) \frac{\partial}{\partial \theta_j} log f(x|\theta) \right) \Big|_{\theta=\beta} \right].$$

Esta aproximación se consigue desarrollando $I(\alpha, \beta)$ en su serie de Taylor con respecto a su segundo argumento alrededor de β y es válida bajo ciertas condiciones de regularidad sobre f siempre y cuando α y β sean lo suficientemente cercanos (en norma Euclídea).

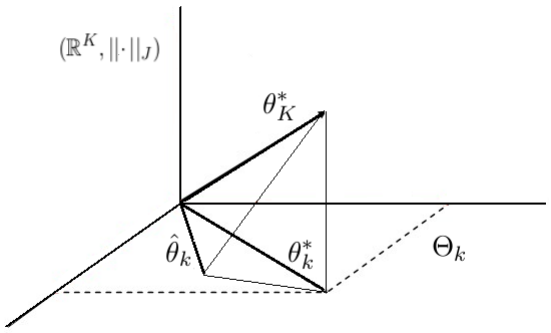
Sea $\underline{y} = (y_1, y_2, ..., y_n)$ una muestra aleatoria de una variable aleatoria con densidad $f(\cdot\,|\theta_K^*)$. Usando la aproximación (4.3) tenemos que

$$I(\theta_K^*, \hat{\theta}_k) \cong \frac{1}{2} ||\theta_K^* - \hat{\theta}_k||_J^2 \tag{4.4}$$

donde J, de dimensión $K \times K$, es la matriz

$$(J)_{ij} = E_{\theta_K^*} \left[\left(\frac{\partial}{\partial \theta_i} log f(x|\theta) \frac{\partial}{\partial \theta_j} log f(x|\theta) \right) \Big|_{\theta=\theta_K^*} \right].$$

Vale la pena notar que no es algo necesariamente evidente que θ_K^* y $\hat{\theta}_k$ estén cerca en norma Euclídea (salvo para $k = K$). Siguiendo las referencias mencionadas continuamos con la deducción asumiendo que la aproximación (4.4) es buena. Usando la aproximación (4.3) para $I(\theta_K^*, \theta_k)$ tenemos que aquel que minimiza la divergencia de Kullback-Leibler (que notamos θ_k^*) es la proyección ortogonal de θ_k^* sobre Θ_k con la métrica inducida por J en $\mathbb{R}^K$.



Retomando (4.4) nos queda que:

$$\begin{aligned} 2I(\theta_K^*, \hat{\theta}_k) &\cong ||\theta_K^* - \hat{\theta}_k||_J^2 \\ &= ||\theta_K^* - \theta_k^*||_J^2 + ||\theta_k^* - \hat{\theta}_k||_J^2 \end{aligned}$$

y por lo tanto

$$\begin{aligned} 2nE[I(\theta_K^*, \hat{\theta}_k)] &\cong E\left[n\ ||\theta_K^* - \theta_k^*||_J^2\right] + E\left[||\sqrt{n}(\theta_k^* - \hat{\theta}_k)||_J^2\right] \\ &= n\ ||\theta_K^* - \theta_k^*||_J^2 + E\left[||\sqrt{n}(\theta_k^* - \hat{\theta}_k)||_J^2\right] \end{aligned} \tag{4.5}$$

Por la normalidad asintótica del estimador de máxima verosimilitud (para una prueba mirar [10]) $||\sqrt{n}(\theta_k^* - \hat{\theta}_k)||_J^2$ tiene distribución asintótica chi-cuadrado con k grados de libertad. Entonces (4.5) queda:

$$\begin{aligned} 2nE[I(\theta_K^*, \hat{\theta}_k)] &\cong n\ ||\theta_K^* - \theta_k^*||_J^2 + k \\ &= \delta + k. \end{aligned} \tag{4.6}$$

Observemos que $\delta = n\ ||\theta_K^* - \theta_k^*||_J^2$ es desconocido pero determinístico. En su trabajo original, Akaike estimó δ usando los resultados de Wald [19] sobre la distribución asintótica del estadístico del test del cociente de máxima verosimilitud . Bajo ciertas condiciones de regularidad el estadístico del test

$$-2log\lambda = -2\sum_{i=1}^{n} log\ \frac{f(y_i|\hat{\theta}_k)}{f(y_i|\hat{\theta}_K)}$$

tiene distribución asintótica chi-cuadrado no central con $\nu = K - k$ grados de libertad y parámetro de no centralidad $\delta = n\ ||\theta_K^* - \theta_k^*||_J^2$. Como $E[-2log\ \lambda - \delta] = \nu$ entonces $-2log\ \lambda - \delta \cong \nu$ y por lo tanto $\delta \cong -2log\ \lambda - nu$. De esta forma (4.6) queda:

$$\begin{aligned} 2nE[I(\theta_K^*, \hat{\theta}_k)] &\cong \delta + k \\ &\cong -2logf(\underline{y}|\hat{\theta}_k) + 2log(\underline{y}|\hat{\theta}_K) - (K-k) + k \\ &= -2logf(\underline{y}|\hat{\theta}_k) + 2k \quad + 2log(\underline{y}|\hat{\theta}_K) - K. \end{aligned}$$

En las aplicaciones prácticas, K puede ser conceptualmente infinito o no estar definido claramente. Como nos interesa hallar el modelo k que minimice $2nE[I(\theta_K^*, \hat{\theta}_k)]$ basta calcular

$$AIC_k = -2logf(\underline{y}|\hat{\theta}_k) + 2k$$

ignorando $2log(\underline{y}|\hat{\theta}_K) - K$ que es un término común a todos los modelos.

Vimos entonces que, si n es suficientemente grande y valen las aproximaciones hechas

$$E\left[2nE[I(\theta_K^*, \hat{\theta}_k)] - \left(2logf(\underline{y}|\hat{\theta}_k) + 2k \quad + 2log(\underline{y}|\hat{\theta}_K) - K\right)\right] = 0. \tag{4.7}$$

En efecto,

$$\begin{aligned} &E\left[2nE[I(\theta_K^*, \hat{\theta_k})] - \left(2logf(\underline{y}|\hat{\theta}_k) + 2k \quad + 2log(\underline{y}|\hat{\theta}_K) - K\right)\right] \\ &= E\left[2nE[I(\theta_K^*, \hat{\theta_k})] - (-2log\lambda - \delta + \delta - (K-k) + k)\right] \\ &= 2nE[I(\theta_K^*, \hat{\theta_k})] - E[-2log\lambda - \delta] - (\delta + k) + (K - k) \\ &= 2nE[I(\theta_K^*, \hat{\theta_k})] - (n\ ||\theta_K^* - \theta_k^*||_J^2 + k) \\ &= 2nE[I(\theta_K^*, \hat{\theta_k})] - E[n\ ||\theta_K^* - \theta_k^*||_J^2 + ||\sqrt{n}(\theta_k^* - \hat{\theta}_k)||_J^2] = 0 \end{aligned}$$

4.4. Valores AIC de referencia

Los valores de AIC son importantes: modelos con valores AIC similares debén ser igualmente considerados. Ahora bien, ¿Qué cantidad constituye una diferencia considerable en los valores de AIC? Burnham y Anderson en [6] proponen lo siguiente:

AIC_k-AIC_{min}	evidencia empírica para el modelo k
0-2	sustancial
4-7	considerablemente menor
> 10	esencialmente ninguna

donde $AIC_k = 2k - 2\ log\ f(\underline{y}|\hat{\underline{\theta}}_k(\underline{y}))$ y $AIC_{min} = \min_k AIC_k$

Capítulo 5

Una aplicación a datos reales

5.1. Dinámica y transporte intracelular

El citoplasma celular puede entenderse como un fluido complejo. A modo de ejemplo uno puede compararlo con una gelatina. Así, si uno pusiera dentro de una gelatina unos granitos chiquitos de arena se quedarían en el lugar donde los pusimos, mientras que en el agua podrían difundir fácilmente. Lo mismo pasa en la célula: compuestos relativamente grandes no se pueden desplazar de un lugar a otro por difusión. Entonces, la célula usa otros mecanismos de transporte: los motores moleculares. Estos motores son proteínas muy interesantes ya que, si uno les da energía suficiente, se mueven dando pasos de entre 8 y 36 nanómetros por el citoesqueleto. El citoesqueleto está compuesto, entre otros filamentos, por microtúbulos y filamentos de una proteína llamada actina, que actúan como "autopistas". Tantos los microtúbulos como los filamentos de actina son los "carriles" que utilizan los motores para moverse. Hay determinados motores que sólo caminan por microtúbulos mientras que otros lo hacen a lo largo de filamentos de actina. Los motores moleculares se unen a la carga que necesita ser transportada y la llevan caminando de un lado a otro por los filamentos del citoesqueleto [15].

Así, el efectivo sistema de transporte de la célula está integrado por filamentos polimerizados (filamentos de actina y microtúbulos) y motores moleculares, proteínas que utilizan energía provista por hidrólisis de ATP para desplazarse a través de los filamentos. Existen 3 familias de motores responsables del transporte de organelas en células: kinesina y dineína, los cuales se desplazan sobre los microtúbulos hacia su extremo positivo y negativo respectivamente, y miosina, que se traslada a través de filamentos de actina. Los motores responsables del transporte a lo largo de microtúbulos (dineína y kinesina) son capaces de movilizar cargas ya sea hacia el centro celular o bien hacia su periferia, asegurando el correcto posicionamiento de las mismas. Es decir, el transporte conducido por dichos motores ocurre de manera bidireccional.

Durante los últimos años diversos autores han propuesto modelos explicativos con el objetivo de comprender la dinámica del transporte. Uno de los modelos sugeridos es el de regulación, en donde tanto motores kinesina y dineína (motores de polaridad opuesta) están unidos a la carga pero solamente un tipo de motor está activo en un momento determinado. Por otro lado, tenemos el modelo de "cinchada"(tug-of-war) de acuerdo al cual motores de polaridad opuesta ejercen fuerza

sobre la carga y el equipo que ejerce más fuerza neta sobre la carga en un momento dado sería el que determina la dirección del transporte.

El grupo de Dinámica y Transporte Intracelular del Departamento de Química Biológica de la FCEN se propone explorar, entre otras cosas, cómo las propiedades biofísicas de los motores afectan el transporte bidireccional en células vivas. La investigación se realizó en células S2 de *Drosophila melanogaster*. Estas celulas se caracterizan por la formación de procesos (mirar 5.1) en presencia de un agente despolimerizante de actina. Vale la pena aclarar que aquí la palabra procesos no hace referencia a algo que transcurre en el tiempo sino que es un término biológico para señalar cierta parte de la célula. Los procesos están constituidos por microtúbulos uniformemente orientados con el extremo positivo hacia la periferia celular y el negativo hacia el centro celular, lo cual permite restringir el estudio sólo al transporte conducido a través de microtúbulos.

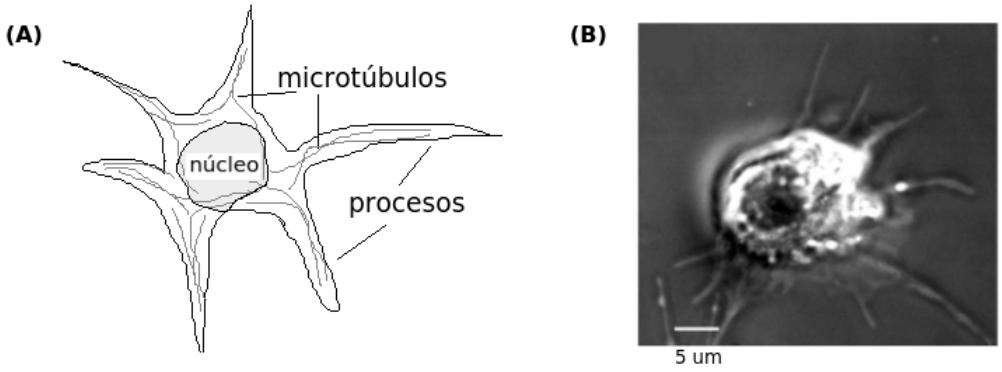
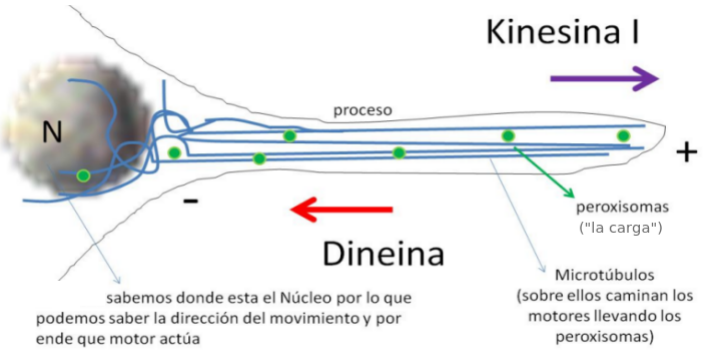


Figura 5.1: (A): esquema de una célula con procesos. (B): imágen real obtenida por microscopía de contraste de fase.

En particular, estudiaron el transporte bidireccional conducido por kineina I y dineína citoplasmática de peroxisomas (una organela) en los procesos. En el caso de estudio los peroxisomas son la carga transportada por los motores. Es decir, serían como el granito de arena en la gelatina, no se desplazan por difusión, necesitan algún motor. Cabe aclarar que la dinámica del transporte se infiere a partir del comportamiento de la carga.

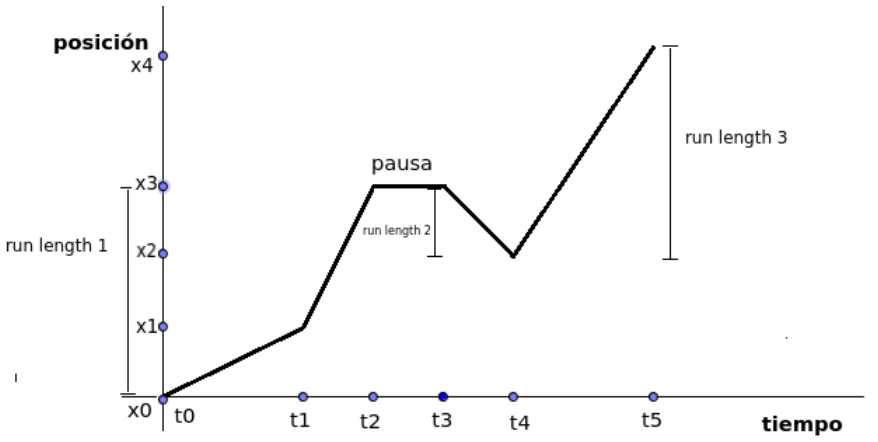
En el laboratorio utilizaron como herramienta la técnica de seguimiento de partícula única (SPT), la cual permite obtener la posición de una partícula individual a lo largo del tiempo, es decir, su trayectoria. Por medio de esta técnica obtuvieron la posición de las organelas con precisión de ~ 5 nm y con una alta resolución temporal (10 ms), por lo cual las trayectorias obtenidas pudieron ser analizadas cuantitativamente para obtener información muy precisa sobre el mecanismo de

movimiento a nivel molecular. Los peroxisomas (la carga) son marcados fluorescentemente y se registra su trayectoria por microscopia de fluorescencia. Todo lo que describe el comportamiento de los motores se infiere a partir de la dinámica de las cargas transportadas por los motores (que son las si se puden observar). Se presenta un resumen del sistema biológico en estudio.



Resumiendo: en células S2 y utilizando la técnica SPT los investigadores del grupo de Dinámica y Transporte Intracelular obtuvieron las trayectorias de los peroxisomas. En las mismas, observaron largos tramos correspondientes al transporte conducido por dineína (transporte hacia el núcleo) o kinesina I (transporte hacia la periferia), presencia de reversiones (cambios en la dirección del movimiento) y periodos de pausas u oscilaciones donde no hay transporte.

Las trayectorias fueron divididas en fragmentos unidireccionales e ininterrumpidos (runs) correspondientes al transporte conducido por dineína y kinesina I así como también se identificaron los periodos de pausas u oscilaciones. A la distancia recorrida durante un run se las llamó run length. Antes de seguir es importante entender qué es considerado un run. Supongamos que la carga transportada tiene un movimiento unidimensional (sólo se mueve a lo largo del eje que determina el microtúbulo) y que su trayectoria viene dada por el siguiente gráfico de posición en función del tiempo:



Aquí vemos que de t_0 a t_1 la carga avanza en sentido positivo con cierta velocidad. En t_1 su velocidad aumenta y de t_2 a t_3 hay una pausa. De t_3 a t_4 avanza en sentido negativo y en t_4 vuelve a cambiar el sentido del movimiento. Finalmente en t_5 la carga se desprende del microtúbulo. De

t_0 a t_2 la carga se mueve ininterrumpidamente en un único sentido y por lo tanto tenemos lo que definimos como un run. La distancia recorrida durante ese tiempo es lo que en la figura está marcado como run length 1. Los run lengths 1 y 3 corresponden al motor kinesina (avance hacia la periferia de la célula, i.e., en sentido positivo) mientras que el run length 2 corresponde al motor dineína (avance en sentido negativo). La simplificación al caso de un movimiento unidimesional fue para entender el concepto de run ya que en la realidad el movimiento es bidimensional. A continuación vemos un gráfico de una trayectoria real. En la figura (A) tenemos un gráfico de y versus x y en (B) está graficada la posición en función del tiempo (para x e y simultaneamente) de un peroxisoma. En rojo se marcaron los *plus end runs* (es decir, los runs que se corresponden con un movimiento hacia la periferia de la célula), en azul los *minus end runs* (se corresponden con un movimiento hacia el núcleo) y en negro los periodos de pausas u oscilaciones donde no se registra transporte.

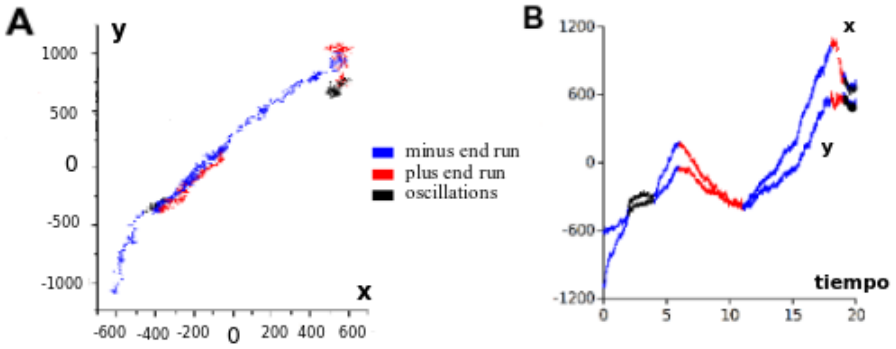


Figura 5.2: Trayectorias reales

En el experimento en células vivas obtuvieron las trayectorias de aproximadamente 350 peroxisomas. Sus trayectorias fueron divididas en runs (entre 1 y 4 runs por trayectoria) y luego se clasificaron los run lengths correspondientes al motor dineína y kinesina. De esta forma obtuvieron 269 run lengths correspondientes al motor dineína y 202 correspondientes al motor kinesina I.

Conocer la distribución de las distancias recorridas a lo largo de los runs (run lengths) respectivos a cada motor puede ayudar a comprender los mecanismos que regulan la actividad de los motores moleculares. La información que proporcionan los run lengths está relacionada con cuánto fue capaz de caminar a lo largo del microtúbulo antes de separarse de él. Vale la pena aclarar que la separación del microtúbulo no es exclusiva de que el(los) motor(motores) se ¨apagaron", puede haberse separado por otros motivos. Así, si la distribución de los run lengths resulta monoexponencial se puede inferir que el(los) motor(motores) se despegaron juntos del microtúbulo con probabilidad constante. Sino, dos o más mecanismos regulan la distancia recorrida. En lo que sigue, utilizaremos las técnicas descriptas en los capítulos 3 y 4 para ajustar las observaciones de los run lengths del motor kinesina I.

5.2. Elección de un modelo

5.2.1. Los modelos candidatos

El problema, en otros términos, es identificar un modelo que ajuste las observaciones de los run length del motor kinesina. Se propusieron los siguientes:

1. **Exponencial** $f(y|\lambda) = \lambda\ e^{-\lambda y} I_{\{y>0\}}$
2. **Biexponencial** $f(y|p,\lambda_1,\lambda_2) = p\ \lambda_1\ e^{-\lambda_1 y} I_{\{y>0\}} + (1-p)\ \lambda_2\ e^{-\lambda_2 y} I_{\{y>0\}}$
3. **Triexponencial** mezcla convexa de 3 exponenciales
4. **Cuatriexponencial** mezcla convexa de 4 exponenciales
5. **Gamma** $f(y|\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda y} y^{\alpha-1} I_{\{y>0\}}$
6. **Bigamma** $f(y|p,\alpha_1,\lambda_1,\alpha_2,\lambda_2) = p\ \frac{\lambda_1^{\alpha_1}}{\Gamma(\alpha_1)} e^{-\lambda_1 y} y^{\alpha_1-1} I_{\{y>0\}} + (1-p)\ \frac{\lambda_2^{\alpha_2}}{\Gamma(\alpha_2)} e^{-\lambda_2 y} y^{\alpha_2-1} I_{\{y>0\}}$
7. **Trigamma** mezcla convexa de 3 gammas
8. **Cuatrigamma** mezcla convexa de 4 gammas

Para realizar la elección se dispone de una muestra de 202 valores. Como todos son mayores a 450, fueron calibrados sustrayendole 450 a cada uno. Además, se eliminó una observación que era un outlier severo. De aquí en más llamamos observaciones a la muestra "calibrada" y depurada.

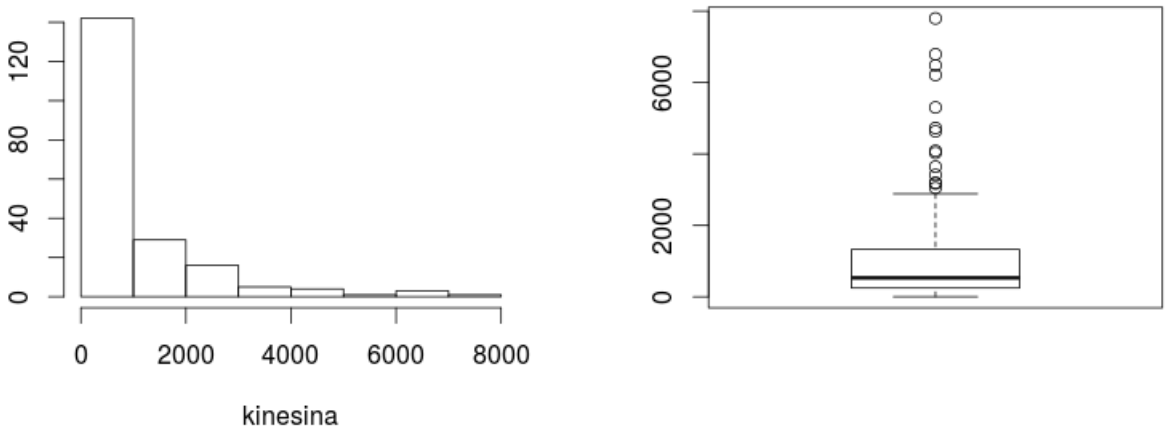


Figura 5.3: Histograma y Boxplot de la muestra calibrada y depurada

5.2.2. Implementación y resultados

Asumimos que la muestra proviene de observaciones independientes de una variable aleatoria con distribución F desconocida. Para cada uno de los modelos paramétricos propuestos llamamos k a la cantidad de parámetros independientes de dicho modelo. Para cada modelo propuesto testeamos

$$H_0 : \ F \in \mathcal{F}_k \qquad \text{vs} \qquad H_1 : F \notin \mathcal{F}_k$$

donde $\mathcal{F}_k$ es la familia paramétrica con k parámetros independientes. Recordemos que el p-valor del test es la probabilidad de haber observado lo que efectivamente se observó (o algo más extremo) bajo H_0. Así, p-valoles altos nos indican que los datos no contradicen la suposición hecha en H_0, es decir, que no hay evidencia en contra de la hipótesis nula y por lo tanto asumimos que el ajuste es bueno. ¿Qué tan grande tiene que ser el p-valor para considerar que el ajuste es bueno? En la literatura hay consenso en pedir un p-valor mayor a 0.2. Cuanto mayor el p-valor, menos evidencia tenemos en contra de la hipótesis nula.

Para poder calcular los p-valores fue necesario computar los estimadores de máxima verosimilitud bajo el respectivo modelo. Para el caso de las familias que son mezcla convexa de gammas se utilizó el paquete mixtools de R que permite calcular estimadores de máxima verosimilitud vía el algorimo Expectation-Maximization. Para el caso de mezcla convexa de exponenciales se utilizó una implementación propia del algorimo EM según los detalles que se dieron en el ejemplo 5 de la página 16. Además de las observaciones y la cantidad de componentes de la mezcla la implementación recibe como input la cantidad máxima de iteraciones y un valor ϵ que será usado como criterio de parada. Los parámetros iniciales con los que se empieza a iterar se construyen de la siguiente manera: si proponemos un modelo con J componentes, la proporción inicial se obtiene normalizando tres observaciones de una variable uniforme en el $[0, 1]$. Según esas proporciones dividimos los datos en J grupos y en cada uno estimamos el parámetro λ de una exponencial como la inversa del promedio de las observaciones en ese grupo. La implementación también permite comenzar a iterar con parámetros dados por el usuario. Para más detalles el código se encuentra en el apéndice.

El test χ^2 supone una elección de ancho y cantidad de intervalos. En este caso se hizo el análisis con dos particiones independientes, que llamaremos Π_1 y Π_2. La primera tiene 10 intervalos mientras que la segunda tiene 15.

El Criterio de Información de Akaike nos proporciona otra forma de evaluar qué tan bien ajusta cierto modelo a los datos observados pero ahora teniendo en cuenta la dimensión del mismo. Para el k-ésimo modelo calculamos:

$$AIC_k = 2k - log \ f(\underline{y}|\hat{\underline{\theta}}_k(\underline{y}))$$

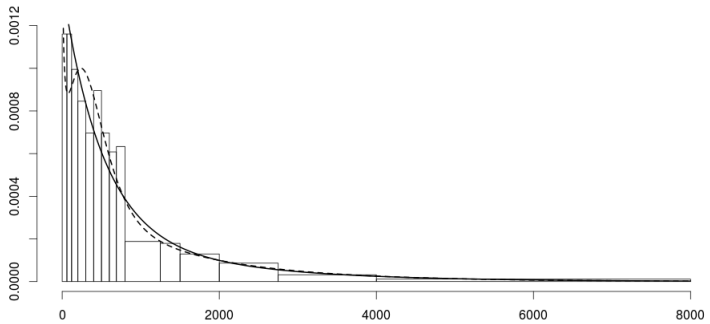
donde recordamos que k es la cantidad de parámetros libres en el modelo, $\underline{y}$ son las observaciones y $\hat{\underline{\theta}}_k(\underline{y})$ es el estimador de máxima verosimilitud bajo el modelo. Podemos interpretar que el modelo con menor valor AIC es aquel que presenta el mejor compromiso entre dimensión y ajuste.

En la siguiente tabla figuran los p-valores de los respectivos tests χ^2 y el valor AIC para cada modelo. Siempre se corroboró que la cantidad de observaciones esperadas (bajo el modelo) para cada intervalo de la partición sea al menos cinco.

Modelo	**k**	**parámetros estimados**	**p-valor** (Π_1)	**p-valor** (Π_2)	**AIC**
Exponencial	1	$\hat{\lambda} = 0{,}000979$	0.01608	0.03605	3189.47
Biexponencial	3	$\hat{p} = (0{,}56834, 0{,}43166)$ $\hat{\lambda} = (0{,}0019997, 0{,}0005855)$	0.1126	0.5378	3178.279
3-exponencial	5	$\hat{p} = (0{,}436, 0{,}1305, 0{,}4334)$ $\hat{\lambda} = (0{,}002012, 0{,}001979, 0{,}0005867)$	0.0356	0.35636	3182.168
4-exponencial	7	$\hat{p} = (0{,}2238, 0{,}1206, 0{,}2231, 0{,}4325)$ $\hat{\lambda} = (0{,}002011, 0{,}002011, 0{,}001988, 0{,}000586)$	0.0057	0.1929	3186.368
Gamma	2	$\hat{\alpha} = 0{,}8141$ $\hat{\lambda} = 0{,}000797$	0.02542	0.1127	3185.443
Bigamma	5	$\hat{p} = (0{,}28868, 0{,}71132)$ $\hat{\alpha} = (3{,}03512, 0{,}73445)$ $\hat{\lambda} = (0{,}0068819, 0{,}000585)$	0.24271	0.50992	3179.119
3-gamma	8	$\hat{p} = (0{,}34657, 0{,}18213, 0{,}4713)$ $\hat{\alpha} = (1{,}0862, 10{,}0728, 1{,}2582)$ $\hat{\lambda} = (0{,}004143, 0{,}01802, 0{,}000715)$	0.10779	0.61595	3181.923
4-gamma	11	$\hat{p} = (0{,}502085, 0{,}19341, 0{,}2858, 0{,}018699)$ $\hat{\alpha} = (1{,}086, 10{,}0642, 3{,}6675, 107{,}8470)$ $\hat{\lambda} = (0{,}003164, 0{,}01751, 0{,}001715, 0{,}01588)$	-	0.25399	3183.989

Vale la pena observar en la tabla la variabilidad que tiene el p-valor según la partición elegida. Por ejemplo, para el modelo 3-exponencial, el p-valor del test es 0.10779 si la partición utilizada es Π_1 mientras que si utilizamos Π_2 es 0.61595. Para el caso de una mezcla convexa de cuatro gammas la cantidad de parámetros independientes (11) supera a la cantidad de intervalos de la partición Π_1 y por lo tanto no podemos realizar el test χ^2.

Vemos que los criterios de selección utilizados sugieren que el modelo biexponencial y bigamma son los más adecuados para ajustar los datos de los run lengths del motor kinesina. Estos modelos presentan los valores AIC más bajos y el p-valor obtenido en los respectivos test χ^2 para la partición Π_2 supera a 0.5. En la siguiente figura vemos estos dos ajustes (en linea punteada el bigamma y en linea continua el biexponencial) sobre el histograma de los datos según Π_2.



En la figura siguiente vemos el histograma de los datos (según la partición dada por Π_1 a la izquierda y según la partición dada por Π_2 a la derecha), junto con la densidad según los parámetros estimados del modelo exponencial (en linea continua negra) y en rojo la frecuencia esperada para los intervalos bajo el modelo. Se presentan las mismas figuras de análisis para los diferentes modelos propuestos.

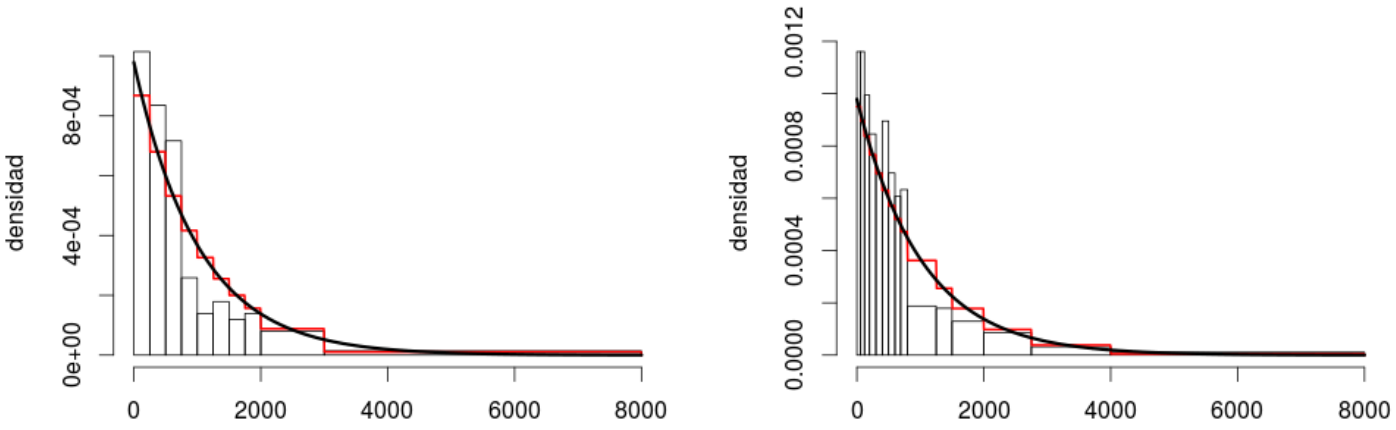


Figura 5.4: Gráficos para el modelo exponencial

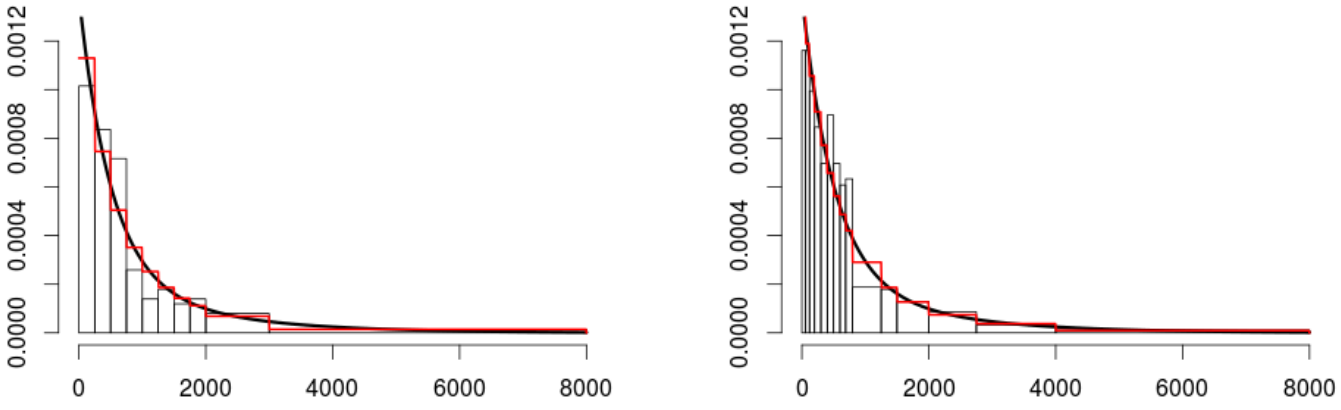


Figura 5.5: Gráficos para el modelo biexponencial

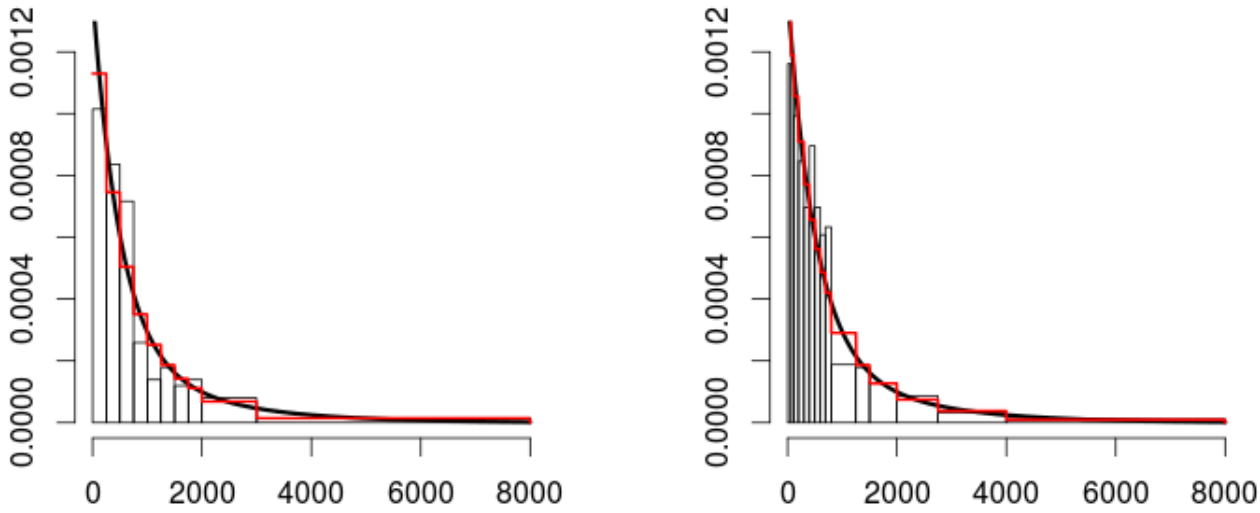


Figura 5.6: Gráficos para el modelo 3-exponencial

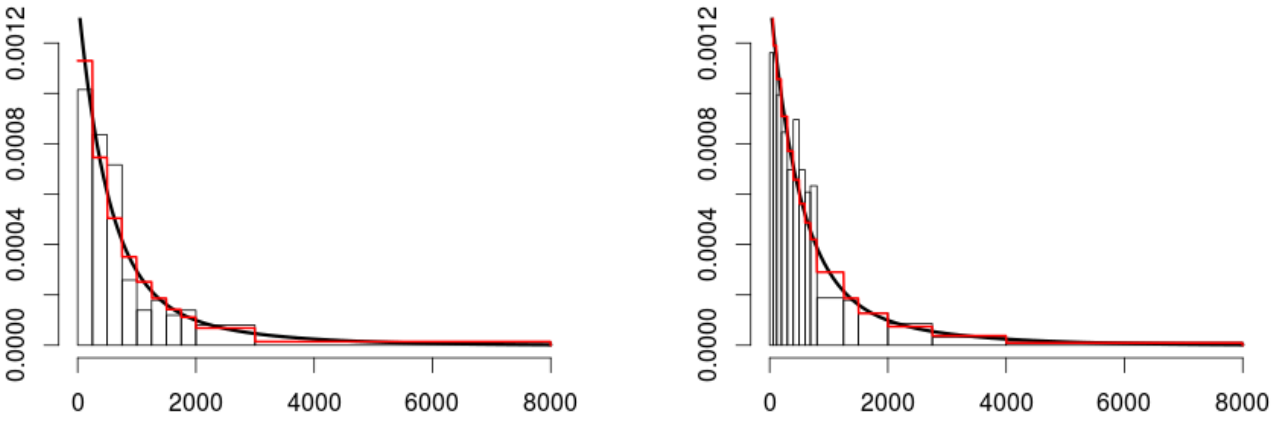


Figura 5.7: Gráficos para el modelo 4-exponencial

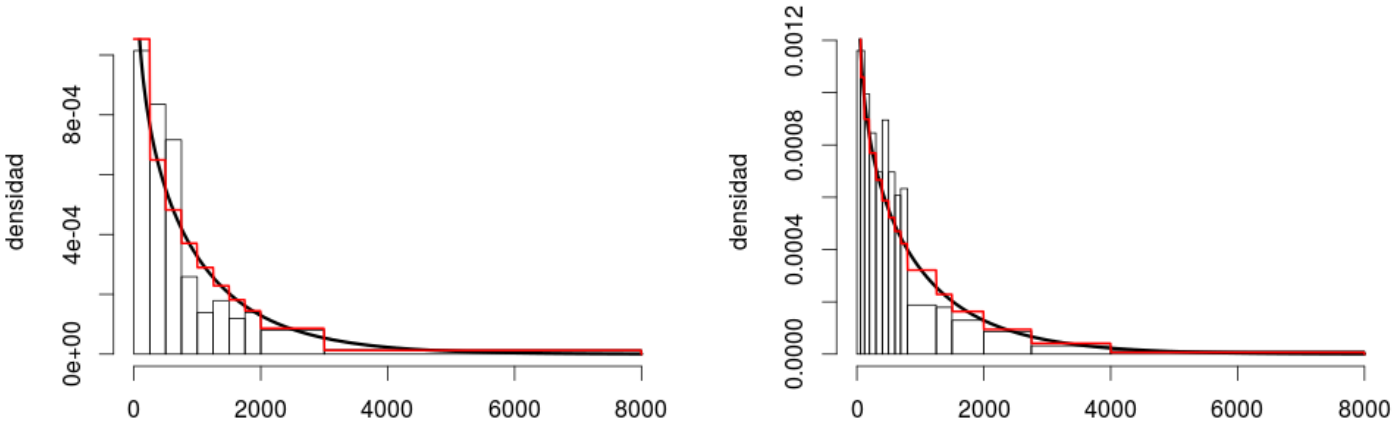


Figura 5.8: Gráficos para el modelo Gamma

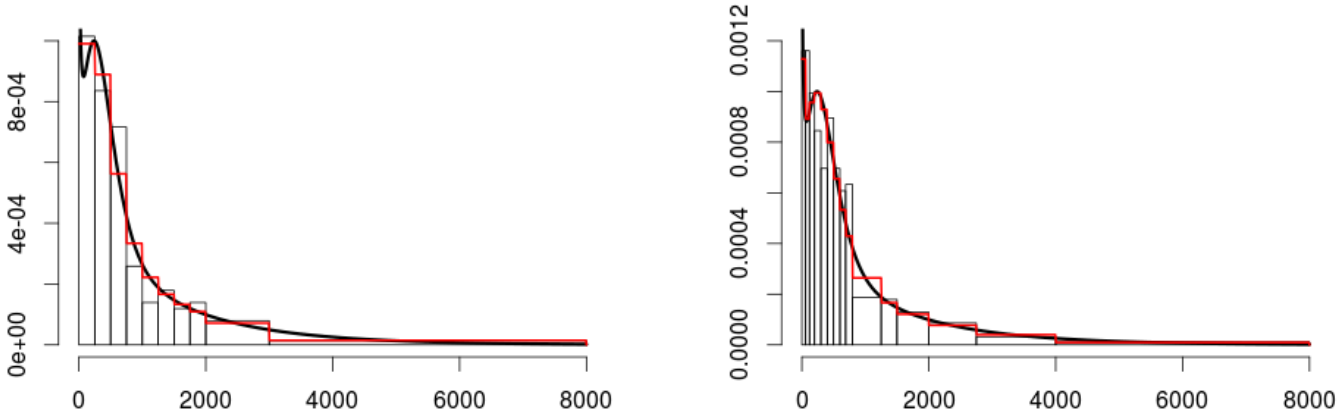


Figura 5.9: Gráficos para el modelo Bigamma

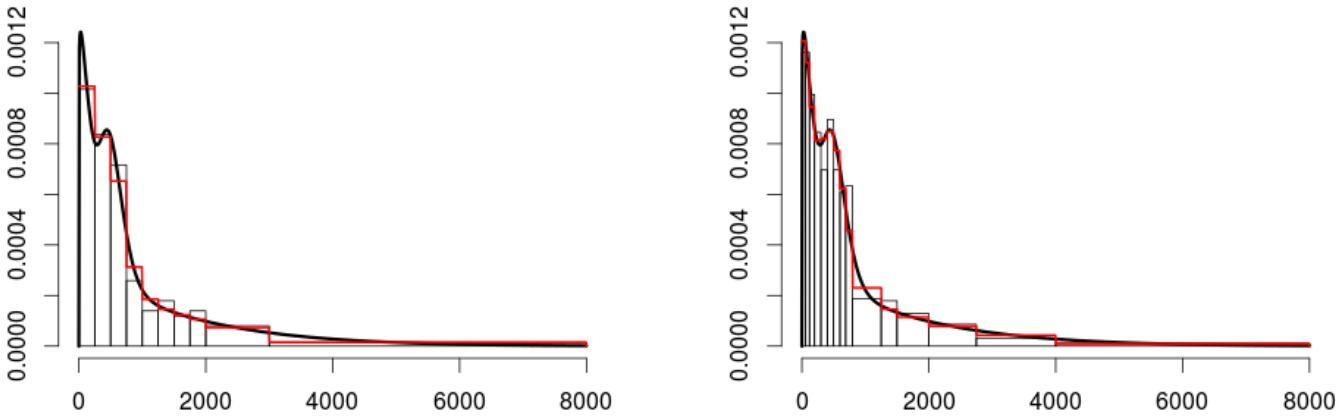


Figura 5.10: Gráficos para el modelo 3-gamma

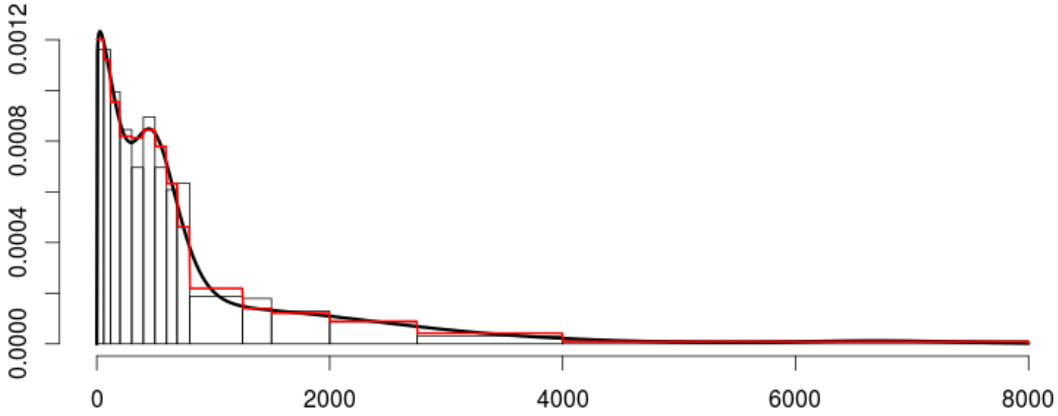


Figura 5.11: Gráfico para el modelo 4-gamma

5.2.3. Simulaciones

¿Qué podemos esperar ver bajo el modelo biexponencial?, ¿que sería razonable observar si en realidad la muestra responde a un modelo exponencial, gamma, bigamma, 3-gamma o 3-exponencial? Para responder estas preguntas simulamos una muestra de 200 de observaciones independientes de una variable biexponencial (resp. exponencial, gamma, bigamma, 3-gamma, 3-exponencial) con los

parámetros del ajuste hecho en la sección anterior y luego las analizamos como lo hicimos con la muestra real. Este procedimiento lo repetimos 200 veces. Reportamos la mediana de los valores AIC obtenidos y la proporción de p-valores mayores 0.2. La particiones utilizadas constan, en todos los casos, de 15 intervalos.

Modelo	$p > 0{,}2$	**AIC**
Exponencial	0.755	3174.053
Biexponencial	0.695	3177.387
3-exponencial	0.5	3181.648
Gamma	0.765	3175.476
Bigamma	0.63	3178.472
3-gamma	0.423	3184.878

Cuadro 5.1: Muestra exponencial

Modelo	$p > 0{,}2$	**AIC**
Exponencial	0.42	3174.107
Biexponencial	0.75	3171.257
3-exponencial	0.585	3175.245
Gamma	0.75	3170.339
Bigamma	0.62	3173.013
3-gamma	0.445	3180.43

Cuadro 5.2: Muestra gamma

Modelo	$p > 0{,}2$	**AIC**
Exponencial	0.13	3171.467
Biexponencial	0.715	3155.617
3-exponencial	0.57	3158.701
Gamma	0.43	3161.72
Bigamma	0.585	3162.194
3-gamma	0.4	3168.194

Cuadro 5.3: Muestra biexponencial

Modelo	$p > 0{,}2$	**AIC**
Exponencial	0.06	3169.811
Biexponencial	0.415	3159.675
3-exponencial	0.65	3163.152
Gamma	0.115	3165.302
Bigamma	0.59	3157.097
3-gamma	0.38	3166.139

Cuadro 5.4: Muestra bigamma

Modelo	$p > 0{,}2$	**AIC**
Exponencial	0.16	3172.045
Biexponencial	0.755	3158.318
3-exponencial	0.585	3162.16
Gamma	0.43	3166.239
Bigamma	0.63	3161.408
3-gamma	0.37	3170.38

Cuadro 5.5: Muestra 3-exponencial

Modelo	$p > 0{,}2$	**AIC**
Exponencial	0.04	3170.43
Biexponencial	0.355	3157.643
3-exponencial	0.22	3161.613
Gamma	0.105	3166.972
Bigamma	0.465	3158.929
3-gamma	0.41	3166.304

Cuadro 5.6: Muestra 3-gamma

En casi todos los escenarios el criterio AIC señala como mejor modelo a aquel que efectivamente generó las observaciones. Las excepciones son los casos de muestras 3-exponenciales y 3-gamma en donde, en ambos casos, el modelo biexponencial es el elegido por el Criterio de Información de Akaike. Asimismo el test χ^2 que coloca como hipótesis nula al verdadero modelo que generó las observaciones tiene una proporción bastante alta de p-valores mayores a 0.2.

Apéndice A

Códigos

A.1. Parámetros iniciales

```
##############################################################################
# expmix.parametrosIniciales
##############################################################################
#INPUT
# x : observaciones
# prop : proporciones inciales
# landaa : estimacion inicial de landa
# k : cantidad de componentes

#OUT
# prop : proporciones iniciales
# landa: estimacion inicial de landa
# k : cantidad de componentes
##############################################################################

expmix.parametrosIniciales<-function (x, prop = NULL, landaa = NULL, k = 2)
{
  n <- length(x)
  if (is.null(prop)) {
    prop = runif(k) # las proporciones las inicializo normalizando una uniforme
    prop = prop/sum(prop)
  }
  else k = length(prop)
  if (k == 1) {
    x.bar = mean(x) }
  else {
    ind = floor(n * cumsum(prop))
    x.part = list()
    x.part[[1]] = x[1:(ind[1] + 1)] #divido los datos en proporciones segun prop
    for (j in 2:k) {
      x.part[[j]] = x[ind[j - 1]:ind[j]]
    }
    x.bar = sapply(x.part, mean)
  }
  if (is.null(landaa)) {
    landaa = 1/x.bar }
  list(prop = prop, landaa = landaa, k = k)
}
```

A.2. Expectation Maximization para combinación convexa de exponciales

```
##############################################################################
#expmixEM
##############################################################################

#INPUT
# Y: observaciones
# p: proporciones iniciales
# landa : estimacion de parametros incicial
# J : cantidad de componentes
# maxit: cantidad maxima de iteraciones
# epsilon : criterio de parada

#OUT
#p : proporciones estimadas
#landa : parametros estimados
#loglik: log verisimilitud
#all.loglik: todas las log-verosimilitudes calculadas
#ft : nombre de la funcion (expmixEM)
##############################################################################

source("exp_init.R")

expmixEM<-function(Y, p = NULL , landa = NULL, J = 2, maxit = 1000, epsilon = 1e-05, graficos=NULL){

n<-length(Y)
densidady<-function(y,LANDA,P){# densidad de y|p, landa
  aux<-0
  for (j in 1:length(P)){ aux <- aux + P[j]*dexp(y,LANDA[j])}
  return(aux)
}

# Parametros Iniciales (eleccion segun expmix.parametrosIniciales)
iniciales<-expmix.parametrosIniciales(x = Y, prop = p, landaa= landa, k=J)
p.estimado<- iniciales$prop
landa.estimado <- iniciales$landaa

p.estimado.new<-rep(NA,J)
landa.estimado.new<-rep(NA,J)

# a l g o r i t m o E M
iter <- 0
diferencia <- epsilon + 1
ll.old<-sum(log(densidady(Y, landa.estimado, p.estimado)))
all.ll<-c(ll.old)

while ( diferencia > epsilon && iter< maxit ){
  iter <- iter + 1
    for(j in 1:J){
    numLanda<-0
    denLanda<-0
    #p.estimado y landa.estimado son las estimaciones actuales de p y landa
       for (i in 1:n){
           pIJmonio<- (p.estimado[j] * dexp(Y[i],landa.estimado[j]))/ densidady(Y[i],landa.estimado,p.estimado)
           denLanda<-denLanda + pIJmonio * Y[i]
            numLanda<-numLanda + pIJmonio
        }
    landa.estimado.new[j]<-numLanda/denLanda
    p.estimado.new[j]<-numLanda/n
    }
  #reasignacin de variables antes de pasar a la proxima iteracion
  landa.estimado<-landa.estimado.new
  p.estimado<-p.estimado.new
```

```
  ll.new<-sum(log(densidady(Y,landa.estimado, p.estimado)))
  diferencia<-ll.new-ll.old
  #agrego la ll con los parametros recien estimados
  ll.old<-ll.new
  all.ll<-c(all.ll,ll.old)
}# fin while

if (iter == maxit){
  cat("\ Cuidado, no converge. Se alcanzo el nmero mximo de iteraciones\ \n ")
}
a = list(p = p.estimado, landa = landa.estimado, loglik = ll.new
        ,all.loglik = all.ll , ft = "expmixEM")
class(a) = "expmixEM"
a
}
```

Bibliografía

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Breakthroughs in Statistics, Vol. I, Foundations and Basic Theory, S. Kotz and N. L. Johnson, eds., Springer-Verlag, New York, 1992, 610-624. (Originally published in Proceedings of the Second International Symposium of Information Theory, B.N. Petrov and F. Caski, eds., Akademia Kiado, Budapest, 1973,267-281)*, 1973.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions of automatic control*, 1974.

[3] D.B. Rubin A.P. Dempster, N.M. Laird. Maximum likelihood from incomplete data via the EM algorithm(with discussion). *Journal of the Royal Statistics Society*, 1977.

[4] H. Bozdogan. Model selection and akaike information criterion (aic): the general theory and its analytical extensions. *Psychometrika*, 1987.

[5] K.P. Burnham and D.R. Anderson. *Information theory and log-likelihood models: a basis for model selection and inference.* Springer, 1998.

[6] K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference:A Practical Information-Theoretic Approach.* Springer, 2002.

[7] H. Chernoff and E. L. Lehmann. The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics*, 1954.

[8] J. deLeeuw. Introduction to Akaike information theory and an extension of the maximum likelihood principle. *Breakthroughs in Statistics, Vol. I, Foundations and Basic Theory, S. Kotz and N. L. Johnson, eds., Springer-Verlag*, 1992.

[9] R.A. Fisher. The conditions under which χ^2 measures the discrepancy between observations and hypothesis. *Annals of Mathematical Statistics*, 1924.

[10] P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, 1967.

[11] I. A. Kieseppá. Akaike information criterion, curve-fitting, and the philosopical problem of simplicity. *The British Journal for the Philosophy of Science*, 1997.

[12] S. Kullback. *Information theory and statistics.* Donver, 1968.

[13] H.B. Mann and A. Wald. On the choice of the number of class intervals in the application of the chi square test. *Annals of Mathematical Statistics*, 1942.

[14] G.D. Murray. Contribution to discussion of paper by Dempster, Laird and Robin. *Journal of the Royal Statistics Society*, 1977.

[15] Patricia Olivella. Autopistas celulares. *El Cable*, 2012.

[16] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosopical Magazine*, 1900.

[17] K. Pearson, R. A. Fisher, and H. F. Inman. K. Pearson and R.A. Fisher on statistical test: A 1935 exchange from Nature. *The American Statistician*, 1994.

[18] A.W. Van Der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[19] A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 1943.

[20] A. Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 1949.

[21] C. A. Williams Jr. On the choice of the numbe and width of classes for the chi-square test of goodnes of fit. *Journal of the American Statistical Association*, 1950.

[22] C.F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 1983.

[23] R. H. Zamar. An introduction to the EM algorithm and its applications. Department of Statistics, University of British Columbia.

[24] W.I. Zangwill. *Nonlinear programming: a unified approach*. Prentice Hall, Englewood Cliffs, New Yersey, 1969.