Introducción a la Probabilidad Bayesiana. Parte 5, Cuadrados mínimos y teoría de ajustes.

Martín Onetto

El viejo péndulo, otra vez.

Consideremos el ejemplo de la medición del período de un péndulo que vimos en la parte 4, nuevamente, pero ahora dentro de la información I vamos a incluir que el período T está determinado por un modelo físico. Vamos a considerar que las oscilaciones eran en ángulos pequeños y utilizar el resultado de que para un péndulo ideal $T = \sqrt{\frac{L}{g}}$ siendo L el largo del péndulo y g la aceleración de la gravedad. En lo que sigue consideramos que conocemos con infinita precisión al largo del péndulo L y **nos vamos concentrar entonces en ver qué podemos decir de g con este modelo y las mediciones**. Respecto a las demás consideraciones, seguimos teniendo una incerteza fija en cada medición que llamamos σ y tomamos a cada $t_i|t_j, T, I = t_i|T, I$ para todo i y j. Para n mediciones la likelihood ahora resulta:

$$\begin{aligned} P(g|D, L, \sigma, I) &\propto P(D|g, L, \sigma, I)P(g|L, \sigma, I) = P(D|g, L, \sigma, I)P(g|I) \\ &\propto \prod_i^n \exp\left\{ \frac{\left(t_i - \sqrt{\frac{L}{g}}\right)^2}{2\sigma^2} \right\} P(g|I) \end{aligned} \tag{1}$$

Vemos que al hacer la regla del producto, la prior en g tiene la dependencia $P(g|L, \sigma, I)$, sin embargo sabemos que el valor de la aceleración de la gravedad sin considerar los datos no estará vinculado a L ni a σ por lo tanto hacemos la reducción $P(g|L, \sigma, I) = P(g|I)$. Consideraremos por simpleza en este ejemplo una prior plana en g, es decir $P(g|I) \propto 1$.

Si trabajamos sobre la expresión de la likelihood, vemos que la posterior resulta:

$$P(g|D, L, \sigma, I) \propto \exp\left\{ -\frac{\left(\bar{t} - \sqrt{\frac{L}{g}}\right)^2}{2\frac{\sigma^2}{n}} \right\} \tag{2}$$

Si miramos esta distribución como función de g podríamos pensar que se parece a una gaussiana, pero no lo es. Si se quiere, se podría decir que es una gausiana en la función $\frac{1}{\sqrt{g}}$. Cómo lo que nos interesa es la distribución de g y no al de alguna función

tenemos que aceptar el resultado y en caso de hacer una estimación proceder con la **aproximación gaussiana**:

$$\log\left[P(\theta|D,I)\right] \sim \log\left[P(\theta_{max}|D,I)\right] + \frac{1}{2}\left(\frac{\partial^2 \log\left[P(\theta|D,I)\right]}{\partial\theta^2}\right)_{\theta=\theta_{max}} (\theta-\theta_{max})^2 \tag{3}$$

donde θ_{max} cumple $\left(\frac{\partial P(\theta|D,I)}{\partial\theta}\right)_{\theta=\theta_{max}} = 0$, recordemos que para que sea valida la aproximación estamos consideran que hay sólo un máximo.

Para nuestro ejemplo la distribución posterior para g con la aproximación es luego:

$$P(g|L,D,I) \sim \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left\{-\frac{(g-\mu_g)^2)}{2\sigma_g^2}\right\} \tag{4}$$

con $\mu_g = g_{max}$ y $\sigma^{-2} = \left(\frac{\partial^2 \log[P(g|D,I)]}{\partial g^2}\right)_{g_{max}}$, para calcular ésto hacemos:

$$\frac{\partial \log\left[P(g|D,I)\right]}{\partial g} = \frac{\partial\ cte}{\partial g} - \frac{\partial}{\partial g}\left(\frac{\left(\bar{t}-\sqrt{\frac{L}{g}}\right)^2}{2\sigma^2/n}\right) = -\frac{1}{2}\frac{(\bar{t}-\sqrt{\frac{L}{g}})}{\sigma^2/n}\sqrt{\frac{L}{g^3}} \tag{5}$$

$$\frac{\partial^2 \log\left[P(g|D,I)\right]}{\partial g^2} = -\frac{1}{4}\frac{1}{\sigma^2/n}\frac{L}{g^3} + \frac{3}{4}\frac{\left(\bar{t}-\sqrt{\frac{L}{g}}\right)}{\sigma^2/n}\sqrt{\frac{L}{g^5}} \tag{6}$$

con la ecuación 5 podemos calcular el máximo igualándola a cero que resulta en

$$\begin{aligned} -\frac{1}{2}\frac{(\bar{t}-\sqrt{\frac{L}{g}})}{\sigma^2/n}\sqrt{\frac{L}{g^3}} &= 0 \\ \mu_g = g_{max} &= \frac{L}{\bar{t}^2} \end{aligned} \tag{7}$$

donde μ_g es el máximo y valor medio de la distribución gaussiana que estamos contruyendo para g.

Luego, evaluando la ecuación 6 en g_{max} tenemos

$$\left(\frac{\partial^2 \log\left[P(g|D,I)\right]}{\partial g^2}\right)_{g_{max}} = -\frac{1}{4}\frac{n\left(\bar{t}^3/L\right)^2}{\sigma^2} \tag{8}$$

Definiendo un σ_g que corresponda al parámetro de dispersión de la aproximación gaussiana de la posterior de g vemos que:

$$\sigma_g^2 = -\frac{1}{\left(\frac{\partial^2 \log[P(g|D,I)]}{\partial g^2}\right)} = \frac{\sigma^2}{n\left[\bar{t}^3/(2L)\right]^2} \tag{9}$$

Como era de esperar los parámetros de la posterior de g serán función de los datos. En particular aquí toda la información de las mediciones está concentrada en dos número n y $\bar{t}$.

Es interesante notar que en la ecuación 5 corresponde a tomar el mínimo sobre el parámetro del exponente. Es decir maximizar algo de la forma $\sum_i (y_i - f(\boldsymbol{\theta})_i)^2$ por lo tanto, la aproximación gaussiana obtiene como valor estimado el mismo que se deriva de usar el método de cuadrados mínimos.

Regressions

Podemos generalizar el resultado anterior para mediciones de la forma $\{y_i, x_i\}$ donde queremos ajustar los parámetros $\boldsymbol{\theta}$ de una función $f(x, \boldsymbol{\theta})$, que corresponde al valor medio de cada y_i, dada una incerteza de medición σ. Si escribimos la likelihood tenemos:

$$P(\{y_i\}|\boldsymbol{\theta}, \{x_i\}, I) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_i \left(y_i - f(x_i, \boldsymbol{\theta})\right)^2\right\} \tag{10}$$

Si procedemos con la apoximación normal para la posterior las ecuaciónes son:

$$\frac{\partial \log\left[P\left(\boldsymbol{\theta}|\{x_i, y_i\}, \sigma, I\right)\right]}{\partial \theta_j} = \frac{1}{\sigma^2}\sum_i \left(y_i - f(x_i, \boldsymbol{\theta})\right)\frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \theta_j} \tag{11}$$

Igualando esta ecuación a cero obtenemos los valores de θ que maximizan la posterior, o likelihood ya que tomamos prior constante, de $\boldsymbol{\theta}$. La solución corresponde a los $\boldsymbol{\theta}$ que satisfacen $f(x_i, \boldsymbol{\theta}) = y_i$. Derivando nuevamente obtenemos,

$$\frac{\partial^2 \log\left[P\left(\boldsymbol{\theta}|\{x_i, y_i\}, \sigma^2, I\right)\right]}{\partial \theta_j \partial \theta_k} = \frac{1}{\sigma^2}\sum_i \left[\left(y_i - f(x_i, \boldsymbol{\theta})\right)\frac{\partial^2 f(x_i, \boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} - \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \theta_j}\frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \theta_k}\right] \tag{12}$$

Si evaluamos la expresión anterior en los $\boldsymbol{\theta}$ en los que satisfacen $f(x_i, \boldsymbol{\theta}) = y_i$, obtenemos

$$\left(\frac{\partial^2 \log\left[P\left(\boldsymbol{\theta}|\{x_i, y_i\}, \sigma^2, I\right)\right]}{\partial \theta_j \partial \theta_k}\right)_{f(x_i, \boldsymbol{\theta}) = y_i} = -\frac{1}{\sigma^2}\sum_i \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \theta_j}\frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \theta_k} \tag{13}$$

Un ajuste con datos discretos

En el caso de que tengamos mediciones de cuentas en un espectrómetro y un modelo que describe cómo se comporta la intensidad para diferentes valores de longitud de onda, el modelo estádistico más usual para la *likelihood* es la distribución de *Poisson* que está gobernada por un sólo parámetro λ

$$P(K = k|\lambda, I) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{14}$$

Para adaptar la distribución al problema de la cantidad de cuantas K en cada canal de energía ν_i escribimos:

$$P\left(K_i = k|\nu_i, \boldsymbol{\theta}, I\right) = \frac{e^{-\lambda(\nu_i,\boldsymbol{\theta})}\left(\lambda(\nu_i,\boldsymbol{\theta})\right)^k}{k!} \tag{15}$$

donde $\lambda(\nu_i, \boldsymbol{\theta})$ denota el modelo físico que tenemos para la intensidad de cada ν_i y gobernada por otros parámetros $\boldsymbol{\theta}$. Por ejemplo si estuviésemos midiendo la radiación de cuerpo negro esos parámetros podrían ser la temperatura T, la constante de Plank h, etc.

Si de nuestras mediciones k_i para algunos ν_i queremos determinar el valor de los parámetros simplemente usamos el teorema de Bayes:

$$P\left(\boldsymbol{\theta}|\{\nu_i, k_i\}, I\right) \propto \exp\left\{-\sum_i \lambda\left(\nu_i, \boldsymbol{\theta}\right)\right\} \prod_i \frac{\lambda\left(\nu_i, \boldsymbol{\theta}\right)^{k_i}}{k_i!} P(\boldsymbol{\theta}|I) \tag{16}$$

Aquí al hacer la aproximación gausiana para la distribución de $\boldsymbol{\theta}$:

$$\begin{aligned} \frac{\partial \log\left[P\left(\boldsymbol{\theta}|\{\nu_i, k_i\}, I\right)\right]}{\partial \theta_j} &= -\sum_i \frac{\partial \lambda}{\partial \theta_j} + \sum_i \frac{k_i}{\lambda_i}\frac{\partial \lambda_i}{\partial \theta_i} + \frac{\partial log\left[P(\theta|I)\right]}{\partial \theta} \\ &= \sum_i \left(\frac{k_i}{\lambda_i} - 1\right)\frac{\partial \lambda_i}{\partial \theta_j} \end{aligned} \tag{17}$$

Podemos ver que la condición para que se anule esta suma es que $\lambda\left(\nu_i, \boldsymbol{\theta}\right) = k_i$. Esto corresponde a que cada medición k_i corresponda al valor de λ en ese ν_i. A su vez, la expresión también se anula para los extremos de λ_i, es decir $\frac{\partial \lambda_i}{\partial \theta_j} = 0$. Sin embargo, esta segunda manera tiene la dificultad de que debe ser cero para cualquier valor de los datos por lo que tomamos $k_i = \lambda_i$ como máximo de $lambda_i$. Para la segunda derivada, tenemos:

$$\frac{\partial^2 \log\left[P\left(\boldsymbol{\theta}|\{\nu_i, k_i\}, I\right)\right]}{\partial \theta_j \partial \theta_k} = \sum_i \left[\left(\frac{k_i}{\lambda_i} - 1\right)\frac{\partial^2 \lambda_i}{\partial \theta_j \theta_k} - \left(\frac{k_i}{\lambda_i^2}\frac{\partial \lambda_i}{\partial \theta_k}\frac{\partial \lambda_i}{\partial \theta_j}\right)\right] \tag{18}$$

Si evaluamos en el máximo que encontramos anteriormente, $k_i = \lambda_i$, tenemos:

$$\left(\frac{\partial^2 \log\left[P\left(\boldsymbol{\theta}|\{\nu_i, k_i\}, I\right)\right]}{\partial \theta_j \partial \theta_k}\right)_{\lambda_{max}} = -\sum_i \frac{1}{k_i}\left(\frac{\partial \lambda_i}{\partial \theta_k}\frac{\partial \lambda_i}{\partial \theta_j}\right)_{(\lambda(\boldsymbol{\theta},\nu_i)=k_i)} \tag{19}$$

Un comentario sobre las varianzas obtenidas

En ambos casos generales de ajuste obtuvimos una expresión para la varianza de la distribución posteriors de los parámetros. En estos casos no usamos más información de los datos que su máximo. Podemos calcular ese objeto teniendo en cuenta toda la distribución de los datos, es decir:

$$\int \prod_i \frac{\partial^2 \log\left[P\left(\boldsymbol{\theta}|x_i, y_i, I\right)\right]}{\partial\theta_j \partial\theta_k} P(y_i|x_i, \boldsymbol{\theta}) dy_i \equiv I(\boldsymbol{\theta}|\{x_i\}) \tag{20}$$

Esta cantidad $I(\boldsymbol{\theta}|\{x_i\})$ se la conoce como la *Información de Fisher*. La cual podemos interpretar en nuestro procedimiento como el valor medio de la incerteza que tenemos de nuestros parámetros, sobre toda la distribución de los datos. Curiosamente si hacemos ambos cálculos explícitamente obtenemos:

Likelihood Gaussiana

$$\prod_i \int \frac{\partial^2 \log\left[P\left(\boldsymbol{\theta}|x_i, y_i, \sigma^2, I\right)\right]}{\partial\theta_j \partial\theta_k} P(y_i|x_i, \boldsymbol{\theta}) dy_i = \tag{21}$$

$$= \int \frac{1}{\sigma^2} \sum_i \left[\left(y_i - f(x_i, \boldsymbol{\theta})\right) \frac{\partial^2 f(x_i, \boldsymbol{\theta})}{\partial\theta_j \partial\theta_k} - \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial\theta_j} \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial\theta_k} \right] \times \tag{22}$$

$$\times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-1}{2\sigma^2} \sum_i \left(y_i - f(x_i, \boldsymbol{\theta})\right)^2 \right\} \prod_i dy_i \tag{23}$$

Aquí estamos integrando sobre la distribución de los datos, por lo que el primer término dentro de la suma desaparecerá ya que el valor medio de la gaussiana es $f(x_i, \boldsymbol{\theta})$. Luego, cómo las f no dependen de y_i resultan constantes ante la integración es así por lo que el resultado final es:

$$I_{jk}(\boldsymbol{\theta}|\{x_i\}) = -\sum_i \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial\theta_j} \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial\theta_k} \frac{1}{\sigma^2} \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i \left(y_i - f(x_i, \boldsymbol{\theta})\right)^2 \right\} \prod_i dy_i \tag{24}$$

$$= -\sum_i \frac{1}{\sigma^2} \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial\theta_j} \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial\theta_k} \tag{25}$$

Es curioso ver que el resultado de evaluarlo en el máximo o pesarlo sobre toda la distribución coincide. Un punto más para la aproximación gaussiana.

Likelihood Poisson

Veamos qué pasa para la Poisson, haciendo el mismo cálculo que antes obtenemos:

$$I_{jk}(\boldsymbol{\theta}|\{\nu_i\}) = \sum_i \sum_{k_i=0}^{\infty} \frac{\partial^2 \log\left[P\left(\boldsymbol{\theta}|\nu_i, k_i, \sigma^2, I\right)\right]}{\partial\theta_j \partial\theta_k} P(k_i|\nu_i, \boldsymbol{\theta}) \tag{26}$$

$$= \sum_i \sum_{k_i=0}^{\infty} \left\{ \left[\left(\frac{k_i}{\lambda_i} - 1 \right) \frac{\partial^2 \lambda_i}{\partial\theta_j \theta_k} - \left(\frac{k_i}{\lambda_i^2} \frac{\partial\lambda_i}{\partial\theta_k} \frac{\partial\lambda_i}{\partial\theta_j} \right) \right] e^{-\sum_i \lambda(\nu_i, \boldsymbol{\theta})} \prod_i \frac{\lambda\left(\nu_i, \boldsymbol{\theta}\right)^{k_i}}{k_i!} \right\} \tag{27}$$

En la expresión anterior tenemos términos lineal es en k_i pesados por su probabilidad, esto por definición es el valor medio de k_i El valor medio de una distribución de Poisson es $E[k_i] = \lambda_i$. Reemplazando obtenemos:

$$I_{jk}(\boldsymbol{\theta}|\{\nu_i\}) = -\sum_i \left(\frac{1}{\lambda_i} \frac{\partial\lambda_i}{\partial\theta_k} \frac{\partial\lambda_i}{\partial\theta_j} \right) \tag{28}$$

En conclusión podemos decir que a la hora de hacer un ajuste, tanto con un modelo discreto de Poisson o continuo Gaussiano, la incerteza que asignamos como $-\left(\frac{\left[P\left(\boldsymbol{\theta}|x_i, y_i, \sigma^2, I\right)\right]}{\partial\theta_j \partial\theta_k} \right)^{-1}$ coincide con la *información de Fisher* de los parámetros, dados los valores de x_i que elegimos.