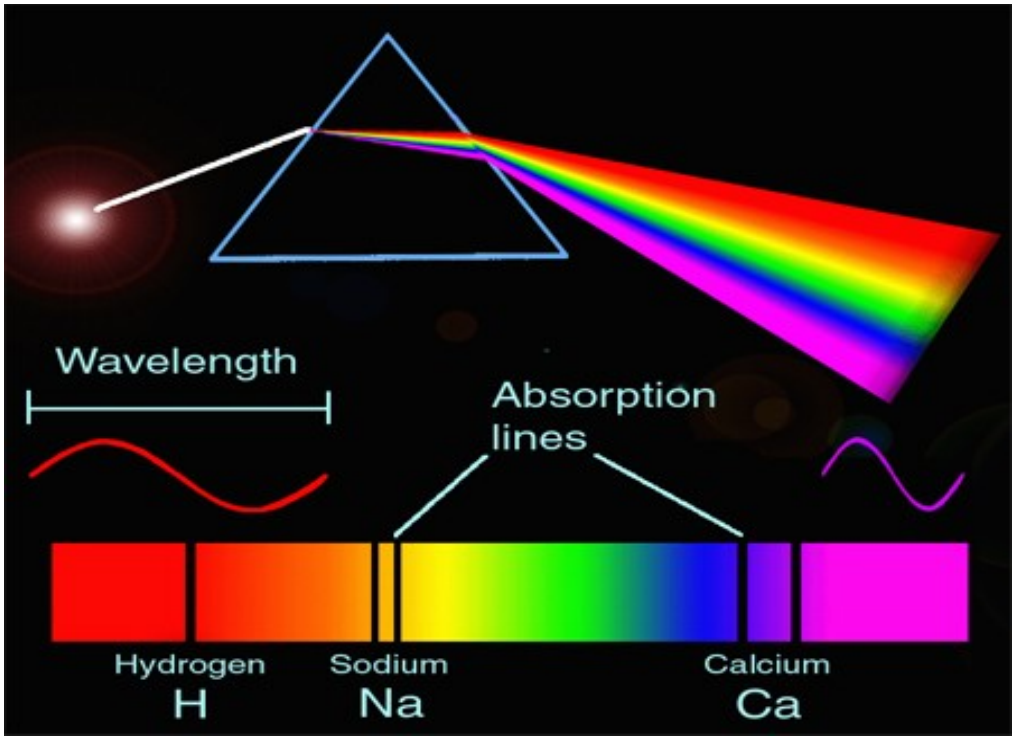
Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas

FI-3104-1 Métodos Numéricos para la Ciencia e Ingeniería

Tarea n°10
"Modelando el espectro de una fuente como función de la longitud de onda"

Benjamín Venegas Ríos
1-12-2015
Santiago,Chile



Resumen

La espectroscopia es el estudio de la interacción entre la radiación electromagnética y la materia, con absorción o emisión de energía radiante. En los últimos años, en la astronomía, el análisis de las composiciones estelares, las temperaturas con la cual emiten radiación las estrellas,,entre otros, ha alcanzado un gran nivel de desarrollo. Sin embargo, a medida que los observatorios a lo largo de todo el mundo han ido recopilando millones de datos astronómicos, ha surgido la necesidad de trabajar con software altamente eficientes y además, de poder modelar los distintos fenómenos físicos que se estudian a partir de la muestras experimentales.

En el presente trabajo, se analizará el espectro de un cuerpo radiante en función de su longitud de onda, cuya forma gráfica indica que existe un continuo para el flujo y luego, unas líneas de absorción. Este espectro permitirá examinar y comprender de mejor forma el comportamineto físico de la fuente radiante. Para eso, es que se buscará modelar la parte continua y la línea de absorción del espectro a partir del archivo espectro.dat mediante 2 modelos distintos pero con la misma cantidad de parámetros. Por un lado, estará el modelo gaussiano y por otra parte, el modelo de Lorentz. Y, finalmente, lo que se buscará será determinar que modelo se aproxima de mejor forma a los datos, (verificación de la hipótesis nula, es decir, que los datos sean consistentes con el modelo) graficándolos cada uno junto a los datos y calculando el nivel de confianza para cada modelo.

Pregunta1

1.1 Introducción

En esta pregunta se pedía modelar el espectro utilizando el modelo gaussiano que consistía en anexar una línea recta menos una distribución normal,con 5 parámetros (a,b,A,mu,sigma) , y de forma similar, con el modelo de Lorentz, utilizando una línea recta menos la distribución de Cauchy con 5 parámetros a determinar (a,b,A,mu,sigma). Es importante mencionar que las dos distribuciones se incluyeron al programa utilizando import scipy.stats.

1.2 Procedimiento

Para realizar esta parte, en primer lugar se definieron 2 funciones, cada una con el modelo a ajustar, es decir, gaussiano y de Lorentz. Posteriormente, se abrió el archivo contenedor de los datos con np.loadtext('name') y se guardaron dos arreglos: El primero fue el arreglo de flujo (en erg/s/cm2/Hz)y el segundo fue el arreglo de longitudes de onda (en Angstrom).

Luego de haber hecho esto, se utilizó la función implementada en librería de scipy, curve_fit, para cada modelo, tomando como parámetros iniciales los siguientes valores:

***a = 3 [erg/s/Hz/cm^2]** --> Por simplicidad se ocupó este valor (también 1 o 0 podían ser). Lo importante era que ambos modelos convergieran a lo observado por los datos.*

***b = (0.04e-16)/200 [erg/s/Hz/cm^2] / [Angstrom]** --> se consideró la pendiente de los datos según dy/dx (donde y=flujo y x=londa)*

***A = 0.1e-16 [erg/s/Hz/cm^2] [Angstrom]** --> Mirando la campana del gráfico de los datos experimentales. Se toma un punto aleatorio basal y otro en la punta de la campana, y la diferencia correspondía a la amplitud (figura 1)*

***Mu (centro) = 6550 [Angstrom]**---> Dado que es el centro, se consideró la parte normal del gráfico y se aproximó al valor mencionado (figura 1) .También pudo haber sido 6563 ,por ejemplo. Lo importante era que el modelo se ajustara bien a los datos.*

***Sigma = 10 [Angstrom]** –-> Viendo la figura 1, la parte normal,la distancia entre el centro de la campana y el punto basal, según la abscisa. (También pudo haber sido 7, por ejemplo)*

Dados estos valores en una primera estimación de los parámetros, al ejecutar la función curve_fit a cada modelo se obtuvieron los valores más probables y de mejor ajuste para cada uno. (Ver Tabla 1 anexada)

Vale la pena recalcar que en el trabajo se pidió no demostrar mediante intervalos

de confianza que efectivamente correspondían a los valores de los parámetros más probables.

Dicho esto, además, se ocupó el test de chi cuadrado para tener una idea de que modelo se ajustaba mejor. Para esto, se creó una función muy sencilla que compararía los valores de los flujos experimentales con los valores obtenidos por cada modelo al cuadrado y sumados. De esta forma, se obtuvieron los valores para chi cuadrado de cada modelo. (Ver Tabla 2 anexada)

Finalmente, se graficaron 3 figuras (ver figuras 1 , 2 , 3) , una figura contenía el modelo gaussiano y los datos experimantales, la otra el modelo de Lorentz más los datos , y por último, una figura con los 2 modelos y los datos para comparar cual de los 2 tenía menos diferencia con el plot de los datos.

1.3 Resultados

A continuación se muestran los parámetros obtenidos y los valores de los chi cuadrado para cada modelo. Además, se agregan 3 figuras que muestran el ajuste a los datos generado por los 2 modelos.

Modelo	*a*	*b*	*A*	*Mu*	*Sigma*
gauss	*8.88e-17*	*7.80e-21*	*8.22e-17*	*6563.22*	*3.26*
Lorentz	*8.81e-17*	*7.92e-21*	*1.11e-16*	*6563.20*	*3.22*

***Tabla 1.** Parámetros obtenidos para cada modelo ocupando curve_fit.*

Donde las unidades para cada parámetro eran:

a = [erg/s/Hz/cm^2]
b = [erg/s/Hz/cm^2] / [Angstrom]
A = [erg/s/Hz/cm^2] [Angstrom]
Mu = [Angstrom]
Sigma = [Angstrom]

Modelo	*Chi cuadrado*
gauss	*5.20e-35*
Lorentz	*5.01e-35*

Tabla 2. *Test de chi cuadrado obtenido para cada modelo.*

Las unidades para Chi cuadrado eran:

chi2= [erg2/s2/Hz2/cm^4]

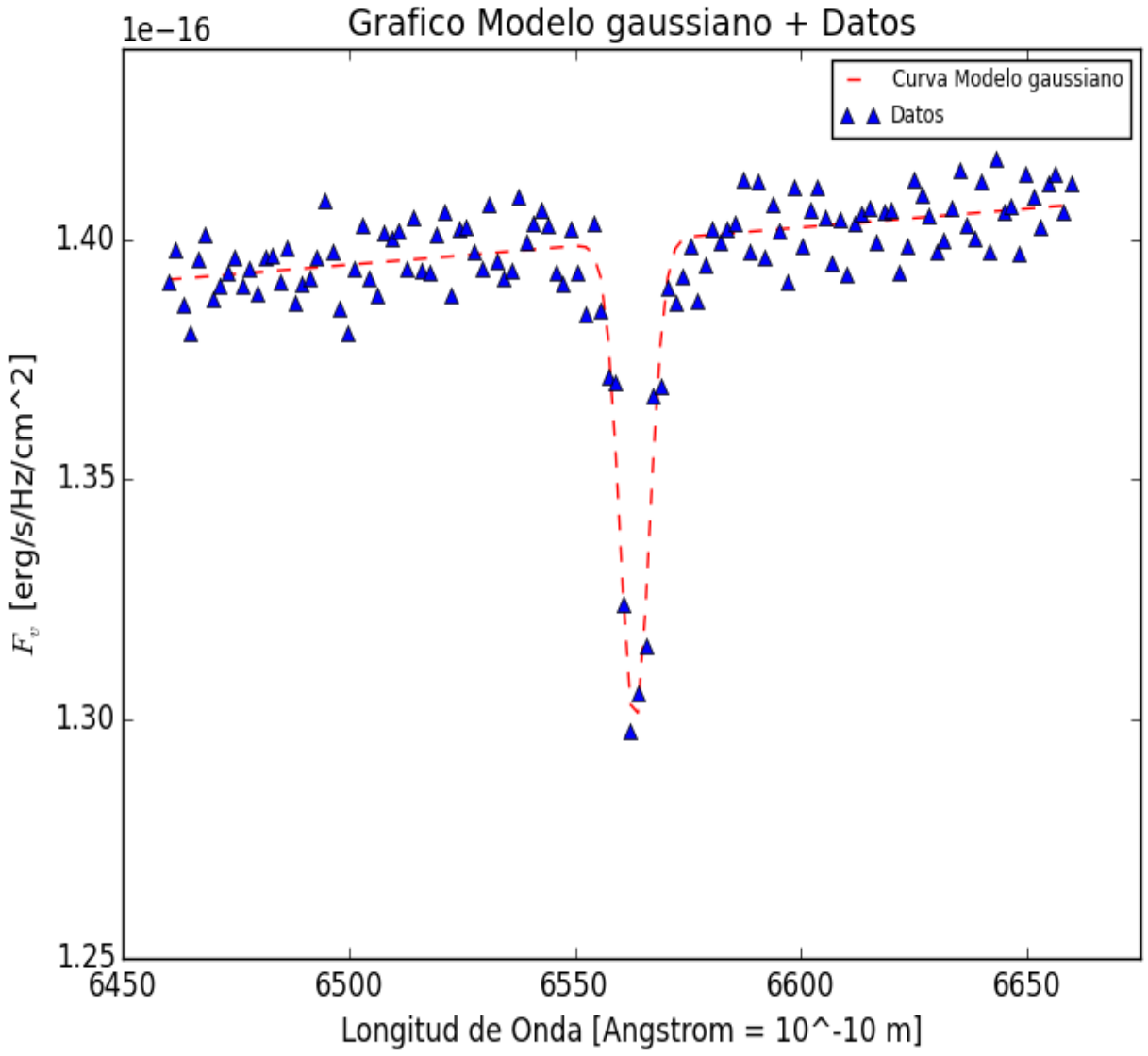


Figura1. *Gráfico del espectro con Flujo v/s Longitud de onda, dado el modelo gaussiano y los datos del archivo espectro.dat.*

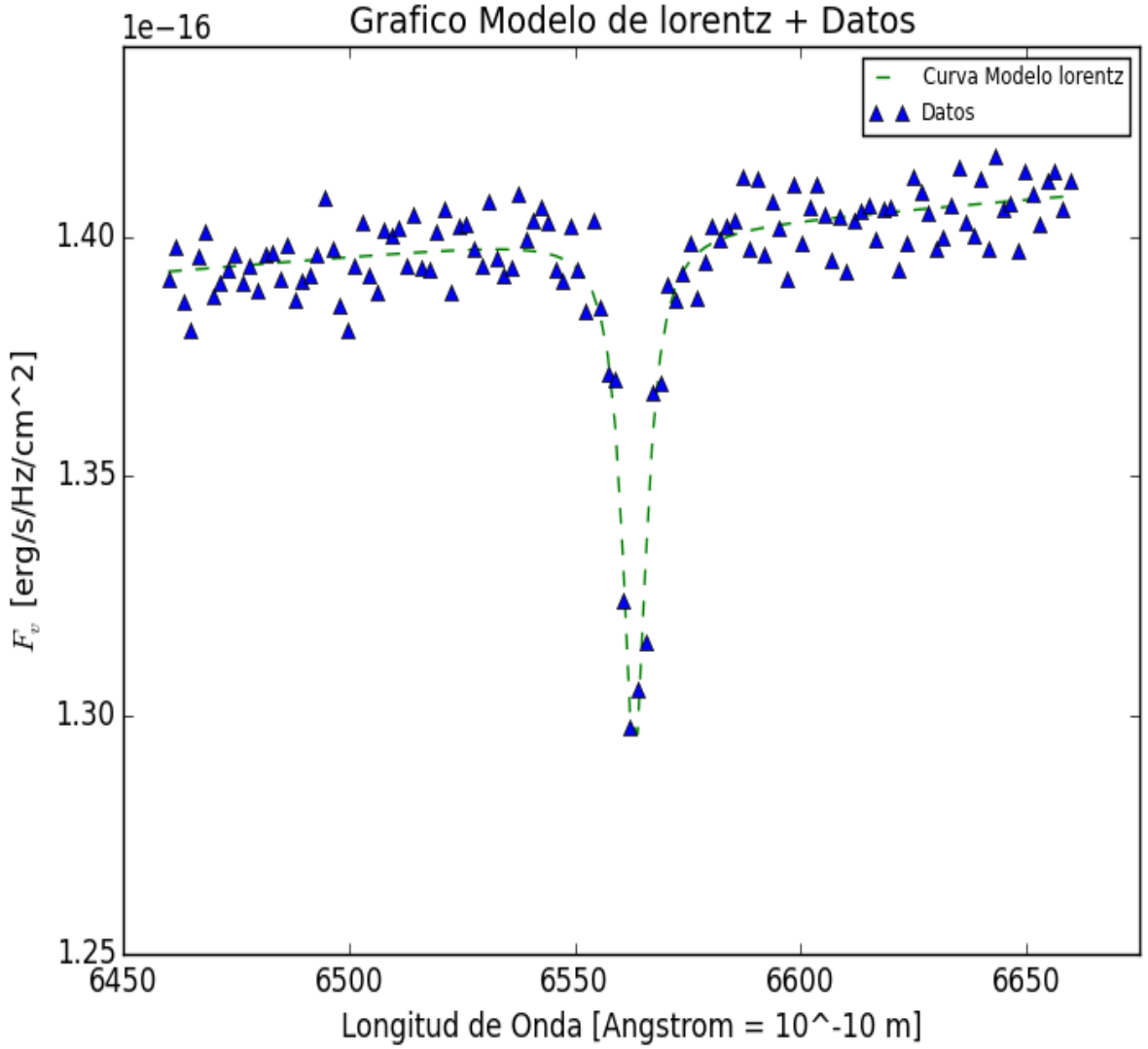


Figura2. *Gráfico del espectro con Flujo v/s Longitud de onda, dado el modelo de Lorentz y los datos del archivo espectro.dat.*

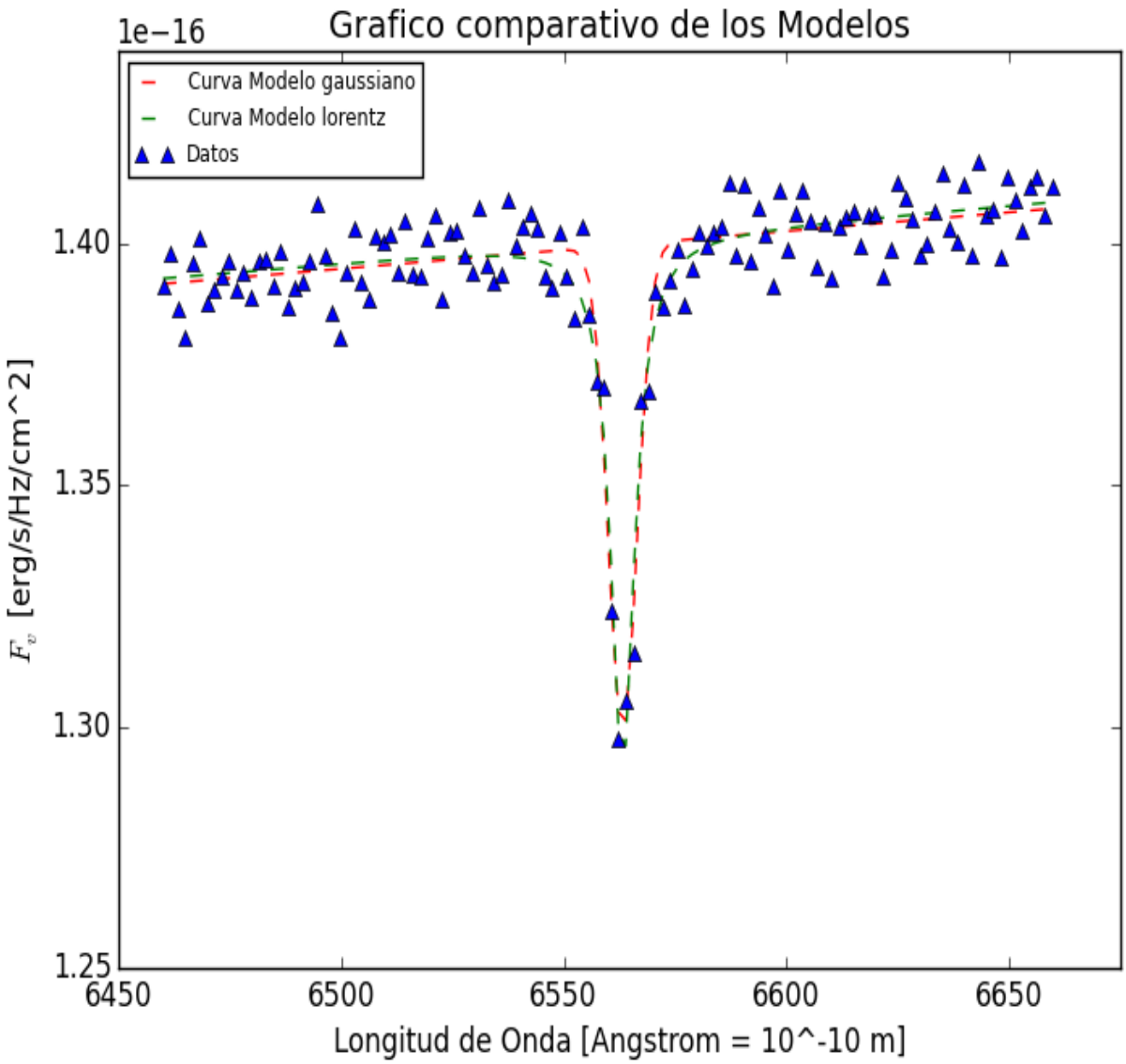


Figura3. *Gráfico comparativo del espectro con Flujo v/s Longitud de onda, con ambos modelos y los datos del archivo espectro.dat.*

Pregunta 2

2.1 Introducción

Los modelos obtenidos en la parte anterior parecen ser una buena aproximación a los datos del archivo espectro.dat. Pero, para comprobar esto, y dado que chi cuadrado no es determinante a la hora de decir que ajuste es mejor que el otro, es que se usará el test de Kolmogórov-Smirnov y así , aparte de determinar el nivel de confianza de cada modelo,es decir, la probabilidad de que los datos sean consistentes con la hipótesis nula(H0), se calculará la direncia máxima entre las funciones de distribucion acumulada del modelo con la de los datos y se comparará con la diferencia crítica dado una tolerancia de error de 0.005, y así se decidirá si es que se rechaza la hipótesis nula o no.

2.2 Procedimiento

Se inicia el test de K-S en primer lugar, creando un arreglo de longitudes de onda de igual tamaño a la muestra (122). Luego, se calculan el mínimo y el máximo para los datos de las longitudes de onda, para así poder calcular , posteriormente, un arreglo ordenado de los flujos para cada modelo ocupando los parámetros calculados en 1 para los dos modelos y el nuevo arreglo.

Así, se procedió a determinar la función de distribución acumulada del modelo, simplemente, enmascarando (los verdaderos son 1 y los falsos son 0) y sumando, para la condición de que si el vector flujo del modelo (ordenado con sort) es menor al valor del flujo de los datos(ordenado), entonces cada vez que haya un True se suma 1, sino solo 0 . De esta forma se creó un arreglo de 122 valores para la distribucion acumulada de cada modelo (el proceso de enmascarar se realizó para todos los flujos de los datos)

Ahora, si bien se determinó la función de distribución acumulada de cada modelo, faltaba restarle por los dos extremos(superior e inferior) la función de distribución acumulada observada y,así,poder determinar el valor de Dn, que no era más que el máximo entre los 2 máximos calculados por cada modelo (se construyeron 2 máximos, debido a que se señaló que existen 2 extremos y cada extremo se comparó con la función de distribución acumulada de un modelo) .

Este desarrollo, si bien fue un poco extenso, también se condensó en una 2 forma, la cual, se anexó al código. Esta forma, requirió de importar desde scipy.stats el método kstest , que era precisamente el test de K-S, y además, el método de kstwobign, de similar utilidad. De esta forma, se calcularon los Dn nuevamente (método de kstest) ,y a su vez, a modo de entender si la hipótesis nula Ho se cumplía,es decir, si es que los datos eran consistentes con el modelo propuesto, se determinó usando el método de kstwobign, dada una tolerancia al error de 0.005 , el Dn [crítico]. Entonces, si Dn>Dn [crítico] se rechazaba Ho, mientras que si Dn <Dn [crítico] se aceptaba Ho.

También, se calculó el nivel de confianza de cada modelo (probabilidad de que los datos sean consistentes con modelo) ocupando kstest y kstwobign (donde N=122)

Para cerrar, se generaron 2 gráficos(figuras 4 y 5) , uno para cada modelo, que mostraban la relación entre las funciones de distribución acumulada observada con la del modelo.

2.3 Resultados

A continuación se muestran 2 tablas con los valores de Dn y de niveles de confianza para cada modelo. Además se agregan las figuras 4, 5 y 6 ya mencionadas anteriormente.

Modelo	*Dn (kstest)*	*Dn (versión extensa)*
gauss	*0.17*	*0.17*
Lorentz	*0.16*	*0.16*

Tabla 3. *Test K-S para cada modelo calculando los respectivos Dn.*

Modelo	*Nivel de confianza (kstest)*	*Nivel de confianza(kstwobign)*
gauss	*0.00125*	*0.00145*
Lorentz	*0.00247*	*0.00284*

Tabla 4. *Nivel de confianza para cada modelo.*

A su vez el Dn crítico para alpha=0.005 según scipy(método kstwobign) arrojó el valor:

Dn [crítico] = 0.12

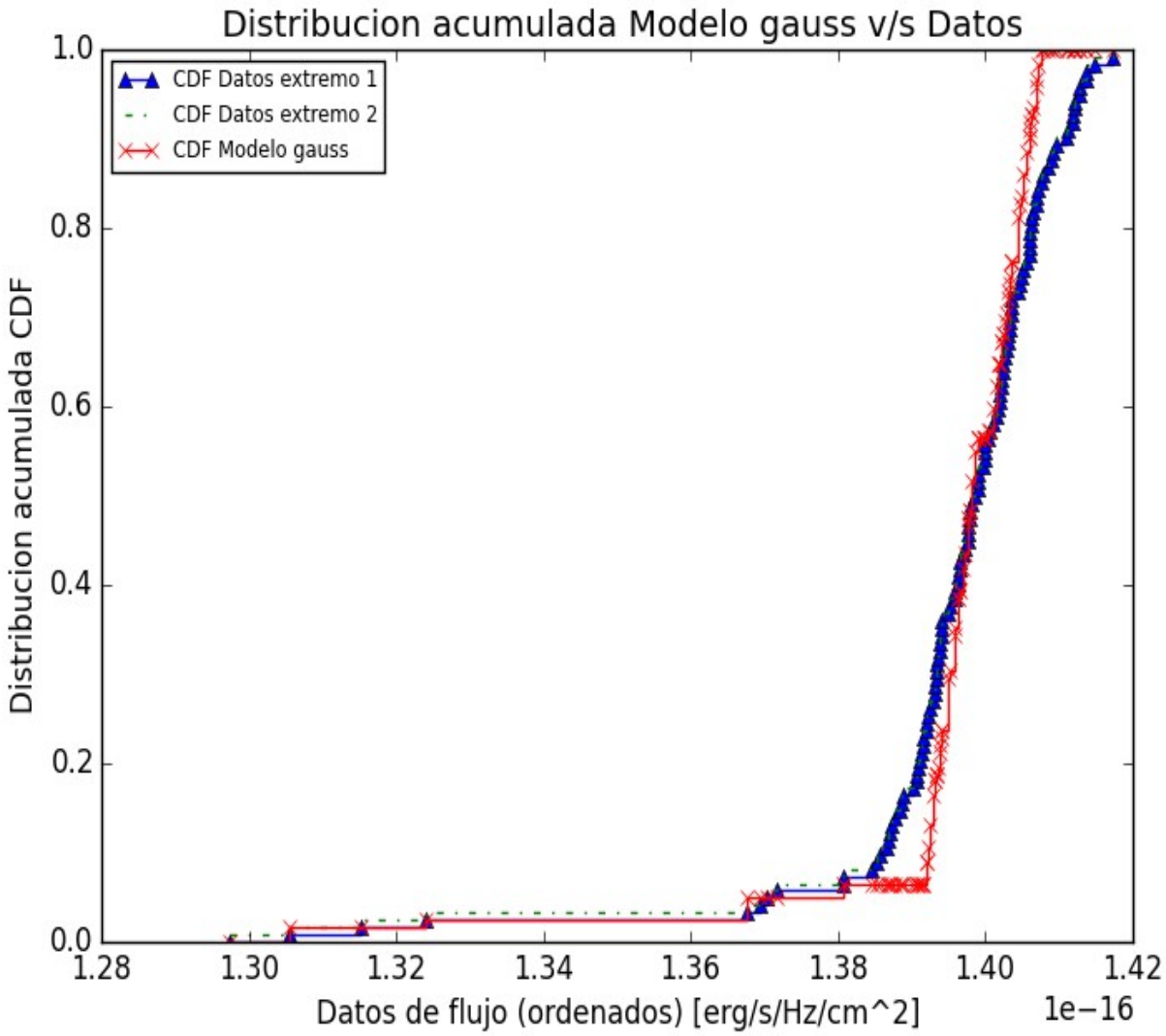


Figura4. *Gráfico de la distribucion acumulada del modelo de gauss v/s la distribución acumulada por extremos de los datos.*

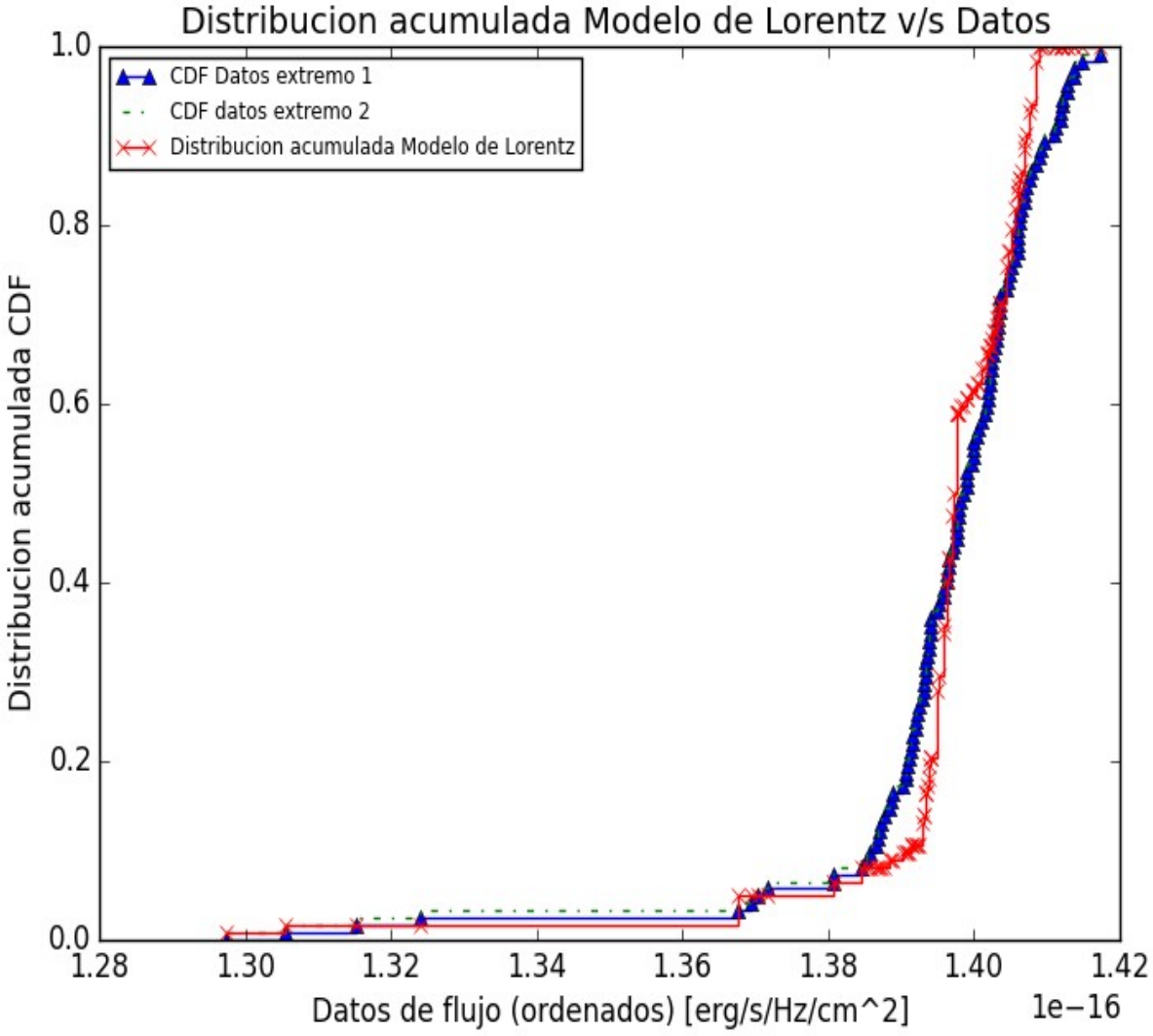


Figura5. *Gráfico de la distribucion acumulada del modelo de Lorentz v/s la distribución acumulada por extremos de los datos.*

Conclusión

Como conclusión se pudo ver que si bien ambos modelos tenían un valor de chi cuadrado relativamente bajo (del orden de los 10^ -35, ver tabla 3) , con una leve inclinación hacia el modelo de Lorentz por haber tenido un valor más bajo, lo que indicaba que los valores asociados al modelo de Lorentz se ajustaban de mejor manera que el modelo gaussiano; esta diferencia no era considerable a la hora de afirmar con 100% de seguridad que modelo era el que mejor se ajustaba a los datos. De hecho, tal como lo indicaba el enunciado, el error asociado a la medición a la medición en cada pixel (correspondiente a cada longitud de onda) no era gaussiano sino que poissoniano (ambas son similares para N grande, pero no iguales). Por ende, el test de chi- cuadrado no era tan significativo a la hora de analizar que modelo era el más adecuado.

De este modo, al haber obtenido en la parte 2 los valores para Dn y Dn[crítico], se logró tener un noción más clara de lo sucedido. Ambos Dn, para los 2 modelos y fuera el que fuera el método ocupado para calcularlo(kstest o kstwobign), resultaron mayores que el Dn [crítico] (0,17>Dn[crítico]=0.12 para modelo gaussiano y 0.16>0.12 para modelo de Lorentz) asociado a una tolerancia de 0.005 (alpha). Esto arrojó evidencias de que los datos no eran consistentes con ninguno de los 2 modelos dado el a0 como parámetro inicial (se rechazaba la hipótesis inicial Ho de consistencia datos-modelo). Y es más, si se añadía la condición de que los niveles de confianza en promedio (dadas leves disimilitudes mostradas en la tabla) para el modelo gaussiano era de 0.135%, mientras que para el modelo de Lorentz un 0.266% (casi el doble que el nivel de confianza para el modelo gaussiano), es evidente de que estos modelos no son una representación exacta de los datos y, aún más, la carencia de confiabilidad se produjo fundamentalmente por la parte continua de lo observado, porque una línea recta en ambos modelos, termina por "suprimir" de cierto modo el grado de dispersión de los datos ,y he aquí donde los 2 modelos terminan por rechazar Ho. De hecho, es muy probable esto, ya que tal como se pudo apreciar en la figura 3, en la parte de las líneas de absorción casi no hay discrepancia entre los modelos y los datos (parte normal de los modelos con los datos). Y lamentablemente, los parámetros iniciales a0 que se dan no logran completar el ruido o dispersión de los datos, así que es difícil modelar ese sector sin lugar a dudas.
Por otra parte, al comparar las distribuciones acumuladas de lo observado con los modelos(figuras 4 y 5) , se pudo apreciar que se asemejan bastante (lo rojo)

en ambos gráficos,sobre todo al inicio entre (1.3 -1.38 [erg/s/Hz,cm2]), no obstante, difieren en ciertos dominios más a la derecha .Esto indica que el grado de dispersión, nuevamente, estaba afectando al ajuste.

Para finalizar, aunque los modelos no se ajusten exactamente igual a la forma de los datos, sobre todo por la difusa forma inicial de estos, el modelo de Lorentz se acerca levemente más (en un doble porcentaje de probabilidad de confianza) a lo observado. Si bien lo importante era que los datos siguieran alguna distribución específica , lo cual en su totalidad no se pudo cumplir, se puede decir que si se acotara una parte del dominio a las líneas de absorción ambos modelos serían una muy buena aproximación y el fenómeno físico, en este caso, espectral, se podría comprender de mejor forma.