Informe 10 - Ajuste de curva y testeo de validez con Kolmogorov-Smirtnov

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 ajuste de datos, junto con un análisis de validez según el test de Kolmogorov-Smirnov.

Los datos experimentales a ajustar provienen de un espectro de emisión con una clara línea de absorción, que pueden verse en la Figura 1. Es posible observar que los datos tienen bastante ruido.

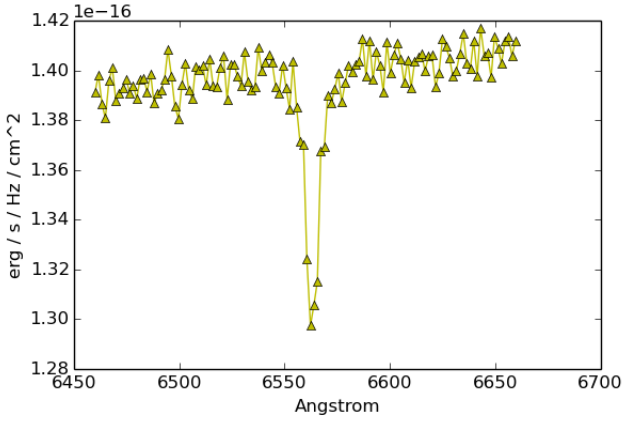

Figura 1: Espectro medido. Es posible observar una linea de absorción.

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.

Se proponen dos modelos distintos: Un ajuste gaussiano y un ajuste lorentziano, correspondientes a las ecuaciones (I.1) y (I.2) respectivamente:

$$f_g(x) = a + bx - \frac{A}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \quad \text{(I.1)}$$

$$f_l(x) = a + bx - \frac{A}{\pi\sigma\left(1+\frac{(x-\mu)^2}{\sigma^2}\right)} \quad \text{(I.2)}$$

Tras encontrar los parámetros a, b, A, μ, σ que mejor ajusten cada modelo, se procederá a someter ambos ajustes al test de Kolmogorov-Smirtnov para verificar que tengamos al menos un 95 % de confianza.

El test de Kolmogorov-Smirtnov determina, en base a la distancia máxima de probabilidad entre las funciones de distribución acumuladas (CDF en adelante), la confiabilidad con que ambas CDF provengan de la misma función. En nuestro caso, deseamos encontrar el modelo que mejor describe los datos observados.

Debido a que los errores asociados a la medición que realizan los pixeles de un CCD no son gaussianos, el test de Kolmogorov-Smirtnov representa una buena opción para calcular la confiabilidad, ya que no necesita información sobre los errores.

El principal parámetro que determinará la confianza será α, que representa la probabilidad de que rechacemos el modelo cuando este es verdadero.

Para agregar un valor extra que nos permita definir alguno de los dos modelos propuesto, se calculará el valor de χ^2 de cada ajuste en base a:

$$\chi^2 = \sum (y - f(x))^2 \quad \text{(I.3)}$$

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.1) y lorentziano (I.2) se buscaron los parámetros a, b, A, μ, σ que mejor ajustasen los datos experimentales. Se utilizó una implementación predefinida de la función gaussiana `norm` [1] y una de la función de lorentz `cauchy` [2], ambas pertenecientes a la librería `scipy.stats`

Para encontrar estos parámetros se utilizó la función `curve_fit` [3] 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.3) para calcular el χ^2 de cada ajuste.

Posteriormente, se utilizó una implementación de Kolmogorov-Smirtnov en `python` perteneciente a la librería `scipy.stats`. El objeto `kstwobign` determina la

distancia máxima que puede haber entre dos CDF para que se tenga una confiabilidad de un $X\,\%$. En este caso se buscaba una confiabilidad de $95\,\%$.

Para calcular el parámetro α se utilizó tanto las funciones pertenecientes al objeto `kstwobign` como una función llamada `kstest`, también de la librería `scipy.stats`. Esta última retorna la distancia máxima entre CDF y el valor $(1-\alpha)$ a partir de los datos y de la función que calcula las CDF.

III. RESULTADOS

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

	Gaussiano	Lorentziano	Unidades
a	$8{,}877 \cdot 10^{-17}$	$8{,}811 \cdot 10^{-17}$	erg / (cm^2 s Hz)
b	$7{,}803 \cdot 10^{-21}$	$7{,}923 \cdot 10^{-21}$	erg / (cm^2 s Hz Å)
A	$8{,}222 \cdot 10^{-17}$	$1{,}114 \cdot 10^{-16}$	erg Å/ (cm^2 s Hz)
μ	6563,22	6563,20	Å
σ	3,258	3,219	Å
χ^2	$5{,}204 \cdot 10^{-35}$	$5{,}006 \cdot 10^{-35}$	erg^2 / (cm^4 s^2 Hz^2)

Cuadro I

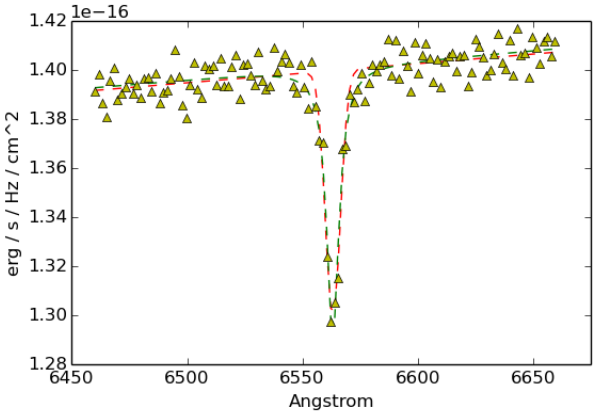

Figura 2: Gráfico de flujo por longitud de onda. Los triángulos amarillos representan los datos experimentales. En rojo el ajuste gaussiano utilizando la función de la ecuación (I.1) y en verde el ajuste lorentziano utilizando la función de la ecuación (I.2), ambos evaluados con los parámetros del Cuadro I.

Para el test de Kolmogorov-Smirtnov se utilizó tanto las funciones de `kstwobign` con la función `kstest`. Para cada uno se obtuvo α_m y α_i respectivamente. Los resultados de la distancia máxima entre las CDF de los modelos D_{max} y la distancia crítica si queremos que el modelo tenga un $95\,\%$ de confiabilidad se tabulan en el Cuadro II.

La comparación entre las CDF experimental y del modelo gaussiano puede verse en la Figura 3. En la Figura 4 se ve la comparación de CDF con el modelo lorentziano.

	Gaussiano	Lorentziano
D_{max}	0.165	0.166
$D_c(95\,\%)$	0.047	0.047
α_m	0.9973	0.9976
α_i	0.9976	0.9979

Cuadro II: D_{max} es la distancia máxima entre CDF de datos y modelo. $D_c(95\,\%)$ es la distancia crítica si se desea un $95\,\%$ de confiabilidad. El parámetro α calculado según cada método se tabula para cada ajuste.

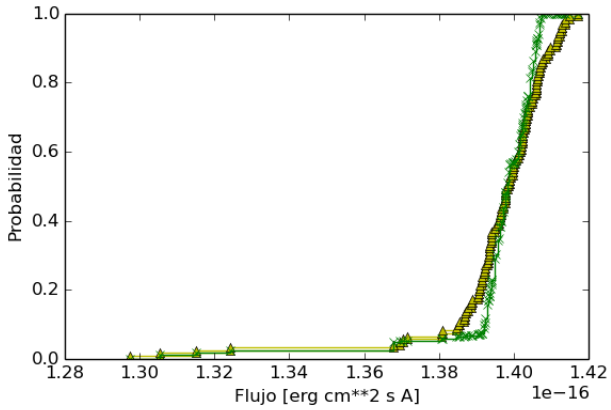

Figura 3: Los triangulos amarillos representan la CDF de datos experimentales. Las cruces verdes representan la CDF del modelo gaussiano.

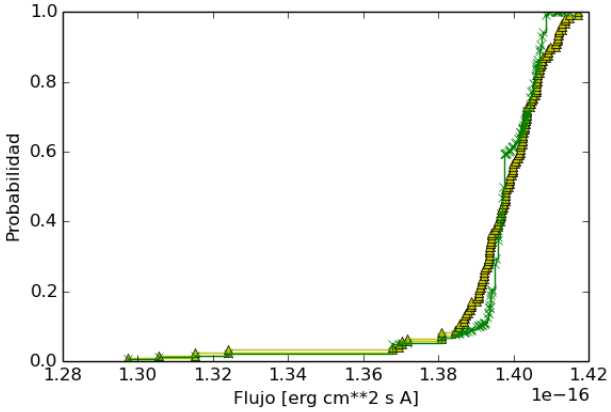

Figura 4: Los triangulos amarillos representan la CDF de datos experimentales. Las cruces verdes representan la CDF del modelo lorentziano.

IV. DISCUSIÓN Y CONCLUSIÓN

De los resultados obtenidos podemos observar que ninguno de nuestros modelos cumple que con un 95 % de certeza estemos modelando bien el fenómeno observado.

En efecto, ambos modelos no cumplen con el criterio del test de Kolmogorov-Smirtnov, pues la distancia máxima entre las CDF de los datos experimentales y de ambos ajustes es mayor a la distancia crítica propuesta por este test para tener el nivel de confiabilidad deseado.

Sin perjuicio de lo anterior, encontramos que los valores α (que representan la probabilidad de que el modelo sea rechazado cuando este es válido) es muy alto para ambos ajustes, cercano a 1, es decir, es altamente probable que los modelos sean rechazados a pesar de ser válidos. Dentro de esta lógica, el que tiene mayor probabilidad de ser rechazado es el ajuste lorentziano. Encontramos buena similitud entre el valor de α calculado utilizando `kstwobign` y `kstest`.

En base a nuestros conocimientos anteriores, podemos decir que el ajuste lorentziano es levemente más acertado que el ajuste gaussiano, debido a que $\chi^2_{lorentz} < \chi^2_{gauss}$.

El mayor problema al momento de modelar estos datos se encuentra en el ruido que presenta la zona lineal de los datos, comprendida en los flujos 1,38 y 1,42 en erg / (cm^2 s Hz). Esta dispersión en los flujos provoca que la CDF de los datos experimentales tenga menor pendiente que la de los modelos, aumentando asi la distancia máxima D_m. Si fuese posible reducir el error en esta zona, entonces podríamos obtener mayor información sobre el mecanismo físico que produce el ensanchamiento de la linea validando alguno de los modelos con el test de Kolmogorov-Smirtnov.

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