

A new method to unveil embedded stellar clusters

Marco Lombardi¹, Charles J. Lada², and João Alves³

¹ University of Milan, Department of Physics, via Celoria 16, 20133 Milan, Italy
e-mail: marco.lombardi@unimi.it

² Harvard-Smithsonian Center for Astrophysics, Mail Stop 72, 60 Garden Street, Cambridge, MA 02138, USA

³ University of Vienna, Türkenschanzstrasse 17, 1180 Vienna, Austria

Received 5 December 2016 / Accepted 25 July 2017

ABSTRACT

In this paper we present a novel method to identify and characterize stellar clusters deeply embedded in a dark molecular cloud. The method is based on measuring stellar surface density in wide-field infrared images using star counting techniques. It takes advantage of the differing *H*-band luminosity functions (HLFs) of field stars and young stellar populations and is able to statistically associate each star in an image as a member of either the background stellar population or a young stellar population projected on or near the cloud. Moreover, the technique corrects for the effects of differential extinction toward each individual star. We have tested this method against simulations as well as observations. In particular, we have applied the method to 2MASS point sources observed in the Orion A and B complexes, and the results obtained compare very well with those obtained from deep *Spitzer* and *Chandra* observations where presence of infrared excess or X-ray emission directly determines membership status for every star. Additionally, our method also identifies unobscured clusters and a low resolution version of the Orion stellar surface density map shows clearly the relatively unobscured and diffuse OB 1a and 1b sub-groups and provides useful insights on their spatial distribution.

Key words. methods: statistical – ISM: clouds – open clusters and associations: general

1. Introduction

Embedded star clusters offer one of the best opportunities to understand star formation. Within these astrophysical laboratories hundreds of stars are formed in volumes below 1 pc^3 . Overall, it is estimated that 80–90% of young stellar objects (YSOs) form in embedded clusters (Lada & Lada 2003; although the exact fraction is somewhat sensitive to the definition of a cluster, see Bressert et al. 2010). It is now established that the fate of these objects is directly linked to the evolution of the molecular gas, which is responsible for most of the gravitational potential that binds the stars in the cluster. Therefore, it is important to study these objects in their early phases, before they are subject to infant mortality (Lada et al. 1984; Geyer & Burkert 2001).

Operationally, clusters are often defined and identified as groups of stars whose stellar surface density exceeds that of field stars of the same physical type (see, e.g., Wilking & Lada 1985; Lada et al. 1991; Carpenter 2000)¹. In this respect, embedded clusters pose special problems because they are buried in dust and gas, and therefore are often not even visible in optical images; moreover, their shape is often elongated and clumpy, reflecting the initial structure of the dense molecular gas (Gutermuth et al. 2005; Megeath et al. 2016). However, even with infrared observations, discovering deeply embedded clusters and measuring their basic parameters (such as surface density and size) can still be a challenge since the associated dust extinguishes both the cluster members and the field stars behind the cloud (see, e.g., Román-Zúñiga et al. 2008). In fact, typical optical or near-infrared observations stellar fields around molecular

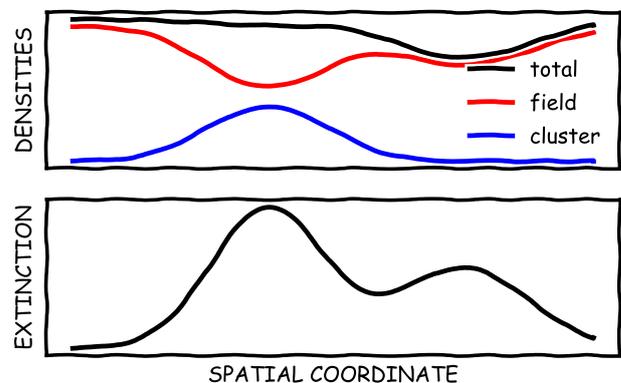


Fig. 1. A one-dimensional sketch figure that illustrates the difficulties encountered in detecting embedded clusters. The bottom black line shows the extinction profile of a molecular cloud. As a result of the extinction, background stars in the field show an underdensity. Even if an embedded cluster is present, the total observed star surface density does not reveal it, making it virtually impossible to detect the cluster without the use of the technique described in this paper.

clouds show underdensities at the location of the clouds because of the effects of extinction on the density of background stars. In such a situation the observed cluster stellar surface density can be comparable or even less than the unobscured field stellar surface density (see Fig. 1).

Different authors have used different techniques to take into account the effects of extinction in the identification and characterization of star clusters. Carpenter (2000) built density maps in the direction of nearby molecular clouds using the 2MASS Point Source Catalog, and corrected the effect of dust extinction using publicly available CO maps (converted into infrared extinction with a constant X-factor). In his technique, the extinction

¹ In this paper we focus on the optical identification of a cluster, and we explicitly ignore issues such as the gravitational stability of overdensities of stars. Therefore, we will call “cluster” any overdensity, irrespective of its boundness.

correction is only applied to field stars: the cluster stellar density is obtained by subtracting an extinction-corrected model of the field stellar density to the observed stellar density. Moreover, the use of CO maps and the largely uncertain X-factor represent a primary source of error.

An opposite approach has been adopted by [Cambrésy et al. \(2013\)](#), who studied star clusters embedded in the Rosette nebula using 2MASS data. In this case, the local extinction was determined directly from the stellar near-infrared colors, and a correction for the effects of extinction was applied to each star. In practice, as noted by the authors themselves, the correction applied would only be correct for background stars, while it is used for all stars (including the foreground, which should not be corrected, and embedded ones, which should be partially corrected). Things are further complicated because molecular clouds are known to have steep gradients in their column densities, and these, if undetected (because of resolution effects, which is surely the case of the Rosette cloud at the 2MASS resolution), would introduce further biases in the extinction map and in the correction to the cluster richness.

[Cambrésy et al. \(2006\)](#) use yet another technique (similar to [Cambrésy et al. 2002](#)) to study the properties of IC 348 in Perseus. The basic idea is that two independent measurements of extinction in molecular clouds are available, one from the star color excess (here applied to $H - K$ color), and one from the star number counts. In absence of contaminants, both measurements should agree. The presence of a cluster, however, will affect both the color excess measurement (in a way that depends on the intrinsic color of the cluster members) and the star count method (in a way that depends on the location of the cluster within the cloud). The difference of the two extinction estimates is a proxy for the cluster density. The various assumptions made when applying this technique (in particular, the degree of embedding of the cluster within the cloud) need to be resolved by calibrating it using independent measurements of cluster richness: this clearly limits its application.

In this paper we present a new methodology to identify and characterize clusters in wide-field or all sky, multi-band, infrared imaging surveys. The method is based on the production of extinction-corrected maps of stellar surface density, and takes advantage of the different H -band luminosity functions of embedded clusters with respect to that of the background population. Additionally, in contrast to the methods described above, it is able to correct for cluster members unidentified because of extinction in a way that takes into account the position of the cluster within the cloud along the line of sight. The technique is based on a rigorous mathematical framework; this provides a clear advantage, in that we can perform a detailed statistical analysis of the method. But the detailed mathematical derivation of this method might not be of as much interest to astronomers as its implementation. However, we stress that those readers not interested in the detailed mathematical aspects of the derivation can still benefit from the method proposed here, because its application is simple and straightforward: Eq. (30), with optionally a second expression for the noise estimate, Eq. (31). We also provide a pseudo-code for it in the appendix.

This paper is organized as follows. In Sect. 2 we present the general framework and discuss the standard technique generally employed to identify clusters. In Sect. 3 we present our new method and we discuss its practical implementation in Sect. 4. Simple numerical tests used to validate the method are described in Sect. 5, while the results of an application to the Orion molecular complex are discussed in Sect. 6. Finally, we summarize the results obtained in this paper in Sect. 7.

2. The standard technique

Traditionally, star clusters are identified as overdensities in the distribution of stars. Although many different options are available to estimate the local angular density of stars σ (see, e.g., [von Hoerner 1963](#); [Casertano & Hut 1985](#); [Gutermuth et al. 2009](#)), we consider in this paper the simple “moving average” estimate²

$$\hat{\sigma}(\mathbf{x}) = \sum_{n=1}^N W(\mathbf{x} - \mathbf{x}_n). \quad (1)$$

Here n runs on the individual stars, $\{\mathbf{x}_n\}$ are the locations of the various stars, and W is a window function, normalized to unity:

$$\int W(\mathbf{x}') d^2x' = 1. \quad (2)$$

So, by construction, W has a unit of one over area. As a simple example, suppose that W is taken to be a top-hat function:

$$W(\mathbf{x}') = \begin{cases} 1/(\pi r^2) & \text{if } |\mathbf{x}'| < r, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

In this case, the sum of Eq. (1) runs just over the N_r stars within an angular distance r from the point of the estimate, and the observed estimates reduces to

$$\hat{\sigma} = \frac{N_r}{\pi r^2}. \quad (4)$$

For a constant density of stars, this quantity is unbiased, in the sense that the mean (ensemble average) E of $\hat{\sigma}$ is just the true density of stars:

$$E[\hat{\sigma}] \equiv \langle \hat{\sigma} \rangle = \frac{E[N_r]}{\pi r^2} = \sigma, \quad (5)$$

and the associated variance is

$$\text{Var}[\hat{\sigma}] = \frac{E[N_r]}{(\pi r^2)^2} = \frac{\sigma}{\pi r^2}. \quad (6)$$

This equation shows that the error associated with the measured star density decreases as $N_r^{-1/2}$: therefore, if it is known that the density of stars is constant, to determine its value it is sensible to use relatively large window functions W .

More generally, if the star density is variable, the average measured density can be shown to be a convolution of the true density with the window function W :

$$E[\hat{\sigma}](\mathbf{x}) = \int W(\mathbf{x} - \mathbf{x}') \sigma(\mathbf{x}') d^2x', \quad (7)$$

and the associated two point correlation function is

$$\begin{aligned} \text{Cov}[\hat{\sigma}](\mathbf{x}, \mathbf{x}') &\equiv E[(\hat{\sigma}(\mathbf{x}) - E[\hat{\sigma}](\mathbf{x}))(\hat{\sigma}(\mathbf{x}') - E[\hat{\sigma}](\mathbf{x}'))] \\ &= \int W(\mathbf{x} - \mathbf{x}'') W(\mathbf{x}' - \mathbf{x}'') \sigma(\mathbf{x}'') d^2x''. \end{aligned} \quad (8)$$

Equation (7) shows that W sets the scale for the resolution of the density map; similarly, Eq. (8) shows that points close enough in the density map will have correlated noise. Therefore, if one aims at finding changes in the star density (such as a star cluster),

² Throughout this paper we will use “hats” to denote measured quantities. Hence, $\sigma(\mathbf{x})$ is the *true* star density at the angular position \mathbf{x} , while $\hat{\sigma}(\mathbf{x})$ is the *measured* star density at the same position.

the window function should have a scale that is comparable or smaller than the typical size of the variations of the star density. However, this is in tension with the noise properties of Eq. (8), because a small window function implies a large noise.

In order to be able to detect a star cluster, the measured density of stars at the location of the cluster must differ from the average density much more than the standard deviation of $\hat{\sigma}$. Hence, the quantity $\text{Cov}[\hat{\sigma}](\mathbf{x}, \mathbf{x})$ sets the sensitivity of the cluster finding algorithm. In general the true density $\sigma = \sigma_{\text{field}} + \sigma_{\text{cluster}}$ is the sum of the field density and of the cluster density, and in some cases $\sigma_{\text{cluster}} \ll \sigma_{\text{field}}$. In these conditions the error is dominated by the shot-noise due to the field star population. In reality, many other effects can prevent the discovery and characterization of star clusters:

- extinction by dark nebulae, which reduces the surface density of background sources;
- the galactic structure, which induces smooth large-scale variations;
- differences in the sensitivity across a mosaic image due to changes in the observational conditions;
- other systematical observational effects, such halos produced by bright objects within the image.

Extinction correction. Among the effects listed above, the first one is particularly important for young clusters, since these objects tend to be deeply embedded and thus can escape a detection; additionally, detected clusters are plagued by large uncertainties in their physical properties (number of stars, mass, and size). For this reason, [Cambrésy et al. \(2013\)](#) developed a technique to perform a simple correction of the results produced by Eq. (1). They noted that the density of background stars observed through a molecular cloud decreases by a factor $10^{-\alpha A}$, where α is the exponential slope of the number counts and A is the extinction, both in the band used for the observation. Therefore, to account for the undetected stars one can just multiply the local density estimate $\hat{\sigma}(\mathbf{x})$ by $10^{\alpha A(\mathbf{x})}$, where $A(\mathbf{x})$ is the local estimate extinction (i.e., the extinction derived from an extinction map at the location of the \mathbf{x}).

The problem with this approach is that it uses the same correction factor for foreground stars, embedded objects, and background stars, which generally results in an overestimate of the local corrected star density. Additionally, the same correction is applied to YSOs and field stars, which however have number count distributions very different. This leaves a large uncertainty on the corrected $\hat{\sigma}(\mathbf{x})$ and, ultimately, on the characterization of each cluster.

3. Maximum-likelihood approach

3.1. Constant extinction, no weighting

As discussed in the previous section, the error associated to $\hat{\sigma}$ is often dominated by shot noise due to the field star population. However, as mentioned, YSOs have photometric properties that differ significantly from those of field stars: the former have a H -band luminosity function (hereafter HLF) that can be approximated by a Gaussian distribution ([Muench et al. 2002](#)), while the latter have a HLF that follows, up to $H \sim 18$ mag, an exponential with slope $\alpha \sim 0.33$ (see, e.g., [Meingast et al. 2016](#)). This suggests that we might use suitable cuts in the H -band to reduce the number of field stars, while essentially retaining all YSOs, thus gaining in terms of noise. However, this naive procedure is difficult to apply, since both YSOs and field stars are

also affected by extinction, which would change the shape and the peaks of the respective HLFs. Therefore, we are forced to use a more systematic approach.

Let us first model the observed HLF for a set of angularly close stars (i.e., in a patch of the sky). We suppose that the true, unextinguished HLF can be described as a mixture of L different HLFs, corresponding to different stellar populations (in the situation considered later on in this paper we will use just two components, one for the field stars and one for the YSOs, but deeper observations might require the inclusion of a third component corresponding to galaxies and following exponential number counts with a slope of 0.6; see [Hubble 1926](#)). The observed HLF differs from the true one because of extinction and because of incompleteness in the faint end; additionally, we will also have photometric errors, but generally these will be small compared to the typical width of the various HLF components, and for the sake of simplicity are therefore neglected here³. We can thus write finally the observed HLF, measured in units of stars per area per magnitude bin, as

$$\sigma(m) = c(m) \sum_{i=1}^L \sigma_i p_i(m - A), \quad (9)$$

where A is the H -band extinction⁴, $c(m)$ is the completeness function (i.e. the probability to detect a star with apparent magnitude m), $p_i(m)$ is the probability distribution for the i th component of the HLF, and σ_i is a coefficient that indicates the predominance of the i th component in the HLF. In order to give σ_i a simple interpretation, it is useful to take p_i suitably normalized. In particular, we assume that

$$\int c(m) p_i(m) dm = 1. \quad (10)$$

With this choice, σ_i can be identified as the observed angular density of stars for the i th component in absence of extinction.

Our goal is to infer the parameters $\{\sigma_i\}$ from a set of observed H -band magnitudes $\{m_n\}$ in a patch of the sky of area S ; we will then repeat this operation in different directions in order to build maps of densities for the various components. This will allow us to identify overdensities in each component, such as in the YSO one. To this purpose, we use Eq. (9) to write the log-likelihood following the prescriptions of [Lombardi et al. \(2013\)](#). The entire analysis is done in the magnitude space, by using $S\sigma$ as the probability distribution density for the magnitudes:

$$\begin{aligned} \ln \mathcal{L} &= \sum_{n=1}^N \ln S \sigma(m_n) - \int S \sigma(m) dm \\ &= \sum_{n=1}^N \left[\ln S c(m_n) \sum_{i=1}^L \sigma_i p_i(m_n - A) \right] \\ &\quad - \int S c(m) \sum_{i=1}^L \sigma_i p_i(m - A) dm. \end{aligned} \quad (11)$$

³ Photometric errors can be easily included in Eq. (9) by replacing $p_i(m - A)$ there with the convolution of this function with the photometric uncertainty. In presence of different uncertainties for each star, one will need to specialize Eq. (9) to each star. A similar technique can be used to include the effects of errors on the extinction measurements.

⁴ In this section we will take the extinction to be identical for a set of angularly close stars; we will then relax this assumption and use the individual extinction in the direction of each star.

This likelihood can be used in Bayes' theorem to infer a posterior probability for the densities $\{\sigma_i\}$. More simply, we can just find the most likely parameters $\{\sigma_i\}$ using a maximum likelihood approach. To this purpose, we consider

$$\frac{\partial \ln \mathcal{L}}{\partial \sigma_i} = \sum_{n=1}^N \frac{c(m_n) p_i(m_n - A)}{\sigma(m_n)} - S \int c(m) p_i(m - A) dm. \quad (12)$$

The maximum likelihood solution is given by the set of densities σ_i that maximize \mathcal{L} or, equivalently, $\ln \mathcal{L}$, i.e. by the zero of Eq. (12); that is, we need to solve

$$\sum_{n=1}^N \frac{c(m_n) p_i(m_n - A)}{\sigma(m_n)} = S \int c(m) p_i(m - A) dm. \quad (13)$$

Unfortunately, the solution of this equation for $L > 1$ cannot be provided in analytic form, and must be obtained by numerical methods. We will discuss below in Sect. 3.3 simple ways to find it.

We can obtain an estimate of the errors associated to the measurements σ_i from the Fisher information matrix (see, e.g., [Feigelson & Babu 2012](#)), that we recall is defined as

$$I_{ij} = E \left[\frac{\partial \ln \mathcal{L}}{\partial \sigma_i} \frac{\partial \ln \mathcal{L}}{\partial \sigma_j} \right] = -E \left[\frac{\partial^2 \ln \mathcal{L}}{\partial \sigma_i \partial \sigma_j} \right]. \quad (14)$$

The Fisher information matrix is related to the minimum covariance matrix that can be attained by an unbiased estimator, as provided by the Cramér-Rao bound:

$$\text{Cov}[\sigma] \geq I^{-1}. \quad (15)$$

Since the maximum-likelihood estimator is asymptotically efficient (i.e. it attains the Cramér-Rao bound when the sample size tends to infinity) and the resulting errors on $\hat{\sigma}$ tend to a multivariate Gaussian distribution, it is interesting to obtain an analytic result for the information matrix.

The Fisher information matrix can be readily evaluated from our data using Eq. (12) of [Lombardi et al. \(2013\)](#):

$$I_{ij} = S \int \frac{c^2(m) p_i(m - A) p_j(m - A)}{\sigma(m)} dm. \quad (16)$$

This relation is interesting from several points of view. First, note that the Fisher matrix contains elements outside the diagonal, unless the probability distributions for the various components do not overlap, i.e. have non-intersecting support: this would basically mean that we could immediately tell to which component belongs a star from its unextinguished magnitude. Second, note that all elements of the matrix are non-negative and therefore, in case of $L = 2$ components, $\text{Cov}[\sigma] \approx I^{-1}$ will have non-positive elements off-diagonal: in other words, the measurements of the two densities σ_1 and σ_2 will be affected by a negative correlation. This is expected and corresponds to a classification error: if p_1 overlaps with p_2 , we cannot uniquely associate stars to each different population and in general an overestimate of one population is done at the expenses of an underestimate of the other population.

It is instructive to consider the form of the information matrix in the special case where $L = 1$. When just one component is used, then I is a scalar and it reduces to

$$I = \frac{S}{\sigma} \int c(m) p(m - A) dm = \frac{S^2}{E[N]}, \quad (17)$$

where $E[N] = S\sigma$ is the average number of stars observed in the area S . Its inverse, I^{-1} , is therefore $E[N]/\sigma^2$, as expected from a simple Poisson statistics.

With $L \geq 2$, in principle we can encounter cases where the Fisher information matrix is singular. Given the analytic form of I , this happens in particular when the two components have the exact same probability distributions within the support of $c(m)$, i.e. $c(m) p_i(m) = c(m) p_j(m)$: in this case, the corresponding rows (or columns) of I are identical. In such a situation, clearly, it is virtually impossible to classify objects as one of the two degenerate components, and therefore the uncertainty on the respective densities σ_i and σ_j are infinite.

For completeness, we also report the expected maximum value of the log-likelihood

$$E[\mathcal{L}] = \frac{L}{2} + S \int \sigma(m) [\ln S \sigma(m) - 1] dm, \quad (18)$$

and the associated variance

$$\text{Var}[\mathcal{L}] = S \int \sigma(m) \ln^2 S \sigma(m) dm. \quad (19)$$

These equations can be used to verify that the chosen model (9) can account well for the data.

3.2. Differential extinction and spatial weight

So far, we have assumed that all stars are subject to the same extinction A ; moreover, we have not weighted stars depending on their angular position as done in Eq. (1). In this section we intend to remove these two limitations and consider the full problem.

The simpler approach to include the effects of differential extinction is to consider the *joint* density in magnitudes and positions. In this framework, we can rewrite Eq. (11) for log-likelihood as

$$\ln \mathcal{L} = \sum_{n=0}^N \ln \sigma(m_n, \mathbf{x}_n) - \int dm \int d^2 x' \sigma(m, \mathbf{x}'). \quad (20)$$

In this equation the quantity $\sigma(m, \mathbf{x}')$ represents the predicted density of stars with magnitude m at the location \mathbf{x}' . Similarly to Eq. (9), we write this quantity as a mixture of different densities, corresponding to different stellar populations:

$$\sigma(m, \mathbf{x}') = c(m) \sum_{i=1}^L \sigma_i(\mathbf{x}') p_i(m - A(\mathbf{x}')), \quad (21)$$

where $\sigma_i(\mathbf{x}')$ represents the density of stars of class i at the sky position \mathbf{x}' . In order to proceed, we need to model these densities. A simple and consistent way of doing this, is to suppose that $\ln \sigma_i(\mathbf{x}')$ can be written as the weighted sum of two terms: one, $\ln \sigma_i(\mathbf{x})$, associated to the density at the point of interest \mathbf{x} (the point where we intend to evaluate the local densities of stellar populations); and one, $\ln \tau_i(\mathbf{x}')$, which parametrizes the local changes of the densities. As a result, we write

$$\sigma_i(\mathbf{x}') = (\sigma_i(\mathbf{x}))^{\omega(\mathbf{x}-\mathbf{x}')} (\tau_i(\mathbf{x}'))^{1-\omega(\mathbf{x}-\mathbf{x}')}. \quad (22)$$

The function ω describes the spatial correlation between densities at different positions and plays a central role in identifying which stars contribute to the density estimate at \mathbf{x} .

We can now insert Eqs. (22) and (21) in Eq. (20) and find the maximum likelihood solution over $\sigma_i(\mathbf{x})$. Calling $A_n \equiv A(\mathbf{x}_n)$ the extinction suffered by the n th star, we find

$$\frac{\partial \ln \mathcal{L}}{\partial \sigma_i(\mathbf{x})} = \sum_{n=1}^N \omega(\mathbf{x} - \mathbf{x}_n) \frac{p_i(m_n - A_n) \sigma_i(\mathbf{x}_n) / \sigma_i(\mathbf{x})}{\sum_j \sigma_j(\mathbf{x}_n) p_j(m_n - A_n)} - \int dm c(m) \int d^2 x' \omega(\mathbf{x}') p_i(m - A(\mathbf{x}')) \sigma_i(\mathbf{x}') / \sigma_i(\mathbf{x}). \quad (23)$$

We now make an important assumption. Because of the form of the parametrization (22), a solution for the maximum likelihood can only be found if we specify the functional form of the functions $\tau_i(\mathbf{x}')$. However, these functions are truly unknown, since they parametrize the local changes of the various densities, which in turn depend on the local density map. Using a statistical approach, however, it is natural to assume that these functions have the same distribution of $\sigma_i(\mathbf{x})$, the quantity we are interested in. As a result, if we take an average of Eq. (23), terms such as $\sigma_i(\mathbf{x}') / \sigma_i(\mathbf{x})$ cancel out:

$$\frac{\partial \ln \mathcal{L}}{\partial \sigma_i(\mathbf{x})} = \sum_{n=1}^N \omega(\mathbf{x} - \mathbf{x}_n) \frac{p_i(m_n - A_n)}{\sum_j \sigma_j(\mathbf{x}_n) p_j(m_n - A_n)} - \int dm c(m) \int d^2 x' \omega(\mathbf{x}') p_i(m - A(\mathbf{x}')). \quad (24)$$

Before proceeding, it is useful to consider the solution of the maximum likelihood approach in the simple case where there is a single population of stars and where the extinction vanishes, $A(\mathbf{x}') = 0$. We find in this case

$$\frac{\partial \ln \mathcal{L}}{\partial \sigma(\mathbf{x})} = \sum_{n=1}^N \frac{\omega(\mathbf{x} - \mathbf{x}_n)}{\sigma(\mathbf{x})} - \int d^2 x' \omega(\mathbf{x}'), \quad (25)$$

where we have used the normalization property (10). The solution of this equation is immediately found as

$$\sigma(\mathbf{x}) = \sum_{n=1}^N W(\mathbf{x} - \mathbf{x}_n), \quad (26)$$

where

$$W(\mathbf{x}) = \frac{\omega(\mathbf{x})}{\int \omega(\mathbf{x}') d^2 x'} \quad (27)$$

We have therefore recovered Eq. (1), with the correct normalization (2) for the weight W .

In the general case, the maximum-likelihood solution of Eq. (24) must be obtained numerically. The errors associated to the solutions can be estimated using the Fisher matrix. In our case, we can obtain the Fisher matrix from

$$\begin{aligned} I_{ij}(\mathbf{x}) &= \iint dm d^2 x' \frac{1}{\sigma(m, \mathbf{x}')} \frac{\partial \sigma(m, \mathbf{x}')}{\partial \sigma_i(\mathbf{x})} \frac{\partial \sigma(m, \mathbf{x}')}{\partial \sigma_j(\mathbf{x})} \\ &= \iint dm d^2 x' c^2(m) \omega^2(\mathbf{x} - \mathbf{x}') \\ &\quad \times \frac{p_i(m - A(\mathbf{x}')) p_j(m - A(\mathbf{x}'))}{\sigma(m, \mathbf{x}')} \frac{\sigma_i(\mathbf{x}') \sigma_j(\mathbf{x}')}{\sigma_i(\mathbf{x}) \sigma_j(\mathbf{x})}. \end{aligned} \quad (28)$$

As before, we can replace \mathbf{x} to \mathbf{x}' in all arguments of σ_i , with the justification that the statistical properties of σ are invariant upon translation. We will discuss in the next section useful expressions to evaluate this quantity in practical cases.

3.3. Implementation

The algorithm proposed in this paper is essentially the search of the solutions of Eq. (24) as a function of the star densities $\{\sigma_i\}$ corresponding to the various components or star populations. The same procedure must be applied to different patches of the sky, so that maps of the star densities can be obtained. These, in turn, will allow us to identify and characterize star clusters, and in particular embedded ones.

Although the practical implementation of the algorithm follows this schema, a number of technical and theoretical aspects must be correctly addressed in order to optimize the detection and make the technique efficient.

First, we note that a simple way to obtain the (positive) solutions of Eq. (24) is through the use of a recursive formula. To this purpose, multiply both members of this equation by $\sigma_i(\mathbf{x})$, and solve for this quantity, thus obtaining the expression

$$\sigma_i(\mathbf{x}) \leftarrow \frac{\sum_{n=1}^N \omega(\mathbf{x} - \mathbf{x}_n) \frac{\sigma_i(\mathbf{x}_n) p_i(m_n - A_n)}{\sum_{j=1}^L \sigma_j(\mathbf{x}_n) p_j(m_n - A_n)}}{\int dm c(m) \int d^2 x' \omega(\mathbf{x}') p_i(m - A(\mathbf{x}'))}. \quad (29)$$

Unfortunately, we are unable to use this equation because we only know the extinction at the locations of the stars $A_n \equiv A(\mathbf{x}_n)$: this prevents us from evaluating the integral over $d^2 x'$ in the denominator. We can, however, move the denominator inside the sum, and evaluate the second integral by replacing $A(\mathbf{x})$ with $A(\mathbf{x}_n)$, the extinction at the direction of each star. Additionally, using the same argument that has been employed in Eq. (23), that is the similarity between σ and τ , it is convenient to replace $\sigma_i(\mathbf{x})$ with $\sigma_i(\mathbf{x}_n)$ in the numerator. This procedure leads to the iteration

$$\sigma_i(\mathbf{x}) \leftarrow \sum_{n=1}^N \frac{W(\mathbf{x} - \mathbf{x}_n) \frac{\sigma_i(\mathbf{x}_n) p_i(m_n - A_n)}{\sum_{j=1}^L \sigma_j(\mathbf{x}_n) p_j(m_n - A_n)}}{\int dm c(m) p_i(m - A_n)}. \quad (30)$$

Equation (30) is the solution proposed in this paper to estimate the local density of stars. As indicated by the left arrow symbol, we can obtain the set of values $\{\sigma_i\}$ by starting with some arbitrary (positive) values for these quantities, and then by calculating updated values of σ_i by applying Eq. (30). The convergence is usually reached within a few tens of iterations.

Note that Eq. (30) has a simple interpretation. Let us ignore for a moment the weight W , i.e. let us assume that all stars have the same weight. The sum in Eq. (30) is carried out over all stars in the patch of the sky, but each star is counted only partially (i.e., contributes with a term between zero and unity in the sum): precisely, each star contributes by the computed probability that the star be associated with the i th component. The way this probability is computed is actually a simple application of Bayes' theorem, where $p_i(m_n - A_n)$ plays the role of the likelihood, $\sigma_i(\mathbf{x}_n)$ is proportional to the prior that the star is of class i , and the sum over j in the denominator is proportional to the evidence. The result of the sum of these terms is divided by the by the result of the integral: this is a correcting factor that takes into account the fact that, because of extinction and incompleteness, we will miss a fraction of stars. Note also that Eq. (30) can be also considered as a special case of a K -means soft clustering algorithm where the only unknown quantities are the classes σ_i (see MacKay 2003).

Before proceeding, it is useful to recall the hypotheses of this algorithm and its strengths. First, we assume that we have some

Table 1. Different two-dimensional spatial functions $\omega(\mathbf{x})$ and corresponding weight functions $W(\mathbf{x})$.

Function	Support	$\omega(\mathbf{x})$	$W(\mathbf{x})$	Eff. area
Top-hat	$ \mathbf{x} \leq s$	1	$1/(\pi s^2)$	πs^2
Conic	$ \mathbf{x} \leq s$	$2(1 - \mathbf{x} /s)$	$3(1 - \mathbf{x} /s)/(\pi s^2)$	$2\pi s^2/3$
Parabolic	$ \mathbf{x} \leq s$	$3(1 - \mathbf{x} ^2/s^2)/2$	$2(1 - \mathbf{x} ^2/s^2)/(\pi s^2)$	$3\pi s^2/4$
Gaussian	\mathbb{R}^2	$2 \exp(- \mathbf{x} ^2/2s^2)$	$\exp(- \mathbf{x} ^2/2s^2)/(2\pi s^2)$	$4\pi s^2$

Notes. All functions considered here are axisymmetric and include a spatial scale s .

knowledge of the HLF for the various populations of stars that are likely to be present in the field. In practice, we will generally use two probabilities, one for the field stars, and one for the YSOs. Second, we assume that we have measured the extinction A_n of each star: note that this is *not* the average extinction at the location of the star, which might be very different because of perspective effects: for example, a foreground star in front of a cloud would have $A_n \simeq 0$, while the average extinction would be significant. This way, the algorithm can directly account for foreground contamination: foreground stars will not be corrected in their counts, since the integral in the denominator of Eq. (30) will evaluate to unity. Similarly, stars within molecular clouds will be corrected only for the amount of material that is really in front of them.

Finally, we stress that the iterative procedure proposed here only find positive solutions for the values σ_i . Although reasonable, nevertheless this choice inevitably introduces biases in the results: for example, in a region where no YSO is present, because of errors we will still measure small positive values for the density of YSOs. However, numerical tests have shown that the bias amount is limited; moreover, a reduction of the bias is associated to a large increase in the scatter. Therefore, we will force the σ_i to be positive and use Eq. (30) for the solution.

The uncertainties associated to Eq. (30) can be computed from the Fisher matrix. For practical applications, it is convenient to rewrite Eq. (28) by replacing the integrals over dm and d^2x' with a sum over the observed objects. This leads to the approximated Fisher matrix expression

$$I_{ij} = \sum_{n=1}^N \frac{\omega^2(\mathbf{x} - \mathbf{x}_n) c^2(m_n) p_i(m_n - A_n) p_j(m_n - A_n)}{\sigma^2(m_n, \mathbf{x}_n)}. \quad (31)$$

In this equation, we take the spatial function ω to be normalized such that

$$\int \omega(\mathbf{x}') d^2x' = \int \omega^2(\mathbf{x}') d^2x'; \quad (32)$$

that is, in terms of W ,

$$\omega(\mathbf{x}) = \frac{W(\mathbf{x})}{\int W^2(\mathbf{x}') d^2x'}. \quad (33)$$

Table 1 reports a few common choices for the spatial function $\omega(\mathbf{x})$ and the corresponding weight function $W(\mathbf{x})$, both correctly normalized. As usual, the covariance matrix associated with the measurements of the densities σ_i can be computed from the inverse of the Fisher matrix, I^{-1} .

4. Simulations

In order to test our method and verify its robustness we have performed a set of simulations. We have considered a small patch

Table 2. Summaries of the results of simulations.

σ_1	σ_2	$\langle(I^{-1})_{11}\rangle$	$\langle(I^{-1})_{12}\rangle$	$\langle(I^{-1})_{22}\rangle$
$\langle\hat{\sigma}_1\rangle$	$\langle\hat{\sigma}_2\rangle$	Var[σ_1]	Cov[σ_1, σ_2]	Var[σ_2]
5.00	20.00	33.38	-19.13	35.27
5.30	19.87	21.37	-12.09	29.60
20.00	5.00	47.44	-19.02	22.34
19.79	5.21	43.31	-13.83	16.33
10.00	10.00	47.44	-19.02	22.34
10.05	9.95	30.63	-13.97	21.04

of the sky with two different stellar populations: field stars, with exponentially increasing number counts $p_1(m) \propto 10^{\alpha m}$, with $\alpha = 0.33 \text{ mag}^{-1}$; and YSOs, with Gaussian number counts $p_2(m) \propto \exp(-(m - m_0)^2/2s^2)$, with $m_0 = 12 \text{ mag}$ and $s = 1.65 \text{ mag}$. We have distributed both populations randomly in the small patch of the sky considered and we have added to each star a random extinction drawn from the probability distribution $p_A(A) \propto A^{-2}$, in the range $A \in [0.1, 2.0] \text{ mag}$. This choice is meant to simulate the effects of differential extinction for objects within molecular clouds. Finally, we have imagined that our observations are complete up to 15 mag, and that no stars can be observed beyond that magnitude: in other words, we have modeled the completeness function as a Heaviside function $c(m) = H(15 - m)$. This way, our final catalog contains, for each star, the angular position in the sky, the H -band magnitude, and the measured extinction. Note that the parameters used here are chosen to simulate a real situation corresponding to the sample application of Sect. 5, that is the Orion molecular cloud complex observed with 2MASS.

We have used these data in our algorithm, represented by Eqs. (30) and (31). As weight function $W(\mathbf{x})$ we have used a Gaussian, and we have chosen the angular units so that

$$\int W^2(\mathbf{x}') d^2x' = \left[\int W(\mathbf{x}') d^2x' \right]^2. \quad (34)$$

This choice guarantees that the effective area of the weight function is unity, i.e. the effective number of stars used for the analysis, in absence of extinction, would be just $\sigma_1 + \sigma_2$. In reality, the presence of the extinction reduces this number by a factor that depends on the relative ratio between σ_1 and σ_2 (typically, by $\sim 20\%$).

We have then performed different simulations, with various choices for σ_1 and σ_2 , to verify the ability of the algorithm to recover the input densities. Figures 2–5 show the observed distributions of $\hat{\sigma}_1$ and $\hat{\sigma}_2$, together with the predicted ones (Gaussian distributions centered on the true values of σ_1 and σ_2 , with the variances predicted from the Fisher matrix I). In general, we can note a few points:

- When one of the input densities is below ~ 7 , there is a clear excess of the corresponding quantity for small measured

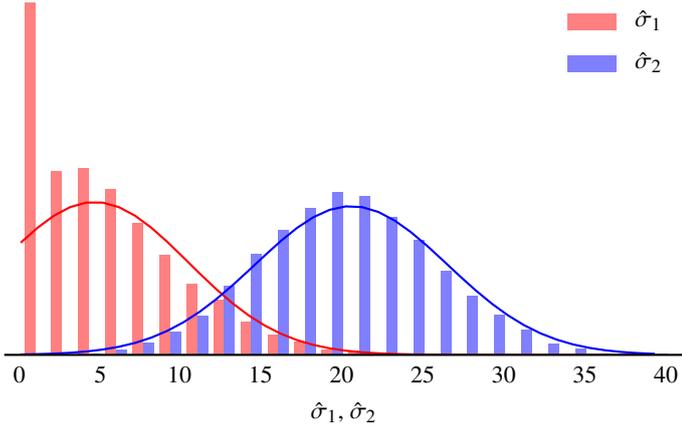


Fig. 2. Distributions of measured densities in a simulation with $\sigma_1 = 5$ and $\sigma_2 = 20$ (histogram), together with the predicted Gaussian distribution obtained according to the Fisher matrix evaluated from Eq. (31). The excess of small values of $\hat{\sigma}_1$ is due to the constraint that $\hat{\sigma} \geq 0$ imposed by the algorithm.

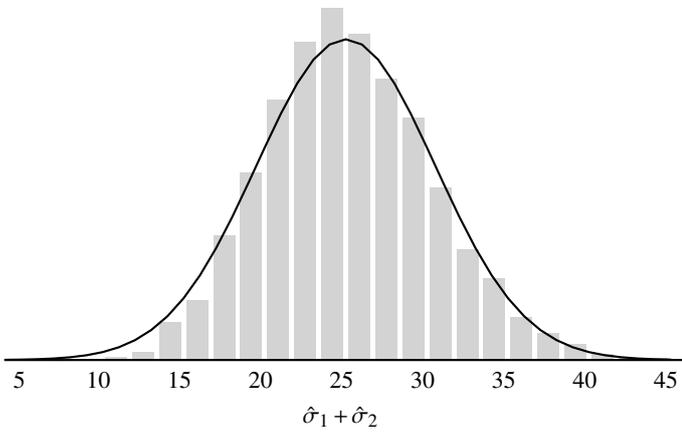


Fig. 3. Distribution of measured total density $\hat{\sigma}_1 + \hat{\sigma}_2$ in a simulation with $\sigma_1 = 5$ and $\sigma_2 = 20$, together with the predicted Gaussian distribution (derived from the Fisher matrix). The distribution is essentially unbiased; moreover, because of the anticorrelation between $\hat{\sigma}_1$ and $\hat{\sigma}_2$, the total density has significantly less scatter than both $\hat{\sigma}_1$ and $\hat{\sigma}_2$.

values. This is a consequence of the fact that the proposed algorithm only returns positive values for the densities.

- Except for the point above, the predicted distributions reproduce very well the measured data. The agreement is particularly good when the input densities are large.
- Overall, the total density $\sigma_1 + \sigma_2$ is better constrained than the individual densities $\sigma_{1,2}$.

Figures 6 and 7 show the biases and the errors on the measured densities $\hat{\sigma}_{1,2}$ for $\sigma_1 = 10$ and σ_2 varying in the interval $[0, 20]$. We can see that there is a measurable bias for $\sigma_2 < 5$, while the results are essentially unbiased above this value. Correspondingly, we observe in the same range $\sigma_2 \in [0, 5]$ a measured scatter in measured densities that is significantly smaller than what predicted from the Fisher matrix. For larger values of the input density the error estimate gets closer to the measured errors, but still remains slightly above. This is actually good, because implies a small overestimate of the error which will make the entire method more robust for cluster detections (that is, it will decrease the number of false positive). In summary, all these simulations confirm that the method works and that the error estimate is quite accurate.

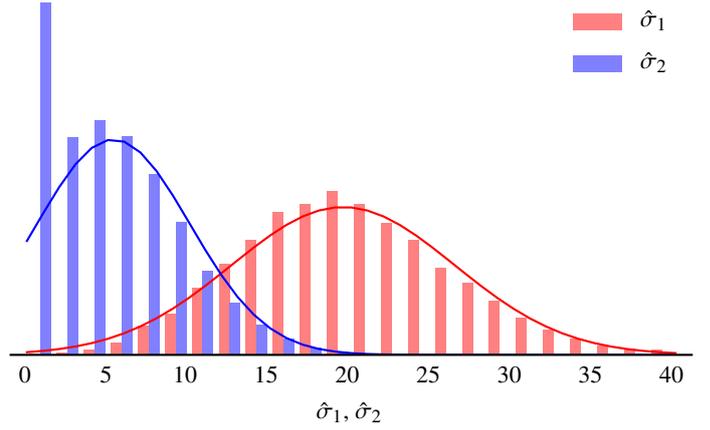


Fig. 4. Distributions of measured densities in a simulation with $\sigma_1 = 20$ and $\sigma_2 = 5$.

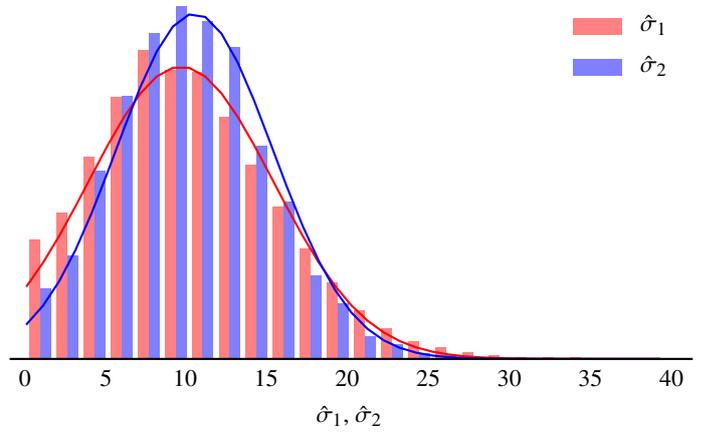


Fig. 5. Distributions of measured densities in a simulation with $\sigma_1 = 10$ and $\sigma_2 = 10$.

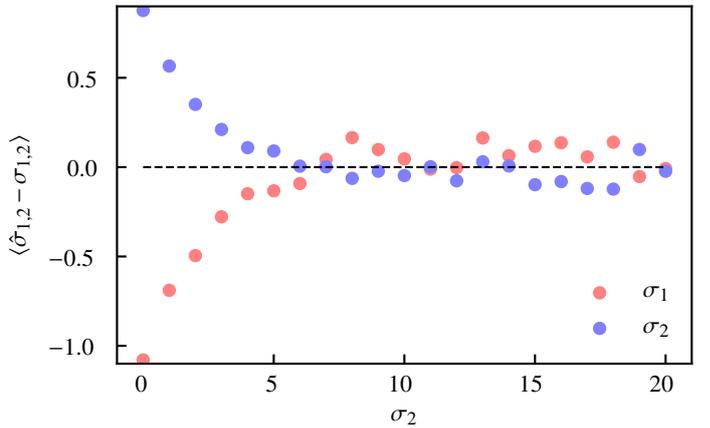


Fig. 6. Biases in the measured densities for $\sigma_1 = 10$ and $\sigma_2 \in [0, 20]$. Note how the bias essentially vanishes for $\sigma_2 > 5$.

5. Sample application: Orion

We have applied the method proposed in this paper to 2MASS data of the Orion A and B molecular cloud complexes. The regions selected ensure that we can verify the reliability of the algorithm proposed here using some of the best studied objects in the sky. In particular, the populations of embedded clusters for both clouds have been the targets of extensive observational campaigns using ground-based, near-infrared (Lada et al. 1991) and millimeter-wave

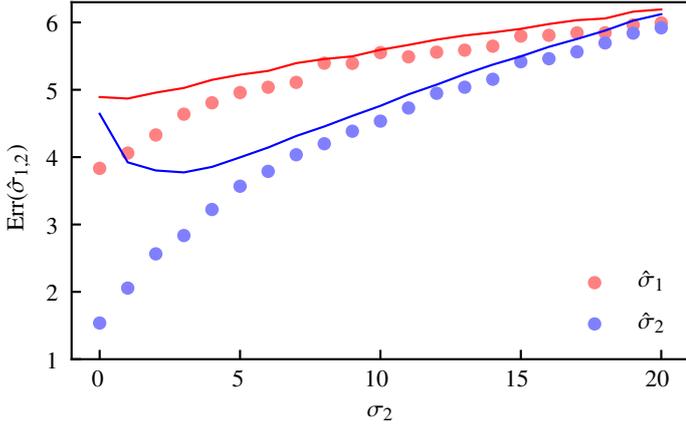


Fig. 7. Scatter on $\hat{\sigma}_1$ and $\hat{\sigma}_2$ for $\sigma_1 = 10$ and $\sigma_2 \in [0, 20]$ (dots), together with the predictions obtained from Eq. (31) (lines).

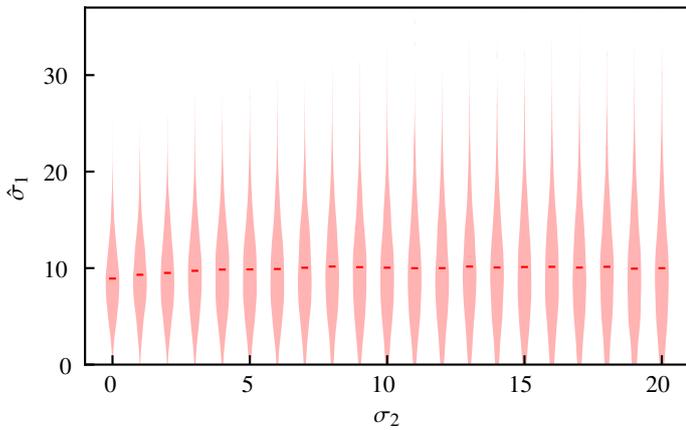


Fig. 8. Violin plot showing the distribution of measured densities $\hat{\sigma}_1$ for $\sigma_1 = 10$ and $\sigma_2 \in [0, 20]$. Each elongated structure corresponds to a different value of σ_2 ; its width is proportional to the distribution of measured values of $\hat{\sigma}_1$, i.e. effectively it is a histogram displayed vertically. The small red dashes indicate the average values.

(Lane et al. 2016) surveys as well as space-based, mid-infrared *Spitzer* Space Telescope (Megeath et al. 2012, 2016), and *Chandra* X-ray (Prisinzano et al. 2008) surveys. Additionally, the distance of these clouds ensures that the 2MASS *H*-band data are, in absence of extinction, complete for YSOs: that is, the cluster HLF at the distance of Orion is essentially entirely within the 2MASS *H*-band limiting magnitude (~ 15 mag).

Figures 10 and 11 show the density of fiducial YSOs measured in Orion A and B. These maps have been produced by our pipeline together with NICER (Lombardi & Alves 2001) and NICEST (Lombardi 2009) extinction maps from the 2MASS Point Source Catalogue (see Lombardi et al. 2011, for details on the data selection and extinction map construction). The algorithm has been run in conditions similar to the simulations described above: that is, we have used two different stellar populations, one associated with the background and characterized by exponential number counts, and one associated with the YSOs and characterized by Gaussian number counts (with parameters consistent with the HLF of Muench et al. 2002). Using an angularly close control field we also measured the distribution of intrinsic colors of stars and the shape of the completeness function: the latter has been modeled using a complementary error function erfc , as described in Appendix A.2. This choice makes it possible to use the entire 2MASS catalogue without any magnitude cut (which would increase the noise in the final data

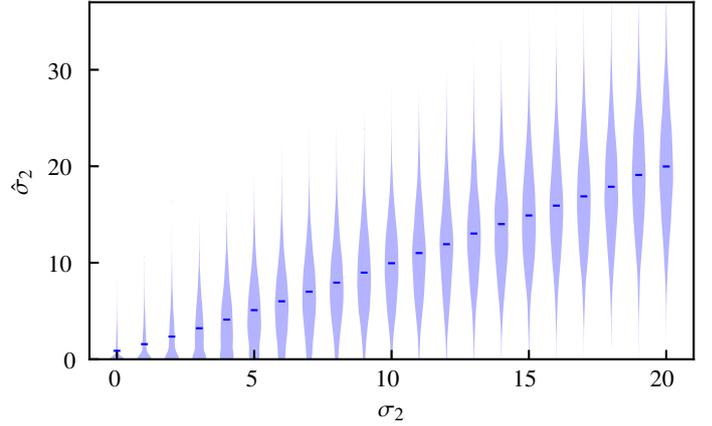


Fig. 9. Violin plot showing the distribution of measured densities $\hat{\sigma}_2$ for $\sigma_1 = 10$ and $\sigma_2 \in [0, 20]$. The small blue dashes indicate the average values.

products). The maps produced have a pixel size of 1.5 arcmin and a weight function $W \propto \omega$ in turn proportional to a Gaussian with $FWHM = 3$ arcmin. We have used a relatively large beam in these maps to increase the sensitivity of our algorithm and to minimize the effects of the biases shown in the simulations described in Sect. 4, while still retaining in most situations the ability to distinguish different clusters (i.e., avoid confusion effects at the distance of Orion).

Since we have at our disposal the covariance map of these measurements, we have assessed the reliability of each density peak in these figures. The red contours in the figures show the areas (larger than 2 pixels) where the local YSO density exceeds 1.5 stars/pixel, corresponding approximately to a signal-to-noise ratio larger than 3.5: note how some regions in black in Fig. 10 do not reach the threshold because of the large error associated with them (mostly due to the high extinction values there).

Table 3 shows the YSO clusters identified in the Orion A and B areas using our simple prescription, together with the most relevant parameters. In some cases we clearly see that angularly close clusters appear as a single contour in our maps: the simple procedure used here to define clusters, the relatively coarse resolution used, and the cluster morphology itself prevent us from deblending some close objects. An extreme case of this situation might be the integral shaped filament (ISF) cluster, where the limitations due to angular resolution would make it difficult to resolve and separate smaller clusters if they exist in such a very populous region. We note that the ISF cluster encompasses M42, the Trapezium and ONC clusters as well as an extended population of YSOs along the ISF.

The radius R reported in the table corresponds to the radius of a circle that would occupy the same area as the identified cluster, i.e. to the connected region of the sky where the inferred density of YSOs exceeds the background by 3σ . At the estimated distance of Orion, 413 pc (Reid et al. 2009), $1'$ corresponds to 0.12 pc: therefore, the clusters identified have radii spanning from ~ 2.4 pc to ~ 0.15 pc.

The well known clusters in these clouds are correctly identified by our procedure. It is interesting to compare Table 3 with clusters identified independently using much more secure data. Among the ones at our disposal, the recent catalog of embedded YSOs obtained by Megeath et al. (2016) using the *Spitzer* Space Telescope and the *Chandra* observatory is probably the most secure and complete: we will therefore focus on this catalog. Since our definition of a cluster is based on an somewhat

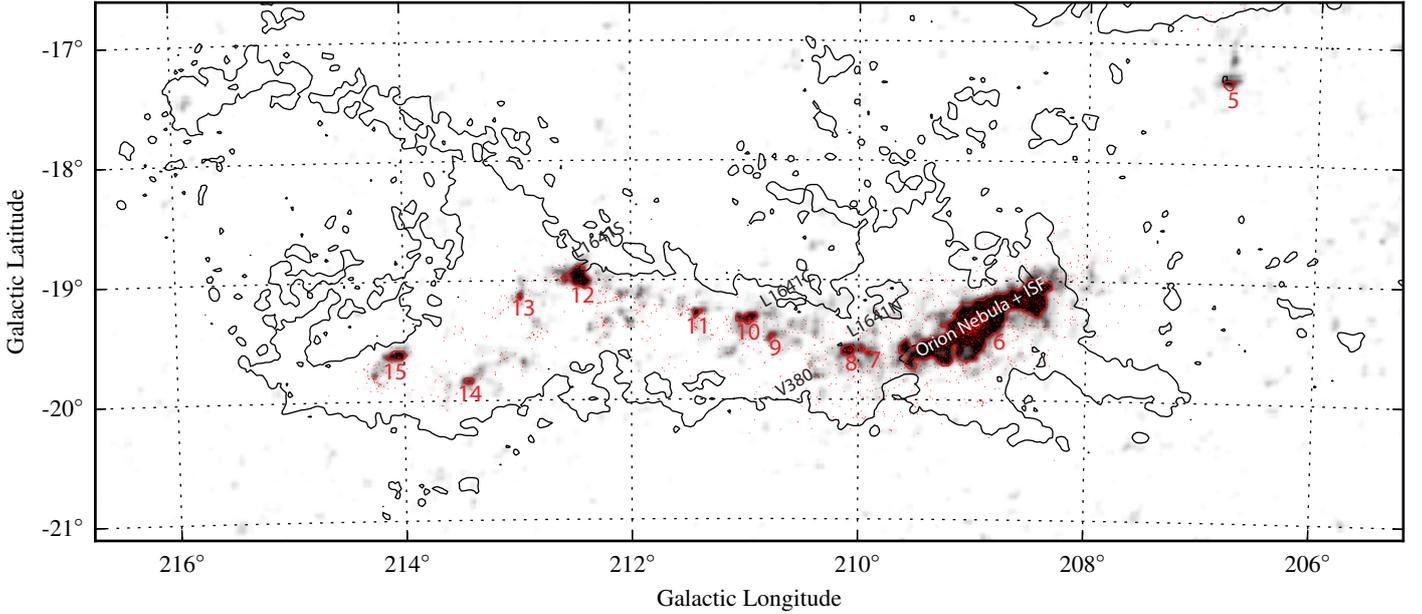


Fig. 10. Results of the cluster finding algorithm in Orion A using the 2MASS Point Source Catalog. The red contours shows all surface density detections 3σ above the background, while the black contour corresponds to $A_K = 0.3$ mag. When displayed in Adobe Acrobat, it is possible to hide the **Grid** lines, the black **Extinction** contours, the red contours corresponding to the **Clusters**, the red dots representing the **Megeath et al. YSOs**, and the clusters' **Names**.

Table 3. YSO clusters identified in Orion A and B.

ID	ℓ	b	RA	Dec	Stars	R	A_K	Cloud	Simbad	Type
	°	°	°	°		'	mag			
1	205.8569	-12.6041	88.4398	+00.4283	88 ± 14	5.2	0.2	Ori B	NGC 2112	OpC
2	205.1234	-14.1127	86.7727	+00.3486	78 ± 13	4.9	1.6	Ori B	NGC 2071	RNe
3	205.2984	-14.3236	86.6667	+00.0995	56 ± 11	4.1	1.4	Ori B	M 78	RNe
4	206.4996	-16.3341	85.4332	-01.8659	188 ± 23	7.1	2.1	Ori B	NGC 2024	Cl*
5	206.8020	-17.3464	84.6710	-02.5928	15 ± 4	2.5	0.1	Ori A	sig Ori	MGr
6	208.9681	-19.3846	83.7998	-05.3541	1601 ± 49	19.8	1.1	Ori A	M 42	HII
7	209.9598	-19.6021	84.0276	-06.2867	8 ± 3	1.9	0.6	Ori A	–	–
8	210.1018	-19.6036	84.0865	-06.4071	27 ± 6	3.1	1.1	Ori A	LDN 1641N	Cl*
9	210.7790	-19.4832	84.4813	-06.9247	5 ± 3	1.5	0.6	Ori A	[BDB2003] G210.80-19.50	Cl*
10	210.9941	-19.3347	84.7061	-07.0406	35 ± 9	3.6	1.7	Ori A	LDN 1641C	Cl*
11	211.4326	-19.2912	84.9301	-07.3924	12 ± 5	2.1	2.2	Ori A	–	–
12	212.4806	-18.9864	85.6458	-08.1475	91 ± 13	5.3	0.6	Ori A	2MASS J05422695-0809173	Y*O
13	212.9849	-19.1519	85.7057	-08.6486	4 ± 2	1.2	1.4	Ori A	[BDB2003] G212.98-19.15	Cl*
14	213.4450	-19.8466	85.2625	-09.3407	10 ± 4	2.1	0.6	Ori A	–	–
15	214.0756	-19.6278	85.7219	-09.7820	32 ± 8	3.4	1.1	Ori A	[BDB2003] G214.06-19.62	Cl*

arbitrary parameters (signal-to-noise threshold, minimum number of pixels, no correction for the stars missed at the boundaries), and since different, more-or-less, arbitrary parameters are also used by Megeath et al. (2016), we find it more appropriate and fair to make a comparison after we first homogenize the data. Specifically, we take Megeath et al. (2016) YSO list and we make out of it a density map using a Gaussian kernel of the same size of the one used for our map. Figures 12 and 13 show the results obtained for Orion A and B, which clearly compares very well with our own maps, derived purely from the 2MASS point source catalog. The most relevant structures are present in both maps and have very similar shapes; the only differences are the noise present in our maps (however at a relatively low level), and the limited coverage of the *Spitzer* derived density maps.

The qualitative similarity of these maps can be quantified if compare clusters identified in both maps using the same

criteria. Table 4 shows a list of clusters identified in the smoothed Megeath et al. (2016) maps using a fix density threshold (set to 1.5 stars/pixel). In this table we compare the number of *Spitzer* YSOs with the number of YSOs predicted from the integral of the σ_{YSOs} over the area of each cluster as defined from Megeath et al. density map, together with the computed 1σ error. It is clear that in almost all cases we find an excellent agreement, although in many cases our estimates are slightly larger than the ones by Megeath et al. We can speculate that this is due to the presence of class III YSO, which likely would be missed by *Spitzer*. Indeed, a comparison of the two panels of Fig. 12 shows that the bottom panel, corresponding to our density map, has spatially more extended clusters than the top panel, corresponding to Megeath et al. density map.

As discussed earlier on, our algorithm is a statistical one and works best when it is applied to a sizeable number of stars.

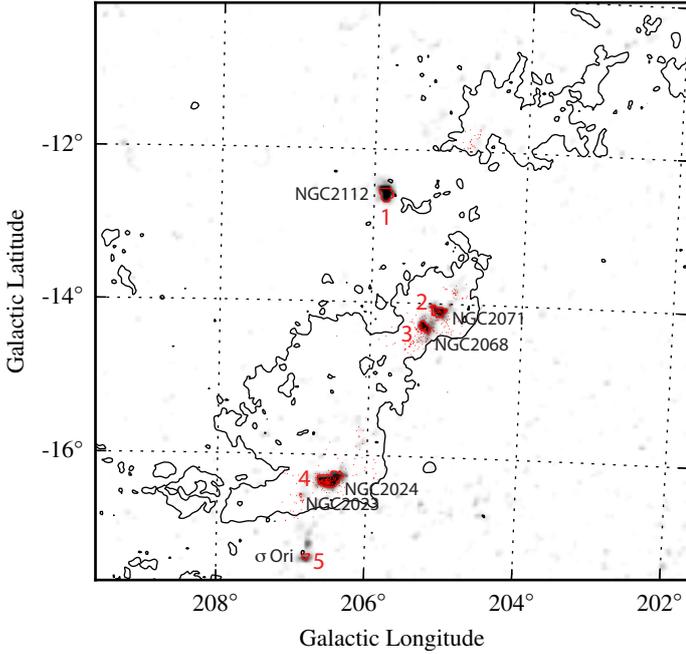


Fig. 11. Results of the cluster finding algorithm in Orion B using the 2MASS Point Source Catalog (see caption of Fig. 10 for the legend). Note that NGC 2023 is below the detection threshold and appears only as a weak smudge in this image. When displayed in Adobe Acrobat, it is possible to hide the Grid, the Extinction, the Clusters, the red dots representing the Megeath et al. YSOs, or the Names.

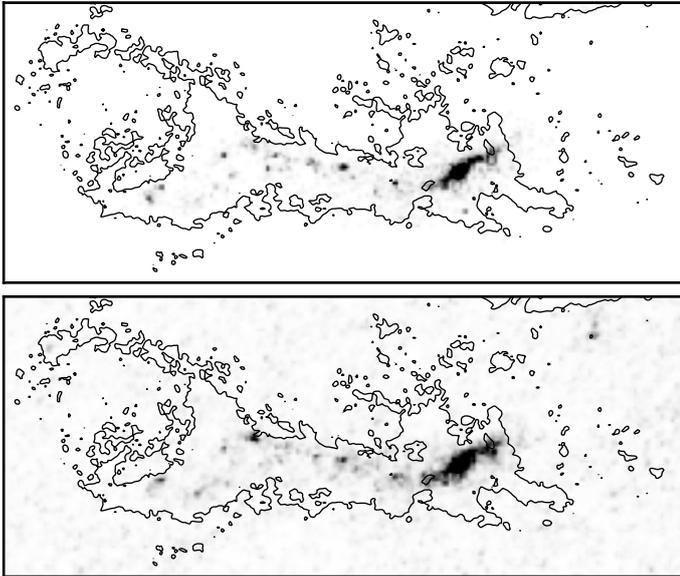


Fig. 12. A Gaussian-kernel smoothed density map of the Megeath et al. (2016) YSO list in Orion A (top) to compare with our density map (bottom).

However, we can also push it and associate to each single star a probability of being a YSO: to this purpose, for the n th star we can compute

$$P_i = \frac{\sigma_i(\mathbf{x}_n)p_i(m_n - A_n)}{\sum_{j=1}^L \sigma_j(\mathbf{x}_n)p_j(m_n - A_n)}. \quad (35)$$

Note how this quantity resembles the term within the outer sum of Eq. (30). Figure 14 shows the distribution of P_{YSO} (that is, the distribution in the probabilities assigned to each object to be a

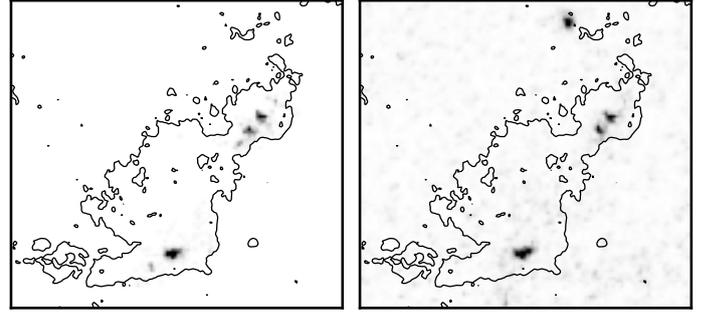


Fig. 13. A Gaussian-kernel smoothed density map of the Megeath et al. (2016) YSO list in Orion B (left) to compare with our density map (right).

Table 4. Correspondence between the clusters identified in this work, marked with the subscript 1, and the ones identified from a smoothed version of the YSO catalog (Megeath et al. 2016).

ℓ	b	Stars ₁	Stars ₂	R	Cloud
214.0472	-19.6402	12 ± 4	8.9	1.9	Ori A
212.9892	-19.1435	10 ± 5	11.5	2.1	Ori A
212.4676	-19.0192	8 ± 3	3.4	1.2	Ori A
210.9717	-19.3367	28 ± 8	39.5	3.3	Ori A
210.0881	-19.5955	23 ± 6	31.1	2.8	Ori A
209.5134	-19.6566	4 ± 2	3.4	1.2	Ori A
209.0159	-19.3924	929 ± 36	873.3	12.9	Ori A
208.5245	-19.1518	68 ± 11	51.6	4.3	Ori A
208.4171	-19.1892	20 ± 5	12.1	2.2	Ori A
208.3900	-19.0734	15 ± 4	10.3	2.1	Ori A
206.5222	-16.3557	135 ± 19	158.5	5.5	Ori B
205.3203	-14.3258	39 ± 9	28.2	3.3	Ori B
205.1356	-14.0971	61 ± 12	55.6	4.2	Ori B

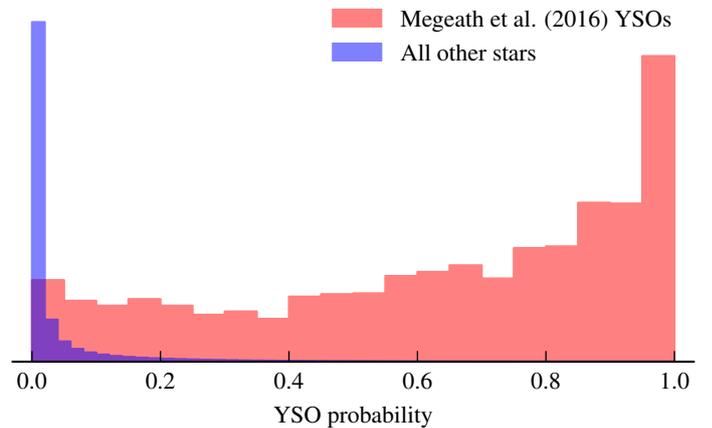


Fig. 14. Distribution of P_{YSO} , the probability that our algorithm assigns to each star to be a YSO, for the Megeath et al. (2016) YSO candidates and for all other objects in the 2MASS point source catalog.

YSO) for the Megeath et al. (2016) YSO candidates and for all the other objects. It is clear how all the other objects have P_{YSO} that is concentrated around zero, while the YSO candidates have a much broader distribution that extends to unity. For these latter objects the distribution, in addition to a substantial peak at $P_{\text{YSO}} = 1$, shows a long tail up to small values of P_{YSO} : this is not unexpected, since our identification is only statistical (and therefore we cannot identify unambiguously YSOs). Note also how the relatively low values of P_{YSO} for some genuine YSOs

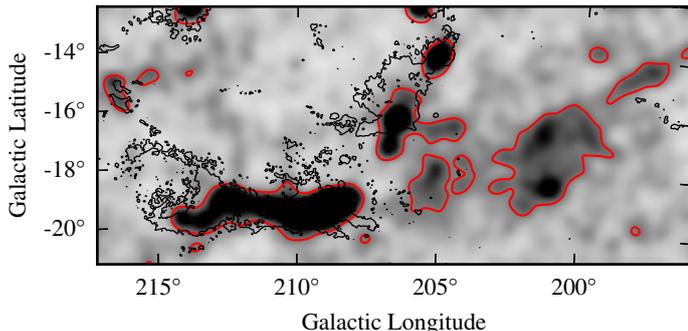


Fig. 15. Low-resolution version of the YSO density in Orion. The grey area at $\ell \approx 205^\circ$ is the Orion OB Ib association, while the lighter area to the right (around $\ell \approx 201^\circ$) is OB Ia, containing the 25 Ori cluster (the grey spot at $\ell \approx 201^\circ$ and $b \approx -18^\circ$).

in our algorithm are compensated by the small tail in the distribution of field stars (this works because there are many more field stars than YSOs, a fact that is accounted for in the algorithm).

Sensitivity to the distributed population. Recently, Kubiak et al. 2017, have identified a rich and well-defined stellar population of about 2000 objects, mostly M stars without extinction or infrared-excesses, as the low-mass counterpart to the Orion OB Ib subgroup (the Orion belt population). This low-mass population is not obviously clustered but instead appears to be distributed across ~ 10 square degrees and the authors speculate that it could represent the evolved counterpart of a Orion nebula-like cluster. While more data is needed to test this scenario, it is clear that much can be learned about the origin of stellar OB associations and the dispersal of clusters into the Galactic field if one is able to trace in a robust manner the distribution of the slightly older and more expanded populations surrounding star-forming molecular clouds.

We now investigate the extent to which the technique proposed here is suitable for detection of looser, more expanded distributions of young stars, in particular the low-mass counterpart to the Orion OB association presented in Kubiak et al. (2017). For this purpose, we have built a lower resolution map of the region, employing a *FWHM* of 30 arcmin.

Figure 15 shows that, surprisingly, we are well able to recover the stellar over-density of the Ori Ib population, and for the first time, the stellar over-density of the Ori Ia population. These over-densities are likely to be created by low-mass stars as 2MASS is still sensitive to the peak of the IMF for the putative distance and age of these subgroups. An analysis of the substructure seen in the distributed population visible in Fig. 15 above the noise pattern is beyond the scope of this paper, but will best addressed once *Gaia* parallaxes are generally available. Of relevance for this paper is that the ability of the method to trace the dispersed population from existing all sky data opens a new window on the unsolved problem of the origins of OB association and cluster dispersal in to the field.

6. Conclusions

The following items summarize the main results presented in this paper:

- We have developed a new method to discover and characterize deeply embedded star clusters.

- The method is able to statistically classify objects as field stars or YSOs and corrects for the effects of differential extinction.
- We have provided expressions for the covariance of the inferred densities and we have validated both the method and the analytic expression for the covariance with a set of simple but realistic simulations.
- We have applied the new method to 2MASS point sources observed in the Orion A and B and we have shown that we can identify and characterize well protostellar clusters in these regions, as well as detect much looser associations such as the OB 1a and 1b subgroups.

Finally, we note that the method proposed here can be easily extended to multi-band observations by using suitable probability distributions $p_i(\mathbf{m})$ for various populations as a function of the magnitude vector \mathbf{m} . Its implementation would be essentially identical, with minor modifications to take into account the different effects of extinction on different bands. The natural use of such an extension would be in the context of techniques such as the one proposed by Juvela & Montillaud (2016) which are able to recover the extinction from a complex analysis of multi-band data.

Acknowledgements. This research has made use of the 2MASS archive, provided by NASA/IPAC Infrared Science Archive, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Additionally, this research has made use of the SIMBAD database, operated at CDS, Strasbourg, France.

References

- Bressert, E., Bastian, N., Gutermuth, R., et al. 2010, *MNRAS*, 409, L54
- Cambr esy, L., Beichman, C. A., Jarrett, T. H., & Cutri, R. M. 2002, *AJ*, 123, 2559
- Cambr esy, L., Petropoulou, V., Kontizas, M., & Kontizas, E. 2006, *A&A*, 445, 999
- Cambr esy, L., Marton, G., Feher, O., T oth, L. V., & Schneider, N. 2013, *A&A*, 557, A29
- Carpenter, J. M. 2000, *AJ*, 120, 3139
- Casertano, S., & Hut, P. 1985, *ApJ*, 298, 80
- Feigelson, E. D., & Babu, G. J. 2012, *Modern Statistical Methods for Astronomy* (Cambridge University Press)
- Geyer, M. P., & Burkert, A. 2001, *MNRAS*, 323, 988
- Gutermuth, R. A., Megeath, S. T., Pipher, J. L., et al. 2005, *ApJ*, 632, 397
- Gutermuth, R. A., Megeath, S. T., Myers, P. C., et al. 2009, *ApJS*, 184, 18
- Hubble, E. P. 1926, *ApJ*, 64
- Juvela, M., & Montillaud, J. 2016, *A&A*, 585, A78
- Kubiak, K., Alves, J., Bouy, H., et al. 2017, *A&A*, 598, A124
- Lada, C. J., & Lada, E. A. 2003, *ARA&A*, 41, 57
- Lada, C. J., Margulis, M., & Dearborn, D. 1984, *ApJ*, 285, 141
- Lada, E. A., Depoy, D. L., Evans, II, N. J., & Gatley, I. 1991, *ApJ*, 371, 171
- Lane, J., Kirk, H., Johnstone, D., et al. 2016, *ApJ*, 833, 44
- Lombardi, M. 2009, *A&A*, 493, 735
- Lombardi, M., & Alves, J. 2001, *A&A*, 377, 1023
- Lombardi, M., Alves, J., & Lada, C. J. 2011, *A&A*, 535, A16
- Lombardi, M., Lada, C. J., & Alves, J. 2013, *A&A*, 559, A90
- MacKay, D. 2003, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press)
- Megeath, S. T., Gutermuth, R., Muzerolle, J., et al. 2012, *AJ*, 144, 192
- Megeath, S. T., Gutermuth, R., Muzerolle, J., et al. 2016, *AJ*, 151, 5
- Meingast, S., Alves, J., Mardones, D., et al. 2016, *A&A*, 587, A153
- Muench, A. A., Lada, E. A., Lada, C. J., & Alves, J. 2002, *ApJ*, 573, 366
- Prinzano, L., Micela, G., Flaccomio, E., et al. 2008, *ApJ*, 677, 401
- Reid, M. J., Menten, K. M., Zheng, X. W., et al. 2009, *ApJ*, 700, 137
- Rom an-Z u niga, C. G., Elston, R., Ferreira, B., & Lada, E. A. 2008, *ApJ*, 672, 861
- von Hoerner, S. 1963, *ZAp*, 57, 47
- Wilkings, B. A., & Lada, C. J. 1985, in *Protostars and Planets II*, eds. D. C. Black, & M. S. Matthews, 297

Appendix A: Implementation

In this section we consider a few details and analytical expressions useful to evaluate some of the expressions needed to implement the method proposed here.

A.1. Algorithm

The algorithm proposed in this paper is essentially an iterative application of Eq. (30) to the set of densities $\{\sigma_i(\mathbf{x})\}$. A few notes, however, are necessary to better implement the algorithm:

- The method assumes that the extinction values in the H band are available for each star used. In our implementation these values are obtained through the use of the NICER algorithm. Other techniques can be used, as long as the extinction is referred to the single star and is *not* an average extinction at the location of the star. In our implementation, this requirement limits the method to stars for which, in addition to the H -band measurements, have at least another band photometry.
- Since the various densities are spatially variable, one needs to repeat the iteration of Eq. (30) to each point in the map.
- Equation (30) itself needs to be iterated a few times (typically around ten, in some cases a little more) before reaching a good convergence. This generally suggests the use of simple analytical models for the various functions involved in this equation (the probability distributions $p_i(m)$ and the completeness function $c(m)$, discussed below).

A.2. H -band probability distributions and completeness function

In this paper we have considered two useful models for the H -band luminosity function: the exponential distribution and the normal (Gaussian) one.

We parametrize the exponential distribution as

$$p(m) = k\beta e^{\beta m} \propto 10^{\alpha m}, \quad (\text{A.1})$$

where k is a normalization constant that will be found later on and $\beta = \alpha \ln 10$.

The normal distribution is parametrized as

$$p(m) = \frac{k}{\sqrt{2\pi}\sigma^2} e^{-(m-m_0)^2/2\sigma^2}, \quad (\text{A.2})$$

where again k is a normalization constant to be investigated later.

Finally, we model the completeness function $c(m)$ in terms of the complementary error function erfc . This can be justified from an empirical point of view since the error function has a shape that resembles the completeness function. It is also reasonable from a statistical point of view if the photometric errors are Gaussian: in this case, one can suppose that the probability that an object be detected is an integral over a relevant part of the Gaussian. Specifically, we assume for $c(m)$ the following functional form:

$$c(m) = \frac{1}{2} \text{erfc} \left(\frac{m - m_c}{\sqrt{2}\sigma_c} \right), \quad (\text{A.3})$$

where m_c is the 50% completeness limit and σ_c sets the width of the completeness function. Figure A.1 shows that our simple model for the control field HLF, the product of $p(m)c(m)$ with $p(m)$ following an exponential distribution and $c(m)$ described in terms of an erfc , reproduces well the data.

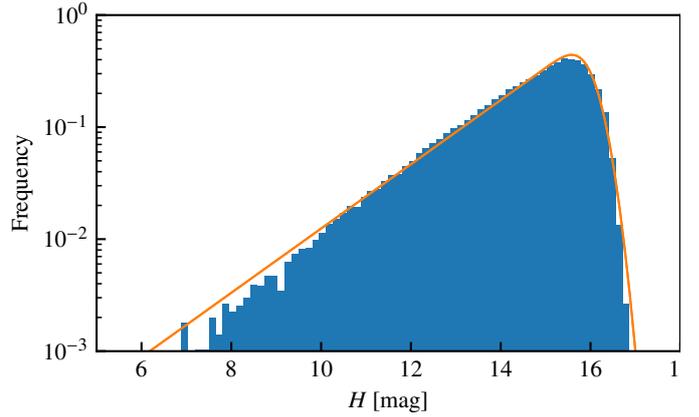


Fig. A.1. Histogram of H -band magnitudes of stars in a control field region, together with its best fit (an exponential with a complementary error function truncation, see the text).

A.3. Statistical errors and completeness function

As noted above, statistical errors can be included in our algorithm by convolving the probability distributions p_i with appropriate kernels. We expect to have two main sources for statistical errors: photometric errors in the H -band magnitude measurements and errors in the extinction measurements. While the formers are relatively simple to characterize (typically, they will be provided in the star catalogue), the latter are a combination of different sources of errors: the scatter in the intrinsic color of sources and the photometric errors in each band used. If extinctions are derived using the NICER algorithm, as assumed here, then the variance associated to each extinction measurement can be computed from the expression

$$\text{Var}(A) \equiv \sigma_A^2 = \frac{1}{\mathbf{k}^T \mathbf{C}^{-1} \mathbf{k}}, \quad (\text{A.4})$$

where, following the notation of Lombardi & Alves (2001), we have called \mathbf{k} is the reddening vector and \mathbf{C} the combined covariance matrix of the observed colors (including both the intrinsic scatter in the color of unextinguished stars and the individual photometric errors).

Looking again at Eq. (30), i.e. the main equation representing our method, we see that in reality the H -band magnitudes and the associated extinction measurements typically enter the problem through the combination $m - A$; moreover, while in the numerator the combination is $m_n - A_n$ and involves therefore measurements of both m and A for individual stars, in the denominator the combination is $m - A_n$, with m integration variable. Therefore, in presence of errors these two parts of the expression must be computed separately.

Let us consider first the denominator. Since there the only argument of p_i with associated errors is A_n , the measured extinction for the n th star, we just need to convolve the magnitude distribution with a Gaussian kernel with variance provided by Eq. (A.4).

In the case of the numerator the situation is slightly more complicated, because there we find the combination $m_n - A_n$. Since A_n is computed from the observed magnitude of each star (including of course the H -band magnitude), m_n and A_n are correlated. A simple calculation shows that the variance associated to the combination $m_n - A_n$ is

$$\text{Var}(m_n - A_n) = k_H^2 \sigma_A^2 + \sigma_H^2 [1 - 2(b_{J-H} - b_{H-K}) k_H]. \quad (\text{A.5})$$

In this expression the last term is related to the correlation between A and m and is equal to σ_H^2 , the square of the photometric error on the H -band magnitude. The factors b_{J-H} and b_{H-K} are quantities that can be computed from Eq. (12) of Lombardi & Alves (2001), while k_H is the reddening vector for the H band, i.e. the ratio A_H/A .

A.4. Convolution integrals

The convolution of the luminosity functions (A.1) and (A.2) with Gaussian kernels representing the statistical errors take simple forms. Calling σ_p^2 the variance of the convolution kernel, given depending on the term by Eq. (A.4) or by Eq. (A.5), and using the tilde to represent the convolved H -band luminosity functions, we find for the exponential distribution

$$\tilde{p}(m) = k\beta e^{\beta m + \sigma_p^2 \beta^2 / 2}. \quad (\text{A.6})$$

Similarly, for the Gaussian model we find

$$\tilde{p}(m) = \frac{k}{\sqrt{2\pi(s^2 + \sigma_p^2)}} e^{-(m-m_0)^2 / 2(s^2 + \sigma_p^2)}, \quad (\text{A.7})$$

A.5. Completeness integrals

Finally, we need to compute the integral in the denominator of Eq. (30), involving the HLF $p(m)$ and the completeness function $c(m)$. A simple change of variable casts this integral (similar to a convolution) into a more convenient form:

$$\int c(m)p(m-A) dm = \int c(m+A)p(m) dm \equiv K(A). \quad (\text{A.8})$$

In this equation we have called $K(A)$ the result of the integral, retaining its explicit dependence on A . We now consider the derivative K' of K :

$$K'(A) = \int c'(m+A)p(m) dm, \quad (\text{A.9})$$

and note that, since c is the error function, c' is a Gaussian function. Since Eq. (A.9) is also (essentially) a convolution, this allows us to compute $K'(A)$ using formulae similar to the ones of the previous section: as a result, we see that $K'(A)$ remains either an exponential or a Gaussian, depending on the function form of $p(m)$.

Finally, we need to integrate back $K'(A)$ to obtain K . The result we obtain from this procedure is, in case of an exponential distribution,

$$K(A) = ke^{\beta(m_c - A) + \sigma_c^2 \beta^2 / 2}. \quad (\text{A.10})$$

For the Gaussian distribution the result contains again the error function:

$$K(A) = \frac{k}{2} \operatorname{erfc} \left(\frac{m_0 - m_c + A}{\sqrt{2(\sigma_c^2 + s^2)}} \right). \quad (\text{A.11})$$

If statistical errors have to be taken into account, these expressions simply change into

$$K(A) = ke^{\beta(m_c - A) + (\sigma_c^2 + \sigma_p^2) \beta^2 / 2}, \quad (\text{A.12})$$

and

$$K(A) = \frac{k}{2} \operatorname{erfc} \left(\frac{m_0 - m_c + A}{\sqrt{2(\sigma_c^2 + s^2 + \sigma_p^2)}} \right). \quad (\text{A.13})$$