# Informe 11 - Estimación de parámetros y credibilidad con métodos Bayesianos

Profesor: Valentino González
Auxiliar: Felipe Pesce
Integrantes: Nicolás Troncoso Kurtovic[1]
[1] *Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Departamento de Física.*

## I. INTRODUCCIÓN

En el siguiente informe se presenta implementación y resultados obtenidos para métodos de estimación de parámetros y credibilidad de estos a partir de la estadística bayesiana

Los datos experimentales a ajustar provienen de un espectro de emisión con una clara línea de absorción H$\alpha$, correspondiente al color Rojo. Esta línea tiene su origen en la transición de un $e^-$ desde el nivel 2 hasta el nivel 3 de un átomo de Hidrógeno, y pertenece a la serie de Balmer. El espectro está compuesto por una línea horizontal de flujo entre las longitudes de onda $\lambda_{\min} = 6460$ y $\lambda_{\max} = 6659{,}6$ y una línea de absorción centrada en $\lambda = 6563$, tal como puede verse en la Figura 1. El nivel base de flujo corresponde a $1 \cdot 10^{-16}$ erg cm$^{-2}$ s$^{-1}$ Å$^{-1}$.

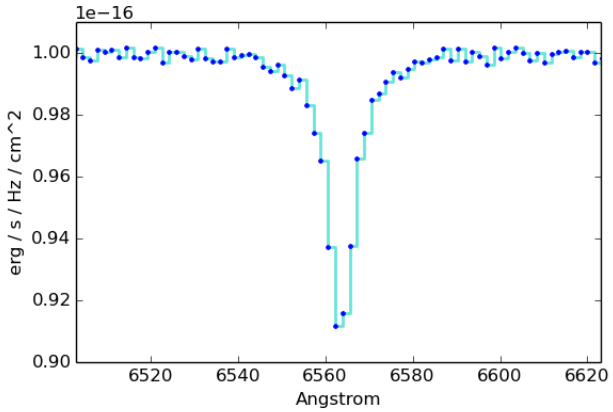


Figura 1: Espectro medido. Es posible observar una linea de absorción en el nivel base centrada en $\lambda = 6563$ Å.

El objetivo de este trabajo es ajustar un modelo teórico a estos datos, lo que nos permitirá deducir el origen físico del ensanchamiento de la línea de absorción. En estrellas, por ejemplo, el ensanchamiento se puede dar debido a colisiones, efecto Stark, efecto Zeeman, Doppler termal y otros mecanismos macroscópicos asociados a turbulencias.

Definimos una función gaussiana como:

$$g(x, A, \mu, \sigma) \;=\; \frac{A}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x \,-\, \mu)^2}{2\sigma^2}\right). \qquad \text{(I.1)}$$

Para ajustar los datos se proponen dos modelos distintos: Un ajuste gaussiano y un ajuste doble gaussiano, correspondientes a las ecuaciones (I.2) y (I.3) respectivamente:

$$f_1(x) \;=\; b - g(x, A_1, \mu, \sigma_1), \qquad \text{(I.2)}$$

$$f_2(x) \;=\; b - g(x, A_{21}, \mu, \sigma_{21}) - g(x, A_{22}, \mu, \sigma_{22}), \qquad \text{(I.3)}$$

con $b = 10^{-16}$ el nivel base de flujo en las unidades trabajadas. Las funciones $f_1$ y $f_2$ comparten el nivel base y el valor $\mu = 6563$. El modelo gaussiano, correspondiente a la ecuación (I.2) posee dos parámetros libres: la amplitud $A_1$ y el ensanchamiento $\sigma_1$, mientras que el modelo doble gaussiano, correspondiente a la ecuación (I.3) es más complejo y posee 4 parámetros libres: dos amplitudes $A_{21}$ y $A_{22}$ y dos desviaciones $\sigma_{21}$ y $\sigma_{22}$.

Sean $\vec{d}$ el conjunto de datos medidos, $\Theta$ el modelo para ajustarlos y $\vec{\theta}$ un set de parámetros. La estadística Bayessiana considera a los parámetros $\vec{\theta}$ como variables aleatorias, y por lo tanto lo que se buscará es la probabilidad de que cada set de parámetros ajuste los datos $\vec{d}$ a partir del modelo $\Theta$. Se cumple que:

$$\mathbb{P}(\vec{\theta}\,|\,\vec{d}, \Theta) \;=\; \frac{\mathbb{P}(\vec{d}\,|\,\vec{\theta}, \Theta) \cdot \mathbb{P}(\vec{\theta}\,|\,\Theta)}{\mathbb{P}(\vec{d}\,|\,\Theta)}, \qquad \text{(I.4)}$$

donde cada probabilidad corresponde a:

- $\mathbb{P}(\vec{\theta}\,|\,\vec{d}, \Theta)$: Probabilidad de que los parámetros $\vec{\theta}$ sean los que ajustan un modelo $\Theta$ a los datos $\vec{d}$.
- $\mathbb{P}(\vec{d}\,|\,\vec{\theta}, \Theta)$: Verosimilitud
- $\mathbb{P}(\vec{\theta}\,|\,\Theta)$: Probabilidad a Priori
- $\mathbb{P}(\vec{d}\,|\,\Theta)$: Factor de normalización

Para encontrar los parámetros se tratará a estos como variables aleatorias y se buscará la esperanza de cada uno. Supondremos que los parámetros tienen un comportamiento gaussiano. Se utilizarán las desviaciones de cada gaussiana correspondiente a los parámetros para estimar el intervalo de credibilidad y así escoger el modelo que mejor ajusta los datos.

Para agregar un valor extra que nos permita estudiar los dos modelos propuestos, se calculará el valor de $\chi^2$ de cada ajuste en base a:

$$\chi^2 \;=\; \sum_i \Big(y_i \;-\; \Theta(x_i, \vec{\theta})\Big)^2. \qquad \text{(I.5)}$$

En la sección **II** se presenta la metodología a utilizar y su implementación. En la sección **III** se muestran los resultados obtenidos. En la sección **IV** se discutirán los resultados y se concluirá sobre ellos.

## II. METODOLOGÍA

En base a los modelos gaussiano (I.2) y doble gaussiano (I.3) se buscaron los parámetros $A_1$, $A_{21}$, $A_{22}$, $\sigma_1$, $\sigma_{21}$ y $\sigma_{22}$ que mejor ajustasen los datos experimentales. Se utilizó una implementación predefinida de la función gaussiana `norm` [1] perteneciente a la librería `scipy.stats`.

Suponemos que los errores de medición son gaussianos y uniformes para toda longitud de onda $\lambda$. Para calcularlos buscamos la desviación estándar $\varepsilon$ de los datos sobre el nivel base de flujo. Para ello se fijan puntos donde se considera que la desviación del flujo no corresponde a error sino que a las gaussianas, y se calcula el error estándar de los puntos fuera de ese conjunto, como se muestra en la Figura 2.

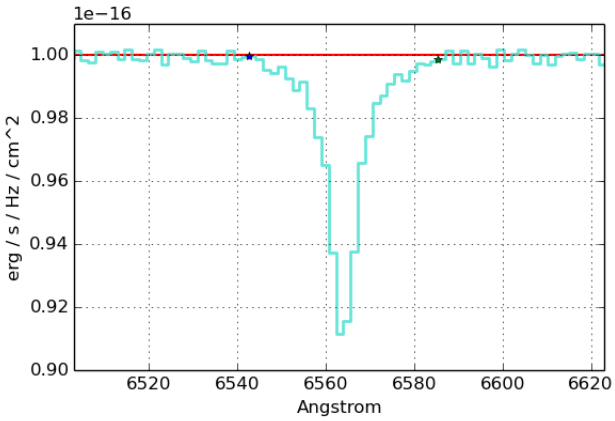


Figura 2: Espectro medido. La línea roja representa el nivel base. Las estrellas en verde representan los límites considerados donde la influencia de la gaussiana es menor que el error de medición. El cálculo de la desviación estándar de los datos se realizó utilizando todos los $\lambda$ que no estuviesen dentro del conjunto de la línea de absorción.

Para el cálculo de (I.4) se debe calcular la verosimilitud y la probabilidad priori. El vector de datos medidos $\vec{d}$ tiene dos componentes: $\vec{x}$ que en este caso corresponde al arreglo de $\lambda$ y $\vec{y}$ que en este caso corresponde al flujo. Dicho esto, para el calculo de la verosimilitud ocupamos la fórmula:

$$\mathbb{P}(\vec{d}\,|\vec{\theta},\Theta) = \frac{1}{(2\pi\varepsilon)^{N/2}} \exp\left(-\frac{1}{2\varepsilon^2}\sum_i y_i - f_j(x_i,\vec{\theta})\right), \tag{II.1}$$

con $j$ dependiendo del modelo. Para la probabilidad a priori tenemos:

$$\begin{aligned} S_i &= \frac{(\beta_i-\mu_i)^2}{\sigma_i)^2} \\ S &= \sum_i \frac{-1}{2S_i} \\ \mathbb{P}(\vec{\theta}\,|\Theta) &= \frac{\exp(S)}{4\pi^2(\prod_i \sigma_i)}, \end{aligned} \tag{II.2}$$

donde se supuso que el vector $\vec{\theta}$ esta compuesto por dos arreglos que representan el valor medio $\mu_i$ y la desviación estándar $\sigma_i$ de cada parámetro (recordemos que se supusieron con errores gaussianos) y $\beta_i$ es el valor esperado a priori para el parámetro $i$.

La multiplicación $\mathbb{P}(\vec{d}\,|\vec{\theta},\Theta)\cdot\mathbb{P}(\vec{\theta}\,|\Theta)$, que llamaremos matriz *post*, es proporcional a la probabilidad buscada, por lo que bastará con sumar todas las probabilidades obtenidas para luego poder normalizar, y así obtener (I.4).

La matriz post tiene la forma $N\times N$ en el caso del ajuste gaussiano simple, con $N$ el número de bins a utilizar, esto debido a que se tienen dos parámetros y queremos conocer la probabilidad de cada combinación, sin embargo en el caso de la doble gaussiana son 4 parámetros a ajustar, por lo que la matriz es de la forma $N\times N\times N\times N$, aumentando enormemente el tiempo que tarda el algoritmo en completarse.

Debido al tratamiento probabilístico que se le da a los parámetros, nos interesará conocer el valor esperado $\mathbb{E}$ de cada parámetro, que tomaremos como el parámetro que entrega el mejor ajuste, además de la desviación estándar $\sigma$ que tomaremos como el intervalo de credibilidad del valor obtenido.

Para obtener el valor esperado se sumará sobre las otras variables y se multiplicará por el ancho de los intervalos equiespaciados sobre los cuales se construyó la grilla. Los valores específicos obtenidos no son de mayor interés, puesto que nos interesa la diferencia relativa entre ellos (en efecto, los valores sólo toman sentido después de la normalización).

Para tener un punto de partida y también de comparación de los parámetros a obtener se utilizó la función `curve_fit` [2] de la librería `scipy.optimize`. Esta función utiliza la minimización de la diferencia de cuadrados para encontrar, a partir de un set de semillas propuestas, los parámetros que mejor ajustan los datos según una función específica. Por separado, se utilizó la ecuación (I.5) para calcular el $\chi^2$ de cada ajuste.

## III. RESULTADOS

El ajuste mediante la función `curve_fit` permitió obtener los parámetros del Cuadro I para cada modelo.

Para la obtención de los parámetros a través del método bayesiano se utilizó una discretización de $N = 200$ y $N = 50$ respectivamente para cada modelo. La razón

| Parámetro | Resultado | Unidades |
|---|---|---|
| $A_1$ | $7{,}617 \cdot 10^{-17}$ | erg Å/ (cm$^2$ s Hz) |
| $\sigma_1$ | 3,702 | Å |
| $\chi^2_1$ | $1{,}149 \cdot 10^{-35}$ | erg$^2$ / (cm$^4$ s$^2$ Hz$^2$) |
| $A_{21}$ | $4{,}106 \cdot 10^{-17}$ | erg Å/ (cm$^2$ s Hz) |
| $A_{22}$ | $4{,}855 \cdot 10^{-17}$ | erg Å/ (cm$^2$ s Hz) |
| $\sigma_{21}$ | 2,444 | Å |
| $\sigma_{22}$ | 8,419 | Å |
| $\chi^2_2$ | $3{,}649 \cdot 10^{-36}$ | erg$^2$ / (cm$^4$ s$^2$ Hz$^2$) |

Cuadro I: Datos obtenidos a partir del ajuste con `curve_fit`.

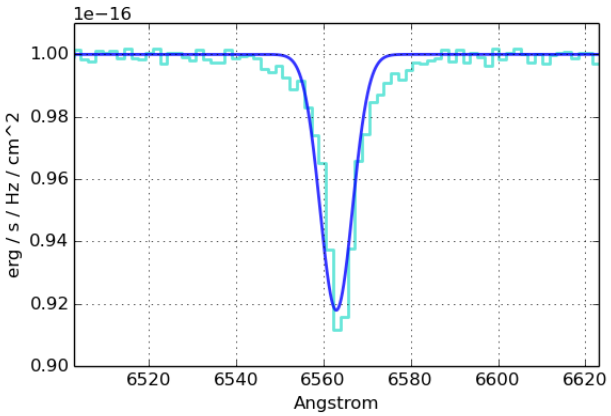


Figura 3: Ajuste de gaussiana a los datos del espectro medido, utilizando los datos de la Cuadro I. Los parámetros obtenidos con el método bayesiano, presentados en el Cuadro II, no presentan una diferencia visible con respecto al ajuste con `curve_fit` en la escala graficada.

de el $N$ menor para el modelo doble gaussiana reside principalmente en el tiempo que tarda el algoritmo en correr.

| Parámetro | Resultado | Unidades |
|---|---|---|
| $A_1$ | $7{,}588 \cdot 10^{-17}$ | erg Å/ (cm$^2$ s Hz) |
| $\sigma_1$ | 3,684 | Å |
| $\chi^2_1$ | $1{,}149 \cdot 10^{-35}$ | erg$^2$ / (cm$^4$ s$^2$ Hz$^2$) |
| $A_{21}$ | $4{,}220 \cdot 10^{-17}$ | erg Å/ (cm$^2$ s Hz) |
| $A_{22}$ | $4{,}567 \cdot 10^{-17}$ | erg Å/ (cm$^2$ s Hz) |
| $\sigma_{21}$ | 3,218 | Å |
| $\sigma_{22}$ | 7,555 | Å |
| $\chi^2_2$ | $8{,}302 \cdot 10^{-36}$ | erg$^2$ / (cm$^4$ s$^2$ Hz$^2$) |

Cuadro II: Datos obtenidos a partir del ajuste con métodos bayesianos.

Para los modelos ajustados a través del método bayesiano se obtuvieron valores de $\mathbb{P}(\vec{d}|\Theta)$ presentados en la ecuación (III.1):

$$\begin{aligned} \mathbb{P}(\vec{d}|f_1) &= 3{,}44 \cdot 10^{-50} \\ \mathbb{P}(\vec{d}|f_2) &= 2{,}76 \cdot 10^{-63}, \end{aligned} \tag{III.1}$$

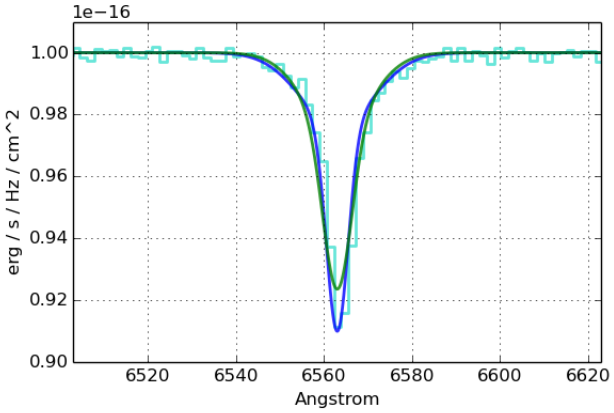


Figura 4: Ajuste de doble gaussiana a los datos del espectro medido. La línea azul representa la curva doble gaussiana con los parámetros entregados por `curve_fit` presentados en el Cuadro I. La línea verde representa la curva doble gaussiana con los parámetros entregados por el método bayesiano presentados en el Cuadro II.

## IV. DISCUSIÓN Y CONCLUSIÓN

A partir de los resultados obtenidos podemos concluir que el método bayesiano presenta una alternativa a los métodos de ajuste de modelos vistos anteriormente en el curso, sin embargo sus ventajas estadísticas por sobre el modelo frecuentista tienen como costo una mayor complejidad de implementación. Además, sus resultados son fuertemente dependientes de los valores priori que le sean entregados, lo cual es un punto desfavorable del método.

Los ajustes realizados con `curve_fit` tampoco están excentos de problemas, en efecto, las semillas entregadas para realizar el ajuste también determinan los resultados finales.

De los Cuadros I y II podemos observar similitud entre los resultados obtenidos, sin embargo el método de `curve_fit` logra mejores resultados visibles en las Figuras 3 y 4, y también apreciables mediante la comparación del parámetro objetivo $\chi^2$.

Ambos métodos indican que el mejor ajuste es el de doble gaussiana, lo cual es abalado por el cálculo de $\chi^2$. La condición que debió haber entregado la ecuación III.1 debió sin embargo haber sido distinta. Errores sistemáticos asociados al cálculo de este valor son muy probables.

En general, este método posee pasos dependientes cada uno del resultado anterior, y basta que en alguna parte del algoritmo se cometa un error para que todo el resultado final pierda interpretación.

[1] http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.norm.html
[2] http://docs.scipy.org/doc/scipy-0.16.0/reference/generated/scipy.optimize.curve_fit.htmll