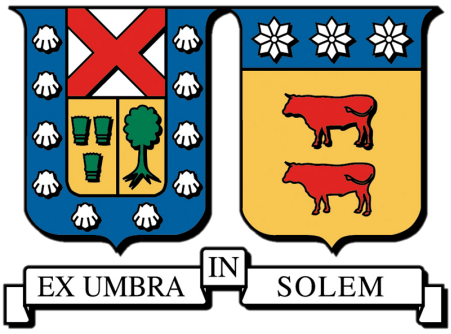

Universidad Técnica Federico Santa María
Departamento de Informática

Profesor Guía: Claudio Torres
Profesor Correferente: Mauricio Araya

IDENTIFICACIÓN DE CLUMPS ASTRONÓMICOS EN NUBES MOLECULARES: UN ENFOQUE DESDE EL CÁLCULO VARIACIONAL

ESTADO DEL ARTE: SEMINARIO DE MEMORIA

Martín Villanueva A.
mavillan@alumnos.inf.utfsm.cl

24 de Junio 2016

1. INTRODUCCIÓN

Uno de los problemas principales existentes hoy en astronomía, es el desarrollo de mecanismo automatizados para la correcta identificación de clumps, y el análisis estructural de estos en las nubes moleculares.

Cada observación de tales nubes puede contener una cantidad considerable *núcleos*, *cluster* y *clumps*. Hace no mucho tiempo, la tarea de identificación de tales estructuras era realizada manualmente por astrónomos, pero esto ya no es posible debido a los siguientes razones:

1. Tales métodos manuales no escalan tan rápido como la generación de nuevos datos, tales como los generados por ALMA fase 2.

2. El tamaño de las imágenes es enorme y puede contener cientos/miles de estructuras de interés. Por lo que puede requerir una cantidad muy alta de tiempo hacerlo manualmente.

3. Los cubos espectroscópicos de datos son difíciles de manipular y visualizar al día de hoy. Adicionalmente la resolución en frecuencia de tales cubos es cada vez más fina.

4. Los resultados de identifación de clumps realizados por astrónomos (manualmente), usualmente no calzan pues entre ellos tienen diferentes concepciones y definiciones acerca de que es un clump y la cómo se relacionan entre ellos (especialmente en los casos de emisiones conjuntas).

La motivación de los métodos computacionales, es poder automatizar estas tareas eliminado el juicio imparcial del astrónomo, y al mismo tiempo poder analizar eficientemente gran cantidad de datos.

En lo que sigue se realiza una descripción formal del problema, definiendo en primer lugar que se entiende por *clump*. Luego se describen las técnicas computacionales ocupadas hoy en día para resolver la problemática de identificación de tales estructuras. Por último se da una breve descripción del enfoque variacional, y como este ha sido aplicado a otras áreas similares con muy buenos resultados.

2. DESCRIPCIÓN FORMAL DEL PROBLEMA

2.1. NUBES MOLECULARES, NÚCLEOS Y CLUMPS

Siguiendo las definiciones dadas por Kennicutt el at. [1], el término **nube molecular** se refiere a una estructura en el medio interestelar (ISM *interstellar medium*) separada de sus alrededores por el rápido cambio de alguna propiedad, tales como la presión, densidad o estados químicos.

Las nubes tienen una estructura compleja, pero los astrónomos teóricos identifican dos estructuras principales: **Clumps**, entendidos como zonas de alta emisión dentro de un *cluster*; **Núcleos**, entendidos como zonas de alta emisión dentro de estrellas individuales o binarias. Tales nubes moleculares están básicamente compuestas de gas y polvo, siendo los gases más abundantes el hidrógeno molecular H_2, CO, H_2O y otras moléculas más complejas.

El gas y polvo interior de tales nubes se distribuye en una componente difusa, al interior de la cual hay filamentos de gas más denso en diversas direcciones, con composición química y propiedades físicas definidas. Dentro de tales filamentos se pueden encontrar los *Núcleos* que se caracterizan por ser poco masivos pero muy densos, y los *Clumps* que se caracterizan por ser regiones de mayor densidad, y están acotados por alguna propiedad.

En las secciones siguientes se explican los principales algoritmos utilizados hoy en día, para la identificación de estructuras del tipo *clump*.

2.2. GAUSSCLUMPS

Este método propuesto por Stutzki et al. [3] y fue uno de los primeros en abordar de forma satisfactoria el problema de la identificación automática de clumps en cubos espectroscópicos. Como se explica y analiza

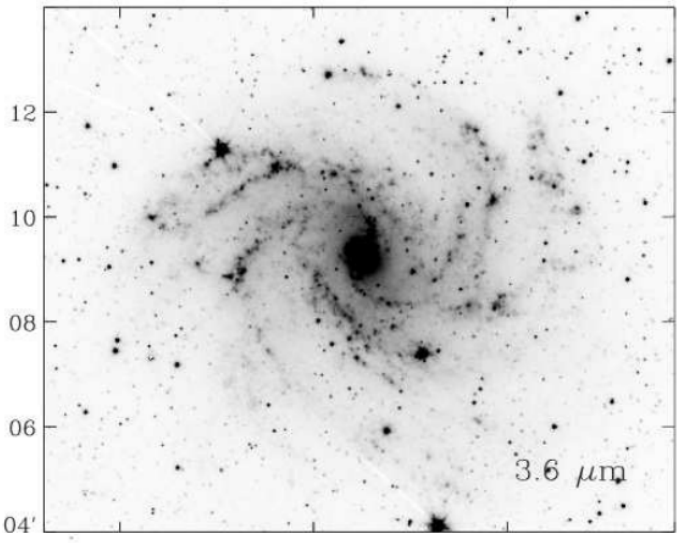

Figura 1: Montaje de imágenes infrarrojas de NGC6946 desde *Spitzer* y *Herschel* (Fuente: [2])

también en Kramer et al. [4], la idea principal de este método es ajustar perfiles Gaussianos que mejor ajusten cada uno de los peaks de emisión presentes en los datos. La razón de ocupar funciones Gaussianas sobre alguna otra opción, es debido a que observaciones de emisiones de algunas de las moléculas más frecuentes en nubes moleculares, han mostrado que su distribución de intensidad tiene forma de campana de Gauss.

El algoritmo iterativamente busca el máximo *peak* de emisión actual, y determina la función gaussiana que mejor se ajusta a tal emisión y la región que lo rodea. Tal procedimiento consiste en determinar la intensidad del peak a_0, el centro $\mu = (\mu_x, \mu_y, \mu_z)$, los anchos por cada eje $(\sigma_x^2, \sigma_y^2, \sigma_z^2)$ y las orientaciones de cada eje $(\theta_x, \theta_y, \theta_z)$ de la Gaussiana, por medio de un proceso de minimización del error cuadrático. La función ajustada es agregada a un catálogo de salida y es restada del cubo de datos original, generando un cubo residual que será la entrada de la próxima iteración. Cómo resultado se obtienen un conjunto de Gaussianas con diferentes forma, posiciones, intensidades y orientaciones, que sumadas al ruido de fondo (residual final del proceso iterativo) permiten reconstruir (aproximadamente) el cubo espectroscópico original.

La función χ^2 (notación utilizada en [3]) a minimizar en cada iteración tiene varios componentes. En su forma más general puede ser escrita como:

$$\chi^2 = \sum \omega_i (Y_i - Y_i^{\text{fit}})^2 + s_0 \sum \omega_i \mathscr{F}(Y_i^{\text{fit}} - Y_i) + s_a (a_0 + b_0 - Y^{\max})^2 + s_c ||\boldsymbol{\mu} - \boldsymbol{\mu}^{\max}||_2, \tag{1}$$

en donde cada uno de los términos tiene el siguiente significado (los parámetros (s_0, s_a, s_c) permiten definir los pesos relativos de cada término dentro de χ^2):

1. $\sum \omega_i (Y_i - Y_i^{\text{fit}})^2$ es el término estándar de error cuadrático, donde Y_i corresponden a las observaciones de intensidad, y Y_i^{fit} a la función a ajustar, e ω_i son los pesos dados a cada observación.

2. $\sum \omega_i \mathscr{F}(Y_i^{\text{fit}} - Y_i)$ penaliza los puntos en donde la función a ajustar sobrepasa a los valores observados, forzando de este modo a las Gaussianas a permanecer por debajo de los valores de la data. La función $\mathscr{F}$ es elegida como $\mathscr{F}(x) = \exp(x)$ en [3], mientras que en [4] se usa $\mathscr{F}(x) = (0 \text{ para } x \leq 0;\ x^2 \text{ para } x > 0)$.

3. En $(a_0 + b_0 - Y^{\max})^2$, a_0 es el valor peak de la Gaussiana y b_0 es el nivel de base (definido como un múltiplo del RMS de la data). El objetivo de este es mantener el *peak de amplitud* de la gaussiana ajustada cercano el peak observado $Y^{\max}$.

4. $||\boldsymbol{\mu} - \boldsymbol{\mu}^{\max}||_2$ es para forzar que el centro del clump observado y el centro la Gaussiana ajustada esten cercanos.

En este algoritmo cada Gaussiana sustraída es considerada un clump de emisión individual, y debido a que entre las Gaussianas puede existir traslape, entonces es posible que un mismo píxel sea asignado a más de un clump. Esta característica distingue a este algoritmo de los otros métodos de identificación de clumps que siguen a continuación.

2.3. CLUMPFIND

Un enfoque totalmente distinto fue propuesto por Williams el al. [2]. Este está inspirado en cómo los humanos detectan estructuras dentro de una imagen bidimensional; El ojo humano humano descompone mapas bidimensionales en estructuras isoladas determinando contornos de nivel, siguiendo de manera gradual desde los contornos de más alto nivel hacia lo más bajos, en donde se mezclan las distintas estructuras.

Del mismo modo *ClumpFind* funciona contorneado la data en múltiplos del RMS del ruido de la observación, esto es, genera una serie de intervalos de intensidades (cada uno de ancho igual a un múltiplo del RMS anterior) en lo cuales se buscan estructuras isoladas (clumps). Este trabaja iterativamente iniciando en el nivel más alto (aquel que contiene al peak) y bajando de forma secuencial hacia los niveles más bajos, buscando en cada nivel las estructuras isoladas, es decir, los conjuntos de píxeles que no están conectados entre sí. Luego si la estructura/clump encontrado a un nivel dado, está conectado con un clump identificado en el nivel superior, entonces ambos corresponden a un mismo clump y se les asigna un mismo identificador. En caso contrario (el clump identificado en este nivel no está conectado con ninguno en el nivel superior) se define este como un nuevo clump encontrado.

Una situación especial se dá cuando un clump encontrado en un nivel está conectado con más de un clump en el nivel superior. Este es el caso de emisiones combinadas o conjuntas de dos clumps cercanos. En tales casos, tal estructura es separada asignando los píxeles que contiene, a cada uno de los clumps de los niveles superiores según un criterio, siendo el más común el conocido *friends of friends*, que asigna un píxel según la cantidad de pixeles vecinos que pertenecen a cada clump (*al con más amigos*).

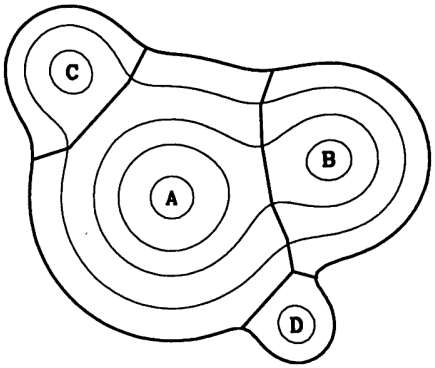

Figura 2: Resultados de ClumpFind en data 2D (Fuente: [2])

Como resultado este algoritmo produce el *Clump Assignment Array* o CAA, que es un arreglo de las mismas dimensiones de los datos de entrada, en el cual cada pixel tiene asignado el identificador del clump al que pertenece, como se muestra en la Figura 2. Una característica importante que distingue ClumpFind del resto de algoritmos, es que no realiza supuestos a priori acerca del perfil de los clumps, y por lo tanto los clumps encontrados pueden tener cualquier forma.

2.4. FELLWALKER

Un algoritmo más reciente corresponde al desarrollado por Berry et al. [5]. Este está basado en el algoritmo de optimización *Hill-Climbing*, aprovechando la propiedad de estancamiento en óptimos locales, para

determinar los peaks de emisión locales y de este modo definir un nuevo clump. Al igual que ClumpFind, este método segmenta el cubo de datos en regiones disjuntas, cada una asociada un *peak* de emisión *significante* e individual.

El procedimiento completo corresponde a un algoritmo de múltiples etapas, las cuales se describen a continuación:

1. (*Thresholding*) En primer lugar se define un nivel base sobre el cual se encuentran las intensidades que son de interés, que permite separar en cierta medida el ruido de la señal. Tal nivel se determina como un múltiplo del RMS de la data. Los pixeles con intensidades inferiores a esta, se identifican como invalidos.

2. (*Removing*) Haciendo uso de un automata celular, se remueven las regiones pequeñas de píxeles isolados (marcandolos como inválidos).

3. (*Walking up*) Se itera sobre todos los pixeles aun no asignado (y válidos) del cubo, computando para cada uno de ellos rutas de ascenso en la dirección de máximo gradiente. Tales rutas tienen dos alternativas (Ver Figura 3): 1) Alcanzar un pixel ya asignado a un clump, y por lo tanto asignar todos los píxeles de la ruta de ascenso actual a tal clump. 2) Alcanzar un *peak* local, encontrando entonces un nuevo clump (En este caso se vuelve a realizar una búsqueda en un vecindad más extensa, para verificar que este es realmente un *peak* local, y no efecto del ruido).

4. (*Merging*) Se analizan luego los clumps encontrados anteriormente. Si la diferencia de altura entre el *peak* de un clump, y la menor intensidad de un pixel de borde con un clump vecino es baja (parámetro definido por el usuario), entonces dichos clumps se unen en un mismo clump.

5. (*Smoothing*) Finalmente un segundo automata celular es utilizado para suavizar los bordes de los clumps encontrados.

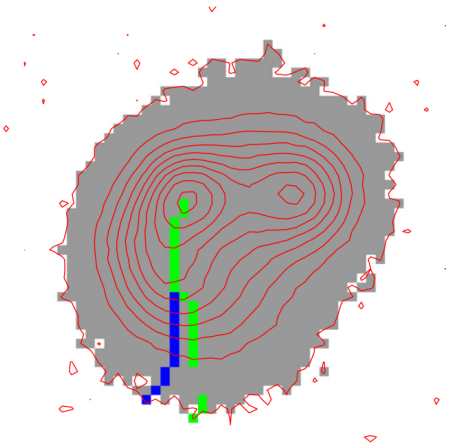

Figura 3: Dos rutas de ascenso de FellWalker en data 2D (Fuente: [5])

En las comparaciones realizadas por Watson [6] y el mismo Berry [5], ambos concluyen que FellWalker es un algoritmo más robusto en cuanto a los resultados que obtiene, en comparación a ClumpFind. En general ClumpFind tiende a ser muy sensible a sus parámetros de entrada y fragmenta los clumps más de los debido.

2.5. DENDROGRAMAS

Como se explicó en la introducción, uno de los aspectos más importantes para comprender el proceso de formación de estrellas en la nubes moleculares, es entender la relaciones jerárquicas existentes entre los

distintos núcleos densos o clumps de la nube. Para ello Rosolowsky et al. [7] propone organizar las clumps encontrados en una estructura de tipo árbol, denominada *dendrograma*.

El enfoque es distinto a los algoritmos de segmentación local como ClumpFind y FellWalker, pues este propone realizar un seguimiento de las estructuras presentes sobre un rango de escalas. De modo similar a ClumpFind, este computa isosuperficies de los cubos de datos que representan a la nube, descomponiendo iterativamente tales isosuperficies en otras de menor tamaño, generando un catálogo de regiones que se almacenan en una estructura de árbol (Figura 8). De esto modo, cada punto del dendrograma corresponde a un volúmen específico en dentro del cubo de datos, definido por la isosuperficie que lo acota.

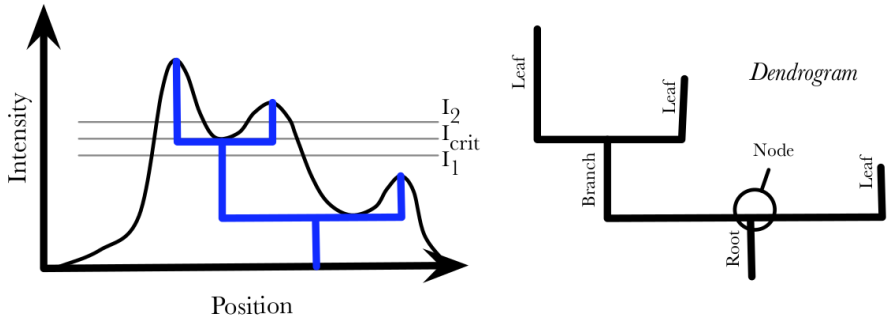

Figura 4: Esquema de funcionamiento de Dendrogramas. *Thresholding* al nivel I_1 produce un sólo objeto, mientras que al hacerlo a I_2 permite identificar dos. (Fuente: [7])

3. DESCRIPCIÓN FORMAL DEL ENFOQUE

3.1. CÁLCULO VARIACIONAL

El cálculo variacional también conocido como cálculo de variaciones, es un campo muy antiguo (y por lo tanto muy estudiado) que trata acerca de la minimización de *Funcionales*. Formalmente un funcional $\mathscr{U}$ es una correspondencia que asigna a cada función $f \in \mathscr{F}(\Omega \to \mathbb{C})$ un valor en un cuerpo escalar $\mathscr{K}$. Los funcionales son usualmente expresados como integrales definidas, involucrando a la función en cuestión y sus derivadas, como por ejemplo:

$$\mathscr{U}(f) = \int_{\Omega} L\left(\mathbf{x}, f(\mathbf{x}), \nabla f(\mathbf{x}), \Delta f(\mathbf{x})\right) d\mathbf{x}, \tag{2}$$

donde la función L es usualmente conocida en este ámbito, como el *Lagrangiano*, así mismo al funcional $\mathscr{U}$ se le llama *Funcional de Energía*.

El cálculo variacional sigue las mismas ideas del cálculo tradicional, esto es, define un equivalente a la derivada de las funciones pero sobre los funcionales, conocidas como *variaciones*. Igualando la primera variación o *variación ge Gateaux* a cero, y bajo ciertas otras restricciones, es posible encontrar la función f_0 que minimiza la energía del funcional, dando como resultado una ecuación diferencial llamada **Ecuación de Euler-Lagrange**.

De este modo se traslada el problema integral a un problema diferencial, para el cuál es posible ocupar toda la maquinaria de métodos numéricos (*Finite differences*, *Finite Element Methods*, *Collocation Methods*, *Spectral Methods*, etc) para intentar resolverla.

Para un tratamiento más acabado y formal de la formulación del problema variacional existen gran cantidad de libros que se pueden consultar. Dos buenas referencias son Logan [8] y Aubert [9].

3.2. APLICACIONES DEL CÁLCULO VARIACIONAL

En las últimas décadas el cálculo variacional ha venido sido utilizado como una potente herramienta para resolver múltiples problemas de minimización. Una de las áreas que más ha explotado esta herramienta es precisamente la del Procesamiento de Imágenes (*Image Processing*). Tal como expone Aubert [9] dependiendo del problema a tratar, es posible construir un funcional acorde, de modo que su minimización permita alcanzar distintos objetivos en cuanto a procesamiento de imágenes. En lo que sigue se describen algunas de las aplicaciones principales en imágenes.

1. DENOISING.
   Objetivo: Encontrar una aproximación *suave* de la imagen, que por medio de este suavizamiento elimine el ruido en las zonas donde este se encuentre presente.

Figura 5: Proceso de *denoising* por cálculo variacional

2. RESTORATION.
   Objetivo: Encontrar una aproximación de una imagen con errores (zonas sin datos por ejemplo), que reconstruya estas zonas con errores de una manera *suave*.

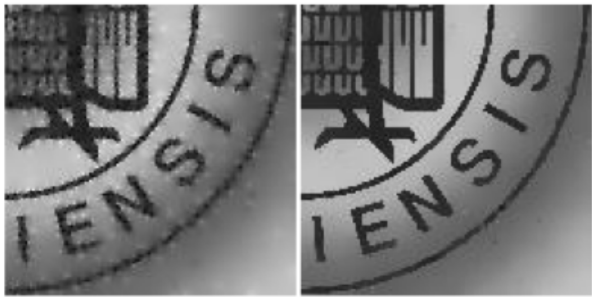

Figura 6: Proceso de *restoration* por cálculo variacional

3. SEGMENTATION.
   Objetivo: Encontrar una curva cerrada *suave*, entre un objeto de la imagen y el fondo.

4. REGISTRATION.
   Objetivo: Encontrar un campo de deformación, que permita hacer calzar una imagen de template y un imagen de entrada.

Para los dos primeros (*denoising* y *restoration*) el funcional de energía a minimizar tiene la forma general siguiente:

$$E(\mathbf{u}) = \int_\Omega \underbrace{D(\mathbf{u})}_{\text{Data term}} + \underbrace{S(\mathbf{u})}_{\text{Smoothing term}} + \underbrace{T(\mathbf{u})}_{\text{Domain specific term}} \, d\mathbf{x}, \tag{3}$$

Figura 7: Proceso de *segmentation* por cálculo variacional

Figura 8: Proceso de *registration* por cálculo variacional

en donde cada término del *Lagrangiano* cumple un objetivo específico: 1) D, que la solución sea cercana a la data de entrada, 2) S, que la solución sea suave, 3) T, que se cumplan restricciones propias del problema.

Las formulaciones del funcional a minimizar para los dos siguientes son más complicadas y requieren de un mayor análisis. Estas pueden ser consultadas en [9].

Referencias

[1] R. Kennicutt Jr and N. Evans II. Star Formation in the Milky Way and Nearby Galaxies. *Anual Review of Astronomy and Astrophysics*, 50:531–608, 2012.

[2] Jonathan P. Williams, Eugene J. de Geus and Leo Blitz. Determining structure in molecular clouds. *The Astrophysical Journal*, 498:693–712, 1994.

[3] J. Stutzki and R. Güsten. High spatial resolution isotopic CO and CS observations on M17 SW: The clumpy structure of the molecular cloud core. *The Astrophysical Journal*, 356:513–533, 1990.

[4] C. Kramer, J. Stutzki, R. Röhrig and U. Corneliussen. Clump mass spectra of molecular clouds. *Astronomy and Astrophysics*, 329:249–264, 1998.

[5] D.S Berry. FellWalker—A clump identification algorithm. *Astronomy and Computing*, 10:22–31, 2015.

[6] Mark E. Watson. Assessing The Performance of Sub-Millimetre Compact Object Detection Algorithms. Master's Thesis, University of Hertfordshire, 2010.

[7] E.W Rosolowsky, J.E Pineda, J. Kauffmann and A. A. Goodman. Structural Analysis of Molecular Clouds: Dendrograms. *The Astrophysical Journal*, 679:1338–1351, 2008.

[8] J. David Logan. *Applied Mathematics*. John Wiley & Sons INC., 3 edition, 2006.

[9] G. Aubert and P. Kornprobst. *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations*. Springer, 2 edition, 2006.

ANEXOS

Actividades	Planificación	Búsqueda de Información	Análisis	Desarrollo	Edición	Total
Tiempos SCT	0.5 hr	1.5 hr	1 hr	6 hr	2 hr	11 hr