Molecular Clump Extraction Algorithm Based on Local Density Clustering

Xiaoyu Luo¹, Sheng Zheng¹, Yao Huang¹, Shuguang Zeng¹, Xiangyun Zeng¹, Zhibo Jiang², Zhiwei Chen²
¹ Center for Astronomy and Space Sciences, China Three Gorges University, Yichang 443000, People’s Republic of China
² Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210023, People’s Republic of China

Abstract The detection and parametrization of molecular clumps is the first step in studying them. We propose a method based on Local Density Clustering algorithm while physical parameters of those clumps are measured using the Multiple Gaussian Model algorithm. One advantage of applying the Local Density Clustering to the clump detection and segmentation, is the high accuracy under different signal-to-noise levels. The Multiple Gaussian Model is able to deal with overlapping clumps whose parameters can be derived reliably. Using simulation and synthetic data, we have verified that the proposed algorithm could characterize the morphology and flux of molecular clumps accurately. The total flux recovery rate in $^{13}$CO (J=1-0) line of M16 is measured as 90.2%. The detection rate and the completeness limit are 81.7% and 20 K km s$^{-1}$ in $^{13}$CO (J=1-0) line of M16, respectively.

Keywords: Molecular Clump, Local Density Clustering, Multiple Gaussian Model

1 INTRODUCTION

The detections of the interstellar molecular hydrogen (H$_2$) by Carruthers (1970) in the ultraviolet band and carbon (CO) by Wilson et al. (1970) at 2.6 mm, creating an exciting new era in the study of the molecular interstellar medium, while the discovery of organic molecules in the medium led to the birth of molecular astrophysics. As one of the fundamental components of the interstellar medium, molecular clouds consist mainly of molecular gas with a mixture of atoms, ions, dust, and other materials (Heyer & Dame, 2015; Heiles et al., 2019). The molecular clouds in the galaxy exhibits the structure over a wide range of scales, from 20 pc or more for giant molecular clouds down to 0.05 pc for dense molecular clumps (Williams et al., 2012; Kauffmann et al., 2013; Lin et al., 2020).

Modern astronomy proved that the formation of stars is inside the molecular clumps (Krumholz & McKee, 2005; Zinnecker & Yorke, 2007; Krumholz et al., 2009). Therefore, the molecular clumps are the keys for theoretical models that aim to reproduce the observed characteristics of star formation in the Galaxy (Rivera-Ingraham et al., 2017; Tang et al., 2019).

As a consequence, several telescopes (e.g., the FCRAO 14 m, the CF 1.2 m, the Bell Laboratories 7 m, the PMO 13.7 m telescopes) have devoted to the CO survey projects (Sanders et al., 1986; Dame et al., 2001; Lee et al., 2001; Zuo et al., 2011). These CO surveys will lead to a better understanding of the evolution of molecular clump, the initial mass function of stars as well as the structure and dynamic evolution of the Milky Way (Heyer & Dame, 2015). With the progresses of the CO survey, it is impractical to process massive amounts of data manually. Therefore, a stable and reliable algorithm for automatically detecting the molecular clumps has become the focus. Several algorithms have been used to detect molecular clumps, such as GaussClumps, FellWalker, ClumpFind and Reinhold (Stutzki & Guesten, 1990; Berry, 2015; Williams et al., 1994; Berry et al., 2007). GaussClumps was first applied to the M17 molecular cloud to detect molecular clumps, and then frequently applied to the detection of clumps in other molecular clouds (Schneider et al., 1998; Dent et al., 2009; Lo et al., 2009). The ClumpFind algorithm was applied to the detection of compact structures in the Rosette molecular clouds. A new giant filament was found by Zhan et al. (2016) with a statistical study on the giant molecular cloud M16 (Sugitani et al., 2002).

Studies shows that the ClumpFind is very sensitive to the initial parameters, and the GaussClumps can only fit a strict elliptic shape. FellWalker exhibit the best performance in detection completeness and parametrization (Li et al., 2020). However, it should be noted that the GaussClumps and the ClumpFind algorithms are affected by the initial parameters, and the algorithms themselves are designed to simulate the “human eye” for molecular clump recognition, which have certain limitations (Rosolowsky et al., 2008). Moreover, for large amounts of molecular cloud data, it is clearly not feasible to rely on algorithms with repeatedly setting parameters by users, although it is possible to achieve satisfactory detection results in certain cases. Therefore, we need to design an algorithm which has fewer parameters or can be adjusted more easily based on the physical properties.

One of the dominant features of molecular clumps with increased local intensity and different shapes is that they are embedded in molecular gas of lower average bulk density (Blitz & Stark, 1986; Lada, 1992). The Local Density Clustering (LDC) algorithm (Alex Rodriguez, 2014) has its basis on assumptions that the cluster centers are surrounded by neighbors with lower local density and they are at a relatively large distance from the points with a higher local density, which...
is similar to the characteristics of molecular clumps. Therefore, we attempt to adopt the LDC in the detection of molecular clumps. In Section 2, the molecular clump detection algorithm based on LDC and parametrization based on Multiple Gaussian Model (MGM) are introduced. In Section 3, the 3D simulated datasets with different number density are described. The performance of the LDC & MGM is compared with traditional algorithms on the datasets. The investigation of the completeness and parametrization of the algorithm in real molecular clump data are in Section 4, while the summary is in Section 5.

2 ALGORITHMS

2.1 The LDC Algorithm

2.1.1 Features Extraction

The algorithm firstly compute three parameters of a point: the local density, the distance, and the gradient. The local density \( \rho_i \) of a point \( p_i \) is defined as:

\[
\rho_i = \sum_j \left( e^{-\frac{(d_{ij}^2)}{2}} \cdot I_j \right)
\]

where the \( d_{ij} \) represents the cut-off distance, \( d_{ij} \) represents the distance between \( p_i \) and \( p_j \), \( I_j \) represents the intensity at \( p_j \).

The distance \( \delta_i \) of a \( p_i \) is defined as:

\[
\delta_i = \min_j \{d_{ij}\} \quad \quad (2)
\]

\( \delta_i \) is measured by computing the minimum distance between \( p_i \) and any other point with higher density. Specially, \( \delta_i \) is set to be the maximum if \( p_i \) has the highest density. The distances \( \delta_i \) are normalized.

The nearest route could be obtained while calculating the distances. The index of the point with the smallest distance among all the data points whose density is greater than the current \( \rho_i \) is recorded in the vector \( N^{(p)} = \{n_1^{(p)}, n_2^{(p)}, \ldots, n_i^{(p)}, \ldots, n_k^{(p)}\} \):

\[
n_i^{(p)} = \left\{ \begin{array}{ll}
0, & \text{if } \delta_i = \max(\delta_j),
\end{array} \right.
\]

where \( n \) represents the total number of data points, the point with longest distance is set as 0 in the vector \( N^{(p)} \).

The gradient \( \nabla_i \) is defined as:

\[
\nabla_i = \frac{\rho_j - \rho_i}{\delta_i} \quad \quad (4)
\]

where \( \rho_j \) and \( \delta_i \) are defined in Formula (2).

2.1.2 The Clump Center Determination

After calculating three parameters, as shown in Figure 1, the distance \( \delta \) is plotted against the density \( \rho \), which is referred to the Decision Graph. The simulated data with 10 clumps is shown in Figure 1 (a), while the detected clumps are shown in Figure 1 (b) with centers marked by red stars. Figure 1 (c) shows the Decision Graph, where the centers of the detected clumps are marked with circles. Whether \( p_i \) is the center point of a clump or not is judged by:

\[
p_i = \left\{ \begin{array}{ll}
p_i^{(C)}, & \text{if } \delta_i \geq \delta_0, \rho_i \geq \rho_0, \\
p_i^{(non-C)}, & \text{else}
\end{array} \right. \quad \quad (5)
\]

where \( p_i^{(C)} \) represents the center point of the \( k_{th} \) clump, \( p_i^{(non-C)} \) represents the point of non-clump. \( \delta_0 \) and \( \rho_0 \) are hyper-parameters of our algorithm, where \( \delta_0 \) represents the minimum distance between the centers of the two clumps, and \( \rho_0 \) represents the minimum peak intensity value of a candidate clump.

Fig. 1: The example of algorithm on 2D simulated data. (a) The 2D data contains 10 simulated clumps. (b) Clustering result. (c) Decision Graph of the data in (a).

2.1.3 Route Clustering

According to the information recorded in the route vector \( N^{(p)} \), the route \( P_k \) end with the \( k_{th} \) center point \( p_{nk}^{(C)} \) of clump can be obtained:

\[
P_k = \{p_1^{(k)}, p_2^{(k)}, \ldots, p_j^{(k)}, \ldots, p_{nk}^{(k)} \} \quad \quad (6)
\]

other points of non-clump are divided into the \( k_{th} \) clump \( C_k \) according to route \( P_k \):

\[
p_j \in C_k, \text{ if } p_j \in P_k \quad j = 1, 2, \ldots, n_k \quad \quad (7)
\]

where \( N \) and \( n_k \) represent the number of clumps and the number of data points in the clump \( C_k \), respectively.

2.1.4 Clump Region Determination

The region of the individual clump \( C_k \) can be determined according to \( \rho \) and \( \nabla \):

\[
p_j^{(k)} \in C_k, \text{ if } \rho_j \geq \rho \text{ or } \nabla_j \geq \nabla_0 \quad j = 1, 2, \ldots, n_k \quad \quad (8)
\]

where \( p_j^{(k)} \) represents the \( j_{th} \) point in clump \( C_k \), \( \rho \) is the average density of the clump \( C_k \), \( \nabla_0 \) is the hyper-parameter. The individual clump could be segregated as \( \rho_j \) greater than \( \rho \) or \( \nabla_j \) greater than \( \nabla_0 \). The morphological image processing is employed to fill in holes among detected clumps and to smooth its boundary.
2.1.5 False Clumps Exclusion

The isolated noise points with high peak intensity value could be recognized as false clumps. The smallest clump should have enough data points to form it. Therefore, the false detected clumps could be eliminated by the following criteria:

\[
C_k = \begin{cases} 
\text{True}, & \text{if } n_k \geq n_0 \\
\text{False}, & \text{else}
\end{cases}
\]  

(9)

where \(n_0\) is the minimum data point number of a clump.

2.1.6 Clump Characterization

The algorithm will provide a pixel mask which is the same shape as the supplied data array. In the mask, the pixel points belonging to the same clumps are marked with an integer, while points that are not assigned are marked with -1. Finally, a table in which each row describes an individual clump is obtained. In each column of the table, \(\text{Peak}_i\) and \(\text{Cen}_i\) represent the position of the clump peak intensity value and centroid on axis \(i\) \((i = 1, 2, 3)\), respectively. \(\text{Size}_i\) represents the size of the clump on axis \(i\). \(\text{Sum}\) and \(\text{Peak}\) represent the total flux and peak intensity value of the clump, respectively. The definition of the centroid is as follows:

\[
\text{Cen}_i = \frac{\sum_{j=1}^{n} (I_j \cdot x_j)}{\sum_{j=1}^{n} I_j}
\]  

(10)

the \(\text{Size}_i\) of the clump \(C_k\) on axis \(i\) is defined as:

\[
\text{Size}_i = \sqrt{\frac{\sum_{j=1}^{n} (I_j \cdot x_j^2) - \left(\frac{\sum_{j=1}^{n} (I_j \cdot x_j)}{\sum_{j=1}^{n} I_j}\right)^2}{\sum_{j=1}^{n} I_j}}
\]  

(11)

where \(I_j\) and \(x_j\) represent the intensity and position of \(p_j\), respectively. For the clump with a Gaussian profile, the size is equal to the standard deviation of the Gaussian.

2.1.7 Algorithm Summarising

The input of algorithm is a 3D (or 2D) data array. \(\delta_0\) and \(\rho_0\) are key hyper-parameters of the algorithm, where \(\delta_0\) represents the minimum distance between the centers of the two clumps, and \(\rho_0\) represents the minimum peak intensity value of a candidate clump. \(n_0\) is the minimum data point number of a clump and \(\nabla_0\) is used to determine the region of a clump. The local density of a point is calculated with the cut-off distance \(d_c\). The input and parameters of the LDC algorithm are listed in Table 1. The outputs include masks indicating the pixels that contribute to each clump, and catalogs holding clump positions, sizes, peak values and total fluxes. The output and parameters of the LDC algorithm are listed in Table 2.

Table 1: The input and parameters of LDC algorithm

| Input   | Description | Default value |
|---------|-------------|---------------|
| Data array | 3D or 2D data array |              |
| \(\rho_0\) | \(3\sigma\) |               |
| \(\delta_0\) | 4           |               |
| \(\nabla_0\) | 0.01        |               |
| \(n_0\) | 27          |               |
| \(d_c\) | 0.8         |               |

2.2 Parametrization Based on MGM

Traditional algorithms are used to segregate overlapping molecular clumps, and there could be large deviation in the parameter estimation of overlapping molecular clumps. Therefore, we adopt MGM to realize the parametrization in this paper.

2.2.1 The 3D Gaussian Model

The observation data of molecular clump is a 3D data array. The first and second dimension of the 3D data array stand for the galactic longitude and latitude, respectively. The third dimension of the 3D data array stands for the velocity. Due to the spatial and velocity are not related, the tilt angles of simulated clumps only appear on the galactic longitude - latitude plane. Therefore, the 3D Gaussian Distribution is described as:

\[
f(x, y, v) = A \exp \left\{ - \left[ \frac{\cos^2 \theta}{2\sigma_x^2} + \frac{\sin^2 \theta}{2\sigma_y^2} \right] (x - x_0)^2 \right. \\
\left. + \left( \frac{\sin 2\theta}{2\sigma_x^2} - \frac{\sin 2\theta}{2\sigma_y^2} \right) (x - x_0) (y - y_0) \right. \\
\left. \left. + \left( \frac{\sin^2 \theta}{2\sigma_x^2} + \frac{\cos^2 \theta}{2\sigma_y^2} \right) (y - y_0)^2 + \frac{(v - v_0)^2}{2\sigma_v^2} \right\}
\]  

(12)

where \((x_0, y_0, v_0)\) represents center point of the distribution, \(\sigma_x, \sigma_y, \sigma_v\) represent the standard deviations in the three axes, respectively. The variable \(A\) represents amplitude of the distribution, and \(\theta\) represents the tilt angle on the \(x - y\) plane.

2.2.2 The 3D MGM

For the scenario where multiple Gaussian components overlap, adopting a single Gaussian distribution to explain will lead to serious deviation. Taking Figure 2 as an example, the black solid line represents the actual data obtained, and the three
dashed lines represent the actual components. The MGM can effectively solve this problem. The MGM is defined as follows:

\[
f_{mix}(x, y, v; \psi) = \sum_{k=1}^{K} f_k(x, y, v; \psi_k)
\]  

(13)

where \(f_k(x, y, v; \psi_k)\) represents \(\psi_k\), 3D Gaussian Distribution described in Formula (12), and \(K\) is the number of Gaussian components. \(\psi = \{\psi_1, \psi_2, \ldots, \psi_k, \ldots, \psi_K\}\) defines the parameters of the model. \(\psi_k\) represents the parameters of the \(k_{th}\) Gaussian Distribution. \(\psi_k\) is specified as :

\[
\psi_k = \begin{bmatrix} A^{(k)} & x^{(k)}_0 & y^{(k)}_0 & v^{(k)}_0 & \sigma_x^{(k)} & \sigma_y^{(k)} & \sigma_v^{(k)} & \theta^{(k)} & \epsilon^{(k)} \end{bmatrix}^T
\]  

(14)

Using the LDC algorithm described in Section 2.1, a series of clumps \((C_1, C_2, \ldots, C_k, \ldots, C_N)\) can be obtained, and the parameters of those clumps calculated in Section 2.1.6 could serve as the initial value \((\psi_1^{(0)}, \psi_2^{(0)}, \ldots, \psi_k^{(0)}, \ldots, \psi_N^{(0)})\) of the model. The clump \(C_1\) and \(C_2\) are considered to overlap each other when \(\|(x_{i0} - x_{j0}, y_{i0} - y_{j0}, v_{i0} - v_{j0})\|_2 \leq \|(\sigma_{ix} + \sigma_{jx}, \sigma_{iy} + \sigma_{jy}, \sigma_{iv} + \sigma_{jv})\|_2 \) \((\| \cdot \|_2\) is the two-norm). Then the parameters \((\psi_1^{(0)}, \psi_2^{(0)}, \ldots, \psi_k^{(0)}, \ldots, \psi_N^{(0)})\) of the overlapping clumps (suppose the number of overlapping clumps is \(m\)) could serve as the initial parameters while using the MGM fit those overlapping clumps. Finally, the catalogue with various clump parameters will be obtained via MGM fitting method.

3 COMPARISON WITH OTHER ALGORITHMS

3.1 Detection Accuracy

3.1.1 3D Simulated Data

The simulated datasets are composed of different number density data with the size of 100 \(\times\) 100 \(\times\) 100, and data at low, medium and high density contain 10, 25 and 100 simulated clumps, respectively. The peak intensity value of the clump take values from 2 to 10, while the size of the clump in velocity axis take values from 3 to 5 and the spital size in the \(x\) and \(y\) axes take values from 2 to 4 (FWHM = 2.35 \(\times\) size). The tilt angles of the simulated clumps on the \(x-y\) plane vary from 0\(^{\circ}\) to 180\(^{\circ}\). Gaussian noise is added to the simulated clumps with a root-mean-square (rms) of 1. For each number density, we generate a total of 10000 simulated clumps. Figure 3 shows the 3D display of one simulated data array.

Combined with the \(\nabla\) and \(\rho\) of each data point, the members and region of the clump \(C_k\) could be determined by Formula (8). (1) Using \(\nabla_0\) as the threshold, the point set \(A_1\) with \(\nabla\) greater than \(\nabla_0\) is the main part of the clump \(C_k\). (2) The average density \(\bar{\rho}\) is calculated based on the point set \(A_1\), then the point set \(A_2\) represents \(\rho\) greater than \(\bar{\rho}\) which is also the main part of the clump. (3) The union of \(A_1\) and \(A_2\) could form the region of an individual clump \(C_k\). The detection results are shown in Figure 5. The region of each clump will determined while the false clump will eliminated by Formula (9). Finally, the parameter estimation in Section 2.1.6 will be performed.

Fig. 3: 3D display of three Gaussian components. The black line is composed of a combination of three gaussian distributions. Dashed lines represent each gaussian components.

Fig. 4: The centers of detected clumps are marked on the integrated intensity maps with red asterisks. From left to right are integral maps of \(x-y\), \(x-v\) and \(y-v\) planes, respectively.

Fig. 5: The intergrated intensity maps of detected clumps are marked with red asterisks. From left to right are integral maps of \(x-y\), \(x-v\) and \(y-v\) planes, respectively.
shows the evaluation of the three algorithms at different peak signal-to-noise ratio (PSNR) levels and different density. The accuracy and completeness of detection are reflected in the precision rate ($P$) and comprehensive score ($F_1$). The $R$, $P$ and $F_1$ are defined as:

$$R = \frac{TP}{TP + FN}, P = \frac{TP}{TP + FP}, F_1 = \frac{2 \cdot P \cdot R}{P + R}$$ (15)

The accuracy and completeness of detection are reflected in $P$ and $R$, respectively. A good detection algorithm should have higher $P$ and $R$. Usually the two indicators will show the opposite trend. The comprehensive performance ability of the algorithm is mainly reflected in $F_1$.

### 3.1.4 Detection Comparison

The GaussClumps, FellWalker and LDC are employed in the detection of simulated clumps. Figure 6 shows the evaluation of indicators $R$, $P$ and $F_1$ of the GaussClumps, FellWalker, and LDC algorithms in different peak signal-to-noise ratio (PSNR) levels and different density. The PSNR is defined as the ratio of the peak intensity value of the simulated clump to the rms of noise. As the PSNR decreases, $R$ of these algorithms at different density levels starts to decrease, especially when the PSNR is less than 4. The FellWalker and LDC algorithms generally have high $P$, while the same indicator of GaussClumps performed worse with the PSNR less than 6. It is obvious that $R$ of those algorithms at different density hold high level with the PSNR above 6, and $P$ of those algorithms hold high level with the PSNR greater than 7, while $P$ of GaussClumps gradually descend with the decrease of the PSNR.

The top panel in Figure 6 shows the $R$, $P$ and $F_1$ of GaussClumps, FellWalker, and LDC algorithms at high density from left to right, respectively. The $R$ and $P$ are above 80% for those algorithms when the PSNR is greater than 5, and $R$ of GaussClumps is lower than the other two algorithms. The middle panel shows the same as above but for medium density. For those clumps in the simulation that overlap heavily or even merge into new clumps, FellWalker and LDC are unable to distinguish these clumps, leading to a decrease in $R$. Since GaussClumps detects the clumps by fitting, it can separate the overlapping clumps from each other, thus improving $R$. The middle panel shows the same as above but for medium number density. $P$ of the three algorithms are essentially the same as in the case of high density, but $R$ of those algorithms have increased and the gap between GaussClumps and the other two algorithms is further reduced with the PSNR above 6. The bottom panel shows the same as above but for low number density. $P$ of the three algorithms are basically the same as in the case of high density, but $R$ are above 90% for the three algorithms when the PSNR is greater than 5, and $R$ of GaussClumps is lower than the other two algorithms.

The experimental results show that $P$ and $R$ of FellWalker and LDC algorithms can be maintained at high level, but $R$ decreases in the case of high density. GaussClumps algorithm has high $R$ and $P$ at the certain PSNR indicating that it is suscepti-

---

**Fig. 6:** The evaluation indicators $R$, $P$ and $F_1$ of the three algorithms are plotted against the PSNR at different number density. *Top panel:* the detection statistics of the three algorithms in high density, from left to right are $R$, $P$ and $F_1$, respectively. *Middle panel:* same as above but for medium density. *Bottom panel:* same as above but for low density.
ble to noise. In terms of the comprehensive performance indicator $F_1$, the FellWalker and the LDC algorithm are essentially the same, both outperforming the GaussClumps algorithm in low PSNR.

### 3.2 Parametrization

#### 3.2.1 Evaluation Indicators

To investigate the performance of the algorithm in terms of parametrization accuracy, various measured parameters are compared with their input values, peak intensity, total flux, tilt angle, size, and position of the clump. For each parameter, the absolute deviation of the position $E(\Delta X)$, angle $E(\Delta \theta)$, and the relative deviation of size $E(\Delta S)$, peak intensity $E(\Delta I)$ and total flux $E(\Delta F)$ are calculated. Those evaluation indicators are defined as:

\[
E(\Delta X) = \frac{1}{N} \sum_{i=1}^{N} (X_i^{(s)} - X_i^{(m)})
\]

\[
E(\Delta \theta) = \frac{1}{N} \sum_{i=1}^{N} (\theta_i^{(s)} - \theta_i^{(m)})
\]

\[
E(\Delta S) = \frac{1}{N} \sum_{i=1}^{N} \frac{S_i^{(s)} - S_i^{(m)}}{S_i^{(s)}}
\]

\[
E(\Delta I) = \frac{1}{N} \sum_{i=1}^{N} \frac{I_i^{(s)} - I_i^{(m)}}{I_i^{(s)}}
\]

\[
E(\Delta F) = \frac{1}{N} \sum_{i=1}^{N} \frac{F_i^{(s)} - F_i^{(m)}}{F_i^{(s)}}
\]

where $N$ represents the number of simulated molecular clumps which are detected correctly by the algorithm. The superscript $s$ and $m$ represent the parameters of the simulated molecular clumps and measured by the algorithm, respectively. $X$, $S$, $I$ and $F$ represent the position, size, peak and total flux of the clump, $\theta$ represents the tilt angle on the $x-y$ plane of the clump.

#### 3.2.2 Performance

We launched statistical experiments to compare the parametrization performance of FellWalker, GaussClumps and LDC & MGM algorithms. The high density simulated data described in Section 3.1.1 are used in the statistical experiments.

Figure 7 shows the relative deviation of peak intensity value as a function of the PSNR. The vertical axis represents the relative deviation of peak intensity between the simulated clump and measured clump, and the horizontal axis represents the PSNR of the clump. The blue, green and red dots represent the relative deviation of clumps detected by FellWalker, GaussClumps and LDC & MGM, respectively. When the dot is above 0, it means that the value of measured by algorithm is less than the simulated, otherwise, the value of measured is greater than the simulated. Error bars represent standard deviation of accuracy. The blue circle, green triangle and red asterisk represent the median of relative deviation measured by the FellWalker, GaussClumps and LDC & MGM algorithm, respectively.

From Figure 7 we can see that as the PSNR of the simulated clump increase, the deviation of the GaussClumps and FellWalker algorithms gradually decrease, while the peak intensity values measured by both algorithms are greater than the simulated. The deviations obtained by the LDC & MGM algorithm are close to 0 with the dispersion decreased gradually, indicating that the peak intensities estimated from the LDC & MGM algorithm is more reliable.

The total flux is an important parameter, which is directly related to the column density and mass of a molecular clump. As can be seen in Figure 8, the total fluxes of GaussClumps and FellWalker algorithms are smaller than the simulated values. The reason is that both algorithms have a cutoff threshold for background noise in detecting molecular clumps and can only detect part of clumps. The most deviation of LDC & MGM does not exceed $\pm 30\%$ with the PSNR greater than 4, indicating that the LDC & MGM is stable in the total flux estimation of molecular clumps.
Figure 9 shows the deviation of tilt angle, the symbols are the same as Figure 7. We can see that the dispersions of the measured deviations are decreased gradually with increasing of the PSNR for the three algorithms, while the deviation is less than 0.5° when the PSNR greater than 4, indicating that the estimation of molecular clump angle by this algorithm is stable.

![Figure 9](Image 45x540 to 289x706)

The position deviations measured by FellWalker and LDC & MGM are almost within 1 pixel and the deviation is no more than 0.5 pixel at the PSNR greater than 4, while the deviation from GaussClumps is greater than the two algorithms. We can see that some horizontal bars appear in the distribution of position measured by GaussClumps in galactic latitude and longitude direction. The reason is that the low spatial resolution of the simulated clumps leading to the centers fitted by Gaussclumps are mainly located on the grid.

Figure 11 shows the absolute deviation of position as a function of the PSNR, from left to right are the deviation on Galactic latitude, Galactic longitude, and velocity, respectively. The position deviations measured by FellWalker and LDC & MGM are almost within 1 pixel and the deviation is no more than 0.5 pixel at the PSNR greater than 4, while the deviation from GaussClumps is greater than the two algorithms. We can see that some horizontal bars appear in the distribution of position measured by GaussClumps in galactic latitude and longitude direction. The reason is that the low spatial resolution of the simulated clumps leading to the centers fitted by Gaussclumps are mainly located on the grid.

Overall, detecting clumps by LDC & MGM at high number density has robust parametrization accuracy in term of position, peak, total flux, size, and tilt angle. The molecular clumps parametrization of the proposed algorithm show less deviation and less dispersion than FellWalker and GaussClumps algorithms with the PSNR above 5.

4 EXPERIMENT IN REAL DATA

4.1 M16 Data

The $^{13}$CO (J=1-0) line of M16, including the region within $15°15' < l < 18°15'$ and $0° < |b| < 1°30'$ from the Milky Way Imaging Scroll Painting (MWISP) survey (Sun et al., 2018), is employed in the molecular clump detection and parametrization. The typical noise level at $^{13}$CO (J=1-0) line is about 0.23 K with the channel width of 0.167 km s$^{-1}$. Figure 12 shows the integrated intensity maps of M16 in $^{13}$CO (J=1-0) line.

Using the M16 data, Zhan et al. (2016) has confirmed the identification of the giant molecular filament (GMF) G18.0-16.8 by Ragan et al. (2014) and find a new giant filament, G16.5-15.8, located in the west 0.8° of G18.0-16.8. Song & Jiang (2017) has calculated the properties of the clump samples under local thermodynamic equilibrium (LTE) assumption. The virial mass and virial parameter are calculated to evaluate whether clumps are bound or unbound. They found the majority of $^{13}$CO clumps are bound, which suggest that those clumps may form stars in the future. Based on their research in detection clump on M16, the $^{13}$CO (J=1-0) line of M16 is used to investigate the performance of our algorithm.

![Figure 12](Image 335x308 to 518x400)

4.2 Clump extraction experiment

After tuning the algorithm parameters, the GaussClumps, FellWalker and LDC & MGM algorithms are applied to detect the $^{13}$CO (J=1-0) line of M16. Figure 13 shows the distribution of peak intensity value of clumps detected by the three algorithms. The observed total flux is defined as the summed flux of those observations above 2×rms of the background. And the recovery rate is defined as the ratio the sum of clumps flux to the observed total flux. The recovery rate of total flux obtained by GaussClumps, FellWalker and LDC are 51.6%, 90.4% and 90.2% in $^{13}$CO emission of M16, respectively.
Fig. 10: The statistics of relative deviation in size as a function of the PSNR. From left to right are the deviation of $Size_1$, $Size_2$, and $Size_3$, respectively ($Size_1$, $Size_2$ represent major and minor size of detected clump in the spatial, $Size_3$ represents the size of detected clump in velocity axis). Two dashed horizontal lines represent the relative deviation of $\pm 10\%$. The blue, green and red dots, special symbols and error bars have the same meaning as Figure 7.

Fig. 11: The statistics of absolute deviation of the position as a function of the PSNR. From left to right are the deviation of galactic latitude, galactic longitude, and velocity, respectively. The blue, green and red dots, special symbols and error bars have the same meaning as Figure 7.

Fig. 13: The distribution of the detected peak intensity values of clumps by GaussClumps, FellWalker, and LDC.

It is can be infered from Figure 13 that the peak intensity values of clumps detected by LDC and FellWalker have a similar distribution with a more flattened peak, while the distribution of the peak intensity values detected by GaussClumps deviates greatly from the other two algorithms. The peak of distribution is about 2 in the FellWalker and LDC algorithm, while the peak intensity value of those simulated clump take values from 2 to 5, while the size of the clump in the velocity axis take values from 2 to 4 and the size in the galactic longitude and latitude axes take values from 0.5 to 2. The clumps detected by GaussClumps, FellWalker and LDC algorithm are matched with the simulated clumps. The number of clumps within each total flux interval are counted, the completeness and the average detection rate above the limitation is obtained.

Figure 14 shows the detection rate of the GaussClumps, FellWalker, and LDC algorithm in $^{13}$CO (J=1-0) line of M16, respectively. As the total flux increases, the detection rate of the GaussClumps grows slowly, while the FellWalker and LDC are able to maintain a relatively high detection rate all the way from the completeness limitation. The detection rate of GaussClumps, FellWalker and LDC are 80.9%, 74.7% and 81.7% above the completeness limitation, respectively. From the detection rate of each algorithm, we can roughly estimate...
that the number density in $^{13}$CO (J=1-0) line of M16 is between the high density and medium density in the simulation datasets described in Section 3.1.1.

![Detection Rate vs Total Flux](image)

Fig. 14: The detection rate of the three algorithms in $^{13}$CO (J=1-0) line of M16 as a function of total flux. The detection rate of GaussClumps, FellWalker and LDC are 80.9%, 74.7% and 81.7%, respectively. The completeness limitation of LDC and FellWalker are 20 K km s$^{-1}$ and 45 K km s$^{-1}$, respectively. While the GaussClumps is 75 K km s$^{-1}$.

Figure 15 shows the statistical histogram of $\Delta I$ for the three algorithms. The simulated clumps could overlap in real observations, leading to the detected peak intensity values of clumps by the three algorithms are systematically larger than those of simulated clumps. While the LDC has the least dispersion of deviations. The long tail at the left side of the peak deviation suggests a relatively high intensity of molecular clumps in M16.

![Peak Deviation](image)

Fig. 15: The histogram of the peak deviation ($\Delta I$) for the GaussClumps, FellWalker, and LDC. The $\Delta I$ is described in Formula (19). The mean deviation of FellWalker, GaussClumps and LDC are -10.8%, -11.0% and -8.9%, respectively. The standard deviation of FellWalker, GaussClumps and LDC are 29.4%, 31.5% and 28.3%, respectively.

5 CONCLUSION

We present a molecular clump detection and parametrization algorithm based on the Local Density Clustering and Multiple Gaussian Model (LDC & MGM). The proposed algorithm is robust and universal in the clump detection. The employed algorithm of LDC in the clump detection and segmentation could achieve high accuracy with different signal-to-noise levels, while the MGM could obtain reliable physical parameters of overlapping clumps.

We applied our method to a simulated data set, and find, (1) detection rate: the recall rate of the algorithm at high, medium and low number density simulated data is greater than 80%, 90%, and 97% with the PSNR above 6, respectively. The algorithm retains a high level of detection accuracy when the PSNR is greater than 3. (2) Accuracy of parameters: the parametrization of the algorithm in simulated data show less deviation and less dispersion with the PSNR above 5. The deviations of peak value and size are almost within 10% with the PSNR above 5, while the deviations of total flux hardly exceed 30% when the PSNR is greater than 4 at the high number density. The deviations of tilt angle on the $x - y$ plane are less than 10$^\circ$ with the PSNR above 4.

We apply our algorithm to the $^{13}$CO (J=1-0) map of the M16 nebula taken by PMO-13.7m telescope. The detection rate of clumps is up to 81.7% with a completeness limitation of 20 K km s$^{-1}$ in $^{13}$CO (J=1-0) line of M16. A total of 658 molecular clumps have been detected by our algorithm and the total flux recovery rate in $^{13}$CO (J=1-0) line of M16 is estimated as 90.2%. The number density in $^{13}$CO (J=1-0) line of M16 may be between the high and medium density in the simulation datasets described in Section 3.1.1.

Acknowledgements

We thank the anonymous referee for his/her suggestive comments that help improve the manuscript a lot. This work was supported by the National Natural Science Foundation of China (U2031202, 11903083, 11873093). This research made use of the data from the MWISP project, which is a multi-line survey in $^{12}$CO/$^{13}$CO/(18)O along the northern galactic plane with PMO-13.7m telescope. We are grateful to all the members of the MWISP working group, particularly the staff members at PMO-13.7m telescope, for their long-term support. MWISP was sponsored by National Key R&D Program of China with grant 2017YFA0402701 and CAS Key Research Program of Frontier Sciences with grant QYZDJ-SSW-SLH047.

References

Alex Rodriguez, A. L. 2014, Science, 344, 1492
Berry, D. S. 2015, Astronomy and Computing, 10, 22
Berry, D. S., Reinhold, K., Jenness, T., & Economou, F. 2007, in Astronomical Society of the Pacific Conference Series, Vol. 376, Astronomical Data Analysis Software and Systems XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell, 425
Blitz, L., & Stark, A. A. 1986, ApJL, 300, L89
Carruthers, G. R. 1970, ApJL, 161, L81
Dame, T. M., Hartmann, D., & Thaddeus, P. 2001, ApJ, 547, 792
Dent, W. R. F., Hovey, G. J., Dewdney, P. E., et al. 2009, MNRAS, 395, 1805
Heiles, C., Li, D., McClure-Griffiths, N., Qian, L., & Liu, S. 2019, Research in Astron. Astrophys. (RAA), 19, 017
Heyer, M., & Dame, T. M. 2015, ARA&A, 53, 583
Kauffmann, J., Pillai, T., & Goldsmith, P. F. 2013, ApJ, 779, 185
Krumschnitz, M. R., & McKee, C. F. 2005, ApJ, 630, 250
Krumschnitz, M. R., McKee, C. F., & Tumlinson, J. 2009, ApJ, 699, 850
Lada, E. A. 1992, ApJL, 393, L25
Lee, Y., Stark, A. A., Kim, H.-G., & Moon, D.-S. 2001, ApJS, 136, 137
Li, C., Wang, H.-C., Wu, Y.-W., Ma, Y.-H., & Lin, L.-H. 2020, Research in Astron. Astrophys. (RAA), 20, 031
Lin, L.-H., Wang, H.-C., Su, Y., Li, C., & Yang, J. 2020, Research in Astron. Astrophys. (RAA), 20, 143
Lo, N., Cunningham, M. R., Jones, P. A., et al. 2009, MNRAS, 395, 1021
Ragan, S. E., Henning, T., Tackenberg, J., et al. 2014, A&A, 568, A73
Rivera-Ingraham, A., Ristorcelli, I., Juvela, M., et al. 2017, A&A, 601, A94
Rosolowsky, E. W., Pineda, J. E., Kauffmann, J., & Goodman, A. A. 2008, ApJ, 679, 1338
Sanders, D. B., Clemens, D. P., Scoville, N. Z., & Solomon, P. M. 1986, ApJS, 60, 1
Schneider, N., Stutzki, J., Winnewisser, G., Poglitsch, A., & Madden, S. 1998, A&A, 338, 262
Song, C., & Jiang, Z. B. 2017, A Census of Dense Clumps in the M16 Giant Molecular Cloud, Master’s thesis, Shanghai Normal University, Shanghai
Stutzki, J., & Guesten, R. 1990, ApJ, 356, 513
Sugitani, K., Tamura, M., Nakajima, Y., et al. 2002, ApJL, 565, L25
Sun, J. X., Lu, D. R., Yang, J., et al. 2018, Acta Astronomica Sinica, 59, 3
Tang, M.-Y., Qin, S.-L., Liu, T., & Wu, Y.-F. 2019, Research in Astron. Astrophys. (RAA), 19, 040
Williams, J. P., Blitz, L., & Mckee, C. F. 2012, Physics, 97
Williams, J. P., de Geus, E. J., & Blitz, L. 1994, ApJ, 428, 693
Wilson, R. W., Jefferts, K. B., & Penzias, A. A. 1970, ApJL, 161, L43
Zhan, X.-L., Jiang, Z.-B., Chen, Z.-W., Zhang, M.-M., & Song, C. 2016, Research in Astron. Astrophys. (RAA), 16, 56
Zhou, P., Xiaoyu, L., Sheng, Z., Zhibo, J., & Shuguang, Z. 2020, ACTA ASTRONOMICA SINICA, 61, 14
Zinnecker, H., & Yorke, H. W. 2007, ARA&A, 45, 481
Zuo, Y.-X., Li, Y., Sun, J.-X., et al. 2011, Chinese Astronomy and Astrophysics, 35, 439