Identificación de Clumps Astronómicos: Un Enfoque Desde el Cálculo Variacional

Martín Villanueva

4 de abril de 2016

1. Introducción

Un cubo espectroscópico de datos puede ser modelado por una función $f(x,y,z)$, donde las variables (x,y) codifican las coordenadas espaciales (*right ascension* y *declination*) respectivamente, y la coordenada z corresponde a la velocidad radial. Esta función f asocia una de estas coordenadas, un valor correspondiente a la intensidad de la emisión captada en tal punto. Sin embargo la función f se conoce en un conjunto discreto de puntos (valores en el cubo de datos).

El problema entonces reside en encontrar una función $g(x,y,z)$ continua, que aproxime la función f, de tal manera que estudiando la forma y propiedades de g, sea posible identificar clumps o regiones de interés en dichas observaciones astronómicas.

2. El modelo Variacional

Sea entonces $f(x,y,z)$ la función de data, y $g(x,y,z)$ la función que pretende aproximar a f, se propone entonces el siguiente funcional:

$$\begin{aligned}\Phi(g) &= \int_{\Omega\subset R^3} L(x,y,z,g,g_x,g_y,g_z)\,d\Omega \\ &= \int_{\Omega\subset R^3} (f-g)^2 + \underbrace{\alpha e^{-(f-g)}}_{\text{Penalty term}} + \underbrace{\beta\,\Psi(g_x,g_y,g_z)}_{\text{Smooth term}}\,d\Omega\end{aligned} \tag{1}$$

a ser minimizado. La inclusión del *Penalty term* y *Smooth term* en el funcional tiene los siguientes motivos:

1. *Penalty term*: Una de los requisitos sobre la función g es que esta nunca *sobrepase* a f. Luego agregando este término se genera una alta penalización sobre el funcional cuando $g(x,y,z) > f(x,y,z)$. Esto a su vez puede ser controlado por el parámetro α.

2. *Smooth term*: Siendo Ψ una función continua, monótona y creciente, el objetivo de este término es agregar *smoothness* a la solución g, permitiendo de esta manera filtrar el ruido o distorsiones presentes en la data. El parámetro β permite controlar el peso de este término, y por lo tanto debe estar directamente relacionado con el SNR (*Signal to Noise Ratio*).

Ocupando entonces los resultados del cálculo variacional, sabemos que el funcional $\Phi(g)$ es minimizado cuando se satisface la siguiente ecuación de Euler-Lagrange:

$$\frac{\partial L}{\partial g} - \frac{d}{dx}\frac{\partial L}{\partial g_x} - \frac{d}{dy}\frac{\partial L}{\partial g_y} - \frac{d}{dz}\frac{\partial L}{\partial g_z} = 0 \quad \forall (x,y,z)\in\Omega \tag{2}$$

3. Resolución del Problema

Como primer acercamiento a la solución de (2), se propone el método de colocación. Sea $\mathcal{N}(\mathbf{x}, \mathbf{x_0}, \sigma)$ una función Gaussiana unitaria, centrada en x_0 con desviación estándar σ, entonces la función g propuesta es como sigue:

$$g(\mathbf{x}) = \sum_{k=1}^{N} \gamma_k \mathcal{N}(\mathbf{x}, \mathbf{x}_k, \sigma_k) \tag{3}$$

que corresponde a una combinación lineal de tales Gaussianas. Cada uno de los componentes de la combinación lineal tiene dos parámetros a determinar (γ_k, σ_k).

Observación: Eventualmente pueden ser usadas Gaussianas elípticas, asumiendo que el número de parámetros a determinar será mayor: Rotaciones de los ejes, y desviaciones estándar a lo largo de cada semieje.

El problema se resume entonces en encontrar la configuración de (γ_k, σ_k), de modo que $\Phi(g)$ sea minimizado. Reemplazando (3) en (2) entrega (en general) una ecuación no lineal en $\{(\gamma_k, \sigma_k)\}_{k=1}^{N}$. Como estas son $2N$ incógnitas, será necesario evaluar (2) en la misma cantidad de puntos en su dominio Ω, generando un sistema no-lineal de $N \times N$.

4. Desafíos

Se listan a continuación un serie de problemáticas en cuanto al modelo y su implementación

1. Penalizar correctamente los casos donde $g > f$.
2. Establecer una relación directa entre el parámetro β y el SNR de la data.
3. Establecer la cantidad N de Gaussianas a ocupar en la construcción de g.
4. Seleccionar los puntos de colocación x_k (Uniforme, Random, Quasi-Random, Ocupando alguna heurística, etc).
5. Seleccionar los puntos objetivos donde evaluar (2).
6. Usualmente para que el método de colocación sea exitoso, requiere valores de N muy grandes. Esto implíca evaluaciones costosas para $g,\ g_x,\ g_y,\ g_z$ las cuales son expresiones que deben computarse en cada evaluación de (3) en algún punto del dominio. Una posible solución a este problema es ocupar FGT (Fast Gauss Transform), pero hay que considerar que las funciones $g_x,\ g_y,\ g_z$ no son precisamente sumas de Gaussianas, por lo cual modificaciones deben hacerse sobre FGT tradicional.
7. Debe hallarse una manera eficiente de resolver el sistema no lineal de $N \times N$ resultante. Posibles opciones son ocupar el método de Newton (con Jacobiano explícito), o las soluciones estables de un sistema dinámico.