El algoritmo EM

Referencias:
Bickel and Docksum: Mathematical Statistics
Casella, G Statistical inference
Lehman, E. Point Estimation

1. Estimación por Máxima Verosimilitud

Uno de los principales objetivos en estadistica es estimar parametros desconocidos de un población usando muestras tomadas de ellas. Más formalmente, sea X una variable aleatoria con función de densidad (o función de probabilidad) f(x;θ) (o en notación bayesiana f(x/θ), donde θ representa el parámetro que se desea estimar usando una muestra aleatoria (X_1,X_2,….X_n). En una muestra aleatoria cada X_i tiene la misma distribiución de X y además son independientes entres si.

Existen muchos métodos de encontrar un estimador $\hat{\theta} = T(X_1,.....X_n)$. Por ejemplo: mínimos cuadrados, momentos, métodos bayesianos, etc. Nosotros solo revisaremos aquí el método de máxima verosimilitud

Definición: Sea **X**= (X_1,X_2,….X_n) una muestra aleatoria de la variable aleatoria X entonces la función de verosimiltud se define por

$$L(\theta) = L(\theta / \mathbf{x}) = f(x_1./\theta).....f(x_n/\theta) = \prod_{i=1}^{n} f(x_i/\theta)$$

Definición: El estimador máximo verosimil (MV) $\hat{\theta}$ del parámetro θ es aquel valor que maximiza la función de verosimilitud. Más formalmente,

$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

Frecuentemente es más fácil maximizar log-lik(θ)=LogLθ) que Log(θ), y debido a que Log es una función monótona el valor de θ donde ambas funciones alcanzan su máximo es el mismo.

Nota: En estadística, por tradición, *log* representa *ln* de cálculo. O sea (log x)'=1/x

Ejemplo1: Hallar el estimador máximo verosimil del parametro p de la distribución Bernoulli

Ejemplo 2: Hallar el estimador MV del parametro λ de una distribución exponencial

También pueden haber varios parámetros desconocidos a estimar. O sea $\hat{\boldsymbol{\theta}} = (\theta_1,...,\theta_k)$

Ejemplo 3: Hallar el estimador MV del vector de parámetros $\theta=(\mu,\sigma^2)$ de una distribución normal.

Por supuesto, los casos más complicado son cuando la distribución es multivariada con varios parámetros a estimar.

2. Familia Exponenciales

En estadistica, muchas veces el proceso de estimación se agrupa según la forma funcional de la densidad o de la función de probabilidad. Los grupos se llaman familias. Existen familias de localización, familias de escala y familias exponenciales entre otras. Nosotros solo revisaremos la estimación de parámetros en familias exponenciales.

Caso uniparamétrico

Sea $f(x/\theta)$, que representa la función de densidad o de probabilidad de una variable aleatoria X, se dice que pertenece a la familia exponencial si existen funciones $a(\theta)$, $T(x)$, $b(\theta)$ y $s(x)$ tales que

$$f(x/\theta) = e^{a(\theta)T(x)+b(\theta)+s(x)} I_A(x)$$

donde I_A representa la función indicadora en el conjunto A, el cual además se supone que no depende del parámetro θ.

Ejemplo. Mostrar que la función de probabilidad Bernoulli pertenece a la familia exponencial

Solución: Una variable aleatoria X tiene una función de probabilidad Bernoulli con parámetro θ si está dada por

$$p(x/\theta) = \theta^x (1-\theta)^{1-x}$$

para x=0, 1.

La función de probabilidad puede ser escrita como

$$p(x/\theta) = \exp[x\ln\theta + (1-x)\ln(1-\theta)]$$

De donde,

$$p(x/\theta) = \exp[x\ln(\theta/(1-\theta)) + \ln(1-\theta)]$$

Luego $a(\theta)=\ln(\theta/(1-\theta))$ y $T(x)=x$. También $b(\theta)=\ln(1-\theta)$

Ejemplo. Mostrar que la función de densidad exponencial pertenece a la familia Exponencial

Algunas veces es más conveniente reparametrizar la familia exponencial usando el parámetro $\eta=a(\theta)$. Así que

$$f(x/\eta) = e^{\eta T(x)+b_0(\eta)+s(x)} I_A(x)$$

donde $b_o(\eta)=b(a^{-1}(\eta))$. Esta es llamada la representación canónica o natural de una familia exponencial.
En el caso de la función de probabilidad de Bernoulli
$\eta=\ln(\theta/(1-\theta))$. Luego, $e^{\eta}=\theta/(1-\theta)$ y $e^{-\eta}+1=1/\theta$. Por lo tanto, $\theta=1/(e^{-\eta}+1)$. De ello se obtiene que $b_o(\eta)=\ln(e^{-\eta}/(e^{-\eta}+1))=-\ln(1+e^{\eta})$

Derivando la función generatriz de momentos de X usando la forma canónica de su densidad se obtiene las siguientes propiedades para variables aleatorias cuyas densidades pertenecen a la familia exponencial

a) Valor Esperado de T(x): $E(T(x)) = -b'_0(\eta)$

En el caso de la distribución de Bernoulli, $b'_0(\eta) = -\dfrac{e^{\eta}}{1+e^{\eta}} = -\theta$

b) Varianza de T(x): $VAR(T(x)) = -b''_0(\eta)$

c) Máxima Verosimiltud de θ. El estimador máximo verosimil de θ, basado en una muestra de tamaño n=1, se obtendrá al resolver la ecuación $E(T(X)/\theta)=T(x)$ con respecto a θ. Aquí X representa la variable aleatoria y x un valor observado de ella. Si se tuviera una muestra aleatoria $(X_1,X_2,\ldots.X_n)$ de la variable aleatoria X entonces la función de densidad conjunta también es de la familia exponencial y el estimador MV de θ será la solución de

$$E(\sum_{i=1}^{n} T(X_i)/\theta) = \sum_{i=1}^{n} T(x_i)$$

Caso Multiparamétrico
Sea $f(x/\boldsymbol{\theta})$, con $\boldsymbol{\theta}=(\theta_1,\ldots.\theta_k)$ que representa la función de densidad o de probabilidad de una variable aleatoria X, se dice que pertenece a la familia exponencial si existen funciones $\mathbf{a(\theta)=(a_1(\theta),\ldots\ldots a_k(\theta))}$, $\mathbf{T(x)=(T_1(x),\ldots.T_k(x))}$, $\mathbf{b(\theta)}$ y $\mathbf{s(x)}$ tales que

$$f(\mathbf{x}/\boldsymbol{\theta}) = e^{\mathbf{a(\theta)T(x)} + \mathbf{b(\theta)} + \mathbf{s(x)}} I_A(\mathbf{x})$$

donde I_A representa la función indicadora en el conjunto A, el cual además se supone que no depende del parámetro $\boldsymbol{\theta}$.

Ejemplo1. La distribución Normal con parametros μ y σ pertenece a la familia exponencial.

Ejemplo 2: Verificar que la distribución multinomial pertenece a la familia exponencial y verificar sus propiedades.

Ejemplo: Máxima verosimitud para un modelo multinomial usado en genética
Supongamos que n individuos se asignan al azar y con reemplazo a 4 categorias y sea

y_i el número de individuos en el i-ésima categoría. Supongamos además que las probabilidades de que una observación pertenezca a cada una de las categorias están dadas respectivamente por

$p_1=1/2+ \theta/4$, $p_2=1/4-\theta/4$, $p_3=1/4-\theta/4$ y $p_4=\theta/4$.

Donde θ es un parámetro desconocido que se desea estimar.
Entonces el vector aleatorio **y**=(y1,y2,y3,y4) se distribuye multinomialmente con función de probabilidad

$$f(\mathbf{y}/\theta)=\frac{(y_1+y_2+y_3+y_4)!}{y_1!y_2!y_3!y_4!}(\frac{1}{2}+\frac{\theta}{4})^{y_1}(\frac{1}{4}-\frac{\theta}{4})^{y_2}(\frac{1}{4}-\frac{\theta}{4})^{y_3}(\frac{\theta}{4})^{y_4}$$

Luego, el logaritmo de la función de verosimiltud será

$$L(\theta/\mathbf{y})=Logf(\mathbf{y}/\theta)=y_1\log(2+\theta)+(y_2+y_3)\log(1-\theta)+y_4\log\theta+o.t$$

donde *o.t* incluye terminos constantes en θ y por lo tanto no interesan. Derivando con respecto a θ e igualando a cero se obtiene

$$\frac{y_1}{2+\theta}-\frac{y_2+y_3}{1-\theta}+\frac{y_4}{\theta}=0\,.$$

Lo cual es equivalente a la ecuación cuadrática

$$n\theta^2+(2y_2+2y_3-y_1+y_4)\theta-2y_4=0$$

donde n=y1+y2+y3+y4.

Esto producirá el número crítico,

$$\hat{\theta}=\frac{-(2y_2+2y_3-y_1+y_4)+\sqrt{(2y_2+2y_3-y_1+y_4)^2+8ny_4}}{2n}$$

que despues de aplicar el criterio de la primera derivada resulta ser el estimador máximo verosimil de θ.

En el caso que se observó $y_1=125$, $y_2=18$, $y_3=20$ y $y_4=34$, resulta que

$$\hat{\theta}=\frac{15+231.968}{394}=0.626822$$

Más adelante hallaremos este mismo valor usando el algoritmo EM.

Descripción del algoritmo EM.

Supongamos que **y** representa el conjunto de datos observados (también es llamado el conjunto de datos incompletos). Suponga que **x** representa el conjunto de datos completos, el cual no es observado directamente sino a traves de **y**. Es decir, existe una función h tal que **y**=h(**x**). Por ejemplo, en un experimento se extraen con reemplazamiento bolas de una urna que contiene bolas rojas, verdes y azules. Uno está interesado en la distribución de los tres tipos de bolas en la muestra. O sea
x_1= numero de bolas rojas
x_2=número de bolas verdes
x_3=número de bolas azules

Pero por alguna razón el individuo que hace las extracciones no puede distinguir entre las bolas rojas y verdes y solo registra
y_1= numero de bolas no azules
y_2=número de bolas azules

En este caso $y_1=x_1+x_2$ y $y_2=x_3$, o sea $h(x_1,x_2,x_3)=(x_1+x_2,x_3)$
Sea el conjunto $X(Y)=\{\mathbf{x}:\mathbf{y}=h(\mathbf{x})\}$. Sea $g(\mathbf{y}/\theta)$ que representa la función de densidad de probabilidad de **y**, la cual posee un parámetro desconocido θ, el cual se desea estimar. Sea $f(\mathbf{x}/\theta)$ que representa la función de probabilidad o de de densidad del conjunto de datos completo. Las densidades g y f se relacionan por

$$g(\mathbf{y}/\theta)=\int_{X(Y)} f(\mathbf{x}/\theta)d\mathbf{x}$$

En el algoritmo EM se asume que es más fácil trabajar con f que con g. La función h no es única, debe ser determinada de tal manera que sea facil trabajar con f(x/θ).
El algoritmo EM es un proceso iterativo para estimar θ. Primero se debe escoger un valor inicial θ_0 para el parámetro tal que $\theta=\theta_0$ y considerar la función

$$Q(\theta/\theta^{(p)})=E[\log f(x/\theta)/y,\theta^{(p)}]=\int \log f(x/\theta)k(x/y,\theta^{(p)})dx$$

Donde $\theta^{(p)}$ representa el estado del parámetro θ en el paso p, y K(x/y, $\theta^{(p)}$)$=\dfrac{f(x/\theta^{(p)})}{g(y/\theta^{(p)})}$. Aquí se esta estimado los datos no observados usando los valores observados y el estado actual del parámetro.
Sea $\theta^{(p)}$, el estimado del parámetro en la p-ésima iteración. Entonces el estimado en la iteración p+1 se encuentra ejecutando los siguientes dos pasos

Paso E (E es por valor esperado)

Se calcula $Q(\theta/\theta^{(p)})$

Paso M (M es por maximización)

Se calcula $\theta^{(p+1)}=\arg\max_{\theta} Q(\theta/\theta^{(p)})$

El proceso iterativo termina cuando $|\theta^{(p+1)}-\theta^{(p)}|<TOL$, donde $TOL=10^{-6}$ o 10^{-7} y el estimado será $\theta^*=\theta^{(p+1)}$

Ejemplo: Aplicar el algoritmo EM para estimar el parámetro θ del ejemplo anterior

Solución:

Primero debemos considerar el modelo completo de la siguiente manera: la primera categoria que tiene probabilidad de ocurrencia ½+θ/4, puede ser considerada como que está formada por dos subcategorias: una de ellas con probabilidad de ocurrencia ½ y la otra con probabilidad de ocurrencia θ/4. Más formalmente, el modelo completo consistiría de x_1,x_2,x_3,x_4, y x_5 donde $y_1=x_1+x_2$, $y_2=x_3$, $y_3=x_4$ y $y_4=x_5$. La función de probabilidad del modelo completo estaría dado por:

$$f(\mathbf{x}/\theta)=\frac{(x_1+x_2+x_3+x_4+x_5)!}{x_1!x_2!x_3!x_4!x_5!}(\frac{1}{2})^{x1}(\frac{\theta}{4})^{x2}(\frac{1}{4}-\frac{\theta}{4})^{x3}(\frac{1}{4}-\frac{\theta}{4})^{x4}(\frac{\theta}{4})^{x5}$$

Notar que,

$$f(\mathbf{y}/\theta)=\sum_{\{\mathbf{x}:\mathbf{y}=h(\mathbf{x})\}} f(\mathbf{x}/\theta)$$

donde $h(x_1,x_2,x_3,x_4,x_5)=(x_1+x_2,x_3,x_4,x_5)=(y_1,y_2,y_3,y_4)$.

Luego,

$$\log f(\mathbf{x}/\theta)=x_2\log\theta+(x_3+x_4)\log(1-\theta)+x_5\log(\theta)+o.t$$

nuevamente o.t incluye terminos que son constantes en θ.
En consecuencia,

$$Q(\theta/\psi)=E(\log f(\mathbf{x}/\theta)/\mathbf{y},\psi)=E[(x_2\log\theta+(x_3+x_4)\log(1-\theta)+x_5\log\theta)/(y_1,y_2,y_3,y_4),\psi]$$

Lo cual se reduce a:

$$Q(\theta/\psi)=E(\log f(\mathbf{x}/\theta)/\mathbf{y},\psi)=\log(\theta)E[(x_2/y_1,\psi)]+(y_2+y_3)\log(1-\theta)+y_4\log\theta$$

Pero, resulta que por construcción.

$x_2/y_1=x_1+x_2 \sim \text{Binomial}(y1, \psi/(2+\psi))$

En consecuencia,

$$E(x_2 / y_1, \psi) = y_1 \frac{\psi}{2+\psi}$$

Paso E. De los resultados anteriores se llega a

$$Q(\theta / \theta^{(p)}) = E(\log f(\mathbf{x}/\theta) / \mathbf{y}, \theta^{(p)}) = y_1 \log(\theta) \frac{\theta^{(p)}}{2+\theta^{(p)}} + (y_2 + y_3) \log(1-\theta) + y_4 \log\theta$$

Paso M: Para maximizar Q con respecto a θ, derivamos primero y se obtiene:

$$Q'(\theta / \theta^{(p)}) = \frac{y_1}{\theta} \frac{\theta^{(p)}}{2+\theta^{(p)}} - \frac{(y_2 + y_3)}{1-\theta} + \frac{y_4}{\theta}$$

Igualando a cero y despejando para θ se obtiene

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

Notar que aqui no resultó una ecuación cuadrática como ocurre en el método de máxima verosimilitud. Como Q''<0 está garantizado que hay un máximo y en consecuencia

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

El algoritmo EM para familias exponenciales

Consideremos una familia exponencial con vector de parametros θ. La función de densidad o de probabilidad viene dado por

$$f(\mathbf{x}/\theta) = \exp[\mathbf{a}(\theta)\mathbf{T}(\mathbf{x}) + \mathbf{b}(\theta) + \mathbf{s}(\mathbf{x})] I_A(\mathbf{x})$$

donde I_A representa la función indicadora.

En este caso el algoritmo EM, tiene la siguiente forma simplificada

Notar que,

$$\log f(\mathbf{x}/\theta) = [\mathbf{a}(\theta)\mathbf{T}(\mathbf{x}) + \mathbf{b}(\theta) + \mathbf{s}(\mathbf{x})] I_A(\mathbf{x})$$

En consecuencia,

$$Q(\boldsymbol{\theta}/\boldsymbol{\theta}^{(p)}) = \mathbf{a}(\boldsymbol{\theta})E(\mathbf{T}(\mathbf{x})/\mathbf{y},\boldsymbol{\theta}^{(p)}) + \mathbf{b}(\boldsymbol{\theta}) + E(\mathbf{s}(\mathbf{x})/y,\theta^{(p)})$$

Sea $\mathbf{T}^{(p)}$=E[$\mathbf{T(x)/y,\theta^{(p)}}$] un estimado del estadistico suficiente **T(x).** La expresión anterior debe ser maximizada en el paso M con respecto a θ, pero el ultimo termino no depende de θ. Asi que el paso E se puede reducir a:

Paso E: Resolver

$$\mathbf{T}^{(p)}=E[\mathbf{T(x)/y,\theta^{(p)}}]$$

Sin perder generalidad podemos considerar que tenemos un solo parámetro. Derivando, con respecto a θ resulta

$$Q'(\theta/\theta^{(p)}) = a'(\theta)E(T(\mathbf{x})/\mathbf{y},\theta^{(p)}) + b'(\theta)$$

Por otro lado por la forma de la familia exponencial se puede mostrar que

$$E(T(\mathbf{x})/\theta) = \int T(\mathbf{x})\exp[a(\theta)t(\mathbf{x}) + b(\theta) + s(\mathbf{x})]d\mathbf{x} = -\frac{b'(\theta)}{a'(\theta)}$$

Luego,

$$Q'(\theta/\theta^{(p)}) = a'(\theta)E(T(\mathbf{x})/\mathbf{y},\theta^{(p)}) - a'(\theta)E(T(\mathbf{x})/\theta)$$

En consecuencia, Q'=0 si y solo sí

$$E(T(\mathbf{x})/\mathbf{y},\theta^{(p)}) - E(T(\mathbf{x})/\theta) = 0$$

Despues del paso E quedaría,

$$T^{(p)} - E(T(\mathbf{x})/\theta) = 0$$

que resulta ser el paso M.

Paso M: Determinar $\boldsymbol{\theta}^{(p+1)}$ como la solución de la ecuación

$$E(\mathbf{T(x)/\theta})=\mathbf{T}^{(p)}$$

El procedimiento aquí es equivalente a estimar primero la parte no observada de la muestra completa y luego aplicar Maxima verosimilitud para estimar el parámetro.
En el ejemplo multinomial genético

T(x)=(x1,x2,x3,x4,x5) y a(θ)=(0,log(θ),log(1-θ),log(1-θ),log(θ))

Luego $E[T(\mathbf{x})/(y_1,y_2,y_3,y_4), \theta^{(p)}]=E\{(x_1,x_2, x_3,x_4,x_5)/(y_1=x_1+x_2,y_2,y_3,y_4), \theta^{(p)}]=T^{(p)}$

producirá el estadístico suficiente $T^{(p+1)}$ con componentes

$$x_1^p = E(x_1 / y_1, \theta^{(p)}) = y_1 \frac{2}{2+\theta^{(p)}}$$

y,

$$x_2^p = E(x_2 / y_1, \theta^{(p)}) = y_1 \frac{\theta^{(p)}}{2+\theta^{(p)}},\ x_3^{p+1} = y_2, x_4^{p+1} = y_3 \text{ y}$$

$$x_5^{p+1} = y_4$$

Similarmente, E[T(x)/θ)=n(1/2,θ/4,(1-θ)/4, (1-θ)/4, θ/4)

Estos estadisticos se sustituyen en la función de densidad de los datos completos y luego se hace la estimación usual por máxima verosimilitud y se obtiene que

$$\theta^{(p+1)} = \frac{x_2^p + y_4}{x_2^p + y_2 + y_3 + y_4}$$

La función en R:

```
em2=function(y1,y2,y3,y4,tol,start)
{#Algoritmo EM par un modelo multinomial usado en genetica
#y1,y2,y3,y4 son las frecuencias observadas
#tol es el criterio de prueba para convergencia
#usualmente 10^-6 o 10^-7
#start es el valor inicial del parametro
n=y1+y2+y3+y4
psiactual=start
psi=psiactual
psilast=0
iter=0
while (abs(psilast-psi)>tol )
   {tempo=estep(psiactual,y1)
   psi=mstep(tempo$y12,tempo$y11,y4,n)
   psilast=psiactual
   psiactual=psi
   iter=iter+1
}
```

```
cat("numero de iteraciones",iter)
psiactual
}
estep=function(psiactual,y1)
 {y11=(2*y1)/(2+psiactual)
 y12=y1-y11
list(y11=y11,y12=y12)
}
mstep=function(y12,y11,y4,n)
{psinuevo=(y12+y4)/(n-y11)
}
```

La siguiente función en Matlab, halla el estimador EM de θ

```
function psi=em2(y1,y2,y3,y4,tol,start)
%Algoritmo EM par un modelo multinomial usado en genetica
%y1,y2,y3,y4 son las frecuencias observadas
% tol es el criterio de prueba para convergencia
%usualmente 10^-6 o 10^-7
%start es el valor inicial del parametro
n=y1+y2+y3+y4;
psiactual=start;
psi=psiactual;
psilast=0;
iter=0;
while (abs(psilast-psi)>tol )
   [y11, y12]=estep(psiactual,y1);
   psi=mstep(y12,y11,y4,n);
   psilast=psiactual;
   psiactual=psi;
   iter=iter+1;
end
disp('numero de iteraciones');
disp(iter);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [y11, y12]=estep(psiactual,y1)
 y11=(2*y1)/(2+psiactual);
 y12=y1-y11;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function psinuevo=mstep(y12,y11,y4,n)
psinuevo=(y12+y4)/(n-y11);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```