UNIVERSIDAD CENTROCCIDENTAL
"LISANDRO ALVARADO"

Decanato de Ciencias y Tecnología
Licenciatura en Ciencias Matemáticas

"Una Versión Estocástica del Algoritmo EM"

Trabajo Especial de Grado presentado por

Br. Emelyn Camacho

como requisito final
para obtener el título de Licenciada
en Ciencias Matemáticas
Área de Conocimiento: Probabilidad y Estadística.
Tutor: Lic. Msc. Jhonny Escalona

Barquisimeto, Venezuela. Febrero de 2012

Universidad Centroccidental
"Lisandro Alvarado"
Decanato de Ciencias y Tecnología
Licenciatura en Ciencias Matemáticas

ACTA
TRABAJO ESPECIAL DE GRADO

Los suscritos miembros del Jurado designado por el Jefe del Departamento de Matemáticas del Decanato de Ciencias y Tecnología de la Universidad Centroccidental "Lisandro Alvarado", para examinar y dictar el veredicto sobre el Trabajo Especial de Grado titulado:

"Una Versión Estocástica del Algoritmo EM"

Presentado por la ciudadana Br. Emelyn Camacho titular de la Cédula de Identidad No. 17.699.744, con el propósito de cumplir con el requisito académico final para el otorgamiento del título de Licenciada en Ciencias Matemáticas.

Luego de realizada la Defensa y en los términos que imponen los Lineamientos para el Trabajo Especial de Grado de la Licenciatura en Ciencias Matemáticas, se procedió a discutirlo con el interesado habiéndose emitido el veredicto que a continuación se expresa:

[1] _____

Con una calificación de _____ puntos.

En fe de lo expuesto firmamos la presente Acta en la Ciudad de Barquisimeto a los _____ días del mes de _____ de _____.

_____	_____
TUTOR	FIRMA
_____	_____
PRINCIPAL	FIRMA
_____	_____
PRINCIPAL	FIRMA

OBSERVACIONES:

_____

[1] Aprobado ó Reprobado

A Dios, a mis padres y a todas aquellas personas que de una u otra manera contribuyeron para alcanzar esta meta.

Agradecimientos

Primeramente agradezco a Dios Todopoderoso por darme la vida, por tener salud y por estar siempre a mi lado en los buenos y malos momentos, por ayudarme a alcanzar esta meta.

A mis padres: Ismael Camacho y Nacarit de Camacho y hermanos por brindarme su apoyo incondicional y su amor sincero.

A mi esposo Leandro Sanchez por su ayuda, por sus consejos y su amor brindado.

Al profesor Jhonny Escalona por su apoyo y ayuda para la realización de este trabajo muchas gracias.

A mis compañeros y amigos: Francisco López, Kissy Àlvarez y Javier Montes por su amistad su compañia y apoyo durante toda la carrera.

Resumen

En este trabajo se desarrollan los aspectos teóricos del algoritmo EM y se estudia una versión estocástica de dicho algoritmo cuando el paso E del algoritmo EM no se pueda calcular de forma cerrada. Se implementan en matlab los algoritmos estudiados para algunos ejemplos particulares.

Índice

Agradecimientos **i**

Resumen **iii**

Introducción **1**

1. Preliminares **2**

1.1. Estimación por Máxima Verosimilitud 2

1.2. Estadísticos Suficientes. 4

1.3. Método de Monte Carlo con Cadenas de Markov. 5

1.3.1. Cadenas de Markov. 6

1.3.2. Distribución de una Cadena de Markov 7

1.4. Métodos MCMC. 8

2. Algoritmo EM **11**

2.1. Propiedades Generales 16

2.2. Algoritmo MCEM 21

3. Algoritmo SAEM **22**

3.1. Método de Aproximaciones Estocásticas. 22

3.2. Pasos del Algoritmo SAEM. 28

3.3. Ventajas y Desventajas del Algoritmo EM y SAEM 33

Apéndice **34**

Referencias **36**

Índice de figuras

2.1. Convergencia del parámetro π 13
2.2. Convergencia del log L 13

3.1. Convergencia del parámetro π 32
3.2. Convergencia del log L 32

INTRODUCCIÓN

El algoritmo EM aparece por primera vez en 1958 en un artículo publicado por Hartley et al. [3]. Pero es en 1977 cuando Dempster, Laird, Rubin et al. [2] dan los aspectos teóricos del algoritmo EM, en dichos trabajos se dan teoremas de convergencia en el caso donde los datos provienen de familias exponenciales. Desde entonces, la popularidad del algoritmo ha ido creciendo debido a su simplicidad y estabilidad. Algunos de los trabajos más recientes basados en el algoritmo EM combinan métodos Monte Carlo con Cadenas de Markov para aproximar el cálculo de la esperanza de la verosimilitud de los datos completos en aquellos casos donde no se puede obtener una expresión cerrada. Para este problema también se ha implementado Aproximaciones Estocásticas [4].

En este trabajo estudiamos en detalle el algoritmo EM siguiendo el artículo de Dempster et al. [2] y parte del artículo de Kuhn et al. [4] donde aparece el algoritmo SAEM, también implementamos los algoritmos EM y SAEM para un modelo estadístico multinomial.

Este trabajo está estructurado de la siguiente manera:

En el primer capítulo mostramos definiciones, propiedades, teoremas que fueron necesarias para la realización de este trabajo.
En el segundo capítulo mostramos las bases teóricas del Algoritmo EM y el Algoritmo MCEM.
En el tercer capítulo mostramos el Algoritmo SAEM y sus Aproximaciones Estocásticas.

Capítulo 1

Preliminares

En este capítulo se presentan fundamentalmente definiciones basadas en el análisis Estadistíco-Matemático, las cuales proporcionan las herramientas para el desarro llo de la investigación, estas son:

- Estimación por Máxima Verosimilitud.
- Estadísticos Suficientes.
- Método de Monte Carlo con cadenas de Markov.

1.1. Estimación por Máxima Verosimilitud

Veamos algunos conceptos previos para la estimación por Máxima Verosimilitud.

Definición 1.1.1. Una **muestra aleatoria** de tamaño n es una sucesión $X_1, \ldots, X_n$ de variables aleatorias independientes idénticamente distribuidas. A la distribución de las X_i la llamaremos **distribución de la población.**

Definición 1.1.2. Sea $X_1, \ldots, X_n$ es una muestra aleatoria y $g : \mathbb{R}^n \to \mathbb{R}$ una función, en la práctica usualmente g es una función continua, entonces al número $g(X_1, \ldots, X_n)$ se le denomina **estadístico**.

Proposición 1.1.1. *Si $X_1, \ldots, X_n$ es una muestra aleatoria de una población con media μ y varianza σ^2, entonces*

$$\mathbb{E}(\bar{X}) = \mu \quad var(\bar{X}) = \frac{\sigma^2}{n}.$$

Los estadísticos se utilizan para estimar una caracterización particular de la distribución de la población. Por ejemplo, la estimación puntual de la media μ de una población toma el resultado de el estadístico $\bar{X}$. Por esta razón, tal estadístico es llamado un **estimador** de μ. Si el estimador $\bar{X}$ de μ tiene la siguiente propiedad:

$$\mathbb{E}(\bar{X}) = \mu$$

Diremos que $\bar{X}$ es un **estimador insesgado** de μ. La teoría de la estimación es una rama importante de la estadística que se basa en la construcción de estimadores que son óptimos en algún sentido.

Definición 1.1.3. Sea $X_1, \ldots, X_n$ una muestra de variables aleatorias con una densidad de probabilidad f. La función de verosimilitud asociada con esta muestra se entiende como la densidad de probabilidad del vector $(X_1, \ldots, X_n)$. En otras pala bras, la función de verosimilitud $L : \mathbb{R}^n \to [0, +\infty)$ es la función definida por

$$L(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n) \quad ((x_1, \ldots, x_n) \in \mathbb{R}^n) \tag{1.1}$$

Si en una determinada región $A \subset \mathbb{R}^n$ la función de verosimilitud toma solo valores pequeños, entonces no es probable que las observaciones $(x_1, \ldots, x_n)$ sea un elemento de A. Esto se puede explicar simplemente escribiendo

$$\mathbb{P}((X_1, \ldots, X_n) \in A) = \int \ldots \int_A L(x_1, \ldots, x_n) dx_1, \ldots, dx_n. \tag{1.2}$$

Dados $(x_1, \ldots, x_n)$ en donde L toma un valor pequeño por lo tanto se dice que es improbable. En una línea directa con esta terminología llamaremos estas observaciones $(x_1, \ldots, x_n)$ probable si $L(x_1, \ldots, x_n)$ asume un valor grande.

A continuación, supongamos que se toma una muestra aleatoria $X_1, \ldots, X_n$ de una población con densidad de probabilidad $f(\bullet, \theta)$ donde $\theta \in \Theta$. La función de verosimilitud depende ahora de $\theta \in \Theta$, por lo tanto, se denota por L_θ en lugar de L. El experimento da como resultado $(x_1, \ldots, x_n) \in \mathbb{R}^n$ de $(X_1, \ldots, X_n)$. Ahora elegimos en Θ un elemento $\hat{\theta}$, que maximize la función

$$\theta \longmapsto L_\theta(x_1, \ldots, x_n). \tag{1.3}$$

Supondremos que dadas las observaciones $(x_1, \ldots, x_n)$, existe un unico θ que maximiza la función de máxima verosimilitud L_θ y lo denotaremos $\hat{\theta}$.

En general, $\hat{\theta}$ depende de las observaciones $(x_1, \ldots, x_n)$. Por lo tanto escribiremos

$$\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n). \tag{1.4}$$

Ejemplo 1. Dada una muestra de variables aleatorias $X_1, \ldots, X_n$, con una distribución de población $N(\mu, \sigma^2)$, donde μ y σ^2 son desconocidos. Dada las observaciones $(x_1, \ldots, x_n)$ del vector estocástico $(X_1, \ldots, X_n)$. Los estimadores de máxima verosimilitud del vector (μ, σ^2) son $\mu = \bar{x}$, $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

1.2. Estadísticos Suficientes.

Definición 1.2.1. Sea $X_1, X_2, ..., X_n$ una variable aleatoria cuya distribución de probabilidad pertenece a una familia de distribuciones dadas por $\mathcal{F} = \{F_\theta | \theta \in \Theta\}$, se dice que el estadístico $U = g(x_1, x_2, ..., x_n)$ es suficiente para θ o para la familia, si y sólo si, la distribución condicionada de $x_1, x_2, ..., x_n | U$ no depende de θ.

En la práctica, para decidir si un estadístico es suficiente (para un parámetro θ) no suele hacerse directamente a partir de la definición anterior, sino que se utiliza un criterio que facilita su comprobación. Dicho criterio recibe el nombre de criterio de factorización de Neyman-Fisher y lo enunciaremos mediante el siguiente teorema:

Teorema 1.2.1. *Dada una muestra aleatoria $X_1, X_2, ..., X_n$ procedente de una población con función de densidad $f(x, \theta)$, diremos que un estadístico U es suficiente para el parámetro θ, si y sólo si, la función de densidad conjunta de la muestra puede factorizarse de la siguiente manera:*

$$f(x_1, x_2, ..., x_n; \theta) = g(u, \theta)h(x_1, x_2, ..., x_n)$$

donde $u = U(x_1, x_2, ..., x_n)$

Ejemplo 2. Dada una muestra de variables aleatorias $X_1, \ldots, X_n$ con $f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$ de parámetro λ.
En efecto, la función de densidad conjunta de la muestra es la siguiente:

$$\begin{aligned} f(x_1, x_2, ..., x_n; \lambda) &= f(x_1; \lambda).f(x_2; \lambda)...f(x_n; \lambda) \\ &= \exp^{-\lambda}\frac{\lambda^{x_1}}{x_1!}.\exp^{-\lambda}\frac{\lambda^{x_2}}{x_2!}...\exp^{-\lambda}\frac{\lambda^{x_n}}{x_n!} \\ &= \exp^{-n\lambda}\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \end{aligned} \tag{1.5}$$

equivalente a:

$$f(x_1, x_2, ..., x_n; \lambda) = \exp^{-n\lambda}\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} = g\left(\sum_{i=1}^n x_i, \lambda\right) h(x_1, x_2...x_n) \tag{1.6}$$

donde el estadístico suficiente para λ es:

$$U(x_1, x_2, ..., x_n) = \sum_{i=1}^n x_i$$

y su factorización es la siguiente:

$$g\left(\sum_{i=1}^n x_i, \lambda\right) = \exp^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}; \quad h(x_1, x_2...x_n) = \prod_{i=1}^n x_i!$$

1.3. Método de Monte Carlo con Cadenas de Markov.

Integración Monte Carlo El método Monte Carlo fué desarrollado por los físicos (Uhlam- Von Newman) durante la segunda guerra mundial, usando una gene ración de números aleatorios para calcular integrales. Suponga que se desea calcular una integral compleja:

$$\int_a^b h(x)dx, \tag{1.7}$$

si se puede descomponer $h(x)$ como el producto de una función $f(x)$ y una densidad de probabilidad $p(x)$ definida en un intervalo (a, b), entonces:

$$\int_a^b h(x)dx = \int_a^b f(x)p(x)dx = E_{p(x)}\left[f(x)\right]. \tag{1.8}$$

También la integral puede ser expresada como la esperanza de $f(x)$ sobre la densidad $p(x)$. Por tanto, si se tiene una muestra grande $x_1, \ldots, x_n$ obtenida de la densidad $p(x)$, entonces por la Ley Fuerte de los Grandes Números se tiene:

$$\int_a^b h(x)dx = E_{p(x)}\left[f(x)\right] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i), \tag{1.9}$$

esto se refiere a la Integración de Monte Carlo.

La Integración Monte Carlo puede ser usada para aproximar la densidad a posteriori (o marginal posterior) requerida por el análisis Bayesiano.

Considere la integral:

$$I(y) = \int f(y|x)p(x)dx, \tag{1.10}$$

con aproximación

$$I(y) = \frac{1}{n}\sum_{i=1}^{n} f(y|x_i), \tag{1.11}$$

donde x_i son muestras para la densidad $p(x)$. La estimación del error estándar Monte Carlo está dado por:

$$SE^2\left[\hat{I}(y)\right] = \frac{1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(f(y|x_i) - \hat{I}(y)\right)^2\right). \tag{1.12}$$

1.3.1. Cadenas de Markov.

Una sucesión $\{X_t : t \in \mathbb{Z}_+\}$ de variables aleatorias con valores en un conjunto E se denomina un *proceso aleatorio a tiempo discreto* con espacio de estados E. En este trabajo, el espacio de estados es finito, y sus elementos son denotados por $E = \{E_1, E_2, \ldots, E_n\}$. Si $X_t = E_i$, se dice que el proceso esta en el estado i a tiempo t.

Definición 1.3.1. Sea $\{X_t : t \in \mathbb{Z}_+\}$ un proceso aleatorio a tiempo discreto con espacio numerable E. Sea $t \in \mathbb{Z}_+$ y estados $E_0, E_1, \ldots, E_i, E_j$,

$$P(X_{t+1} = E_j | X_t = E_i, \ldots, X_0 = E_0) = P(X_{t+1} = E_j | X_t = E_i) \tag{1.13}$$

Cuando ambos lados de la ecuación (1.13) están bien definidos, este proceso es denominado una *cadena de Markov*. Si el lado derecho de la ecuación (1.13) es independiente de t entonces la cadena se dice que es *homogénea.*

La ecuación (1.13) es usualmente denominada la *propiedad de Markov*. La matriz $A = \{a_{ij}\}_{i,j \in E}$, donde

$$a_{ij} = P(X_{t+1} = E_j | X_t = E_i)$$

es denominada la *matriz de transición* de la cadena de Markov. La matriz de transición tiene las siguientes propiedades:

- $a_{ij} \geq 0, \forall\ E_i, E_j \in E$.
- $\sum_{k \in E} a_{ik} = 1$.

Una matriz A indexada en E y satisfaciendo las propiedades anteriores es llamada una matriz aleatoria.

Para simplificar la notación, utilizaremos de ahora en adelante i_t en vez de E_{i_t}.

1.3.2. Distribución de una Cadena de Markov

La variable aleatoria X_0 es llamado el *estado inicial*, y su distribución π,

$$\pi(i) = P(X_0 = i) \tag{1.14}$$

es la *distribución inicial*. De la propiedad de Markov se tiene

$$\begin{aligned} P(X_0 = i_0, \ldots, X_t = i_t) &= P(X_0 = i_0) P(X_1 = i_1 | X_0 = i_0) \cdots \\ &\quad \times P(X_t = i_t | X_{t-1} = i_{t-1}, \ldots, X_0 = i_0) \\ &= P(X_0 = i_0) P(X_1 = i_1 | X_0 = i_0) \cdots P(X_t = i_t | X_{t-1} = i_{t-1}) \end{aligned}$$

y en términos de la notación definida anteriormente,

$$P(X_0 = i_0, \ldots, X_t = i_t) = \pi(i_0)a_{i_0 i_1} \cdots a_{i_{t-1} i_t} \tag{1.15}$$

La ecuación (1.15) para todo $k \geq 0$ y estados $i_0, i_1, \ldots, i_k$ es la *distribución* de una Cadena de Markov homogénea. Por tanto se tiene el siguiente resultado.

Las ecuaciones de Chapman-Kolmogorov proporcionan un método para el cálculo de las probabilidades de transición de n pasos. Estas ecuaciones son:

$$P_{i,j}^{n+m} = \sum_{k=0}^{\infty} p_{i,k}^{n} P_{k,j}^{m} \quad \forall n, m \geq 0, \quad \forall i, j$$

donde $p_{i,k}^{n} P_{k,j}^{m}$ representa la probabilidad que a partir del estado i el proceso pasará al estado j en $n + m$ transiciones.

1.4. Métodos MCMC.

En ocasiones es difícil simular el valor de un vector aleatorio X cuyas variables aleatorias componentes son independientes. A continuación presentamos un método llamado Método de Monte Carlo con Cadenas de Markov que permite generar un vector cuya distribución es aproximadamente a la de X y tiene la ventaja adicional de que la función de masa (o densidad) de X puede estar dada salvo una constante multiplicativa, lo que es de gran importancia en la aplicaciones.

Sea X un vector aleatorio discreto cuyo conjunto de valores posibles es x_j, con $j \geq 1$. Sea la función masa de probabilidad de X dada por $P\{X = x_j\}$, $j \geq 1$, y supongamos que estamos interesados en calcular

$$\theta = E[h(X)] = \sum_{j=1}^{\infty} h(x_j) P\{X = x_j\}$$

para alguna función específica h. En ocasiones cuando es difícil evaluar la función $h(x_j)$ con $j \geq 1$, es necesasio recurrir a la simulación este método es llamado **método de Monte Carlo con Cadena de Markov** es usado para generar una secuencia parcial de números aleatorios independientes y vectores aleatorios idénticamente ditribuidas $X_1, X_2, \ldots, X_n$ con una función de masa $P\{X = x_j\}$, $j \geq 1$.

Sean $b(j)$, $j = 1, ..., m$ números positivos, y sea $B = \sum_{j=1}^{m} b(j)$. Suponga que m es grande, que B es difícil de calcular y que queremos similuar una variable aleatoria(o una sucesión de varibles aleatorias) con función de masa de probabilidad

$$\pi(j) = b(j)/B, \quad j = 1, ..., m$$

Una forma de simular una sucesión de varibles aleatorias cuyas ditribuciones convergen a $\pi(j)$, $j = 1, ..., m$, consiste en determinar una Cadena de Markov que sea fácil de simular y cuyas probabilidades límites sean las $\pi(j)$. **El algoritmo de Hastings-Metropolis** proporciona un método para realizar esta tarea. A continuación presentamos el algoritmo de Metropolis-Hastings que permite generar una cadena de Markov reversible en el tiempo, cuyas probabilidades limites son $\pi(j) = b(j)/B$ $j = 1, ..., m$.

(a) Elegir una matriz Q de probabilidades de transición, de Markov, irreducible, con probabilidades de transición $q(i, j)$,$i, j = 1, ..., m$. Además, elegir algún valor entero k entre 1 y m.

(b) Sean $n = 0$ y $X_o = k$.

(c) Generar una variable aleatoria X tal que $P\{X = j\} = q(X_n, j)$ y generar un número aleatorio U.

(d) Si $U < [b(X)Q(X, X_n)]/[b(X_n)q(X_n, X)]$, entonces $NS = X$; en caso contrario, $NS = X_n$.

(e) $n = n + 1$, $X_n = NS$.

(f) Ir al paso 3.

Sea $X = (X_1, ..., X_n)$ un vector aleatorio con función de masa de probabilidad (o función de densidad de probabilidad en el caso continuo) $p(x)$, la cual solo esta determinada salvo una costante multiplicativa, y supongamos que queremos generar

un vector aleatorio cuya distribución condicional de X, dado que $X \in A$ para algún conjunto A. Es decir, queremos generar un vector aleatorio con función de masa

$$f(x) = \frac{p(x)}{P\{X \in A\}}, \quad para \quad x \in A$$

El muestreador de Gibbs supone que para cualquier i, $i = 1, ..., n$ y cualesquiera valores x_j, $j \neq i$, podemos generar una variable aleatoria X con la función de masa de probabilidad

$$P\{X = x\} = P\{X_i = x | X_j = x_j, j \neq i\}$$

En resumen, vemos que la Cadena de Markov reversible en el tiempo, con proba bilidad estacionarias dadas por f, generada por el muestreador de Gibbs es la siquiente:

(a) Sea $x = (x_1, ..., x_n)$ un vector en A para el cual $p(x) > 0$.

(b) Sea I una variable aleatoria que toma unos de los valores $1, ..., n$, donde cada valor tiene la misma probabilidad de ocurrir.

(c) Si $I = i$, generar el valor de una variable aleatoria X tal que

$$P\{X = x\} = P\{X_i = x | X_j = x_j, j \neq i\}$$

(d) Si $X = x$ y $(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_n) \in A$, entonces el nuevo valor de x_i es igual a x. En caso contrario, se mantiene el valor de x_i.

(e) Regresar al paso 2.

Capítulo 2

Algoritmo EM

El algoritmo EM es una técnica utilizada para encontrar el estimador de máxima verosimilitud de una distribución de un conjunto de datos incompletos. El término "datos incompletos" en su forma general implica la existencia de dos muestras aleatorias $\mathcal{X}$, $\mathcal{Y}$ y una función de $\mathcal{X}$ a $\mathcal{Y}$. Sean y los datos observados de la muestra $\mathcal{Y}$, el correspondiente x de $\mathcal{X}$ no es directamente observable, pero si lo es indirectamente a traves de y. Específicamente, supondremos que existe una función $x \to y(x)$ de $\mathcal{X}$ a $\mathcal{Y}$ y x es conocido en $\mathcal{X}(y)$, el subconjunto de $\mathcal{X}$ determinado por la ecuación $y = y(x)$ donde y son los datos observados. A x lo llamaremos datos completos.
Observemos que

$$g(y|\Phi) = \int_{\mathcal{X}(y)} f(x|\Phi)dx$$

El algoritmo EM encuentra el valor de Φ que maximiza a $g(y|\Phi)$ comenzando con un valor inicial para Φ. En cada iteración p el algoritmo realiza dos pasos, en el primer paso se calcula la esperanza $Q(\Phi|\Phi^{(p)}) = E\left[\log(f(x|\Phi))|y, \Phi^{(p)}\right]$, donde $\Phi^{(p)}$ es el estimado de Φ en la iteración p y en el segundo paso maximizamos $Q(\Phi|\Phi^{(p)})$. Para fijar ideas, mostraremos un ejemplo indicando los pasos del algoritmo.

Ejemplo 3. 197 individuos estan distribuidos multinomialmente en cuatro categorías, los datos observados son:

$$y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

supongamos que las probabilidades de que una observación este en cada una de las categorías es respectivamente

$$(p_1, p_2, p_3, p_4) = \left(\frac{1}{2} + \frac{\pi}{4}, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{\pi}{4}\right)$$

para algún $\pi \in [0,1]$ desconocido y el cual se desea estimar. Así

$$g(y|\pi) = \frac{(y_1+y_2+y_3+y_4)!}{y_1!y_2!y_3!y_4!}\left(\frac{1}{2}+\frac{\pi}{4}\right)^{y_1}\left(\frac{1}{4}(1-\pi)\right)^{y_2}\left(\frac{1}{4}(1-\pi)\right)^{y_3}\left(\frac{\pi}{4}\right)^{y_4}$$

Para ilustrar el algoritmo EM, representamos a y como los datos incompletos de una población multinomial de 5 categorías y las probabilidades de que un individuo pertenezca a cada categoría son $\left(\frac{1}{2}+\frac{\pi}{4}, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{\pi}{4}\right)$, la idea es dividir la primera de las cuatro categorías en dos categorías. Así, los datos completos son $x = (x_1, x_2, x_3, x_4, x_5)$, donde $y_1 = x_1+x_2$, $y_2 = x_3$, $y_3 = x_4$, $y_4 = x_5$ y la verosimilitud de los datos completos es

$$f(x|\pi) = \frac{(x_1+x_2+x_3+x_4+x_5)!}{x_1!x_2!x_3!x_4!x_5!}\left(\frac{1}{2}\right)^{x_1}\left(\frac{\pi}{4}\right)^{x_2}\left(\frac{1}{4}(1-\pi)\right)^{x_3}\left(\frac{1}{4}(1-\pi)\right)^{x_4}\left(\frac{\pi}{4}\right)^{x_5}$$

aplicando log tenemos

$$\log f(x|\pi) = K + (x_2+x_5)\log\left(\frac{\pi}{4}\right) + (x_3+x_4)\log\left(\frac{1}{4}(1-\pi)\right)$$

así,

$$E\left[\log f(x|\pi)|y, \pi^p\right] = K + E\left(x_2|y_1, \pi^p\right)\log\left(\frac{\pi}{4}\right) + y_4\log\left(\frac{\pi}{4}\right) + (y_3+y_2)\log\left(\frac{1}{4}(1-\pi)\right)$$

donde K no depende de π. Ahora, utilizando el hecho de que x_2 y y_1 tienen distribución binomial obtenemos

$$E\left(x_2|y_1, \pi\right) = y_1\frac{\frac{\pi^p}{4}}{\frac{1}{2}+\frac{\pi^p}{4}}$$

por lo tanto el primer paso del algoritmo EM es

$$Q(\pi|\pi^{(p)}) = E\left[\log f(x|\pi)|y, \pi\right] = K + y_1\frac{\frac{\pi^p}{4}}{\frac{1}{2}+\frac{\pi^p}{4}}\log\left(\frac{\pi}{4}\right) + y_4\log\left(\frac{\pi}{4}\right) + (y_3+y_2)\log\left(\frac{1}{4}(1-\pi)\right)$$

En el segundo paso maximizamos $Q(\pi|\pi^{(p)})$, entonces derivando con respecto a π obtenemos

$$Q^{'}(\pi|\pi^{(p)}) = 0 + y_1\frac{\frac{\pi^p}{4}}{\frac{1}{2}+\frac{\pi^p}{4}}\frac{1}{\pi} + y_4\frac{1}{\pi} - (y_3+y_2)\frac{1}{1-\pi} = 0$$

Despejando π obtenemos

$$\pi = \pi^{p+1} = \frac{(y_1 + y_4)\pi^p + 2y_4}{(y_1 + y_2 + y_3 + y_4)\pi^p + 2(y_2 + y_3 + y_4)}$$

En la siguiente tabla se presentan los resultados obtenidos del Algoritmo EM con 8 iteraciones y con una tolerancia de 10^{-6}.

p	$\pi^{(p)}$	$\log L$
1	0.5000	-10.3030
2	0.6082	-7.6126
3	0.6243	-7.5498
4	0.6265	-7.5487
5	0.6268	-7.5487
6	0.6268	-7.5487
7	0.6268	-7.5487
8	0.6268	-7.5487

TABLA 2.1: ITERADOS DEL ALGORITMO EM

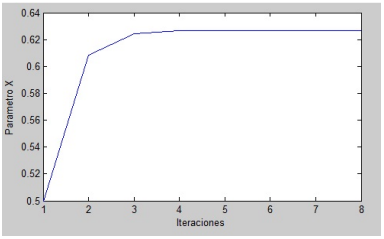

Figura 2.1: Convergencia del parámetro π

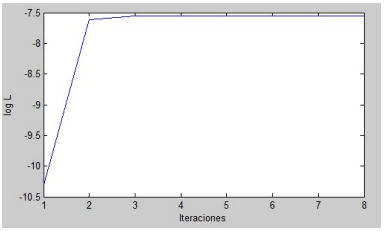

Figura 2.2: Convergencia del log L

Las definiciones precisas de los pasos EM mencionados anteriormente y su interpretación heurística se presentan a continuación:

Definición 2.0.1. Diremos que una función de densidad de probabilidad f pertenece a la familia exponencial si tiene la siguiente forma:

$$f(x|\Phi) = b(x)\exp(\Phi t(x)^T)/a(\Phi) \tag{2.1}$$

Donde Φ denota el vector parámetros $1\times r$, $t(x)$ denota un vector $1\times r$ de estadísticos suficientes de datos completos y T denota la matriz transpuesta.

A continuación se presenta una caracterización simple del algoritmo EM que por lo general se puede aplicar cuando la función de verosimilitud de los datos completos pertenece a la familia exponencial. Supongamos que $\Phi^{(p)}$ indica el valor actual de Φ después de p iteraciones del algoritmo. La siguiente iteración se puede describir en dos etapas, como sigue:
Paso-E: estimación de los estadísticos suficientes de los datos completos $t(x)$ mediante la búsqueda

$$t^{(p)} = E(t(x)|y, \Phi^{(p)}) \tag{2.2}$$

Paso-M: determinar $\Phi^{(p+1)}$ como la solución de las ecuaciones

$$E(t(x)|\Phi) = t^{(p)} \tag{2.3}$$

La ecuación (2.3) es la forma familiar de las ecuaciones de probabilidad para la estimación de los datos de máxima verosimilitud dados de una familia exponencial. Es decir, si tuviéramos que suponer que $t(x)$ representa el número estadísticos suficientes que se calcula a partir de los datos observados x de (2.1), entonces la ecuación (2.3) suele definir el estimador de máxima verosimilitud de Φ. Teniendo en cuenta que para x dado, maximizando $\log f(x|\Phi) = -\log a(\Phi) + \log b(x) + \Phi t(x)^T$ es equivalente a maximizar

$$-\log a(\Phi) + \Phi t(x)^T$$

donde x depende solamente de $t(x)$. Por lo tanto es fácil ver que la ecuación (2.3) define la condición usual para maximizar $-\log a(\Phi) + \Phi t^{(p)T}$ si no se calcula $t^{(p)}$ a partir de (2.2) que representa un valor de $t(x)$ asociado con cualquier x en $\mathcal{X}$.

Una dificultad con el paso-M es que la ecuación (2.3) no siempre tiene solución para Φ en Ω. En tales casos, los valores de Φ que maximizan estan en la frontera de Ω y una definición más general, como la que daremos a continuación debe ser utilizada. Sin embargo, si la ecuación (2.3) puede ser resuelta para Φ en Ω, entonces la solución es única debido a la propiedad de convexidad de logaritmo de la verosomilitud para las familias exponenciales.

Antes de proceder a los casos menos restringidos, explicaremos por qué los pasos E y M del algoritmo EM permiten obtener el valor Φ^* de Φ que maximizan

$$L(\Phi) = \log g(y|\Phi), \tag{2.4}$$

donde $g(y|\Phi)$ se define a partir de $g(y|\Phi) = \int_{\mathcal{X}(y)} f(x|\Phi)dx$ y de la ecuación (2.1). En primer lugar, introducimos la notación de la densidad condicional de x dado y y Φ, concretamente,

$$k(x|y,\Phi) = f(x|\Phi)/g(y|\Phi) \tag{2.5}$$

de modo que (2.4) puede escribirse como

$$L(\Phi) = \log f(x|\Phi) - \log k(x|y,\Phi) \tag{2.6}$$

Para las familias exponenciales, notemos que

$$k(x|y,\Phi) = b(x)\exp(\Phi t(x)^T)/a(\Phi|y) \tag{2.7}$$

donde

$$a(\Phi|y) = \int_{\mathcal{X}(y)} b(x)\exp(\Phi t(x)^T)dx \tag{2.8}$$

Así, vemos que $f(x|\Phi)$ y $k(x|y,\Phi)$ ambos representan a las familias exponenciales con el mismo parámetro natural Φ y el mismo estadístico suficiente $t(x)$, pero que se definen sobre diferentes espacios simples $\mathcal{X}$ y $\mathcal{X}(y)$. Ahora podemos escribir (2.6) de la siguiente forma

$$L(\Phi) = -\log a(\Phi) + \log a(\Phi|y), \tag{2.9}$$

donde la ecuación paralela a (2.8) es

$$a(\Phi) = \int_{\mathcal{X}} b(x)\exp(\Phi t(x)^T)dx. \tag{2.10}$$

derivando (2.10) y (2.8) se obtiene, denotando $t(x)$ por t,

$$\mathbf{D}\log a(\Phi) = (\partial/\partial\Phi)\log a(\Phi) = E(t|\Phi) \tag{2.11}$$

y, similarmente,

$$\mathbf{D}\log a(\Phi|y) = E(t|y,\Phi), \tag{2.12}$$

donde

$$\mathbf{D}L(\Phi) = -E(t|\Phi) + E(t|y,\Phi) \tag{2.13}$$

La fórmula (2.13) es la clave para entender el E paso del algoritmo EM, ya que si el algoritmo converge a Φ^*, de modo que en el límite de $\Phi^{(p)} = \Phi^{(p+1)} = \Phi^*$, entonces la combinación (2.2) y (2.3) conduce a $E(t|\Phi^*) = E(t|y,\Phi^*)$ o $\mathbf{D}L(\Phi) = 0$ en $\Phi = \Phi^*$ lo cual indica que Φ^* es un punto estacionario de L.

2.1. Propiedades Generales

En esta sección daremos algunos resultados básicos del algoritmo EM y además introduciremos una nueva función

$$Q(\Phi`|\Phi) = E(\log f(x|\Phi`)|y,\Phi), \tag{2.14}$$

suponiendo que existen para todos los pares $Q(\Phi`|\Phi)$. En particular, se supone que$f(x|\Phi) > 0$ en casi todo $\mathcal{X}$ para todo $\Phi \in \Omega$. Ahora se define la iteración EM $\Phi^{(p)} \mapsto \Phi^{(p+1)}$ de la siguiente manera:

Paso-E: calcular $Q(\Phi|\Phi^{(p)})$.

Paso-M: elegir $\Phi^{(p+1)}$ a un valor de $\Phi \in \Omega$ que maximiza $Q(\Phi|\Phi^{(p)})$.

La idea heurística es que nos gustaría elegir Φ^* para maximizar $\log f(x|\Phi)$. Dado que no conocemos $\log f(x|\Phi)$, podemos maximizar las expectativas actuales dada la información y y el ajuste actual $\Phi^{(p)}$. En el caso especial de las familias exponenciales

$$Q(\Phi|\Phi^{(p)}) = -\log a(\Phi) + E(b(x)|y,\Phi^{(p)}) + \Phi t^{(p)T},$$

de modo que la maximización de $Q(\Phi|\Phi^{(p)})$ es equivalente a la maximización de $-\log a(\Phi) + \Phi t^{(p)T}$, al igual que en las definiciones más especializadas del paso-M. La familia exponencial del paso-E dado por (2.2) es, en principio, más sencillo que el paso-E general. En el caso general, de $Q(\Phi|\Phi^{(p)})$ debe ser calculado para todo $\Phi \in \Omega$, mientras que para la familia exponencial sólo tenemos que calcular las expectativas de la r-componentes de $t(x)$.
Tambien utilizaremos la siguiente notación

$$H(\Phi^{'}|\Phi) = E(\log K(X|y, \Phi^{'})|y, \Phi)$$

Lema 2.1.1. *Para cualquier par* $Q(\Phi`|\Phi)$ *en* $\Omega \times \Omega$

$$H(\Phi`|\Phi) \leq H(\Phi|\Phi), \tag{2.15}$$

con igualdad si y sólo si $k(x|y, \Phi`) = k(x|y, \Phi)$ *en casi todas partes.*

Demostración:
Probemos que $H(\Phi`|\Phi) - H(\Phi|\Phi) \leq 0$
En efecto

$$\begin{aligned}
H(\Phi`|\Phi) - H(\Phi|\Phi) &= E(\log(K(X|y, \phi^{'}))|y, \phi) - E(\log(K(X|y, \phi))|y, \phi) \\
&= E[(\log(K(X|y, \phi^{'}))|y, \phi) - (\log(K(X|y, \phi))|y, \phi)] \\
&= E\left[(\log \frac{(K(X|y, \phi^{'}))}{(K(X|y, \phi))}|y, \phi)\right] \\
&\leq \log E\left[\frac{(K(X|y, \phi^{'}))}{(K(X|y, \phi))}|y, \phi)\right] \\
&= \log \int \frac{(K(X|y, \phi^{'}))}{(K(X|y, \phi))}.(K(X|y, \phi))d_x \\
&= \log \int (K(X|y, \phi^{'}))d_x \\
&= \log(1) = 0
\end{aligned}$$

Por lo tanto se tiene que $H(\Phi`|\Phi) \leq H(\Phi|\Phi)$.

Para definir un caso particular de un algoritmo iterativo requiere de una lista de sucesiones de valores $\Phi^{(0)} \to \Phi^{(1)} \to \Phi^{(2)}...$ partiendo de $\Phi^{(0)}$. Sin embargo, el término **algoritmo iterativo** se refiere a una norma aplicable a cualquier punto de partida es decir un mapeo de $\Phi \to M(\Phi)$ de Ω a Ω de tal manera que cada paso de $\Phi^{(p)} \to \Phi^{(p+1)}$ se define por

$$\Phi^{(p+1)} = M(\Phi^{(p)}). \tag{2.16}$$

Definición 2.1.1. Un algoritmo iterativo con asignación $M(\Phi)$ es un algoritmo EM generalizado (un algoritmo GEM) si

$$Q(M(\Phi)|\Phi) \geq Q(\Phi|\Phi) \tag{2.17}$$

para todo $\Phi \in \Omega$. Notemos que las definiciones del algoritmo EM dadas anteriormente requieren de lo siguiente

$$Q(M(\Phi)|\Phi) \geq Q(\Phi`|\Phi) \tag{2.18}$$

para todo par $(\Phi`|\Phi)$ en $\Omega \times \Omega$, es decir $\Phi` = M(\Phi)$ maximiza $Q(\Phi`|\Phi)$.

Teorema 2.1.1. *Para cada algoritmo GEM se cumple que*

$$L(M(\Phi)) \geq L(\Phi), \tag{2.19}$$

para todo $\Phi \in \Omega$, donde la igualdad se cumple si y sólo si ambos

$$Q(M(\Phi|\Phi)) = Q(\Phi|\Phi), \quad y \quad k(x|y, M(\Phi)) = k(x|y, \Phi) \tag{2.20}$$

casi en todas partes.

Demostración:
Probemos que $L(M(\Phi)) - L(\Phi) \geq 0$

$$\begin{aligned} L(M(\Phi)) - L(\Phi) &= \{Q(M(\Phi)|\Phi) - H(M(\Phi)|\Phi) - Q(\Phi|\Phi)\} + \{H(\Phi|\Phi)\} \\ &= (Q(M(\Phi)|\Phi) - Q(\Phi|\Phi)) + (H(\Phi|\Phi) - H(M(\Phi)|\Phi)) \geq 0 \end{aligned}$$

ya que por definición del algoritmo GEM las diferencias de las funciones Q es ≥ 0 y por el lema 1 la diferencia de las funciones de H son mayores e iguales a cero con igualdad si y solo si $k(x|y, \Phi) = k(x|y, M(\Phi))$ en casi todas partes.

Corolario 2.1.1. *Supongamos que para cualquier* $\Phi^* \in \Omega$, $L(\Phi^*) \geq L(\Phi)$ *para todo* $\Phi \in \Omega$. *Entonces para cada algoritmo GEM se cumple*

(a) $L(M(\Phi^*)) = L(\Phi^*)$.

(b) $Q(M(\Phi^*)|\Phi^*) = Q(\Phi^*|\Phi^*)$

(c) $k(x|y, M(\Phi^*)) = k(x|y, \Phi^*)$

en casi todas partes.

Corolario 2.1.2. *Si para algún* $\Phi^* \in \Omega$, $L(\Phi^*) \geq L(\Phi)$ *para todo* $\Phi \in \Omega$ *tal que* $\Phi \neq \Phi^*$, *entonces para cada algoritmo GEM se cumple*

$$M(\Phi^*) = \Phi^*. \tag{2.21}$$

Teorema 2.1.2. *Supongamos que* $\theta^{(p)}$ *para* $p = 0, 1, 2, ...$ *son los iterados del algoritmo GEM tal que:*

(a) La suseción $L(\theta^{(p)})$ *es acotada y*

(b) $Q(\theta^{(p+1)}|\theta^{(p)}) - Q(\theta^{(p)}|\theta^{(p)}) \geq \lambda(\theta^{(p+1)} - \theta^{(p)})(\theta^{(p+1)} - \theta^{(p)})'$ *para algún escalar* $\lambda > 0$ *y para todo p Entonces la sucesión* $\theta^{(p)}$ *converge a algún* θ^* *en la clausura de* Ω.

Demostración:

Por hipótesis en (a) y teorema (2.1.1) se tiene que la sucesión es convergente. Así, $\forall \epsilon > 0$, existe $p(\epsilon)$ tal que $allp \geq p(\epsilon)$ y $r \geq 1$ se tiene:

$$\sum_{j=1}^{r} L(\theta^{(p+j)}) - L(\theta^{(p+j-1)}) = L(\theta^{(p+r)}) - L(\theta^{(p)}) < \epsilon \tag{2.22}$$

Por Lema(2.1.1) y la ecuación (2.19) tenemos

$$0 \leq Q(\theta^{(p+j)}|\theta^{(p+j-1)}) - Q(\theta^{(p+j-1)}|\theta^{(p+j-1)}) \leq L(\theta^{(p+j)}) - L(\theta^{(p+j-1)}), \tag{2.23}$$

Para $j \geq 1$, y por (2.22) tenemos

$$\sum_{j=1}^{r} Q(\theta^{(p+j)}|\theta^{(p+j-1)}) - Q(\theta^{(p+j-1)}|\theta^{(p+j-1)}) < \epsilon, \tag{2.24}$$

$\forall p \geq p(\epsilon)$ y $\forall r \geq 1$ donde cada término en la suma es no-negativo.

Aplicando la parte (b) en el teorema para $p, p+1, p+2, ..., p+r-1$ y sumando, tenemos por (2.24)

$$\epsilon > \lambda \sum_{j=1}^{r} (\theta^{(p+j)} - \theta^{(p+j-1)})(\theta^{(p+j)} - \theta^{(p+j-1)})' \tag{2.25}$$

cuando

$$\epsilon > \lambda (\theta^{(p+r)} - \theta^{(p)})(\theta^{(p+r)} - \theta^{(p)})' \tag{2.26}$$

por lo tanto se ha probado la convergencia de $\theta^{(p)}$ para algún θ^*.

2.2. Algoritmo MCEM

En el transcurso de la presente sección, seguiremos usando el modelo de datos incompletos introducidos en el capítulo anterior. Recordemos que el paso-E del algoritmo EM consiste en calcular la esperanza mediante $Q(\Phi|\Phi^{(p)}) = E\left[\log(f(x|\Phi))|y, \Phi^{(p)}\right]$.

Debemos considerar que en algunos casos la evaluación númerica directa de esta esperanza es complicada. En el trabajo propuesto por Wei y Tanner (1990) et al.[7] se utilizan Métodos Monte Carlo para aproximar los insolubles paso-E con una media empírica sobre la base de datos simulados:

$$\widehat{Q}(\Phi|\Phi^{p}) = \frac{1}{m}\sum_{j=1}^{m}\log f(\xi^{j}|\Phi), \tag{2.27}$$

donde $\xi^1, \ldots, \xi^m$ son idénticamente distribuidas con una función de probabilidad $p(x|\Phi)$.

El algoritmo (MCEM) es empleado de la siguiente manera:
Paso-E: reemplazar $Q(\Phi|\Phi^{(p)})$ por $\widehat{Q}(\Phi|\Phi^{p})$ la cual consiste en calcular una aproximación de Monte Carlo.
Paso-M: maximizar la función obtenida en el paso anterior.

Dos consideraciones importantes con respecto a la implementación de este algoritmo de Monte Carlo EM (MCEM) es la convergencia y la especificación de m. En cuanto a la especificación de m, se considera poco eficiente comenzar con un gran va lor de m cuando la aproximación actual de Φ está lejos del valor real. Por el contrario, se recomienda aumentar el valor de m a medida que los iterados se van acercando al verdadero valor del estimador de máxima verosimilitud. Podemos monitorear la convergencia del algoritmo graficando en cada iteración la función de verosimilitud o simplemente examinando la tabla con los valores de $\Phi^{(p)}$ con cada iteración p. A partir de cierto número de iteraciones, la gráfica puede revelar que el proceso se ha estabilizado. En este punto, se puede terminar el algoritmo o continuar con un mayor valor de m lo que podria disminuir la variabilidad del sistema.

Capítulo 3

Algoritmo SAEM

El algoritmo SAEM propuesto por Lavielle y Moulines et al. [4] es una poderosa alternativa para el algoritmo EM cuando el paso-E no se pueda calcular, consiste en reemplazar dicho paso por medio de aproximaciones estocásticas. Antes de presentar los pasos del algoritmo SAEM daremos una breve explicación acerca de las aproximaciones estocásticas.

3.1. Método de Aproximaciones Estocásticas.

Sea $M(x)$ una función dada y α una constante tal que la ecuación

$$M(x) = \alpha \tag{3.1}$$

tiene una única raíz $x = \theta$. Para determinar el valor de θ por aproximaciones sucesivas se comienza con la elección de uno o más valores de $x_1, ..., x_r$ arbitrariamente, y luego, sucesivamente, obtener nuevos valores de x_n. Si

$$\lim_{n\to\infty} x_n = \theta \tag{3.2}$$

independientemente de los valores iniciales arbitrarios $x_1, ..., x_r$, el método es eficaz para la función particular $M(x)$ y el valor α. La velocidad de la convergencia en (3.2) y la facilidad con la que x_n se puede calcular determina la utilidad práctica del método. Consideramos que la generalización estocástica del problema anterior en el que la naturaleza de la función $M(x)$ es desconocido para el experimentador. En su lugar, se supone que a cada valor de x corresponde una variable aleatoria $Y = Y(x)$ con función de distribución $Pr[Y(x) \leq y] = H(y|x)$ tal que

$$M(x) = \int_{-\infty}^{\infty} y dH(y|x) \tag{3.3}$$

donde Y es el valor esperado dado x. Ni la naturaleza exacta de $H(y|x)$ ni la de $M(x)$ es conocida por el experimentador, pero se supone que la ecuación (3.1) tiene una única raíz θ, y se desea estimar θ haciendo observaciones sucesivas Y en los niveles de $x_1, x_2....$ que determine de forma secuencial en acuerdo con algún procedimiento experimental definido.

En lo que sigue vamos a dar un procedimiento especial para la estimación de θ que es consistente con ciertas restricciones sobre la naturaleza de $H(y|x)$.

Teoremas de convergencia

Suponemos que a partir de ahora $H(y|x)$ es que para cada x, una función de distribución en y y que existe una constante positiva C tal que

$$Pr[|Y(x)| \leq C] = \int_{-C}^{C} dH(y|x) = 1 \tag{3.4}$$

para todo x. En particular, para cada x el valor esperado $M(x)$ definido por la ecuación (3.3) existe y es finito. Suponemos, además, que existen constantes finitas α y θ tales que

$$M(x) \leq \alpha \quad para \quad x < 0, \quad M(x) \geq \alpha \quad para \quad x > 0 \tag{3.5}$$

Sea $\{a_n\}$ una suceción fija de constante positiva tal que

$$0 < \sum_{1}^{\infty} a_n^2 = A < \infty \tag{3.6}$$

Definamos una cadena de Markov (no estacionarias) $\{x_n\}$, tomando x_1 una constante arbitraria y definiendo

$$x_{n+1} = x_n + a_n(\alpha - y_n), \tag{3.7}$$

donde $\sum_1^\infty a_n = \infty$ y y_n es una variable aleatoria, tal que

$$Pr[y_n \leq y|x_n] = H(y|x_n). \tag{3.8}$$

Sea

$$b_n = E(x_n - \theta)^2. \tag{3.9}$$

demostraremos que

$$\lim_{n\to\infty} b_n = 0 \tag{3.10}$$

independiente del el valor inicial de x_1. Es bien es conocido, (3.10) implica la convergencia de probabilidad de x_n a θ. De (3.7), se tiene

$$\begin{aligned} b_{n+1} &= E(x_{n+1}-\theta)^2 \\ &= E[E(x_{n+1}-\theta)^2|x_n]] \\ &= E\left[\int_{-\infty}^{\infty}\{(x_n-\theta)-a_n(y-\alpha)\}^2 dH(y|x_n)\right] \\ &= b_n + a_n^2 E\left[\int_{-\infty}^{\infty}(y-\alpha)^2 dH(y|x_n)\right] - 2a_n E[(x_n-\theta)(M(x_n)-\alpha)] \end{aligned}$$

colocando

$$d_n = E[(x_n-\theta)(M(x_n)-\alpha)], \tag{3.11}$$

$$e_n = E\left[\int_{-\infty}^{\infty}(y-\alpha)^2 dH(y|x_n)\right], \tag{3.12}$$

podemos escribir

$$b_{n+1} - b_n = a_n^2 e_n - 2a_n d_n. \tag{3.13}$$

Notemos que de (3.5)

$$d_n \geq 0,$$

mientras que de (3.4)

$$0 \leq e_n \leq [C+|\alpha|]^2 < \infty,$$

Ahora de (3.6), implica que la serie de términos positivos $\sum_1^\infty a_n^2 e_n$ converge. Sumando (3.13) obtenemos

$$b_{n+1} = b_1 + \sum_{j=1}^{n} a_j^2 e_j - 2\sum_{j=1}^{n} a_j d_j. \tag{3.14}$$

desde $b_{n+1} \geq 0$ resulta que

$$\sum_{j=1}^{n} a_j d_j \leq \frac{1}{2}\left[b_1 + \sum_{1}^{\infty} a_n^2 e_n\right] < \infty. \tag{3.15}$$

Por lo tanto los términos positivos de la serie

$$\sum_{1}^{\infty} a_n d_n \tag{3.16}$$

converge. De ello se desprende de (3.14) que

$$\lim_{n\to\infty} b_n = b_1 + \sum_{1}^{\infty} a_n^2 e_n - 2\sum_{1}^{\infty} a_n d_n = b \tag{3.17}$$

existe, con $b \geq 0$. Ahora supongamos que existe una sucesión k_n de constantes no negativas tal que

$$d_n \geq k_n b_n, \qquad \sum_{1}^{\infty} a_n k_n = \infty. \tag{3.18}$$

De la primera parte de (3.18) y la convergencia de (3.16) se deduce que

$$\sum_{1}^{\infty} a_n k_n b_n < \infty. \tag{3.19}$$

De (3.19) y de la segunda parte de (3.18) se deduce que para cualquier $\epsilon > 0$ existe un número infinito de valores n tal que $b_n < \epsilon$. Ya que sabemos que $b = \lim_{n\to\infty} b_n$ existe, se deduce que $b = 0$.

Lema 3.1.1. *Si una suceción $\{k_n\}$ de constantes no negativas existe satisfaciendo (3.18) entonces $b = 0$*

Demostración:

Sea

$$A_n = |x_1 - \theta| + [C + |\alpha|](a_1 + a_2 + ... + a_{n-1}). \tag{3.20}$$

Entonces de (3.4) y (3.8) nos queda que

$$Pr[|x_n - \theta| \leq A_n] = 1. \tag{3.21}$$

Ahora el conjunto

$$\bar{k_n} = inf\left[b_1 + \sum_1^\infty a_n^2 e_n\right] \quad para \quad 0 < |x-\theta| \leq A_n. \tag{3.22}$$

De (3.5) se tiene que $\bar{k_n} \geq 0$. Además denotando $P_n(x)$ la distribución probabilidad de x_n, teniendo

$$d_n = \int_{|x-\theta|\leq A_n} (x-\theta)(M(x)-\alpha)dP_n(x) \qquad \geq \int_{|x-\theta|\leq A_n} \bar{k_n}|x-\theta|^2 dP_n(x) = \bar{k_n} b_n. \tag{3.23}$$

En particular la sucesión $\{\bar{k_n}\}$ definida por (3.22) satisface la primera parte de (3.18). Para establecer la segunda parte de (3.18) supondremos que:

$$\bar{k_n} \geq \frac{K}{A_n} \tag{3.24}$$

para cualquier constante $K > 0$ y un n suficientemente grande, y

$$\sum_{n=2}^\infty \frac{a_n}{(a_1 + ... + a_{n-1})} = \infty \tag{3.25}$$

de la ecuación (3.25) se tiene

$$\sum_1^\infty a_n = \alpha. \tag{3.26}$$

así para un n suficientemente grande

$$2[C + |a|](a_1 + ... + a_{n-1}) \geq A_n \tag{3.27}$$

Implica que de (3.25) para un n suficientemente grande

$$a_n \bar{k_n} \geq a_n \frac{K}{A_n} \geq \frac{a_n K}{2[C + |a|](a_1 + ... + a_{n-1})}, \tag{3.28}$$

Así, de la primera parte de (3.18), (3.28) y de (3.25) ha sido probado el lema.

Lema 3.1.2. *Si* $\bar{k_n} \geq \frac{K}{A_n}$ *y* $\sum_{n=2}^\infty \frac{a_n}{(a_1+...+a_{n-1})} = \infty$ *entonces se tiene que b=0.*

Demostración:

Por las hipotesis de (3.6) y (3.25) sobre $\{a_n\}$ son satisfechas para la sucesión $a_n = \frac{1}{n}$, de aqui

$$\sum_1^\infty \frac{1}{n^2} = \frac{\pi^2}{6}, \quad \sum_{n=2}^\infty \left[\frac{1}{n\left(1 + \frac{1}{2} + ... + \frac{1}{n-1}\right)}\right] = \infty$$

En general, cualquier sucesión $\{a_n\}$ tal que existen 2 constantes positivas c' y c'' para lo cual

$$\frac{c'}{n} \leq a_n \leq \frac{c''}{n} \tag{3.29}$$

pudiera satisfacer (3.6) y (3.25). Denominaremos a cualquier sucesión $\{a_n\}$ que satis facen (3.6), (3.25) y (3.24), una sucesión de tipo $\frac{1}{n}$. Si $\{a_n\}$ es una sucesión de tipo $\frac{1}{n}$ es facil encontrar la funciones $M(x)$ que satisfacen (3.5) y (3.24). Supogamos por ejemplo, que $M(x)$ satisface la siguiente condición de (3.24): para algún $\delta \geq 0$.

$$M(x) \leq \alpha - \delta \quad para \quad x < \theta, \quad M(x) \geq \alpha + \delta \quad para \quad x > \theta.$$

Entonces para $0 < |x - \theta| \leq A_n$ se tiene que

$$\frac{M(x) - \alpha}{x - \theta} \geq \frac{\delta}{A_n}, \tag{3.30}$$

así

$$\bar{K_n} \geq \frac{\delta}{A_n}, \tag{3.31}$$

lo cual (3.24) con $K = \delta$.

Teorema 3.1.1. *Si* $\{a_n\}$ *es de tipo* $\frac{1}{n}$*, si* $Pr[|Y(x)| \leq C] = \int_{-C}^{C} dH(y|x) = 1$ *se cumple, y si* $M(x)$ *satisface* $M(x) \leq \alpha - \delta \quad para \quad x < \theta, \quad M(x) \geq \alpha + \delta \quad para \quad x > \theta$*. entonces* $b = 0$*.*

Demostración:

Un caso interesante ocurre cuando $M(x)$ satisface las siguientes condiciones:

$$M(x) \quad es \quad no \quad decreciente \tag{3.32}$$

$$M(\theta) = \alpha \tag{3.33}$$

$$M'(\theta) > 0 \tag{3.34}$$

Vamos a demostrar que (3.24) tambien tiene este caso. De (3.33) se tiene

$$M(x) - \alpha = (x - \theta)[M'(\theta) + \epsilon(x - \theta)], \tag{3.35}$$

donde $\epsilon(t)$ es una función tal que

$$\lim_{t\to 0} \epsilon(t) = 0, \tag{3.36}$$

De aqui existe una constante positiva $\delta > 0$ tal que

$$\epsilon(t) \geq -\frac{1}{2}M'(\theta) \quad para \quad t \leq \delta, \tag{3.37}$$

de modo que

$$\frac{M(x)-\alpha}{x-\theta} \geq \frac{1}{2}M'(\theta) > 0 \quad para \quad |x-\theta| \leq \delta, \tag{3.38}$$

aquí, para $\theta + \delta \leq x \leq \theta + A_n$, como $M(x)$ es no decreciente,

$$\frac{M(x)-\alpha}{x-\theta} \geq \frac{M(\theta+\delta)-\alpha}{A_n} \geq \frac{\delta M'(\theta)}{2A_n}. \tag{3.39}$$

mientras que para $\theta - A_n \leq x \leq \theta - \delta$

$$\frac{M(x)-\alpha}{x-\theta} = \frac{\alpha - M(x)}{\theta - x} \geq \frac{\alpha - M(\theta-\delta)}{A_n} \geq \frac{\delta M'(\theta)}{2A_n}. \tag{3.40}$$

Así, podemos suponer sin perdida de generalidad que $\frac{\delta}{A_n} \leq 1$

$$\frac{M(x)-\alpha}{x-\theta} \geq \frac{\delta M'(\theta)}{2A_n} \quad para \quad 0 < |x-\theta| \leq A_n. \tag{3.41}$$

de modo que (3.24) se mantiene con $K = \frac{\delta M'(\theta)}{2} > 0$

3.2. Pasos del Algoritmo SAEM.

La idea del Algoritmo SAEM es similar al del Algoritmo MCEM, la idea básica es dividir el paso-E en dos pasos: un paso de aproximación y un paso de simulación. El paso de simulación consiste en generar realizaciones de los datos no observados y el paso de integración consiste en aproximar la integral que aparece en el paso-E del Algoritmo EM por un promedio dado por:

$$\widehat{Q}(\Phi|\Phi^p) = \frac{1}{m}\sum_{j=1}^{m} \log f(\xi^j|\Phi). \tag{3.42}$$

La ecuación (3.42) es utilizada en el Algoritmo SAEM en un paso de aproximaciones estocásticas dado por

$$\tilde{Q}(\Phi|\Phi^{p+1}) = \tilde{Q}(\Phi|\Phi^p) + \gamma_p((\widehat{Q}(\Phi|\Phi^p) - \tilde{Q}(\Phi|\Phi^p)) donde$$

γ_p es una sucesión de tamaño de paso que satisface $\gamma_p \in [0,1]$, $\sum_{p=1}^{\infty}\gamma_p = \infty$ y $\sum_{p=1}^{\infty}\gamma_p^2 < \infty$, por ejemplo $\gamma_p = \frac{1}{p}$ satisface dichas condiciones.

En resumen los pasos del Algoritmo SAEM puede ser escrito de la siguiente manera:

(a) **Simulación:** Generar muestras $z^{(1)}, \ldots, z^{(m(p))}$ de $p(z|\Phi^{(p)}, y)$.

(b) **Aproximación Estocástica:** Actualizar $\tilde{Q}(\Phi, \Phi^{(p)})$

$$\tilde{Q}(\Phi, \Phi^{(p)}) = \tilde{Q}(\Phi, \Phi^{(p-1)}) + \gamma_p \left(\frac{1}{m(p)} \sum_{j=1}^{m(p)} \log(p(\Phi|z^{(j)}, y)) - \tilde{Q}(\Phi, \Phi^{(p-1)}) \right)$$

(c) **Maximización:** Maximizar $\tilde{Q}(\Phi, \Phi^{(p)})$ en Ω. Esto es, hallar $\Phi^{(p+1)} \in \Omega$ tal que

$$\tilde{Q}(\Phi, \Phi^{(p+1)}) \geq \tilde{Q}(\Phi, \Phi^{(p)}) \qquad \forall \Phi \in \Omega.$$

El Algoritmo SAEM se puede escribir en términos de los estadísticos suficientes cuando la función de verosimilitud de los datos completos pertenece a la familias exponenciales:

Simulación: usar ξ^{p-1}, para generar ξ^p de probabilidad $p(x|\Phi)$.

Aproximación estocástica: actualizar t_{p-1} de acuerdo a $t_p = t_{p-1} + \gamma_p(E(t(x)|\Phi) - t_{p-1})$.

Maximizacion: actualizar Φ_p de acuerdo a $\Phi_{p+1} = \hat{\Phi}(t_p)$.

Para ilustrar el algoritmo SAEM presentaremos el siguiente ejemplo

Ejemplo 4. 197 individuos estan distribuidos multinomialmente en cuatro categorías, los datos observados son:

$$y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

supongamos que las probabilidades de que una observación este en cada una de las categorías es respectivamente

$$(p_1, p_2, p_3, p_4) = \left(\frac{1}{2} + \frac{\pi}{4}, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{\pi}{4}\right)$$

para algún $\pi \in [0, 1]$ desconocido y el cual se desea estimar. Así

$$g(y|\pi) = \frac{(y_1 + y_2 + y_3 + y_4)!}{y_1! y_2! y_3! y_4!} \left(\frac{1}{2} + \frac{\pi}{4}\right)^{y_1} \left(\frac{1}{4}(1-\pi)\right)^{y_2} \left(\frac{1}{4}(1-\pi)\right)^{y_3} \left(\frac{\pi}{4}\right)^{y_4}$$

$\left(\frac{1}{2} + \frac{\pi}{4}, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{\pi}{4}\right)$, la idea es dividir la primera de las cuatro categorías en dos categorías. Así, los datos completos son $x = (x_1, x_2, x_3, x_4, x_5)$, donde $y_1 = x_1 + x_2$, $y_2 = x_3$, $y_3 = x_4$, $y_4 = x_5$ y la verosimilitud de los datos completos es

$$f(x|\pi) = \frac{(x_1 + x_2 + x_3 + x_4 + x_5)!}{x_1! x_2! x_3! x_4! x_5!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\pi}{4}\right)^{x_2} \left(\frac{1}{4}(1-\pi)\right)^{x_3} \left(\frac{1}{4}(1-\pi)\right)^{x_4} \left(\frac{\pi}{4}\right)^{x_5}$$

aplicando log tenemos

$$\log f(x|\pi) = K + (x_2 + x_5) \log\left(\frac{\pi}{4}\right) + (x_3 + x_4) \log\left(\frac{1}{4}(1-\pi)\right).$$

Dado π_0, los pasos del algoritmo SAEM en la iteración $p + 1$ son los siguientes:

Paso de simulación

Tomar $Z_p^{(1)}, ..., Z_p^{(M)}$ de la distribución binomial de parámetros $N = 125$ y $P = \pi^{(p)}/(\pi^{(p)} + 2)$.

Aproximación Estocástica

$$\hat{Z}_{p+1} = \hat{Z}_p + \gamma_k(\bar{Z}_p - \hat{Z}_p)$$

Paso-E

$$Q_{p+1}(\pi|\pi^{(p)}) = \frac{1}{m}\sum_{j=1}^{m} log(P(\pi|Z^{(j)}, y))$$

que también puede escribir como

$$Q_{p+1}(\pi|\pi^{(p)}) = (\hat{Z}_{p+1} + x_5)log(\pi) + (x_3 + x_4)log(1-\pi))$$

donde el estadístico suficiente es

$$\bar{Z}_p = \frac{1}{m}\sum_{j=1}^{m} Z_p^{(j)}$$

Maximización

$$\hat{\pi} = \frac{\hat{Z}_{p+1} + x_5}{x_3 + x_4 + \hat{Z}_{p+1} + x_5}$$

En la siguiente tabla se presentan los resultados obtenidos del Algoritmo SAEM con 8 iteraciones y con una tolerancia de 10^{-6}.

p	$\pi^{(p)}$	$\log L$
1	0.5000	-7.5699
2	0.6162	-7.5649
3	0.6175	-7.5714
4	0.6158	-7.5600
5	0.6190	-7.5549
6	0.6211	-7.5538
7	0.6216	-7.5512
8	0.6232	-7.5502

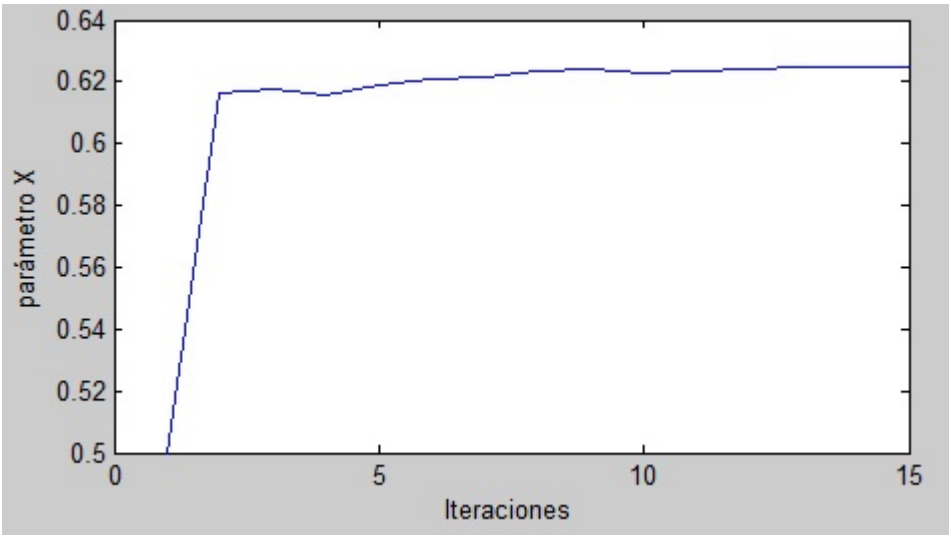

Figura 3.1: Convergencia del parámetro π

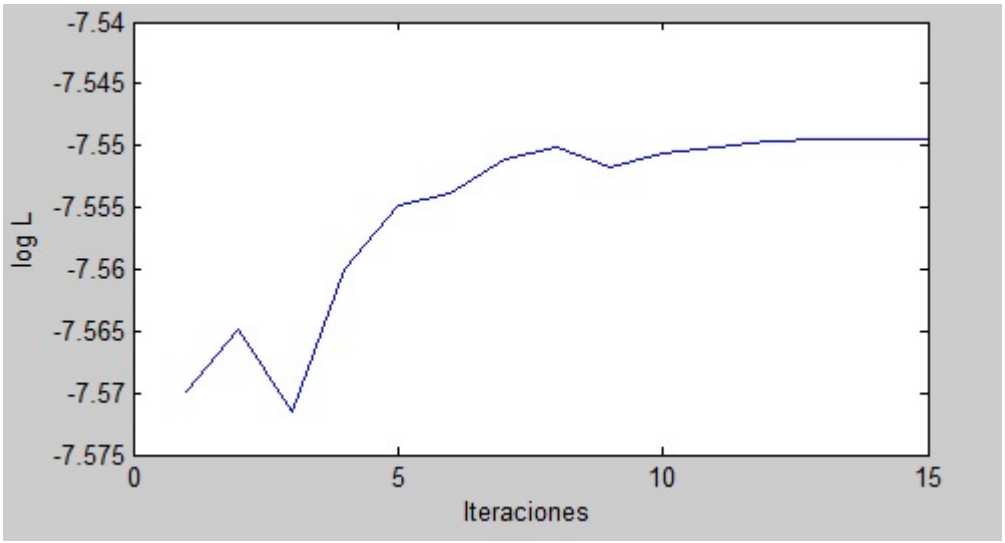

Figura 3.2: Convergencia del log L

3.3. Ventajas y Desventajas del Algoritmo EM y SAEM

- Los Algoritmos son numericamente estables.
- Bajo condiciones generales, los algoritmos alcanzan convergencia global. Es decir, iniciando de un punto arbitrario $\pi^{(0)}$ en el espacio de parámetros, la convergencia es casi siempre a un máximo local, salvo una muy mala elección de $\pi^{(0)}$ o alguna patología local en el logaritmo de la función de verosimilitud.
- Los algoritmos son faciles de implementar.
- Los algoritmos son generalmente faciles de programar, ya que no hace evaluaciones de la función de verosimilitud ni de sus derivadas.
- Los algoritmos requieren poco espacio de almacenamiento.
- El algoritmo SAEM ayuda a reducir el número de simulaciones en comparación con el algoritmo MCEM.

Apéndice

CODIGOS EN MATLAB

Algoritmo EM

```
function [X,L]=em(x0,y,niter,tol)
% [X,L]=em(0.5,[125 18 20 34],15,10^(-6))
done=0;
k=1;
while done==0 && k<niter
    x=((y(1)+y(4))*x0+2*y(4))/((y(1)+y(2)+y(3)+y(4))...
    ...*x0+2*(y(2)+y(3)+y(4)));
    s=x-x0;
    if abs(s)<tol
        done=1;
    end
    X(k)=x0;
    L(k)=sum(log(2:sum(y)))-sum(log(2:y(1)))-sum(log(2:y(2)))...
    ...-sum(log(2:y(3)))-sum(log(2:y(4)))+y(1)*log(0.5+x0/4)...
    ...+(y(2)+y(3))*log(0.25*(1-x0))+y(4)*log(0.25*x0);
    x0=x;
    k=k+1;
end
    X(k)=x0;
    L(k)=sum(log(2:sum(y)))-sum(log(2:y(1)))-sum(log(2:y(2)))...
    -sum(log(2:y(3)))-sum(log(2:y(4)))+y(1)*log(0.5+x0/4)...
    ...+(y(2)+y(3))*log(0.25*(1-x0))+y(4)*log(0.25*x0);
```

Algoritmo SAEM

```
function [X,L]=saem(x0,y,niter,tol)
% [X,L]=mcem(0.5,[125 18 20 34],15,10^(-6))
done=0;
k=1;
g=1;
st=0;
while done==0 && k<niter
    Z=binornd(125,x0/(x0+2),10,1);
    st=st+g*(mean(Z)-st);
    x=(st+y(4))/(st+y(2)+y(3)+y(4));
    s=x-x0;
    if abs(s)<tol
        done=1;
    end
    X(k)=x0;
    x0=x;
     L(k)=sum(log(2:sum(y)))-sum(log(2:y(1)))-sum(log(2:y(2)))...
     -sum(log(2:y(3)))-sum(log(2:y(4)))+y(1)*log(0.5+x0/4)...
      ...+(y(2)+y(3))*log(0.25*(1-x0))+y(4)*log(0.25*x0);
    k=k+1;
    g=1/k;
end
    X(k)=x0;
     L(k)=sum(log(2:sum(y)))-sum(log(2:y(1)))-sum(log(2:y(2)))...
     -sum(log(2:y(3)))-sum(log(2:y(4)))+y(1)*log(0.5+x0/4)+(y(2)+y(3))...
      ...*log(0.25*(1-x0))+y(4)*log(0.25*x0);
```

Referencias

[1] Cappé, O. Moulines, E. y Rydén, T. (2007). *Inference in Hidden Markov Models.* Chapter 11.

[2] Demptster, A.P, Laird, N.M y Rubin, D.B. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, Vol. 39, pp. 1-38.

[3] Hartley, H.O (1958). *Maximum Likelihood Estimation from Incomplete Data.* Biometries. 14, pp. 174-194.

[4] Kuhn, E. Lavielle, M. (2003). *Coupling a Stochastic Approximation Versión of EM with a MCMC procedure.* ESAIM probab, Statist. 8, pp. 115-131.

[5] Pestman, W. (1998). *Mathematical Statistics. An Introduction.* Walter de Gruyter.

[6] Robbins, H. Monro, S. (1951). *A Stochastic Aproximation Method.* The Annals of Mathematical Statistics, Vol.22, pp. 400-407.

[7] Wei, C.G, Tanner, M.A. (1990). *A Monte Carlo Implementation of the EM Algorithm and the Door Man's Data Argumentation Algorithms.* Journal of the American Statistical Association, Vol.85, pp. 699-704.