Métodos numéricos para la Ciencia e Ingeniería
Tarea 11: Ajuste de Curvas Bayesiano

Felipe Toledo B.

December 14, 2015

1 Introducción

En el presente trabajo se estudian modelos orientados a describir una línea de absorción de una observación espectroscópica, presentada en la Figura 1.

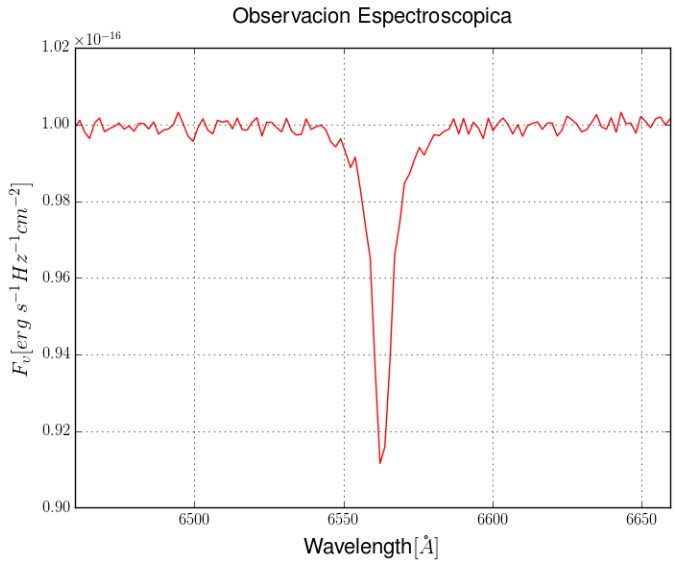

Figura 1: Datos originales observados. El nivel del continuo y la longitud de onda central de la línea de absorción son conocidos, con valores $1 \cdot 10^{-16}[\frac{erg}{s\ Hz\ cm^{-2}}]$ y $6563[\AA]$ respectivamente.

Para describir la línea se probarán dos modelos:

1. Línea Gaussiana Simple: A partir de la curva se observa que puede resultar conveniente describirla asumiendo que es de forma Gaussiana. En la ecuación (1) se explicita el modelo propuesto.

$$M_1(\lambda; A, b) = C_0 - Ae^{\frac{-(\lambda-\lambda_0)^2}{2b^2}} \tag{1}$$

   Éste posee dos parámetros libres, A y b. El valor de λ_0 y C_0 se asume conocido, donde λ_0 es la longitud de onda central de la línea de absorción y C_0 corresponde al nivel del continuo.

2. Línea Gaussiana Doble: El modelo está explicitado en la ecuación (2). Se tienen los mismos parámetros conocidos λ_0 y C_0, pero en éste hay cuatro parámetros libres: A_1, A_2, b_1, b_2.

$$M_2(\lambda; A_1, A_2, b_1, b_2) = C_0 - A_1 e^{\frac{-(\lambda-\lambda_0)^2}{2b_1^2}} - A_2 e^{\frac{-(\lambda-\lambda_0)^2}{2b_2^2}} \tag{2}$$

Los objetivos son:

1. Definir las distribuciones a Priori para cada parámetro.
2. Estimar el valor de los parámetro de cada modelo junto a su intervalo de credibilidad al 68%, usando métodos Bayesianos.
3. Escoger el modelo que entregue una mejor representación de los datos utilizando técnica Bayesianas.

2 Metodología

Para obtener una estimación de los parámetros $\mathbf{\Theta}$ para cada modelo M, se utiliza la regla de Bayes (3) y los datos empíricos $\mathbf{x} = (\lambda, \mathbf{F})$.

$$P(\Theta|\mathbf{x}, M) = \frac{P(\mathbf{x}|\Theta, M)P(\Theta|M)}{P(\mathbf{x}|M)} \tag{3}$$

El término $P(\Theta|\mathbf{x}, M)$ indica la probabilidad a posteriori de que los parámetros representen a los datos $\mathbf{x}$ para el modelo M. Se calcula utilizando la distribución de probabilidad de la derecha, donde $P(\mathbf{x}|\Theta, M)$ es la verosimilitud, que mide la cercanía de los datos respecto a $M(\Theta)$. La probabilidad $P(\Theta|M)$ corresponde a la distribución a priori de los parámetros, la que se fija arbitrariamente. Finalmente, el término $P(\mathbf{x}|M)$ es difícil de calcular por lo que se utiliza como constante de normalización, de forma que (3) queda como (4).

$$P(\Theta|\mathbf{x}, M) = K \cdot P(\mathbf{x}|\Theta, M)P(\Theta|M) \tag{4}$$

Para todos los casos se utiliza la función de verosimilitud (5), donde M representa el modelo a evaluar. El valor de σ corresponde al error asociado a cada punto, el cual se estima usando la desviación estándar de los puntos lejanos a la línea de absorción sobre el continuo ($\approx$ error de medición).

$$P(\mathbf{x}|\Theta, M) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{N}(F_i - M(\lambda_i, \Theta))^2} \tag{5}$$

La probabilidad a priori de los parámetros de cada distribución dependerá de cada modelo y se detalla en las secciones 2.1 y 2.2. Con todo lo anterior, se puede calcular $P(\Theta|\mathbf{x}, M)$, la que permite estimar los valores óptimos[1] y sus intervalos de credibilidad.

2.1 Parámetros modelo 1: Línea Gaussiana Simple

El modelo descrito por (1) posee dos parámetros libres, A y b. Para utilizar técnicas Bayesianas de estimación es necesario definir una probabilidad a priori para el valor de cada parámetro. Dado que ambos valores están asociados a datos de un fenómeno continuo (mediciones físicas), serán modelados utilizando funciones Gaussianas.

[1] En este caso se utilizará como valor óptimo Θ^* al valor $\mathbf{E}(\Theta)$ calculado usando $P(\Theta|\mathbf{x}, M)$.

2.1.1 Parámetro *A*

A priori se decide que la distribución de A es:

$$P(A) = \frac{1}{\sqrt{2\pi}\sigma_A} e^{\frac{-(A-\bar{A})^2}{2\sigma_A^2}} \tag{6}$$

Para determinar el valor de cada parámetro se debe tener en cuenta que A representa la amplitud en Flujo de la curva Gaussiana del modelo. Para determinarlos se presumirá que las mediciones no están muy alejadas de los valores medios esperados. Con todo esto, se decide lo siguiente:

- $\bar{A}$ será la amplitud máxima medida respecto al contínuo.
- σ_A corresponderá al error de medición de Flujo. Se utilizará el mismo valor que el σ de la verosimilitud, ya que corresponde al mismo fenómeno.

2.1.2 Parámetro *b*

Análogamente, la distribución escogida para b es:

$$P(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{\frac{-(b-\bar{b})^2}{2\sigma_b^2}} \tag{7}$$

- $\bar{b} = 0.8493|\lambda_0 - \lambda(\frac{\bar{A}}{2})|$
- $\sigma_b = \Delta$ Se estima a partir de la resolución en longitud de onda (resolución horizontal) del modelo.

El valor de $\bar{b}$ se determina usando la aproximación *Full Width at Half Maximum* (FWHM), donde $\lambda(\frac{\bar{A}}{2})$ es uno de los valores de longitud de onda en que la línea de absorción cae a la mitad en amplitud[2].

2.2 Parámetros modelo 2: Línea Gaussiana Doble

En este caso la distribución de probabilidad de cada parámetro también será Gaussiana y de forma análoga a los de la Sección 2.1. Al determinar los parámetros a Priori se aprovecha la disponibilidad de dos Gaussianas en el modelo, para utilizar una en el ajuste del extremo delgado de la línea y la otra para la base ancha. Entonces, en resumen:
Para las variables de la curva delgada, A_1 y b_1:

- $\bar{A_1} = \frac{2\bar{A}}{3}$ - La altura media de la línea de absorción debe ser $\approx \bar{A}$ al sumar las exponenciales.
- $\sigma_{A_1} = \frac{2\sigma_A}{3}$ - El error en F recibe el mismo escalamiento que A_1.
- $\bar{b_1} = \bar{b}$
- $\sigma_{b_1} = \sigma_b$ - Error asociado al muestreo.

Y para la curva ancha, A_2 y b_2:

- $\bar{A_2} = \frac{\bar{A}}{3}$
- $\sigma_{A_2} = \frac{\sigma_A}{3}$
- $\bar{b_2} = 3\bar{b}$ - Se estima que la base es aproximadamente tres veces mas ancha que la zona delgada.
- $\sigma_{b_2} = \sigma_b$

[2]Se asume que la línea de absorción es aproximadamente simétrica.

3 Selección de Modelo

Para escoger el modelo más apropiado se comparan sus funciones de verosimilitud (8) evaluadas en los datos originales.

$$P(\mathbf{x}|M) = \int P(\mathbf{x}|\Theta, M_i) d\Theta \tag{8}$$

Así, se escogerá como mejor modelo a aquel con el valor de (8) más elevado.

4 Resultados

Para operar numéricamente con los datos, y evitar *overflows* o divisiones por cero, se escalan los valores del espectro multiplicándolo por $2.5 \cdot 10^{18}$. Esta corrección es considerada e invertida al momento de retornar los valores calculados.
El error de los datos, calculado usando el estimador Gaussiano de varianza[3] respecto al continuo, es $\sigma^2 = 0.1695$.
A continuación se presentan en detalle los valores de cada término introducido en la sección Metodología para su modelo correspondiente, junto al ajuste logrado.

4.1 Modelo 1: Línea Gaussiana Simple

Los parámetros estimados a priori son:

- $\bar{A} = 8.848 \cdot 10^{-18} [\frac{erg}{s\ Hz\ cm^{-2}}]$
- $\sigma_A = 1.646 \cdot 10^{-19} [\frac{erg}{s\ Hz\ cm^{-2}}]$
- $\bar{b} = 4.0[\mathring{A}]$
- $\sigma_b = 1.6[\mathring{A}]$

Las distribuciones de probabilidad a posteriori para cada estimador pueden observarse en las Figuras 2 y 3. A partir de ellas se determinan los valores Θ^*.
Para los intervalos de credibilidad se integra cada distribución desde su valor medio hacia afuera simétricamente, deteniéndose en los valores extremos tales que el valor del área sea 0.68. El enfoque es válido debido a que la distribución a posteriori para cada parámetro está asociada a una densidad de probabilidad.
Así, en resumen, se obtiene:

- $A^* = 8.846 \cdot 10^{-18} [\frac{erg}{s\ Hz\ cm^{-2}}]$
- Invervalo de credibilidad de A al 68%: $A \in [8.618,\ 9.043] \cdot 10^{-18} [\frac{erg}{s\ Hz\ cm^{-2}}]$
- $b^* = 4.2782$
- Invervalo de credibilidad de b al 68%: $b \in [4.109,\ 4.417] \cdot 10^{-18} [\frac{erg}{s\ Hz\ cm^{-2}}]$

Con los valores determinados a posteriori para el Modelo 1, se obtiene el ajuste de la Figura 4.

[3] Se calcula usando los datos con $\lambda < 6525[\mathring{A}]$ y $\lambda > 6600[\mathring{A}]$. El estimador usado es $\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{N} (F_i - C_0)$.

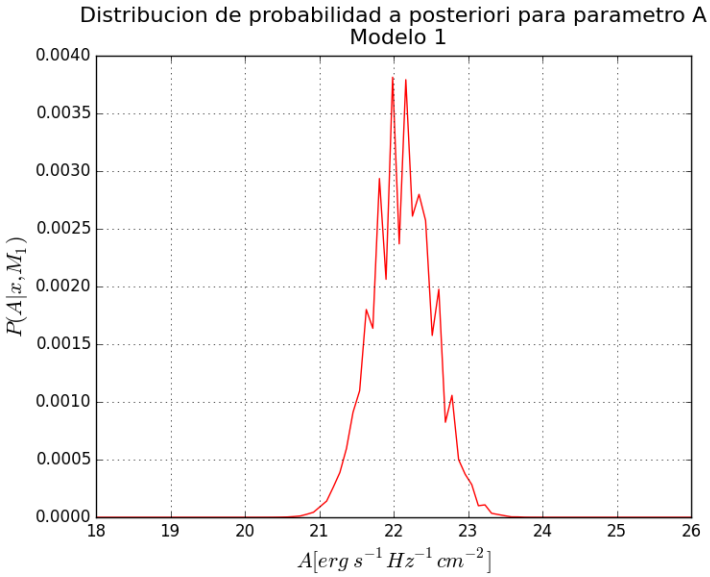

Figura 2: Distribución de probabilidad a posteriori para parámetro A del Modelo 1.

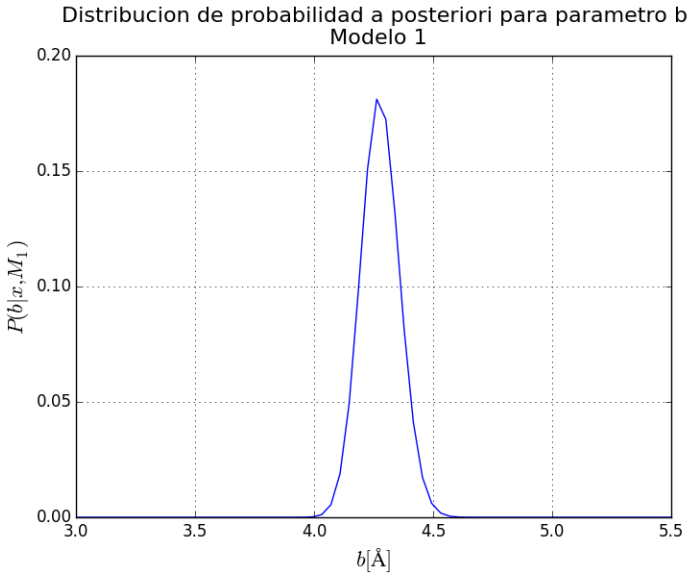

Figura 3: Distribución de probabilidad a posteriori para parámetro b del Modelo 1.

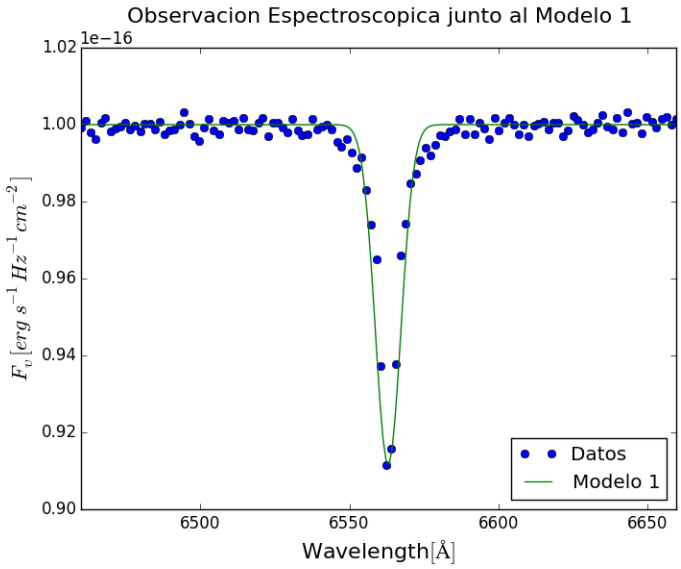

Figura 4: Ajuste bayesiano del Modelo 1 junto a los datos originales.

5 Conclusiones

Se concluye que el método Bayesiano de ajuste de curvas es una buena herramienta para verificar la capacidad que tiene un modelo para describir un fenómeno, mediante la herramienta de funciones de verosimilitud. Esta metodología resulta especialmente atractiva para verificar las capacidades predictivas de modelos teóricos mediante contraste con resultados experimentales. También tiene potenciales aplicaciones en inteligencia artificial. Un sistema inteligente podría inventarse muchos modelos e ir discriminandólos y ajustándolos usando sus observaciones y criterios Bayesianos, de forma que se converja a una *mejor descripción* de su sistema en estudio.
El requerimiento de definir distribuciones a priori para cada parámetro demanda una comprensión, al menos superficial, del fenómeno a describir. Lo mínimo que se necesita conocer es el tipo de distribución. En caso de que no se conozcan los parámetros para alguna variable, se puede lograr un ajuste mediante iteraciones con distintos parámetros en las distribuciones a priori, por ejemplo usando un algoritmo tipo Monte Carlo. Se escogerían entonces aquellos parámetros que maximicen la verosimilitud del modelo. Este tipo de procedimiento queda limitado por la disponibilidad de tiempo de cálculo, que como se sabe puede resultar muy elevado para este tipo de algoritmos. El disponer de mayor información sobre un sistema permite, entonces, obtener tiempos de cálculo más reducidos.
Finalmente, este trabajo ilustra sobre la importancia de implementar métodos de optimización en operaciones matemáticas muy repetitivas. Si se hubiese dispuesto de más series de datos el tiempo de ejecución del programa hubiese aumentado considerablemente debido a la cantidad de integrales que se calculan para obtener el valor de cada parámetro. En caso de dedicarse al trabajo en esta línea, o con algoritmos de monte carlo, se recomienda estudiar técnicas de optimización de código tanto a nivel de algoritmos como conscientes de la arquitectura del procesador (por ejemplo utilizar *multithreading)*, ya que tanto esta experiencia como las anteriores indican que el tiempo de cómputo es una limitante crucial.