Research in Astron. Astrophys. **20XX** Vol. **X** No. **XX**, 000–000
http://www.raa-journal.org *http://www.iop.org/journals/raa*
(LaTeX: 2021-0272.tex; printed on October 21, 2021; 10:08)

Research in Astronomy and Astrophysics

Received 20XX Month Day; accepted 20XX Month Day

Molecular Clump extraction algorithm based on Local Density Clustering *

Xiaoyu Luo[1], Sheng Zheng[1], Yao Huang[1], Shuguang Zeng[1], Xiangyun Zeng[1], Zhibo Jiang[2], Zhiwei Chen[2]

[1] Center for Astronomy and Space Sciences, China Three Gorges University, Yichang 443000, People's Republic of China
xyzeng2018@163.com
[2] Purple Mountain Observatory, National Astronomical Observatories, Chinese Academy of Sciences, Nanjing 210008, 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.

Key words: 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 CfA 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 domainant 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

* Supported by the National Natural Science Foundation of China.

distance from the points with a higher local density, which is similar to the characteristics of melecular 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 melocular 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 ρ_i of a point p_i is defined as:

$$\rho_i = \sum_j (e^{-(d_{ij}/d_c)^2} \cdot I_j) \tag{1}$$

where the d_c 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 δ_i of a p_i is defined as:

$$\delta_i = \min_{j:\rho_j > \rho_i} d_{ij} \tag{2}$$

δ_i is measured by computing the minimum distance between p_i and any other point with higher density. Specially, δ_i is set to be the maximum δ if p_i with highest density. The distances δ_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 p_i is recorded in the vector $N^{(p)} = \{n_1^{(p)}, n_2^{(p)}, \cdots, n_i^{(p)}, \cdots, n_n^{(p)}\}$:

$$n_i^{(p)} = \begin{cases} 0, \text{ if } \delta_i = \max(\delta_i) \\ j, \text{ if } \delta_i = d_{ij} \end{cases} \tag{3}$$

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

The gradient ∇_i is defined as:

$$\nabla_i = \frac{\rho_j - \rho_i}{\delta_i} \tag{4}$$

where ρ_j and δ_i are defined in Formula (2).

2.1.2 *The Clump Center Determination*

After calculating three parameters, as shown in Figure 1, the distance δ is plotted against the density ρ, 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:

$$p_i = \begin{cases} p_k^{(C)}, \text{ if } \delta_i \geq \delta_0, \rho_i \geq \rho_0 \\ p^{(non-C)}, \text{ else} \end{cases} \quad i = 1, 2, \cdots, n \tag{5}$$

where $p_k^{(C)}$ represents the center point of the k_{th} clump, $p^{(non-C)}$ represents the point of non-clump. δ_0 and ρ_0 are hyper-parameters of our algorithm, where δ_0 represents the minimum distance between the centers of the two clumps, and ρ_0 represents the minimum peak intensity value of a candidate clump.

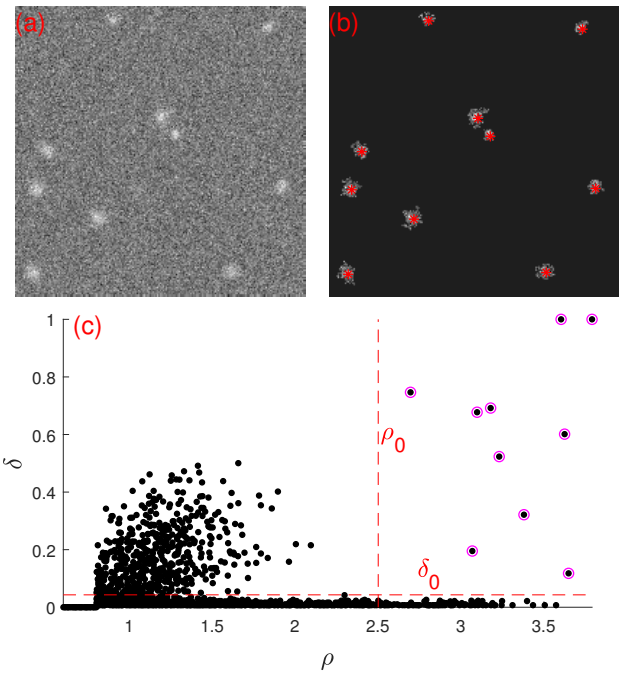

Fig. 1: The example of algorithm on 2D simulated data. (a) The 2D data contains 10 simulated clumps. (b) Clustering result. (c) Decision Graph of the data in (a).

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 $p_{n_k}^{(C)}$ of clump can be obtained:

$$P_k = \left\{p_1^{(k)}, p_2^{(k)}, \cdots, p_j^{(k)}, \cdots, p_{n_k}^{(C)}\right\} \; k = 1, 2, \cdots, N \tag{6}$$

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

$$p_j \in C_k, \text{ if } p_j \in P_k \; j = 1, 2, \cdots, n_k \tag{7}$$

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 ρ and ∇:

$$p_j^{(k)} \in C_k, \text{ if } \rho_j \geq \bar{\rho} \text{ or } \nabla_j \geq \nabla_0 \; j = 1, 2, \cdots, n_k \tag{8}$$

where $p_j^{(k)}$ represents the j_{th} point in clump C_k, $\bar{\rho}$ is the average density of the clump C_k, ∇_0 is the hyper-parameter. The individual clump could be segregated as ρ_j greater than $\bar{\rho}$ or ∇_j greater than ∇_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:

$$\mathrm{C}_k = \begin{cases} \text{True, if } n_k \geq n_0 \\ \text{False, else} \end{cases} \tag{9}$$

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, $Peak_i$ and Cen_i represent the position of the clump peak intensity value and centroid on axis i ($i = 1, 2, 3$), respectively. $Size_i$ represents the size of the clump on axis i. Sum and $Peak$ represent the total flux and peak intensity value of the clump, respectively. The definition of the centroid is as follows:

$$Cen_i = \frac{\sum_{j=1}^{n_k}(I_j \cdot x_j)}{\sum_{j=1}^{n_k} I_j} \tag{10}$$

the $Size_i$ of the clump C_k on axis i is defined as:

$$Size_i = \sqrt{\frac{\sum_{j=1}^{n_k}(I_j \cdot x_j^2)}{\sum_{j=1}^{n_k} I_j} - \left(\frac{\sum_{j=1}^{n_k}(I_j \cdot x_j)}{\sum_{j=1}^{n_k} I_j}\right)^2} \tag{11}$$

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. δ_0 and ρ_0 are key hyper-parameters of the algorithm, where δ_0 represents the minimum distance between the centers of the two clumps, and ρ_0 represents the minimum peak intensity value of a candidate clump. n_0 is the minimum data point number of a clump and ∇_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	
Parameters	ρ_0	3σ
	δ_0	4
	∇_0	0.01
	n_0	27
	d_c	0.8

Table 2: The output of LDC algorithm

Output	Description	Explanation
Mask	Data array	The same shape as the input data
Parameters	$Peak_i$ ($i = 1, 2, 3$)	The position of the clump peak value on axis i
	Cen_i ($i = 1, 2, 3$)	The position of the clump centroid on axis i
	$Size_i$ ($i = 1, 2, 3$)	The size of the clump on axis i
	$Peak$	The peak value of the clump
	Sum	The total flux of the clump

The detection of the algorithm is not affected by neither the shape of the clumps nor the dimensionality of the space they embedded in. The detection results of the LDC in defferent number density and different PSNR are shown in the Table 3. The size of a simulated data is $100 \times 100 \times 100$.

Table 3: The performance of the algorithm

Number density levels	High	Medium	Low
Number of clumps in $100 \times 100 \times 100$ data array	100	25	10
Recall rate (PSNR ≥ 6)	$> 80\%$	$> 91\%$	$> 97\%$
Precision rate (PSNR ≥ 6)	$> 90\%$	$> 96\%$	$> 98\%$
F_1 (PSNR ≥ 6)	$> 86\%$	$> 94\%$	$> 98\%$

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 longtitude 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 longtitude - latitude plane. Therefore, the 3D Gaussian Distribution is described as:

$$\begin{aligned} f(x, y, v) = A \exp\Bigg\{ - \Bigg[& \left(\frac{\cos^2\theta}{2\sigma_x^2} + \frac{\sin^2\theta}{2\sigma_y^2}\right)(x - x_0)^2 \\ & + \left(\frac{\sin 2\theta}{2\sigma_y^2} - \frac{\sin 2\theta}{2\sigma_x^2}\right)(x - x_0)(y - y_0) \\ & + \left(\frac{\sin^2\theta}{2\sigma_x^2} + \frac{\cos^2\theta}{2\sigma_y^2}\right)(y - y_0)^2 + \frac{(v - v_0)^2}{2\sigma_v^2} \Bigg] \Bigg\} \end{aligned} \tag{12}$$

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 θ 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:

$$f_{mix}(x,y,v;\psi) = \sum_{k=1}^{K} f_k\left(x,y,v;\psi_k\right) \tag{13}$$

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

$$\psi_k = \left\{A^{(k)}, x_0^{(k)}, \sigma_x^{(k)}, y_0^{(k)}, \sigma_y^{(k)}, \theta^{(k)}, v_0^{(k)}, \sigma_v^{(k)}\right\}^T \tag{14}$$

Using the LDC algorithm described in Section 2.1, a series of clumps $(C_1, C_2, \cdots, C_k, \cdots, C_N)$ can be obtained, and the parameters of those clumps calculated in Section 2.1.6 could serve as the initial value $(\psi_1^{(0)}, \psi_2^{(0)}, \cdots, \psi_k^{(0)}, \cdots, \psi_N^{(0)})$ of the model. The clump C_i and C_j are considered to overlap each other when $\|(x_{i0}-x_{j0}, y_{i0}-y_{j0}, v_{i0}-v_{j0})\|_2 \leq \|(\sigma_{ix}+\sigma_{jx}, \sigma_{iy}+\sigma_{jy}, \sigma_{iv}+\sigma_{jv})\|_2$ ($\|\cdot\|_2$ represents two-norm). Then the parameters $(\psi_1^{(0)}, \psi_2^{(0)}, \cdots, \psi_k^{(0)}, \cdots, \psi_m^{(0)})$ 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.

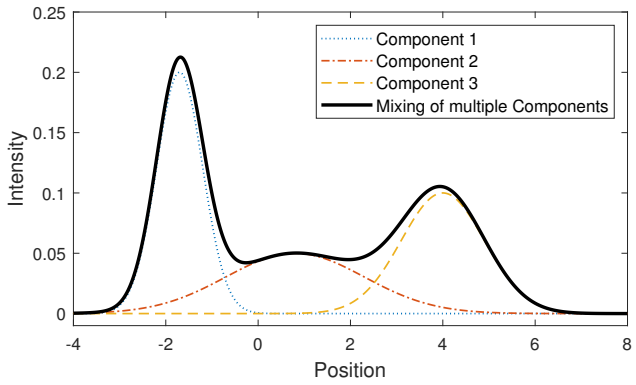

Fig. 2: Overlapping of three Gaussian components. The black line is composed of a combination of three gaussian distributions. Dashed lines represent each gaussian components.

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 $0°$ to $180°$. 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.

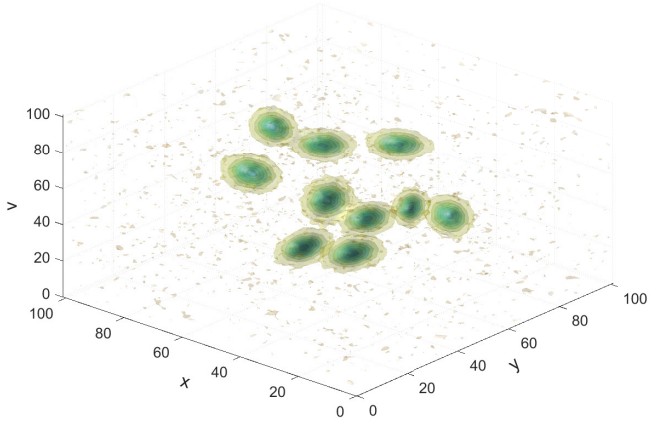

Fig. 3: 3D display of simulated clumps.

3.1.2 Detection based on LDC Algorithm

As shown in Figure 4, from left to right are the integral maps on the three planes of $x-y$, $x-v$ and $y-v$, respectively. The center points of clumps are marked with red asterisks on the integral graphas.

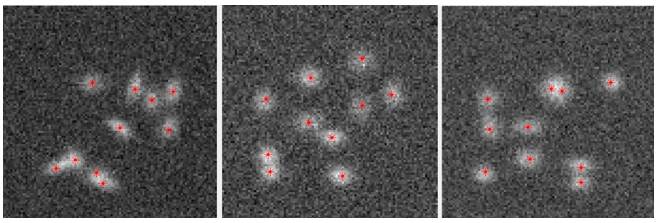

Fig. 4: The centers of detected clumps are marked on the intergrated intensity maps with red asterisks. From left to right are integral maps of $x-y$, $x-v$ and $y-v$ planes, respectively.

Combined with the ∇ and ρ of each data point, the members and region of the clump C_k could be determined by Formula (8). (1) Using ∇_0 as the threshold, the point set A_1 with ∇ greater than ∇_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 ρ 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.

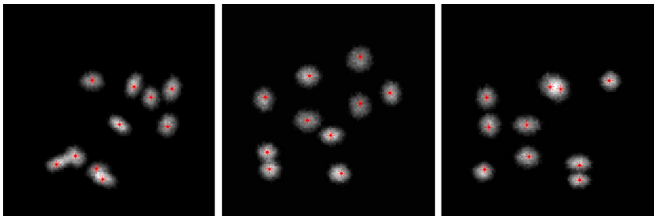

Fig. 5: The intergrated intensity maps of detected clumps are marked with red asterisks. From left to right are integral maps of the $x-y$, $x-v$ and $y-v$ planes, respectively.

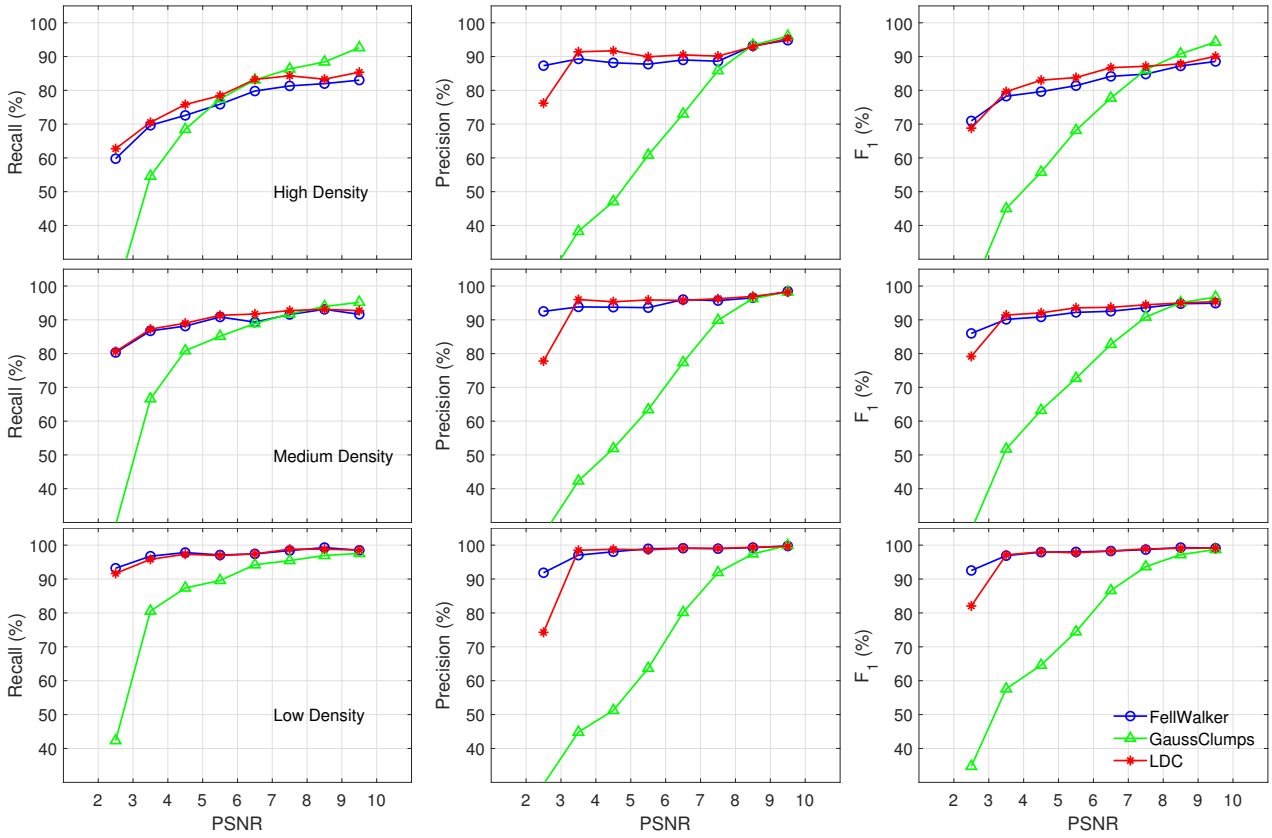

Fig. 6: The evaluation indicators R, P and F_1 of the three algorithms are plotted against the PSNR at different number density. *Top panel*: the detection statistics of the three algorithms in high density, from left to right are R, P and F_1, respectively. *Middle panel*: same as above but for medium density. *Bottom panel*: same as above but for low density.

3.1.3 Evaluation Indicators

The detection of molecular clumps is considered to be correct if the euclidean distance between the center of the detected clump and the center of the simulated is less than 2 pixels in the three axes.

Four statistics are obtained by the detection results as follows: True-Positive (TP), True-Negative (TN), False-Positive (FP), and False-Negative (FN) (Zhou et al., 2020). The evaluation indicators for the algorithm include: $recall\ rate$ (R), $precision\ rate$ (P) and comprehensive score (F_1). The R, P and F_1 are defined as:

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

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 algolrithms 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 7. While P of GaussClumps is greater than FellWalker and LDC in the case of the high PSNR, and R of GaussClumps is lower than the two algorithms in low PSNR. 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-

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 pararmeters 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:

$$E(\Delta X) = \frac{1}{N}\sum_{i=1}^{N}(X_i^{(s)} - X_i^{(m)}) \tag{16}$$

$$E(\Delta\theta) = \frac{1}{N}\sum_{i=1}^{N}(\theta_i^{(s)} - \theta_i^{(m)}) \tag{17}$$

$$E(\Delta S) = \frac{1}{N}\sum_{i=1}^{N}\frac{S_i^{(s)} - S_i^{(m)}}{S_i^{(s)}} \tag{18}$$

$$E(\Delta I) = \frac{1}{N}\sum_{i=1}^{N}\frac{I_i^{(s)} - I_i^{(m)}}{I_i^{(s)}} \tag{19}$$

$$E(\Delta F) = \frac{1}{N}\sum_{i=1}^{N}\frac{F_i^{(s)} - F_i^{(m)}}{F_i^{(s)}} \tag{20}$$

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, θ represents the tilt anlge 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.

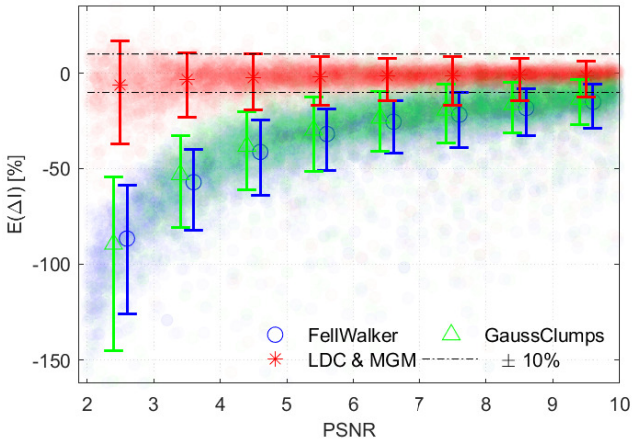

Fig. 7: The statistics of relative deviation in peak intensity by the three algorithms as a function of the PSNR. The blue, green and red dots show the distribution of the individual measurements. The Special symbols and error bars represent the median and standard deviation of accuracy, respectively. Two dashed horizontal lines represent the relative deviation of $\pm 10\%$.

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 simualted. 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.

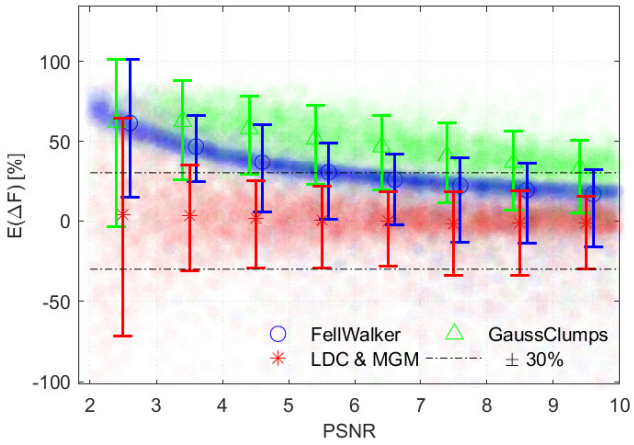

Fig. 8: The statistics of the relative deviation in the total flux as a function of the PSNR. Two dashed horizontal lines represent the relative deviation of $\pm 30\%$. The blue, green and red dots, special symbols and error bars have the same meaning as Figure 7.

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 ±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 10° when the PSNR greater than 4, indicating that the estimation of molecular clump angle by this algorithm is stable.

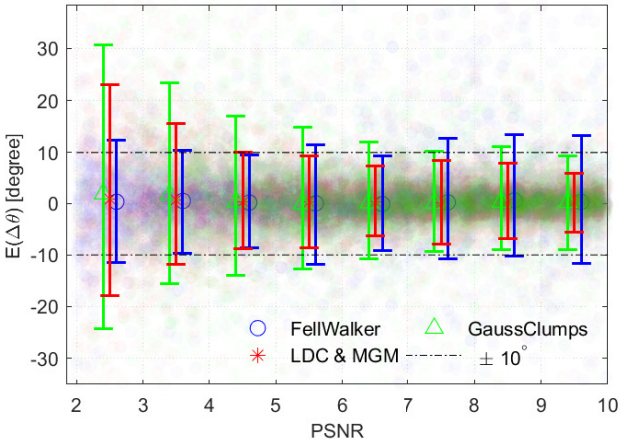

Fig. 9: The statistics of absolute deviation in the tilt angle as a function of the PSNR. The tilt angle on the $x-y$ plane of the molecular clump vary from 0° to 180°. The minimum ratio of the major axis to the minor axis in these clumps is 1.4. The blue, green and red dots, special symbols and error bars have the same meaning as Figure 7.

The size of the molecular clump can be used to describe the different shapes of them, which is a very important parameter for the classification of the molecular clump. From left to right, the panels of Figure 10 show the statistics relative deviation in $Size_1$, $Size_2$, and $Size_3$, respectively. In Figure 10, we can see that the size obtained by GaussClumps exhibit a large deviation. The measured size of GaussClumps and FellWalker algorithms are lower than the simualted size. With increase of the PSNR, the deviations from the GaussClumps and FellWalker algorithms gradually decrease, while the deviations from the LDC & MGM algorithm are closed to zeros. The deviation of LDC & MGM is less than 10% with the PSNR above 4, indicating that the size of clump obtained from the algorithm is reliable.

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 longititude, 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 longititude 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 parametration 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 intergrated 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.

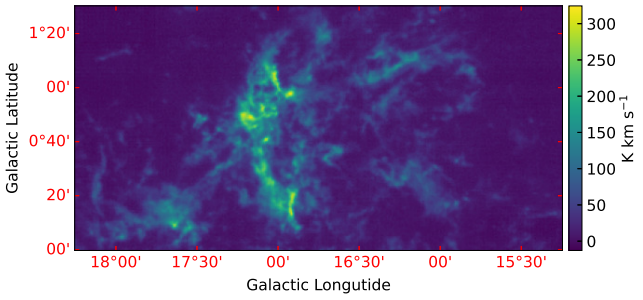

Fig. 12: The intergrated intensity maps of M16 in ^{13}CO (J=1-0) line with a volecity range of $15.93-27.06$ km s^{-1}.

4.2 Clump extraction experiment

After tuning the algorithm parameters, the GaussClumps, FellWalker and LDC 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 $2\times rms$ 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.

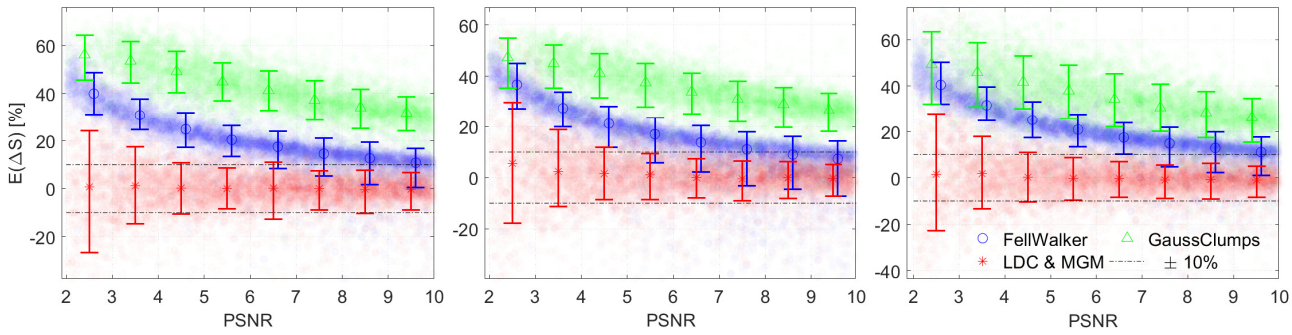

Fig. 10: The statistics of relative deviation in size as a function of the PSNR. From left to right are the deviation of $Size_1$, $Size_2$, and $Size_3$, respectively ($Size_1$, $Size_2$ represent major and minor size of detected clump in the spatial, $Size_3$ represents the size of detected clump in 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.

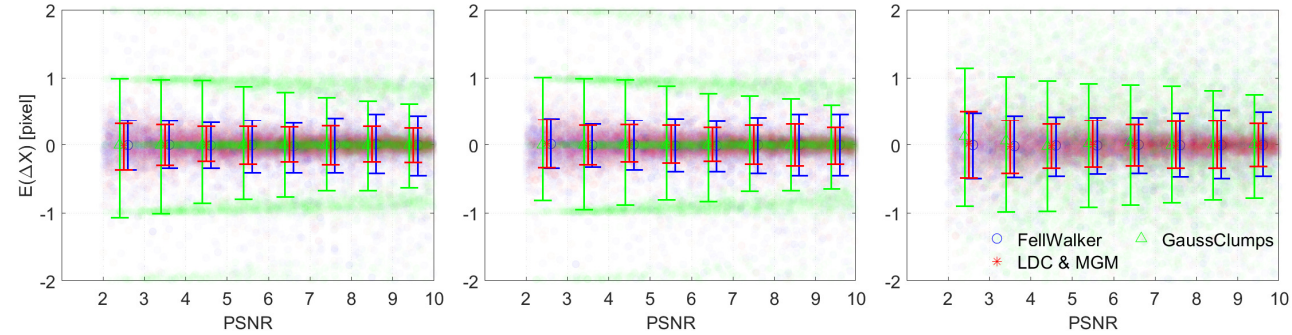

Fig. 11: The statistics of absolute deviation of the position as a function of the PSNR. From left to right are the deviation of galactic latitude, galactic longitude, and velocity, respectively. The blue, green and red dots, special symbols and error bars have the same meaning as Figure 7.

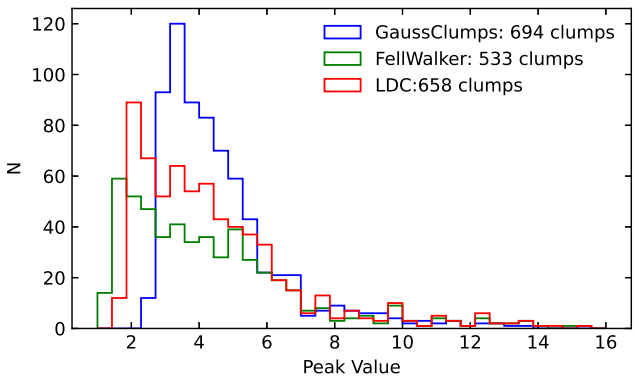

Fig. 13: The distribution of the detected peak intensity values of clumps by GaussClumps, FellWalker, and LDC.

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 flatted peak, while the distribution of the peak intensity values detected by GaussClumps deviates greatly from the other two algorithms. The peak of distributon is about 2 in the FellWalker and LDC algorithm, while the GaussClumps is for 3.4. Combined with the minimum peak intensity value (about 2.1 K) of clumps detected by GaussClumps and the noise level (0.23 K) at ^{13}CO (J=1-0) line, it shows that the PSNR of clumps detected by GaussClumps are greater than 9, while the recall rate of the algorithm in the Section 3.1.4 can be maintained a certain level with the PSNR above 5. It may be the Gaussclumps algorithm tends to fit a clump with a strict elliptic shape, and it fails to fit a clump with weaker peak intensity value in the real data.

4.3 Completeness

The limitation of the telescope sensitivity causes low quality clump being ignored. Other indicators of the algorithm are the completeness and the detection rate above the limition. The "completeness limit" here refers to the total flux or mass above which a clump can be detected at certain level with an algorithm. The smaller and weaker molecular clumps, the less likely they are to be detected.

We designed the dataset by randomly inserting simulated clumps into the ^{13}CO (J=1-0) line of M16. 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 longtitude 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, FellWlaker, 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.

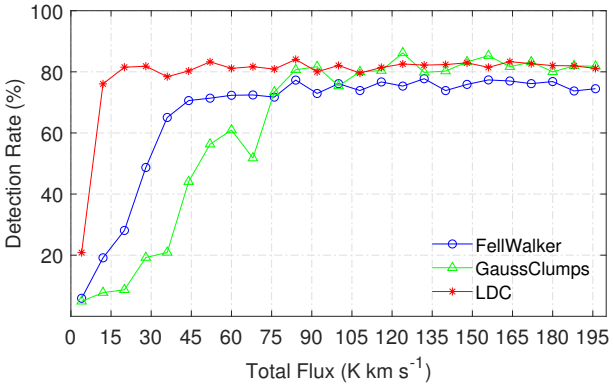

Fig. 14: The detection rate of the three algorithms in ^{13}CO (J=1-0) line of M16 as a function of total flux. The detection rate of GaussClumps, FellWalker and LDC are 80.9%, 74.7% and 81.7%, respectively. The completeness limitation of LDC and FellWalker are 20 K km s^{-1} and 45 K km s^{-1}, respectively.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

Figure 15 shows the statistical histogram of ΔI 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.

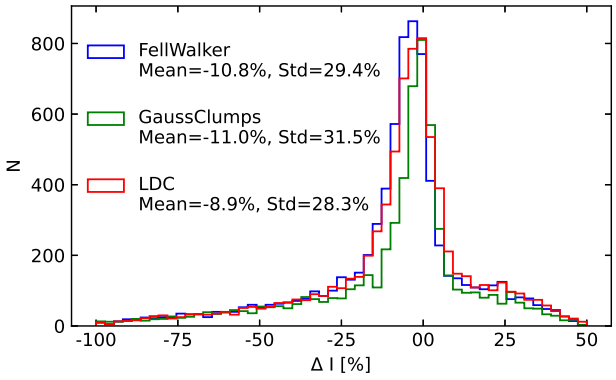

Fig. 15: The histogram of the peak deviation (ΔI) for the GaussClumps, FellWalker, and LDC. The ΔI is described in Formula (19). The mean deviation of FellWalker, GaussClumps and LDC are -10.8%, -11.0% and -8.9%, respectively. The standard deviation of FellWalker, GaussClumps and LDC are 29.4%, 31.5% 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 accurary 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 titl angle on the $x-y$ plane are less than 10° 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/C^{18}O along the northern galactic plane with PMO-13.7m telescope. We are grateful to all the members of the MWISP working group, particulally the staff mambers 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 1
Berry, D. S. 2015, Astronomy and Computing, 10, 22 1
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 1
Blitz, L., & Stark, A. A. 1986, ApJL, 300, L89 1
Carruthers, G. R. 1970, ApJL, 161, L81 1
Dame, T. M., Hartmann, D., & Thaddeus, P. 2001, ApJ, 547, 792 1
Dent, W. R. F., Hovey, G. J., Dewdney, P. E., et al. 2009, MNRAS, 395, 1805 1
Heiles, C., Li, D., McClure-Griffiths, N., Qian, L., & Liu, S. 2019, Research in Astron. Astrophys. (RAA), 19, 017 1
Heyer, M., & Dame, T. M. 2015, ARA&A, 53, 583 1
Kauffmann, J., Pillai, T., & Goldsmith, P. F. 2013, ApJ, 779, 185 1
Krumholz, M. R., & McKee, C. F. 2005, ApJ, 630, 250 1
Krumholz, M. R., McKee, C. F., & Tumlinson, J. 2009, ApJ, 699, 850 1
Lada, E. A. 1992, ApJL, 393, L25 1
Lee, Y., Stark, A. A., Kim, H.-G., & Moon, D.-S. 2001, ApJS, 136, 137 1
Li, C., Wang, H.-C., Wu, Y.-W., Ma, Y.-H., & Lin, L.-H. 2020, Research in Astron. Astrophys. (RAA), 20, 031 1

Lin, L.-H., Wang, H.-C., Su, Y., Li, C., & Yang, J. 2020, Research in Astron. Astrophys. (RAA), 20, 143 1
Lo, N., Cunningham, M. R., Jones, P. A., et al. 2009, MNRAS, 395, 1021 1
Ragan, S. E., Henning, T., Tackenberg, J., et al. 2014, A&A, 568, A73 7
Rivera-Ingraham, A., Ristorcelli, I., Juvela, M., et al. 2017, A&A, 601, A94 1
Rosolowsky, E. W., Pineda, J. E., Kauffmann, J., & Goodman, A. A. 2008, ApJ, 679, 1338 1
Sanders, D. B., Clemens, D. P., Scoville, N. Z., & Solomon, P. M. 1986, ApJS, 60, 1 1
Schneider, N., Stutzki, J., Winnewisser, G., Poglitsch, A., & Madden, S. 1998, A&A, 338, 262 1
Song, C., & Jiang, Z. B. 2017, A Census of Dense Clumps in the M16 Giant Molecular Cloud, Master's thesis, Shanghai Normal University, Shanghai 7
Stutzki, J., & Guesten, R. 1990, ApJ, 356, 513 1
Sugitani, K., Tamura, M., Nakajima, Y., et al. 2002, ApJL, 565, L25 1
Sun, J. X., Lu, D. R., Yang, J., et al. 2018, Acta Astronomica Sinica, 59, 3 7
Tang, M.-Y., Qin, S.-L., Liu, T., & Wu, Y.-F. 2019, Research in Astron. Astrophys. (RAA), 19, 040 1
Williams, J. P., Blitz, L., & Mckee, C. F. 2012, Physics, 97 1
Williams, J. P., de Geus, E. J., & Blitz, L. 1994, ApJ, 428, 693 1
Wilson, R. W., Jefferts, K. B., & Penzias, A. A. 1970, ApJL, 161, L43 1
Zhan, X.-L., Jiang, Z.-B., Chen, Z.-W., Zhang, M.-M., & Song, C. 2016, Research in Astron. Astrophys. (RAA), 16, 56 1, 7
Zhou, P., Xiaoyu, L., Sheng, Z., Zhibo, J., & Shuguang, Z. 2020, ACTA ASTRONOMICA SINICA, 61, 14 5
Zinnecker, H., & Yorke, H. W. 2007, ARA&A, 45, 481 1
Zuo, Y.-X., Li, Y., Sun, J.-X., et al. 2011, Chinese Astronomy and Astrophysics, 35, 439 1