UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

EL USO DEL ALGORITMO E.M. PARA INFERIR SECUENCIAS DE A.D.N

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A :

MARIA ISABEL OROZCO TONCHEZ

DIRECTOR DE TESIS: DRA. ELIANE REGINA RODRIGUEZ CALONI

2001

MAT. MARGARITA ELVIRA CHÁVEZ CANO
Jefa de la División de Estudios Profesionales
P r e s e n t e

Comunicamos a usted que hemos revisado el trabajo de Tesis:

EL USO DEL ALGORITMO E.M. PARA INFERIR SECUENCIAS DE A.D.N.

realizado por María Isabel Orozco Tónchez

Con número de cuenta 9550326-2 . pasante de la carrera de Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

A t e n t a m e n t e

Director de tesis Propietario	Dra. Eliane Regina Rodrigues Caloni	
Propietario	Dr. Mogens Bladt Petersen	
Propietario	M. en A. P. María del Pilar Alonso Reyes	
Suplente	M. en C. José Antonio Flores Díaz	
Suplente	M. en C. Ruth Fuentes García	

Consejo Departamental de Matemáticas.
M. en C. Jose Antonio Flores Díaz

Agradecimientos.

A mi asesora de Tesis, la Dra. Eliane por su paciencia, sabiduría y generosidad en la enseñanza y dirección de este trabajo.

A mis sinodales: Dr. Mogens Bladt, M. en A.P. Pilar Alonso, M. en C. Jose Antonio Flores y M. en C. Ruth Fuentes por sus comentarios y ayuda en la revisión de esta tesis.

A mis profesores de carrera y muy especialmente al Act. Javier Fernandez García.

Dedicado

A la persona que con su constancia me enseñó a ser responsable, con su fortaleza a ser perseverante, con su humildad a ser sincera, con su grandeza a ser juiciosa y con su amor a aprender a vivir después de su muerte. A ti padre, que sacrificaste lo innimaginable por vernos concluir nuestras metas, te agradezco tu entrega y compañía durante todo este tiempo.

A mi tío José, sin el cual este sueño no hubiese podido ser realidad, a mi madre Isabel, mis hermanos Carlos y Víctor por su eterno apoyo, comprensión, ayuda y aceptación de este esfuerzo conjunto que nos mantuvo unidos a pesar de la distancia.

A todas las personas que de alguna forma ayudaron a que este esfuerzo se viera concluido:

Mis amigos: Jack, Laurita, Nadia, Rosa, Lizette, Yuri, Virginia, Ericka, Candi, Anto, Gise, Juan Carlos, Ivan y Jorge por su cariño y consejos.

A Leo, Era, Judy, Luisito, Chucho, Julisa por el ejemplo de trabajo conjunto.

A Paty, Miriam, Ere, Lupita, Mati, Ale, Paola, Nancy, Nora. Mis tíos: Roscelia, Metodio, Alfonso, Kokis, Lupe, Bertha y Carlos. Mis primos Juan, Ross, Freddy y Anel. A todos ellos por su compañía y ayuda incondicional cuando más la necesitaba.

A Roberto por su amor incondicional, Camilo y Angel por su pasión por la vida.

A Luis Lara por su acertado comentario que hizo que cambiara mi perspectiva ante la vida.

El uso del algoritmo E.M. para inferir secuencias de A.D.N.

María Isabel Orozco Tónchez

Directora de tesis: Dra. Eliane Regina Rodriguez Caloni

Índice General

Introducción

"El ser es eterno, porque existen leyes para conservar los tesoros de la vida, de los cuales el universo extrae su belleza."
Goethe.

Mentiría al decir que sólo existió una única razón para realizar una tesis acerca de este tema en particular. No obstante, la razón principal podría ser mi constante curiosidad por el mundo y la vida que se encuentra en él, así como los grandes avances en materia biológica que se pueden obtener con la utilización de métodos de inferencia estadística. Alguna vez me llegué a preguntar, ¿De dónde provengo?, y las escasas respuestas que obtenía no eran muy claras. ¡Qué maravilla!, llegaba a pensar, sería el conocer nuestros orígenes, y no sólo hablo de la especie humana, sino de la vida en general.

Dentro de cada una de nuestras células se encuentran los cromosomas, en lo que se puede llamar nuestra información genética, conformados por moléculas del A.D.N. (Acido Desoxirribonucléico). Cada uno de los nucleótidos de que se conforma esta molécula de A.D.N. el cual está escrito en forma de una secuencia de letras (o secuencia de nucleótidos), son llamadas bases nitrogenadas, las cuales pueden ser consideradas como caracteres del "texto" de la información genética. Existen cuatro nucleótidos diferentes que contienen un residuo de desoxirribosa (un azúcar), un fosfato y una base de pirimidina o bien de purina. Las bases de la pirimidina son la Timina (T) y la Citocina (C), y las bases de la purina son la adenina (A) y la guanina (G).

Desde los años ochenta se propuso para la comunidad científica en su totalidad el reto de conocer lo que se llama el mapa genético del ser humano. A este proyecto se le dio el nombre de "Proyecto GENOMA". Sin duda alguna, el proyecto genoma es un trabajo conjunto e internacional que tiene como principales propósitos el de poder descifrar de forma confiable todos y cada uno de los genes presentes en los 46 cromosomas del ser humano (22 cromosomas homólogos y 2 cromosomas del sexo), y no solo decodificar la información que esta contenida, sino también conocer su funcionamiento.

A mi parecer, éste es uno de los retos más importantes en la biología, el poder descifrar la procedencia de todo organismo vivo y descubrir de qué forma todos y cada uno de nosotros procede de un mismo patrón que la evolución ha ido transformando hasta convertirnos en lo que ahora somos. Se puede considerar que gracias al descubrimiento efectivo de las secuencias de A.D.N. se infieren aspectos básicos de la evolución en diversas especies actuales, incluyendo la nuestra, ya que gracias a la información que poseen los cromosomas a nivel molecular se pueden elaborar árboles genealógicos y esquemas que nos dirijan de forma precisa al descubrimiento de nuestros orígenes como especie humana o bien de el funcionamiento de la evolución y descubrir su base fundamental de mutaciones.

Con base a estos métodos genéticos se pueden llegar a inferir mutaciones que dieron lugar a nuevos organismos. El llegar a conocer lo antes posible el mapa genético del ser humano puede acarrearnos un sin fin de posibilidades de aplicaciones a la medicina que antes se nos estaban vedadas. Y no sólo esto, gracias a los métodos genéticos de decodificación de A.D.N. se pueden dar pasos agigantados en diversas terapias génicas para tratamientos médicos; la creación de antibióticos, hormonas y sustancias biológicas activas en el ramo de la industria biotecnológica.

Asimismo, se puede hablar acerca de la tecnología del A.D.N. recombinante, del cual explicaré su proceso más adelante, por su gran ayuda para la realización del clonaje de genes. Y dentro

de esta rama se abre un sin fin de posibilidades para la ayuda, tanto al ser humano como a las especies animales. Utilizando esta tecnología se pueden obtener grandes poblaciones de bacterias para su uso experimental; se pueden producir compuestos con fines antibacterianos, antivirales y antitumorales. Y eso que no se ha hablado acerca de su inmensa utilidad en la creación de vacunas (aún en período de evaluación), la obtención de aditivos en la industria alimentaria y el empleo de "mapas genéticos" que permiten detectar la localización precisa de los genes que están causando las enfermedades para su pronto diagnóstico.

Gran parte de la ingeniería genética, la tecnología del A.D.N. recombinante y los métodos genéticos en general, se han dedicado ampliamente a la resolución de todos estos problemas, cabe tan sólo mencionar que este proceso es en extremo lento, ya que las pruebas genéticas que se deben realizar en los laboratorios se repiten una gran cantidad de veces para estar seguros de su veracidad.

El trabajo del genetista no es nada fácil, de hecho para llegar a inferir las secuencias de A.D.N. se deben realizar muchos experimentos para poder estar seguros de que la secuencia es la correcta, y cada uno de éstos debe ser repetido un número determinado de veces. Es realmente aquí en donde entra el trabajo de las herramientas probabilísticas y estadísticas, ya que por medio de ellas se puede validar teóricamente el trabajo de laboratorio y de esta manera evitar pérdida de tiempo en el desarrollo de esta labor maratónica.

Existen varios métodos a seguir y difícil sería decir cuál es el que conviene más para su aplicación. Sin embargo, algunos métodos podrían combinarse para dar lugar a un menor error de predicción (errores que de ser posible no se deben cometer), minimizarlos lo más posible, ya que de estos métodos depende el poder avanzar de forma más rápida y eficiente en el descubrimiento de "mapas genéticos" de la especie humana, de organismos multicelulares y de toda un área extensa que a veces pareciera estancada por la falta de rapidéz en sus experimentos.

Siempre ha existido una profunda curiosidad dirigida hacia el conocimiento de las ciencias biológicas en general, de hecho gran parte de los ejemplos aplicados a las matemáticas de una u otra forma se dirigen hacia probabilidades biológicas, los cuales constituyen un constante reto para las ciencias exactas.

El propósito fundamental del presente trabajo es el de utilizar métodos probabilísticos y estadísticos que ayuden a inferir adecuadamente las secuencias de A.D.N. Dentro del estudio genético en laboratorios existen métodos para realizar este trabajo; no obstante, llega a suceder que si bien ayudan a dar una idea de la secuencia que se busca muchas veces la información es perdida por cuestiones meramente experimentales, para lo cual deben introducirse métodos analíticos que ayuden a inferir esta información, considerando para esto el menor error posible para poder atacar los problemas biológicos con mayor precisión. Para este efecto se utilizarán métodos como el de la verosimilitud, el de aproximación numérica, algoritmo E.M. (Esperanza-Maximización), el cual es un método iterativo y el de cadenas de Markov, que cae dentro del área de procesos estocásticos; estos métodos serán utilizados posteriormente en forma conjunta como un solo método para la inferencia de secuencias de A.D.N.

De tal modo que globalmente esta tesis se presenta de la siguiente forma:

- En el **Capítulo 1** se introducirán términos generales en genética, así como una explicación del mecanismo utilizado en trabajo de laboratorio para decodificar secuencias de A.D.N., que servirán para el trabajo matemático que se realiza más adelante.

- El **Capítulo 2** trata algunos métodos de inferencia estadística y de probabilidad con el propósito fundamental de encontrar lo que se conoce como parámetros de funciones, los cuales en este caso representarían una descripción del comportamiento de la secuencia de A.D.N., si bien estos métodos son útiles cuando se tiene información completa no son eficientes cuando, como en

nuestro caso, hace falta información de la secuencia de una molécula de A.D.N.

- Para superar este problema en el **Capítulo 3** se describe la forma en que un algoritmo iterativo llamado Esperanza-Maximización trabaja para poder inferir información faltante a partir de información observada así como una introducción a su área teórica. Este método iterativo se utilizará en el modelo matemático para obtener los estimadores máximo verosímiles del modelo que describe la obtención de las secuencias de A.D.N. y que se encuentra en el siguiente capítulo.

- En el **Capítulo 4** se construye un modelo matemático para poder obtener las secuencias que llamamos verdaderas de las secuencias observadas en el laboratorio por los genetistas, utilizando para este efecto el material visto anteriormente, los cuales de una u otra forma son parte primordial para que el modelo tenga validéz teórica.

- Y por último, en el **Apéndice** se encuentran los apartados que son mencionados a lo largo de este trabajo y que en su momento no era necesario incluirlos en el cuerpo de la tesis, como son, en términos biológicos: la morfología de los cromosomas, las enfermedades producidas por mutaciones, los procesos de la meiosis y mitosis, las abreviaturas de los aminoácidos, etc.; mientras que en términos matemáticos se encuentran: la desigualdad de Jensen, la distribución multinomial y el cálculo de estimadores del modelo mencionado en el **Capítulo 5** entre otros temas útiles para el seguimiento de esta tesis.

Capítulo 1.

Introducción a la genética.

"Todo lo que existe en el universo es fruto del azar y la necesidad."
Demócrito.

1.1 Introducción

En este capítulo se presentan algunos conceptos básicos en genética, asimismo se plantea el problema que se estudia con las técnicas de los **CAPÍTULOS 2** y **3**.

Se empieza por definir qué es una secuencia de A.D.N., su relación con los genes y cromosomas para de esta manera justificar la importancia de su composición molecular para futuros estudios.

1.2 ¿Qué es el A.D.N.?

Es importante tener claro que el A.D.N. (ácido desoxirribonucléico) constituye un depósito de información genética, y que es el principal componente genético de la célula, y el que transmite la información codificada de una célula a otra y por ende de un organismo a otro. En la célula eucariótica[1] el A.D.N. no se halla libre, sino formando complejos en algo parecido a una madeja enredada conocida como cromatina[2]; cada molécula de

[1] Una célula eucariótica es aquella que tiene una clara separación entre su núcleo y el citoplasma, así como la existencia de organelos con funciones específicas y división celular mitótica, es decir, división de células que no sean gametos.

[2] La cromatina es el complejo de A.D.N.: proteína que forma el cromosoma. Actualmente se considera a la cromatina como la expresión interfásica de los cromosomas antes que se organicen y contraigan para la siguiente división celular.

A.D.N. es representada comúnmente por un modelo de doble-hélice propuesto por Watson y Crick (1953), está compuesta en su totalidad por combinaciones de pares de únicamente cuatro tipos de bases diferentes llamadas nucleótidos. Estas bases se dividen en purinas o pirimidinas. Las bases nitrogenadas de la purina son la adenina (A) y la guanina (G), mientras que las bases de la pirimidina son la timina (T) y la citosina (C).

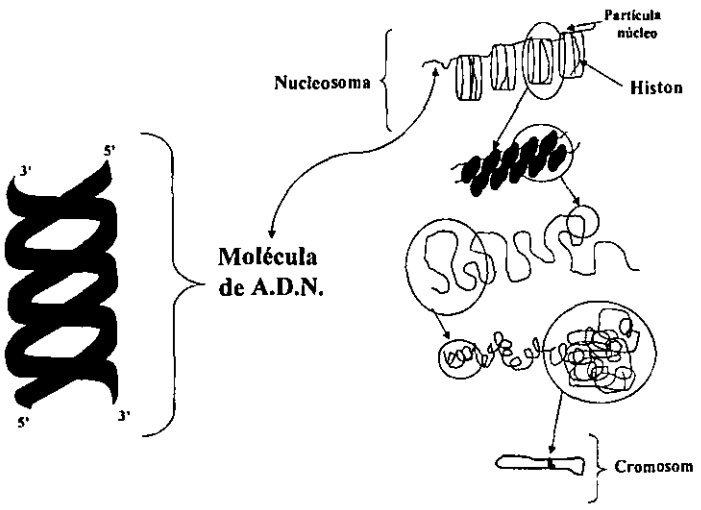


Fig.1.1 Visión general de una molécula de A.D.N. y su relación con los cromosomas.

Si observa el modelo de la doble hélice (Fig. 1.2) se puede dar cuenta de que ésta consta de "peldaños" y "lados" como en una escalera; los "peldaños" están hechos de las bases y los "lados" de la escalera consisten de residuos de desoxirribosa[3] unidos por medio de fosfatos, estos "lados" están unidos por lazos de hidrógeno entre las bases (pirimidinas y purinas).

Es importante mencionar que la adenina siempre está aparejada con la timina, y la guanina con la citosina; este hecho es muy importante de recordar cuando se hable, más adelante, del número de combinaciones a los cuales se reduce el espacio de

[3] La desoxirribosa es el azúcar que se encuentra en la estructura del A.D.N.

probabilidad a considerar. Por convención, las partes finales de las "tiras" de la hélice doble que forman la molécula se designan como 5' y 3', estas "tiras" son antiparalelas, es decir, una cadena de nucleótidos van de 5' a 3' y la otra en la dirección 3'a 5'. Solamente cadenas antiparalelas forman hélices dobles estables (ver Watson y Crick (1953)).

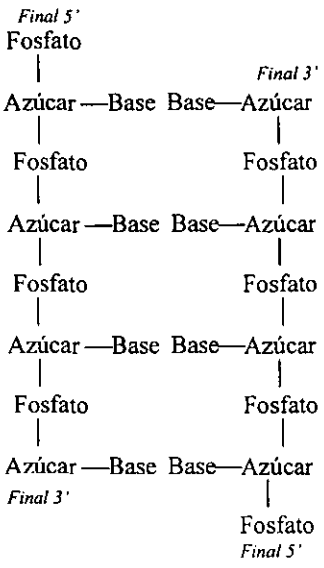


Fig. 1.2 Composición de una molécula de A.D.N.

En las células eucarióticas, el A.D.N. se halla fundamentalmente en el núcleo, aunque también existen pequeñas cantidades en el citoplasma, dentro de las mitocondrias (en el caso de células animales) o bien en los cloroplastos (en el caso de vegetales). Este A.D.N., se conoce como A.D.N. mitocondrial, y se representa como un modelo circular (ver Friedman, Dill, et.al. (1992, p.8)), distinto al modelo de doble-hélice de Watson y Crick, ya mencionado anteriormente.

La información genética que se analizará para efectos de esta tesis es el A.D.N. nuclear de las células. Los genes son parte de los cromosomas y el A.D.N. es el material del cual están formados los genes. Los cromosomas se encuentran formados por una sola

molécula de A.D.N. organizada en diferente orden y que pareciera una madeja enredada[4]. El genoma humano contiene 22 cromosomas autosomales y un par de cromosomas del sexo que consisten en 3.6 x 10^4 pares de bases.

El gen es una unidad hereditaria que contiene la información necesaria para poder elaborar una cadena polipeptídica[5] concreta, estos polipéptidos son unidos en un orden específico para producir proteínas que darán origen a organismos vivos.

Los genes tienen una organización especial, no en toda la banda formada por la molécula de A.D.N. existe información que codifica las proteínas. Un gen está constituido por:

(1) **Exones**, son la porción funcional de las secuencias de A.D.N. en los genes, que codifican para las proteínas, es en donde se puede decir que se encuentra la información a duplicar.

(2) **Intrones**, son las secuencias de A.D.N. que no codifican, no tienen una función específica, el número y tamaño de éstos depende del gen que se trate. Algunas veces, la relación evolutiva de genes puede ser inferida a partir de la organización de estos intrones y de su localización, de esto se hablará más adelante.

(3) Los intrones y exones se intercalan en el gene. La **frontera** entre ellos se encuentra claramente definida. Un intrón siempre empieza inmediatamente después de la aparición de la pareja de bases GT e inmediatamente después del término de un intrón se encuentra la pareja de bases AG. Inmediatamente antes del primer exón se encuentra el tripleto ATG, el cual da la referencia de inicio para la síntesis de la proteína y la referencia para el final de la síntesis es indicada por uno de los tripletos TGA, TAA o TAG (ver Fig.1.3).

El final de la izquierda del gen se indica por 5' e indica la posición más cercana al inicio de donde empieza la síntesis de

[4] Para mayor información acerca de los cromosomas y su morfología referirse al ***Apéndice B.1.***
[5] Una cadena polipeptídica es una cadena formada por proteínas o porciones de una proteína que consisten de dos o más moléculas de aminoácidos, los cuales son compuestos de tres nucleótidos y son la base química de las proteínas. Existen 20 tipos de aminoácidos.

proteína, mientras que 3' se escribe como el final de la derecha de la molécula e indica la posición más cercana al final de la síntesis de proteína.

Fig. 1.3. Organización de un gen humano.

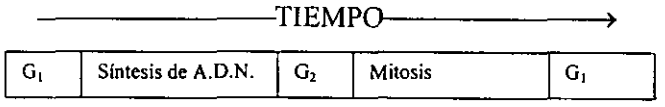
1.3 La síntesis del A.D.N. y la división celular.

La reproducción del A.D.N. es el proceso que se conoce como síntesis, de este proceso depende la información transmitida de una célula a otra; no obstante, es más importante aún conocer de forma clara cómo se lleva a cabo cada uno de los procesos por separado que llevan a la célula a reproducirse para tener una visión más general de lo que sucede a nivel molecular.

Todas las células de nuestro organismo se encuentran en un constante cambio, o lo que se conoce como ciclo celular, este ciclo comprende principalmente dos períodos generales: la interfase y la división celular, las células pasan la mayor parte de su "vida" en interfase, durante el cual aumentan su tamaño y su complemento cromosómico.

Durante la división celular, la cual se llama mitosis (para células somáticas) o meiosis (para células de los gametos:

espermatozoides u óvulos)[6], ocurre no sólo la duplicación de una célula madre en dos células hijas idénticas, sino que también el contenido genético se replica o sintetiza. La división celular es sólo la fase final de un cambio básico que tuvo lugar a nivel molecular durante la interfase. La síntesis de A.D.N. tiene lugar totalmente durante una parte limitada de la interfase, denominada período S o sintético, que a su vez es precedido y seguido por dos espacios o períodos de la interfase (G_1 y G_2) en los que no hay síntesis de A.D.N. G_2 es el intervalo entre el final de la síntesis de A.D.N. y el comienzo de la mitosis (meiosis). Durante G_2 la célula contiene el doble de la cantidad de A.D.N. (cromosomas) que se encuentra en la célula original. Después de la mitosis cada célula hija tendrá el número original de la célula madre, mientras que en la meiosis se producen por cada célula madre cuatro células hijas, las cuales tendrán la cuarta parte de la célula madre.



El proceso de síntesis de una molécula de A.D.N. en forma esquemática es :

$$\text{A.D.N.} \xrightarrow{\text{Transcripción}} \text{A.R.N.}^{7} \xrightarrow{\text{Traducción}} \text{Proteína.}$$

La molécula de A.D.N. se separa en dos tiras (ver Fig.1.4), dando lugar a dos nuevas moléculas de A.D.N. que se forman siguiendo el patrón que les corresponde, es decir, en donde quede sola una A (adenina) se apareará con una T (timina), o bien una G (guanina) con una C (citosina) y viceversa.

[6] Para mayor información acerca del proceso de la mitósis y la meiosis referirse al ***Apéndice B.2.***

[7] A.R.N. es ácido ribonucléico auxiliar en la síntesis de proteína, se localiza tanto en el núcleo, donde es sintetizado, como en el citoplasma, donde tiene lugar la síntesis de proteínas.

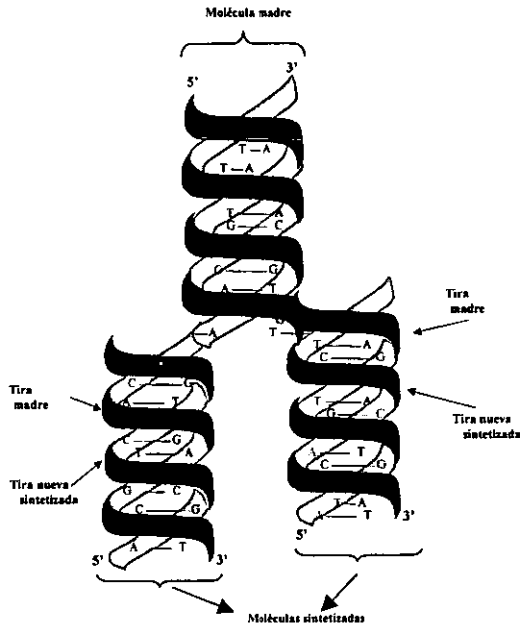


Fig. 1.4 Síntesis de una molécula de A.D.N.

El A.D.N. actúa como un patrón claro para el A.R.N., el cual transporta la información dada por el A.D.N. del núcleo al citoplasma en forma de aminoácidos, los cuales más tarde darán lugar a las proteínas. Este proceso se conoce como transcripción, al cual le sigue el proceso de traducción, en donde los ribosomas[8] (que se encuentran en el citoplasma de la célula de forma dispersa) trasladan la información codificada por el A.R.N. en una cadena de tripletos que forman lo que se conoce como una cadena polipeptídica (se darán detalles en seguida), dependiendo de la

[8] Los ribosomas son organelos celulares y el sitio de la síntesis de la proteína durante la traducción.

información que lleve impresa la cadena de A.D.N. original sobre la cual está sucediendo todo este proceso. Durante el proceso de traducción es en donde se pueden llegar a dar anomalías genéticas, acerca de las cuales se hablará más adelante.

Una cadena polipeptídica está formada por una serie de tripletos de bases, cada tripleto forma un aminoácido. Para el A.R.N. mensajero (el que transporta la información de la cadena de A.D.N. del núcleo al citoplasma) la timina (T) se convierte en uracilo (U), y de este modo en la traducción la adenina (A) se une al uracilo (U) en el A.R.N. mensajero (ver Figura 1.5). Las primeras dos bases en el tripleto son las que especifican el aminoácido que les corresponde, ya que la tercera base se agrega sin tener implicación alguna, es decir, el aminoácido prolina es sintetizado para las combinaciones CCU,CCC,CCA o bien CCG (ver Tabla 1.1).

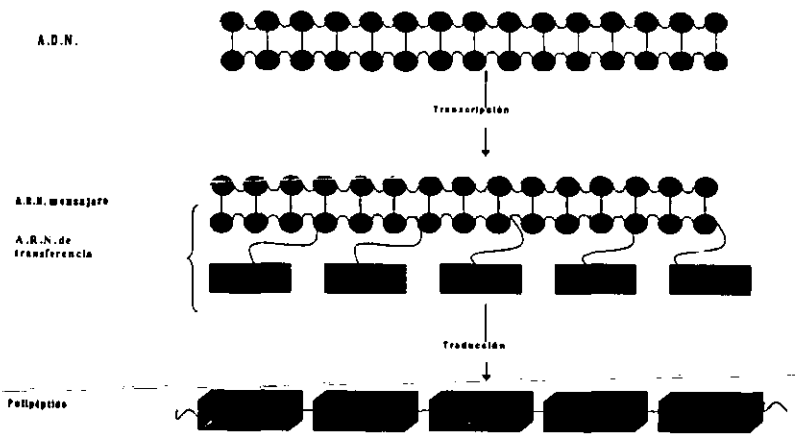


Fig. 1.5 Proceso de transcripción y traducción en la síntesis de A.D.N.

Primera posición	Segunda posición				Tercera posición
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	PARAR	PARAR	A
	Leu	Ser	PARAR	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Tabla 1.1 Formación de los aminoácidos [9].

1.4 Anomalías.

En el código genético pueden existir diversos tipos de anomalías, los cuales pueden ser a nivel cromosomal o bien de genes. Dentro de las anomalías cromosómicas se encuentran las numéricas y las estructurales[10] (la adición o pérdida de cromosomas completos o bien un reordenamiento del material genético, respectivamente). Por otro lado, a nivel de bases que forman el A.D.N. se puede hablar fundamentalmente de cuatro tipos de anomalías: las deleciones, las inserciones, las duplicaciones y las sustituciones. Estas anomalías se conocen como mutaciones. Las mutaciones no solamente tienen lugar en el núcleo de la célula. Como se vio anteriormente, existe también el A.D.N. mitocondrial,

[9] Los nombres de cada uno de los aminoácidos se encuentran en el ***Apéndice B.3.***
[10] Para mayor detalle acerca de las anomalías cromosómicas referirse al ***Apéndice B.4.***

el cual también puede dar lugar a mutaciones.

Las mutaciones pueden ser divididas en mutaciones puntuales y reordenaciones, según si afectan a un par de bases o a una región más amplia, respectivamente. La frecuencia de las mutaciones puede ser aumentada por la acción de radiaciones ionizantes, la luz ultravioleta y agentes químicos (ver Pellón (1996)). Las mutaciones son un cambio en la secuencia de A.D.N. que influye en su posterior traducción por el A.R.N. mensajero, de tal forma que se ve afectada la función y expresión del producto protéico. Este tipo de mutaciones (a nivel gen) son las que interesan particularmente para efectos de la tesis, ya que gracias a estos cambios se puede conocer algo más acerca de la influencia del genoma en las característica fenotípicas[11] de cualquier organismo.

La misma estructura del A.D.N. sólo permite cuatro tipos de alteraciones a nivel del gen:

(1) **La sustitución de un nucleótido por otro**: se conoce como transición, si se sustituye una base de purina (Adenina o Guanina) por otra de purina, o una pirimidina (Timina o Citosina) por otra de pirimidina y transversiones si se sustituye una pirimidina por una purina o viceversa.

(2) **El borrado de uno o más nucleótidos**. El borrado o deleción puede ser detectado por medio de la tecnología del A.D.N. recombinante (del cual se hablará más adelante en mayor detalle) por medio de la falta de una sección del fragmento de A.D.N. Las deleciones genéticas no son causas comúnes de mutación en el genoma humano a excepción de unas pocas enfermedades. Ejemplos de estas enfermedades son la α-Thalassemia[12], que es causada por la deleción del grupo genético de la α-globina, la deficiencia en la hormona del crecimiento y la hipercolesterolemia familiar (HF), resultante por el defecto en la baja densidad del gen receptor lipoprotéico. Las enfermedades causadas por deleciones o

[11] Las características fenotípicas de un individuo son las características observables o bien la expresión funcional de un gen. En general, el fenotipo es la expresión del genotipo (constitución genética específica de un organismo), por ejemplo, color de los ojos, síntomas de una enfermedad, entre otros.

[12] Para mayor detalle acerca de las enfermedades que se enunciarán en adelante referirse al ***Apéndice B.5.***

borrado de bases son en general pocas en relación con las enfermedades que involucran anomalías cromosómicas (Síndrome de Down o Fibrosis quística).

(3) **La inserción de uno o más nucleótidos**. Son casos raros de mutación en el genoma humano; de cualquier modo, la transposición de A.D.N. es común en el genoma aunque usualmente no involucra a la parte que codifica proteína. Cuando la transposición ocasionalmente se inserta en un gen, la expresión genética defectuosa se lleva a cabo. Un ejemplo clínico de este tipo de anomalía es la Hemofilia, la cual es transmitida únicamente por las madres a hijos hombres y se caracteriza por alteraciones en la coagulación de la sangre (ver Friedman, Dill, et.al (1992)).

(4) **La duplicación** de secuencias de A.D.N. son comunes en la evolución y pueden ser causados por desaparejamiento entre secuencias homólogas de A.D.N. que se encuentren muy cercanas, con duplicación de material genético que es contenido dentro del gen. El mecanismo de duplicación es similar al visto con α-Thalassemia, las duplicaciones alteran el mecanismo de lectura del A.R.N. mensajero. Las enfermedades HF y la Distrofia Muscular de Duchenne son ejemplos de desórdenes causados por duplicaciones.

Gracias a la organización de los genes se puede llegar a notar que existen algunos espacios no codificadores de proteína (intrones), los cuales pueden ser más largos que la cantidad de espacios codificadores de proteínas (exones) dependiendo del gen que se trate, de hecho algunos investigadores piensan que es aquí, en los intrones, en donde radica el principio de la evolución[13].

Una vez que se ha dado lugar a una mutación, el código genético se replica dando por resultado que esta mutación continúe heredada a su progenie; este es el modo mas común de cómo los padres pueden transmitir a sus hijos diversas mutaciones genéticas que no han sido estudiadas aún.

La incidencia de enfermedades genéticas causadas por mutaciones es de 79/1000 nacimientos después de que la madre

[13] Para mayor conocimiento acerca de éste principio consultar: Chen(1976), Moya (1990), Freeman (1998), Oliva (1996).

cumple los 25 años de edad (ver Friedman, Dill, et.al (1992)). Existen diversas formas en las que se puede dar lugar a mutaciones, de hecho las mutaciones puntuales se pueden dar lugar de forma espontánea durante la replicación del A.D.N. y el consecuente error de los mecanismos que editan la información para corregirlo. Asimismo ocurren mutaciones por diversos factores externos como la radiación o el humo del cigarrillo, y la incidencia de enfermedad se ve aumentada si se posee determinado gen que se vea activado por estas causas.

Algo interesante de mencionar es el hecho de la existencia de determinados grupos étnicos que tienen mayor incidencia en ciertas enfermedades, un ejemplo de ésto son algunas poblaciones al sureste de Asia que poseen la anomalía cromosómica para desarrollar la enfermedad α-Thalassemia, mientras que por otro lado algunas personas que habitan el África poseen el gen-específico sickle-globina que los protege contra la malaria, ésto es lo que se conoce como una ventaja selectiva (ver Friedman, Dill, et.al. (1992)).

1.5 Clonaje de secuencias.

El proyecto genoma es un proyecto que involucra tanto el clonaje, como el secuenciamiento y armado de las moléculas de A.D.N., para que sobre la marcha se vaya descubriendo el funcionamiento de cada una de las secuencias que se pudo decodificar. A partir de ahora este trabajo se enfocará por completo en lo que se refiere a los pasos a seguir dentro de este proyecto a gran escala.

El clonaje es apenas el primer nivel del problema y su resultado forma el pilar del proyecto mencionado. Si bien el término clonación ha sido muy utilizado para difundir a la muy conocida oveja "Dolly", el clonaje se había visto desde hace tiempo y desde otra perspectiva. Clonar significa reproducir de forma asexual una línea de células, organismos o determinados segmentos de A.D.N. para que se obtengan réplicas genéticamente iguales al original.

En general, el proceso de clonación se puede dividir en cuatro pasos básicos, los cuales son:

1. **La fragmentación del A.D.N.** de un organismo sin afectar genéticamente al gen de interés que se quiera clonar. Existen diversas formas para la realización de los cortes de un gen, la mayor parte de ellos son de manera aleatoria, aunque desde 1970 con el descubrimiento de las enzimas de restricción se pudieron hacer algunos avances, ya que este tipo de enzimas llegan a reconocer determinadas secuencias y cortan el gen en donde encuentren la secuencia que se desea (ver Suzuki y Knudtson, (1989)). El objetivo principal de esta parte del clonaje es el de dividir la molécula de A.D.N. poco a poco hasta quedarse con fragmentos más pequeños y manejables.

2. **La combinación del fragmento de A.D.N. con un vector genético**. Una vez que se tienen estos fragmentos se deben escoger aquellos que se deseen clonar, los cuales se conocen como genes aislados, se une este A.D.N. fragmentado que contiene el gen aislado con el A.D.N. de un vector genético u hospedero (comúnmente es el A.D.N. de un virus o una bacteria), de esto se obtiene como resultado una molécula nueva de A.D.N., la cual es llamada recombinante.

3. **Cultivar las moléculas recombinadas en un hospedador**. Esta molécula de A.D.N. recombinante unida al vector genético es transportada posteriormente al tubo de ensaye en el cual se recombinó a un hospedador, que es un medio de cultivo, donde se replicará.

4. **Selección de réplicas**. Se seleccionan las réplicas del hospedador, que ahora se conocen como clones, que hayan producido la molécula de A.D.N. recombinante que contenía el gen específico que se deseaba. Algunas veces las moléculas se pueden expresar con colores diferentes al que presenta en general el hospedador, indicando así que tiene la posibilidad de producir la proteína que codifica para el gen específico que fue insertado en el hospedero.

Este proceso un tanto largo y tedioso también se conoce como la técnica del A.D.N. recombinante y es el que se ha estado utilizando como primer paso para el descubrimiento de mapas genéticos de varios organismos multicelulares. En definitiva, el hecho de realizar este proceso una y otra vez hasta que no queden dudas de las localizaciones del gen específico lleva a pensar que el trabajo del genetista en su laboratorio es bastante forzoso y que debían existir formas más rápidas y eficientes para la realización de este proceso.

Es importante mencionar que el objetivo principal de este método es el de no afectar en modo alguno a los genes cuando se parten en pequeños fragmentos, aunque de todas formas durante el proceso se puede dar lugar a mutaciones que no pueden evitarse, como puede llegar a suceder durante la traducción, cuando se sintetiza la proteína, o bien durante la multiplicación del A.D.N. cuando se encuentra en el hospedador (ver Fig.1.6).

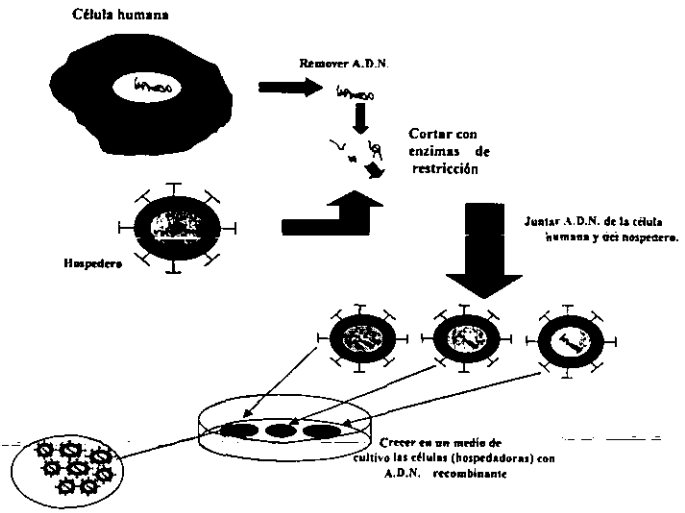


Fig.1.6 Proceso de la recombinación de A.D.N.

1.6 Secuenciamiento de A.D.N.

El secuenciamiento, también llamado decodificación, es el

proceso mediante el cual se intenta conocer el orden de las bases de la secuencia de A.D.N. reproducida en el paso anterior (clonaje) para que de este modo se pueda obtener el código del gen en específico.

Una vez que un clon en particular ha sido seleccionado para la secuenciación, es necesario partirlo en subclones más pequeños. Existen principalmente dos estrategias para generar subclones para el secuenciamiento: la *estrategia aleatoria*, mucho más rápida, es en donde la molécula de A.D.N. es rota en fragmentos de forma aleatoria, de este modo la secuencia tiene puntos de partida y orientaciones aleatorias; la *estrategia dirigida* es mucho más lenta pero se puede decir que más segura, en esta estrategia se utiliza la información de una secuencia anterior para generar el siguiente subclón de tal forma que coincidan en un 25% ó 50% con respecto al subclón anterior (ver Churchill(1995)).

Se pueden determinar cientos de bases de los subclones de una simple reacción de secuenciamiento. Existen dos métodos de decodificación de A.D.N.: el método de Maxam y Gilbert (método químico de decodificación) y el método de Sanger (método enzimático de decodificación), (ver Maxam y Gilbert (1977) y Sanger (1977)). Inicialmente, las moléculas de A.D.N. eran secuenciadas con el método de Maxam y Gilbert, pero debido al desarrollo que ha tenido el método enzimático, ahora lo convierten en la mejor opción para la mayor parte de los problemas de secuenciamiento. La aplicación del método químico se utiliza aún para el estudio de las interacciones entre el A.D.N. y las proteínas, así que habría que darle su importancia. Después del secuenciamiento de los genes se lleva a cabo lo que se puede llamar una inferencia empírica de la secuencia decodificada. No obstante, dado que se pueden originar fallas durante este proceso la secuencia empírica representará la secuencia verdadera, pero puede tener un cierto número de bases donde no coinciden (ver Zachary (1998)).

En términos generales el proceso de secuenciamiento se desarrolla de la siguiente forma:

1. Una vez que se tienen clonadas las moléculas de A.D.N. y escogido el gen específico que se quiere determinar como recombinante, se introducen estos fragmentos en cuatro reacciones distintas con propiedades tintóreas por separado de tal forma que estas reacciones producen que el final de cada fragmento de A.D.N. termine en una base determinada: la adenina (A), la timina (T), la guanina (G) y la citosina (C). Cada una de las bases tendrá un color diferente y de esta forma se podrá reconocer a una base de otra. Así cada reacción tiene copias parciales del gen que termina en cada base en un color distinto (ver Fig. 1.7).

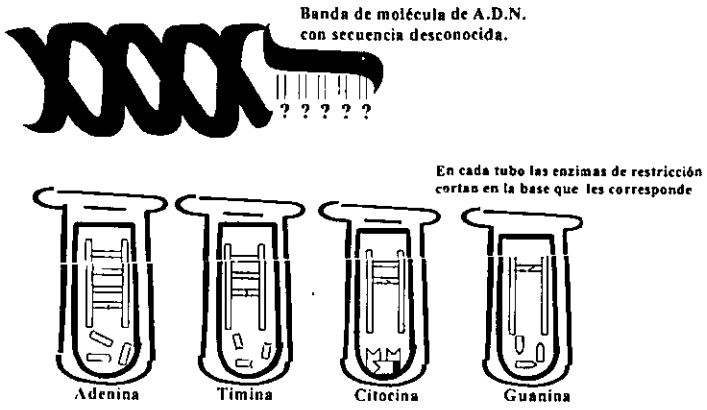


Fig.1.7 Fragmentación del gen finalizado en cada base.

2. Cada una de estas reacciones, en donde se encuentran las moléculas de A.D.N., con sus finales en cada base, se introduce en un gel llamado de electroforesis, posteriormente se les aplica una corriente eléctrica para producir una reacción.

3. Las moléculas del A.D.N. se moverán a través del gel y dependiendo de su tamaño se aglutinarán en algunos lugares, para esto se debe tener una cantidad considerable de fragmentos del mismo tamaño que terminen en una determinada base, en cierta posición, para que de esta forma se puedan detectar las huellas de la secuencia de A.D.N. Como los cortes en la secuencia de A.D.N. son aleatorios, es posible que para algunos tamaños no se tenga un número suficiente de fragmentos (entonces se tiene una falla, que se llamará una mutación, donde una base es borrada). Los fragmentos más pequeños se moverán más rápido, y de este modo se pueden observar las posiciones en las que se encuentra cada base por separado de una sola secuencia de A.D.N. cortada en fragmentos pequeños (ver Fig. 1.8).

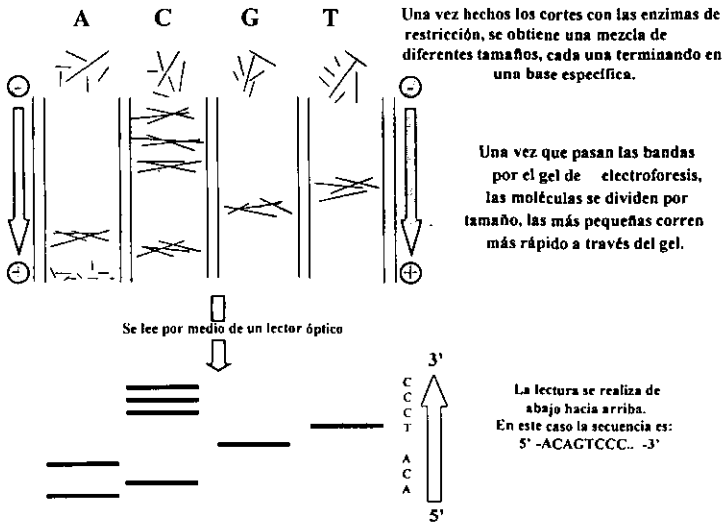


Fig. 1.8 Secuenciamiento del gen utilizando el gel de electroforesis.

4. Una vez hecho esto, la secuencia que queda dibujada en el gel de electroforesis se pasa por un lector óptico para que capte la

mayor cantidad de bases en el gel, precisamente por esta razón es que se requiere la mayor cantidad de fragmentos del mismo tamaño que terminen en la misma base, ya que la precisión del lector será aún mejor. Puede suceder que una o más bases de la secuencia no sean detectadas, posiblemente por falta de fluorescencia de las bases, y por lo tanto no son incluidas en el secuenciamiento. En este caso, se dice que ocurrió una mutación (deleción), o bien, por errores de lectura una base puede ser sustituida por otra (sustitución) o una base ser insertada (inserción) (ver Fig. 1.9). Por ejemplo, para el caso de mayor grosor de línea se supondría que en esa posición lo más seguro es que exista la base indicada, no obstante, podría pasar que esta misma línea impida verificar que en el lugar de una sola base finalizada haya dos (inserción) en donde existe un espacio vacío (deleción) (caso (3) y (4)). Por otro lado, las líneas punteadas significarían muy poca precisión, ya que no se estaría completamente seguro de la existencia de una base o de otra (caso (1) y (2)) y se incurriría nuevamente en una inserción.

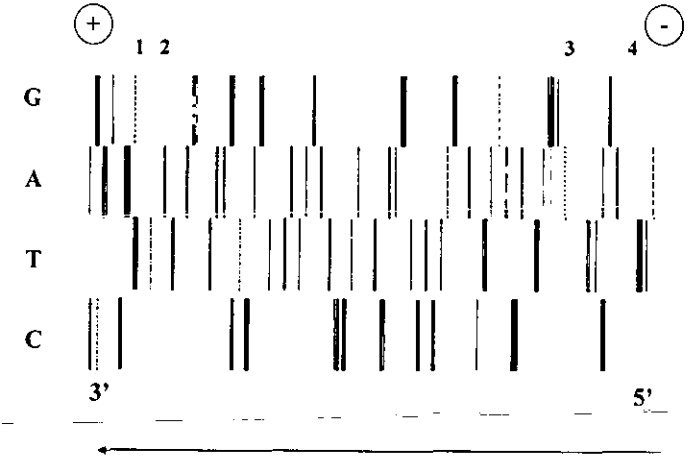


Fig.1.9 Forma de una secuencia de A.D.N. una vez observada por lector óptico

Es un gran triunfo el saber que después de todo este tipo de procesos se haya llegado a algo, ya que se dio a conocer el mapa genético de la bacteria Esclerichia Coli. En definitiva este es un paso gigantesco, ya que esto nos motiva a pensar que con el proyecto genoma se pueda llegar a hablar dentro de unos cuantos años (de hecho se tiene como meta el año 2005) del mapa genético del humano. Conocer con precisión la localización y decodificación de cada uno de los genes de nuestro organismo y de este modo tener aplicaciones diversas a la salud y a la tecnología en general.

Dentro de los problemas que enfrentan estos procesos de secuenciamiento se encuentra el hecho de que el trabajo en los laboratorios llega a ser tardado y a veces poco confiable, ya que muchas veces no se tiene seguridad completa de que el gen específico que se utilizó para clonar y después decodificar sea el mismo y no tenga mutaciones, lo cual conlleva a la realización del proceso repetidas veces hasta asegurar el menor margen de error posible.

Una vez que se tienen secuenciadas por medio de un sensor óptico cada una de las bases se debe proceder al armado de las secuencias.

1.7 Armado de las secuencias.

Se puede decir que este es el proceso crucial dentro de este análisis; ya que en el armado de los fragmentos del gen será donde se utilice la inferencia para deducir qué base está en cada una de las posiciones en la secuencia de A.D.N. Dentro del armado de las secuencias es en donde se pueden ver aplicaciones claras a la estadística, probabilidad y procesos estocásticos, ya que ahora se empezará a hablar de variables aleatorias que se quieren ajustar, de ser posible, a una distribución de probabilidad.

Una vez que se tienen decodificadas las bases de la secuencia de A.D.N. se procede a armarla. Como se mencionó anteriormente, los cortes de los clones se realizan por lo general en forma aleatoria, ya que es un método más rápido. No obstante, por esta razón no se tiene seguridad alguna del lugar que ocupa dentro de la molécula de A.D.N. cada fragmento del gen. Lo que se hace en la práctica es

acomodar los fragmentos dependiendo de su incidencia, es decir, acomodar de tal modo los fragmentos de A.D.N. (como se repiten la mayor parte de los experimentos existe evidencia para poder compararlos) para que se puedan observar en un solo plano claramente (ver Fig.1.10), y de este modo poder inferir la base más probable en cada posición. Una vez que se tiene la mayor parte de la secuencia de A.D.N. es común que se tengan pequeños vacíos entre fragmentos. Lo que comúnmente se hace es una combinación entre las dos estrategias mencionadas anteriormente (aleatoria y dirigida), es decir, se corta en forma aleatoria y después de armar los pedacitos se usa la estrategia dirigida para llenar los huecos que puedan aparecer durante el armado de la estrategia aleatoria. La figura denota a la secuencia verdadera en la parte superior de la tabla y a la observada por el lector óptico en la parte inferior y donde N denota que no fue posible saber qué base se encuentra en esa posición y – denota un espacio vacío.

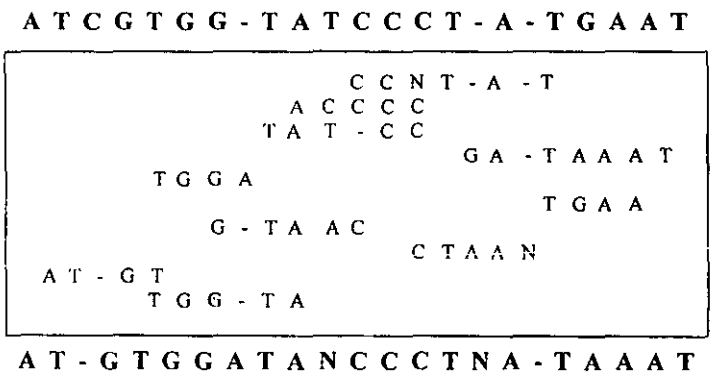


Fig. 1.10 Armado de la secuencia de A.D.N.

Dentro de el armado de las piezas de un rompecabezas hay que ser en extremo precisos para que el resultado que se obtenga sea el esperado, esto es lo mismo que sucede con las moléculas de A.D.N., hay que ser en extremo precavidos y cometer el menor error posible para que éstas no tengan repercusión posterior.

En el próximo capítulo se hablará de forma más extensa de los métodos probabilísticos y estadísticos por medio de los cuales se puede tener un acercarmiento más veraz del proceso de clonaje,

secuenciamiento y posterior armado de moléculas de A.D.N., minimizando lo más posible el error intrínseco en estos procesos. De ahora en adelante se considerará que la secuencia ha sido armada y se trabajará con la secuencia final obtenida en todo este proceso.

1.8 Referencias de las Figuras de este capítulo.

Fig. 1.1 Visión general de una molécula de A.D.N. y su relación con los cromosomas.
Friedman, Dill, et. al. (1992) Fig.1-3

Fig. 1.2 Composición de una molécula de A.D.N.
Alphey (1997) Fig.1.2

Fig. 1.3 Organización de un gen humano.
Friedman, Dill, et. al. (1992) Fig.1-3

Fig. 1.4 Síntesis de una molécula de A.D.N.
Friedman, Dill, et. al. (1992) Fig.1-2

Fig. 1.5 Proceso de transcripción y traducción en la síntesis de A.D.N.
Monod (1997)

Fig. 1.6 Proceso de la recombinación de A.D.N.
Suzuki y Knudtson (1989) Fig.5.3

Fig. 1.7 Fragmentación del gen finalizado en cada base.
Suzuki y Knudtson (1989) Fig.5.4

Fig. 1.8 Secuenciamiento del gen utilizando el gel de electroforesis.
Alphey (1997) Fig.6.1

Fig. 1.10 Armado de la secuencia de A.D.N.
Churchill. (1989) Fig 7

Capítulo 2.

Resultados Preliminares.

"Así como la luz se manifiesta a sí misma y a la oscuridad, la verdad es la medida de sí misma y del error. "
Spinoza, Ético, P.III Prop.43

2.1 Introducción.

Como se vio en el capítulo anterior, gran parte de los proyectos realizados para encontrar las bases fundamentales de la aparición de la especie humana tienen que ver en gran medida con los métodos de secuenciamiento de A.D.N. Cabe recordar que los métodos de Sanger y de Maxam & Gilbert son los más utilizados en este momento para este efecto. No obstante, se puede pensar que este estudio no puede ser fructífero si no hay forma de medir la confiabilidad de la información que poco a poco se van obteniendo con los métodos mencionados.

El estudio de las secuencias de A.D.N. es amplio en términos estadísticos y se puede considerar que se ha venido avanzando desde hace mucho tiempo en este tópico. Existen métodos novedosos para el estudio del secuenciamiento hecho a las cadenas de A.D.N., la mayor parte de los cuales se basan en la utilización de algoritmos probabilísticos que incluyen probabilidades condicionales para el cálculo de las probabilidades de aparición de determinados segmentos de A.D.N. en cada secuencia que se quiera predecir.

A continuación se dará una explicación acerca de los métodos que se utilizan para analizar el armado de las secuencias de A.D.N. para posteriormente englobarlos en un único método utilizando para esto una aplicación en la práctica.

Este capítulo se organiza de la siguiente forma: en la sección 2.2 se presentan algunos resultados elementales de probabilidad y estadística para hacer más accesible el lenguaje manejado a través de este trabajo. La sección 2.3 presenta una introducción al tema de los procesos estocásticos; que se utilizarán más adelante en la construcción del modelo en el **Capítulo 4**, posteriormente se trata el tema de la función de verosimilitud en la sección 2.4 como una introducción a la sección 2.5, que trata al respecto del método de máxima verosimilitud, un proceso muy útil para encontrar estimadores. Y por último, la sección 2.6 es dedicada a un ligero análisis de métodos de aproximación al máximo de la función de verosimilitud utilizados comúnmente cuando se tienen observaciones completas.

2.2 Resultados Elementales.

La estadística es una ciencia experimental. Trata de ordenar, estudiar y predecir el comportamiento de ciertas características de los elementos de un conjunto que existe en la naturaleza. Para este estudio se utilizan principalmente fundamentos provenientes de la teoría de probabilidad.

Dentro del área estadística se encuentra la inferencia estadística: el que a partir de ciertos datos de la población que se quiere estudiar se trate de averiguar la función de probabilidad que los rige o algún otro aspecto de nuestro interés; tema que se tratará con mayor profundidad en la sección 2.4.1.

Se comienza dando algunas definiciones que serán útiles para entender con mayor claridad el avance sobre el capítulo.

Definición 2.1 Se llama *espacio muestral* asociado a un experimento, al conjunto de todos los resultados posibles de dicho experimento. En general se indica este conjunto por Ω.

Ejemplo:

(i) $\Omega = \{\omega : \omega =$ las bases de la pirimidina (T y C) y la purina (A y G)$\} = \{T, A, C, G\}$

(ii) $\Omega = \{\omega : \omega =$ los diferentes tipos de aminoácidos$\}$
$=$ {Alanina, Arginina, Asparagina, Ácido aspártico, Cysteina, Glutamina, Ácido glutamínico, Glicina, Histidina, Isoleucina, Leucina, Licina, Metionina, Fenilalanina, Prolina, Serina, Treonina, Triptophano, Tirosina, Valina}

Definición 2.2 Un *evento* es cualquier subconjunto del espacio muestral Ω.

Por ejemplo, si el espacio muestral fuese Ω = {AA, AT, AC, AG, TA, TT, TC, TG, CA, CT, CC, CG, GA, GT, GC, GG} y se esté interesado en el evento A = las secuencias formadas únicamente por purinas, se obtiene que AG y GA cumplen con esta condición y decir que el evento "las secuencias formadas solamente por purinas" ocurre, es lo mismo que decir que el experimento tiene como resultado uno de los elementos del conjunto {AG, GA}.

Definición 2.3 Una σ*–álgebra* $\Im$ en un conjunto Ω es una clase de subconjuntos de Ω que cumplen las siguientes condiciones:

(i) $\emptyset \in \Im$ y $\Omega \in \Im$

(ii) Si $B_1, B_2, \ldots \in \Im$ entonces $\bigcup_i B_i \in \Im$

(iii) Si $B \in \Im$ entonces $B^c = (\Omega \setminus B) \in \Im$

donde $\emptyset$ es el conjunto vacío, B^c es el complemento del subconjunto B y $\Omega \setminus B$ indica el conjunto Ω sin los elementos de B.

Por ejemplo $\Im = \{\emptyset, \Omega\}$ es la σ*–álgebra* más trivial.

Definición 2.4 Una *variable aleatoria* (sobre Ω) es una función X que asocia $\omega \in \Omega$, a un valor $X(\omega)$ donde $X(\omega)$ es un número real.

Ejemplo:

Tómese a $\Omega = \{\omega : \omega =$ distintos tripletos formados por las bases A, C, G y T$\}$ donde Ω contendrá a 64 distintas combinaciones, ahora defina a cada una de las bases con un número, es decir, A = 1, C = 2, G = 3 y T = 4 y a X la variable aleatoria cuyo valor para cualquier elemento de Ω es el valor que resulta de sumar los valores de las bases de cada tripleto, es decir:

X(TTT) = 12, X(TTC) = 10, X(TTA) = 9, X(TTG) = 11, ... y así sucesivamente.

Definición 2.5 Un *espacio de probabilidad* finito está formado por un conjunto finito no vacío Ω, una σ–álgebra $\Im$ y una función $P(\cdot) : \Im \rightarrow [0,1]$, llamada una función de probabilidad.

Definición 2.6 La *función de probabilidad* $P(\cdot)$ es una función, en la familia $\Im$, de eventos del espacio muestral Ω que satisface:

(i) $P(\varnothing) = 0$ y $P(\Omega) = 1$

(ii) Si $B \subset \Omega$ entonces $P(B) \geq 0$

(iii) $P(D \cup B) = P(D) + P(B)$ si los eventos D y B son mutuamente excluyentes, es decir $D \cap B = \varnothing$

(iv) Si $B_1, B_2, \ldots$ son tales que $B_i \cap B_j = \varnothing$ para $i \neq j$, entonces $P(\cup B_i) = \sum_i P(B_i)$

Definición 2.7 La *probabilidad* de que un evento ocurra $B \subset \Omega$ está dado por $P(B) = \Sigma_{\omega \in B}\, p(\omega)$. Donde $p(\cdot)$ es una función de probabilidad sobre Ω. Si $B = \{\omega\}$ entonces se tiene un evento elemental y $P(B) = p(\omega)$.

Definición 2.8 Sea H un evento tal que $P(H) > 0$. Para cualquier evento B defina la siguiente probabilidad

$$P(B \mid H) = \frac{P(B \cap H)}{P(H)}$$

es llamada la *probabilidad condicional* de B dado H.

2.2.1 Teorema de Bayes.

La fórmula de Bayes expresa la llamada "probabilidad de las causas" y resuelve el problema de que suponiendo que un suceso D puede producirse como consecuencia de cualquiera B_1, B_2, ..., B_n sucesos y sabiendo que D se produjo, averiguar cuál es la probabilidad de que haya sido por causa del suceso B_i, o bien, formalmente:

Teorema 2.1 Sea B_1, B_2, ..., B_n una serie de sucesos del espacio muestral Ω, donde $P(B_i) > 0$, i = 1, 2, ..., n, se cumple $B_i \cap B_j = \emptyset$, i $\neq$ j y además $\cup_i B_i = \Omega$. Para cualquier evento $D \subset \Omega$, con $P(D) > 0$ sucede que

$$P(B_i \mid D) = \frac{P(B_i)\,P(D \mid B_i)}{\sum_{j=1}^{n} P(B_j)P(D/B_j)}$$

Demostración:

Dado que $P(D) > 0$ se puede utilizar la **Definición 2.8**. Entonces se puede escribir

$$P(B_i|D) = \frac{P(D \cap B_i)}{P(D)} = \frac{P(B_i)\,P(D|B_i)}{P(D)} \qquad \ldots.(2.1)$$

la última igualdad porque

$$P(D|B_i) = \frac{P(D \cap B_i)}{P(B_i)}$$

entonces

$$P(D \cap B_i) = P(B_i)\,P(D|B_i)$$

dado que $\bigcup_{j=1}^{n} B_j = \Omega$ se tiene

$$P(D) = P(D \cap \Omega) = P(D \cap (\bigcup_{j=1}^{n} B_j))$$

$$= P(\bigcup_{j=1}^{n}(D \cap B_j)) = \sum_{j=1}^{n} P(D \cap B_j) \qquad ...(2.2)$$

la última igualdad porque, dado que $B_i \cap B_j = \emptyset$ para $i \neq j$ se tiene que $(D \cap B_i) \cap (D \cap B_j) = \emptyset$ para $i \neq j$, además se tiene que para B_1, B_2, ..., B_n disjuntos

$$P(\bigcup_{j=1}^{n} B_j) = \sum_{j=1}^{n} P(B_j)$$

Por lo tanto de (2.1) y (2.2) se tiene:

$$P(B_i|D) = \frac{P(D \cap B_i)}{\sum_{j=1}^{n} P(D \cap B_j)} = \frac{P(B_i)\ P(D|B_i)}{\sum_{j=1}^{n} P(B_j)P(D/B_j)}$$

■

Definición 2.9 Se dice que n variables aleatorias $X_1,...,X_n$ tienen una *probabilidad discreta conjunta* si el vector aleatorio $(X_1,...,X_n)$ puede tomar solamente un número finito o una sucesión numerable de valores distintos posibles $(x_1,...,x_n)$ en R^n.

Definición 2.10 La *función de densidad de probabilidad conjunta* de $X_1,...,X_n$ se define como la función f tal que un punto $(x_1,...,x_n)$ en R^n cumple

$$f(x_1,...,x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$$

en notación vectorial, se dice que el vector aleatorio X tiene una distribución discreta y que su función de probabilidad en un punto $x \in R^n$ está dado por:

$$f(x) = P(X = x)$$

donde $X = (X_1,...,X_n)$, $x = (x_1,...,x_n)$ y $P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$ y para cualquier subconjunto $D \subset R^n$ sucede que $P(X \in D) = \Sigma_{x \in D} f(x)$

Definición 2.11 Se dice que n variables aleatorias $X_1,...,X_n$ tienen una *probabilidad continua conjunta* si existe una función no negativa f definida sobre R^n tal que para cualquier subconjunto $D \subset R^n$ sucede que

$$P[(X_1,...,X_n) \in D] = \int_D f(x_1,...,x_n)dx_1...dx_n.$$

La función f se denomina la función de densidad conjunta de X_1 ,..., X_n. En notación vectorial, $f(x)$ denota la funcion de densidad conjunta del vector aleatorio X y lo anterior se puede escribir de la siguiente forma :

$$P[X \in D] = \int_D f(x)\, dx$$

donde $X = (X_1,...,X_n)$, $x = (x_1,...,x_n)$ y $dx = dx_1\, dx_2 ...dx_n$

Definición 2.12 La *función de distribución conjunta* de n variables aleatorias independientes $X_1,..., X_n$ se define como la función $F : R^n \rightarrow [0,1]$ cuyo valor en un punto $(x_1,...,x_n)$ del espacio n-dimensional R^n está dada por:

$$\begin{aligned} F(x_1,...,x_n) &= P(X_1 \le x_1, X_2 \le x_2,..., X_n \le x_n) \\ &= \sum_{\{(y_1,...,y_n) : y_i \le x_i, i=1,...,n\}} P(X_1 \le x_1)...P(X_n \le x_n) \\ &= \left[\sum_{y_1 \le x_1} P(X_1 = y_1)\right]...\left[\sum_{y_n \le x_n} P(X_n = y_n)\right] \end{aligned}$$

si X_i es una variable aleatoria discreta, i = 1,2,...,n y

$$P(X_1 \le x_1, X_2 \le x_2,..., X_n \le x_n) = \int_{y_1 \le x_1} ... \int_{y_n \le x_n} f(y_1,..., y_n)dy_1...dy_n$$

si X_i es una variable aleatoria continua, i = 1,2,...,n.

Algunas propiedades de la función de distribución conjunta son:

1) F es una función no decreciente

2) $\lim_{x\to\infty} F(x) = 1$ donde $x \to \infty$ significa $x_i \to \infty$, $i = 1,2,...,n$

3) $\lim_{x\to-\infty} F(x) = 0$

4) $F(x)$ es continua por la derecha, es decir

$$\lim_{x\to x_0^+} F(x) = F(x_0)$$ donde $x \to x_0^+$ significa que $x_i \to x_{i_0}$ bajo la condición de que $x_i > x_{i_0}$

Si se conoce la densidad conjunta de n variables aleatorias $X_1,...,X_n$ entonces se puede obtener la distribución marginal de cualquier variable aleatoria X_i a partir de esta densidad conjunta.

Definición 2.13 La *función de densidad marginal* f_i de X_i para cualquier valor x_i está dada por:

$$f_i(x_i) = \int_{x_1}\cdots\int_{x_{i-1}}\int_{x_{i+1}}\cdots\int_{x_n} f(x_1,...,x_{i-1},x_i,x_{i+1},...,x_n)dx_1...dx_{i-1}dx_{i+1}...dx_n$$

para el caso específico de una densidad de probabilidad continua, y la *función de densidad marginal* f_i de X_i para cualquier valor x_i está dada por

$$f_i(x_i) = \sum_{x_1}\cdots\sum_{x_{i-1}}\sum_{x_{i+1}}\cdots\sum_{x_n} f(x_1,...,x_{i-1},x_i,x_{i+1},...,x_n)$$

para el caso específico de una densidad de probabilidad discreta.

2.3 Cadenas de Markov.

2.3.1 ¿Qué es un proceso estocástico?

Un proceso estocástico es una sucesión de variables aleatorias con valores en un mismo conjunto $E \subset R$. Se puede describir un proceso de la siguiente forma: se toma una familia de variables aleatorias $\{X_t\}$, en un mismo conjunto llamado espacio de estados donde X_t mide, en el instante t, el aspecto del proceso bajo consideración.

Por ejemplo: X_t podría ser el número de personas que llegan a hacer cola en el instante t en la caja de un supermercado, conforme transcurre el tiempo el número de personas aumentará o disminuirá, siendo atendidas por la encargada de la caja, y dependiendo de si llega una persona a formarse en la cola el valor de X_t aumentará en 1, por el contrario si la persona ya fue atendida, se va y no ha llegado nadie, el valor de X_t disminuirá también en una unidad. Es decir, en cualquier instante t, X_t toma valores del subconjunto 0,1,2,...; y t puede ser cualquier valor en el tiempo, o bien, del subconjunto $(-\infty, \infty)$. Los valores que puede tomar X_t son llamados sus *estados* y los cambios en el valor de X_t reciben el nombre de *transiciones* de sus estados. Lo que se acaba de describir representa la base de la elaboración de un proceso estocástico de una cola que incluye intervalos aleatorios entre las llegadas de las personas a la cola y períodos aleatorios en el punto de servicio. De este modelo sencillo se derivan una cantidad diversa de modelos más complicados.

Se puede observar que un proceso estocástico se encuentra en varias partes, en cualquier proceso que comprenda variabilidad al azar con el transcurso del tiempo: en la conservación de las especies naturales, en geofísica se han usado para predecir la magnitud y localización de los terremotos y en la industria para predecir las duraciones de una huelga.

Un proceso estocástico es una familia de variables aleatorias $\{X_t\}$ donde t es un punto en el espacio parametral T, y en donde para cada $t \in T$, X_t es un punto en un espacio E, llamado espacio de estados. Ahora bien, la teoría de la probabilidad considera de interés

establecer las relaciones entre las X_t, para los diferentes valores fijos de t.

2.3.2 Introducción a las cadenas de Markov.

Definición 2.14 Un *proceso de Markov* es un proceso estocástico $\{X_t : t \in T \subseteq (-\infty,\infty)\}$ para el cual, dado el valor de X_t, la distribución de X_s ($s > t$) no depende en forma alguna de un conocimiento de X_u ($u < t$), es decir, el comportamiento del proceso en un tiempo futuro, cuando se conoce el estado presente del proceso, no se altera por el conocimiento adicional acerca de su comportamiento pasado. Esto se traduce, formalmente, de la siguiente forma, para

$$\lambda_0 < \lambda_1 < \ldots < \lambda_k < t_1 < t_2 \ldots < t_n$$

se tiene

$$P(X_{t_1},\ldots,X_{t_n} / X_{\lambda_0} = x_0,\ldots,X_{\lambda_k} = x_k) = P(X_{t_1},\ldots,X_{t_n} / X_{\lambda_k} = x_k)$$

es decir, si se tiene

$$t_0 < t_1 < t_2 < \ldots < t_n < t < t'$$

entonces

$$P(X_{t'} = x / X_{t_0} = x_0,\ldots,X_{t_n} = x_n, X_t = y) = P(X_{t'} = x / X_t = y)$$

con $x, y, x_i \in E$, i = 0, …, n lo anterior se conoce como propiedad de Markov.

Definición 2.15 Una *cadena de Markov discreta* es un proceso de Markov $X = \{ X_n : n \in T \}$ con un espacio de tiempo discreto $T = \{0, 1, 2, \ldots\}$ y un espacio de estados E finito o contable infinito.

Las probabilidades de transición de una cadena de Markov pueden o no ser independientes de n. Si la probabilidad de transición, denotada por P_{ij} es independiente de n entonces la cadena de Markov se dice que es homogénea en el tiempo (o tiene probabilidad de transición estacionarios) y se tiene que

$$P_{ij}^{(n,n+m)} = P(X_{n+m} = j / X_n = i) = P(X_{n'+m} = j / X_{n'} = i)$$

tomando a n' cualquier paso en el tiempo, con i, j $\in E$. Y se escribe

$$P_{ij}^{(n,n+m)} = P_{ij}^{(m)}$$

donde $\{P_{ij}^{(m)} : i, j \in E\}$ es llamada la matriz de transición en ***m*** pasos.

Para ***m*** = 1 se tiene que $\{P_{ij} : i, j \in E\}$ es llamada la matriz de transición de la cadena ***X***. Indíquese por $P = \{P_{ij} : i, j \in E\}$ a esta matriz de transición. La cual es de N x N si E es finito y tiene N elementos. Si E es infinito entonces la matriz P es infinita. Todas las entradas de la matriz P deben ser no-negativas y

$$\sum_{j \in E} P_{ij} = 1$$

de este modo se dice que P es una *matriz estocástica.*

Asimismo se considera una distribución de probabilidad inicial sobre E, $\pi_0(\cdot)$, que es una colección de números no negativos $\pi_0 = \{\pi_0(i) ; i \in E\}$ que suman uno, es decir $\pi_0(\cdot) = P(X_0 = \cdot)$ es tal que

$$0 \leq \pi_0(i) \leq 1 \quad \text{para todo } i \in E \qquad \text{y} \qquad \sum_{i \in E} \pi_0(i) = 1$$

La sucesión aleatoria ***X*** es llamada una cadena de Markov con matriz de probabilidad de transición P y distribución inicial π_0. La definición de ***X*** especifica que x_0 es tomada en E dependiendo de la distribución de probabilidad π_0 y que si $x_n = i$ y si ***X*** es homogénea en el tiempo entonces $x_{n+1} = j$ con probabilidad P_{ij}, independientemente de los valores que X_m asuma para m < n.

A tiempo discreto, la posición en el espacio de estados del proceso, llamado el estado de la cadena de Markov, es registrada en cada unidad de tiempo, mientras que en tiempo continuo el estado es observado continuamente.

A continuación se muestra que π_0 y P determinan completamente la distribución conjunta de n valores sucesivos de la cadena de Markov.

Lema 2.1 Para todo $n \geq 0$ y $i_0, i_1, \ldots, i_n \in E$, se tiene

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \pi_0(i_0) P_{i_0 i_1} P_{i_1 i_2} \cdot \ldots \cdot P_{i_{n-1} i_n}$$

Demostración

(1) Se demuestra para n = 1.

$$P(X_0 = i_0, X_1 = i_1) = P(X_1 = i_1 / X_0 = i_0) P(X_0 = i_0) = \pi_0(i_0) P_{i_0 i_1}$$

(2) Se supone cierto para n = k

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_k = i_k) = \pi_0(i_0) P_{i_0 i_1} P_{i_1 i_2} \ldots P_{i_{k-1} i_k}$$

Y se demuestra para n = k+1

$$\begin{aligned} P(X_0 = i_0, X_1 = i_1, \ldots, X_{k+1} = i_{k+1}) &= P(X_{k+1} = i_{k+1} / X_0 = i_0, X_1 = i_1, \ldots, X_k = i_k) \\ &\quad P(X_0 = i_0, X_1 = i_1, \ldots, X_k = i_k) \\ &\overset{(1)}{=} P(X_{k+1} = i_{k+1} / X_0 = i_0, \ldots, X_k = i_k) \\ &\quad \pi_0(i_0) P_{i_0 i_1} \ldots P_{i_{k-1} i_k} \\ &\overset{(2)}{=} P(X_{k+1} = i_{k+1} / X_k = i_k) \pi_0(i_0) P_{i_0 i_1} \ldots P_{i_{k-1} i_k} \\ &= \pi_0(i_0) P_{i_0 i_1} \ldots P_{i_{k-1} i_k} P_{i_k i_{k+1}} \end{aligned}$$

(1) Por hipótesis de inducción.
(2) Por propiedad de cadena de Markov.

■

Definición 2.16 Sea P una matriz de transición en el espacio de estados E. La matriz de transición P es *irreducible* si es posible que una cadena de Markov con matriz de transición P pueda moverse de cualquier estado i a otro j en un tiempo finito. Una cadena de Markov con matriz de probabilidad de transición P se dice que es irreducible si su matriz P es irreducible.

El modelo de cadenas de Markov se utiliza más adelante, en el **Capítulo 4**, para modelar el comportamiento de las secuencias de A.D.N. Para tener esto un poco más claro, recuerde que en el método de secuenciación de A.D.N. se pasaba a través de un gel de electroforesis una corriente eléctrica que hacía que los pedacitos de cromosomas con terminaciones en las bases A, T, G y C corrieran sobre el gel y se ubicaran en una posición determinada, de tal modo que al pasar un lector óptico se podían observar las posiciones de las bases (ver Fig.1.9). No obstante, puede suceder que uno o dos pedacitos del gen no se noten claramente, lo que induciría al "ruido". Este tipo de alteraciones puede ser estimado por medio de el método de cadenas de Markov Ocultas.

Existen dos posibilidades para el uso del método. Una de ellas es considerar que la cadena de Markov modela la sucesión de bases en la secuencia de A.D.N. En este caso se considera que la posición de las bases sea $\{X_k\}_{k\geq 0}$ una cadena de Markov, una vez utilizado el método de electroforesis. Ahora bien, como intervienen varias variables que pueden afectar nuestra observación se toma en cuenta que lo que observa el investigador es una sucesión $\{Y_k\}_{k\geq 0}$ (la cual se ve afectada por el ruido). El caso en que $X_k = Y_k$; para $k \geq 0$, vendría siendo el estado en el que no se encuentra oculta la cadena de Markov $\{X_k\}_{k\geq 0}$, es decir, lo que observa el investigador es exactamente igual a la secuencia de A.D.N., en cuyo caso no existe ningún tipo de ruido.

Otra posibilidad es suponer que toda la secuencia es un estado de la cadena de Markov y modelar por cadena de Markov las posibles modificaciones (mutaciones) que pueden ocurrir en una posición específica. Acerca de esto se habla de forma más extensa en el **Capítulo 4** donde se considera esta situación para modelar la secuencia de A.D.N. que se obtiene después de su armado.

2.4 La función de Verosimilitud.

Dentro del análisis que se lleva a cabo en este trabajo será de vital importancia manejar claramente el concepto de verosimilitud. La verosimilitud de un modelo es una función que depende de los datos observados y de los parámetros o distribución que los rige, asimismo, provee una medida comparativa de que tan bien se pueden predecir las observaciones que ya se han realizado una vez que se tienen los modelos. La función de verosimilitud es vista como una función de los parámetros del modelo dado que los datos son observados y lo que se quiere es estimar los parámetros que dan origen a las observaciones.

Para empezar se dará una introducción al principio de verosimilitud para posteriormente pasar al concepto de máxima verosimilitud como tal.

2.4.1 Principio de verosimilitud

Definición 2.17 Cualquier característica con valor desconocido de la distribución que genera los datos experimentales se llama *parámetro* de la distribución. En general se denota el parámetro por θ (que puede ser un vector o una única incógnita).

Por ejemplo, para el caso de un modelo regido por la distribución multinomial (ver **A.3**) los parámetros que tienen un valor desconocido, son p_i para $i = 1, ..., k$, donde k son los diferentes resultados del experimento multinomial y p_i con $i = 1, ...,$ k son las probabilidades respectivas de los resultados.

Asuma que se ha desarrollado un modelo para un proceso físico y que se desea determinar los valores más deseables para algunos parámetros de este modelo. La teoría de utilizar datos experimentales para estimar estos parámetros es conocida como la teoría de la estimación. Dentro de la teoría de la estimación, es importante que al trabajar con modelos se pueda asegurar que se utiliza un modelo que realmente se acerque a las observaciones. Es decir, un modelo es más verosímil o plausible que otro, si al considerar únicamente las observaciones bajo la función de verosimilitud los datos son más probables.

Definición 2.18 Se utiliza la palabra *estadística* para describir una función de las variables aleatorias con la distribución de probabilidad con la que se trabaja. La media y la varianza son las estadísticas más comunes.

Por ejemplo, para la distribución multinomial descrita en **A.3**

$$P(Y_1 = y_1, ..., Y_k = y_k) = \frac{n! p_1^{y_1} p_2^{y_2} ... p_k^{y_k}}{y_1! y_2! ... y_k!}$$

$$n = 1, 2, ...;\ 0 < p_j < 1 ; \sum_{j=1}^{k} p_j = 1 ;\ y_j = 0, 1, ..., n ; \sum_{j=1}^{k} y_j = n$$

Las estadísticas son la media o esperanza $E(y_j) = n\, p_j$ y la varianza $Var(y_j) = n\ p_j\ (1- p_j)$ con $j = 1, 2, ..., k$.

Definición 2.19 Un *estimador* $\hat{\boldsymbol{\theta}}$ para un parámetro desconocido θ es una función de $X_1, X_2, ..., X_n$ (variables aleatorias) que representa a los valores observados bajo el modelo con parámetro θ. Se puede pensar en $\hat{\boldsymbol{\theta}}$ como una variable aleatoria con sus propias leyes de probabilidad.

Ahora bien, el principio de verosimilitud hace explícito el hecho de que sólo son importantes los datos actuales observados y que éstos mismos son los que nos aportan evidencia relevante acerca del parámetro que se quiere estimar (ver Florens, Mouchart, et.al.(1990)). El concepto preciso que existe en el *Principio de Verosimilitud* está íntimamente ligado con la función de verosimilitud (**Definición 2.20**).

Definición 2.20 Para datos observados, $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, la función $L(\theta) \propto f(x;\theta)$ considerada como una función de θ, es llamada la *función de verosimilitud* de θ, donde $f(\cdot\ ;\theta)$ es la densidad de $\boldsymbol{X} = (X_1, ..., X_n)$ bajo el parámetro θ.

Por ejemplo, para el caso de la distribución multinomial, su función de densidad vendría siendo

$$f(y;\theta) = \frac{n! p_1^{y_1} p_2^{y_2} \ldots p_k^{y_k}}{y_1! y_2! \ldots y_k!}$$

y su función de verosimilitud se puede considerar como

$$L(\theta) \propto \frac{p_1^{y_1} p_2^{y_2} \ldots p_k^{y_k}}{y_1! y_2! \ldots y_k!}$$

Note que la diferencia estriba en una constante que no afecta a la función de verosimilitud, ya que ésta sólo depende del parámetro con el que se está trabajando.

- El principio de verosimilitud dice que si dos conjuntos distintos de valores observados tienen la misma función de verosimilitud entonces se obtendrá el mismo estimador máximo verosímil θ^* de θ para ambos conjuntos. Esto se debe al hecho de que θ^* depende de los datos solamente a través de la función de verosimilitud. De la misma forma se tiene que para $L_i(\theta)$ (con i = 1,2), funciones de verosimilitud proporcionales entre sí, $L_1(\theta)$ y $L_2(\theta)$ contendrán la misma información acerca de θ, ya que son proporcionales como funciones de θ, sin embargo no es necesario que sean iguales.

En otras palabras, como las dos funciones para el experimento sólo se diferencian por un factor que depende de las muestras obtenidas de los conjuntos, y no del parámetro θ, entonces sucede que ambas funciones proporcionan exactamente la misma información acerca de θ y se puede tomar tanto a una función $L_1(\theta)$ como a $L_2(\theta)$ indistintamente para calcular la estimación del parámetro θ; esto es lo que se conoce como el *Principio de Verosimilitud.*

Al hacer inferencia acerca de θ después de haber observado x, toda información experimental relevante está contenida en la función de verosimilitud para los datos observados x. Más aún, dos funciones de verosimilitud contienen la misma información acerca de θ si son proporcionales entre sí (como función de θ), y un estadístico obtendrá la misma información de θ a partir de cualquiera de las dos funciones de verosimilitud.

Definición 2.21 El conjunto Θ de todos los valores posibles de un vector de parámetros $\theta = (\theta_1, \theta_2, ..., \theta_k)$ se llama *espacio paramétrico.*

Un problema de inferencia estadística, en rasgos generales, intenta determinar dónde es más probable que se encuentre el verdadero valor del parámetro desconocido θ en el espacio paramétrico Θ, partiendo desde un principio de las observaciones x y de la función de verosimilitud $L(\theta)$.

2.5 Máxima verosimilitud.

En 1921 Sir R. A. Fisher propuso un método para estimar parámetros y puntualizó algunas razones por los cuales era mejor método que otros utilizados para el mismo fin. El procedimiento propuesto por Fisher es llamado de máxima verosimilitud y es conocido por ser el más utilizado y conveniente para estimar parámetros. El método de máxima verosimilitud indica que debe examinarse la función de verosimilitud de los valores de la muestra y tomar como estimados de los parámetros desconocidos aquellos valores que maximicen la función de verosimilitud.

El método de estimación de máxima verosimilitud es un método puntual muy conocido y usado dentro de la inferencia estadística clásica. Este método consiste en tomar como estimador del parámetro θ el valor que haga máxima la probabilidad de ocurrencia de la muestra dado el parámetro desconocido θ. Es decir, se toma una función de densidad $f(\cdot\,;\theta)$ y se extrae una muestra independiente $x_1, x_2, ..., x_n$; la función de densidad de la muestra (que es proporcional a la verosimilitud del problema) será el producto:

$$f(x_1;\theta)f(x_2;\theta)...f(x_n;\theta) \propto L(\theta) \qquad .....(2.3)$$

En este contexto se obtendrá el estimador θ^* para el cual la probabilidad de haber obtenido la muestra $x_1, x_2, ..., x_n$ sea máxima y por el principio de verosimilitud se tendrá el valor que maximiza $L(\theta)$.

Dado que $L(\theta)$ en general es proporcional a un producto de funciones de θ, entonces para obtener el valor θ^* se trabaja con $\log L(\theta)$. La función logarítmica es creciente, de modo que encontrar el máximo para $L(\theta)$ es equivalente a encontrar el máximo para $l(\theta) = \log L(\theta)$. De este modo el producto de funciones que se tiene en (2.3) se convierte en una suma y se determina a θ^* a través de:

$$L(\theta) \propto \frac{g_\theta(x_1)}{c} \cdots \frac{g_\theta(x_n)}{c} = \frac{1}{c^n} g_\theta(x_1) \cdots g_\theta(x_n)$$

tome $L(\theta) = g_\theta(x_1) \ldots g_\theta(x_n)$, donde $g_\theta(x_i)$ es proporcional a la función de densidad de la variable aleatoria X_i.

$$l(\theta) = \log g_\theta(x_1) + \ldots + \log g_\theta(x_n)$$

$$l(\theta) = \sum_{i=1}^{n} \log g(x_i ; \theta) \qquad \text{.....(2.4)}$$

Observación. Note que se puede tomar

$$L(\theta) = f(x_1; \theta) \ldots f(x_n; \theta)$$

en este caso

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i ; \theta)$$

y de ahora en adelante se tomará de esta forma a la función de verosimilitud $L(\theta)$ y la de log-verosimilitud $l(\theta)$.

Ahora bien, una de las formas de obtener el valor θ^* que maximiza $l(\theta)$ es derivando la ecuación en relación a θ, igualando a cero la ecuación (2.4) (que puede ser un sistema de ecuaciones si θ es multivariado) y se encuentra la solución θ^*, lo que se conocerá como el estimador máximo verosímil del parámetro θ desconocido.

Definición 2.22 Todos aquellos puntos x en donde la tangente a la curva $y = f(x)$ es horizontal, es decir los puntos correspondientes a

las raíces ξ de la ecuación $f'(\xi) = 0$ se llaman puntos críticos o puntos estacionarios de f.

Definición 2.23 Un estimador máximo verosímil de θ^* es un valor de θ que maximiza la verosimilitud $L(\theta)$ o equivalentemente la log-verosimilitud $l(\theta)$.

Por ejemplo, para el caso de la distribución multinomial, para k = 3 se tendría

$$L(\theta) = \frac{p_1^{y_1} p_2^{y_2} (1-p_1-p_2)^{y_3}}{y_1!\, y_2!\, y_3!}$$

Donde $\theta = (p_1, p_2)$. Tomando en cuenta el hecho de que $p_1 + p_2 + p_3 = 1$, primero se aplica la función logaritmo a la función de verosimilitud:

$$l(\theta) = \log L(\theta) = y_1 \log p_1 + y_2 \log p_2 + y_3 \log (1 - p_1 - p_2),$$

después se obtiene la derivada con respecto al o los parámetros que se quieran estimar, en este caso los parámetros son p_1 y p_2, ya que p_3 se encuentra en función de éstas dos. Derivando con respecto a p_1 se tiene

$$\frac{\partial l(\theta)}{\partial p_1} = \frac{y_1}{p_1} - \frac{y_3}{1-p_1-p_2}$$

y derivando con respecto a p_2 se obtiene

$$\frac{\partial l(\theta)}{\partial p_2} = \frac{y_2}{p_2} - \frac{y_3}{1-p_1-p_2}$$

Igualando a cero se obtiene un sistema de dos ecuaciones con dos incógnitas:

$$\frac{y_1}{p_1} - \frac{y_3}{1-p_1-p_2} = 0$$

$$\frac{y_2}{p_2} - \frac{y_3}{1-p_1-p_2} = 0$$

Resolviendo el sistema de ecuaciones se obtiene

$$p_1^* = \frac{y_1}{y_1 + y_2 + y_3} \quad \text{y} \quad p_2^* = \frac{y_2}{y_1 + y_2 + y_3}$$

y por lo tanto $$p_\theta^* = \left(\frac{y_1}{y_1 + y_2 + y_3}, \frac{y_2}{y_1 + y_2 + y_3} \right)$$

2.6 Algunos métodos numéricos de aproximación.

2.6.1 Introducción.

Anteriormente se vio que la función de verosimilitud $L(\theta)$, que es proporcional a un producto de funciones de densidades de las variables aleatorias $f(\cdot\,; \theta)$, puede tener una forma en la cual se dificulte encontrar θ^* de forma usual: derivando $L(\theta)$ con respecto a θ y encontrar a θ de forma explícita, o lo que es lo mismo encontrar solución a

$$\frac{\partial L(\theta)}{\partial \theta} = 0$$

que maximice $L(\theta)$. Debido a esto se trabaja con el logaritmo de la función de verosimilitud $l(\theta) = \log L(\theta)$. No obstante, esta forma puede ser todavía difícil de trabajar y no se puede resolver el problema de forma analítica. Una vez que se encuentra éste obstáculo lo único que resta hacer es tratar de aproximar el valor de θ^* por otros métodos. Y no se piense que esto es poco frecuente, se presenta en la mayor parte de los problemas estadísticos que tengan que ver con distribuciones de probabilidad más complejas con las cuales se puede modelar de forma más precisa algún fenómeno en particular.

Existen diversos métodos numéricos de aproximación para estimar parámetros, la mayor parte de los cuales se basan principalmente en encontrar los ceros de la función. Dentro del análisis numérico se pueden mencionar varias técnicas para encontrar los ceros de una función específica, como lo pueden ser el método de Newton-Raphson, los métodos quasi-Newton, métodos de Newton modificados y el método de puntuación de Fisher, que se encuentra dentro de los métodos de Newton modificados y a su vez su versión modificada utilizando la matriz de información empírica en lugar de la matriz de esperanza (ver Mclachlan y Krishnan (1997)).

Se explicarán brevemente dos de los más importantes y conocidos métodos numéricos: el método de Newton-Raphson y el método de puntuación (scoring) de Fisher; para posteriormente hablar de forma más extensa del Algoritmo E.M., el cual es un método iterativo de gran utilidad.

2.6.2 El método de Newton-Raphson.

Existen diversas formas de explicar el método Newton-Raphson, no obstante, la forma que se utilizará para dar una idea general del método será la explicación del método gráfico, después de lo cual se adentrará en la construcción del método utilizando una aproximación por expansión de Taylor para una función de variable real y posteriormente se realizará la extensión a una función en el espacio n-dimensional.

Comience definiendo a $g(\cdot)$ una función de variable real a la cual se le quiere estimar el valor del argumento que la maximice, que en el caso que interesa vendría siendo

$$g(\theta) = \frac{\partial l(\theta)}{\partial \theta}$$

El método de Newton-Raphson, también llamado simplemente método de Newton es un método iterativo basado en la aproximación de la derivada, para el caso real, o el vector gradiente, en el espacio n dimensional, de la función $g(\cdot)$. Nótese que el hecho

de utilizar la transformación de logaritmo se realiza para evitar cargas pesadas en la computadora, además de que la función logaritmo es útil como transformación por tener un comportamiento "bonito" en sentido matemático.

Tomando en cuenta la Fig. 2.1 considere $y = g(x)$ una función con raíz en x^* y x_0 la primera aproximación a la raíz x^*, que debe cumplir con que la distancia entre x^* y x_0 debe ser suficientemente pequeña. Asimismo, la función $g'(x_i) \neq 0$ para todo i = 1, ..., n, y $g'(x)$ no debe estar muy próximo a cero para que el método converja, es decir, no debe haber raíces cercanas unas de otras en términos relativos (ver Fig.2.2).

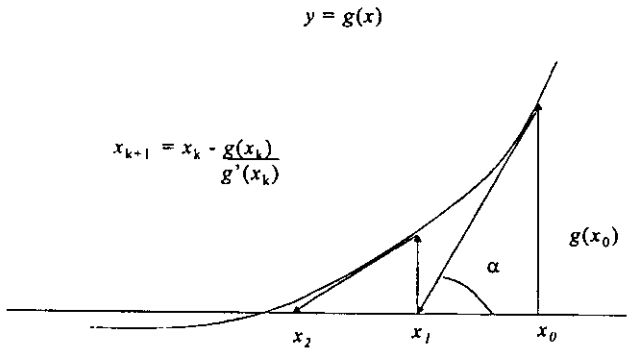


Fig.2.1 Método de Newton-Raphson para $g(x) = 0$

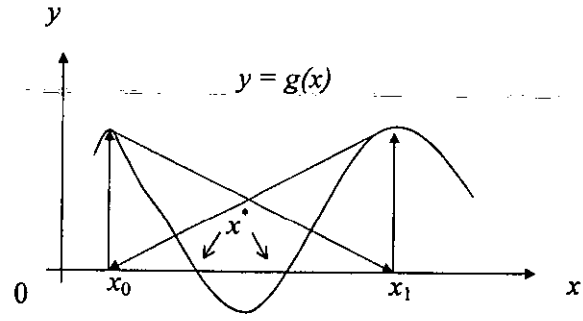


Fig.2.2 Caso de raíces cercanas. (ver Luthe, Olivera, et.al. (1980) Fig. 3-8)

Como se había mencionado anteriormente, el método de Newton se basa en una aproximación lineal a la función $g(\theta)$ con que se trabaja para encontrar sus raíces o ceros de la función, para lo cual utiliza la línea tangente a $g'(\theta)$. El primer paso consiste en trazar una vertical desde la primera aproximación θ_0, que no debe estar lejos de la raíz θ^*, hasta encontrar la curva de la ecuación $y = g(\theta)$; por el punto de corte $(\theta_0, g(\theta_0))$ trazar una tangente a la curva $g(\theta)$ hasta intersectar el eje x, en este punto de intersección se tendrá la nueva aproximación θ_1 para θ^* y se repetirá el proceso tantas veces sea necesario hasta que se obtenga una aproximación suficientemente confiable a la raíz θ^* de la ecuación. De tal forma que lo que se obtiene de forma analítica es:

$$\tan \alpha = \frac{\operatorname{sen} \alpha}{\cos \alpha} = \frac{g(\theta_0)}{(\theta_0 - \theta_1)} = g'(\theta_0)$$

de donde se obtiene al despejar θ_1 que

$$\theta_1 = \theta_0 - \frac{g(\theta_0)}{g'(\theta_0)} \qquad \text{con } g'(\theta_0) \neq 0$$

y de igual forma

$$\theta_2 = \theta_1 - \frac{g(\theta_1)}{g'(\theta_1)} \qquad \text{con } g'(\theta_1) \neq 0$$

y de forma mas general

$$\theta_k = \theta_{k-1} - \frac{g(\theta_{k-1})}{g'(\theta_{k-1})} \qquad \text{con } g'(\theta_{k-1}) \neq 0$$

Ahora bien, la otra forma de deducir el algoritmo recursivo de Newton-Raphson de forma analítica está basada en polinomios de Taylor.

Sea g una función doblemente derivable en un intervalo (a,b). Sea $\theta^{(n-1)}$ una aproximación a θ^*, la raíz de la función g, es decir $g(\theta^*) = 0$, $g'(\theta^{(i)}) \neq 0$ con i = 1,...,n-1 y $|\theta^{(n-1)} - \theta^*|$ sea "pequeña". Considere el polinomio de Taylor para $g(\theta)$ expandido para $\theta^{(n-1)}$.

$$g(\theta) = g(\theta^{(n-1)}) + (\theta - \theta^{(n-1)})\, g'(\theta^{(n-1)}) + \frac{(\theta - \theta^{(n-1)})^2}{2} g''(\xi(\theta))$$

donde $\xi(\theta)$ se encuentra entre θ y $\theta^{(n-1)}$. Como $g(\theta^*) = 0$, esta ecuación para $\theta = \theta^*$es:

$$0 = g(\theta^{(n-1)}) + (\theta^* - \theta^{(n-1)})\, g'(\theta^{(n-1)}) + \frac{(\theta^* - \theta^{(n-1)})^2}{2} g''(\xi(\theta^*))$$

El método de Newton se deduce asumiendo que como $|\theta^* - \theta^{(n-1)}|$ es pequeña entonces el término $|\theta^* - \theta^{(n-1)}|^2$ es todavía más pequeño y se puede hacer:

$0 \approx g(\theta^{(n-1)}) + (\theta^* - \theta^{(n-1)})\, g'(\theta^{(n-1)})$ de tal forma que resolviendo para θ^* se obtiene:

$$\theta^* \approx \theta^{(n-1)} - \frac{g(\theta^{(n-1)})}{g'(\theta^{(n-1)})}$$

lo cual establece el método de Newton-Raphson.

Cabe recordar que se empieza con una aproximación inicial $\theta^{(0)}$ y se genera la sucesión $\{\theta^{(n)}\}$ definida por:

$$\theta^{(n)} = \theta^{(n-1)} - \frac{g(\theta^{(n-1)})}{g'(\theta^{(n-1)})} \qquad \text{con } g'(\theta^{(n-1)}) \neq 0 \text{ para } n \geq 1, \ldots.(2.5)$$

Bajo determinadas condiciones se puede afirmar que el método iterativo de Newton-Raphson converge[14] al menos localmente a una raíz de la ecuación $g(\theta) = 0$, es decir, que si $\theta^{(0)}$ es la primera aproximación a la raíz de $g(\theta) = 0$, debe cumplirse que

1. $\theta^{(0)}$ debe estar suficientemente cercana a la raíz de $g(\theta) = 0$.

2. $g(\theta)$ debe ser cóncava[15] en el intervalo (a,b) para que se cumpla la monotonicidad de la función.

[14] Para mayor información acerca de la convergencia del método referirse a: Luthe, Olivera, et.al.(1980).

[15] La definición de concavidad se encuentra en el *Apéndice A.1*

3. $g \in C^2$ (a,b)

4. $g'(\theta^*) \neq 0$

5. $g'(\theta)$ no debe estar muy próximo a cero, esto quiere decir que no debe haber raíces cercanas unas de otras. En la Fig. 2.2 se muestra un problema de no convergencia que puede presentarse al tener raíces relativamente cercanas.

Ahora bien, una vez que se ha tratado el método de Newton-Raphson para el caso real, se realizará una extensión para el caso de más de una variable.

Tómese $\boldsymbol{X}$ un vector aleatorio k-dimensional con función de densidad de probabilidad $f(\boldsymbol{x};\theta)$ en $\mathbf{R}^k$, donde $\theta = (\theta_1, \theta_2, ..., \theta_n)^t$ [16] es el vector que contiene a los parámetros desconocidos, dentro de la f.d.p. para $\boldsymbol{X}$. Donde el espacio de parámetros es denotado por Θ, $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ denota un vector aleatorio continuo, que representa a las observaciones.

La función de verosimilitud para θ formada en base a los datos observados $\boldsymbol{x}$ está dada por:

$$L(\theta) = f(\boldsymbol{x};\theta)$$

como ya se había visto anteriormente, el estimador de θ puede ser obtenido a partir de igualar a cero la ecuación:

$$\frac{\partial L(\theta)}{\partial \theta} \text{ o bien de forma equivalente a } \frac{\partial l(\theta)}{\partial \theta}$$

denótese a $$M(\boldsymbol{x};\theta) = \frac{\partial l(\theta)}{\partial \theta} = \left[\frac{\partial l(\theta)}{\partial \theta_1}, ..., \frac{\partial l(\theta)}{\partial \theta_n}\right]^t \quad .....(2.6)$$

Ahora bien, el método de Newton-Raphson utiliza para resolver (2.6) el aproximar el vector gradiente $M(\boldsymbol{x};\theta)$ de la función de log verosimilitud $l(\theta)$ por una expansión de serie de Taylor lineal en una vecindad de $\theta^{(k)}$, la k-ésima aproximación. Considerando:

[16] De aquí en adelante $(x_1,x_2,...,x_n)^t$ denotará a la transpuesta de un vector $(x_1,x_2,...,x_n)$.

$$I(\theta^{(k)}; x) = \frac{-\partial^2 \log L(\theta)}{\partial\theta^{(k)}\partial\theta^{(k)\,t}} = -\begin{pmatrix} \frac{\partial^2 l(\theta)}{\partial\theta_1^{(k)}\partial\theta_1^{(k)}}, \ldots, \frac{\partial^2 l(\theta)}{\partial\theta_1^{(k)}\partial\theta_n^{(k)}} \\ \vdots \\ \frac{\partial^2 l(\theta)}{\partial\theta_n^{(k)}\partial\theta_1^{(k)}}, \ldots, \frac{\partial^2 l(\theta)}{\partial\theta_n^{(k)}\partial\theta_n^{(k)}} \end{pmatrix}$$

se tiene

$$M(x;\theta) \approx M(x;\theta^{(k)}) - I(\theta^{(k)}; x)(\theta - \theta^{(k)}) \qquad \text{.....(2.7)}$$

o bien

$$\begin{pmatrix} \frac{\partial l(\theta)}{\partial\theta_1} \\ \vdots \\ \frac{\partial l(\theta)}{\partial\theta_n} \end{pmatrix} \approx \begin{pmatrix} \frac{\partial l(\theta)}{\partial\theta_1^{(k)}} \\ \vdots \\ \frac{\partial l(\theta)}{\partial\theta_n^{(k)}} \end{pmatrix} - \begin{pmatrix} \frac{\partial^2 l(\theta)}{\partial\theta_1^{(k)}\partial\theta_1^{(k)}}, \ldots, \frac{\partial^2 l(\theta)}{\partial\theta_1^{(k)}\partial\theta_n^{(k)}} \\ \vdots \\ \frac{\partial^2 l(\theta)}{\partial\theta_n^{(k)}\partial\theta_1^{(k)}}, \ldots, \frac{\partial^2 l(\theta)}{\partial\theta_n^{(k)}\partial\theta_n^{(k)}} \end{pmatrix} \begin{pmatrix} \theta_1 - \theta_1^{(k)} \\ \vdots \\ \theta_n - \theta_n^{(k)} \end{pmatrix}$$

La matriz $I(\theta\ ;\ x)$ es llamada la *matriz de información observada*, que como se puede notar es cuadrada, y debe cumplir que su determinante sea distinto de cero para poder ser invertible, condición que más adelante será de mucha utilidad. Para obtener $\theta^{(k+1)}$ se iguala a cero la parte derecha de la ecuación (2.7):

$$M(x;\theta^{(k)}) - I(\theta^{(k)};x)(\theta - \theta^{(k)}) = 0$$

$$\begin{pmatrix} \frac{\partial l(\theta)}{\partial\theta_1^{(k)}} \\ \vdots \\ \frac{\partial l(\theta)}{\partial\theta_1^{(k)}} \end{pmatrix} - \begin{pmatrix} \frac{\partial^2 l(\theta)}{\partial\theta_1^{(k)}\partial\theta_1^{(k)}}, \ldots, \frac{\partial^2 l(\theta)}{\partial\theta_1^{(k)}\partial\theta_n^{(k)}} \\ \vdots \\ \frac{\partial^2 l(\theta)}{\partial\theta_n^{(k)}\partial\theta_1^{(k)}}, \ldots, \frac{\partial^2 l(\theta)}{\partial\theta_n^{(k)}\partial\theta_n^{(k)}} \end{pmatrix} \begin{pmatrix} \theta_1 - \theta_1^{(k)} \\ \vdots \\ \theta_n - \theta_n^{(k)} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

donde reagrupando términos se tiene

$$M(x;\theta^{(k)}) = I(\theta^{(k)}; x)(\theta - \theta^{(k)})$$

$$\begin{pmatrix} \frac{\partial l(\theta)}{\partial \theta_1^{(k)}} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial l(\theta)}{\partial \theta_1^{(k)}} \end{pmatrix} = - \begin{pmatrix} \frac{\partial^2 l(\theta)}{\partial \theta_1^{(k)} \partial \theta_1^{(k)}} , \ldots , \frac{\partial^2 l(\theta)}{\partial \theta_1^{(k)} \partial \theta_n^{(k)}} \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \frac{\partial^2 l(\theta)}{\partial \theta_n^{(k)} \partial \theta_1^{(k)}} , \ldots , \frac{\partial^2 l(\theta)}{\partial \theta_n^{(k)} \partial \theta_n^{(k)}} \end{pmatrix} \begin{pmatrix} \theta_1 - \theta_1^{(k)} \\ \cdot \\ \cdot \\ \cdot \\ \theta_n - \theta_n^{(k)} \end{pmatrix}$$

suponga que la matriz de información observada tiene inversa

$$M(x;\theta^{(k)})\, I^{-1}(\theta^{(k)};x) = (\theta - \theta^{(k)})$$

despejando θ y haciendo $\theta = \theta^{(k+1)}$ se tiene

$$\theta^{(k+1)} = \theta^{(k)} + I^{-1}(\theta^{(k)};x)\, M(x;\theta^{(k)}) \qquad .....(2.8)$$

es decir

$$\begin{pmatrix} \frac{\partial l(\theta)}{\partial \theta_1^{(k)}} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial l(\theta)}{\partial \theta_1^{(k)}} \end{pmatrix} \begin{pmatrix} \frac{\partial^2 l(\theta)}{\partial \theta_1^{(k)} \partial \theta_1^{(k)}} , \ldots , \frac{\partial^2 l(\theta)}{\partial \theta_1^{(k)} \partial \theta_n^{(k)}} \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \frac{\partial^2 l(\theta)}{\partial \theta_n^{(k)} \partial \theta_1^{(k)}} , \ldots , \frac{\partial^2 l(\theta)}{\partial \theta_n^{(k)} \partial \theta_n^{(k)}} \end{pmatrix} + \begin{pmatrix} \theta_1^{(k)} \\ \cdot \\ \cdot \\ \cdot \\ \theta_n^{(k)} \end{pmatrix} = \begin{pmatrix} \theta_1^{(k+1)} \\ \cdot \\ \cdot \\ \cdot \\ \theta_n^{(k+1)} \end{pmatrix}$$

Si la función de log verosimilitud es cóncava y unimodal, entonces la sucesión de iteraciones $\{\theta^{(k)}\}$ converge al estimador máximo verosímil θ^*; cuando la función de log verosimilitud no es cóncava, el método de Newton-Raphson no garantiza la convergencia desde cualquier punto inicial, como en el caso de una sola variable, este hecho también se extiende para el vector θ.

Si bien el método converge bastante rápido al estimador máximo verosímil (cuando se cumplen las condiciones adecuadas), un problema con este método es que en cada iteración se requiere del cálculo de la matriz de información (n x n dimensional) y la resolución de un sistema de n ecuaciones lineales, lo que llega a ser bastante caro en términos computacionales mientras la dimensión n se vuelve cada vez más grande.

En determinadas ocasiones este método llega a aproximarse a

puntos silla y a mínimos locales tanto como a los máximos, es decir, no tiene un comportamiento monótono (creciente o decreciente).

2.6.3 El método de puntuación (scoring) de Fisher.

El método de puntuación de Fisher es uno de los métodos de Newton-Raphson modificado, en donde la matriz de información observada $I(\theta^{(k)};x)$ para θ es reemplazada por $\mathcal{I}(\theta^{(k)})$, la que se conoce como la matriz de información esperada evaluada en $\theta^{(k)}$ o como se puede encontrar en la literatura acerca de inferencia estadística, la matriz de información de Fisher (Fisher 1922a), esta matriz mide el promedio de información sobre todas las posibles observaciones.

La matriz de información esperada está dada por:

$$\mathcal{I}(\theta^{(k)}) = - E_\theta \{ I(\theta^{(k)}; X)\}$$

donde la notación significa:

$$\begin{pmatrix} \mathcal{I}(\theta_1^{(k)}) \\ \\ \mathcal{I}(\theta_n^{(k)}) \end{pmatrix} = - \begin{pmatrix} E_\theta\left(\frac{\partial^2 l(\theta^{(k)})}{\partial\theta_1^{(k)}\partial\theta_1^{(k)}}\right) & \cdots & E_\theta\left(\frac{\partial^2 l(\theta^{(k)})}{\partial\theta_1^{(k)}\partial\theta_n^{(k)}}\right) \\ \vdots & & \vdots \\ E_\theta\left(\frac{\partial^2 l(\theta^{(k)})}{\partial\theta_n^{(k)}\partial\theta_1^{(k)}}\right) & \cdots & E_\theta\left(\frac{\partial^2 l(\theta^{(k)})}{\partial\theta_n^{(k)}\partial\theta_n^{(k)}}\right) \end{pmatrix}$$

Como comúnmente llega a ser muy difícil o tedioso calcular la matriz de información esperada para datos independientes e idénticamente distribuidos, se calcula la matriz de información empírica $I_e(\theta; x)$, la cual está dada por:

$$I_e(\theta ; x) = \sum_{j=1}^{n} \left[m(x_j;\theta)\, m^{t}(x_j;\theta) - n^{-1} M(x;\theta)\, M^{t}(x;\theta) \right]$$

evaluada en $\theta^{(k)}$, es decir,

$$I_e(\theta^{(k)}; x) = \sum_{j=1}^{n} \left[m(x_j;\theta^{(k)})\, m^{t}(x_j;\theta^{(k)}) - n^{-1} M(x;\theta^{(k)})\, M^{t}(x;\theta^{(k)}) \right]$$

(ver Mclachlan y Krishnan (1997))

donde $m(x_j;\theta)$ denota la función puntuación basada en una sola observación x_j de la muestra x, es decir:

$$m(x_j;\theta) = \frac{\partial \log f(x_j;\theta)}{\partial \theta} = \left(\frac{\partial \log f(x_j;\theta)}{\partial \theta_1}, \ldots, \frac{\partial \log f(x_j;\theta)}{\partial \theta_n} \right)$$

y $M(x;\theta)$ dada en (2.6), denotada en la literatura como la función puntuación de la muestra, se transforma en :

$$M(x;\theta) = \sum_{j=1}^{n} m(x_j;\theta)$$

la cual es la estadística de la función puntuación para la muestra completa $x = (x_1, \ldots, x_n)$. De tal modo que la matriz de información empírica resulta:

$$I_e(\theta; x) = \sum_{j=1}^{n} \left[m(x_j;\theta)\, m^{t}(x_j;\theta) \right] - \frac{1}{n} \left[\sum_{j=1}^{n} m(x_j;\theta) \right] \left[\sum_{j=1}^{n} m(x_j;\theta) \right]^{t}$$

para la k-ésima iteración se tendría

$$I_e(\theta^{(k)}; x) = \sum_{j=1}^{n} \left[m(x_j;\theta^{(k)})\, m^{t}(x_j;\theta^{(k)}) \right] - \frac{1}{n} \left[\sum_{j=1}^{n} m(x_j;\theta^{(k)}) \right] \left[\sum_{j=1}^{n} m(x_j;\theta^{(k)}) \right]^{t} \quad \ldots.(2.9)$$

Una vez que se evalúa en $\theta = \theta^{*}$, el estimador máximo verosímil, la matriz de información empírica se convierte en:

$$I_e(\theta^{*}; x) = \sum_{j=1}^{n} m(x_j;\theta^{*})\, m^{t}(x_j;\theta^{*})$$

Esto sucede debido a que θ^{*} es un cero de la función puntuación $m(x_j;\theta)$. De esta forma, (2.9) es la matriz de información que se utilizará en el lugar de $I(\theta^{(k)}; x)$ en la versión modificada

del método de Newton-Raphson, las iteraciones quedarían de la siguiente forma:

$$\theta^{(k+1)} = \theta^{(k)} + I_c^{-1}(\theta^{(k)}; x)\, M(x;\theta^{(k)})$$

$$= \theta^{(k)} + I_c^{-1}(\theta^{(k)}; x) \sum_{j=1}^{n} m(x_j;\theta^{(k)})$$

Otro método muy utilizado para estimar el valor θ^* es el método de Monte Carlo vía cadenas de Markov (ver Churchill (1992)). Este método consiste en construir una cadena de Markov con ciertas características; simular valores de esta cadena por determinado tiempo y utilizar los valores obtenidos al final para la obtención de un valor estimado para θ^*. Dentro de los métodos más conocidos se encuentran el método de Metropolis y el de Gibbs, métodos en los cuales no se entrará en detalles para efectos de esta tesis, no obstante por su importancia se mencionan algunas referencias para el lector interesado: Gamerman (1997), Gibbs et.al.(1996) y Ross (1996).

Como el lector recordará, este capítulo hizo mención a la inferencia de estimadores cuando se conocen todas las observaciones de una muestra, no obstante, en la práctica esto normalmente no ocurre, por el contrario se llega a carecer de información acerca de la muestra por diversas razones, éste es el hecho que se tratará de forma teórica en el capítulo siguiente.

Capítulo 3.

El algoritmo Esperanza-Maximización.

"Perdonad lo largo de esta carta, no tuve tiempo de escribir otra más corta"
Blasie Pascal (1623-1662)

3.1 Introducción.

Cuando se habla de las secuencias de A.D.N. y de los métodos de armado por lo general se maneja como supuesto que la secuencia que se observa es la verdadera; no obstante, en la mayor parte de los experimentos, los datos con que se trabajan no están completos, no se sabe cuál es la secuencia verdadera (datos faltantes) pero se tiene una estimación de cual es esta secuencia.

De este modo, deben existir métodos que ayuden a encontrar una estimación para la secuencia verdadera, suponiendo que la que es observada por el lector óptico pueda tener algún tipo de mutación, la cual puede ser causada por el proceso de lectura o por el proceso de decodificación. De este modo los datos con los cuales se trabaja son los datos observados (sucesión leída) y la secuencia verdadera (no observada o datos faltantes).

El presente capítulo trata de un algoritmo que es aplicable en diversas ramas de la estadística, ya que tiene como objetivo fundamental encontrar estimadores máximo verosímiles para los parámetros de una determinada distribución cuando se tienen datos incompletos, lo cual en la mayor parte de los experimentos, no necesariamente biológicos, sucede.

Se empezará dando una introducción a la teoría del algoritmo para posteriormente indicar algunos resultados importantes en relación a la convergencia del mismo.

3.2 Teoría del Algoritmo E.M.

El algoritmo Esperanza-Maximización (E.M.) fue utilizado y mencionado por primera vez en un artículo publicado en 1977 por Dempster, Laird y Rubin llamado "*Maximun Likelihood from Incomplete Data vía the E.M. Algorithm.*". Este algoritmo está enfocado a resolver problemas con datos incompletos en donde la estimación de máxima verosimilitud resulte problemática debido a la ausencia de parte de los datos colectados en la muestra. De esta forma el algoritmo asocia un problema de datos incompletos a uno de datos completos para el cual el cálculo del estimador máximo verosímil sea posible de encontrar de forma explícita o por medio de algún método de aproximación utilizando un paquete de computadora (ver Dempster, et. al. (1977)).

La asociación entre los problemas de datos incompletos y el de datos completos consiste en reformular el problema de los datos incompletos en términos del problema de datos completos. Por lo tanto se establece una relación entre sus respectivas verosimilitudes para trabajar directamente con el estimador máximo verosímil del problema de datos completos. De tal forma que al obtener el resultado para los datos completos también se obtendrá el resultado para los datos incompletos.

En cada iteración del algoritmo E.M., como se verá más adelante, se realizan dos pasos fundamentales llamados: paso Esperanza (E) y paso Maximización (M); los cuales, en adelante se denotan como E.M.

Si bien el algoritmo E.M. tiene una diversa variedad de ventajas como lo son su estabilidad numérica con respecto a otros métodos iterativos también hay que mencionar algunos problemas que deben ser resueltos aún; como lo son el caso de que dentro de el espacio muestral Θ exista más de un candidato a ser estimador máximo verosímil. De este modo el resultado final dependerá del valor inicial dado en el paso E para acercarse ya sea a uno o a otro

estimador. También es posible que no se converja a un máximo global, problema del que sufren todos los métodos iterativos de aproximación. Asimismo, el algoritmo puede llegar a converger lentamente en casos de problemas muy fáciles o bien en aquellos en donde se carezca de demasiada información.

Todos estos problemas pueden llegar a verse disminuidos por la estabilidad que ofrece el paso M al incrementar la verosimilitud en cada iteración (exceptuando puntos fijos). En condiciones óptimas se puede asegurar la convergencia global. En cierta forma, el trabajo analítico es más fácil, ya que sólo se requiere del cálculo de la esperanza condicional de la log-verosimilitud para los datos completos. Algo muy importante dentro de este estudio es que este método provee valores estimados de los datos faltantes.

Como se había mencionado anteriormente, el algoritmo E.M. resuelve el problema de la ecuación de verosimilitud de datos incompletos aplicando de forma iterativa el método de máxima verosimilitud para los datos completos de la siguiente forma: los datos no observables son reemplazados por su esperanza condicional dado los datos observados, y, utilizando una actualización del estimador máximo verosímil que se busca. De este modo se calcula el estimador máximo verosímil para el caso donde se tienen los datos completos y se repite el cálculo de esperanza con la nueva expresión del estimador máximo verosímil y se prosigue así sucesivamente.

Formalmente se puede escribir lo siguiente. Denótese a $\boldsymbol{Y}$ el vector aleatorio que corresponde a los datos observados (que pueden ser incompletos) y θ el vector de los parámetros desconocidos que forman el modelo considerado. Sea Θ el espacio paramétrico de θ. Denote por $f(\cdot;\theta)$ la función de densidad de $\boldsymbol{Y}$ bajo θ. Sea $\boldsymbol{X}$ el vector aleatorio correspondiente al vector de lo que serían los datos completos, es decir, este vector esta formado por los datos observados $\boldsymbol{Y}$ y las variables aleatorias que corresponden a los datos faltantes. Indique por $f_c(\cdot;\theta)$ la densidad de $\boldsymbol{X}$ bajo θ.

Sea $f_c(\cdot;\theta)$ la función de verosimilitud para los datos completos, es decir, la verosimilitud que se podría obtener si se

pudiesen observar completamente los datos. De este modo se tiene que $L_c(\theta) \propto f_c(\boldsymbol{x}; \theta)$ es la función de verosimilitud para θ formada con base a los datos completos $\boldsymbol{x}$. Sin embargo, como se mencionó en el capítulo anterior se puede tomar $L_c(\theta) = f_c(\boldsymbol{x}; \theta)$ y de ahora en adelante se utilizará este hecho. Así se tiene que

$$L(\theta) = f(\boldsymbol{y}; \theta) \qquad \text{y} \qquad l(\theta) = \log L(\theta)$$

$$L_c(\theta) = f_c(\boldsymbol{x}; \theta) \qquad \text{y} \qquad l_c(\theta) = \log L_c(\theta)$$

De este modo, se tienen dos espacios muestrales $\boldsymbol{Y}$ y $\boldsymbol{X}$, donde el espacio $\boldsymbol{X}$ corresponde al espacio de los datos completos y $\boldsymbol{Y}$ es el espacio de los datos incompletos. Se tiene también un mapeo $\mathbf{h}$ que asocia un elemento de $\boldsymbol{X}$ con un elemento de $\boldsymbol{Y}$, de tal modo que el conjunto $\mathbf{h}_y$ viene siendo el conjunto de los elementos $\boldsymbol{x} \in \boldsymbol{X}$ tal que $\boldsymbol{y} = \mathrm{h}(\boldsymbol{x})$, es decir $\mathbf{h}_y$ es el conjunto $\{\boldsymbol{x} \in \boldsymbol{X} \mid \boldsymbol{y} = \mathrm{h}(\boldsymbol{x})\}$.

Por ejemplo, suponga que la dimensión de $\boldsymbol{Y}$ y $\boldsymbol{X}$ son 4 y 5 respectivamente, entonces $\mathbf{h}_y = \{\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5) = (x_1, x_2, y_2, y_3, y_4) \in \boldsymbol{X} : \boldsymbol{y} = (y_1, y_2, y_3, y_4) = \mathbf{h}(\boldsymbol{x})\}$ y donde $y_1 = x_1 + x_2$, entonces $\mathbf{h}(\boldsymbol{x}) = (x_1 + x_2, y_2, y_3, y_4)$.

De donde se obtiene que la función de densidad de $\boldsymbol{x}$ bajo θ es igual que la integral sobre $\mathbf{h}_y$ de la función de densidad de los datos completos. De este modo se tiene: para $f_c(\boldsymbol{x}; \theta) = L_c(\theta)$ y $f(\boldsymbol{y}; \theta) = L(\theta)$ que

$$f(\boldsymbol{y};\theta) = \int_{h_y} f_c(\boldsymbol{x};\theta)\, d\boldsymbol{x}$$

De este modo el valor de θ que maximiza $f_c(\boldsymbol{x}; \theta)$ es el mismo que maximiza $f(\boldsymbol{y}; \theta)$, ya que aunque tengan dimensión distinta, bajo $\mathbf{h}_y$ se cumple que $f_c(\boldsymbol{x}; \theta)$ se mapea en $f(\boldsymbol{y}; \theta)$. Como se puede notar, la condición anterior es inmediata cuando se trabaja con transformaciones sencillas, como lo pueden ser la suma de dos variables (ver sección 4.1), y aún si se trabaja con transformaciones mas complicadas la condición se sigue cumpliendo, aunque el trabajo analítico se dificulte un poco mas.

Ahora se describe el funcionamiento del algoritmo E.M.:

Primera iteración.

Paso E.

Tome $\theta^{(0)}$ un valor inicial para el parámetro θ. En la primera iteración del paso E se requiere del cálculo de la esperanza condicional de la función $l_c(\theta) = \log f_c(x;\theta)$ dados los datos observados y, es decir se debe obtener:

$$G(\theta;\theta^{(0)}) = E_\theta^{(0)} \{ l_c(\theta) | y \} = E_\theta^{(0)} \{ \log f_c(X;\theta) | y \}$$

donde $E_\theta^{(0)} \{ l_c(\theta) | y \}$ es la esperanza de $l_c(\theta)$ dado los datos observados y y evaluada en $\theta = \theta^{(0)}$ donde X es el vector aleatorio representando a los datos completos, es decir, parte de X son los datos observados y la otra parte son las variables aleatorias representando los datos faltantes.

Note que en este paso se obtendrá el valor esperado bajo $\theta^{(0)}$ de cantidades que involucran las variables aleatorias representando los datos faltantes.

Paso M.

El siguiente paso corresponde al paso M donde se buscará en Θ el valor del parámetro θ que maximice $G(\theta;\theta^{(0)})$, esta θ será llamada $\theta^{(1)}$ y que cumple con lo siguiente:

$$G(\theta^{(1)};\theta^{(0)}) \geq G(\theta;\theta^{(0)}) \ ; \text{para todo } \theta \in \Theta$$

De esta forma se procede iterativamente regresando al paso E con el nuevo valor $\theta^{(1)}$ en lugar de $\theta^{(0)}$. Note que en este paso se obtiene el estimador máximo verosímil de θ utilizando los datos observados y la esperanza de las expresiones que involucran las variables aleatorias representando los datos faltantes.

En la **n-ésima iteración**, los pasos E y M se ven de la siguiente forma:

Paso E.
Se calcula $G(\theta;\theta^{(n-1)})$ donde

$$G(\theta;\theta^{(n-1)}) = E_{\theta^{(n-1)}} \{ l_c(\theta) | y \} = E_{\theta^{(n-1)}} \{ \log f_c(X|\theta) | y \} \quad ....(3.1)$$

Paso M.
Tomar $\theta^{(n)}$ como el θ en el espacio de los parámetros Θ que maximice $G(\theta;\theta^{(n-1)})$, es decir:

$$G(\theta^{(n)};\theta^{(n-1)}) \geq G(\theta;\theta^{(n-1)}) \ ; \text{para todo } \theta \in \Theta \quad .....(3.2)$$

o en otras palabras $\theta^{(n)}$ pertenece al conjunto

$$K(\theta^{(n-1)}) = \{\arg\max_{\theta\in\Theta} G(\theta;\theta^{(n-1)})\}$$
$$= \{ \theta \in \Theta : \text{maximiza } G(\theta;\theta^{(n-1)})\}$$

que es el conjunto de puntos θ que maximizan a $G(\theta;\theta^{(n-1)})$.

Se verá más adelante, en la sección 3.5 que la condición (3.2) es suficiente para que $l(\theta^{(n)}) \geq l(\theta^{(n-1)})$.

3.3 El algoritmo E.M. generalizado.

Note que algunas veces puede no ser tan fácil encontrar el valor de θ que maximiza de forma global a la función $G(\theta;\theta^{(n)})$. De este modo Dempster, Laird y Rubin (1977) presentan una generalización del algoritmo E.M. donde a cada iteración n para el paso M se requiere que el nuevo parámetro $\theta^{(n)}$ sea seleccionado de tal forma que cumpla:

$$G(\theta^{(n)};\theta^{(n-1)}) \geq G(\theta^{(n-1)};\theta^{(n-1)}) \quad .....(3.3)$$

De este modo basta encontrar el valor $\theta \in \Theta$ tal que el valor de la función $G(\cdot\ ;\theta^{(n-1)})$ evaluada en este θ es mayor o igual que el valor de la función evaluada en $\theta^{(n-1)}$. Esto facilita mucho los cálculos porque ya no se tiene que conseguir un máximo global para $G(\cdot\ ;\theta^{(n-1)})$.

Observación: Tanto en el algoritmo E.M. como en el algoritmo E.M. generalizado se puede imponer la condición de paro: $l(\theta^{(n)}) - l(\theta^{(n-1)}) < \varepsilon$ o bien $|\theta^{(n-1)} - \theta^{(n)}| < \varepsilon$ para un $\varepsilon > 0$ fijado previamente, que dependerá del grado de acercamiento que el investigador desee al estimador máximo verosímil, ya que, como más adelante se mostrará, se cumple la monotonicidad en la función de log-verosimilitud y también la convergencia para sucesiones de estimadores.

3.4 Monotonicidad del algoritmo E.M.

Parte fundamental del trabajo que se realiza en el algoritmo es el de asegurar que la verosimilitud no decrece, ya que únicamente se estaría divergiendo de los estimadores máximo verosímiles que se buscan, de este modo se debe asegurar que la diferencia $l(\theta^{(n+1)}) - l(\theta^{(n)}) \geq 0$ siempre se cumpla. En otras palabras que la log-verosimilitud para los datos incompletos no decrece después de las iteraciones realizadas por medio del algoritmo E.M. y por la monotonicidad de la función logarítmica se cumple que $L(\theta^{(n+1)}) \geq L(\theta^{(n)})$.

Teorema 3.1 Cuando se usa el algoritmo E.M. o la versión generalizada del algoritmo E.M. se tiene que para n = 0, 1, 2, ... la función de log-verosimilitud $l(\theta)$ de los datos observados es tal que $l(\theta^{(n+1)}) \geq l(\theta^{(n)})$.

Demostración:
(a) Algoritmo E.M.

Sea $g(\boldsymbol{x} \mid \boldsymbol{y}; \theta)$ la densidad condicional de los datos completos $\boldsymbol{x}$ dado los datos observados $\boldsymbol{y}$ entonces para $f_c(\boldsymbol{x}; \theta) = L_c(\theta)$ y $f(\boldsymbol{y}; \theta) = L(\theta)$. Se tiene que:

$$g(\boldsymbol{x} \mid \boldsymbol{y};\theta) = \frac{f_c(\boldsymbol{x};\theta)}{f(\boldsymbol{y};\theta)} = \frac{L_c(\theta)}{L(\theta)}$$

entonces para $g(\boldsymbol{x} \mid \boldsymbol{y};\theta)$ se puede escribir la log-verosimilitud de los datos incompletos de la siguiente forma:

$$l(\theta) = \log L(\theta) = \log\left(\frac{L_c(\theta)}{g(x/y;\theta)}\right) = \log L_c(\theta) - \log(g(\boldsymbol{x} \mid \boldsymbol{y};\theta))$$

Tomando la esperanza a la ecuación anterior, condicionando sobre los datos observados $\boldsymbol{y}$ y evaluando en $\theta^{(n)}$ se tiene

$$l(\theta) = E_{\theta^{(n)}}\{l_c(\theta) \mid \boldsymbol{y}\} - E_{\theta^{(n)}}\{\log(g(\boldsymbol{X} \mid \boldsymbol{y};\theta)) \mid \boldsymbol{y}\}$$

$$= G(\theta;\theta^{(n)}) - E_{\theta^{(n)}}\{\log(g(\boldsymbol{X} \mid \boldsymbol{y};\theta)) \mid \boldsymbol{y}\}$$

por definición de $G(\theta;\theta^{(n)})$

Considere la siguiente notación:

$$H(\theta;\theta^{(n)}) = E_{\theta^{(n)}}\{\log(g(\boldsymbol{X} \mid \boldsymbol{y};\theta)) \mid \boldsymbol{y}\}$$

De este modo se puede escribir:

$$l(\theta) = G(\theta;\theta^{(n)}) - H(\theta;\theta^{(n)})$$

En la (n+1)-ésima iteración del paso M del algoritmo E.M. se toma $\theta^{(n+1)} \in \Theta$ de modo que $G(\theta^{(n+1)};\theta^{(n)}) \geq G(\theta;\theta^{(n)})$ para todo $\theta \in \Theta$. De este modo, se tiene que $G(\theta^{(n+1)};\theta^{(n)}) - G(\theta^{(n)};\theta^{(n)}) \geq 0$ para todo $n = 0, 1, 2, \ldots$

Note que:

$$l(\theta^{(n+1)}) = [G(\theta^{(n+1)};\theta^{(n)}) - H(\theta^{(n+1)};\theta^{(n)})]$$

y que

$$l(\theta^{(n)}) = [G(\theta^{(n)};\theta^{(n)}) - H(\theta^{(n)};\theta^{(n)})]$$

de tal forma que

$$l(\theta^{(n+1)}) - l(\theta^{(n)}) = [G(\theta^{(n+1)};\theta^{(n)}) - H(\theta^{(n+1)};\theta^{(n)})] - [G(\theta^{(n)};\theta^{(n)}) - H(\theta^{(n)};\theta^{(n)})]$$

$$= [G(\theta^{(n+1)};\theta^{(n)}) - G(\theta^{(n)};\theta^{(n)})] - [H(\theta^{(n+1)};\theta^{(n)}) - H(\theta^{(n)};\theta^{(n)})].$$

Dado que la función logarítmica es creciente y $G(\theta^{(n+1)};\theta^{(n)}) - G(\theta^{(n)};\theta^{(n)}) \geq 0$ para todo n = 0, 1, 2, ..., para mostrar que $l(\theta^{(n+1)}) - l(\theta^{(n)}) \geq 0$ para todo n = 0, 1, 2, ... , se tiene que mostrar que :

$$- [H(\theta^{(n+1)};\theta^{(n)}) - H(\theta^{(n)};\theta^{(n)})] \geq 0$$

o bien

$$[H(\theta^{(n+1)};\theta^{(n)}) - H(\theta^{(n)};\theta^{(n)})] \leq 0 \qquad \text{.....(3.4)}$$

Por la definición de $H(\cdot\,;\cdot)$ se tiene que para todo $\theta \in \Theta$ sucede que:

$$H(\theta^{(n+1)};\theta^{(n)}) - H(\theta^{(n)};\theta^{(n)}) =$$

$$= E_{\theta}{}^{(n)} \{\log (g(\boldsymbol{X} \mid \boldsymbol{y};\theta^{(n+1)})) \mid \boldsymbol{y}\} - E_{\theta}{}^{(n)} \{\log (g(\boldsymbol{X} \mid \boldsymbol{y};\theta^{(n)})) \mid \boldsymbol{y}\}$$

$$= E_{\theta}{}^{(n)} \{[\log (g(\boldsymbol{X} \mid \boldsymbol{y};\theta^{(n+1)})) - \log (g(\boldsymbol{X} \mid \boldsymbol{y};\theta^{(n)}))] \mid \boldsymbol{y}\}$$

$$= E_{\theta^{(n)}}\left[\log\left(\frac{g(X / y;\theta^{(n+1)})}{g(X / y;\theta^{(n)})}\right) \Big/ y\right] \qquad \text{.....(3.5)}$$

Como la función logarítmica es cóncava se puede utilizar la desigualdad de Jensen[17] y se tiene para (3.5) que :

$$E_{\theta^{(n)}}\left[\log\left(\frac{g(X / y;\theta^{(n+1)})}{g(X / y;\theta^{(n)})}\right) \Big/ y\right] \leq \log\left[E_{\theta^{(n)}}\left(\frac{g(X / y;\theta^{(n+1)})}{g(X / y;\theta^{(n)})}\right) \Big/ y\right]$$

y de donde:

$$\log\left[E_{\theta^{(n)}}\left(\frac{g(X / y;\theta^{(n+1)})}{g(X / y;\theta^{(n)})}\right) \Big/ y\right] = \log\left[\int_{h_y} \frac{g(x / y;\theta^{(n+1)})g(x / y;\theta^{(n)})}{g(x / y;\theta^{(n)})}\, dx\right] \text{....(3.6)}$$

[17] Ver *Apéndice A.2.*

dado que por definición

$$E_{\theta^{(n)}}\left(\frac{g(X/y;\theta^{(n+1)})}{g(X/y;\theta^{(n)})}/y\right)$$

es la esperanza de $g(X \mid y;\theta^{(n+1)}) / g(X \mid y;\theta^{(n)})$ con respecto a la función de densidad $g(x \mid y;\theta)$ evaluada en $\theta = \theta^{(n)}$.

De (3.6) se tiene que

$$\log\left[\int_{h_y} \frac{g(x/y;\theta^{(n+1)})g(x/y;\theta^{(n)})}{g(x/y;\theta^{(n)})}dx\right] = \log\left[\int_{h_y} g(x/y;\theta^{(n+1)})dx\right] = \log(1) = 0$$

ya que se integra sobre todos los datos x tales que $y = h(x)$ y g es la densidad condicional de X dado $Y = y$. De este modo dado que:

$$[H(\theta^{(n+1)};\theta^{(n)}) - H(\theta^{(n)};\theta^{(n)})]$$
$$\leq \log\left[\int_{h_y} g(x/y;\theta^{(n+1)})dx\right]$$

se tiene que

$$[\, H(\theta^{(n+1)};\theta^{(n)}) - H(\theta^{(n)};\theta^{(n)}) \,] \geq 0$$

Entonces se cumple (3.6) y de este modo

$$l(\theta^{(n+1)}) - l(\theta^{(n)}) = [G(\theta^{(n+1)};\theta^{(n)}) - G(\theta^{(n)};\theta^{(n)})] - [H(\theta^{(n+1)};\theta^{(n)}) - H(\theta^{(n)};\theta^{(n)})] \geq 0.$$

es decir, $l(\theta^{(n+1)}) \geq l(\theta^{(n)})$ que es lo que se quería demostrar.

Ahora bien, por la monotonicidad de la función logarítmica se tiene $L(\theta^{(n+1)}) \geq L(\theta^{(n)})$ para todo n = 0, 1, 2, ... De este modo $L(\theta^{(n)})$ es monótona creciente en n.

(b)Versión generalizada del algoritmo E-M.

De (a) se ve que

$$l(\theta^{(n+1)}) - l(\theta^{(n)}) = [G(\theta^{(n+1)};\theta^{(n)}) - G(\theta^{(n)};\theta^{(n)})] - [H(\theta^{(n+1)};\theta^{(n)}) - H(\theta^{(n)};\theta^{(n)})]$$

también de (a) se ve que $H(\theta^{(n+1)};\theta^{(n)}) - H(\theta^{(n)};\theta^{(n)}) \leq 0$ y de la definición del algoritmo E.M. generalizado se tiene que $G(\theta^{(n+1)};\theta^{(n)}) \geq G(\theta^{(n)};\theta^{(n)})$.

De este modo se tiene que $l(\theta^{(n+1)}) - l(\theta^{(n)}) \geq 0$ y por la monotonicidad de la función logarítmica se tiene que $L(\theta^{(n+1)}) \geq L(\theta^{(n)})$ para n = 1, 2, ...

■

3.5 Condiciones de regularidad de Wu (1983).

Ya se ha visto que la (n+1)-ésima iteración del paso M del algoritmo E.M. consiste en seleccionar un elemento de $K(\theta^{(n)})$, de la misma forma en el algoritmo E.M. generalizado el conjunto $K(\theta^{(n)})$ debe ser tomado de tal forma que:

$$K(\theta^{(n)}) = \{ \theta \in \Theta; G(\theta ;\theta^{(n)}) \geq G(\theta^{(n)} ;\theta^{(n)})\}.$$

Note que $K(\cdot)$ es una función que asocia un punto a un conjunto, ya que pueden existir uno o más parámetros $\theta \in \Theta$ que maximicen a $G(\theta ;\theta^{(n)})$.

En esta sección se presentan condiciones que Wu (1983) establece y que aseguran la convergencia de la sucesión de valores de la log-verosimilitud $\{l(\theta^{(n)})\}$ a un valor estacionario[18] de $l(\theta)$. Las suposiciones de Wu (1983) son las siguientes:

1. Θ es un subconjunto en el espacio euclideano n-dimensional R^n.

2. Sea $\Theta_{\theta^{(0)}} = \{\theta \in \Theta: l(\theta) \geq l(\theta^{(0)})\}$, supóngase que $\Theta_{\theta^{(0)}}$ es compacto para cualquier $l(\theta^{(0)}) > -\infty$

[18] **Definición 3.1** Sea f una función de n variables, el punto en donde el vector gradiente de f es el vector cero o no existe se llama punto crítico o estacionario de f.

3. $L(\theta)$ es continua en Θ y diferenciable en el interior de Θ.

4. Para $\theta^{(n+1)}$ una solución de la ecuación:

$$\frac{\partial G(\theta ; \theta^{(n)})}{\partial \theta} = 0$$

$\theta^{(n+1)}$ está en el interior de Θ.

Observaciones.

(i) De las condiciones 1,2 y 3 se obtiene que cualquier sucesión $\{L(\theta^{(n)})\}$ está acotada por arriba para cualquier $\theta^{(0)} \in \Theta$ donde se asume que el punto $\theta^{(0)}$ satisface que $L(\theta^{(0)}) > -\infty$.

(ii) La condición 4 se da por hecho si en determinado caso $\Theta_{\theta^{(0)}}$ está en el interior de Θ, para todo $\theta^{(0)} \in \Theta$.

3.6 Teoremas de convergencia para las sucesiones E.M.

En esta sección se mencionará el teorema principal de convergencia para una sucesión de parámetros generada por el algoritmo E.M. Note que para el caso del algoritmo E.M. generalizado valen resultados similares. La demostración de este teorema de convergencia se basa en otro teorema más general que se encuentra en Zangwill (1969). La demostración para la sucesión E.M. se encuentra en el artículo de Wu (1983). Para efectos de esta tesis no se presentará la demostración del teorema de Wu, pero puede ser encontrada en detalle en Wu (1983) pags. 97-98.

Teorema 3.2 (Wu(1983) y Zangwill(1969)). Sea $\mathbf{C}$ un conjunto cualquiera, $\mathbf{M}(\cdot)$ una aplicación que asocia un punto de $\mathbf{C}$ a un subconjunto de $\mathbf{C}$, y $\{\mathbf{x}_k\}_{k\geq 0}$ una sucesión generada por $\mathbf{x}_{k+1} \in \mathbf{M}(\mathbf{x}_k)$. Considere un conjunto $\Gamma \subset \mathbf{C}$, tal que

(i) todos los puntos $\mathbf{x}_k$ están en un conjunto compacto $\mathbf{B} \subset \mathbf{C}$

(ii) **M** es cerrado sobre el complemento de Γ, es decir, puede pasar que $\mathbf{M(x')} \cap \Gamma^c \neq \emptyset$ o bien $\mathbf{M(x')} \cap \Gamma \neq \emptyset$ entonces $\mathbf{M(x')} \cap \Gamma^c$ = D(cerrado).

(iii) Existe una función continua $\alpha(\cdot)$ en **C** tal que
(a) Si $\mathbf{x} \notin \Gamma$, $\alpha(\mathbf{y}) > \alpha(\mathbf{x})$ para todo $\mathbf{y} \in \mathbf{M(x)}$
(b) Si $\mathbf{x} \in \Gamma$, $\alpha(\mathbf{y}) \geq \alpha(\mathbf{x})$ para todo $\mathbf{y} \in \mathbf{M(x)}$

Entonces todos los puntos límite de $\{\mathbf{x_k}\}_{k\geq 0}$ están en el conjunto Γ y $\alpha(\mathbf{x_k})$ converge de forma monótona a $\alpha(\mathbf{x})$ para algún $\mathbf{x} \in \Gamma$.

Trabajando con la notación que ya se ha definido anteriormente, el **Teorema 3.2**, para el caso que interesa se convierte en el teorema siguiente

Teorema 3.3 Sea Θ el espacio paramétrico, $K(\theta') = \{\theta \in \Theta$: maximiza $G(\theta;\theta')\}$, que es el conjunto de puntos θ que maximizan a $G(\theta;\theta')$, y $\{\theta^{(k)}\}_{k\geq 0}$ una sucesión generada por $\theta^{(n+1)} \in K\{\theta^{(n)}\}$, es decir, si se tiene $\theta^{(1)}$, ..., $\theta^{(n)}$, entonces $\theta^{(n+1)}$ será elemento de $K\{\theta^{(n)}\}$. Considere el conjunto

$$S = \left\{ \theta'' \in \text{int}\,\Theta \,;\, \left. \frac{\partial l(\theta)}{\partial \theta} \right|_{\theta = \theta''} = 0 \right\}$$

el conjunto de puntos estacionarios de $l(\cdot)$ que están en el interior de Θ de $l(\cdot)$ y $\Theta_{\theta^{(0)}} = \{\theta \in \Theta \,;\, l(\theta) \geq l(\theta^{(0)})\}$, suponga que

(i) $\Theta_{\theta^{(0)}}$ es compacto.

(ii) todos los puntos $\theta^{(n)}$ están contenidos en $\Theta_{\theta^{(0)}}$.

(iii) $\mathbf{K}(\cdot)$ es cerrado sobre el complemento de S, $\mathbf{K}(\theta^+) \cap S^c \neq \emptyset$ o bien $\mathbf{K}(\theta^+) \cap S \neq \emptyset$ entonces $\mathbf{K}(\theta^+) \cap S^c$ = D(cerrado).

(iv) La función de logverosimilitud $l(\cdot)$ es tal que
(a) Si $\theta^* \notin S$, $l(\theta') > l(\theta^*)$ para todo $\theta' \in K(\theta^*)$
(b) Si $\theta^* \in S$, $l(\theta') \geq l(\theta^*)$ para todo $\theta' \in K(\theta^*)$

Entonces todos los puntos límite de $\{\theta^{(n)}\}_{n\geq 0}$ están en el conjunto S y $l(\theta^{(n)})$ converge de forma monótona a $l(\theta^*)$ para algún $\theta^* \in S$.

El teorema anterior es un teorema que de forma teórica asegura la convergencia de las iteraciones en el algoritmo E.M. y del algoritmo E.M. generalizado, no obstante, para casos prácticos se nota como más adecuado el teorema que a continuación se menciona aunque las condiciones de regularidad se siguen conservando.

Teorema 3.4 (Wu (1983) Teorema Principal de Convergencia) Sean $\{\theta^{(n)}\}$ una sucesión de parámetros producidos por las iteraciones del algoritmo E.M. generado por $\theta^{(n+1)} \in K(\theta^{(n)})$ y S el conjunto de puntos estacionarios de $l(\theta)$ que están en el interior de Θ. Si

(i) $K(\theta^{(n)})$ es cerrado sobre el complemento de S.

(ii) $l(\theta^{(n+1)}) > l(\theta^{(n)})$ para toda $\theta^{(n)} \notin S$.

Entonces bajo las condiciones de regularidad de Wu todos los puntos límite de $\{\theta^{(n)}\}$ son puntos estacionarios de $l(\theta)$ y $l(\theta^{(n)})$ converge de forma monótona a $l^* = l(\theta^*)$ para algún punto estacionario $\theta^* \in S$.

Hay que mencionar que es necesario que $G(\theta;\psi)$ sea continua en θ y ψ para que el mapeo de la sucesión E.M. sea cerrado.

Proposicion 3.1 (Wu (1983)) Si $G(\theta;\psi)$ es continua en θ y ψ entonces $K(\cdot)$ es cerrado sobre el complemento de S.

El siguiente corolario es un caso particular del **Teorema 3.4**

Corolario 3.1 Bajo las condiciones de regularidad de Wu, supóngase que $G(\theta;\psi)$ satisface la condición de continuidad sobre θ

y ψ. Entonces todos los puntos límite de cualquier sucesión $\{\theta^{(n)}\}$ del algoritmo E.M. son puntos estacionarios de $l(\theta)$, y $l(\theta^{(n)})$ converge de forma monótona a algún valor $l^* = l(\theta^*)$ para algún punto estacionario θ^*.

Idea de la demostración.

De la **Proposicion 3.1** se tiene que la condición ***(i)*** del **Teorema 3.4** se satisface, es decir $K(\theta^{(n)})$ es cerrado sobre el complemento de S. Por otro lado de la dinámica del algoritmo E.M. se tiene que la condición ***(ii)*** del **Teorema 3.4** se satisface, es decir $l(\theta^{(n+1)}) > l(\theta^{(n)})$ para toda $\theta^{(n)} \notin S$. Por lo tanto, se aplica directamente el **Teorema 3.4** para concluir que, todos los puntos límite de cualquier sucesión $\{\theta^{(n)}\}$ del algoritmo E.M. son puntos estacionarios de $l(\theta)$ y $l(\theta^{(n)})$ converge de forma monótona a $l^* = l(\theta^*)$ para algún punto estacionario $\theta^* \in S$.

■

A continuación se presenta el enunciado y la idea de la demostración de un teorema que tienen que ver con la convergencia de la sucesión de parámetros $\{\theta^{(n)}\}$ a algún punto θ^*. Defina $S(a) \subset S$ de la siguiente forma $S(a) = \{\theta \in \Theta: l(\theta) = a\}$.

Teorema 3.5 Sea $\{\theta^{(n)}\}$ una sucesión del algoritmo E.M. generalizado generado por $\theta^{(n+1)} \in K(\theta^{(n)})$ y satisfaciendo:

(i) $K(\theta^{(n)})$ cerrado sobre el complemento de S.

(ii) $l(\theta^{(n+1)}) > l(\theta^{(n)})$ para todo $\theta^{(n)} \notin S$.

Supóngase que $S(l^*)$ consiste en un sólo punto θ^*, donde l^* es el límite de $l(\theta^{(n)})$. Entonces $\theta^{(n)}$ converge a θ^*.

Idea de la demostración.

Note que las condiciones ***(i)*** y ***(ii)*** corresponden a las condiciones ***(i)*** y ***(ii)*** del **Teorema 3.4**. De este modo se tiene que todos los puntos límite de $\{\theta^{(n)}\}$ son estacionarios para $l(\theta)$ y $l(\theta^{(n)})$ converge de forma monótona para $l^* = l(\theta^{(n)})$ para algún punto estacionario $\hat{\boldsymbol{\theta}}$. Pero como $S(l^*) = \{\theta^*\}$, pasa que $\hat{\boldsymbol{\theta}} = \theta^*$, (esto quiere decir que no pueden haber dos puntos estacionarios

diferentes con el mismo valor l^*), y dado que todos los puntos límite de $\{\theta^{(n)}\}_{n\geq 0}$ son estacionarios se tiene que $\lim_{n\to\infty} \theta^{(n)} = \theta^*$.

■

3.7 Convergencia a un valor estacionario de una sucesión $\{l(\theta^{(n)})\}$ con n = 0, 1, 2, ... obtenida por el algoritmo E.M.

Parte fundamental de la teoría del algoritmo E.M. es el hecho de poder asegurar que después de un cierto número de iteraciones realizadas se tenga la convergencia a un valor específico, lo cual se trató en la sección anterior. En determinadas ocasiones el valor al que se llega no es el estimador máximo global que se busca, sino uno local, lo cual dependerá la mayor parte de las veces del valor inicial $\theta^{(0)}$ que se de para empezar a iterar. Asimismo, hay que considerar la posibilidad de que la función de verosimilitud tenga un punto silla y que la sucesión converja a éste.

Teorema 3.6 Toda sucesión monótona acotada converge.

Demostración
Tome una sucesión $\{\theta^{(n)}\}$ acotada y creciente. Entonces el conjunto de términos de la sucesión tiene una cota superior y por la propiedad de completés esta sucesión tiene a τ su cota superior más chica, se tendrá que demostrar que $\{\theta^{(n)}\} \to \tau$. Tome $\varepsilon > 0$, entonces $\tau - \varepsilon < \tau$, de tal modo que $\tau - \varepsilon$ no es una cota superior de $\{\theta^{(n)}\}$. Entonces, algún término $\theta^{(n)}$ es más grande que $\tau - \varepsilon$ y como $\{\theta^{(n)}\}$ es creciente, $\theta^{(n)} > \tau - \varepsilon$ para todo $n > k$ y dado que $\{\theta^{(n)}\}$ es acotada superiormente por τ se tiene que $\theta^{(n)} \leq \tau < \tau+\varepsilon$ para todo n. De tal modo, $|\theta^{(n)} - \tau| < \varepsilon$ para $n > k$, y por tanto $\theta^{(n)} \to \tau$. De forma análoga, si $\{\theta^{(n)}\}$ es acotada y decreciente, $\theta^{(n)}$ converge a su cota inferior más grande (ver Garner (1988)).

■

En la sección anterior se vio que $\{L(\theta^{(n)})\}_{n\geq 0}$ es una sucesión creciente, por lo tanto si $\{L(\theta^{(n)})\}_{n\geq 0}$ es una sucesión acotada superiormente, entonces $L(\theta^{(n)})$ converge a algún valor l^* donde $l^* = L(\theta^*)$ para algún punto θ^* que cumpla con que

$$\left.\frac{\partial L(\theta)}{\partial \theta}\right|_{\theta=\theta^*} = 0 \qquad \text{o bien} \qquad \left.\frac{\partial \log L(\theta)}{\partial \theta}\right|_{\theta=\theta^*} = 0$$

El valor θ^* es por lo general un máximo local o bien un punto silla, aunque de cualquier forma, existen casos en que un pequeña perturbación de θ causará que el algoritmo diverja cuando el punto al que converge la sucesión no sea máximo global de $L(\theta)$, que sea por ejemplo, un punto de inflexión o punto silla.

Por lo general sucede que $L(\theta)$ tiene varios puntos estacionarios y la convergencia de la sucesión $\{\theta^{(n)}\}$ dependerá del punto inicial $\theta^{(0)}$; y por ejemplo, en el caso óptimo en que en el espacio paramétrico sólo exista un estimador máximo verosímil cualquier sucesión va a converger a él sin tener que depender del valor inicial que se tome.

Proposición 3.2 Si $\tilde{\sigma}$ es un punto silla para $L(\theta)$, $\theta \in \Theta$ entonces la sucesión $\{\theta^{(n)}\}_{n\geq 0}$ puede converger para el punto silla $\tilde{\sigma}$, si $\tilde{\sigma}$ maximiza globalmente $H(\theta;\tilde{\sigma})$, $\theta \in \Theta$, es decir, $\{L(\theta^{(n)})\}_{n\geq 0}$ puede converger para $L(\tilde{\sigma})$.

Demostración.
De la demostración del **Teorema 3.1** se tiene que

$$l(\theta) = G(\theta;\theta^{(n)}) - H(\theta;\theta^{(n)}) \qquad \text{.....(3.7)}$$

donde $\theta^{(n)}$ es el valor de θ obtenido en la iteración n-1 del paso E del algoritmo E.M. tal que cumple con lo siguiente: sea $\theta^{(n)}$ tal que $\theta^{(n)}$ maximiza globalmente $H(\theta;\theta^{(n)})$, es decir $H(\theta;\theta^{(n)}) \leq H(\theta^{(n)};\theta^{(n)})$ para todo $\theta \in \Theta$.

Derivando (3.7) con respecto a θ se tiene:

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\partial G(\theta;\theta^{(n)})}{\partial \theta} - \frac{\partial H(\theta;\theta^{(n)})}{\partial \theta}$$

De este modo por definición de punto máximo:

$$\frac{\partial H(\theta;\theta^{(n)})}{\partial \theta}\Bigg|_{\theta=\theta^{(n)}} = 0 \qquad \text{.....(3.8)}$$

Dado que $\theta^{(n)} \in \Theta$, entonces $\theta^{(n)} = \theta'$ para algún $\theta' \in \Theta$. De este modo

$$\frac{\partial l(\theta)}{\partial \theta}\Bigg|_{\theta=\theta'} = \frac{\partial G(\theta;\theta')}{\partial \theta}\Bigg|_{\theta=\theta'} - \frac{\partial H(\theta;\theta')}{\partial \theta}\Bigg|_{\theta=\theta'}$$

$$= \frac{\partial G(\theta;\theta')}{\partial \theta}\Bigg|_{\theta=\theta'} \qquad \text{.....(3.9)}$$

Sea $\tilde{\sigma}$ un punto estacionario de $L(\theta)$, entonces por definición de punto estacionario y por la derivada del logaritmo se tiene:

$$\left[\frac{1}{L(\theta)}\frac{d\,L(\theta)}{d\theta}\right]\Bigg|_{\theta=\delta} = \frac{1}{L(\tilde{\sigma})}\left[\frac{d\,L(\theta)}{d\theta}\right]\Bigg|_{\theta=\delta} = 0 \qquad \text{....(3.10)}$$

de (3.8) y (3.9) se nota que si θ' es un punto silla para $L(\theta)$ entonces por definición θ' es un punto estacionario para $L(\theta)$, y se tiene de (3.10) que

$$\frac{\partial l(\theta)}{\partial \theta}\Bigg|_{\theta=\theta'} = 0$$

es decir, el punto silla θ' puede ser el límite producido por el algoritmo E.M. para la sucesión $\{\theta^{(n)}\}_{n\geq 0}$ dado que es una solución de $\frac{\partial l(\theta)}{\partial \theta} = 0$.

■

Capítulo 4.

Aplicaciones del algoritmo E.M. en genética.

"Sería muy singular el que toda la naturaleza, todos los planetas obedecieran las leyes eternas, y de que hubiese un pequeño animal, unos metro y medio de alto, quien, en desobediencia de esas leyes pudiese actuar como quisiese, solamente de acuerdo a su capricho"
Francois Marie Arouet Voltaire (1694-1778)

4.1 El algoritmo E.M. para inferir frecuencias de genes.

En esta sección del capítulo se usará el algoritmo E.M. para el análisis de ligamento entre genes en un cromosoma. Este problema ha sido utilizado en varias ocasiones para ilustrar el uso de algoritmos recursivos en general (ver Mclachlan y Krishnan (1997)) además de que es útil en modelos de genética para la estimación de frecuencias de genes. Esta aplicación es la que interesará en esta sección.

En el modelo original que se toma para describir este problema, el vector de los datos observados es el vector de la frecuencia de cuatro clases de genes que se representa por $y = (y_1, y_2, y_3, y_4)^t$ donde y_i es la frecuencia de la expresión del i-ésimo gen, y $Y = (Y_1, Y_2, Y_3, Y_4)$ denota a la variable aleatoria correspondiente a las observaciones y. Ahora bien, durante el estudio que se ha realizado anteriormente en laboratorios se observaron un total de n genes y se consideró que los valores de las frecuencias provenían de

una distribución multinomial[19] con cuatro celdas con sus respectivas probabilidades:

$$\tfrac{1}{2} + \tfrac{1}{4}\theta\ ,\ \tfrac{1}{4}(1-\theta),\ \tfrac{1}{4}(1-\theta)\ \text{y}\ \tfrac{1}{4}\theta \qquad ;\ \text{con}\ 0 \le \theta \le 1$$

donde θ es el parámetro desconocido.

De este modo la función de probabilidad de los datos observados es:

$$f(y;\theta) = \frac{n!}{y_1!y_2!y_3!y_4!}\left(\tfrac{1}{2} + \tfrac{1}{4}\theta\right)^{y_1}\left(\tfrac{1}{4}(1-\theta)\right)^{y_2}\left(\tfrac{1}{4}(1-\theta)\right)^{y_3}\left(\tfrac{1}{4}\theta\right)^{y_4} \qquad \ldots(4.1)$$

note que (4.1) se puede escribir

$$f(y;\theta) = \frac{n!}{y_1!y_2!y_3!y_4!}\left(\tfrac{1}{4}\right)^{y_1+y_2+y_3+y_4}(2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4}$$

Se sabe que $L(\theta) \propto f(y;\theta)$. Entonces se toma

$$L(\theta) = f(y;\theta) = (2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4}$$

y por lo tanto, tomando el logaritmo de $L(\theta)$ se tiene

$$l(\theta) = y_1\log(2+\theta) + (y_2+y_3)\log(1-\theta) + y_4\log\theta$$

Suponga ahora que la clase con frecuencia y_1 está formada por dos subclases de las cuales no se saben sus respectivas frecuencias, lo único que se puede observar es que la suma de las frecuencias de estas subclases totaliza y_1. Sean Y_{11} y Y_{12} las variables aleatorias que representan a las frecuencias respectivas de las subclases de la clase con frecuencia y_1. De éste modo el vector de datos completos sería $x = (y_{11}, y_{12}, y_2, y_3, y_4)$ donde y_{11} y y_{12} son los posibles valores no observados de las variables Y_{11} y Y_{12} y que cumplen con la condición de que $y_1 = y_{11} + y_{12}$. Por lo tanto, los datos y que se observan no están completos.

[19] Para mayor información acerca de la distribución multinomial consultar el *Apéndice A.3*

De este modo se tiene que los espacios $\boldsymbol{Y}$ y $\boldsymbol{X}$ considerados en la sección 3.2 son:

$$\boldsymbol{Y} = \{\, y = (y_1, y_2, y_3, y_4), \sum_{i=1}^{4} y_i = 1 \}$$

y

$$\boldsymbol{X} = \{\, x = (y_{11}, y_{12}, y_2, y_3, y_4) \; ; y_1 = y_{11} + y_{12} \,\}$$

y donde $h(\boldsymbol{x}) = \{(y_{11} + y_{12}, y_2, y_3, y_4) = (y_1, y_2, y_3, y_4) \mid h(\boldsymbol{x}) = y\}$

Tome la función de probabilidad del vector de datos completos como una función multinomial con parámetros respectivos como sigue:

$$(½),\ (¼\,\theta),\ (¼(1-\theta)),\ (¼\,(1-\theta))\ \text{y}\ (¼\,\theta)$$

es decir

$$f_c(x;\theta) = \frac{n!}{y_{11}!y_{12}!y_2!y_3!y_4!}\left(\frac{1}{2}\right)^{y_{11}}\left(\frac{\theta}{4}\right)^{y_{12}}\left(\frac{1-\theta}{4}\right)^{y_2+y_3}\left(\frac{\theta}{4}\right)^{y_4}$$

Esto proporciona que $f(y;\theta)$ sea dada por (4.1). Para ver esto note que:

$$\sum_{\{(y_{11},y_{12}):\, y_{11}+y_{12}=y_1\}} f_c(x;\theta) =$$

$$=^{20} \frac{n!}{y_2!y_3!y_4!}\left(\frac{\theta}{4}\right)^{y_4}\left(\frac{1-\theta}{4}\right)^{y_2+y_3}\sum_{y_{11}=0}^{y_1}\frac{1}{y_{11}!(y_1-y_{11})!}\left(\frac{1}{2}\right)^{y_{11}}\left(\frac{\theta}{4}\right)^{y_1-y_{11}}$$

$$=^{21} \frac{n!}{y_2!y_3!y_4!}\left(\frac{\theta}{4}\right)^{y_4}\left(\frac{1-\theta}{4}\right)^{y_2+y_3}\frac{1}{y_1!}\sum_{y_{11}=0}^{y_1}\frac{y_1!}{y_{11}!(y_1-y_{11})!}\left(\frac{1}{2}\right)^{y_{11}}\left(\frac{\theta}{4}\right)^{y_1-y_{11}}$$

$$= \frac{n!}{y_1!y_2!y_3!y_4!}\left(\frac{\theta}{4}\right)^{y_4}\left(\frac{1-\theta}{4}\right)^{y_2+y_3}\left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}$$

La última igualdad considerando el hecho que

[20] Teniendo en cuenta que $y_{11} + y_{12} = y_1$ y por la expresión dada para $f_c(x;\theta)$.
[21] Multiplicando y dividiendo por $y_1!$

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$$

Lo anterior quiere decir que el sumar sobre todos los valores faltantes lleva a la densidad de los datos observados y y bajo el parámetro θ.

Por otro lado note que

$$f_c(x;\theta) = \frac{n!}{y_{11}!y_{12}!y_2!y_3!y_4!}\left(\frac{1}{2}\right)^{y_{11}}\left(\frac{1}{4}\right)^{y_{12}+y_2+y_3+y_4}\theta^{y_{12}}(1-\theta)^{y_2+y_3}\theta^{y_4}$$

y por lo tanto se puede tomar

$$L_c(\theta) = f_c(x,\theta) = (1-\theta)^{y_2+y_3}\,\theta^{y_{12}+y_4}$$

es la función de verosimilitud de los datos completos bajo θ y el logaritmo de ésta es:

$$l_c(\theta) = (y_{12}+y_4)\log\theta + (y_2+y_3)\log(1-\theta)$$

Dado que $y_1 = y_{11} + y_{12}$ es suficiente con realizar el trabajo bajo una sola de las variables faltantes, suponiendo y_{12}, ya que la otra, en este caso y_{11}, se obtiene cuando se resta y_{12} de y_1 una vez que se conoce el valor de y_1.

Suponiendo que se conoce y_{12} se puede calcular el estimador máximo verosímil para el parámetro θ por los métodos tradicionales, es decir, se obtiene la derivada con respecto al parámetro θ de la ecuación anterior y se iguala a cero para encontrar de forma explícita el estimador máximo verosímil de θ, es decir, derivando $l(\theta)$ con respecto a θ se obtiene

$$\frac{\partial l_c(\theta)}{\partial \theta} = \frac{y_{12}+y_4}{\theta} - \frac{(y_2+y_3)}{1-\theta}$$

Igualando a cero para encontrar de forma explícita el estimador máximo verosímil se tiene:

$$(y_{12}+y_4)(1-\theta) - (y_2+y_3)\,\theta = 0$$

$$y_{12}+y_4 - y_{12}\theta - y_4\theta - (y_2+y_3)\,\theta = 0$$

$$y_{12}+y_4 - \theta\,(y_{12}+y_2+y_3+y_4) = 0$$

$$\theta\,(y_{12}+y_2+y_3+y_4) = y_{12}+y_4$$

$$\theta^* = \frac{y_{12}+y_4}{y_{12}+y_2+y_3+y_4} \qquad \text{.....(4.2)}$$

Si se supiera quién es y_{12} entonces (4.2) sería el estimador máximo verosímil, pero como no es así entonces se utiliza el algoritmo E.M. para resolver el problema.

PASO E.

El paso E consiste en calcular el valor esperado

$$G(\theta;\theta^{(n)}) = E_{\theta^{(n)}}\{l_c(\theta)\,|\,\boldsymbol{y}\}$$

De este modo

$$G(\theta;\theta^{(n)}) = E_{\theta^{(n)}}[(\boldsymbol{Y}_{12}+y_4)\log\theta\,|\,\boldsymbol{y}] + E_{\theta^{(n)}}[(y_2+y_3)\log(1-\theta)\,|\,\boldsymbol{y}]$$
$$= E_{\theta^{(n)}}[\boldsymbol{Y}_{12}\mid y_1]\log\theta + y_4\log\theta + (y_2+y_3)\log(1-\theta)$$

la última igualdad dado que $\boldsymbol{Y}_{12}$ depende solamente de y_1, y por lo tanto basta con calcular $E_{\theta^{(n)}}[\boldsymbol{Y}_{12}\mid y_1]$.

Para encontrar la esperanza condicional se hace la pregunta de ¿Cuál es la proporción de experimentos que le corresponden tanto a y_{11} como a y_{12} dado que se hicieron y_1 experimentos?

Se sabe que y_1 es la frecuencia del gene 1, recuerde que y_1 está formado por la suma de dos subclases, y_{11} y y_{12}, de las cuales no se conocen sus respectivas frecuencias, es decir, y_{11} es la frecuencia de una de las expresiones del gen 1 y y_{12} es la frecuencia de la otra expresión del gen 1. Donde la expresión del gen puede ser

por ejemplo una enfermedad o una determinada expresión fenotípica[22]. Tómese como hipótesis que en base a observaciones se nota que la expresión fenotípica 1 ocurre con probabilidad ½ y que la expresión 2 ocurre con probabilidad (¼)θ. Se considera un éxito cada vez que la expresión 1 ocurre. Esto pasa con probabilidad $\left(\frac{1}{2}\right)$ por probabilidad clásica; y un fracaso en caso de ocurrencia de la expresión 2, el cual ocurre con probabilidad $\left(\frac{1}{4}\theta\right)$. Como la ocurrencia de cada evento es independiente se tiene que dado $Y_1 = y_1$ sucede lo siguiente

$$P(Y_{11}=y_{11}|Y_1=y_1)=\binom{y_1}{y_{11}}\left[\frac{1}{2}\right]^{y_{11}}\left[\frac{1}{4}\theta\right]^{(y_1-y_{11})}$$

$$P(Y_{12}=y_{12}|Y_1=y_1)=\binom{y_1}{y_{12}}\left[\frac{1}{2}\right]^{(y_1-y_{12})}\left[\frac{1}{4}\theta\right]^{y_{12}}$$

es decir, que dado $Y_1 = y_1$,

Y_{11} tiene distribución binomial con parámetros y_1 y $\left[\frac{1}{2}\right]$

y además

Y_{12} tiene distribución binomial con parámetros y_1 y $\left[\frac{1}{4}\theta\right]$

[22] La expresión fenotípica son las características observables en un individuo.

De este modo se obtiene

$$E_{\theta^{(n)}}[\mathbf{Y}_{12} \mid y_1] = \left[\frac{\theta^{(n)}}{4} y_1 \right] + \left[\frac{\theta^{(n)}}{4} \right]$$

PASO M.

Encontrar θ que maximiza a $G(\theta;\theta^{(n)})$, es decir θ que maximice

$$G(\theta^{(n+1)};\theta^{(n)}) = \left[\frac{\theta^{(n)}}{4} y_1 \right] \log\theta + y_4 \log\theta + (y_2+y_3)\log(1-\theta) \qquad \text{.....(4.3)}$$

Note que (4.3) corresponde a la expresión para $l_c(\theta)$ con $E_{\theta^{(n)}}[\mathbf{Y}_{12} \mid y_1]$ en el lugar de y_{12}. Así, el valor de θ que maximiza (4.3) está dado por

$$\theta^{(n+1)} = \frac{E_{\theta^{(n)}}[\mathbf{Y}_{12} \mid y_1] + y_4}{E_{\theta^{(n)}}[\mathbf{Y}_{12} \mid y_1] + y_2 + y_3 + y_4}$$

Ahora denótese a $E_{\theta^{(n)}}[\mathbf{Y}_{12} \mid y_1]$ por $y_{12}^{(n)}$ y se tiene que

$$\theta^{(n+1)} = \frac{y_{12}^{(n)} + y_4}{y_{12}^{(n)} + y_2 + y_3 + y_4}$$

Es precisamente aquí en donde empieza el algoritmo a trabajar. Sea $\theta^{(0)}$ el valor inicial presentado para θ. Así se realiza la primera iteración del paso E donde se obtiene el valor $y_{12}^{(0)}$ y con este valor se obtiene $\theta^{(1)}$ en el paso M.

Una vez que se tiene a $\theta^{(1)}$ se procede a recurrir iterativamente de la siguiente forma:

Segunda iteración del paso E.

$$y_{12}^{(1)} = E_{\theta^{(1)}}(Y_{12} \mid Y = y\} = E_{\theta^{(1)}}(Y_{12} \mid Y_1 = y_1) = y_1 \left[\frac{\theta^{(1)}}{2 + \theta^{(1)}} \right]$$

Segunda iteración del paso M.

$$\theta^{(2)} = \frac{y_{12}^{(1)} + y_4}{y_{12}^{(1)} + y_2 + y_3 + y_4}$$

y en general para la n-ésima iteración del paso E:

$$y_{12}^{(n-1)} = E_{\theta^{(n-1)}}(Y_{12} \mid Y = y\}$$

$$= E_{\theta^{(n-1)}}(Y_{12} \mid Y_1 = y_1) = y_1 \left[\frac{\theta^{(n-1)}}{2 + \theta^{(n-1)}} \right]$$

y la n-ésima iteración del paso M:

$$\theta^{(n)} = \frac{y_{12}^{(n-1)} + y_4}{y_{12}^{(n-1)} + y_2 + y_3 + y_4}$$

y así de forma sucesiva hasta encontrar la n tal que la diferencia entre $\theta^{(n)}$ y $\theta^{(n+1)}$ sea suficientemente pequeña como para asegurar en términos prácticos la convergencia al estimador máximo verosímil θ^*.

4.2 El algoritmo E.M. para inferir secuencias de A.D.N.

Esta sección trata específicamente acerca del uso del algoritmo E.M. en una de sus aplicaciones para poder inferir las secuencias de A.D.N.

El modelo que se analiza es un modelo estocástico con espacio de estados y observaciones finito. En la aplicación directa a la sucesión de A.D.N., las observaciones (eventos) corresponden a la sucesión de las bases (A, C, G y T) de las secuencias. No obstante, el modelo puede comportarse de forma flexible siempre y cuando se tenga cuidado con la definición clara del espacio de estados correspondiente a las observaciones.

El problema que se resolverá con este modelo es el de inferir los estados (secuencia verdadera **S**) a partir del vector de observaciones **Y**. Los estados serán inferidos a partir de un sistema de ecuaciones y de una fórmula recursiva. Como hipótesis adicionales se tendrá una ecuación de observación y la forma en que se desarrolla el sistema de ecuaciones asociado con el comportamiento de los estados del proceso estocástico.

Sea $\mathbf{S} = (s_1, s_2, ..., s_m)$ la secuencia de estados no observable (que vendría siendo la secuencia de A.D.N. verdadera) y $\mathbf{Y} = (y_1, y_2, ..., y_m)$ la sucesión de A.D.N. observada y también la variable aleatoria con distribución dependiente de **S**.

Asuma que los estados se obtendrán a partir de un sistema de ecuaciones denotado por $P_\theta(s_t \mid s_1, s_2, ..., s_{t-1})$, es decir, la distribución de probabilidad del estado al tiempo t dado que se conocen los estados anteriores. Suponga que la secuencia **S** es una cadena de Markov y que por lo tanto cumple con lo siguiente: $P_\theta(s_t \mid s_1, s_2, ..., s_{t-1}) = P_\theta(s_t \mid s_{t-1})$, que vendría siendo la probabilidad de transición en un paso para la cadena de Markov **S**.

Asimismo, se denota como la ecuación de observación a $P_\theta(y_t \mid s_t, y_1, ..., y_{t-1})$.

Este trabajo se concentrará únicamente en aplicaciones a sucesiones lineales por lo que se considera, además de la ecuación de observación, que dependerá del estado actual y de la observación anterior $P_\theta(y_t \mid s_t, y_{t-1})$ para $t \geq 2$, la probabilidad en $t = 1$ que dependerá sólo del estado actual, es decir, $P_\theta(y_1 \mid s_1)$ es la observación al tiempo $t = 1$ dado el estado presente y $P_\theta(s_1)$ es la probabilidad inicial de la cadena **S**.

Ahora bien, para encontrar a la ecuación de observación $P_0(y_t \mid s_t, y_{t-1})$, $t \geq 2$, y $P_0(y_1 \mid s_1)$ recuerde que una secuencia de A.D.N. está formada por combinaciones diversas de únicamente 4 bases (A, C, G y T); utilizando este principio defina un vector O = (A, C, G, T) = (O_1, O_2, O_3, O_4) para asociarlo a cada observación y_t y a cada estado respectivo s_t con t = 1, 2, ..., m.

Tome a $y_t = (y_{t,1}, y_{t,2}, y_{t,3}, y_{t,4})$ el vector que representa la t-ésima observación y $s_t = (s_{t,1}, s_{t,2}, s_{t,3}, s_{t,4})$ el vector que representa el t-ésimo estado, donde se cumplirá lo siguiente para k = 1, 2, 3, 4:

$$y_{t,k} = \begin{cases} 1 & y_t = O_k \\ 0 & \text{en otro caso} \end{cases} \quad \text{y} \quad s_{t,k} = \begin{cases} 1 & s_t = O_k \\ 0 & \text{en otro caso.} \end{cases}$$

Esto quiere decir que al cumplirse $y_t = O_k = s_t$ se llega a que el estado que se busca s_t es exactamente igual al que se observa y_t.

De ahora en adelante se utilizará la notación k = 1, 2, 3, 4 para indicar a cada uno de los valores A, C, G, T respectivamente. Cada una de las probabilidades de transición tendrá un modelo multinomial (ver **A.3**) y esto es por la forma en la que se esta definiendo tanto vector de observaciones como el de estados. Nótese que se tienen k = 4 distintas posibilidades para cada uno de las entradas en el vector y_t y s_t. De este modo, se tiene que para $P_{(ji)k}$ la probabilidad de transición esta definida por lo siguiente.

$P_{(ji)k}$ = P(observar k | observación anterior fue i y el estado actual es j)

cumpliéndose que:

$$\sum_{k=1}^{4} P_{(ji)k} = 1.$$

Ahora bien, utilizando todo lo anterior la ecuación de observación resulta de la siguiente forma:

$$P_\theta(y_t \mid s_t, y_{t-1}) = \prod_{k=1}^{3} \prod_{j=1}^{4} \prod_{i=1}^{4} P_{(ji)k}^{y_{t,k}\, y_{t-1,i}\, s_{t,j}} \left(1 - \sum_{k=1}^{3} P_{(ji)k}\right)^{y_{t,4}\, y_{t-1,i}\, s_{t,j}} . \quad \ldots(4.4)$$

Una vez que se tiene a la ecuación de observación se encuentra la ecuación de estados $P_\theta(s_t \mid s_{t-1})$. Defínase a λ_{ij} con i,j $\in \{A, C, G, T\}$ como los componentes de la matriz de transición de los estados, de tal modo que

$$\lambda_{ij} = P(\text{ir al estado i} \mid \text{se encuentra en el estado j})$$

sujeto a

$$\sum_{j=1}^{4} \lambda_{ij} = 1$$

y de este modo la ecuación de estados estaría dada por:

$$P_\theta(s_t \mid s_{t-1}) = \prod_{i=1}^{4} \prod_{j=1}^{3} \lambda_{ij}^{s_{t-1,i}\, s_{t,j}} \left(1 - \sum_{j=1}^{3} \lambda_{ij}\right)^{s_{t-1,i}\, s_{t,4}} . \quad \ldots\ldots(4.5)$$

Además defina

$$P_{ij} = P(\text{observación inicial es j} \mid \text{estado inicial es i})$$

sujeto a

$$\sum_{j=1}^{4} P_{ij} = 1$$

y también

$$\pi_i = P(\text{estado inicial es i})$$

sujeto a

$$\sum_{i=1}^{4} \pi_i = 1.$$

De este modo se tiene que

$$P_\theta(s_1) = \prod_{i=1}^{3} \pi_i^{s_{1,i}} \left(1 - \sum_{i=1}^{3} \pi_i\right)^{s_{1,4}} . \quad \ldots\ldots\ldots(4.6)$$

y

$$P_\theta(y_1 \mid s_1) = \prod_{i=1}^{4}\prod_{j=1}^{3} P_{ij}^{s_{1,i}\, y_{1,j}} \left(1 - \sum_{j=1}^{3} P_{ij}\right)^{s_{1,i}\, y_{1,4}}. \qquad \text{.....(4.7)}$$

donde $\theta = (\pi_i, \lambda_{ij}, P_{ij}, P_{(ji)k}$, i, j, k $\in \{1, 2, 3, 4\})$ es el vector de parámetros que se quiere estimar.

Con la ecuación de observación $P_\theta(y_t \mid s_t, y_{t-1})$, la ecuación de estados $P_\theta(s_t \mid s_{t-1})$ y las probabilidades iniciales: $P_\theta(y_1 \mid s_1)$ y $P_\theta(s_1)$ se construirá la función de verosimilitud del modelo de la siguiente forma:

Teorema 4.1. Sean $s^m = (s_1, s_2, ..., s_m)$ y $y^m = (y_1, y_2, ..., y_m)$ sucesiones lineales. Suponga que $P_\theta(s_t \mid s_1, ..., s_{t-1}) = P_\theta(s_t \mid s_{t-1})$ con t = 2, 3, ..., m, que y_t dependa solamente de y_{t-1} y s_t, y las distribuciones iniciales dadas por $P_\theta(y_1 \mid s_1)$ y $P_\theta(s_1)$. Entonces se tiene que

$$P_\theta(y_1, y_2, ..., y_m, s_1, s_2, ..., s_m)$$
$$= \left(\prod_{t=2}^{m} P_\theta(y_t \mid s_t, y_{t-1})\right) P_\theta(y_1 \mid s_1) \left(\prod_{t=2}^{m} P_\theta(s_t \mid s_{t-1})\right) P_\theta(s_1)$$

donde $\theta = \{(\pi_i, \lambda_{ij}, P_{ij}, P_{(ji)k}); i, j, k \in \{1, 2, 3, 4\} = \{A, C, G, T\}\}$

Demostración
Se demuestra por inducción.

(i) Se demuestra para m = 3.

$$P_\theta(y^3, s^3) = P_\theta(y_1, y_2, y_3, s_1, s_2, s_3)$$
$$= P_\theta(y_3 \mid y_1, y_2, s_1, s_2, s_3)\, P_\theta(y_1, y_2, s_1, s_2, s_3)$$
$$= P_\theta(y_3 \mid y_1, y_2, s_1, s_2, s_3)\, P_\theta(y_2 \mid y_1, s_1, s_2, s_3)$$
$$P_\theta(y_1, s_1, s_2, s_3)$$
$$= P_\theta(y_3 \mid y_1, y_2, s_1, s_2, s_3)\, P_\theta(y_2 \mid y_1, s_1, s_2, s_3)$$
$$P_\theta(y_1 \mid s_1, s_2, s_3)\, P_\theta(s_1, s_2, s_3)$$

$$= P_\theta(y_3 \mid y_1, y_2, s_1, s_2, s_3) P_\theta(y_2 \mid y_1, s_1, s_2, s_3)$$
$$P_\theta(y_1 \mid s_1, s_2, s_3)\, P_\theta(s_3 \mid s_1, s_2)\, P_\theta(s_2 \mid s_1)\, P_\theta(s_1)$$

Por propiedad de Markov para s^3 y propiedad de la secuencia de observaciones y^3 se tiene:

$$P_\theta(y^3, s^3) = P_\theta(y_3 \mid y_2, s_3)\, P_\theta(y_2 \mid y_1, s_2)\, P_\theta(y_1 \mid s_1)\, P_\theta(s_3 \mid s_2)$$
$$P_\theta(s_2 \mid s_1) P_\theta(s_1)$$

$$= \left(\prod_{t=2}^{3} P_\theta(y_t \mid s_t, y_{t-1}) \right) P_\theta(y_1 \mid s_1) \left(\prod_{t=2}^{3} P_\theta(s_t \mid s_{t-1}) \right) P_\theta(s_1)$$

(ii) Asuma que el resultado se cumple para m = k

$$P_\theta(y^k, s^k) = P_\theta(y_1, y_2, \ldots, y_k, s_1, s_2, \ldots, s_k) =$$

$$= \left(\prod_{t=2}^{k} P_\theta(y_t \mid s_t, y_{t-1}) \right) P_\theta(y_1 \mid s_1) \left(\prod_{t=2}^{k} P_0(s_t \mid s_{t-1}) \right) P_\theta(s_1)$$

(iii) Se demuestra para m = k+1

$$P_\theta(y_1, y_2, \ldots, y_{k+1}, s_1, s_2, \ldots, s_{k+1}) =$$

$$= P_\theta(y_{k+1} \mid y_1, y_2, \ldots, y_k, s_1, s_2, \ldots, s_{k+1})\, P_\theta(y_1, y_2, \ldots, y_k, s_1, s_2, \ldots, s_{k+1})$$

$$= P_\theta(y_{k+1} \mid y_k, s_{k+1})\, P_\theta(y_1, y_2, \ldots, y_k, s_1, s_2, \ldots, s_{k+1})$$

$$= P_\theta(y_{k+1} \mid y_k, s_{k+1})\, P_\theta(s_{k+1} \mid y_1, y_2, \ldots, y_k, s_1, s_2, \ldots, s_k)$$
$$P_\theta(y_1, y_2, \ldots, y_k, s_1, s_2, \ldots, s_k)$$

$$= P_\theta(y_{k+1} \mid y_k, s_{k+1})\, P_\theta(s_{k+1} \mid s_k)\, P_\theta(y_1, y_2, \ldots, y_k, s_1, s_2, \ldots, s_k)$$

$$= {}^{23} P_\theta(y_{k+1} \mid y_k, s_{k+1})\, P_\theta(s_{k+1} \mid s_k)$$

$$\left(\prod_{t=2}^{k} P_\theta(y_t \mid s_t, y_{t-1}) \right) P_\theta(y_1 \mid s_1) \left(\prod_{t=2}^{k} P_\theta(s_t \mid s_{t-1}) \right) P_\theta(s_1)$$

$$= \left(\prod_{t=2}^{k+1} P_\theta(y_t \mid s_t, y_{t-1}) \right) P_\theta(y_1 \mid s_1) \left(\prod_{t=2}^{k+1} P_\theta(s_t \mid s_{t-1}) \right) P_\theta(s_1)$$

De este modo se cumple para k+1 y por lo tanto para todo m, con lo cual se concluye la demostración del teorema.

■

Lo que seguiría es encontrar los estimadores máximo verosímiles para el parámetro $\theta = \{ (\pi_i, \lambda_{ij}, P_{ij}, P_{(ji)k})$ con $i, j, k \in \{1, 2, 3, 4\}$ para posteriormente aplicar el algoritmo E.M. y poder estimar de forma recursiva a $\mathbf{s}^m$.

4.2.1 Cálculo de estimadores máximo verosímiles.

El cálculo de los estimadores se realiza para el caso donde $O = (O_1, O_2)$. Para O con dimensión distinta a 2 el procedimiento es similar. Por el **Teorema 4.1** se tiene que

$$f_c(y^m, s^m; \theta) = \left(\prod_{t=2}^{m} P_\theta(y_t \mid s_t, y_{t-1}) \right) P_\theta(y_1 \mid s_1) \left(\prod_{t=2}^{m} P_\theta(s_t \mid s_{t-1}) \right) P_\theta(s_1)$$

donde las probabilidades de transición que aparecen en la expresión anterior estan dados por (4.4), (4.5), (4.6) y (4.7); $y^m = (y_1, y_2, ..., y_m)$ es el vector de observaciones y $s^m = (s_1, s_2, ..., s_m)$ es el vector de estados.

En este caso se tiene $y_t = (y_{t,1}, y_{t,2})$; $s_t = (s_{t,t}, s_{t,2})$, $t = 1, 2, ..., m$; donde por definición

$$\sum_{j=1}^{2} y_{t,j} = \sum_{j=t}^{2} s_{t,j} = 1; \; t = 1, 2, \ldots, m$$

[23] Por hipótesis de inducción.

La log-verosimilitud está dada por:

$$\log f_c(y^m, s^m;\theta) =$$

$$= \log \left(\left(\prod_{t=2}^{m} P_\theta(y_t \mid s_t, y_{t-1}) \right) P_\theta(y_1 \mid s_1) \left(\prod_{t=2}^{m} P_\theta(s_t \mid s_{t-1}) \right) P_\theta(s_1) \right)$$

$$= \log P_\theta(s_1) + \sum_{t=2}^{m} \log P_\theta(s_t \mid s_{t-1}) + \log P_\theta(y_1 \mid s_1) + \sum_{t=2}^{m} \log P_\theta(y_t \mid s_t, y_{t-1})$$

Para resolver de forma más sencilla la derivada y la obtención de los estimadores se separan las ecuaciones con las que se trabajan de la siguiente forma:

$$P_\theta(s_1) = \prod_{i=1}^{2} \pi_i^{s_{1,i}} = \pi_1^{s_{1,1}} (1-\pi_1)^{s_{1,2}}$$

$$P_\theta(s_t \mid s_{t-1}) = \prod_{i=1}^{2} \prod_{j=1}^{2} \lambda_{ij}^{s_{t-1,i}\, s_{t,j}} = \prod_{i=1}^{2} \lambda_{i1}^{s_{t-1,i}\, s_{t,1}} (1-\lambda_{i1})^{s_{t-1,i}\, s_{t,2}}$$

$$P_\theta(y_1 \mid s_1) = \prod_{i=1}^{2} \prod_{j=1}^{2} P_{ij}^{s_{1,i}\, y_{1,j}} = \prod_{i=1}^{2} P_{i1}^{s_{1,i}\, y_{1,1}} (1-P_{i1})^{s_{1,i}\, y_{1,2}}$$

$$P_\theta(y_t \mid s_t, y_{t-1}) = \prod_{i=1}^{2} \prod_{j=1}^{2} \prod_{k=1}^{2} P_{(ji)k}^{s_{t,i}\, y_{t-1,j}\, y_{t,k}} = \prod_{i=1}^{2} \prod_{j=1}^{2} P_{(ji)1}^{s_{t,i}\, y_{t-1,j}\, y_{t,1}} (1-P_{(ji)1})^{s_{t,i}\, y_{t-1,j}\, y_{t,2}}$$

de donde el logaritmo de cada una vendría siendo:

$$\log P_\theta(s_1) = s_{1,1} \log \pi_1 + s_{1,2} \log(1-\pi_1) \qquad \text{.....(4.8)}$$

$$\log \prod_{t=2}^{m} P_\theta(s_t \mid s_{t-1}) = \sum_{t=2}^{m} \sum_{i=1}^{2} [s_{t,1} s_{t-1,i} \log \lambda_{i1} + s_{t,2} s_{t-1,i} \log(1-\lambda_{i1})] \qquad \text{.....(4.9)}$$

$$\log P_\theta(y_1 \mid s_1) = \sum_{i=1}^{2} [s_{1,i} y_{1,1} \log P_{i1} + s_{1,i} y_{1,2} \log(1-P_{i1})] \qquad \text{.....(4.10)}$$

$$\log \prod_{t=2}^{m} P_\theta(y_t \mid s_t, y_{t-1}) = \sum_{t=2}^{m} \sum_{i=1}^{2} \sum_{j=1}^{2} [s_{t,i}\, y_{t-1,j} y_{t,1} \log P_{(ji)1} + s_{t,i}\, y_{t-1,j} y_{t,2} \log(1-P_{(ji)1})] \qquad \text{.....(4.11)}$$

Se obtiene la derivada con respecto a cada uno de los parámetros que se quieren encontrar, que en este caso vendrían siendo:

$$\pi_i, \lambda_{ij}, P_{ij}, P_{(ji)k} \text{ con } i, j, k \in \{1, 2\}$$

Para (4.8) se tiene:

$$\frac{\partial \log P_\theta(s_1)}{\partial \pi_1} \qquad \frac{s_{1,1}}{\pi_1} - \frac{s_{1,2}}{(1-\pi_1)} = 0$$

y una vez resuelta la ecuación se tiene que $\pi_1 = s_{1,1}$ y como $\pi_1 + \pi_2 = 1$ entonces $\pi_2 = s_{1,2}$ de forma análoga se obtienen los demás parámetros; de tal modo que el vector de parámetros resulta[24]:

$$\theta = (\pi_i, \lambda_{ij}, P_{ij}, P_{(ji)k}) \text{ con } i, j, k \in \{1, 2\} \text{ donde}$$

$$\pi = (\pi_1, \pi_2) = (s_{1,1}, s_{1,2})$$

$$\lambda_{ij} = \frac{\sum_{t=2}^{m} s_{t,j}\, s_{t-1,i}}{\sum_{t=2}^{m} s_{t-1,i}}$$

$$P_{ij} = y_{1,j}$$

$$P_{(ji)k} = \frac{\sum_{t=2}^{m} s_{t,i}\, y_{t-1,j}\, y_{t,k}}{\sum_{t=2}^{m} s_{t,i}\, y_{t-1,j}}$$

.....(4.12)

[24] Los cálculos para obtener los estimadores se encuentran en el *Apéndice A.4*.

Nótese que los valores que hace falta conocer son los estados s_t para conocer al estimador máximo verosímil, pero como no es así se utiliza el algoritmo E.M. para superar este problema.

PASO E

Se toma un valor $\theta^{(0)} = (\pi^{(0)}, \lambda_{ij}^{(0)}, P_{ij}^{(0)}, P_{(ji)k}^{(0)}, i, j, k \in \{1, 2, 3, 4\})$ como un valor inicial y se calcula el valor esperado de s_t condicionando sobre los datos observados y^m.

Primera iteración.

$$G(\theta,\theta^{(0)}) = E_{\theta^{(0)}}(f_c(\theta) \mid y^m) = E_{\theta^{(0)}}\left[\log P(y^m, s^m;\theta) \mid y^m\right]$$

$$= E_{\theta^{(0)}}\left[\log P_\theta(s_1) + \sum_{t=2}^{m} \log P_\theta(s_t \mid s_{t-1}) + \log P_\theta(y_1 \mid s_1) + \sum_{t=2}^{m} \log P_\theta(y_t \mid s_t, y_{t-1}) \mid y^m\right]$$

$$= E_{\theta^{(0)}}[\log P_\theta(s_1) \mid y^m] + \sum_{t=2}^{m} E_{\theta^{(0)}}[\log P_\theta(s_t \mid s_{t-1}) \mid y^m]$$

$$+ E_{\theta^{(0)}}[\log P_\theta(y_1 \mid s_1) \mid y^m] + \sum_{t=2}^{m} E_{\theta^{(0)}}[\log P_\theta(y_t \mid s_t, y_{t-1}) \mid y^m]$$

Por comodidad se toma a cada una de las esperanzas por separado:

$$E_{\theta^{(0)}}[\log P_\theta(s_1) \mid y^m]$$
$$= E_{\theta^{(0)}}[S_{1,1}\log\pi_1^{(0)} + S_{1,2}\log(1-\pi_1^{(0)}) \mid y^m]$$
$$= E_{\theta^{(0)}}[S_{1,1}\log\pi_1^{(0)} \mid y^m] + E_{\theta^{(0)}}[S_{1,2}\log(1-\pi_1^{(0)}) \mid y^m]$$
$$= \log\pi_1^{(0)} E_{\theta^{(0)}}(S_{1,1} \mid y^m) + \log(1-\pi_1^{(0)}) E_{\theta^{(0)}}(S_{1,2} \mid y^m)$$

así que

$$E_{\theta^{(0)}}[\log P_\theta(s_1) \mid y^m]$$
$$= \log\pi_1^{(0)} E_{\theta^{(0)}}[S_{1,1} \mid y^m] + \log(1-\pi_1^{(0)}) E_{\theta^{(0)}}[S_{1,2} \mid y^m] \quad ....(4.13)$$

donde $S_{t,i}$ es la variable aleatoria que asume valores $s_{t,i}$, t = 1, 2, ..., m, i = 1, 2.

De este modo se usa $E_{\theta^{(0)}}[S_{1,i} \mid y^m]$, i = 1, 2 en la verosimilitud en el lugar de $s_{1,i}$.

$$E_{\theta^{(0)}}[\log \prod_{t=2}^{m} P_\theta(s_t \mid s_{t-1}) \mid y^m] = \sum_{t=2}^{m} \sum_{i=1}^{2} E_{\theta^{(0)}}[S_{t,1}S_{t-1,i}\log\lambda_{i1}^{(0)}$$

$$+ S_{t,2}S_{t-1,i}\log(1-\lambda_{i1}^{(0)}) \mid y^m]$$

$$= \sum_{t=2}^{m} \sum_{i=1}^{2} \{E_{\theta^{(0)}}[S_{t,1}S_{t-1,i}\log\lambda_{i1}^{(0)} \mid y^m] + E_{\theta^{(0)}}[S_{t,2}S_{t-1,i}\log(1-\lambda_{i1}^{(0)}) \mid y^m]\}$$

$$= \sum_{t=2}^{m} \sum_{i=1}^{2} \{\log\lambda_{i1}^{(0)} E_{\theta^{(0)}}[S_{t,1}S_{t-1,i} \mid y^m] + \log(1-\lambda_{i1}^{(0)}) E_{\theta^{(0)}}[S_{t,2}S_{t-1,i} \mid y^m]\}$$

y por lo tanto

$$E_{\theta^{(0)}}[\log \prod_{t=2}^{m} P_\theta(s_t \mid s_{t-1}) \mid y^m] =$$

$$= \sum_{t=2}^{m} \sum_{i=1}^{2} \{\log\lambda_{i1}^{(0)} E_{\theta^{(0)}}[S_{t,1}S_{t-1,i} \mid y^m] + \log(1-\lambda_{i1}^{(0)}) E_{\theta^{(0)}}[S_{t,2}S_{t-1,i} \mid y^m]\} \quad \text{.....(4.14)}$$

De este modo se usa en la verosimilitud $E_{\theta^{(0)}}[S_{t,j}, S_{t-1,i} \mid y^m]$, en el lugar de $s_{t,j}$, $s_{t-1,i}$, i, j = 1, 2.

$$E_{\theta^{(0)}}[\log P_\theta(y_t \mid s_t) \mid y^m] = \sum_{i=1}^{2} E_{\theta^{(0)}}[S_{i,1}y_{1,1}\log P_{i1}^{(0)} + S_{i,1}y_{1,2}\log(1-P_{i1}^{(0)}) \mid y^m]$$

$$= \sum_{i=1}^{2} \{E_{\theta^{(0)}}[S_{i,1}y_{1,1}\log P_{i1}^{(0)} \mid y^m] + E_{\theta^{(0)}}[S_{i,1}y_{1,2}\log(1-P_{i1}^{(0)}) \mid y^m]\}$$

$$= \sum_{i=1}^{2} \{y_{1,1}\log P_{i1}^{(0)} E_{\theta^{(0)}}[S_{i,1} \mid y^m] + y_{1,2}\log(1-P_{i1}^{(0)}) E_{\theta^{(0)}}[S_{i,1} \mid y^m]\}$$

$$E_{\theta^{(0)}}[\log P_\theta(y_1 \mid s_1) \mid y^m] =$$

$$= \sum_{i=1}^{2} \{(y_{1,1}\log P_{i1}^{(0)} + y_{1,2}\log(1-P_{i1}^{(0)})) E_{\theta^{(0)}}[S_{i,1} \mid y^m]\} \quad \text{.....(4.15)}$$

De este modo se usa en la verosimilitud $E_{\theta^{(0)}}[S_{i,1} \mid y^m]$, i = 1, 2 en el lugar de $s_{i,1}$.

$$E_\theta^{(0)}\left[\log\prod_{t=2}^{m} P_\theta(y_t \mid S_t, y_{t-1}) \mid y^n\right]$$

$$= \sum_{t=2}^{m}\sum_{i=1}^{2}\sum_{j=1}^{2} E_\theta^{(0)}\left[S_{t,i}y_{t-1,j}y_{t,1}\log P_{(ji)1}^{(0)} + S_{t,i}y_{t-1,j}y_{t,2}\log(1-P_{(ji)1}^{(0)}) \mid y^m\right]$$

$$= \sum_{t=2}^{m}\sum_{i=1}^{2}\sum_{j=1}^{2}\{E_\theta^{(0)}\left[S_{t,i}\, y_{t-1,j}y_{t,1}\log P_{(ji)1}^{(0)} \mid y^m\right]$$

$$+ E_\theta^{(0)}\left[S_{t,i}\, y_{t-1,j}y_{t,2}\log(1-P_{(ji)1}^{(0)}) \mid y^m\right]\}$$

$$= \sum_{t=2}^{m}\sum_{i=1}^{2}\sum_{j=1}^{2}\{\, y_{t-1,j}y_{t,1}\log P_{(ji)1}^{(0)}\, E_\theta^{(0)}\left[S_{t,i} \mid y^m\right]$$

$$+ y_{t-1,j}y_{t,2}\log(1-P_{(ji)1}^{(0)})E_\theta^{(0)}\left[S_{t,i} \mid y^m\right]\}$$

y por lo tanto

$$E_\theta^{(0)}\left[\log\prod_{t=2}^{m} P_\theta(y_t \mid s_t, y_{t-1}) \mid y^m\right] =$$

$$= \sum_{t=2}^{m}\sum_{i=1}^{2}\sum_{j=1}^{2}\{(y_{t-1,j}y_{t,1}\log P_{(ji)1}^{(0)} + y_{t-1,j}y_{t,2}\log(1-P_{(ji)1}^{(0)}))E_\theta^{(0)}\left[S_{t,i} \mid y^m\right]\} \quad ...(4.16)$$

De este modo se usa en la verosimilitud $E_\theta^{(0)}\left[S_{t,i} \mid y^m\right]$, i = 1, 2 en el lugar de $s_{t,i}$. Y por lo tanto lo que se deben encontrar son las siguientes cantidades:

$$E_\theta^{(0)}\left[S_{t,i} \mid y^m\right] \qquad \text{y} \qquad E_\theta^{(0)}\left[S_{t,j}, S_{t-1,i} \mid y^m\right]$$

con t, i, j = 1, 2, ..., m

Nótese que $s_{t,i}$ es cero o uno. Por lo tanto calcular sus esperanzas corresponde a obtener $P_\theta[S_{t,i} \mid y^m]$ y $P_\theta[S_{t,j}, S_{t-1,i} \mid y^m]$. De tal modo que se tiene:

❖ $E_\theta^{(0)}\left[S_{t,i} \mid y^m\right]$

$$P_\theta(S_{t,i} \mid y^m) = \frac{P_\theta(S_{t,i}, y^m)}{P_\theta(y^m)}$$

$$= \sum_{s_{t+1}} \frac{P_\theta(s_{t,i}, s_{t+1}, y^m)}{P_\theta(y^m)}$$

$$= \sum_{s_{t+1}} P_\theta(s_{t,i} \mid s_{t+1}, y^m)\, P_\theta(s_{t+1} \mid y^m)$$

$$=^{25} \sum_{s_{t+1}} P_\theta(s_{t,i} \mid s_{t+1}, y^t)\, P_\theta(s_{t+1} \mid y^m)$$

$$=\sum_{s_{t+1}} \frac{P_\theta(s_{t,i}, s_{t+1}, y^t)}{P_\theta(s_{t+1}, y^t)}\, P_\theta(s_{t+1} \mid y^m)$$

$$=\sum_{s_{t+1}} \frac{P_\theta(s_{t,i}, s_{t+1} \mid y^t)}{P_\theta(s_{t+1} \mid y^t)}\, P_\theta(s_{t+1} \mid y^m)$$

$$=\sum_{s_{t+1}} \frac{P_\theta(s_{t+1} \mid s_{t,i}, y^t)\, P_\theta(s_{t,i} \mid y^t)}{P_\theta(s_{t+1} \mid y^t)}\, P_\theta(s_{t+1} \mid y^m)$$

$$=\sum_{s_{t+1}} \frac{P_\theta(s_{t+1} \mid s_{t,i})\, P_\theta(s_{t,i} \mid y^t)}{P_\theta(s_{t+1} \mid y^t)}\, P_\theta(s_{t+1} \mid y^m)$$

$$=\sum_{j} \frac{P_\theta(s_{t+1,j} \mid s_{t,i})\, P_\theta(s_{t,i} \mid y^t)}{P_\theta(s_{t+1,j} \mid y^t)}\, P_\theta(s_{t+1,j} \mid y^m)$$

De este modo se tiene

$$P_\theta(s_{t,i} \mid y^m) = \sum_{j} \frac{\lambda_{ij}\, P_\theta(s_{t,i} \mid y^t)\, P_\theta(s_{t+1,j} \mid y^m)}{P_\theta(s_{t+1,j} \mid y^t)}$$

Ahora se necesita encontrar expresiones para $P_\theta(s_{t,i} \mid y^t)$ y $P_\theta(s_{t,j} \mid y^{t-1})$, note que en $P_\theta(s_{t,i} \mid y^m)$ hay una dependencia en $P_\theta(s_{t,i} \mid y^t)$ y $P_\theta(s_{t,j} \mid y^{t-1})$. De forma conjunta con la fórmula para $P_\theta(s_{t,i} \mid y^m)$ se presentarán las fórmulas para $P_\theta(s_{t,j} \mid y^t)$ y $P_\theta(s_{t,j} \mid y^{t-1})$, que formarán las bases del algoritmo recursivo que se usará para calcular las respectivas esperanzas en el paso E.

[25] Porque s_t y y_l, $l > t$ son independientes dado s_{t+1}.

- $P_\theta(s_{t,i} \mid y^t)$

Para t=1

$P_\theta(s_{t,i} \mid y^t) = P_0(s_{1,i} \mid y_1)$ el cual será uno de los parámetros iniciales del algoritmo.

Para t≥2

$$P_\theta(s_{t,i} \mid y^t) = \frac{P_\theta(s_{t,i}, y_t, y^{t-1})}{P_\theta(y^t)}$$

$$= \frac{P_\theta(y_t \mid s_{t,i}, y^{t-1})\, P_\theta(s_{t,i} \mid y^{t-1})\, P_\theta(y^{t-1})}{P_\theta(y^t)}$$

$$= \frac{P_\theta(y_t \mid s_{t,i}, y^{t-1})\, P_\theta(s_{t,i} \mid y^{t-1})\, P_\theta(y^{t-1})}{\sum_{s_t} P_\theta(y_t \mid s_t, y^{t-1})\, P_\theta(s_t \mid y^{t-1})\, P_\theta(y^{t-1})}$$

$$= \frac{P_\theta(y_t \mid s_{t,i}, y_{t-1})\, P_\theta(s_{t,i} \mid y^{t-1})}{\sum_{s_t} P_\theta(y_t \mid s_t, y_{t-1})\, P_\theta(s_t \mid y^{t-1})}$$

$$= \frac{P_\theta(y_t \mid s_{t,i}, y_{t-1})\, P_\theta(s_{t,i} \mid y^{t-1})}{\sum_{l} P_\theta(y_t \mid s_{t,l}, y_{t-1})\, P_\theta(s_{t,l}, y^{t-1})}$$

$$= \sum_k \frac{P_0(y_{t,k} \mid s_{t,i}, y_{t-1})\, P_\theta(s_{t,i} \mid y^{t-1})}{\sum_{l} P_\theta(y_{t,l} \mid s_{t,l}, y_{t-1})\, P_\theta(s_{t,l} \mid y^{t-1})}$$

$$= \sum_j \sum_k \frac{P_\theta(y_{t,k} \mid s_{t,i}, y_{t-1,j})\, P_\theta(s_{t,i} \mid y^{t-1})}{\sum_{l} P_\theta(y_{t,l} \mid s_{t,l}, y_{t-1,j})\, P_\theta(s_{t,l} \mid y^{t-1})}$$

De donde sigue que

$$P_0(s_{t,i} \mid y^t) = \sum_j \sum_k \frac{P_{(ji)k}\, P_\theta(s_{t,i} \mid y^{t-1})}{\sum_l P_{(jl)k}\, P_\theta(s_{t,l} \mid y^{t-1})}$$

- $P_\theta(s_{t,j} \mid y^{t-1})$

Para t=1

$P_0(s_{1,j} \mid y^0) = P_0(s_{1,j})$ que será uno de los parámetros iniciales del algoritmo.

Para t≥2

$$P_\theta(s_{t,j} \mid y^{t-1}) = \frac{P_\theta(s_{t,j}, y^{t-1})}{P_\theta(y^{t-1})} \sum_{s_{t-1}} \frac{P_\theta(s_{t,j} \mid s_{t-1}, y^{t-1})}{P_\theta(y^{t-1})}$$

$$= \sum_{s_{t-1}} P_\theta(s_{t,j} \mid s_{t-1}, y^{t-1})\, P_\theta(s_{t-1} \mid y^{t-1})$$

$$= \sum_j P_\theta(s_{t,j} \mid s_{t-1,i})\, P_\theta(s_{t-1,i} \mid y^{t-1})$$

$$= \sum_j \lambda_{ij}\, P_\theta(s_{t-1,i} \mid y^{t-1})$$

De esta forma se obtiene un sistema de ecuaciones recursivo que depende del parámetro θ para t ≥ 2 que es el siguiente:

$$\begin{aligned}
&(i) P_\theta(s_{t,i} \mid y^m) = \sum_j \frac{\lambda_{ij}\, P_\theta(s_{t,i} \mid y^t)\, P_\theta(s_{t+1,j} \mid y^m)}{P_\theta(s_{t+1,j} \mid y^t)} \\
&(ii) P_\theta(s_{t,i} \mid y^t) = \sum_j \sum_k \frac{P_{(ji)k}\, P_\theta(s_{t,i} \mid y^{t-1})}{\sum_l P_{(jl)k}\, P_\theta(s_{t,l} \mid y^{t-1})} \\
&(iii) P_\theta(s_{t,j} \mid y^{t-1}) = \sum_i \lambda_{ij}\, P_\theta(s_{t-1,i} \mid y^{t-1})
\end{aligned} \qquad \text{...(4.17)}$$

Ahora se calcula

- $E_\theta^{(0)}[S_{t,j},S_{t-1,i} \mid y^m]$. Como en la esperanza anterior se trabaja con $P_\theta(S_{t,j},S_{t-1,i} \mid y^m)$ para $t \geq 2$.

$$P_\theta(S_{t,j},S_{t-1,i} \mid y^m) = \frac{P_\theta(S_{t,j},S_{t-1,i},y^m)}{P_\theta(y^m)}$$

$$= P_\theta(S_{t-1,i} \mid S_{t,j},y^m)\, P_\theta(S_{t,j} \mid y^m)$$

$$=^{(26)} P_\theta(S_{t-1,i} \mid S_{t,j},y^{t-1})\, P_\theta(S_{t,j} \mid y^m)$$

$$= \frac{P_\theta(S_{t-1,i},S_{t,j},y^{t-1})\, P_\theta(S_{t,j} \mid y^m)}{P_\theta(S_{t,j},y^{t-1})}$$

$$= \frac{P_\theta(S_{t-1,i},S_{t,j} \mid y^{t-1}) P_\theta(S_{t,j} \mid y^m)}{P_\theta(S_{t,j} \mid y^{t-1})}$$

$$= \frac{P_\theta(S_{t,j} \mid S_{t-1,i},y^{t-1}) P_\theta(S_{t-1,i} \mid y^{t-1})\, P_\theta(S_{t,j} \mid y^m)}{P_\theta(S_{t,j} \mid y^{t-1})}$$

$$= \frac{P_\theta(S_{t,j} \mid S_{t-1,i})\, P_\theta(S_{t-1,i} \mid y^{t-1})\, P_\theta(S_{t,j} \mid y^m)}{P_\theta(S_{t,j} \mid y^{t-1})}$$

$$= \frac{\lambda_{ij}\, P_\theta(S_{t-1,i} \mid y^{t-1})\, P_\theta(S_{t,j} \mid y^m)}{P_\theta(S_{t,j} \mid y^{t-1})}$$

[26] Dado que s_t, s_{t-1} dependen solamente de $y^t = (y_1, y_2, \ldots, y_t)$

En donde las cantidades que aparecen en esta expresión son las dadas por el parámetro θ y por la solución de las ecuaciones recursivas (4.17). De este modo se obtienen los valores de $E_{\theta^{(0)}}[S_{t,i} \mid y^{m}]$ y $E_{\theta^{(0)}}[S_{t,j}, S_{t-1,i} \mid y^{m}]$ y se utilizan estos valores en el paso M del algoritmo que consiste en utilizar los valores esperados encontrados en sustitución de S_t y de $S_t S_{t-1}$ obteniéndose de esta forma los valores de $\theta^{(1)}$ que son utilizados en la siguiente iteración del algoritmo E.M.

4.2.2 Caso particular del algoritmo recursivo.

Para dejar un poco más claro el mecanismo que se sigue para encontrar las probabilidades a través del sistema de ecuaciones recursivas, tómese el caso para el tamaño de una secuencia m = 2 para el primer paso del algoritmo E.M.

Desde un principio se asume conocido $P(s_1) = \pi^{(0)}$ y el parámetro $\theta^{(0)}$ como primera aproximación del paso esperanza. Ahora bien, para encontrar $P_{\theta}(S_{t,i} \mid y^{m})$ para t = 1, 2 y m = 2 se tiene de (4.17, ***(i)***) para t = 1

$$P_{\theta^{(0)}}(S_{1,i} \mid y^{2}) = \sum_{j} \frac{\lambda_{ij}^{(0)} P_{\theta^{(0)}}(S_{1,i} \mid y_1) P_{\theta^{(0)}}(S_{2,j} \mid y^{2})}{P_{\theta^{(0)}}(S_{2,j} \mid y_1)} \quad .....(4.18)$$

nótese que hace falta conocer $P_{\theta^{(0)}}(S_{1,i} \mid y_1)$, $P_{\theta^{(0)}}(S_{2,j} \mid y^{2})$ y $P_{\theta^{(0)}}(S_{2,j} \mid y_1)$, tome a $P_{\theta^{(0)}}(S_{1,i} \mid y_1)$ para empezar; utilizando (4.17, ***(ii)***) de las ecuaciones recursivas, para t = 1 se obtiene

$$P_{\theta^{(0)}}(S_{1,i} \mid y_1) = \sum_{j} \sum_{k} \frac{P_{(ji)k}^{(0)} P_{\theta^{(0)}}(S_{1,i} \mid y^{0})}{\sum_{l} P_{(jl)k}^{(0)} P_{\theta^{(0)}}(S_{1,l} \mid y^{0})}$$

note que $P_{\theta^{(0)}}(S_{1,i} \mid y^{0})$ y $P_{\theta^{(0)}}(S_{1,l} \mid y^{0})$ son dados al inicio del algoritmo donde $P_{0^{(0)}}(S_{1,i}) = \pi_i^{(0)}$ con i = 1, 2, de donde resulta que

$$P_{\theta^{(0)}}(s_{1,i}\,|\,y_1) = \sum_j \sum_k \frac{P^{(0)}_{(ji)k}\ \pi_i^{(0)}}{\sum_l P^{(0)}_{(jl)k}\ \pi_l^{(0)}} \qquad \text{.....(4.19)}$$

Tome a (4.17, ***(ii)***) para encontrar

$$P_{\theta^{(0)}}(s_{2,j}\,|\,y^2) = \sum_r \sum_u \frac{P^{(0)}_{(rj)u}\ P_{\theta^{(0)}}(s_{2,j}\,|\,y_1)}{\sum_s P^{(0)}_{(rs)u}\ P_{\theta^{(0)}}(s_{2,s}\,|\,y_1)} \qquad \text{.....(4.20)}$$

note que sólo hace falta encontrar $P_{\theta^{(0)}}(s_{2,j}\,|\,y_1)$ que se encuentra tanto en (4.18) como en (4.20). Tome a (4.17, ***(iii)***) para encontrar $P_{\theta^{(0)}}(s_{2,j}\,|\,y_1)$ es decir

$$P_{\theta^{(0)}}(s_{2,j}\,|\,y_1) = \sum_i \lambda_{ij}^{(0)}\ P_{\theta^{(0)}}(s_{1,i}\,|\,y_1)$$

y como $P_{\theta^{(0)}}(s_{1,i}\,|\,y_1)$ se conoce de (4.19) se tiene que

$$P_{\theta^{(0)}}(s_{2,j}\,|\,y_1) = \sum_i \lambda_{ij}^{(0)} \left(\sum_j \sum_k \frac{P^{(0)}_{(ji)k}\ \pi_i^{(0)}}{\sum_l P^{(0)}_{(jl)k}\ \pi_l^{(0)}} \right) \qquad \text{.....(4.21)}$$

y (4.20) resulta

$$P_{\theta^{(0)}}(s_{2,j}\,|\,y^2) = \sum_r \sum_u \frac{P^{(0)}_{(rj)u} \left(\sum_i \lambda_{ij}^{(0)} \left(\sum_j \sum_k \frac{P^{(0)}_{(ji)k}\ \pi_i^{(0)}}{\sum_l P^{(0)}_{(jl)k}\ \pi_l^{(0)}} \right) \right)}{\sum_s P^{(0)}_{(rs)u} \left(\sum_i \lambda_{is}^{(0)} \left(\sum_s \sum_k \frac{P^{(0)}_{(si)k}\ \pi_i^{(0)}}{\sum_l P^{(0)}_{(sl)k}\ \pi_l^{(0)}} \right) \right)}$$

de tal modo que (4.18) resulta

$$P_{\theta}^{(0)}(s_{1,i}|\, y^2) =$$

$$\frac{\displaystyle\sum_{j} \lambda_{ij}^{(0)} \left(\sum_{l} \sum_{k} \frac{P_{(ji)k}^{(0)}\, \pi_i^{(0)}}{\displaystyle\sum_{l} P_{(jl)k}^{(0)}\, \pi_l^{(0)}} \right) \left(\sum_{r} \sum_{u} P_{(rj)u}^{(0)} \left(\sum_{i} \lambda_{ij}^{(0)} \left(\sum_{j} \sum_{k} \frac{P_{(ji)k}^{(0)}\, \pi_i^{(0)}}{\displaystyle\sum_{l} P_{(jl)k}^{(0)}\, \pi_l^{(0)}} \right) \right) \right) \left(\sum_{s} P_{(ru)s}^{(0)} \left(\sum_{i} \lambda_{is}^{(0)} \left(\sum_{s} \sum_{k} \frac{P_{(si)k}^{(0)}\, \pi_i^{(0)}}{\displaystyle\sum_{l} P_{(sl)k}^{(0)}\, \pi_l^{(0)}} \right) \right) \right)}{\displaystyle\sum_{i} \lambda_{ij}^{(0)} \left(\sum_{j} \sum_{k} \frac{P_{(ji)k}^{(0)}\, \pi_i^{(0)}}{\displaystyle\sum_{l} P_{(jl)k}^{(0)}\, \pi_l^{(0)}} \right)}$$

Por último se calcula $P_{\theta}^{(0)}(s_{t,j},\, s_{t-1,i}|\, y^m)$

$$P_{\theta}(s_{t,j},\, s_{t-1,i}|\, y^m) = \frac{\lambda_{ij}^{(0)}\, P_{\theta}(s_{t-1,i}\,|\, y^{t-1})\, P_{\theta}(s_{t,j}\,|\, y^{(m)})}{P_{\theta}(s_{t,j}\,|\, y^{t-1})}$$

para t = 2

$$P_{0}^{(0)}(s_{2,j},\, s_{1,i}|\, y^2) = \frac{\lambda_{ij}^{(0)}\, P_{\theta}^{(0)}(s_{1,i}\,|\, y_1)\, P_{\theta}^{(0)}(s_{2,j}\,|\, y^2)}{P_{0}^{(0)}(s_{2,j}\,|\, y_1)} \qquad \text{.....(4.22)}$$

y como ya se conocen los valores de $P_{\theta}^{(0)}(s_{1,i}\,|\, y_1)$ de (4.19), $P_{\theta}^{(0)}(s_{2,j}\,|\, y^2)$ de (4.20) y $P_{0}^{(0)}(s_{2,j}\,|\, y_1)$ de (4.21), (4.22) resulta

$$P_{\theta}^{(0)}(s_{2j}\, s_{1,i} \mid y^2) =$$

$$\frac{\lambda_{ij}^{(0)}\left(\sum_{j}\sum_{k}\frac{P_{(ji)k}^{(0)}\,\pi_i^{(0)}}{\sum_{l} P_{(jl)k}^{(0)}\,\pi_l^{(0)}}\right)\left(\sum_{r}\sum_{u} P_{(ru)}^{(0)}\left(\sum_{i}\lambda_{ij}^{(0)}\left(\sum_{j}\sum_{k}\frac{P_{(ji)k}^{(0)}\,\pi_i^{(0)}}{\sum_{l} P_{(jl)k}^{(0)}\,\pi_l^{(0)}}\right)\right)\sum_{s} P_{(ru)s}^{(0)}\left(\sum_{i}\lambda_{is}^{(0)}\left(\sum_{s}\sum_{k}\frac{P_{(si)k}^{(0)}\,\pi_i^{(0)}}{\sum_{l} P_{(sl)k}^{(0)}\,\pi_l^{(0)}}\right)\right)\right)}{\sum_{i}\lambda_{ij}^{(0)}\left(\sum_{j}\sum_{k}\frac{P_{(ji)k}^{(0)}\,\pi_i^{(0)}}{\sum_{l} P_{(jl)k}^{(0)}\,\pi_l^{(0)}}\right)}$$

Note que las ecuaciones que se obtuvieron se encuentran en términos de valores que ya se conocen, fueron dados al principio, con el algoritmo de actualización y la primera iteración del paso E. Con esto se termina de ilustrar la forma de encontrar cada una de las esperanzas que se buscan en el paso E. Ahora se continua con el paso M del algoritmo E.M. en general para inferir la secuencia verdadera de A.D.N.

Paso M.

Maximizar $G(\theta;\theta^{(0)})$, es decir:

Obtener la derivada con relación a θ, igualar a cero y se encontrarán los estimadores máximo verosímiles. En este caso, ya se obtuvieron tanto las derivadas como los estimadores en la sección 4.2.1 y que están dados por (4.12).

De este modo se reemplazan los valores respectivos de $\theta^{(0)}$ en la fórmula para $G(\theta;\theta^{(0)})$ y de ahí se obtiene el vector $\theta^{(1)}$ que será utilizado en el paso E de la próxima iteración del algoritmo E.M. Así se tiene que $\theta^{(1)} = \{\pi^{(1)},\ \lambda_{ij}^{(1)},\ P_{ij}^{(1)}, P_{(ji)k}^{(1)},\ i, j, k \in \{1, 2, 3, 4\}\}$ esta dado por

$$\pi^{(1)} = (E_0^{(0)}[S_{t,i} \mid y^m], E_\theta^{(0)}[S_{t,j}, S_{t-1,i} \mid y^m])$$

$$\lambda_{ij}^{(1)} = \frac{\sum_{t=2}^{m} E_0^{(0)}[S_{t,j}, S_{t-1,i} \mid y^m]}{\sum_{t=2}^{m} E_\theta^{(0)}[S_{t-1,i} \mid y^m]}$$

$$P_{ij}^{(1)} = y_1$$

$$P_{(ji)k}^{(1)} = \frac{\sum_{t=2}^{m} y_{t-1,j}\, y_{t,k}\, E_\theta^{(0)}[S_{t,i} \mid y^m]}{\sum_{t=2}^{m} y_{t-1,j}\, E_\theta^{(0)}[S_{t,i} \mid y^m]}$$

donde $E_\theta^{(0)}[S_{t,i} \mid y^m] = P_\theta^{(0)}(s_{t,i} \mid y^m)$ o bien:

$$P_\theta^{(0)}(s_{t,i} \mid y^2) =$$

$$\frac{\sum_i \lambda_{ij}^{(0)} \left(\sum\sum_{j\ k} \frac{P_{(ij)k}^{(0)}\, \pi_i^{(0)}}{\sum_l P_{(jl)k}^{(0)}\, \pi_l^{(0)}} \right) \left(\sum\sum_{r\ u} P_{(rj)u}^{(0)} \left(\sum_i \lambda_{ij}^{(0)} \left(\sum\sum_{j\ k} \frac{P_{(ji)k}^{(0)}\, \pi_i^{(0)}}{\sum_l P_{(jl)k}^{(0)}\, \pi_l^{(0)}} \right) \right) \sum_s P_{(rs)s}^{(0)} \left(\sum_i \lambda_{is}^{(0)} \left(\sum\sum_{s\ k} \frac{P_{(si)k}^{(0)}\, \pi_i^{(0)}}{\sum_l P_{(sl)k}^{(0)}\, \pi_l^{(0)}} \right) \right) \right)}{\sum_i \lambda_{ij}^{(0)} \left(\sum\sum_{j\ k} \frac{P_{(ij)k}^{(0)}\, \pi_i^{(0)}}{\sum_l P_{(jl)k}^{(0)}\, \pi_l^{(0)}} \right)}$$

y $E_{\theta^{(0)}}[S_{t,j},S_{t-1,i} \mid y^m] = P_{\theta^{(0)}}(S_{t,j},S_{t-1,i} \mid y^m)$ o bien:

$$P_{\theta}^{(0)}(s_{t,j}\, s_{t-1,i} \mid y^2) =$$

$$\frac{\lambda_{ij}^{(0)}\left(\sum_{j}\sum_{k}\frac{P_{(ij)k}^{(0)}\,\pi_i^{(0)}}{\sum_{l}P_{(jl)k}^{(0)}\,\pi_l^{(0)}}\right)\left(\sum_{r}\sum_{u}P_{(\eta)u}^{(0)}\left(\sum_{l}\lambda_{ij}^{(0)}\left(\sum_{j}\sum_{k}\frac{P_{(jl)k}^{(0)}\,\pi_i^{(0)}}{\sum_{l}P_{(jl)k}^{(0)}\,\pi_l^{(0)}}\right)\right)\sum_{s}P_{(ru)s}^{(0)}\left(\sum_{i}\lambda_{is}^{(0)}\left(\sum_{s}\sum_{k}\frac{P_{(sij)k}^{(0)}\,\pi_i^{(0)}}{\sum_{l}P_{(sl)k}^{(0)}\,\pi_l^{(0)}}\right)\right)\right)}{\sum_{i}\lambda_{ij}^{(0)}\left(\sum_{j}\sum_{k}\frac{P_{(ij)k}^{(0)}\,\pi_i^{(0)}}{\sum_{l}P_{(jl)k}^{(0)}\,\pi_l^{(0)}}\right)}$$

donde $\pi_i = P(s_{1,i})$

En general para la n-ésima iteración del paso E y M se tiene:

n-ésima iteración del paso E.

Se obtienen los valores de $E_{\theta^{(n-1)}}[S_{t,i} \mid y^m] = P_{\theta^{(n-1)}}[S_{t,i} \mid y^m]$ y de $E_{\theta^{(n-1)}}[S_{t,j}\ S_{t-1,i} \mid y^m] = P_{\theta^{(n-1)}}[S_{t,j}\ S_{t-1,i} \mid y^m]$ que estarían dados por el sistema de ecuaciones recursivo y el algoritmo de actualización, como en la primera iteración del algoritmo E.M., y se utiliza este valor para obtener $\theta^{(n)}$ en la

n-ésima iteración del Paso M

$$\pi^{(n)} = (E_{\theta^{(n-1)}}[S_{1,1} \mid y^m], E_{\theta^{(n-1)}}[S_{1,2} \mid y^m])$$

$$\lambda_{ij}^{(n)} = \frac{\sum_{t=2}^{m} E_0^{(n-1)}[S_{t,j}, S_{t-1,i} \mid y^m]}{\sum_{t=2}^{m} E_0^{(n-1)}[S_{t-1,i} \mid y^m]}$$

$$P_{ij}^{(n)} = y_1$$

$$P_{(ji)k}^{(n)} = \frac{\sum_{t=2}^{m} y_{t-1,j}\, y_{t,k}\, E_\theta^{(n-1)}[S_{t,i} \mid y^m]}{\sum_{t=2}^{m} y_{t-1,j}\, E_\theta^{(n-1)}[S_{t,1} \mid y^m]}$$

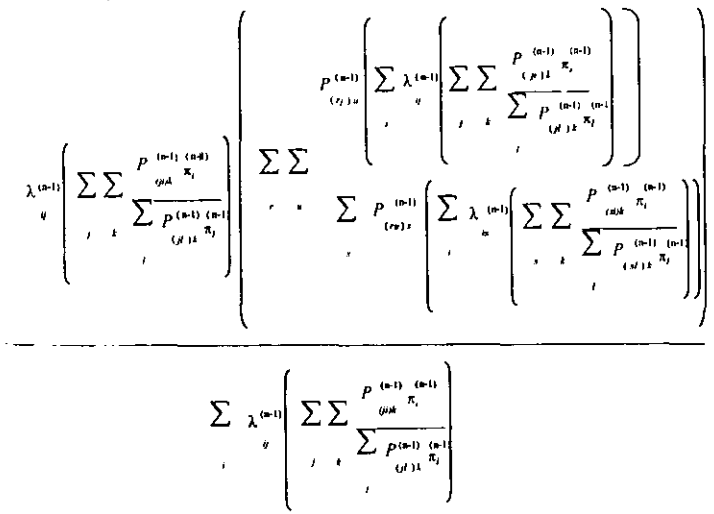
donde $E_\theta^{(n-1)}[S_{t,i} \mid y^m]$ está dada por

$$P_\theta^{(0)}(s_{t,i} \mid y^2) =$$

$$\frac{\sum_{j} \lambda_{ij}^{(0)} \left(\sum_{j} \sum_{k} \frac{P_{(ji)k}^{(0)}\, \pi_t^{(0)}}{\sum_{l} P_{(jl)k}^{(0)}\, \pi_l^{(0)}} \right) \left(\sum_{r} \sum_{u} P_{(\eta)u}^{(0)} \left(\sum_{i} \lambda_{ij}^{(0)} \left(\sum_{j} \sum_{k} \frac{P_{(ji)k}^{(0)}\, \pi_t^{(0)}}{\sum_{l} P_{(jl)k}^{(0)}\, \pi_l^{(0)}} \right) \right) \right)}{\sum_{i} \lambda_{ij}^{(0)} \left(\sum_{j} \sum_{k} \frac{P_{(ji)k}^{(0)}\, \pi_t^{(0)}}{\sum_{l} P_{(jl)k}^{(0)}\, \pi_l^{(0)}} \right)}$$

y $E_0^{(n-1)}[S_{t,j}, S_{t-1,i} \mid y^m]$ por

$$P_{\theta}^{(n-1)}(s_{t,j}, s_{(t-1),i} \mid y^m) =$$



Repitiendo estos pasos sucesivamente se consigue el estimador θ^* que maximiza a la verosimilitud. Una vez que se encuentra θ^* se utiliza para obtener una estimación de la sucesión de A.D.N. $S^n = (s_1, ..., s_m)$ para lo cual se utiliza el siguiente algoritmo de actualización:

4.2.3 Algoritmo de actualización.

(a) empiece con una distribución inicial $P(s_1) = \pi^*$

(b) utilice este resultado para obtener $P(s_t \mid y^{t-1})$, $P(s_t \mid y^t)$ $t \geq 2$ del algoritmo recursivo.

(c) use $P(s_m \mid y^m)$ y $P(s_m \mid y^{m-1})$ para calcular $P(s_{m-1} \mid y^m)$.

(d) aplicar el paso (c) sucesivamente hasta completar la secuencia, es decir, hasta obtener $P(s_1 \mid y^m)$.

Observación:

(1) Los valores utilizados en el algoritmo de actualización son el máximo verosímil y los obtenidos a través del algoritmo recursivo son los parámetros dados por el estimador máximo verosímil θ^*.

Apéndice A.

"Y lo que oscila en apariencias fluctuantes, fijadlo en ideas duraderas"
Jaques Monod, El azar y la necesidad (1971)

A.1 Concavidad

Definición A.1 Una función f diferenciable en un intervalo (a,b) se dice que es *cóncava hacia arriba* en (a,b) si y sólo si f'es creciente en (a,b). La función f es *cóncava hacia abajo* en (a,b) si y sólo si $f$es decreciente en (a,b).

Teorema A.1 Sea f una función doblemente derivable en (a,b). Entonces, si
a) $f''(x) > 0$ para toda $x \in$ (a, b) implica que f es cóncava hacia arriba en (a, b).
b) $f''(x) < 0$ para toda $x \in$ (a, b) implica que f es cóncava hacia abajo en (a, b).

Demostración. Como $f'' > 0$ en (a, b), entonces f'es creciente en (a, b) y por la definición anterior en el inciso a), se infiere que f es cóncava hacia arriba en (a, b). El inciso b) se demuestra de forma análoga.

A.2 Desigualdad de Jensen.

Si $f(x)$ es una función cóncava hacia abajo y X una variable aleatoria, entonces $E(f(X)) \leq f(E(X))$, si las esperanzas existen y son finitas.

Demostración.

Expandiendo $f(x)$ como serie de Taylor, en una vecindad de $\mu = E(X)$.

$$f(x) = f(\mu) + f'(\mu)(x-\mu) + \frac{f''(\xi)(x-\xi)^2}{2!} \qquad \text{donde } \xi \in (x, \mu)$$

Como $f''(\xi) \leq 0$, se obtiene:

$$f(x) \leq f(\mu) + f'(\mu)(x - \mu)$$

entonces

$$f(X) \leq f(\mu) + f'(\mu)(X - \mu)$$

tomando esperanzas:

$$E(f(X)) \leq E(f(\mu)) + E(f'(\mu)(x - \mu))$$

$$E(f(X)) \leq f(\mu) + f'(\mu)\, E(X - \mu)$$

$$E(f(X)) \leq f(\mu) + f'(\mu)\,[E(X) - E(\mu)]$$

como $E(X) = \mu$ entonces

$$E(f(X)) \leq f(\mu) + f'(\mu)\,[\mu - \mu]$$

$$E(f(X)) \leq f(\mu)$$

$$E(f(X)) \leq f(E(X))$$

que es lo que se quería demostrar.

■

A.3 Distribución multinomial

La distribución multinomial es una extensión natural de la distribución Bernoulli, el cual consiste en experimentos que sólo

tienen dos diferentes posibilidades de ocurrir, llamados éxito o fracaso. En el caso de la multinomial un solo experimento puede producir uno de $\mathbf{k} \geq 3$ diferentes resultados. Algunos ejemplos de experimentos multinomiales serían:

1. El lanzamiento de un dado; que tiene $\mathbf{k} = 6$ diferentes resultados.

2. La respuesta de una persona a la pregunta de ¿por quién votó en estas elecciones? está, por ejemplo, en $\mathbf{k} = 7$ diferentes posibilidades: (a) Votó por Labastida, (b) Votó por Fox, (c) Votó por Muñoz Ledo, (d) Votó por Cárdenas, (e) Votó por Rincón Gallardo, (f) No votó y (g) Anuló su voto.

Un único experimento multinomial tiene **k** diferentes posibles resultados, en la notación comúnmente usada **k** parámetros son utilizados para describir un experimento multinomial y estos estan definidos como p_i = P(resultado i ocurra) , i = 1,2,...,**k** donde $p_1 + p_2 + ... + p_k = 1$.

Ahora suponga que un experimento consiste en **n** intentos independientes de acuerdo con un modelo multinomial, para cada uno de los intentos los parámetros son $p_1, p_2, ..., p_k$. Para el caso general, se realizan **n** intentos independientes de tipo multinomial, Toman probabilidades $p_1, p_2, ..., p_k$ donde los diferentes resultados posibles de suceso y el vector aleatorio multinomial es (Y_1, Y_2, ...,Y_k) donde Y_j = el número de veces que el resultado j ocurre en los **n** intentos, j = 1, 2, ..., **k**.

La función de probabilidad para (Y_1,Y_2, ...,Y_k) es

$$P(Y_1 = y_1, Y_2 = y_2, ..., Y_k = y_k) = \frac{n! p_1^{y_1} p_2^{y_2} ... p_k^{y_k}}{y_1! y_2! ... y_k!}$$

donde $\mathbf{n} = 1, 2, 3, ... ; 0 < p_j < 1 ; \sum_{j=1}^{k} p_j = 1; \quad y_j = 0, 1, ..., \mathbf{n}.; \sum_{j=1}^{k} y_j = \mathbf{n}.$

La función de probabilidad marginal para un y_j debe ser binomial con parámetros **n** y p_j. De donde se tiene que:

$$E(Y_j) = \mathbf{n}\, p_j, \text{ Var } (Y_j) = \mathbf{n}\, p_j (1 - p_j) \text{ con } j = 1, 2, ..., \mathbf{k}.$$

Por ejemplo, tome el caso para k = 3, donde y_1, y_2, y_3, se distribuyen de forma trinomial con parámetros p_1, p_2 y p_3 respectivamente. Y_1, Y_2, Y_3 variables aleatorias; $n = y_1 + y_2 + y_3$, entonces se tiene que

$$f(y_1, y_2) = \frac{n!}{y_1! y_2! (n - y_1 - y_2)!} p_1^{y_1} p_2^{y_2} (1 - p_1 - p_2)^{n - y_1 - y_2}$$

$$= \frac{n!}{y_1!(n - y_1)!} p_1^{y_1} \frac{(n - y_1)!}{y_2!(n - y_1 - y_2)!} p_2^{y_2} (1 - p_1 - p_2)^{n - y_1 - y_2}$$

ahora sumando sobre y_2 para encontrar la marginal de y_1 se tiene:

$$P(Y_1 = y_1) = \sum_{y_2=0}^{n-y_1} \frac{n!}{y_1!(n - y_1)!} p_1^{y_1} \frac{(n - y_1)!}{y_2!(n - y_1 - y_2)!} p_2^{y_2} (1 - p_1 - p_2)^{n - y_1 - y_2}$$

$$= \frac{n!}{y_1!(n - y_1)!} p_1^{y_1} \sum_{y_2=0}^{n-y_1} \frac{(n - y_1)!}{y_2!(n - y_1 - y_2)!} p_2^{y_2} (1 - p_1 - p_2)^{n - y_1 - y_2}$$

$$= \frac{n!}{y_1!(n - y_1)!} p_1^{y_1} (1 - p_1)^{n - y_1}$$

donde $P(Y_1 = y_1)$ se distribuye de forma binomial con parámetro n y p_1. De forma análoga se demuestra para Y_2.

A.4 Cálculo de estimadores máximo verosímiles.

$$P(y^m,s^m) = P(s_1)\prod_{t=2}^{m} P(s_t \mid s_{t-1})\, P(y_1 \mid s_1) \prod_{t=2}^{m} P(y_t \mid y_{t-1}, s_{t-1})$$

$s_t = (s_{t,1}, s_{t,2})$

$y_t = (y_{t,1}, y_{t,2})$

(i) $P(s_1) = \prod_{j=1}^{2} \pi_j^{s_{1,j}} = \pi_1^{s_{1,1}} (1-\pi_1)^{s_{1,2}}$

(ii) $\prod_{t=2}^{m} P(s_t \mid s_{t-1})$

$$= \prod_{i=1}^{2} \prod_{j=1}^{2} \lambda_{ij}^{s_{t,j} s_{t-1,i}} = \prod_{i=1}^{2} \lambda_{i1}^{s_{t,1} s_{t-1,i}} (1 - \lambda_{i1})^{s_{t,2} s_{t-1,i}}$$

(iii) $P(y_1 \mid s_1)$

$$= \prod_{i=1}^{2} \prod_{k=1}^{2} p_{ik}^{s_{1,i} y_{1,k}} = \prod_{i=1}^{2} p_{i1}^{s_{1,i} y_{1,1}} (1 - p_{i1})^{s_{1,i} y_{1,2}}$$

(iv) $\prod_{t=2}^{m} P(y_t \mid y_{t-1}, s_{t-}) = \prod_{i=1}^{2} \prod_{j=1}^{2} \prod_{k=1}^{2} p_{(ij)k}^{s_{t,i} y_{t-1,j} y_{t,k}} =$

$$\prod_{i=1}^{2} \prod_{j=1}^{2} p_{(ij)1}^{s_{t,i} y_{t-1,j} y_{t,1}} (1 - p_{(ij)1})^{s_{t,i} y_{t-1,j} y_{t,2}}$$

de donde se obtienen los logaritmos de cada uno por separado, para *(i)* la primera ecuación:

$$\log P(s_1) = s_{1,1} \log \Pi_1 + s_{1,2} \log(1-\Pi_1)$$

$$\prod_{t=2}^{m} \log P(s_t \mid s_{t-1}) = \sum_{i=1}^{2} [s_{t,1} s_{t-1,i} \log \lambda_{i1} + s_{t,2} s_{t-1,i} \log(1 - \lambda_{i1})]$$

$$\log P(y_1 \mid s_1) = \sum_{i=1}^{2} [s_{1,i} y_{1,1} \log \; p_{i1} + s_{1,i} y_{1,2} \log(1 - p_{i1})]$$

$$\prod_{t=2}^{m} \log P(y_t \mid y_{t-1}, s_{t-1}) =$$

$$\sum_{i=1}^{2} \sum_{j=1}^{2} (s_{t,i} y_{t-1,j} y_{t,1} \log \; p_{(ij)1} + s_{t,i} y_{t-1,j} y_{t,2} \log(1 - p_{(ij)1}))$$

Se obtienen las derivadas con respecto a los parámetros a encontrar:

$$\frac{d \log P(s_1)}{d\Pi_1} = \frac{s_{1,1}}{\Pi_1} - \frac{s_{1,2}}{(1-\Pi_1)}$$

igualando a cero la ecuación:

$$(1-\Pi_1)s_{1,1} - \Pi_1 s_{1,2} = 0$$

$$-\Pi_1(s_{1,1} + s_{1,2}) = -s_{1,1}$$

$\Pi_1(s_{1,1} + s_{1,2}) = s_{1,1}$ de donde se obtiene que $\Pi_1 = s_{1,1}$ dado que $s_{1,1} + s_{1,2} = 1$. Como $\Pi_2 = 1 - \Pi_1$, entonces $\Pi_2 = s_{1,2}$ y que

$$\Pi = (\Pi_1, \Pi_2) = (s_{1,1}, s_{1,2}) = s_1 \qquad \text{.....(A.4.1)}$$

Para *(ii)* la segunda ecuación se tiene:

$$\log \prod_{t=2}^{m} P(s_t \mid s_{t-1}) = \sum_{t=2}^{m} (s_{t,1} s_{t-1,1} \log \lambda_{11} + s_{t,2} s_{t-1,1} \log(1 - \lambda_{11}) + s_{t,1} s_{t-1,2} \log \lambda_{21} + s_{t,1} s_{t-1,2} \log(1 - \lambda_{21})$$

$$\frac{d \log \prod_{t=2}^{m} P(s_t / s_{t-1})}{d\lambda_{1,1}} = \sum_{t=2}^{m} \left(\frac{s_{t,1} s_{t-1,1}}{\lambda_{11}} - \frac{s_{t,2} s_{t-1,1}}{1 - \lambda_{11}} \right)$$

igualando a cero la ecuación y resolviéndola se tiene:

$$(1-\lambda_{11})\sum_{t=2}^{m} s_{t,1}s_{t-1,1} - \lambda_{11}\sum_{t=2}^{m} s_{t,2}s_{t-1,1} = 0$$

$$-\lambda_{11}[\sum_{t=2}^{m} s_{t,1}s_{t-1,1} + \sum_{t=2}^{m} s_{t,2}s_{t-1,1}] = -\sum_{t=2}^{m} s_{t,1}s_{t-1,1}$$

y por lo tanto

$$\lambda_{11} = \frac{\sum_{t=2}^{m} s_{t,1}s_{t-1,1}}{\sum_{t=2}^{m} s_{t-1,1}(s_{t,1}+s_{t,2})}$$

Como $s_{t,1} + s_{t,2}=1$ se tiene que

$$\lambda_{11} = \frac{\sum_{t=2}^{m} s_{t,1}s_{t-1,1}}{\sum_{t=2}^{m} s_{t-1,1}}$$

$$\lambda_{12} = 1-\lambda_{11} = 1-\frac{\sum_{t=2}^{m} s_{t,1}s_{t-1,1}}{\sum_{t=2}^{m} s_{t-1,1}} = \frac{\sum_{t=2}^{m} s_{t-1,1} - \sum_{t=2}^{m} s_{t,1}s_{t-1,1}}{\sum_{t=2}^{m} s_{t-1,1}}$$

$$= \frac{\sum_{t=2}^{m} s_{t-1,1}(1-s_{t,1})}{\sum_{t=2}^{m} s_{t-1,1}}$$

y como $1-s_{t,1}= s_{t,2}$. De este modo se tiene:

$$\lambda_{12} = \frac{\sum_{t=2}^{m} s_{t-1,1}s_{t,2}}{\sum_{t=2}^{m} s_{t-1,1}}.$$

De las expresiones para λ_{11} y λ_{12} se puede escribir

$$\lambda_{1j} = \frac{\sum_{t=2}^{m} s_{t-1,1}s_{t,j}}{\sum_{t=2}^{m} s_{t-1,1}}. \qquad \text{.....(A.4.2)}$$

Ahora se obtiene la derivada con respecto a λ_{21}, así

$$\frac{d \log \prod_{t=2}^{m} P(s_t / s_{t-1})}{d\lambda_{21}} = \sum_{t=2}^{m} \left(\frac{s_{t,1} s_{t-1,2}}{\lambda_{21}} - \frac{s_{t,2} s_{t-1,2}}{1 - \lambda_{21}} \right).$$

Igualando a cero la ecuación anterior y resolviéndola se tiene lo siguiente:

$$(1 - \lambda_{21}) \sum_{t=2}^{m} s_{t,1} s_{t-1,2} - \lambda_{21} \sum_{t=2}^{m} s_{t,2} s_{t-1,2} = 0$$

$$-\lambda_{21} \left[\sum_{t=2}^{m} s_{t,1} s_{t-1,2} + \sum_{t=2}^{m} s_{t,2} s_{t-1,2} \right] = -\sum_{t=2}^{m} s_{t,1} s_{t-1,2}$$

$$\lambda_{21} = \frac{\sum_{t=2}^{m} s_{t,1} s_{t-1,2}}{\sum_{t=2}^{m} s_{t-1,2} (s_{t,1} + s_{t,2})}$$

Como $s_{t,1} + s_{t,2} = 1$ entonces

$$\lambda_{21} = \frac{\sum_{t=2}^{m} s_{t,1} s_{t-1,2}}{\sum_{t=2}^{m} s_{t-1,2}},$$

Y de este modo

$$\lambda_{22} = 1 - \lambda_{21} = \frac{\sum_{t=2}^{m} s_{t-1,2} - \sum_{t=2}^{m} s_{t,1} s_{t-1,2}}{\sum_{t=2}^{m} s_{t-1,2}} = \frac{\sum_{t=2}^{m} s_{t-1,2} (1 - s_{t,1})}{\sum_{t=2}^{m} s_{t-1,2}}$$

y como $1 - s_{t,1} = s_{t,2}$ entonces se tiene:

$$\lambda_{22} = \frac{\sum_{t=2}^{m} s_{t-1,2} s_{t,2}}{\sum_{t=2}^{m} s_{t-1,2}}$$

En resumen el estimador λ_{2j} es dado por

$$\lambda_{2j} = \frac{\sum_{t=2}^{m} s_{t-1,2} s_{t,j}}{\sum_{t=2}^{m} s_{t-1,2}} \qquad \text{.....(A.4.3)}$$

y de (A.4.2) y (A.4.3) se tiene:

$$\lambda_{ij} = \frac{\sum_{t=2}^{m} s_{t-1,i} s_{t,j}}{\sum_{t=2}^{m} s_{t-1,i}} \qquad \text{.....(A.4.4)}$$

Para ***(iii)*** la tercera ecuación se tiene:

$$\log P(y_1 \mid s_1) = s_{1,1} y_{1,1} \log P_{11} + s_{1,1} y_{1,2} \log(1 - P_{11}) + s_{1,2} y_{1,1} \log P_{21} + s_{1,2} y_{1,2} \log(1 - P_{21})$$

Note que si $s_{1,1} = 0$, entonces P_{11} y P_{12} pueden ser cualquier valor que satisfaga $P_{11} + P_{12} = 1$. Análogamente, si $s_{1,2} = 0$ entonces P_{21} y P_{22} pueden ser cualquier valor que satisfaga $P_{21} + P_{22} = 1$. Ahora bien, note que si $s_{1,1} = 0$ entonces $s_{1,2} = 1$ y viceversa, de este modo suponga que $s_{1,1} \neq 0$.

$$\frac{d \log P(y_1 / s_1)}{dP_{11}} = \frac{s_{1,1} y_{1,1}}{P_{11}} - \frac{s_{1,1} y_{1,2}}{1 - P_{11}}$$

Igualando la ecuación anterior a cero y resolviéndola se tiene que:

$$(1 - P_{11}) s_{1,1} y_{1,1} - P_{11} s_{1,1} y_{1,2} = 0$$
$$- P_{11} (s_{1,1} y_{1,1} + s_{1,1} y_{1,2}) = - s_{1,1} y_{1,1}$$

$$P_{11} = \frac{s_{1,1} y_{1,1}}{s_{1,1}(y_{1,1} + y_{1,2})}$$

y como $y_{1,1} + y_{1,2} = 1$ entonces se tiene que los estimadores P_{11} y P_{12} son: $P_{11} = y_{1,1}$ y como $P_{12} = 1 - P_{11} = 1 - y_{11}$ se tiene que $P_{12} = y_{1,2}$

De donde el estimador P_{ij} es dado por

$$P_{1j} = y_{1,j} \qquad \text{.....(A.4.5)}$$

Ahora se obtiene la derivada con respecto a P_{21} y se tiene:

$$\frac{d \log P(y_1 / s_1)}{dP_{21}} = \frac{s_{1,2} y_{1,1}}{P_{21}} - \frac{s_{1,2} y_{1,2}}{1 - P_{21}}$$

igualando la ecuación anterior a cero y resolviéndola se obtiene:

$$(1 - P_{21}) s_{1,2} y_{1,1} - P_{21} s_{1,2} y_{1,2} = 0$$
$$- P_{21} (s_{1,2} y_{1,1} + s_{1,2} y_{1,2}) = - s_{1,2} y_{1,1}$$

$$P_{21} = \frac{s_{1,2} y_{1,1}}{s_{1,2}(y_{1,1} + y_{1,2})}$$

y como $y_{1,1} + y_{1,2} = 1$ entonces se tiene que los estimadores P_{21} y P_{22} son: $P_{21} = y_{1,1}$ y como $P_{22} = 1 - P_{21} = 1 - y_{11}$ se tiene que $P_{22} = y_{1,2}$

De donde el estimador P_{2j} es

$$P_{2j} = y_{1,j} \qquad \text{.....(A.4.6)}$$

De (A.4.5) y (A.4.6) se tiene que el estimador P_{ij} es dado por:

$$P_{ij} = y_{i,j} \qquad \text{.....(A.4.7)}$$

Para *(iv)* la última ecuación se tiene:

$$\log \prod_{t=2}^{m} P(y_t \mid y_{t-1}, s_t)$$

$$= \sum_{t=2}^{m} \sum_{i=1}^{2} (s_{t,i} y_{t-1,1} y_{t,1} \log P_{(1i)1} + s_{t,i} y_{t-1,1} y_{t,2} \log(1 - P_{(1i)1})$$

$$+ s_{t,i} y_{t-1,2} y_{t,1} \log P_{(2i)1}) + s_{t,i} y_{t-1,2} y_{t,2} \log(1 - P_{(2i)1}))$$

$$= \sum_{t=2}^{m} [s_{t,1} y_{t-1,1} y_{t,1} \log P_{(11)1} + s_{t,1} y_{t-1,1} y_{t,2} \log(1 - P_{(11)1} + s_{t,2} y_{t-1,1} y_{t,1} \log P_{(21)1}$$

$$+ s_{t,2} y_{t-1,1} y_{t,2} \log(1 - P_{(21)1} + s_{t,1} y_{t-1,2} y_{t,1} \log P_{(12)1} + s_{t,1} y_{t-1,2} y_{t,2} \log(1 - P_{(12)1}$$

$$+ s_{t,2} y_{t-1,2} y_{t,1} \log P_{(22)1} + s_{t,2} y_{t-1,2} y_{t,2} \log(1 - P_{(22)1}].$$

Ahora se obtienen las derivadas para cada uno de los estimadores que se quieren encontrar:

$$\frac{d \log \prod_{t=2}^{m} P(y_t / y_{t-1}, s_t)}{dP_{(11)1}} = \sum_{t=2}^{m} \left[\frac{s_{t,1} y_{t-1,1} y_{t,1}}{P_{(11)1}} - \frac{s_{t,1} y_{t-1,1} y_{t,2}}{1 - P_{(11)1}} \right].$$

Igualando a cero la ecuación para después resolverla se obtiene:

$$(1 - P_{(11)1}) \sum_{t=2}^{m} s_{t,1} y_{t-1,1} y_{t,1} - P_{(11)1} \sum_{t=2}^{m} s_{t,1} y_{t-1,1} y_{t,2} = 0$$

$$- P_{(11)1} [\sum_{t=2}^{m} s_{t,1} y_{t-1,1} y_{t,1} + \sum_{t=2}^{m} s_{t,1} y_{t-1,1} y_{t,2}] = - \sum_{t=2}^{m} s_{t,1} y_{t-1,1} y_{t,1}$$

de donde se tiene que el primer parámetro $P_{(11)1}$ es:

$$P_{(11)1} = \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,1} y_{t,1}}{\sum_{t=2}^{m} s_{t,1} y_{t-1,1} (y_{t,1} + y_{t,2})}$$

y como $y_{t,1}+y_{t,2}= 1$ se tiene :

$$P_{(11)1} = \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,1} y_{t,1}}{\sum_{t=2}^{m} s_{t,1} y_{t-1,1}}$$

y

$$P_{(11)2} = 1 - P_{(11)1} = 1 - \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,1} y_{t,1}}{\sum_{t=2}^{m} s_{t,1} y_{t-1,1}}$$

$$= \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,1} - \sum_{t=2}^{m} s_{t,1} y_{t-1,1} y_{t,1}}{\sum_{t=2}^{m} s_{t,1} y_{t-1,1}} = \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,1} (1 - y_{t,1})}{\sum_{t=2}^{m} s_{t,1} y_{t-1,1}}$$

y como $y_{t,2} = 1 - y_{t,1}$ el segundo parámetro vendría siendo:

$$P_{(11)2} = \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,1} y_{t,2}}{\sum_{t=2}^{m} s_{t,1} y_{t-1,1}}$$

y en general $P_{(11)k}$ se comporta:

$$P_{(11)k} = \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,1} y_{t,k}}{\sum_{t=2}^{m} s_{t,1} y_{t-1,1}} \qquad \text{.....(A.4.8)}$$

El tercer parámetro por encontrar vendría siendo $P_{(21)1}$ así que se deriva con respecto a él para obtener:

$$\frac{d \log \prod_{t=2}^{m} P(y_t / y_{t-1}, s_t)}{dP_{(21)1}} = \sum_{t=2}^{m} \left(\frac{s_{t,2} y_{t-1,1} y_{t,1}}{P_{(21)1}} - \frac{s_{t,2} y_{t-1,1} y_{t,2}}{1 - P_{(21)1}} \right).$$

Igualando a cero la ecuación anterior para después resolverla se obtiene:

$$(1-P_{(21)1})\sum_{t=2}^{m} s_{t,2}y_{t-1,1}y_{t,1} - P_{(21)1}\sum_{t=2}^{m} s_{t,2}y_{t-1,1}y_{t,2} = 0$$

$$-P_{(21)1}[\sum_{t=2}^{m} s_{t,2}y_{t-1,1}y_{t,1} + \sum_{t=2}^{m} s_{t,2}y_{t-1,1}y_{t,2}] = -\sum_{t=2}^{m} s_{t,2}y_{t-1,1}y_{t,1}$$

de donde se tiene que el tercer parámetro $P_{(21)1}$ es:

$$P_{(21)1} = \frac{\sum_{t=2}^{m} s_{t,2}y_{t-1,1}y_{t,1}}{\sum_{t=2}^{m} s_{t,2}y_{t-1,1}(y_{t,1}+y_{t,2})}$$

y como $y_{t,1} + y_{t,2} = 1$ se tiene :

$$P_{(21)1} = \frac{\sum_{t=2}^{m} s_{t,2}y_{t-1,1}y_{t,1}}{\sum_{t=2}^{m} s_{t,2}y_{t-1,1}}$$

y

$$P_{(21)2} = 1 - P_{(21)1} = 1 - \frac{\sum_{t=2}^{m} s_{t,2}y_{t-1,1}y_{t,1}}{\sum_{t=2}^{m} s_{t,2}y_{t-1,1}}$$

$$= \frac{\sum_{t=2}^{m} s_{t,2}y_{t-1,1} - \sum_{t=2}^{m} s_{t,2}y_{t-1,1}y_{t,1}}{\sum_{t=2}^{m} s_{t,2}y_{t-1,1}} = \frac{\sum_{t=2}^{m} s_{t,2}y_{t-1,1}(1-y_{t,1})}{\sum_{t=2}^{m} s_{t,2}y_{t-1,1}}$$

y como $y_{t,2} = 1 - y_{t,1}$ el cuarto parámetro vendría siendo:

$$P_{(21)2} = \frac{\sum_{t=2}^{m} s_{t,2}y_{t-1,1}y_{t,2}}{\sum_{t=2}^{m} s_{t,2}y_{t-1,1}}$$

En general $P_{(21)k}$ se comporta de la siguiente forma:

$$P_{(21)k} = \frac{\sum_{t=2}^{m} s_{t,2}\, y_{t-1,1}\, y_{t,k}}{\sum_{t=2}^{m} s_{t,2}\, y_{t-1,1}} \qquad \text{.....(A.4.9)}$$

de (A.4.8) y (A.4.9) se tiene que el estimador $P_{(i1)k}$ es:

$$P_{(i1)k} = \frac{\sum_{t=2}^{m} s_{t,i}\, y_{t-1,1}\, y_{t,k}}{\sum_{t=2}^{m} s_{t,i}\, y_{t-1,1}} \qquad \text{.....(A.4.10)}$$

El quinto parámetro por encontrar vendría siendo $P_{(12)1}$ así que se deriva con respecto a él para obtener:

$$\frac{d \log \prod_{t=2}^{m} P(y_t / y_{t-1}, s_t)}{dP_{(12)1}} = \sum_{t=2}^{m} \left(\frac{s_{t,1} y_{t-1,2} y_{t,1}}{P_{(12)1}} - \frac{s_{t,1} y_{t-1,2} y_{t,2}}{1 - P_{(12)1}} \right)$$

Igualando a cero la ecuación anterior para después resolverla se tiene:

$$(1 - P_{(12)1}) \sum_{t=2}^{m} s_{t,1} y_{t-1,2} y_{t,1} - P_{(12)1} \sum_{t=2}^{m} s_{t,1} y_{t-1,2} y_{t,2} = 0$$

$$- P_{(12)1} \left[\sum_{t=2}^{m} s_{t,1} y_{t-1,2} y_{t,1} - \sum_{t=2}^{m} s_{t,1} y_{t-1,2} y_{t,2} \right] = - \sum_{t=2}^{m} s_{t,1} y_{t-1,2} y_{t,1}$$

de donde se tiene que el quinto parámetro $P_{(12)1}$ es:

$$P_{(12)1} = \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,2} y_{t,1}}{\sum_{t=2}^{m} s_{t,1} y_{t-1,2} (y_{t,1} + y_{t,2})}$$

y como $y_{t,1} + y_{t,2} = 1$ se tiene :

$$P_{(12)1} = \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,2} y_{t,1}}{\sum_{t=2}^{m} s_{t,1} y_{t-1,2}}$$

y

$$P_{(12)2} = 1 - P_{(12)1} = 1 - \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,2} y_{t,1}}{\sum_{t=2}^{m} s_{t,1} y_{t-1,2}}$$

$$= \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,2} - \sum_{t=2}^{m} s_{t,1} y_{t-1,2} y_{t,1}}{\sum_{t=2}^{m} s_{t,1} y_{t-1,2}} = \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,2} (1 - y_{t,1})}{\sum_{t=2}^{m} s_{t,1} y_{t-1,2}}$$

y como $y_{t,2} = 1 - y_{t,1}$ el sexto parámetro vendría siendo:

$$P_{(21)1} = \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,2} y_{t,2}}{\sum_{t=2}^{m} s_{t,1} y_{t-1,2}}$$

En general $P_{(12)k}$ se comporta:

$$P_{(12)k} = \frac{\sum_{t=2}^{m} s_{t,1} y_{t-1,2} y_{t,k}}{\sum_{t=2}^{m} s_{t,1} y_{t-1,2}} \qquad \text{.....(A.4.11)}$$

El séptimo parámetro por encontrar vendría siendo $P_{(22)1}$ así que se deriva con respecto a él y se obtiene:

$$\frac{d \log \prod_{t=2}^{m} P(y_t / y_{t-1}, s_t)}{dP_{(22)1}} = \sum_{t=2}^{m} \left(\frac{s_{t,2} y_{t-1,2} y_{t,1}}{P_{(22)1}} - \frac{s_{t,2} y_{t-1,2} y_{t,2}}{1 - P_{(22)1}} \right).$$

Igualando a cero la ecuación anterior para después resolverla se obtiene:

$$(1-P_{(22)1})\sum_{t=2}^{m} s_{t,2}y_{t-1,2}y_{t,1} - P_{(22)1}\sum_{t=2}^{m} s_{t,2}y_{t-1,2}y_{t,2} = 0$$

$$-P_{(22)1}\left[\sum_{t=2}^{m} s_{t,2}y_{t-1,2}y_{t,1} - \sum_{t=2}^{m} s_{t,2}y_{t-1,2}y_{t,2}\right] = -\sum_{t=2}^{m} s_{t,2}y_{t-1,2}y_{t,1},$$

de donde se tiene que el séptimo parámetro $P_{(22)1}$ es:

$$P_{(22)1} = \frac{\sum_{t=2}^{m} s_{t,2}y_{t-1,2}y_{t,1}}{\sum_{t=2}^{m} s_{t,2}y_{t-1,2}(y_{t,1}+y_{t,2})}$$

y como $y_{t,1} + y_{t,2} = 1$ se tiene :

$$P_{(22)1} = \frac{\sum_{t=2}^{m} s_{t,2}y_{t-1,2}y_{t,1}}{\sum_{t=2}^{m} s_{t,2}y_{t-1,2}}$$

y

$$P_{(22)2} = 1 - P_{(22)1} = 1 - \frac{\sum_{t=2}^{m} s_{t,2}y_{t-1,2}y_{t,1}}{\sum_{t=2}^{m} s_{t,2}y_{t-1,2}}$$

$$= \frac{\sum_{t=2}^{m} s_{t,2}y_{t-1,2} - \sum_{t=2}^{m} s_{t,2}y_{t-1,2}y_{t,1}}{\sum_{t=2}^{m} s_{t,2}y_{t-1,2}} = \frac{\sum_{t=2}^{m} s_{t,2}y_{t-1,2}(1-y_{t,1})}{\sum_{t=2}^{m} s_{t,2}y_{t-1,2}}$$

y como $y_{t,2} = 1 - y_{t,1}$ el octavo parámetro vendría siendo:

$$P_{(22)2} = \frac{\sum_{t=2}^{m} s_{t,2}y_{t-1,2}y_{t,2}}{\sum_{t=2}^{m} s_{t,2}y_{t-1,2}}$$

y en general $P_{(22)k}$ se comporta:

$$P_{(i2)k} = \frac{\sum_{t=2}^{m} s_{t,2}\, y_{t-1,2}\, y_{t,k}}{\sum_{t=2}^{m} s_{t,2}\, y_{t-1,2}} \qquad \text{.....(A.4.12)}$$

de (A.4.11) y (A.4.12) se tiene que el estimador $P_{(i2)k}$ es:

$$P_{(i2)k} = \frac{\sum_{t=2}^{m} s_{t,i}\, y_{t-1,2}\, y_{t,k}}{\sum_{t=2}^{m} s_{t,i}\, y_{t-1,2}} \qquad \text{.....(A.4.13)}$$

De tal modo que de (A.4.10) y de (A.4.13) se tiene que el estimador $P_{(ij)k}$:

$$P_{(ij)k} = \frac{\sum_{t=2}^{m} s_{t,i}\, y_{t-1,j}\, y_{t,k}}{\sum_{t=2}^{m} s_{t,i}\, y_{t-1,j}} \qquad \text{.....(A.4.14)}$$

Resumiendo se tienen ahora de (A.4.1), (A.4.4), (A.4.7) y (A.4.14) que son:

$$\Pi = (\Pi_1, \Pi_2) = (s_{1,1}, s_{1,2}) = s_1$$

$$\lambda_{ij} = \frac{\sum_{t=2}^{m} s_{t-1,i} s_{t,j}}{\sum_{t=2}^{m} s_{t-1,i}}$$

$$P_{ij} = \begin{cases} \text{a) } s_{1,1} \neq 0\ ; \\ P_{1,i} = y_{1,i} \text{ con } i = 1, 2. \\ \quad P_{2,i} \text{ cualquier valor tal que } P_{21} + P_{22} = 1 \\ \text{b) } s_{1,2} \neq 0\ ; \\ P_{2,i} = y_{1,i} \text{ con } i = 1, 2. \\ \quad P_{1,i} \text{ cualquier valor tal que } P_{11} + P_{12} = 1 \end{cases}$$

$$P_{(ij)k} = \frac{\sum_{t=2}^{m} s_{t,i}\, y_{t-1,j}\, y_{t,k}}{\sum_{t=2}^{m} s_{t,i}\, y_{t-1,j}}$$

Apéndice B

"El fuego descansa cambiando"
Heráclito

B.1. Los cromosomas y su morfología.

Se dice que una célula humana es diploide cuando consta de 46 cromosomas, la cantidad normal en un ser humano. Hay 22 pares de autosomas y los cromosomas que designan el sexo (género) son indicados por X y Y. La constitución del género masculino es XY y la del femenino es XX. En la metafase de la mitosis[27] cada cromosoma consta de dos cromátidas idénticas que se separan en fases ulteriores para transformarse cada una de ellas en uno de los 46 cromosomas de una de las dos células hijas. Las dos cromátidas se encuentran unidas entre sí por el cinetocoro, también llamado centrómero o constricción primaria (ver Fig.B.1). Un locus o loci es el sitio o localización en un cromosoma ocupado por un gen, es decir, por un específico conjunto de alelos. Un alelo es uno de dos o más formas alternativas de un gen que ocurre en el mismo locus, por ejemplo, los alelos para el tipo de sangre en el ser humano son: A, B, O y uno de ellos es el que se encontrará en un locus de un determinado cromosoma para expresarse en el genotipo del individuo.

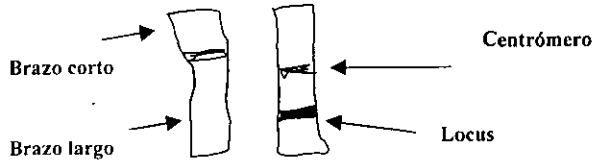


Fig. B.1. Morfología de un cromosoma.

[27] El proceso de la mitosis se explica en el *Apéndice B.2.*

B.2 Proceso de la mitosis y la meiosis.

La **mitosis** es la división celular por medio de la cual una célula con 46 cromosomas produce dos células hijas, las cuales a su vez tendrán 46 cromosomas heredados como material genético.

La mitosis es la porción más corta del ciclo celular, y se divide en tres etapas fundamentales:

(1) Antes de que la mitosis termine, la célula entra en un estado que se conoce como G_1, durante el cual el A.D.N. se encuentra en un estado no replicativo. G_1 termina cuando la síntesis del A.D.N. comienza.

(2) La síntesis de A.D.N. tiene lugar durante un período conocido como S, durante el cual la molécula de A.D.N. se reproduce en unidades llamadas replicones, cada una de las cuales realiza una copia de sí misma por replicación semiconservativa de cada tira de A.D.N. de la doble-hélice.

(3) Al terminar esta fase, cada cromosoma ha sido copiado y tiene dos cromátidas (ver Fig. B.2). En la metafase cada cromosoma está formado por dos componentes simétricos, las cromátidas, cada una de las cuales contiene una sola molécula de A.D.N. Las cromátidas sólo están unidas entre sí a nivel del centrómero y se separan al comienzo de la anafase (fase siguiente de la metafase). El período entre el final de la síntesis de A.D.N. y el comienzo de la mitosis se llama G_2.

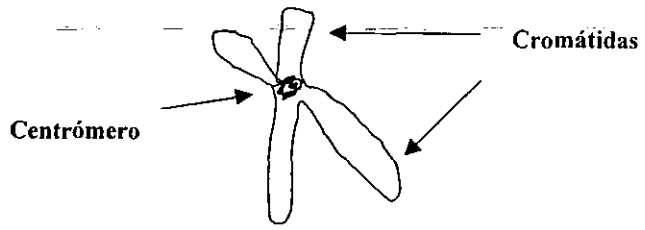


Fig. B.2 Unión de las cromátidas por medio del centrómero.

La parte del ciclo celular entre cada mitosis se conoce como interfase.

G_1	S	G_2	Mitosis	G_1	S	G_2	Mitosis	G_1

--------Interfase---------

Asimismo, la mitosis es un proceso continuo que por lo general se divide en cuatro etapas (ver Fig. B.3 (Fig 1- 6 Genethics (1992) Friedman, Dill)), las cuales son:

(1) **Profase**: Esta etapa empieza cuando termina la interfase, y los cromosomas se condensan y son visibles al microscopío. Durante la profase la membrana nuclear desaparece, en este momento cada cromosoma consta de dos líneas paralelas – las cromátidas hermanas – que son sostenidas por el centrómero.

(2) **Metafase**: Los cromosomas se contraen completamente y se mueven al centro de la célula. Se extienden fibras de los husos de los centrómeros de cada cromosoma a dos centríolos que se encuentran localizados en polos opuestos de la célula.

(3) **Anafase**: El centrómero de cada cromosoma se divide y las cromátidas hijas se separan, convirtiéndose en dos cromosomas hijas que comienzan a moverse a los polos opuestos de la célula.

(4) **Telofase**: Cada una de las cromosomas hijas llegan a sus polos correspondientes y comienzan a descondensarse (fase en la cual no son visibles claramente en el microscopio). El citoplasma se divide y la membrana nuclear se vuelve a formar para cada una de las nuevas células hijas. Y la interfase comienza para cada una de las células hijas.

Proceso de la meiosis.

La **meiosis** es la división celular en la cual un gameto diploide precursor produce gametos haploides[28] (ver Fig. B.4 (Fig 1- 7 Genethics (1992) Friedman, Dill)). Cada uno de los gametos que se producen tiene 23 cromosomas en lugar de 46. La meiosis es precedida por la síntesis de A.D.N. y consiste de dos divisiones celulares.

Primera división meiótica

La **meiosis I** se conoce como la división reductiva, ya que reduce el número de cromosomas de 46 a 23.

(1) Al principio de la **meiosis I profase (profase 1)**, cada cromosoma ha completado su replicación y ahora consiste de dos cromátidas hermanas unidas por un centrómero. Esta fase se divide en diversos estadíos secuenciales:

(a) *Leptonene*: los cromosomas se vuelven visibles a la luz del microscopio como delgadas tiras.

(b) *Zygotene*: los cromosomas homólogos[29], uno de los cuales lo da la madre y el otro el padre, se aparean a todo lo largo de las tiras.

(c) *Paquitene*: los cromosomas se condensan aún más y se lleva a cabo la recombinación entre cromátidas individuales de cromosomas homólogos a través del crossing-over[30].

(d) *Diplotene*: el par de cromosomas homólogos empiezan a separarse pero están aún unidos por los lugares de crossing-over.

---

[28] Se dice que una célula es haploide cuando únicamente consta de la mitad (23) de los cromosomas normales (46). Esto sucede en las células gametos y en algunos estadios del ciclo de vida de algunas plantas.

[29] Cromosomas homólogos son pares de cromosomas que son prácticamente idénticos en forma, tamaño y función y que se aparean durante la división celular meiótica.

[30] Crossing-over es el intercambio de segmentos cromosomales entre cromátidas durante la división celular.

(e) *Diakinesis*: los cromosomas alcanzan su máxima condensación.

(2) En la **metafase 1** la membrana nuclear desaparece y las líneas bivalentes de cromosomas se centran en la célula. Un huso conecta los centrómeros con los centríolos de la célula en polos opuestos.

(3) En la **anafase 1** los cromosomas homólogos se separan y se mueven a polos opuestos de la célula. Las cromátidas hermanas de cada cromosomas permanecen unidas por el centrómero, las cuales no se dividen en la meiosis 1.

(4) Durante la **telofase 1** los dos conjuntos de células cromosomales haploides llegan a polos opuestos de la célula y se divide el citoplasma.

Segunda división meiotica.

Esta segunda etapa es precedida por un pequeño estadío de interfase en donde no se lleva a cabo síntesis alguna de A.D.N.

La **meiosis II** es similar a la mitosis, en donde los cromosomas, que consisten de dos cromátidas hijas unidas por un centrómero, se alinean en el centro de la célula, se forma un huso que las conecta a los polos opuestos de la célula para posteriormente separarse e irse cada cromosoma hija a polos opuestos de la célula. La diferencia radica en que en la meiosis II existen sólo 23 cromosomas en la célula madre original que dará lugar a dos células hijas con igual cantidad de cromosomas, mientras que en la mitosis eran 46.

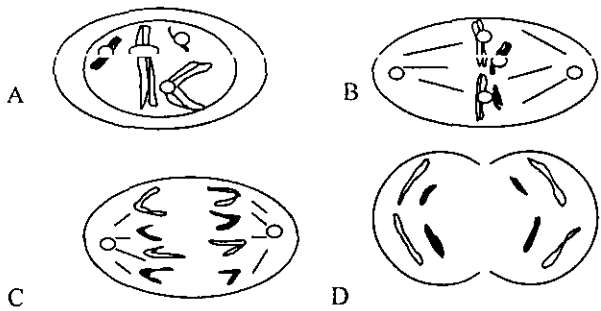


Fig.B.3 Proceso de la mitosis. (A) Profase, (B)Metafase, (C) Anafase y (D)Telofase.

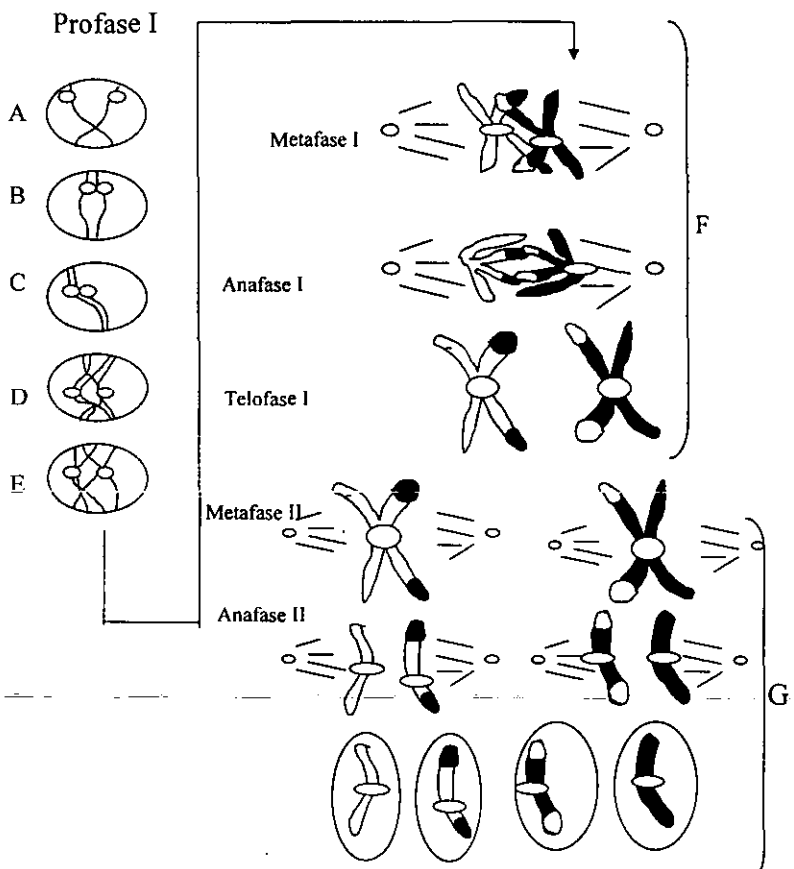


Fig. B.4 (A)Leptonene de la Profase I, (B)Zygotene de la Profase I, (C) Pakytene de la Profase I, (D) Diplotene de la Profase I, (E) Diakinesis, (F)Continuación de la primera división meiótica y (G) Segunda división meiótica.

B.3 Abreviaturas de aminoácidos.
(De Tabla 1.1)

Ala	Alanina
Arg	Arginina
Asn	Asparagina
Asp	Ácido aspártico
Cys	Cisteina
Gln	Glutamina
Glu	Ácido glutamínico
Gly	Glicina
His	Histidina
Ile	Isoleucina
Leu	Leucina
Lys	Licina
Met	Metionina
Phe	Fenilalanina
Pro	Prolina
Ser	Serina
Thr	Treonina
Trp	Triptofano
Tyr	Tirosina
Val	Valina

B.4 Anomalías cromosómicas.

Anomalías numéricas. Este tipo de anomalías se llega a dar cuando se modifica el número de cromosomas del ser humano disminuyéndose o aumentándose. Los casos más comunes de anomalías son:

- Poliploides: el aumento de series haploides completas de cromosomas. La mayor parte de los casos terminan en abortos espontáneos o en muerte durante la infancia.

- Trisomía: es la presencia de tres copias de un cromosoma en lugar de las dos copias normales. El ejemplo médico más común es el Síndrome de Down (Trisomía 21).

- Monosomía: es la presencia de un solo miembro del par de cromosomas en el kariotipo[31]. Un ejemplo médico es el Síndrome de Turner.

- Mosaiquismo: es la presencia de dos o más líneas de células con diferentes kariotipos en un paciente. Las características que se presentan a nivel clínico dependen fundamentalmente en el cromosoma específico que se vea afectado.

Anomalías estructurales. Son un nuevo arreglo del material genético entre cromosomas o de él mismo. Que pueden ser un balance genético, cuando no hay cambio en la cantidad de material genético, o bien inbalanceado cuando hay un aumento o pérdida de segmentos esenciales de cromosomas.

B.5 Enfermedades producidas por mutaciones.

α-Thalasemia. Es causada por la deleción del conjunto de genes de la α-globina, de hecho el problema se origina principalmente en la etapa del cross-over de la meiosis, debido a que las secuencias de A.D.N. se encuentran demasiado próximos el uno con el otro. Los fetos que carecen de la α-globina usualmente mueren por un edema generalizado ya sea durante la etapa de gestación o muy poco después de nacer.

Deficiencia en la hormona de crecimiento. Así como en la α-Thalasemia, el problema se origina por una deleción en el conjunto de genes que controlan el crecimiento por estar demasiado cerca una secuencia de otra, lo cual causa enanismo en personas homocigotas[32].

Distrofia muscular de Duchenne. El gen anómalo se encuentra en el brazo corto del cromosoma X. Su incidencia aproximada es de

[31] El kariotipo es una micrografía tomada a los cromosomas que se encuentran en células preparadas y después cortarlas para emparejar cromosomas homólogos, vendrían siendo como un arreglo de todos los cromosomas en la célula de un individuo, los cuales se observan en microscopio durante la metafase.

[32] Homocigota es la persona que posee los mismos alelos en un determinado locus.

1/4000 varones nacidos vivos. Las mujeres heterocigotos[33] para este gen son asintomáticas, aunque algunas presentan pequeñas anomalías como pseudohipertrofia de pantorrillas o ligera debilidad. La coincidencia de este gen en ambos cromosomas X es letal en el periodo intrauterino. Las distrofias musculares en general son un grupo heterogéneo de enfermedades genéticamente determinadas en las que se produce una degeneración progresiva del músculo esquelético. El sustrato patógeno de la Distrofia de Duchenne es la ausencia congénita de una proteína estructural del músculo denominada distrofina, lo que determina una permeabilidad anormal de las membranas y la necrosis o muerte progresiva de las fibras musculares. El cuadro completo incluye una miopatía esquelética, una miocardiopatía y un coeficiente intelectual bajo. La debilidad y la atrofia muscular son progresivas y simétricas, se afectan todos los músculos esqueléticos incluidos los de la cara, aunque predominan sobretodo en el inicio, la musculatura pélvica y del tronco. La muerte se produce alrededor de los 20 años a causa de la insuficiencia respiratoria o, en el 10% de los casos, de la miocardiopatía.

Fibrosis quística. Dilatación anormal y permanente de los bronquios debido a la destrucción de los componentes musculares y elásticos de la pared bronquial. Es la anomalía genética más frecuente en la raza blanca. Su incidencia es de 1/2500 nacidos vivos. Actualmente se puede realizar, mediante marcación genética, el diagnóstico prenatal.

Hemofilia. Heredopatía recesiva X-relacionada, es decir, transmitida por el cromosoma X, afecta exclusivamente a los varones y se caracteriza por alteraciones en la coagulación de la sangre.

Hipercolesterolemia Familiar (HF). Esta enfermedad resulta en el defecto del gen receptor de la baja densidad lipoprotéica, lo cual causa una anormalidad en el control de grasas en el cuerpo. La deleción en este caso es de sólo una parte del gen receptor.

[33] Heterocigoto es la posesión de diferentes alelos (uno de dos o más formas alternativas de un gen que se manifiesta en un locus) en un determinado locus.

Síndrome de Down. La causa es una posible translocación[34] o mosaicismo que da lugar a tres cromosomas 21, por lo cual también se le conoce a ésta enfermedad como Trisomía 21. Se caracteriza por manos y dedos cortos, pies cortos y anchos, talla baja, retraso mental, perfil facial plano, orejas de implantación baja y anormales, boca entreabierta y lengua prominente. La tasa de ocurrencia para bebés nacidos vivos con esta trisomía está muy relacionada con la edad de la madre (ver Tabla B.1 (Tabla 2-2. Genethics (1992) Friedman, Dill)).

Tabla B.1 Riesgo de Síndrome de Down.

Edad de la madre	Nacidos vivos.
24	1/1250
25	1/1210
26	1/1250
27	1/1210
28	1/1180
29	1/1140
30	1/1000
31	1/830
32	1/630
33	1/490
34	1/350
35	1/282
36	1/220
37	1/170
38	1/130
39	1/100
40	1/80
41	1/60
42	1/48
43	1/38
44	1/30

[34] Translocación es un tipo de anomalía estructural que se caracteriza por la fusión de brazos enteros de cromosomas acrocéntricos.

Síndrome de Turner. El individuo es de aspecto femenino, pero prácticamente no hay gónadas[35], las cuales constan sólo de dos bandas de tejido conjuntivo. Tales individuos tienen 44 autosomas, un solo cromosoma X y ningún cromosoma Y, lo que resulta en un total de 45 cromosomas, uno menos de lo normal.

[35] Las gónadas son los órganos de la reproducción: testículos en los machos y ovarios en la hembra. En ambos sexos, las gónadas tienen función doble: (1) producción de las células de la reproducción: espermatozoides y oocitos y (2) secreción de las hormonas sexuales.

Comentarios

Si bien el objetivo principal de esta tesis es el mostrar el uso del algoritmo Esperanza Maximización en una aplicación a la genética, también es importante notar algunos puntos que se han ido mencionando a lo largo de este trabajo y que es de particular importancia desglosarlos con mayor claridad. A continuación se comentarán algunas ventajas y desventajas del algoritmo Esperanza Maximización, método utilizado a lo largo de este trabajo, así como resultados relevantes del mismo.

Ventajas.

- Una de las principales ventajas del algoritmo consiste en su estabilidad numérica, es decir, que a cada paso o iteración realizada, se puede asegurar que la verosimilitud aumenta o se mantiene, es decir $l(\theta^{(n-1)}) \geq l(\theta^{(n)})$, (ver sección 3.4).

- Recuerde que se empezó a trabajar con datos observados que se consideraban incompletos, de tal modo que aunque no se tenían todos los valores observados se pudo converger a un estimador máximo verosímil θ^*, utilizando para esto un plano de cadena de Markov Oculta. Hay que hacer énfasis, si en el espacio paramétrico existe mas de un candidato a ser estimador máximo verosímil, en el hecho de que la convergencia a uno u otro estimador dependerá del valor inicial $\theta^{(0)}$ que se tome al iniciar la iteración, aunque bajo determinadas condiciones se puede asegurar la convergencia local (ver sección 3.6 **Teorema 3.4**). En caso contrario, si el espacio paramétrico es unimodal, se convergerá a este único estimador (ver sección3.6 **Teorema 3.5**).

- Ayudados por el estimador máximo verosímil θ^* y utilizando el algoritmo de actualización se infiere el vector de estados **S** que se buscaba, es decir, la secuencia de A.D.N. verdadera, a partir del vector de observaciones **Y**. Tomando en cuenta que esta es un aplicación al área biológica, hay que notar que el modelo utilizado (ver **Teorema 4.1**) puede ser modificado y extendido a otras áreas, ya que la diferencia estribará en la definición acertada de los vectores observados y por encontrar. Un ejemplo que se puede mencionar es el caso en que se utilicen en el lugar de las bases A, C, G y T a los tripletos de los aminoácidos, base fundamental de las proteínas, para poder inferir secuencias más largas de genes.

- Como la base del modelo es un plano de cadena de Markov Oculta las aplicaciones a esta área se ven extendidas a todo aquello en donde se vea involucrada la hipótesis de inferencia de información perdida o con observaciones con ruido.

- El algoritmo es fácil de programar en una computadora. Programas existentes en la literatura pueden ser obtenidos vía internet, por ejemplo:

a) http://linkage.rockefeller.edu/ott/eh.htm (un programa para la estimación de la frecuencia del haplotipo)
b) http://www-stat.stanford.edu/~susan/course/b494/index/node101.html trata el ejemplo de estimar las frecuencias de genes.
c) http://sato-www.cs.titech.ac.jp/prism : programa que trabaja en general para cualquier tipo de aplicación.
d) http://www.hds.utc.fr/~mdang/Progs/prognem.html: programa para clasificación espacial.
e) http://www.ibc.wustl.edu/~josec/DISTRIB/FORTRAN_EM/

- El trabajo analítico es más sencillo que otros métodos, ya que sólo se calcula la esperanza condicional de la log verosimilitud para el problema de datos completos que es el que necesita ser maximizado. Y aun si el trabajo analítico puede ser amplio no es complicado en la mayor parte de las aplicaciones.

- El algoritmo provee valores estimados de datos perdidos a partir de la sustitución de estos valores por su esperanza condicionada a los datos observados (como se notó al encontrar la secuencia verdadera de A.D.N. a partir de la secuencia observada en el laboratorio).

DESVENTAJAS

- El Algoritmo E.M. puede converger lentamente si hay demasiada información perdida o en el caso de información completa si el problema a resolver resulta ser muy sencillo puede resultar más rápido utilizar un método que se adecúe de una mejor forma.

- No se garantiza la convergencia global cuando hay muchos máximos locales en el espacio paramétrico. Este es un problema de todos los métodos numéricos de aproximación.

- En casos donde el trabajo analítico llega a ser demasiado difícil de trabajar, existe la posibilidad de utilizar el algoritmo E.M. en aproximación de Monte Carlo vía Cadenas de Markov.

- Dentro de las desventajas que se notaron al algoritmo E.M. en su presentación en 1977 se encontraba el hecho de que como método numérico, contrario a los métodos de Newton, no se produce un estimado de la matriz de covarianza del estimador máximo verosímil. No obstante, Meilijson en 1989 propuso un método numérico basado en los cómputos que se realizan a lo largo de los pasos E y M para encontrar esta matriz de covarianza, así como un método para acelerar la convergencia al estimador máximo verosímil.

OTROS COMENTARIOS.

Dentro de los resultados que se pueden mencionar de la aplicación al modelo utlizado en la tesis se pueden encontrar en el área de simulación para el A.D.N. mitocondrial (se trabaja con sucesiones circulares), trabajo con un fragmento de cromosoma X humano y la secuencia del bacteriófago λ (la cual ya ha sido

E.M. en la estimación a posteriori como parte de un problema bayesiano, así como el de Monte Carlo vía cadenas de Markov.

Bibliografía.

[1] Alphey Luke. (1997) *DNA sequencing from experimental methods to bioinformatics*. Springer Verlag. New York, Inc. New York.

[2] Berger James O. (1985) *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer Verlag Inc. E.U.A.

[3] Brown, T.A. (1992) *Genetics, a molecular approach*. Chapman and Hall. 2nd ed.

[4] Chen Kang Chai. (1976) *Genetic Evolution*. The University of Chicago Press. U.S.A.

[5] Churchill Gary A. (1995) (Enero 90-158) *Accurate Restoration of DNA Sequence."Case studies in Bayesian Statistics"*. Gatsaris,C., Hodges, J.S. Kars, R.E.and Singpur-Walla, N.D. Springer-Verlag. New York.

[6] Churchill Gary A. (1992) *Hidden Markov Chains and the Analysis of Genome Structure*. Vol.16 No.2 pp 107-115.

[7] Churchill Gary A. (1989) *Stochastic models for heterogeneous DNA Sequences Bulletin of Mathematical Biology.*. Vol.51. No.1 pp.79-94.

[8] Coleman Rodney (1986) *Procesos estocásticos*. Editorial Limusa. Vol.14, México.

[9] DeGroot Morris H. (1998) *Probabilidad y Estadística*. 2a.ed. Adisson-Wesley Iberoamericana.E.U.A. Trad. José M. Bernardo.

[10] Dempster A.P., Laird N.M. and Rubin D.B. (1977) *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society Ser. B. Vol. 39; 1-38.

[11] Dobzhansky Theodosius. (1982) *Genetics and the origins of species.* The Columbia Classics in Evolution Series. Niles Eldredge & Stephen Jay Gould. Columbia University Press.U.S.A.

[12] Florens Jean Pierre, Mouchart Michel. (1990) *Elements of Bayesian Statistics.* Marcel Dekker Inc. E.U.A.

[13] Freeman Scott, Herron Jon C. (1998) *Evolutionary Analysis.* Prentice Hall. U.S.A.

[14] Friedman J.M., Dill F.J., Hayden M.R, McGillivray B.C. (1992) *Genetics.* The National Medical Series for Independent Studies Williams and Wilkins E.U.A.

[15] Gamerman Dani. (1997) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Chapman and Hall. London.

[16] Garner Lynn E. (1988) *Calculus and Analytic Geometry.* Dellen Publishing Company, U.S.A.

[17] Gelman Andrew, Carlen John B., Stern Hal S., Rubin Donald B. (1994) *Bayesian Data Analysis.* (Libro en preparación).

[18] Grobstein Cliford. (1973) *La estrategia de la vida.* Editorial Blume. España.

[19] Larson J. Harold. (1982) *Introduction to Probability Theory and Statistical Inference.* John Wiley & Sons, Inc.

[20] Lee Clark Randolph, W. Cumley Rusell. (1981) *El libro de la salud.* Compañía Editorial Continental. 2ª impresión.

[21] Lindley D. V. (1972) *Bayesian Statistics, a review.* Regional Conference series in applied mathematics. Society for Industrial and Applied Mathematics. England.

[22] Lindsey J. K. (1996) *Parametric Statistical Inference.* Clarendon Press, Oxford.

[23] Luthe, Olivera, Schutz. (1980) *Métodos numéricos.* Editorial Limusa, México.

[24] Maynard Smith John. (1993) *The Theory of Evolution.* Cambridge University Press, U.S.A.

[25] Maxam A.M. and Gilbert W. (1997) *A new method for sequencing DNA.* Proc. Natl. Acad. Sci. U.S.A. 74. 560-564

[26] Medhi J. (1994) *Stochastic Processes.* 2nd.Edition. John Wiley & Sons. India.

[27] Mckusick Victor A. (1967) *Genética Humana.* Editorial Rabasa.

[28] McLachlan Geoffrey J., Krishnan Thriyambakam. (1997) *The E.M. Algorithm and Extensions.* Wiley Series in Probability and Statistics. U.S.A.

[29] Monod Jaques. (1971) *El azar y la necesidad.* Ensayo sobre la filosofía natural de la biología moderna. Monte Ávila Editores, C.A. Caracas.

[30] Moya Andrés. (1989) *Sobre la estructura de la teoría de la evolución.* Edit. del Hombre. España.

[31] Oliva Virgili Rafael. (1996) *Genoma Humano.* Masson, S.A. España.

[32] Ostrowski,A.M. (1966) *Solution of equations and systems of equations.* 2nd ed. Academic. New York.

[33] Pellón José R. (1986) *La ingeniería genética y sus aplicaciones.* Edit. Acribia, S.A. España

[34] Rodés Teixidos Joan, Guardia Massó Jaime. (1992) *El manual de medicina.* Ediciones Científicas y Técnicas, S.A. España.

[35] Ross Sheldon (1996) *Simulation.* Academic Press. U.S.A.

[36] Santaló Luis A. (1975) *Probabilidad e Inferencia Estadística.* Argentina.

[37] Schleif Robert. (1983) *Genetics and Molecular Biology.* The Johns Hopkins University Press. E.U.A.

[38] Schrödinger Erwin (1947) *¿Qué es la vida?.* Compañía Editorial Espasa – Calpe. Argentina.

[39] Suzuki D. and Knutdson P. (1989) *The Clash between the new genetics and human values.* Harvard University Press. E.U.A.

[40] Tanner Martin A. (1996) *Tools for Statistical Inference. Methods for the Exploration of Posterior Distributions and Likelihood Functions.* 3rd ed. Springer Series in Statistics. E.U.A.

[41] Thorne Jeffrey L. and Churchill Gary A. (1995) Marzo. *Estimation and Reliability of Molecular Sequence Alignments.* Biometrics 51. 100-113.

[42] Walrand Jean, Varaiya Pravin. (1996) *High Performance Comunication Networks*. Morgan Kaufman Publishers, Inc.

[43] Watson J.D. and Crick F.H.C. (1953) *Molecular structure of nucleic acids: a structure for deoxypentose nucleic acid.* Nature 171, 737.

[44] Weir B.S. (1993) *Analysis of DNA sequences.* Statistical Methods in Medical Research 2; 225-239.

[45] Weir B.S. (1988) *Statistical Analysis of DNA Sequences.* Vol.80 No.6 Mayo 18.

[46] Wu C.F. Jeff. (1983) *On the convergence properties of the E.M. Algorithm.* The Annals of Statistics. Vol 11, No.1, 95-103.

[47] Zangwill (1969) *Nonlinear programming: A unified aproach.* Prentice Hall, Englewood Cliffs. New York.