Using Gaia to derive distances to embedded objects

Hajanirina Erick Randriamanantena

Department of Physics and Astronomy

University of Leeds

A thesis submitted for the degree of

Master of Science

November 29, 2021

Acknowledgements

I would like to thank my primary supervisor, Dr Stuart Lumsden, for his guidance, support and patience throughout my MSc. I appreciate his kind advice, precious comment, constructive suggestions, and careful review, which greatly improve the quality of this work. I would also like to thank Prof Melvin Hoare for his assistance.

A big thank you goes to the DARA Big Data project led by Prof Anna Scaife at the University of Manchester, for all the funding I received for the full duration of my MSc programme. I also acknowledge the School of Physics and Astronomy at the University of Leeds for hosting this research.

Finally, and most importantly, I would like to thank my whole family for their unconditional love and encouragement despite me being so far away.

Abstract

In this thesis, we present distance measurements to embedded stars by using the wealth of astrometric data from *Gaia* DR2. Our methodology is based on the Bayesian techniques and the Markov chain Monte Carlo (MCMC) sampling by modelling the extinction towards the region of interest to infer the distance to the target source. We model the A_G extinction in the line of sight to provide reliable distance measurements. We also use the A_V extinction derived by Anders *et al.* (2019) to see the improvement in the distances when using additional catalogues. The distance is subsequently inferred from the jump point on the extinction from the Off-cloud to On-cloud stars as each extinction measurement has its corresponding distance.

We inferred distances to Young Stellar Objects (YSOs) selected from the literature and to the sub-regions of the high mass star formation region, Cygnus X (DR20, DR21, DR22, DR23, and W75N). We found that Gaia can provide a reliable distance to an object associated with a molecular cloud with moderate-sized extinction, showing a small systematic uncertainty of less than 5%. For dark clouds, however, our extinction models inferred lower distances compared to maser distances, kinematic distances, and to the extinction distances of Foster *et al.* (2012). This is because there are multiple extinction breakpoints towards those selected regions, and our models provide distances to the first jump. We also found that the sub-regions of Cygnus X are located at a similar distance of $\sim$ 1kpc according to A_G, and at $\sim$1.6

kpc according to A_V. This suggests that the idea of using additional photometric data with *Gaia* in the A_V model improves the distance as it added many input stars for the models. Our methodology failed to measure distances to object in a cloud with complex extinction distribution that differs from our simple dust screen model. We stress, however, that the advent of the full Gaia Data Release 3 will significantly improve our distance measurements as many more data will be available.

Abbreviations

2MASS	2 Micron All-Sky Survey
APSIS	Astrophysical Parameters Inference System
ADQL	Astronomical Data Query Language
A_G	Extinction in the G band
A_V	Extinction in the V band
BNCE	Blue Number Count Extinction
CNM	Cold Neutral Medium
CoMRS	Centre-of-Mass Reference System
CO	Carbon Monoxide
ES	Evolved Star
ESA	European Space Agency
ESO	European Southern Observatory
FLAME	Final Luminosity, Age, and Mass Estimator
GAIA	Graphical Astronomy and Image Analysis Tool
GAIA DR2	GAIA Data Release Two
GMC	Giant Molecular Cloud
GCRS	Geocentric Celestial Reference System
H_2	Molecular Hydrogen
HCHII	Hyper-compact HII region
HIM	Hot Ionized Medium
ICRS	International Celestial Reference System
IMF	Initial Mass Function
IR	Infrared
ISM	Interstellar Medium
KDA	Kinematic Distance Ambiguity
kpc	kiloparsec
LRS	Local Standard of Rest

MASER	Microwave Amplification by Stimulation Emission of Radiation
MCMC	Markov chain Monte Carlo
MLE	Maximum Likelihood Estimation
MS	Main Sequence
MSX	Midcourse Space Experiment
NICER	Near-Infrared Colour Excess Revisited
NIR	Near-infrared
PAH	Polycyclic aromatic hydrocarbon
Pan-STARRS	Panoramic Survey Telescope and Rapid Response System
pc	parsec
PSC	Point Source catalogue
RAVE	RAdial Velocity Experiment
RGE	Red Giant Extinction
RMS	Red MSX survey
RUWE	Renormalised Unit Weight Error
SED	Spectral Energy Distribution
SKA	Square Kilometre Array
SNR	Signal-to-Noise Ratio
T_{eff}	effective temperature
UCHII	Ultra-compact HII region
UKIDSS	UKIRT Infrared Deep Sky Survey
UV	Ultraviolet
VLBI	Very Long Baseline Interferometry
WIM	Warm Ionized Medium
WISE	Wide-field Infrared Survey
WNM	Warm Ionized Medium
WTTS	Weak Line T Tauri Star
YSO	Young Stellar Object

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Looking at the night sky on a cloud-free might make us wonder how far away those stars are. Determining distances to the stars is one of the significant challenges in astronomy and astronomers are still trying to find the appropriate way to calculate them by applying several methodologies. Knowing the distances of stars leads to solving the obstacle to our understanding of the Milky Way and the understanding of galaxies in general. For example, combining the distance with the apparent brightness gives us its true luminosity (see Eq. 1.1).

$$L = 4\pi D^2 B \tag{1.1}$$

where B is the brightness of the star and D is the distance.

Apart from the luminosity, knowing distance to stars also allows us to infer different stellar parameters such as its mass from the orbital motions, its true motions through space and indeed its physical size. The brightness of a star can be measured directly. By combining the measured brightness from different wavelengths with the colour of the star, without considering the extinction, its surface temperature can be determined. When the luminosity of a star and its surface temperature is known, they can be plotted against each other in the Hertzsprung-Russel diagram (H-R diagram) to understand the mass, size, age and evolution of the star.

With the H-R diagram, we can recognise if the star is a main-sequence star or

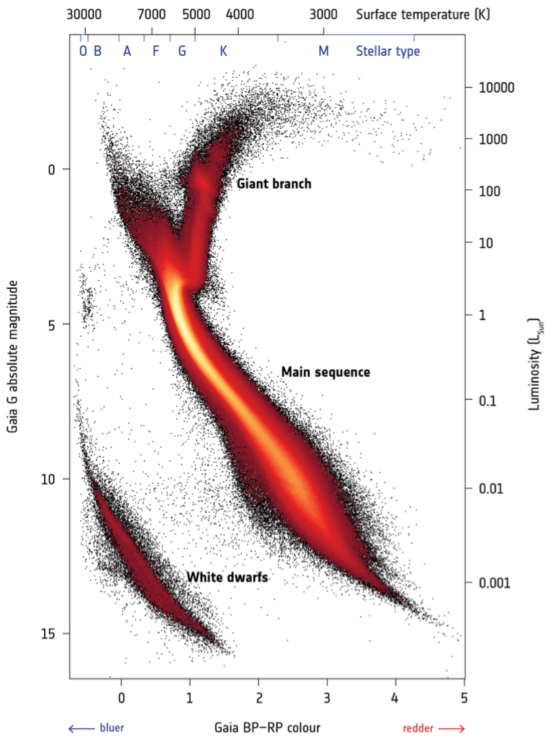

Figure 1.1: Gaia's Hertzsprung-Russell diagram. (Credit: ESA)

a super-giant star, as shown in Fig. 1.1. Most of the stars appear in the narrow band of the diagram, called the *main sequence* (MS), which means that stars of the same temperature have essentially the same luminosity and therefore have the same size. Some stars appear above the main sequence, meaning that they are more luminous than main-sequence stars of the same temperature (Giant branch). Some stars are located below the main sequence (White dwarfs), which are typically 10 *mag* fainter than main-sequence stars of the same temperature.

Measuring distances to stars depends on their environment during and after their formation. When stars are ejected from the centre of the place that they formed and can be observed visually, their distances can be calculated from their measured parallax. However, at their early stage of formation, observing their distances is arduous as they are embedded at the centre of the molecular cloud where they formed. Even if nuclear fusion might have taken place at their core,

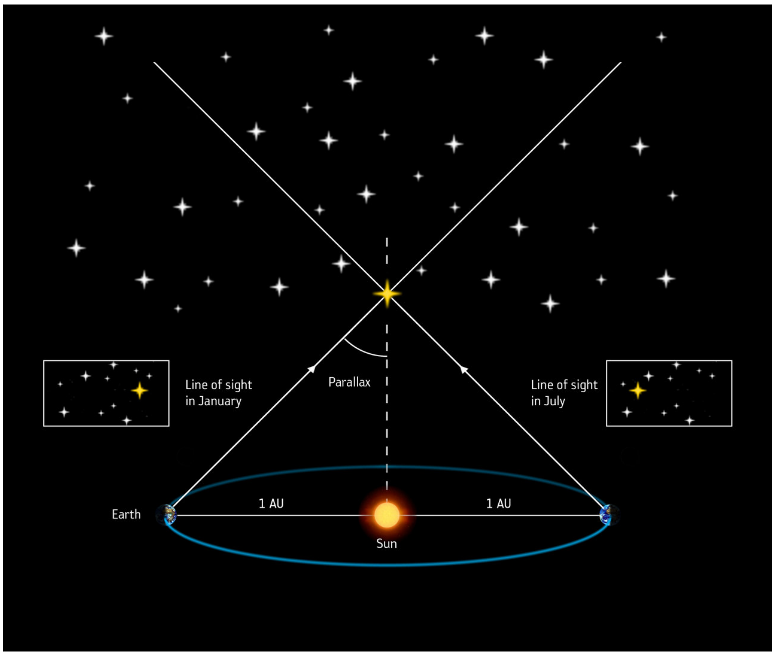

Figure 1.2: Simple illustration of trigonometric parallax method (Credit: ESA).

they are still embedded by their natal cloud, so their distance is still hard to measure. Those objects are pre-main sequence and known as young stellar objects (YSOs), which are only observable at long-wavelength (infrared and radio) (see Sect. 1.7).

1.1 Parallax

Parallax is conceptually the most fundamental distance measurement technique for stars. The parallax of an object is its observed angular displacement with respect to a reference frame due to the movement of one observer over a baseline. This is a pure geometric measurement of distance as it does not make assumptions about the intrinsic properties of the star (Bailer-Jones, 2015). As shown in Fig. 1.2, the Earth-Sun distance is used twice as a baseline to measure the parallax of an object. After solving the simple trigonometric problem, the distance r of a star is equal to the quantity $1/\omega$, where ω is its observed parallax.

The first stellar parallax was obtained by Bessel (1838), who measured $\omega =$

0.29" for the star 61 Cygni. After that, it was used in the Hipparcos mission to derive distances to $\sim 10^5$ stars down to a V-band magnitude of around 12 (Perryman *et al.*, 1997). Although this method is more effective for nearby stars, it becomes handy for distance measurement since the advent of the Gaia mission (Gaia Collaboration *et al.*, 2016b).

Inverting parallax to derive distances is only valid in the absence of noise, and it is unreliable for the faintest, most distant object. Results from Perryman *et al.* (1997) show that parallax distances using Hipparcos were accurate only out to distances of around 200 pc. This lower precision is due to dust attenuation and low SNR as noise increases at a higher distance. However, with Gaia, the uncertainty associated with each measurement is provided, which should be taken into account for better estimates of distances (Astraatmadja & Bailer-Jones (2016), Bailer-Jones *et al.* (2018)).

1.2 The Gaia mission

Gaia is a space observatory dedicated to astrometry. It was launched by the European Space Agency (ESA) in late 2013 and reached the Lagrange L2 point located about 1.5 million kilometres from Earth, one of the advantages of which is to provide an extremely stable thermal environment. There, it describes a Lissajous-type orbit to avoid eclipses of the Sun by the Earth, in order to be able to power its solar panels. Its primary goal is to map the stars of the Milky Way with greater precision. Compared to its predecessor, the Hipparcos mission (1989-1993) (Perryman *et al.*, 1997), Gaia observes a larger number of stars with unprecedented precision in the range of microarcsecond. More than 1.7 billion sources were observed in the second release of its data collection, while Hipparcos observed only 120,000 sources. Before its end, the mission has already brought a huge improvement in our understanding of the Milky Way.

1.2.1 Gaia astrometric instruments

The spacecraft, which weighs about 2000 kg, carries two astrometric telescopes separated by a fixed angle of 106.7 degrees. The two telescopes are mounted in the hexagonal optical bench with a ∼3m diameters and merged into a common focal plane (see Fig. 1.3). The aperture of the two telescopes is 1.45m × 0.4m and the focal length is 35m.

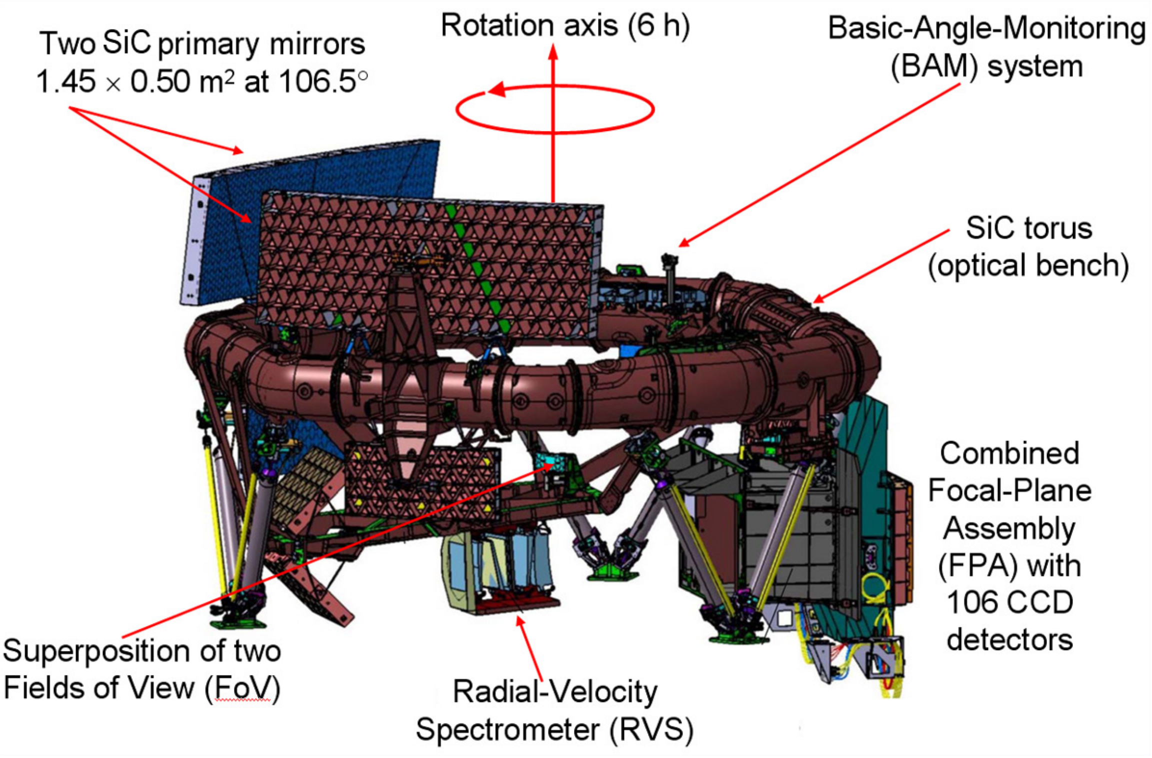
Two SiC primary mirrors
1.45 × 0.50 m^2 at 106.5°
Rotation axis (6 h)
Basic-Angle-Monitoring (BAM) system
SiC torus (optical bench)
Combined Focal-Plane Assembly (FPA) with 106 CCD detectors
Superposition of two Fields of View (FoV)
Radial-Velocity Spectrometer (RVS)

Figure 1.3: Gaia payload module. The figure illustrates the two telescopes separated by a wide-angle of 106.5 degrees, which are mounted on the hexagonal optical bench. The scientific payload has three instruments: the astrometric instrument, the broad-band photometer and the radial velocity spectrometer. The combined focal plane consists of 106 CCDs on which are split between astrometric, photometric and radial velocity spectrograph. (Credit: ESA).

The focal plane is split into 5 CCDs categories arranged in 7 rows and 17 columns (see Fig. 1.4). The first column contains the Basic Angle Monitor (BAM) and WaveFront Sensor (WFS), the Sky Mapper (SM) is located in the two next

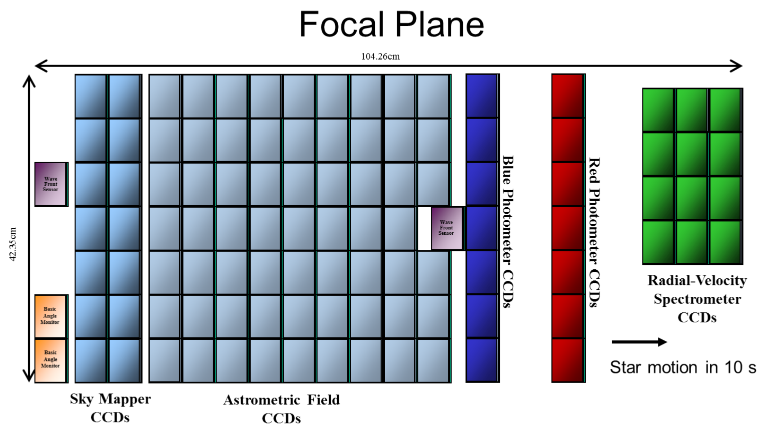

Figure 1.4: Layout of the focal plane assembly. Stars move from left to right of the diagram. The skymappers provide source image detection and two-dimensional position estimation. The astrometric field provides accurate AL measurements and (sometimes) AC positions. The additional CCDs are the blue and red photometers (BP, RP), the radial-velocity spectrometer (RVS), wavefront sensing (WFS), and basic-angle monitoring (BAM). (Credit: EADS Astrium).

columns, which detects objects passing the field of view, allowing the CCDs in the focal plane to read them. The Astrometric Field (AF), which consists of 62 CCDs are formed by the next nine columns. The next two columns indicate the Blue and Red Photometers (BP, RP), and the final three columns are the Radial Velocity Spectrograph (RVS). In total, Gaia focal plane consists of 106 CCDs, and each of them is 4500 by 1996 pixels in size with pixels of 10 micrometres along scan × 30 micrometres across scan size (59 mas x 107 mas). In total the focal plane has 938,000,000 pixels.

1.2.2 Gaia scanning law

The Gaia scanning law is a key aspect of the astrometric performance, designed to optimise delivered final astrometric accuracy (see Fig. 1.5). Gaia simultaneously observes two directions of sight by rotating continuously with a slight precession, and while maintaining the same angle (45°) to the Sun. By precisely measuring the relative positions of the objects of the two viewing directions separated by a wide-angle, the high rigidity of the reference system is obtained. Each object will

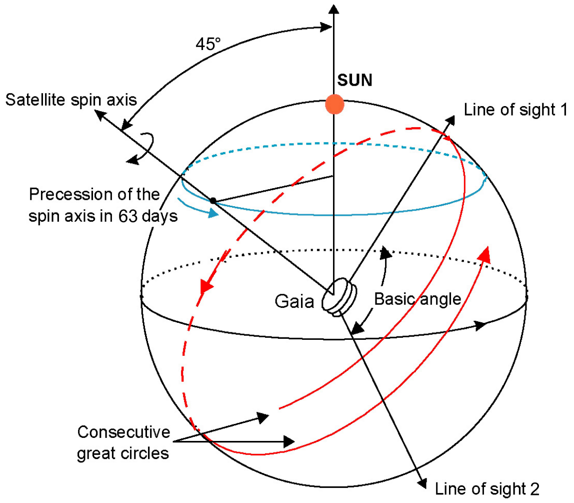

Figure 1.5: The Gaia scanning law. (Credit: ESA).

be observed on average around 70 times during the mission, which is expected to last for 5 years. The three-dimensional motion of Gaia is defined by the combination of four motions which are: the translation around the Sun (a period of one year as that of the Earth), the orbit that Gaia describes around the Lagrangian L2 point with the rotation of Gaia around its own spin axis, and the precession motion (the change in the direction of the Gaia spin axis following a circle).

These measurements will allow the determination of the astrometric parameters of the stars: the projected position in the sky (2 angles), the proper motion (2 values for their derivative with respect to time) and the trigonometric parallax (which provides the distance to each source). In addition, it provides optical broad-band photometry in the G band and colour information in the form of apparent brightness in the G_{BP} (330 - 680nm) and G_{RP} (630 - 1050nm) bands. It also measures radial velocity and high-resolution spectral data in the narrow band (847 - 874 nm).

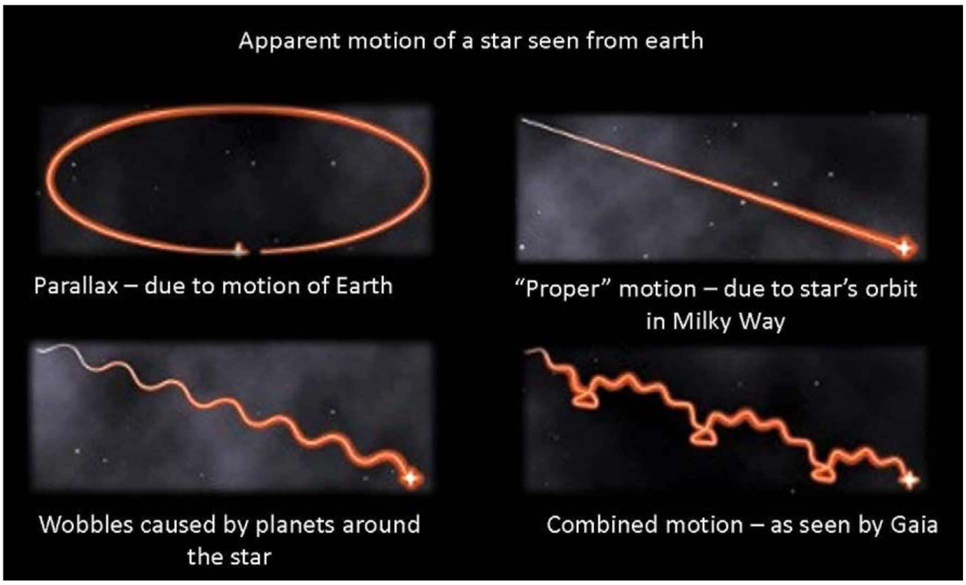
Apparent motion of a star seen from earth
Parallax – due to motion of Earth
"Proper" motion – due to star's orbit in Milky Way
Wobbles caused by planets around the star
Combined motion – as seen by Gaia

Figure 1.6: Illustration of the combination of apparent path of a star across the sky. Figure from Gilmore (2018).

1.2.3 Stellar motion seen by Gaia

Fig. 1.6 illustrates the path of a single star on the sky observed by Gaia. The first motion is the apparent motion of the star or parallax (see Sect. 1.1). The second motion is the proper motion of a star, which tells us how a star is orbiting in the Milky Way and where the star came from. The third motion is the wobble motion caused by planets around a star. The combination of these three motions makes it challenging for Gaia to deal with, and hence a repeated observation of a single star is needed in order to observe the stellar motions (around 80 times). In order to model this complex path, one requires at minimum six parameters (the five global astrometric solutions and the radial velocity), plus sufficient parameters to model multiplicity/planets, and clearly enough precise data over a sufficiently long time to allow a robust fit (Gilmore, 2018).

1.2.4 The Gaia astrometric solution

As described by O'Mullane *et al.* (2011), the central part of the data processing for Gaia is the system known as the Astrometric Global Iterative Solution or AGIS, which was carried out by the Gaia Data Processing and Analysis Consortium (DPAC;Collaboration *et al.* (2016)). All the mathematical tools of the basic observation model were put in the AGIS software (see Lindegren *et al.* (2012)). It is worth to note that there are two different astrometric solutions: the full (five-parameter) solutions with positions, parallaxes, and proper motions; and the fall-back (two-parameter) solutions with only positions.

The astrometric principles for Gaia were outlined already in the Hipparcos Catalogue (ESA 1997, Vol. 3, Ch. 23). The general principle of a global astrometric data analysis was succinctly formulated as the minimization problem:

$$\min_{s,n} \left\| f^{\mathrm{obs}} - f^{\mathrm{calc}}(\boldsymbol{s}, \boldsymbol{n}) \right\|_{\mathcal{M}} \tag{1.2}$$

Where s is the vector of unknowns describing a star's barycentric motion, and n is a vector of *nuisance parameters* describing the instrument and other incidental factors which are not of direct interest for the astronomical problem but are nevertheless required for realistic modelling of the data. g^{obs} represents the vector of all measurements and g^{calc} represents the vector of detector coordinates calculated from the astrometric parameters. The norm is calculated in a metric $\mathcal{M}$ defined by the statistics of the data, this is classically referred to as error weighting (see studies of O'Mullane *et al.* (2011), Lindegren *et al.* (2012), Lindegren *et al.* (2018a) for an in-depth review).

The AGIS software is based on several models which are fully described in the study done by Lindegren *et al.* (2012) including the reference model, the astrometric model, the attitude model, the geometric instrument model, and the signal model. For the reference system, the Barycentric Celestial Reference System (BCRS) is used to model the orbit of Gaia and the light propagation from the source to Gaia. The BCRS is aligned with the International Celestial Reference System (ICRS). The Centre-of-Mass Reference System (CoMRS) is needed for the co-moving celestial reference system having its origin at the centre of mass of the satellite and a coordinate time equal to the proper time at Gaia, and the Geocentric Celestial Reference System (GCRS) for a massless particle. The astrometric model is for the calculation of the proper direction to a source at an arbitrary time of observation. The attitude model describes the instantaneous orientation of the Gaia instrument in the celestial reference frame. The geometric instrument model defines the precise layout of the CCDs. The fitting of the CCD data models produces *observations* that are the input to the astrometric core solution.

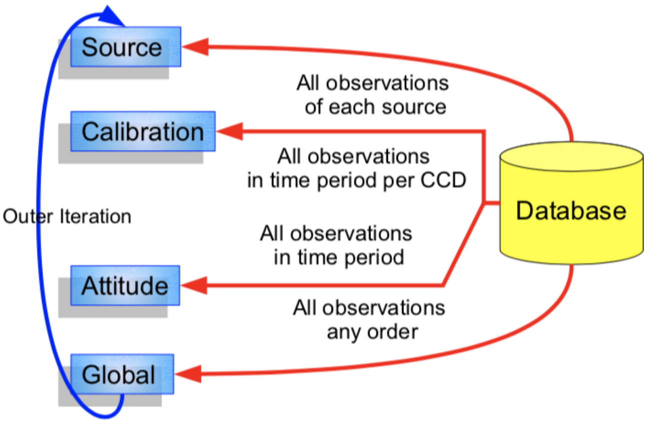
Source
Calibration
Attitude
Global
Outer Iteration
Database
All observations of each source
All observations in time period per CCD
All observations in time period
All observations any order

Figure 1.7: Illustration of the access pattern of AGIS. Figure from O'Mullane *et al.* (2011).

The global solution consist of four main blocks, which are:

- Source: All observations of a given source - spatial.
- Attitude: All observations within a given time period - temporal.
- Calibration: All observations within a given time period falling on a given CCD-temporal/spatial.

- Global: All observations in any order.

The approach goes through the data once for each block, iterate them internally, and then perform the next outer iteration. This is illustrated in Fig. 1.7.

The measurements with Gaia carry a statistical error and a systematic error. As described in the study of Lindegren *et al.* (2018a), the uncertainty in parallax and position at the reference epoch J2015.5 is about 0.04 mas for bright (G < 14 mag) sources, 0.1 mas at G = 17 mag, and 0.7 mas at G = 20 mag for sources with the five astrometric solutions. In the proper motion components the corresponding uncertainties are 0.05, 0.2, and 1.2 mas yr^{-1}, respectively. The model used for the total error in parallax for Gaia DR2 is

$$\sigma_{\text{ext}} = \sqrt{k^2\sigma_i^2 + \sigma_s^2} \tag{1.3}$$

For this, external data must be used to calibrate the model by estimating ω_0, k and σ_0 (see Lindegren *et al.* (2018b)).

1.3 The Interstellar Medium

The space between the stars of the galaxy is not empty. It is filled with gas (atoms, molecules, ions and electrons) of very low density in which small solid particles are mixed. This medium constitutes approximately 5% of the mass of stars in the galaxy which is itself composed mainly of hydrogen (more than 70% of its mass). Different phases characterise the ISM according to the temperature, the fractional volumes and the densities of hydrogen present: molecular clouds, H II regions, the cold neutral medium (CNM), the warm neutral medium (WNM), the warm ionised medium (WIM), and the hot ionised medium (HIM). The characteristics of these phases are explained below and summarised in Table 1.1.

Table 1.1: Components of the different phase of the ISM according to McKee & Ostriker (1977)

Component	Temperature	Volume	n_H
	(K)	(%)	cm^{-3}
Molecular clouds	20 - 50	< 1	$10^2 - 10^6$
H II regions	10^4	~ 10	$\sim 10^3$
Cold Neutral Medium (CNM)	50 - 100	1 - 5	20 - 50
Warm Neutral Medium (WNM)	6000 - 10000	10 - 20	0.2 - 0.5
Warm Ionized Medium (WIM)	~ 8000	20 - 50	0.2 - 0.6
Hot Ionized Medium (HIM)	$\sim 10^6$	30 - 70	~ 0.0065

Molecular clouds

This phase occupies a small part of the ISM as it is only concentrated in clouds. Following Van Dishoeck *et al.* (1988) and Ferriere (2001), the molecular cloud can be classified into three categories based on the visual extinction, A_V, along the line of sight towards the cloud. i) Diffuse clouds are marked by $A_V < 1\ mag$ and made of cold atomic gas ($T \sim 100\ K$). ii) Dark clouds are identified by $A_V > 5\ mag$, and made of icy gas ($T \sim 10 - 20\ K$). iii) translucent clouds are characterised by an intermediate value of A_V, made both from molecular and atomic gas. A molecular cloud is the birthplace of stars. The dark part of the molecular cloud contains the pre-stellar heart, which is protected from the outside environment. This dark part is identified by using the lines of carbon monoxide molecule (CO) rather than using the molecular hydrogen (H_2) as it is very difficult to observe. The reason for the difficulty of observing H_2 molecules is that they are perfectly symmetric and homonuclear diatomic molecules, so the spectral lines of H_2 are extremely weak. Because of the impact of gravity, molecular cloud divides into hierarchical substructures of clumps and core. This gravitational collapse is the beginning of star formation.

H II regions

H II regions are regions that are composed of ionised hydrogen in which massive star formation has recently taken place. H II regions are emission nebulae created when young main-sequence OB stars ionise nearby gas clouds with intense UV radiation. In our Galaxy, H II regions distribute in a similar way as the molecular clouds and can be seen across the entire Galactic disk (Balser *et al.*, 2011).

Cold Neutral Medium (CNM)

The cold neutral medium is composed of clouds and filaments (Heiles & Troland, 2003). Compared to the molecular cloud, the cold neutral medium occupies a larger part of the ISM but is less voluminous. CNM is not easy to observe as the molecular cloud. However, the existence of the CNM is revealed by emission in the 21cm hyperfine transition of H I, with a median temperature of approximately 70 K, determined by a comparison of emission to absorption spectra.

Warm Neutral Medium (WNM)

The warm neutral medium occupies a large volume of the ISM, but it is not very dense. This phase is the warmest neutral phase of the ISM, which can be observed by using the 21 cm HI emission line. The main source of the excitation of the CNM and the WNM is the photoelectric emission of PAHs.

Warm Ionized Medium (WIM)

After the birth of new massive stars in the ISM, a warm ionised gas is created around them. The WIM is created by the leakage of photons out of the HII regions. The WIM is approved as a major component of the interstellar medium of the Milky Way and other disk galaxies (Haffner *et al.*, 2009).

Hot Ionized Medium

The hot ionised medium surrounds the molecular cloud, H II regions, WIM, WNM, and CNM. The propagation of shocks from supernovae is the main responsible for the increase of the temperature for this phase of ISM (Ferriere, 2001). Shocks propagate more easily in a sparse environment and can distribute energy over large volumes in the Galaxy. The cooling process may remain millions of years, and it makes this phase occupy a large volume in a Galaxy.

1.4 Low Mass Stars

A review of the process of star formation produced by Larson (2003) and Kennicutt Jr & Evans (2012), together with previous work on star formation is summarised in this work. Understanding the formation of a star is of major importance in this study to derive their distance.

The process of star formation is summarised as follows:

Stars are born from the material of the ISM (see Sect. 1.3), the Giant Molecular Cloud (GMCs). The birth of a star is a result of the action of gravity, which makes a large number of hydrogen atoms combine. Charged and neutral particles of GMCs are supported by thermal and turbulent pressure to resist gravity. The ionised particles are linked to the magnetic field of the cloud, therefore, prevented from collapse, while neutral atoms are not. The process is known as *ambipolar diffusion*. Ambipolar diffusion results in a loss of turbulent and magnetic support for the cloud because the fractional ionisation of the GMCs is quite low. It is assumed that after the absence of magnetic and turbulent support, the only force preventing collapse due to gravity is the thermal pressure. At this stage, the beginning of collapse is described by the virial theorem:

$$2K + U = 0 \tag{1.4}$$

where K is the internal kinetic energy of the cloud, and U is the gravitational potential energy of the spherically symmetric uniform cloud. Eq. 1.4 shows the condition for a stable system, assuming that the external pressure ≈ 0.

This equilibrium does not last long as certain factors can break the balance whether to trigger a gravitational collapse or a cloud dissipation. This disturbance can be caused by either the explosion of a supernova or the passage of the cloud in an area of a high density of matter. For the first factor, a supernova gives rise to a tremendous shock wave which violently compresses the regions it crosses and can, therefore, cause the gravitational collapse of a GMC. And for the second factor, when a giant molecular cloud crosses one of the dense areas in our Galaxy, it undergoes a compressive force which can cause a gravitational collapse as our Galaxy does not have a uniform distribution of matter, but contains denser areas than the average (Roman-Duval *et al.*, 2016).

Once this stability is broken, GMCs start to contract. This process of contraction was studied by the British physicist *James Jeans* who showed that a cloud of gas subjected to the opposite demands of the force of gravitation and the internal pressure ends up contracting if its mass is higher than a certain threshold called Jeans' mass.

$$M_J = \left(\frac{5kT}{G_{\mu m_H}}\right)^{3/2} \left(\frac{3}{4\pi\rho}\right)^{1/2} \tag{1.5}$$

where k is the Boltzmann's constant, T is the temperature, μ is the mean molecular weight, m_H is the mass of a hydrogen atom, G is the gravitational constant, and ρ is the density in kg/cm^{-3}.

Once Jeans' mass has been exceeded, the collapsing material in the molecular cloud is essentially in a free-fall (Shu, 1977). The collapse is isothermal since the cloud is optically thin, and so Jeans' mass decrease as the density increase. Since then, the cloud will not only contract, but it also begins to fragment into smaller blocks, in a process which is known as *fragmentation.* A new series of fragmentation begins, and each of the blocks subdivides itself into smaller and denser clouds, and Jeans' mass, therefore, continues to decrease and so on.

A series of divisions are unfolded and give rise from a giant cloud to a large quantity of smaller and smaller fragments. The fragmentation of the cloud leads to a clumpy structure in the molecular cloud. However, the fragmentation process eventually stops when the temperature of the cloud begins to rise, which increases Jeans' mass. The smallest clouds, which appeared when the critical threshold was at their lowest, are then too small to fragment and the whole process of fragmentation stops.

When the fragmentation stops, each small cloud of gas continues to contract and heat up by converting its gravitational energy into thermal energy to form a "protostar". At temperature $\approx$ 2000 K, hydrogen molecules (H_2) begin to dissociate. The energy lost from this dissociation leads to a decrease in the gas pressure, which makes the gravity dominate again and free-fall collapse reoccurs. An accretion disc will form around the protostar due to the overall angular momentum vector at the initial rotating core. Once the protostar has become a hydrogen-burning star, where the temperature is hot enough, a strong stellar wind/jet forms along the axis of rotation (Fig. 1.8). This feature is easily seen by radio telescopes. Stellar jets eventually dissipate the surrounding materials as they widen, leaving a remnant disc from which planets may eventually form. The star begins its main sequence phase once the accretion of materials stops.

1.5 Massive stars

Massive stars are defined as stars that have a typical mass $M \geq 8M_\odot$ (Guzmán *et al.* (2016), Billington *et al.* (2019), Urquhart *et al.* (2015)), which imply that they already begun hydrogen-burning before accretion stage has finished. This causes a difficulty in the study of the processes of high-mass star formation since we cannot observe separately the luminosity of the star due to accretion and the intrinsic luminosity of the protostar (Ward-Thompson & Whitworth, 2011).

However, studying massive stars are of great importance in astronomy be-

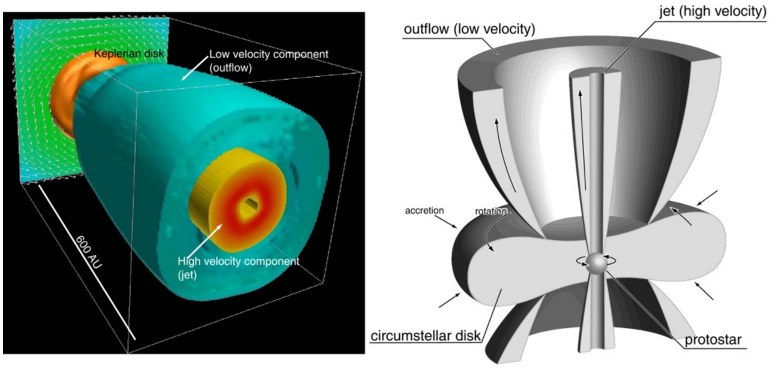

Figure 1.8: Protostellar outflow and jet diagram from Machida (2017). The image on the left shows a three-dimensional view of outflow, jet and circumstellar disk in the mass accretion phase while the Schematic view around a protostar is shown in the right.

cause of their feedback effect on the evolution of the universe. They are dominant sources of energy in the ISM, as a form of mechanical (powerful stellar winds, outflows, supernova shock waves), and radiation (by creating HII regions) (Schilke, 2017). They enrich the ISM with heavy elements (aluminium, gold, iron, etc ...) through supernova explosion (Maund *et al.*, 2017). In addition, studying massive stars are essential in the study of galaxy emission and evolution in the distant universe (Stanway, 2016). Since their luminosity dominates the stellar population in the galaxy, the observed spectra and colour of the evolution of the galaxy are dominated by the evolution of its high mass stars (Hartquist, 2011).

Despite the importance of understanding massive stars, there are still a few factors that make them arduous to study. They are located very far away as they are usually formed in distant clusters (Beltrán *et al.*, 2006). They evolve much faster than their low-mass counterpart. High-mass stars are very rare and short-lived, being usually deeply embedded into their natal environment throughout their very early evolutionary stages (Roman-Lopes, 2013). Thus, observing them requires high-resolution instruments at IR or a longer wavelength. Moreover, massive stars do not appear to form in isolation, they almost always appear to form in clusters (De Wit *et al.*, 2005), making a full study of a single massive star difficult.

Consequently, the physical processes that dominate high mass star formation, especially in their early stage of evolution, are not yet fully understood and still under active study. Questions remain unanswered whether their formation is significantly different from their low-mass counterparts or whether they form in the same way by triggering a gravitational collapse. Different concept have been introduced in the literature to analyse the origin of massive star formation (e.g.: Krumholz & Bonnell (2007), Zinnecker & Yorke (2007), Schilke (2015)). So far, there are three well-known models of massive star formation, which are i) monolithic collapse and disk accretion, ii) competitive accretion and runaway growth, and iii) stellar collisions and mergers (see the review by Zinnecker & Yorke (2007) for details).

1.6 Star clusters

A stellar cluster is a group of stars born from the same molecular cloud and which are still linked by gravity. The stars that make up a cluster, therefore, have the same age and the same chemical composition (Krumholz *et al.*, 2019). Star clusters exist in different forms, ranging from the fragile association of a few dozen members to the dense aggregate of millions of stars. Following Krumholz *et al.* (2019), stellar clusters are classified into two different types depending on the conditions in which they were formed: open clusters and globular clusters. Open clusters are groups of a few tens to a few thousand stars that form in molecular clouds of the galactic plane. They are generally young (≤ 1 Gyr) and have low mass ($\leq 10^5 M_\odot$) (Krause *et al.*, 2020). Their mass is not large enough for the stars to remain clustered for more than a few million to a few hundred million years. Open clusters are typical objects of the Galaxy disk mainly composed of stars of population I. Globular clusters, on the other hand, are made up of a few tens of thousands to a few million stars that are gravitationally bound. They have a dense spherical structure and have a higher stellar density towards their centre. They are generally old (>1 Gyr) and massive ($\geq 10^4 M_\odot$) (Krause *et al.*, 2020). Globular clusters are typical objects of the halo of the Galaxy and composed of stars of population II.

Since large fractions of star formation happen in a clustered environment (Lada & Lada, 2003), understanding star clusters are important in astronomy to study the origin and early evolution of stars and planetary systems. In addition, some of the star clusters have a long life (> 13 Gyr), making them helpful for the analysis of the age of the Universe (O'Malley *et al.*, 2017). Another importance of the study of star clusters also is the effect of their feedback on the universe. Massive stars that form in clusters are the main responsible for feedback in galaxies (Goldbaum *et al.*, 2016). This latter study showed that feedback suppresses galaxies' star formation rates and also leads to the formation of a multi-phase atomic and molecular interstellar medium. The enormous amounts of ionizing radiation from those young stellar clusters have a huge impact on the galaxy and its surroundings. Thus understanding the massive star clusters leads to a better understanding of the evolution of the universe.

Clusters form in massive dense cores of molecular gas that are strongly self-gravitating (Lada & Lada, 2003). Cluster formation has three phases according to Krause *et al.* (2020). The first phase is the creation of the highly transient and inhomogeneous molecular cloud structure due to the action of supersonic turbulence. The second phase is the formation of individual stars in clusters and association from the cloud structure in the first phase. The last phase is characterised by the stellar feedback that makes the disappearance of the remaining gas core.

Table 1.2: Summary of the YSOs classification

Class	Emission	Average stage duration
Class 0	Submillimeter	10^4 yrs
Class I	Far-infrared	10^5 yrs
Class II	Near-infrared	10^6 yrs
Class III	Visible	10^7 yrs

1.7 Young stellar objects (YSOs)

YSOs are stars that are in their early evolutionary stage. They are pre-main sequence stars and characterised by the presence of circumstellar material. There are four main classes of low-mass YSOs, which are classified according to the peak and shape of their spectral energy distribution (SED). Table 1.2 summarizes the different classes of YSOs and their characteristics.

The very early protostar stage is called “Class 0”, which was discovered by André (1994). Class 0 are YSOs that are extremely obscured by a large circumstellar material (radius $\sim$ 10,000 AU). It is characterised by a strong submillimeter emission, not a near-infrared or mid-infrared emission. “Class 0” protostars are rarer and correspond to the youngest protostar stage known to date (probable age $\sim 10^4$ yrs), and only the blackbody emission of the cool dust in the envelope is visible.

The other three classes of protostars (“Class I”, “Class II” and “Class III”) were found by Lada (1987). They can be distinguished based on the slope of their Spectral Energy distributions (SEDs) in the mid-infrared. “Class I” protostars are supposed to be formed at age $\sim\ 10^5\ yrs$, and they are characterised by the positive SED slope at near and mid-IR wavelength. As the gravity continues to shrink the molecular cloud, it makes the circumstellar spin faster which flattens into a disk with a central bulge. The radius of the circumstellar disc for “Class II” protostar is estimated to be around $\sim$ 500-1000 AU, which is 10 - 20 times the size of our solar system. “Class II” protostar eventually become optically

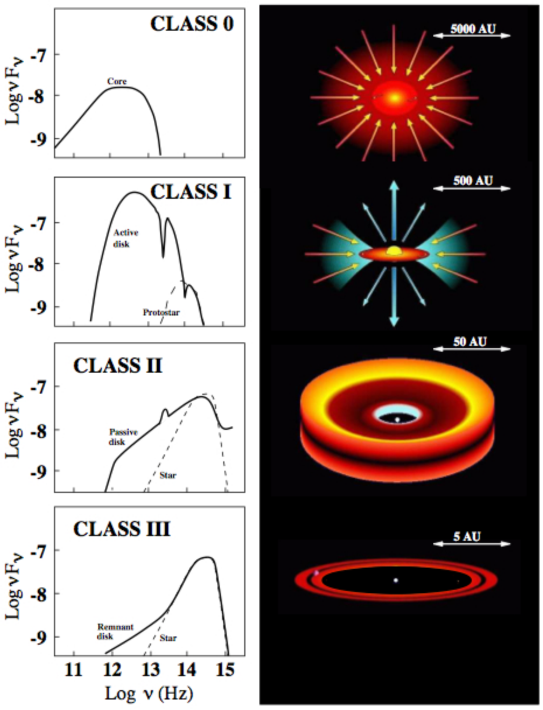

Figure 1.9: Evolutionary classification of YSOs. Source: Isella (2006)

visible on the stellar birth line as pre-main-sequence stars when the envelope material disappears, and only the disk remains around the protostar. In terms of their SED, as they are no longer obscured by the dusty envelope, the observed flux will become more noticeable, but the dust within the disk will continue to provide excess at near-infrared wavelength. At this stage, we have the classical T Tauri. “Class III” protostars are stars that have no surrounding disks. This corresponds to weak-line T Tauri star (WTTS), and also called “naked”. The transition between “Class II” and “Class III” stages is called “transition-disk”, and it appears to take place between $\sim 10^6$ and $\sim 10^7$ *yrs* (Beskin *et al.*, 2003). This evolutionary sequence and the typical spectral energy distribution (SED) of a protostar is shown in Fig. 1.9.

1.8 Stellar Extinction

Extinction is the combined effects of scattering and absorption of electromagnetic radiation emitted from an object. The incoming photon is not destroyed in scattering, but they are changed in direction. Contrary, in the absorption, the incoming photon is destroyed, and its energy remains in the dust grain. Interstellar extinction is quantified as the number of magnitudes by which a molecular cloud dims starlight passing through it.

In the presence of the extinction, the observed intensity of light, I_λ, at wavelength, λ, is given by:

$$I_\lambda = I_{\lambda 0} exp(-\tau_\lambda) \tag{1.6}$$

where $I_{\lambda 0}$ is the intensity of that would be received at the Earth in the absence of interstellar extinction along the line of sight and τ_λ is the optical depth at the observed wavelength.

Stellar extinction varies as a function of wavelength (Draine, 2003). The extinction is stronger for short wavelength, which has the effect of making the sources redder. The dependence of the extinction on the wavelength is described by the *extinction law*. The variation of extinction with wavelength λ is usually expressed as a ratio of colour excesses, $E(\lambda - V)/E(B - V)$, or of the absolute extinction A_λ/A_V, where B and V refer to the optical bands. Cardelli *et al.* (1989) characterise the extinction curve by a single parameter, which is the ratio of the total extinction over the selective extinction:

$$R_V = \frac{A_V}{E(B - V)} \tag{1.7}$$

R_V is the total-to-selective extinction ratio that describes the variations in UV/optical bands (Fig. 1.10). The visible part of the spectrum is roughly proportional to λ^{-1}, while the UV is characterised by the hump in each curve. In almost all regions, the value of R_V is close to 3.1 (Draine, 2003).

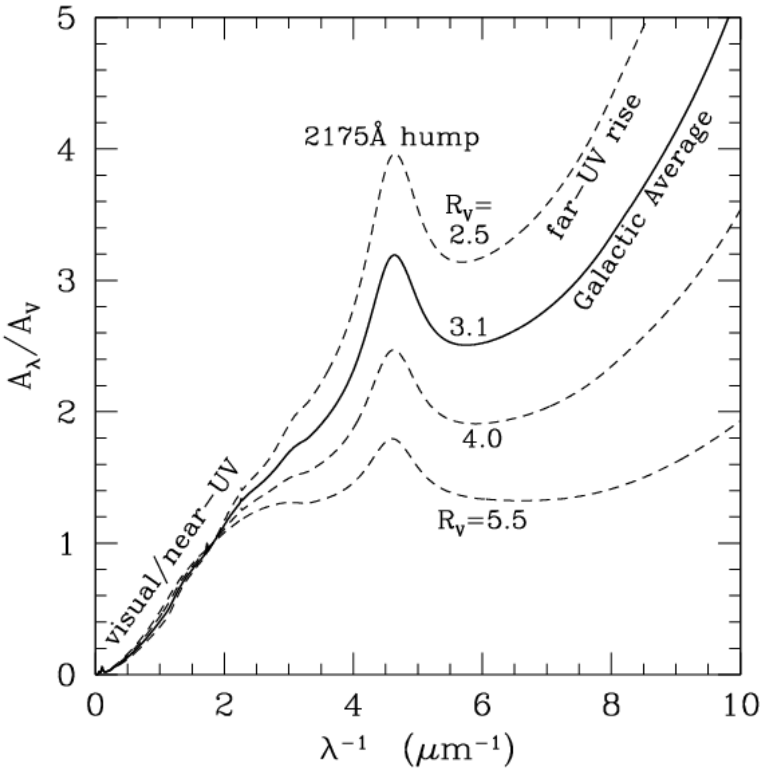

Figure 1.10: Interstellar extinction curves of the Milky Way ($R_V = 2.5$, 3.1, 4.0, 5.5) from Li & Mann (2012). It indicates that dust grains on different sightlines have different size distributions. The regional variations in the Galactic optical/UV extinction curves is characterized by the total-to-selective extinction ratio R_V.

1.8.1 Determination of stellar extinction

Determination of extinction is essential in any field of astrophysics for the study of objects at distances greater than a few dozen parsecs. Although the estimation of extinction of a single star is not within the scope of this work, it is important to provide a brief overview of the technique that can be used for the determination of the extinction. One of the oldest techniques used is the *star counting* method. This provides an extinction value from the comparison of the number of stars in magnitude intervals in an extinguished field, and the number of stars in a near reference field assumed without extinction. The star counting method has been used by different studies such as Dobashi *et al.* (2005), Stead & Hoare (2010), and Foster *et al.* (2012). Another technique is the *colour excess* method, which is a measure of the reddening of a star due to interstellar dust. The colour excess is the difference between the observed star colour and its intrinsic colour corresponding to the star's spectral type. This can be written as $E_{\lambda_1-\lambda_2} = A_{\lambda_1}-A_{\lambda_2} = (m_{\lambda_1}-m_{\lambda_2})-(m_{\lambda_1}-m_{\lambda_2})_{int}$, where $(m_{\lambda_1}-m_{\lambda_2})_{int}$ denotes the intrinsic colour of the star, $(A_{\lambda_1}, A_{\lambda_2})$ the total extinctions in the given photometric band, and $(m_{\lambda_1} - m_{\lambda_2})$ the observed colour. This is the most used method in the literature. For instance, Lada *et al.* (1994) and Alves *et al.* (1998) used this technique by mapping and measuring the distribution of dust extinction through a molecular cloud. The method was then generalized by Lombardi & Alves (2001) and Lombardi (2009). Lombardi *et al.* (2006) also calculated the extinction on the $K_S - band$ with a high resolution 8° × 6° extinction map of the Pipe nebula using 4.5 million stars from 2MASS PSC (Skrutskie *et al.*, 2006).

1.8.2 The extinction A_G

The extinction in the G band (A_G) that we will be using in this work is one of the stellar parameters estimated by the Gaia astrophysical parameters inference system (Apsis) (Bailer-Jones *et al.*, 2013). The main goal of the Apsis data

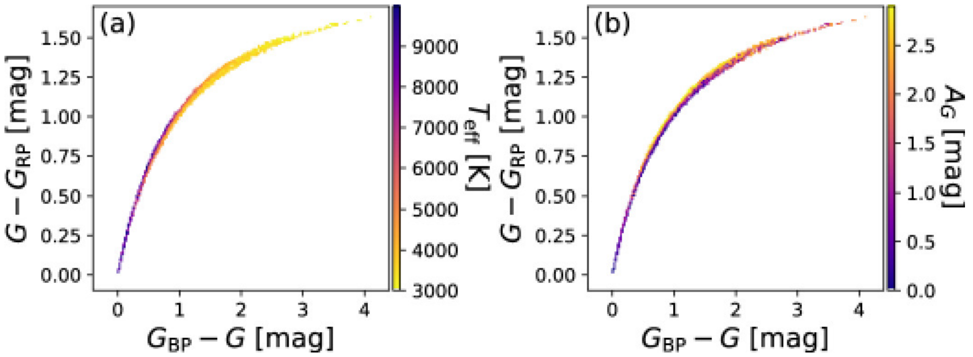

Figure 1.11: Colour–colour diagrams for stars from the PARSEC 1.2S models with an extinction law from Cardelli *et al.* (1989) and [Fe/H] = 0. Panel (a) represents the dominant factor T_{eff} in the Gaia colour, which degenerate with A_G extinction (panel b). (Credit: Andrae *et al.* (2018))

processing pipeline is to classify and estimate astrophysical parameters for the Gaia sources using Gaia data. It consists of two algorithms called Priam and FLAME. Priam is used to infer T_{eff}, A_G, and $E(BP - RP)$, while FLAME is used to estimate stellar luminosities, masses and ages of stars. The summary of how they obtained A_G using the Priam algorithm is provided below:

Gaia has three photometric bands, the first one is the G band, and the two other bands (G_{BP} and G_{RP}) were obtained from the integration of the Gaia prism spectra. Andrae *et al.* (2018) indicated that the colour is strongly influenced by T_{eff}, so it is impossible to derive the extinction using only the colours (see Fig. 1.11). Therefore they combined the three bands photometry with parallax ϖ to infer the line-of-sight extinction and colours $E(BP-RP)$ of a star. As they used magnitudes and parallax, rather than the colours, the available signal is primarily the dimming of the sources due to absorption (Andrae *et al.*, 2018).

The estimation of A_G was performed using a machine learning algorithm with a univariate output called ExtraTrees (Geurts *et al.* (2006)). Once they obtained the distance by simply inverting the parallax, they computed the quantity $M_X + A_X$ using the equation

$$M_G = G - 5\ log_{10}\ r + 5 - A_G \tag{1.8}$$

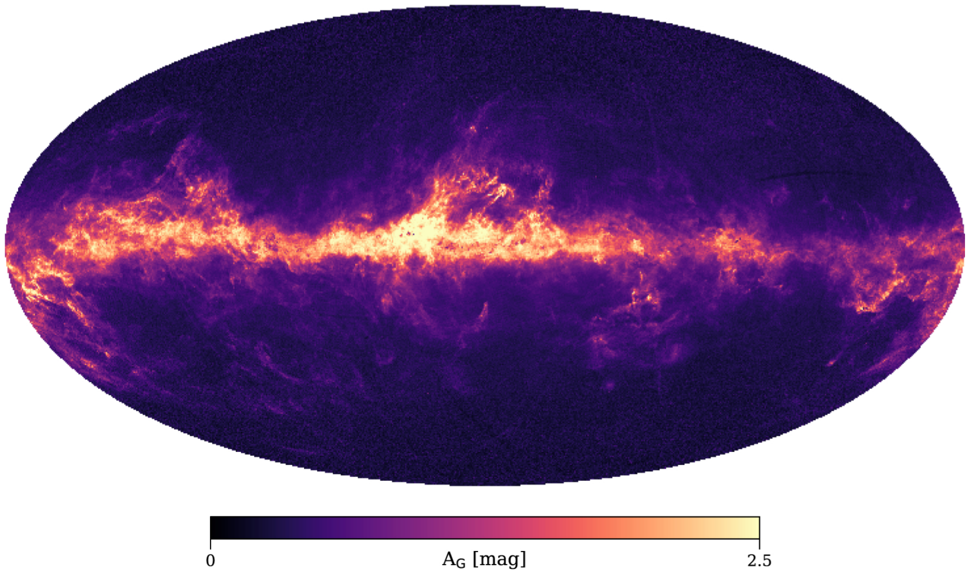

Figure 1.12: Distribution of A_G in Galactic coordinates. The map is centered on the Galactic Center, with longitudes increasing towards the left. Figure from Andrae *et al.* (2018).

Where M_G is the absolute G-band magnitude and $r = \frac{1}{\varpi}$ is the distance.

Andrae *et al.* (2018) also mentioned that there are very few reliable literature estimate of the extinction, and published estimates are of A_V and/or $E(B - V)$ rather than A_G and $E(BP - RP)$. Consequently, they had to train on synthetic stellar spectra to estimate the A_G. The features they used for EXTRATREES were the three observables $M_G + A_G$, $M_{BP} + A_{BP}$, and $M_{RP} + A_{RP}$. Since EXTRATREES cannot extrapolate from the training data range, it does not estimate a negative results for A_G. This non-negativity results of A_G means that it cannot be Gaussian, but it should be truncated Gaussian. We also follow this truncated Gaussian distribution to model the A_G extinction in this work (see Sect. 3.3.1).

The distribution of A_G in Galactic coordinates (Mollweide projection) is presented in Fig. 1.12, showing that the mean extinction is reliable. Although they used parallax to reinforce the three optical bands for the line-of sight-extinctions estimate, the results they obtained may not very accurate. The results cannot be compared to the literature as the literature infer A_B and A_V, not A_G. In addition,

A_G suffers from large uncertainties (0.46 mag) so the use of an individual star is limited. For this reason, they advised that the A_G estimates from their work should be used statistically, and with a group of stars. The mismatch between the training sets (synthetic data) and the features (real Gaia data) they used for Extratrees is only of the order of ∼0.1mag or less in the zeropoints (Evans *et al.*, 2018). Thus, it did not lead to obvious systematic error.

1.8.3 The extinction A_V

Here we summarise the method used by Anders *et al.* (2019) to derive stellar parameters, extinction and distances for stars brighter than G = 18. They used a python code called StarHorse developed by Queiroz *et al.* (2018). It is a robust code that improves the distances derived by Bailer-Jones *et al.* (2018) and other stellar parameters. StarHorse basically compared the observed set of astrometric, photometric, and spectroscopic data to stellar evolutionary models to infer stellar parameters, extinction and distances. They used a prior that contains a stellar initial mass function, density laws for the main components of the Milky Way (thin disc, thick disc, bulge, and halo), the metallicity, and age prior for those components. If Y is the set of measured parameter that they used, Y can be written as $Y = \{T_{eff}, log\ g, [M/H], m_\lambda, \pi\}$, where T_{eff}, $log\ g$, $[M/H]$, m_λ, and π is the effective temperature, surface gravity, metallicity, the apparent magnitude, and parallax respectively. Given those observed quantities, StarHorse inferred the posterior distribution of the set of parameter θ that contains $\theta = \{m_*, \tau, d, A_V\}$ where m_*, τ, d, A_V is the mass, age, distance and visual extinction respectively. StarHorse also produced a whole 3D tomographic of the Galaxy, which is better than the extinction derived from star counts method (Queiroz *et al.*, 2018). It provided the posterior probability distribution of a star over a grid of stellar models, distances, and extinction (Queiroz *et al.*, 2018). The algorithm is summarized in Fig. 1.13.

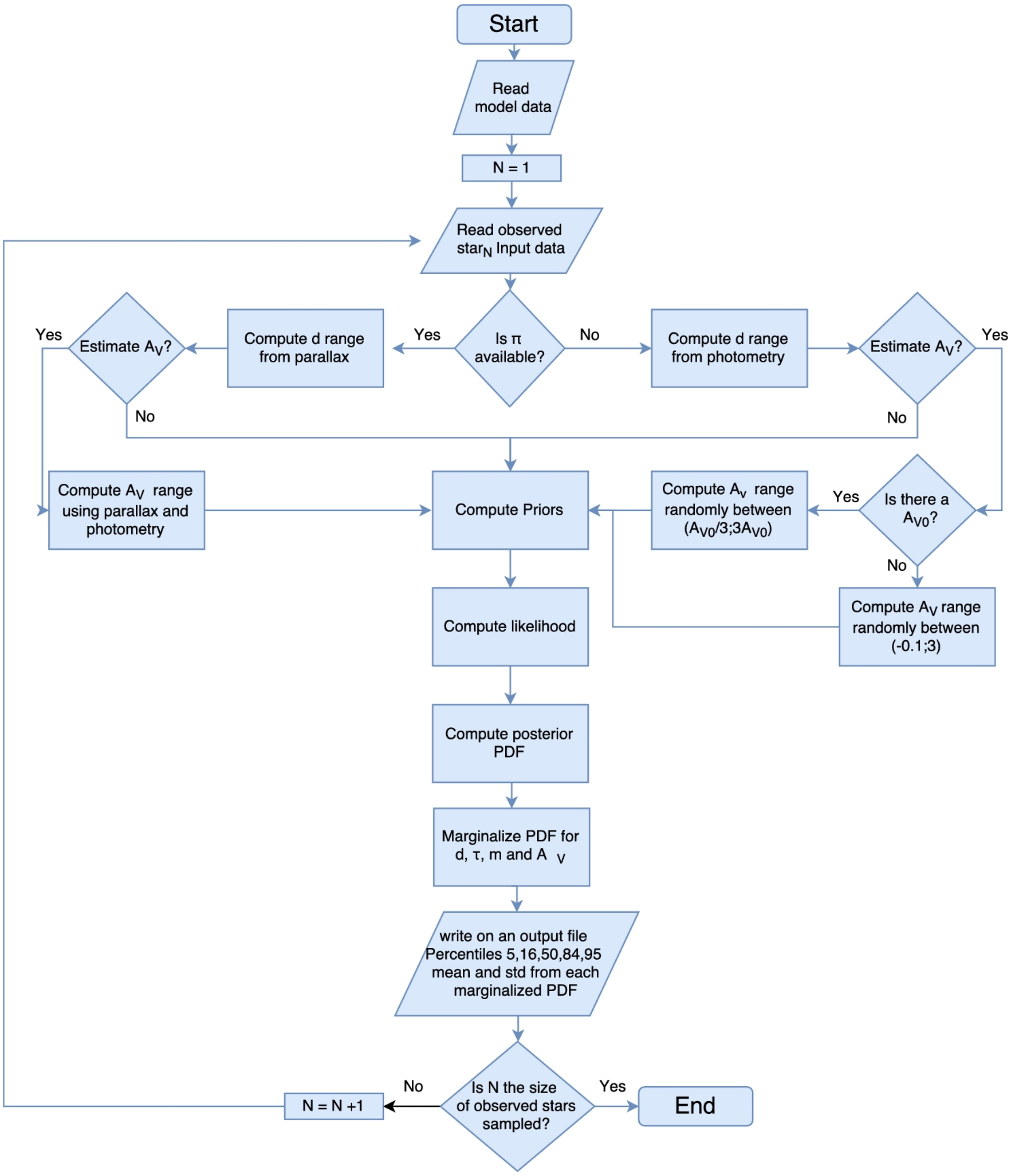
Start
Read model data
N = 1
Read observed star$_N$ Input data
Is π available?
Yes
No
Compute d range from parallax
Compute d range from photometry
Estimate A_V?
Yes
No
Estimate A_V?
Yes
No
Compute A_V range using parallax and photometry
Compute Priors
Compute A_v range randomly between (A_{V0}/3;3A_{V0})
Yes
Is there a A_{V0}?
No
Compute likelihood
Compute A_V range randomly between (-0.1;3)
Compute posterior PDF
Marginalize PDF for d, τ, m and A V
write on an output file Percentiles 5,16,50,84,95 mean and std from each marginalized PDF
No
N = N +1
Is N the size of observed stars sampled?
Yes
End

Figure 1.13: StarHorse flow diagram Queiroz *et al.* (2018). The diagram illustrates the algorithm used to infer stellar parameters, distance and extinction with the StarHorse method.

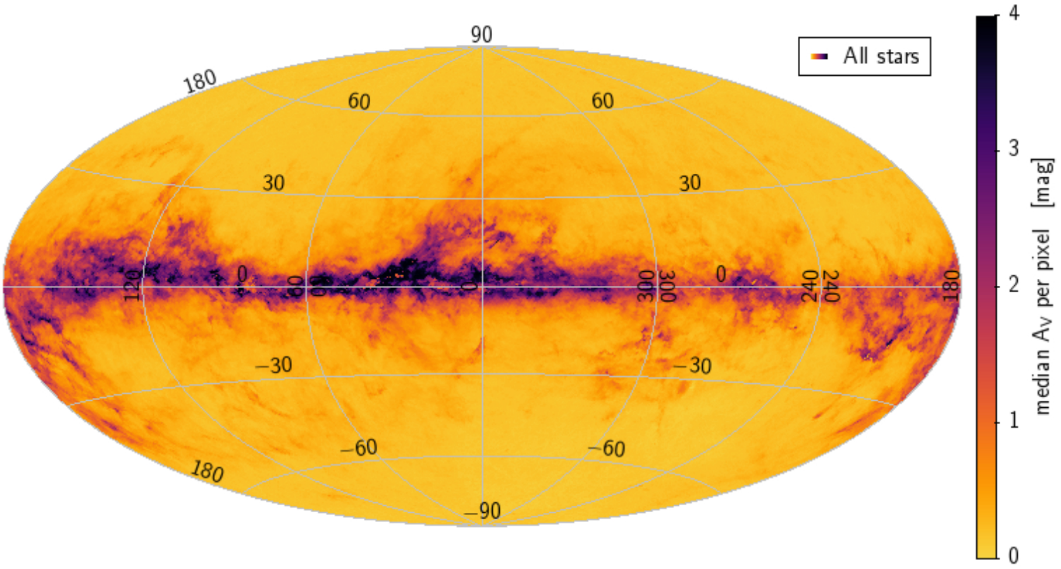

Figure 1.14: All-sky median *StarHorse* extinction map (Aitoff projection). Figure from (Anders *et al.*, 2019).

As Anders *et al.* (2019) used a broad-band optical Gaia passband, they improved the StarHorse code by taking a better account in the extinction when comparing synthetic and observed photometry catalogues. They pointed out that the dust-attenuated photometry of very broad photometric passbands such as the Gaia DR2 should take into account that passband extinction coefficient A_i/A_V varies as a function of its source spectrum F_λ (T_{t_eff}) and the extinction A_V. Therefore, they took into account the coefficients A_i/A_V for each stellar models and the extinction in the StarHorse code. The passband extinction coefficient is given by the relation

$$\frac{A_i}{A_V} = \frac{2.5}{A_V} \cdot \log_{10} \frac{\int F_\lambda \cdot T_\lambda^i \mathrm{d}\lambda}{\int F_\lambda \cdot T_\lambda^i \cdot 10^{-0.4 a_\lambda \cdot A_V} \mathrm{d}\lambda} \tag{1.9}$$

where T_λ^i is the transition curve, and a_λ is the extinction law that they adopted from work done by Schlafly *et al.* (2016). For the stellar model discussed above, they used the Kurucz grid of stellar spectra Kurucz (1993) for the computation of the bolometric corrections (as a function of T_{eff} and A_V) and for the default extinction law. The StarHorse-derived extinction is shown in Fig. 1.14. The code reached different regions of the Milky Way such as the bulge, halo, and outer disc, which were not achieved with the use of Gaia data alone.

1.9 Bayesian astrostatistics

1.9.1 The MCMC based Bayesian analysis in distance estimates

The Monte Carlo based Bayesian analysis is one of the most advanced techniques used in several disciplines during the last decades. The Bayesian statistics is a theory that allows us to analyse observed data while the Markov chain Monte Carlo is used to sample from a distribution. The combination of those two techniques leads to an impressive result in all disciplines of science, including astronomy (Sharma, 2017).

Numerous studies have used Bayesian techniques and MCMC to infer distances from the measured photometric and spectroscopic. The spectro-photometric distances using the Bayesian technique was introduced by Breddels *et al.* (2010). They derived absolute magnitude and distances for RAVE stars using stellar models and spectroscopic data from RAVE's survey. This method was then extended by Binney *et al.* (2013), Schlafly *et al.* (2014), and Santiago *et al.* (2016) by taking account of extinction to their work.

The use of Bayes's theorem dramatically increased since the coming of parallax measurements from the Gaia mission, which is a survey dedicated to astrometry. Anderson *et al.* (2018), for instance, used the Extreme Deconvolution (XD) algorithm to improve Gaia parallax precision with the use of the photometric data from 2MASS.

Recently, (Leistedt & Hogg, 2017) used a hierarchical Bayesian estimate to derive distances to 1.4 million stars of the Gaia DR2. Their method focuses on the use of the available data from Gaia itself, and without the use of a stellar model. Besides, Hawkins *et al.* (2017) used Gaia to derive the intrinsic magnitude, absolute magnitude and dispersion of helium burning RC stars. Bailer-Jones *et al.* (2018) also use Bayesian inference techniques to derive distances to 1.3 billion stars, which are provided in the Gaia catalogues. The spectro-photometric distances cited above was improved by Anders *et al.* (2019) by adding the astrometric data from Gaia with additional photometric and spectroscopic data from several surveys (described in Sect. 2).

For a molecular cloud, the MCMC based Bayesian analysis also is used in several works in the literature. Zucker *et al.* (2018), Zucker *et al.* (2019), Yan *et al.* (2019a), and Yan *et al.* (2019b) performed distance measurements to several clouds with the use of Gaia parallaxes. For the use of Gaia parallax, the choice of the likelihood that can be used for the parameter of inference is summarised by Hogg (2018).

The method used in this thesis is based on Bayesian statistics. A brief history and the basic theory of Bayesian data analysis is described in this section.

Back in history, the term "*Bayesian*" refers to Thomas Bayes who is an English Presbyterian minister, statistician, and mathematician (1702-1761). The term "Bayesian" is used after the publication of his work *"An Essay towards solving a Problem in the Doctrine of Chances (1763) "*.

Currently, Bayesian analysis is becoming popular and is widely used in many different fields of research such as astronomy, public health and economics. According to Loredo (2013), the Bayesian approach was first used in astronomy in the late 1970s. In a nutshell, Bayesian statistics allows us to update our knowledge of physical parameters using a new set of observations. It is often used to estimate parameters and their uncertainties from a set of model parameters without knowing the shape or scale of their respective distributions.

1.9.2 Bayes' Rule

The simple equation of Bayes's theorem is

$$p(\Theta|data) = \frac{p(\Theta)p(data|\Theta)}{p(data)} \tag{1.10}$$

where Θ is the parameter of interest to be inferred, $p(\Theta)$ is the called "prior" (see Sect. 1.9.4), $p(data|\Theta)$ is the *likelihood* (see Sect. 1.9.3), and $p(\Theta|data)$ is the posterior distribution of the parameter given a set of observation.

The final goal of the Bayesian data analysis is to produce a posterior probability distribution of the parameter of interest or the event. It considers first the prior uncertainty about the model parameters with a probability distribution and updates that prior uncertainty with new data.

1.9.3 Likelihood

The likelihood function is the probability of obtaining a set of N observations, given a known model and its set of model parameters. Otherwise, likelihood is a model that allows us to presume how likely the data point is, under all the possible true measurements. The likelihood is not a probability but is proportional to the probability, so we cannot sample from the likelihood. The likelihood takes into consideration the importance of the different possibilities of outcome in an experiment in the Bayesian framework. Several distributions can be used as a model descriptor for our data, such as the normal distribution, binomial distribution, chi-square distribution, and Poisson distribution. Knowledge of the distribution of the particular parameter of interest needs to be taken into consideration before adopting one of the models.

To illustrate the likelihood function, let's use the well-known chi-square χ^2 function as an example. The χ^2 is given by the relation

$$\chi^2 = p(D|\Theta) = \sum \frac{(f_i - f_{\Theta,i})^2}{\sigma_i^2} \tag{1.11}$$

This shows the probability of obtaining a set of N observations of the quantity f_i with its error σ_i given a theoretical model $f_{\Theta,i}$ at each point.

1.9.4 Prior

Prior is the probability distribution that summarises what is known about the particular parameter or a particular event before making new observations. It also can be defined as a prior belief about the parameter or the event until new data is obtained. The choice of the prior distribution is the key feature in a Bayesian approach as it plays a significant role in the inference. There are different types of prior that can be used in a Bayesian analysis such as the uniform priors, Jeffrey's priors, reference priors, informative priors, and maximum entropy priors. The mathematics of the Bayesian inference suggests that the prior information of an event or a parameter should be chosen before making a new observation.

In practice, Gelman *et al.* (2017) suggested that the prior also should be characterised according to the form of the likelihood function rather than only the philosophical interpretation of the initial information of the parameter. They concluded that the choice of the prior needs the understanding of the problem to have a proper posterior. A prior such as the Jeffreys is used to make the posterior distribution of the parameter more sensitive to the data, and it can be used depending on the concept of individual data points (Kass & Wasserman, 1996). However, another type of prior such as the uniform prior is used for high dimensional problems as it requires some conditions on the likelihood function to have a proper posterior. In our case, we use a uniform prior between the minimum and maximum distances in our distance samples to estimate the distance of our embedded objects (see Sect. 3.3.2 and Sect. 3.4.2), as we have a high dimensional space in our problem.

1.9.5 Posterior

The desired distribution of the parameter is given by the posterior, which is the product of the likelihood and the prior. The posterior distribution contains all of the possible distribution of the parameter and it is difficult to solve. A common way to solve the posterior is to draw samples to characterise the shape of the distribution.

1.9.6 Maximum Likelihood Estimation

The Maximum Likelihood Estimation (MLE) is a frequentist method that can be used to estimate the statistical model. The MLE techniques estimate the set of all possible values of the parameter of inference Θ, and it maximises the likelihood function $p(data|\Theta)$. Maximising the likelihood function is equivalent to minimising the likelihood. In the initial guess of the parameter for the MCMC sampling, the use of the MLE techniques is more convenient for the initialisation of the chain (Sect. 3.5). In our case, we use the MLE to find the most probable distance from our set of distances and use it as a starting point for the inference to make the chain converge rapidly.

1.9.7 EMCEE

Emcee is a python MCMC module produced by Foreman-Mackey *et al.* (2013), and is based on the affine-invariant ensemble sampler (Goodman & Weare, 2010). Emcee allows us to draw N samples Θ_i from the posterior density, which can also written by the equation

$$p(\Theta, \alpha|D) = \frac{1}{Z} p(D|\Theta, \alpha) p(D|\Theta, \alpha) \tag{1.12}$$

where $p(\Theta, \alpha)$ is the prior distribution and $p(D|\Theta, \alpha)$ is the likelihood function, which can be relatively easily computed for any particular value of (Θ_i, α_i). $Z = p(D)$ is the normalisation factor.

The advantage of using Emcee algorithm is that it is simple to use once we have the Bayesian model. It also handy in our problems and allows us to perform MCMC rapidly rather than writing an MCMC algorithm from scratch.

After solving the Bayesian model with Emcee, a corner plot produced by Foreman-Mackey (2016) is used to visualise the result. A corner plot is used as it is a way to represent the probability density of the samples in a multi-dimensional space. An example of the corner plot used in this work is shown in Fig. 1.15. In these visualisations, each one- and two-dimensional projection of the sample is plotted to reveal covariances. The 16^{th}, 50^{th}, and 84^{th} percentiles of the distributions are depicted by the dashed vertical lines in the histogram of the three parameters.

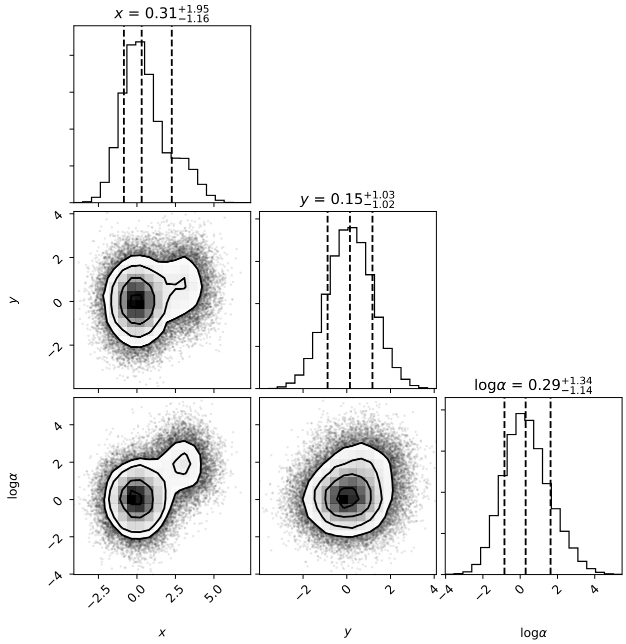

Figure 1.15: Corner plot example from Foreman-Mackey (2016). This shows the marginalized distribution for each parameter independently in the histograms along the diagonal and then the marginalized two dimensional distributions in the other panels.

1.9.8 Aim of this thesis - Outline

The major aim of this thesis is to estimate distances to embedded stars which are only observed at a longer wavelength. Those sources share common launching mechanisms in stars of different types, whether young forming stars or evolved compact dying stars. They are totally obscured by dust and do not have visible data. Consequently, a real study of one of them is challenging and their distance is often deduced from the distance of the molecular cloud associated with them (see Sect. 3.1 for more details). Kinematic distance is one of the methods used to estimate the distance of a molecular cloud, derived by measuring the local standard of rest (LSR) velocity, V_{LSR}, of an object and assuming a model of Galactic rotation. However, as indicated by Wenger *et al.* (2018), this method suffers from large uncertainties and the kinematic distance ambiguity (KDA). Another approach to estimate molecular cloud distance is the use of maser parallax measurements provided by the VLBI telescopes. Maser is often associated with protostellar objects in star-forming regions, and radio telescopes can observe maser parallax with an accuracy of better than 10 μas. The problem with this approach is that it cannot be used towards a cloud that lacks masers, and it also only generally provides one distance for a whole cloud.

With the advent of the second release of Gaia data, the issue concerning the distance determination indicated above is being solved. As shown in the previous section, an advanced statistical method such as the Bayesian analysis sounds promising and give a reliable distances measurement of the region associated with those embedded objects. In this thesis, we use the MCMC based Bayesian analysis with the unprecedented astrometric data provided by Gaia DR2 to reproduce distance measurements to YSOs from the literature. This thesis will infer distances to YSOs and ES selected from the Leeds RMS Source Survey (RMS, Lumsden *et al.* (2013)), and we also derive distances towards the sub-structures of Cygnus X (DR20, DR21, DR22, DR23, and W75N).

This thesis is organised as follows. Chapter 2 will provide a summary of the Gaia DR2 content, the StarHorse Gaia DR2 catalogues (Andrae *et al.*, 2018), and the sample selection. The Bayesian models (A_G and A_V) for the distance inference is described in Chapter 3. Chapter 4 presents all the findings obtained from this work, following a comparison of the results to those obtained from the literature. In Chapter 5, the conclusions of this work are summarized.

Chapter 2

Data analysis

2.1 Gaia data release 2

Gaia DR2 is the second release of the Gaia mission (Gaia Collaboration *et al.*, 2016b). It started in April 2018 and contains celestial positions for more than 1.7 billion sources based on observations collected during the first 22 months of the mission since July 2014 (Gaia Collaboration *et al.*, 2018). Gaia DR2 provides more than 1.3 billion parallaxes and proper motions for objects that have a magnitude limit of G = 21 mag and a bright limit of G = 3 mag. The typical astrometric uncertainty is around 0.7 mas for the faintest stars and 0.04 mas for the bright limit. Gaia DR2 also provides line-of-sight extinction in the G band, A_G, for 88 million sources calculated by Andrae *et al.* (2018), which we combine with parallaxes to infer distances to the region of interests. This latter work has been discussed in Sect. 1.8.2. As indicated by Yan *et al.* (2019b), the extinction A_G is capable of detecting molecular clouds despite its large uncertainty. The overall content of Gaia DR2 is summarised in Table 2.1.

2.2 StarHorse Gaia DR2 catalogues

In this work, we also used a visual extinction in the Johnson V band, A_V, and the set of distances derived by Anders *et al.* (2019) to see the improvement on the

Table 2.1: Overall content of Gaia DR2 from Mignard (2019).

Data product or source	Number of sources
Total (excluding solar system)	1 692 919 135
Five-parameter astrometry (position, parallax, proper motion)	1 331 909 727
Two-parameter astrometry (position only)	361 009 408
ICRF3 prototype sources (link to radio reference frame)	2 820
Gaia-CRF2 extra-galactic sources (optical reference frame)	556 869
G-band (330–1050 nm)	1 692 919 135
G_{BP}-band (330–680 nm)	1 381 964 755
G_{RP}-band (630–1050 nm)	1 383 551 713
Median radial velocity over 22 months	7 224 631
Classified as variable	550 737
Variable type estimated	363 969
Detailed characterisation of light curve	390 529
Effective temperature Tef	161 497 595
Extinction A_G	87 733 672
Colour excess $E(G_{BP} - G_{RP})$	87 733 672
Radius	76 956 778
Luminosity	76 956 778
Solar system object epoch astrometry and photonetry	14 099

derived distance obtained from A_G. The following surveys were used with Gaia in the StarHorse code: the Two Micron All Sky Survey (2MASS; Skrutskie *et al.* (2006)), the Pan-STARRS1 *grizy* (Scolnic *et al.* (2015)) and AllWISE (Cutri *et al.* (2014)).

Anders *et al.* (2019) derived stellar parameters, distances, and extinction for 265 million of the 285 million objects brighter than G=18 mag. After cleaning the results, the final catalogues contains 137 millions sample stars that gave a median precision of 5% in distance, 0.20 mag in V-band extinction, and 245 K in effective temperature for $G \leq 14$, degrading to fainter magnitudes 12%, 16%, and 14% in distance, 0.20 mag, 0.23 mag, and 0.24 mag in V-band extinction, and 245 K, 260 K, and 230 K in effective temperature for $14 < G \leq 16$, $16 < G \leq 17$, and $17 < G \leq 18$, respectively. Table 2.2 shows the comparison of the work done by Anders *et al.* (2019) to all of the currently available astro-photometric distances and extinctions derived based on the Gaia astrometric solution.

2.3 Sample selection

2.3.1 Gaia DR2

For the selection of our Gaia star, we applied all the criteria needed to reduce contamination from stars with poor astrometric data. We started by drawing a box region centred on our object of interest, and we only considered Gaia stars in this box. Gaia stars were selected according to their parallaxes and A_G extinction. We considered stars that have positive parallaxes ($\omega > 0$). As suggested by Schönrich *et al.* (2019), the ratio $\omega/\delta\omega$ needs to be greater than 10 for a safe selection of Gaia stars. However, keeping stars that have $\omega/\delta\omega > 10$ diminished the number of stars because it removed stars that have a higher extinction in the G band ($A_G > 2.5$). Thus, we applied the RUWE (Lindegren *et al.*, 2018b) condition associated to each Gaia source, which is mentioned in the Gaia archive. The RUWE is not yet in the Gaia archive, but it basically describes the useful

Table 2.2: Comparision of some of the currently available astro-spectro-photometric distances and extinctions based on Gaia data to those provided by Anders *et al.* (2019), which we used in this work.

Reference	Survey(s)	mag_limits (mag)	Objects	σ_d/d (%)	σ_{AV} (mag)	σ_{Teff} (K)
Queiroz *et al.* (2018)	Gaia DR1 + spectroscopy	–	1.5 M	15	0.07	–
Mints & Hekker (2017)	Gaia DR1 + spectroscopy	–	3.8 M	15	–	–
Sanders & Das (2018)	Gaia DR2 + spectroscopy	–	3.1 M	3	0.01	40
Santiago et al. (in prep)	Gaia DR2 + spectroscopy	–	2 M	5	0.07	40
Bailer-Jones *et al.* (2018)	Gaia DR2	$G \lesssim 21$	1330 M	25	–	–
McMillan (2018)	Gaia DR2	$G \lesssim 13$	7 M	6	–	–
Andrae *et al.* (2018)	Gaia DR2	$G \leq 17$	80 M	–	0.46	324
Anders *et al.* (2019)	Gaia DR2 + photometry	$G < 18$	137 M	13	0.22	250
Final samples of Anders *et al.* (2019) (after cleaning the results)						
$G \leq 14$			14,432,712	5	0.20	245
$14 < G \leq 16$			49,171,794	12	0.20	245
$16 < G \leq 17$			43,398,790	16	0.23	260
$17 < G \leq 18$			29,602,832	14	0.24	230

NOTE: *The first three columns represent the reference, the surveys, and the magnitude limits used, respectively. The last three columns refer to the median precision in relative distance, V-band extinction, and effective temperature, respectively.*

statistical implicit for Gaia stars.

The RUWE can be computed by using Eq. 2.1

$$\mathrm{RUWE} = \frac{\sqrt{\chi^2(AL)/nobs(AL) - m}}{f(G, G_{BP-G_{RP}})} \tag{2.1}$$

where m is the number of parameter solved and f is a renormalising function. AL refers to Gaia along-scan direction.

As mentioned by Lindegren *et al.* (2018b), the RUWE is expected to be around 1 for an observed single star. A RUWE greater than 1 indicates a non-single star or a problematic solution for the astrometry. For these reasons, we applied RUWE ≤ 1.4 for the selection criteria.

For the set of distances of the surrounding objects, we could not use the reciprocal of the parallax, $1/\omega$, as it is not reliable for distance measurement (Bailer-Jones *et al.*, 2018). The inversion of parallax gives us the exact distance if we use the true parallax of the star, $r = \frac{1}{\omega_{true}}$, which is not possible yet as measurement errors are always present in astronomy. In addition, the inversion of the parallax is a biased measurement as the distribution of the resulting distance from this method is highly asymmetrical (Luri *et al.*, 2018). Another issue is that Gaia provides negative parallax measurements, so the inversion of the negative parallax yield a negative distance, which does not have a physical meaning (Astraatmadja & Bailer-Jones, 2016). However, those negative parallaxes cannot be removed from the samples as they contain important information. Therefore, an advanced method must be taken to consider all of the observed information with measurement errors to have a better estimation of the distance.

To have a reliable estimate on the distance of a star, Bailer-Jones *et al.* (2018) proposed a Bayesian analysis that finds the best estimates of the distance of a single star given its measured parallax ω and its error σ_ω. In their study, they used a simpler exponentially decreasing space density prior that follows a galactic model for the distance estimates (see Eq. 2.2, Bailer-Jones (2015)).

$$P(r) = \frac{1}{2L^3} r^2 e^{-r/L},\ for\ r > 0 \tag{2.2}$$

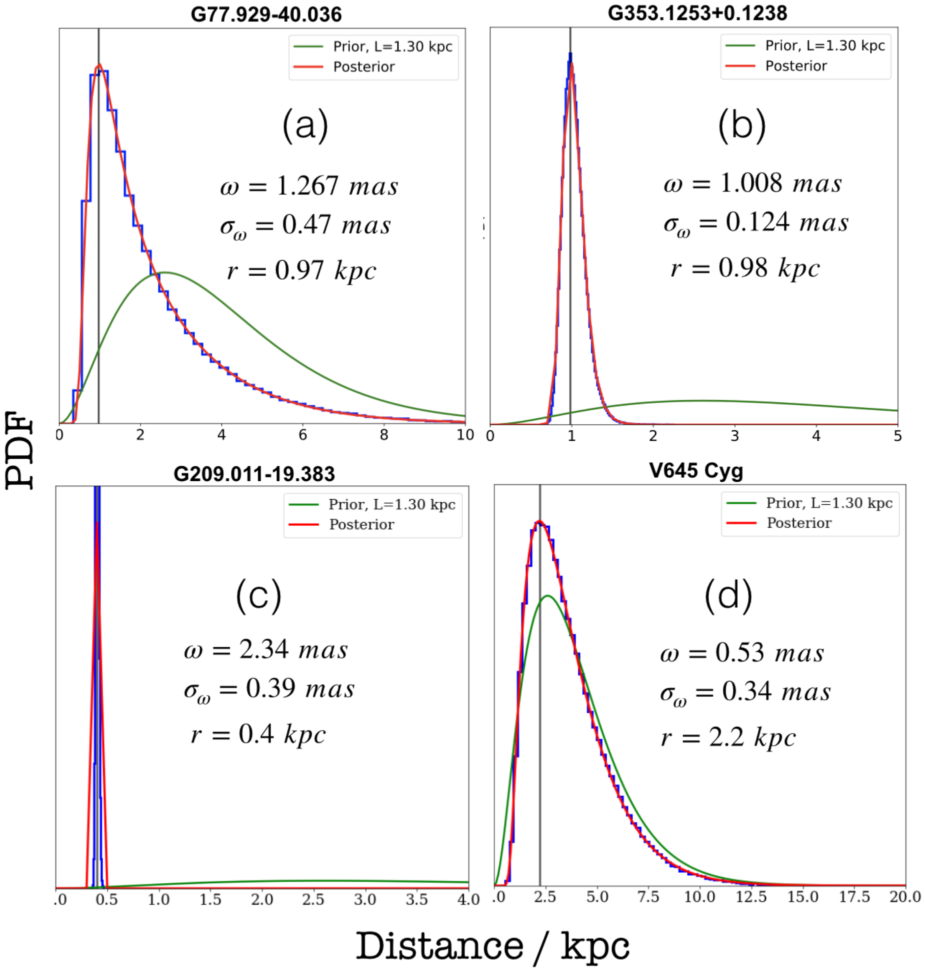

Figure 2.1: Example of Gaia distances by using Bayesian inference (Bailer-Jones *et al.* (2018)). The Bayesian analysis demonstrates the impact of the parallax error on the distances. The blue histogram and the red curve shows the posterior distribution of the distance and its best fit respectively and the green curve represents the prior distribution of the distance. Figure (a) and Figure (b) shows that the two stars are estimated to have a similar distance even if their observed parallax are different. The prior on the distance is underestimated when the error on the parallax is too small (Figure (b) and Figure (c)), while the distance estimates follow the prior when the star has a significant error on the parallax (Figure (a) and Figure (d)).

where L is the length scale, and r is the true distance to the star.

Fig. 2.1 shows an example of Bayesian distance estimates proposed by Bailer-Jones *et al.* (2018) that illustrates the importance of using Bayesian techniques to handle the parallax error from Gaia. When a star has a small error on the parallax, the reciprocal of the parallax is still used to derive its distances. For that, the exponential prior was underestimated and was differentiated from the posterior distribution (see the green curve in (b) and (c) in Fig. 2.1). However, when the error on the parallax is significant, the posterior distribution of the distance follows the prior, which demonstrated that the error on the parallax plays a vital role in distance estimates.

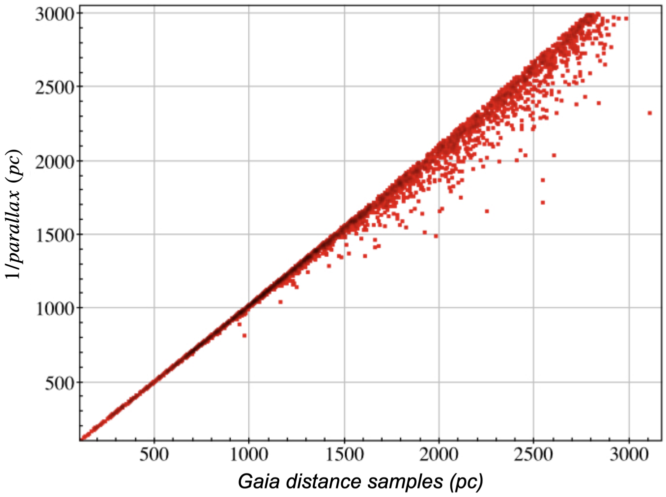

Figure 2.2: Comparison of the Bayesian estimation of individual star provided in the Gaia catalogues (Bailer-Jones *et al.* (2018)) to the distance obtained by inverting the parallax. The distance obtained from the inversion of parallax begin to differ from 500 *pc* where we start to have a significant error on the parallax $\delta\omega/\omega \geq 0.5$

For all reasons cited above, we decided to use the distances provided by Bailer-Jones *et al.* (2018), which is also included in the Gaia DR2 catalogue. Fig. 2.2 depicts the comparison between Gaia DR2 distance estimates and the naive inversion of the parallax towards G010.3844+02.2128. Gaia distances begin to differ from the distances obtained through simple inversion of parallax for sources with large errors, $\delta\omega/\omega \geq 0.5$.

For the extinction, we use Gaia stars that have A_G extinction greater than zero ($a\_g\_val\ > 0$). The error in the extinction ΔA_G is estimated by

$$\Delta A_G = \frac{1}{2}(A_G^{upper} - A_G^{lower}) \tag{2.3}$$

where A_G^{upper} and A_G^{lower} are the 84^{th} and 16^{th} percentiles of A_G provided by the Gaia catalog.

2.3.2 Starhorse stars

For the extinction and set of distance for the A_V model, we applied the same criteria as the selection for Gaia stars. Although they only derived distances and extinction for stars brighter than $G = 18$, their results can be used for the purposes of this work. The process for the star selection starts by drawing a box region centred in our object of interest, and we consider all stars inside that box. All the stars in their catalogues have extinction and their corresponding distance measurements.

For the distance samples and A_V extinction, we used the 50^{th} percentiles from the StarHorse results. The error in the distance was estimated using its standard deviation, while the error in the extinction was estimated using its 16^{th} percentiles and 84^{th} percentiles, given by the relation

$$\Delta A_V = \frac{1}{2}(A_V^{84th} - A_V^{16th}) \tag{2.4}$$

We also applied RUWE $\leq$ 1.4 associated with each star.

Chapter 3

Implementation of the Bayesian models

3.1 Overview

As discussed in Sect. 1.9.8, embedded stars do not have reliable visible data. The most probable approach to calculate their distances is to study the region associated with them by considering all of the available stars in the sight-line. A molecular cloud is the birthplace of stars, and all stars in the same molecular cloud are located at a similar distance. We assume here that our stars are normally distributed around the molecular cloud associated with those objects. Therefore, our study is based on the molecular cloud distance, which is related to our objects of interest.

The use of parallaxes of the available stars in a molecular cloud accompanied by an extinction map of the region was introduced in the study done by Lombardi *et al.* (2008). They derived a reliable distance of the Ophiucus and Lupus cloud complexes using parallax measurements provided by Hipparcos (Perryman *et al.*, 1997) and extinction maps of these regions(NICER, Lombardi & Alves (2001)). Their analysis is based on a rigorous maximum-likelihood approach.

As in Sect. 1.9.1, several studies have been made for the derivation of molecular cloud distances since the advent of Gaia. Yan *et al.* (2019b) demonstrated that the A_G extinction derived directly from Gaia three-band photometry and Gaia parallaxes are capable of deriving reliable distance measurements to molecular clouds. They built a Bayesian model of A_G extinction and derived reliable distances to 59 molecular clouds at high Galactic latitudes ($|b| > 10^o$). Later on, Yan *et al.* (2019b) performed another measurements of molecular cloud distances towards the Galactic plane ($209.75° \leq 1 \leq 219.75°$ and $|b| \leq 5°$) using the baseline subtraction method and Bayesian statistics.

Based on the work cited in Sect. 1.9.1 and those discussed above, we implemented the Bayesian model of the line-of-sight extinction towards our region of interests, and infer their distances. For that, the A_G extinction and the A_V extinction were used. The two models are discussed in Sect. 3.3.1 and Sect. 3.4.1, respectively.

3.2 Basic Analysis

Our technique for deriving distances to the molecular cloud is based on a general fact where stars observed through excessive column densities must exhibit a high reddening even when observed in projection towards dense regions of the cloud (Lombardi *et al.* (2008)). In other words, the molecular cloud increases the extinction of all-stars behind them. To illustrate the approach, we used the region of Barnard 68 (Fig. 3.1). Stars that are associated with the molecular cloud are opaque at visible-light wavelength as the light coming from those stars are totally absorbed by dust. The near-infrared image reveals all of the objects embedded in the dust cloud and shows that those stars appear redder than stars located around the edge of the molecular cloud. The change in colour is the jump point on the extinction that we inferred here. The terms *off-cloud* and *on-cloud* were used to design stars that have lower and higher extinction values, respectively. The bottom panel of Fig. 3.1 depicts the line-of-sight extinction profile towards the region of our objects. The diagram shows a simple dust-screen model that

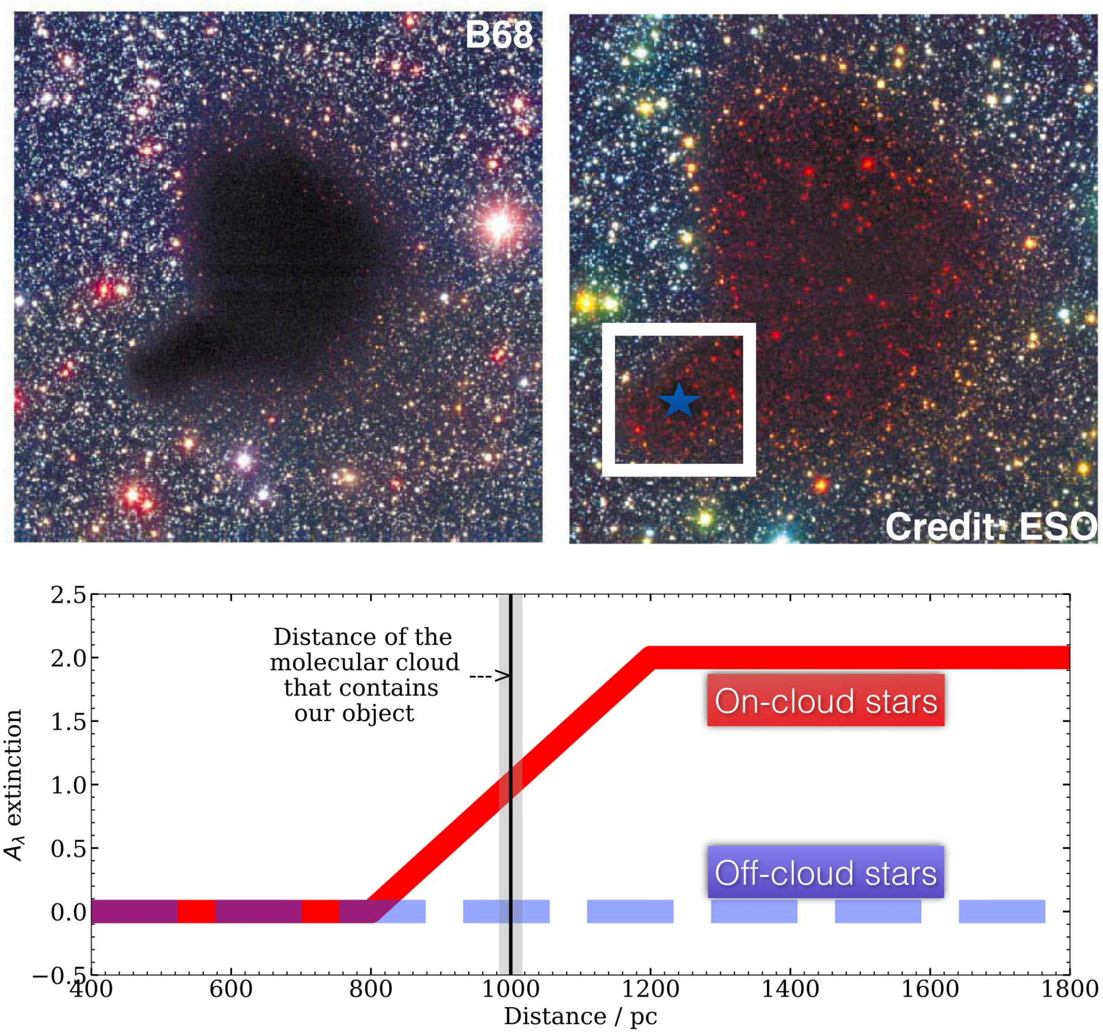

Figure 3.1: The two figures on the top panel show an optical image and near-infrared image of Barnard 68 (hereafter B68) respectively, which is a good region to illustrate our approach. The blue star in the second image shows an example of a star that we are going to measure its distance and the box is the region that we consider to infer the jump point in the extinction in that region. The bottom panel shows a simple extinction profile that shows our dust-screen model in a given line of sight towards our object. The blue line depicts the extinction of off-cloud stars while the red line shows the variation of extinction for on-cloud stars. The distance of the molecular cloud that contains our object is marked by the black vertical line and its error (grey line). (Image credit: "ESO")

illustrates the jump in the extinction for $On-cloud$ stars when passing across the molecular cloud. By taking into account the extinction of *Off-cloud* stars and the extinction of *On-cloud* stars, the Bayesian models provided the most probable distances to the region.

In summary, the Bayesian models involve five parameters: the distance D of the region, the extinction for foreground stars μ_1, the error on the extinction for foreground stars σ_1, the overall extinction μ_2 for background stars, and the error on the extinction for background stars σ_2.

3.2.1 Classification of star

As in Sect 3.2, when going through the molecular cloud along the line of sight, the extinction moderately increases from *off-cloud* stars to *on-cloud* stars. With respect to Yan *et al.* (2019b), the probability of a star to be located in front of the cloud or behind the cloud is indicated by the CDF of the normal distribution below

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{x} e^{-t^2/2} \mathrm{d}t \tag{3.1}$$

And given the distance D of the molecular cloud associated with our object and the range of distances d_i that surround our object, the likelihood of a star to be located in front of the molecular cloud is

$$f_i = \phi\left(\frac{\mathrm{D} - \mathrm{d_i}}{\Delta \mathrm{d}_i}\right) \tag{3.2}$$

and the likelihood of a star to be located behind the cloud is

$$1 - f_i \tag{3.3}$$

In practice, the extinction of foreground stars is estimated by taking the minimum distance D_{min} and $(D_{min} + r\ pc)$ in our distance catalogues and calculate the average extinction in the distance range $[D_{min}, D_{min} + r\ pc]$. The value of r is chosen according to the distance cut that we use for the inference given the prior knowledge that we obtained from the literature.

3.3 Distance estimates using A_G

3.3.1 Basic model

The approach used in this work is based on Yan *et al.* (2019a) with some modifications. As discussed in Section 1.8.2, ExtraTrees cannot extrapolate beyond the training data range, and does not derive negative value of A_G. This non-negativity means that the likelihood for A_G cannot be Gaussian, but the most appropriate way of modelling this is by using the truncated Gaussian distribution (Andrae *et al.*, 2018). We also followed this for our A_G model. The truncated Gaussian distribution, with the mean μ and standard deviation σ is written as

$$p\left(A_{\mathrm{G}i}|\mu,\sigma\right)=\begin{cases}\dfrac{\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{A_{\mathrm{G}i}-\mu}{\sigma}\right)^{2}\right)}{\frac{1}{2}\left(\mathrm{erf}\left(\frac{A_{\mathrm{G}}^{\max}-\mu}{\sqrt{2}\sigma}\right)+\mathrm{erf}\left(\frac{\mu-A_{\mathrm{G}}^{\min}}{\sqrt{2}\sigma}\right)\right)}\\ 0\end{cases} \tag{3.4}$$

Where (A_G^{min}) is the minimum value of A_G extinction and (A_G^{max}) is the maximum value of A_G extinction, which is equal to 0 and 3.6 mag respectively. A_{Gi} is the range of extinction of stars that we consider in our study. The erf in the denominator indicates the error function of (A_G^{max}) and (A_G^{min}) and is given by the relation:

$$\mathrm{erf}(z)=\frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\mathrm{d}t \tag{3.5}$$

In order to consider the measured error in the A_G extinction, we convolved the standard deviation of A_G extinction with σ_1 and σ_2, which are the error for the extinction for foreground stars and the error for background stars, respectively. Thus, the likelihood of measuring A_{Gi} for foreground stars is given by

$$PF_i=p\left(A_{\mathrm{Gi}}|\mu_1,\sqrt{\sigma_1^2+\Delta A_{\mathrm{Gi}}^2}\right) \tag{3.6}$$

Identically, the likelihood of background stars is

$$PB_i=p\left(A_{\mathrm{G}i}|\mu_2,\sqrt{\sigma_2^2+\Delta A_{\mathrm{G}i}^2}\right) \tag{3.7}$$

The total likelihood is the product of all background stars. By taking into account all of our parameters and the Eq. 3.6 and Eq. 3.7, the form of our total likelihood is

$$p\left(A_{\mathrm{Gi}}|\mu_1, \sigma_1, \mu_2, \sigma_2, \mathrm{D}\right) = f_i PF_i + \left(1 - f_i\right) PB_i \tag{3.8}$$

3.3.2 Prior

For the prior, we chose a uniform prior for the distance D. Prior knowledge in the distance is essential in our model, so we set a smart prior for a better result. The distance prior was set to be uniform between D_{min} and D_{max} from our dataset. For the foreground and background extinction, we chose an exponential distribution as the extinction value increase as a function of the distance. In summary, our prior for the five parameters are

$$\begin{cases} \mathrm{D} & \sim \mathcal{U}\left(\mathrm{D_{min}}, \mathrm{D_{max}}\right) \\ \mu_1 & \sim \mathcal{E}(\mu_f) \\ \mathrm{I}\sigma_1 & \sim \mathcal{E}(2) \\ \mu_2 & \sim \mathcal{E}\left(\mu_b\right) \\ \mathrm{I}\sigma_2 & \sim \mathcal{E}\left(\mathrm{I}\sigma_b\right) \end{cases} \tag{3.9}$$

where $\mathcal{U}$ and $\mathcal{E}$ represent the uniform and exponential distributions, respectively, μ_b and $I\sigma_b$ are the mean and reciprocal standard deviation of extinction A_G of background stars with distances $> D_{max} - r\ pc$. The initial guess for μ_f and μ_b were derived from the mean extinction of *off-cloud* stars and *on-cloud* stars, respectively. The 2 mag for $I\sigma_1$ is the reciprocal of 0.45 mag, which is the typical extinction A_G standard deviation of clustering stars (Andrae *et al.*, 2018).

3.4 Distance estimates using A_V

3.4.1 A_V model

The model used for A_V extinction was pretty much the same as the model used for A_G (see Sect. 3.3.1). The only difference here is that the truncated Gaussian distribution is only used for A_G due to the limitation of A_G ($A_G \in [0, 3.609]$ mag) (Andrae *et al.*, 2018). For A_V, we use the full Gaussian distribution for objects in all regions in this study.

The full Gaussian distribution of A_V with the mean μ and standard deviation σ is indicated by the relation

$$p\left(A_{\mathrm{V}i}|\mu,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{A_{\mathrm{V}i}-\mu}{\sigma}\right)^2\right) \tag{3.10}$$

where A_{Vi} is the range of A_V extinctions of our star.

As included in Sect. 3.3.1, to consider the measurement error in the A_V extinction, we convolved the standard deviation of A_V extinction with σ_1 and σ_2. Consequently, the likelihood of measuring A_{Vi} for foreground stars is given by

$$PF_i = p\left(A_{\mathrm{Vi}}|\mu_1, \sqrt{\sigma_1^2 + \Delta A_{\mathrm{Vi}}^2}\right) \tag{3.11}$$

and the likelihood for background stars is

$$PB_i = p\left(A_{\mathrm{V}i}|\mu_2, \sqrt{\sigma_2^2 + \Delta A_{\mathrm{V}i}^2}\right) \tag{3.12}$$

The total likelihood of the A_V extinction is given by the relation

$$p\left(A_{\mathrm{Vi}}|\mu_1,\sigma_1,\mu_2,\sigma_2,\mathrm{D}\right) = f_i PF_i + \left(1-f_i\right) PB_i \tag{3.13}$$

3.4.2 Prior

For the choice of prior, we used the same prior as for A_G except the inference of the error for foreground stars $I\sigma_1$. The prior used for A_V are

$$\begin{cases} \mathrm{D} & \sim \mathcal{U}\left(\mathrm{D_{min}}, \mathrm{D_{max}}\right) \\ \mu_1 & \sim \mathcal{E}(\mu_f) \\ \mathrm{I}\sigma_1 & \sim \mathcal{E}\left(\mathrm{I}\sigma_f\right) \\ \mu_2 & \sim \mathcal{E}\left(\mu_b\right) \\ \mathrm{I}\sigma_2 & \sim \mathcal{E}\left(\mathrm{I}\sigma_b\right) \end{cases} \tag{3.14}$$

where $\mathcal{U}$ and $\mathcal{E}$ the uniform and exponential distributions, respectively. μ_b and $I\sigma_b$ are the mean and reciprocal standard deviation of A_V extinction of background stars, while μ_f and $I\sigma_f$ are the mean and reciprocal standard deviation of the A_V extinction of the foreground stars.

3.5 MCMC sampling

Finally, we solved the posterior distribution with Monte Carlo Markov Chain sampling (MCMC) techniques. The MCMC is a process for generating a random walk in the parameter space and drawing a representative sample from the posterior distribution. We applied the same process for A_G and A_V. As introduced in Sect. 1.9.7, we used *emcee* to sample the joint posterior probability. We initialized 15 walkers for the sampling, which means we calculated 15 independent chains. After setting up the walkers, we ran 1000 steps from the posterior distribution of parameters, this step is called the "burn-in" phase. Then, we ran 2000 steps again for the production phase. As this work is a high dimensional inference, *emcee* comes with some outliers and takes a large number of steps to converge until it finds a good starting point for the sampling. For this, we speeded up the convergence with the "Maximum Likelihood Estimation (MLE)" technique. MLE provided the most probable value for the starting guess of the sample based on the data that we used. We also increased the number of steps to observe a clear convergence of the chain.

Chapter 4

Results

This section deals with the derived distances obtained from this work and the comparison of the results to the literature. In Sect. 4.1, we test the two models with the Orion A region to examine the reliability of the methodology. After testing the model, in Sect. 4.2.1, we examine distances to YSOs taken from the RMS survey (Lumsden *et al.*, 2013). We also derive distance towards the active star forming region Cygnus X in Sect. 4.2.2. Finally, Sect. 4.3 compares the results to several distance measurements in the literature.

4.1 Testing the models

4.1.1 The Orion A structure

The Orion A cloud is the closest massive star-forming population in the Galaxy. By using *Gaia* DR2, Großschedl *et al.* (2018) studied the 3D shape of Orion and identified that it consists of two separate parts, which are the northern part of the cloud or *head* (box (d) in Figure 4.1) and the southern part of the cloud or *tail* (box (b) and (c) in Figure 4.1). This recent study also showed that the *head* of Orion A is bent in regards to the *tail*, following a finding of the *head* to be located on the plane of the sky and the *tail* not far from the line of sight. Those parts of the cloud contain rich clusters of young stars and active sites of massive star formation (Rezaei Kh. *et al.*, 2020), not a flat filament in the plane of sky

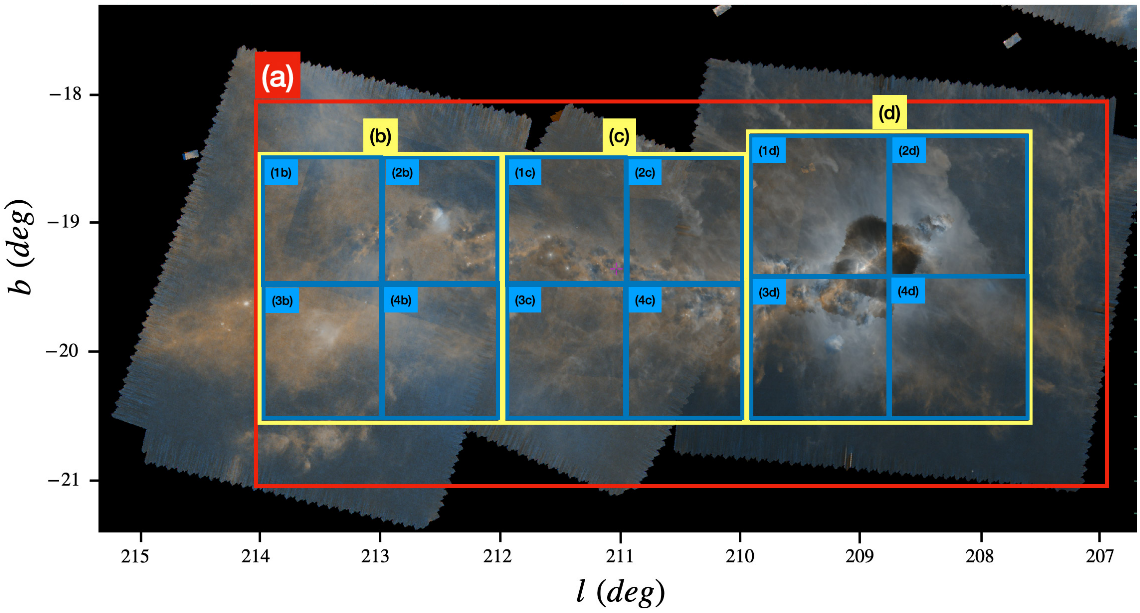

Figure 4.1: Illustration of the Orion A sub-regions used for the test of the effectiveness our methodology. All the boxes in the figure represent different parts of the Orion A, in which we used to analyse all the parameters involved in our models. (Credit: ESASky (Martí *et al.*, 2016), Herschel PACS RGB 70, 160 micron)

(Schilke, 2015).

The Orion A is suitable to test the models by a number of factors. It has a well known distance derived from the literature $\sim$ 400 pc (Großschedl *et al.* (2018), Menten *et al.* (2007), and Schlafly *et al.* (2015)). The understanding of the distance gradient across the Orion A indicated by those authors also is of significant importance for the test to analyse several distances when shifting from the *head* to the *tail*. In addition, this region is well observed by *Gaia* due to its location nearby the Sun. Thus, it contains many datapoints with high quality parallaxes and extinctions in the *Gaia* DR2, making it the best region to test the models.

4.1.2 Orion A distance analysis

Variation of number of stars

We divided the Orion A into four boxes which are named [(a), (b), (c), (d)], and we derived distances to each of them. After that, we subdivided (b), (c), and (d) into four sub-regions each and inferred their distances. Fig. 4.1 illustrates all the sub-regions used. By using those different sub-regions, the number of input stars decreased compared to the big three boxes. This helps to understand the effect of using fewer stars into our models, and also help to identify the distance gradient across the Orion A.

Extinction cut-off

As outlined in Sect. 3, the technique used in this work relies on the extinction of stars and their corresponding distances. We used a simple dust screen model of the extinction of foreground stars and background stars to infer the most reliable distances to the star-forming region. Here we applied a lower cutoff in the extinction (1.5 mag for A_G and 2 mag for A_V using box (c)) to see if we can still obtain distances. This is important to test the sensitivity of the model when inferring distances to a region with a small jump in the extinction.

Distance cut-off

We also extended the distance cutoff to 2 kpc for the box (c) to test the effect of including many background stars into the models. Some of the objects we studied in this work have an estimated distance in the literature. Hence, by varying the distance cutoff, we are able to understand the reliable cutoff than can be applied for other objects according to the literature value. More importantly, reducing the distance cutoff removes stars that are located very far away, which are not part of the region under study.

At the end part of the test, we added an additional error in the distance samples to our datasets by $\sim$ 50%. This is just a quick check of the importance of the error in our data samples and its effect in the results.

4.1.3 Results and discussions

In general, the two models of extinction detected distances to all the regions drawn with respect to the extinction breakpoint. The box (a) of Fig. 4.1 was first used for the distance measurement. It represents the whole region of the Orion A including the *head* and the *tail*. It also contains data from all part of the Orion A, so it provides an average distance towards this region. For that, we obtained a distance of 399^{+8}_{-8} pc with A_G, and 427^{+5}_{-4} pc with A_V, respectively. The one obtained with A_V is $\sim$28 pc higher than the A_G distance. This slight difference is caused by the large number of on-cloud stars obtained with A_V data.

We also derived several distances to the other three boxes, shifting from the *head* to the *tail* (see Fig. 4.2 and Fig. 4.3). For the head, which is represented by (d), we obtained 340^{+13}_{-10} pc with A_G, and 349^{+8}_{-7} pc with A_V. For the *tail*, we used the box (b) and (c) to derive its distance. For (b), we obtained 455^{+21}_{-20} pc with A_G, and 413^{+11}_{-10} pc with A_V. For (c), we obtained 410^{+19}_{-12} pc with A_G, and 407^{+11}_{-10} pc with A_V. The results indicate that the tail is located further away than the head, which are $\sim$ 115 pc more distant for A_G, and $\sim$ 64 pc away for A_V.

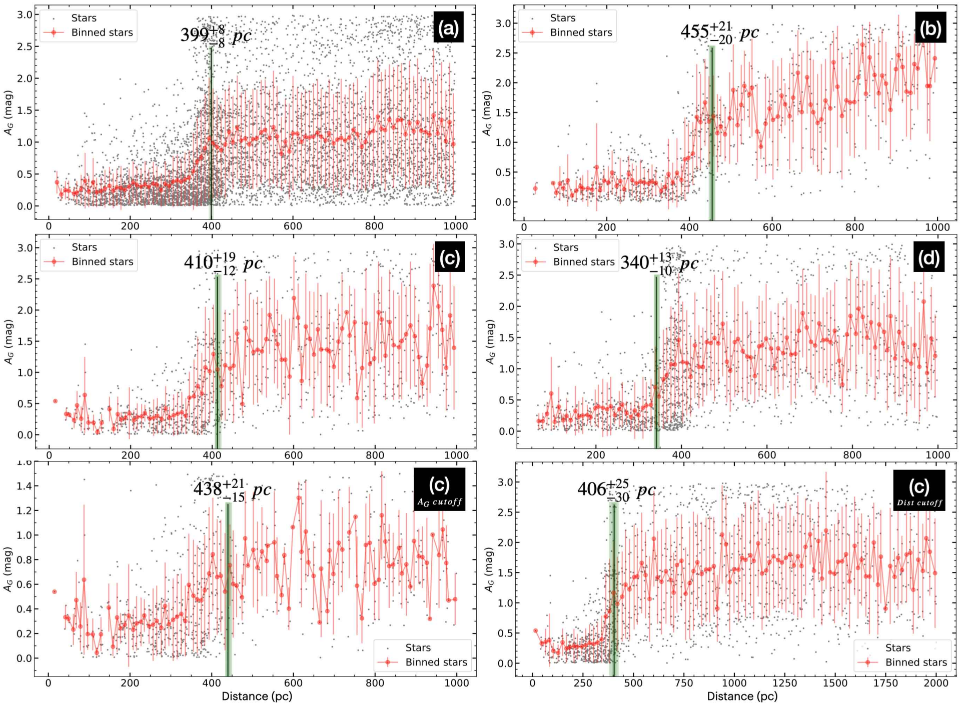

Figure 4.2: Distances towards the Orion A with A_G obtained from the test. The Orion A is located at a distances ∼ 400 pc. The grey points represents the selected stars for each box, while the red points show their corresponding binned data averaged in every 10 pc. The derived distances is indicated by the green vertical line. The vowel in brackets for each figure indicates the part of the Orion A used for the test.

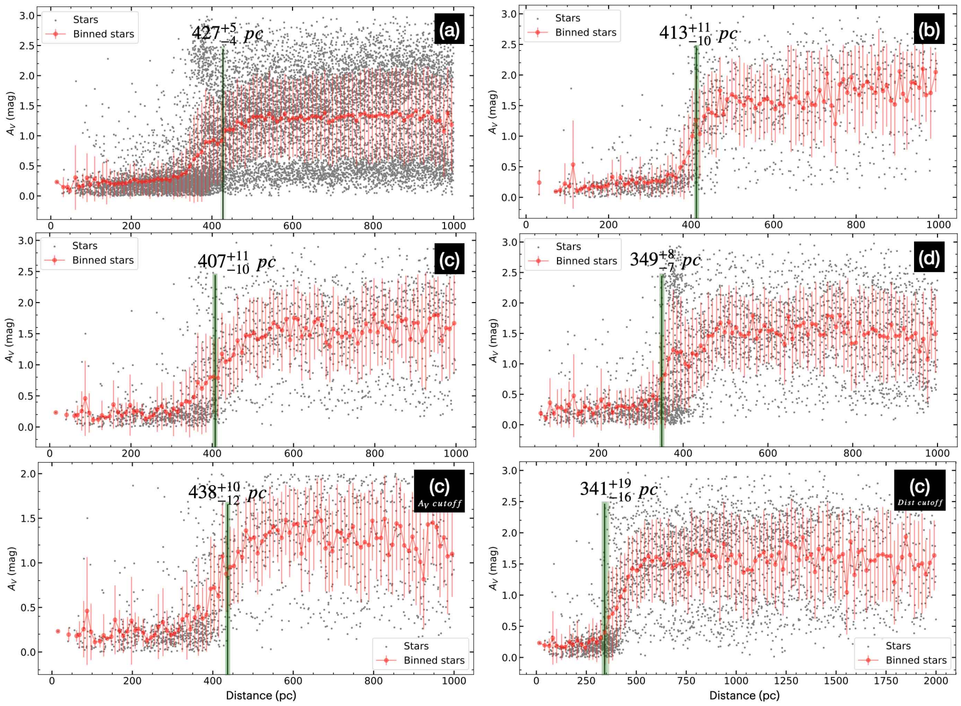

Figure 4.3: Test of the A_V model to derive distances to the Orion A. See the caption of Fig. 4.2 for more details.

This demonstrates the additional ~ 70 pc on the distances of the tail stated by Schlafly *et al.* (2015).

Variation of the number of stars

To further understand the reliability of our methodology, we derived distances to the sub-regions of (b), (c), and (d). For those sub-regions, we obtained data within 1 degree square across the region. Compared to the number of stars that has A_V extinction, those sub-regions contain fewer star with A_G extinction, making them very difficult to deal with the A_G model. In addition, the distribution of A_G across those sub-regions is very complicated, and we failed to derive a reliable distances to some of the sub-regions drawn. For instance, the distances of the sub-regions (1c) and (2c) obtained using the A_G model were very large (see Fig. 4.4). We can clearly see the location of the jump in the A_G extinction, but the algorithm inferred larger distance, which are 521^{+29}_{-45} pc and 488^{+28}_{-39} pc respectively. In the opposite way, for (4c), we obtained a lower distance 282^{+46}_{-38} pc. In those sub-regions, the jump point in the extinction was not confirmed because of implementing fewer stars in our model.

As in Sect. 4.1.2, we used a simple dust screen model, so our methodology depends on the behaviour of foreground and background extinction. Thus, having more datapoints does not always provide distances. For instance, the sub-region (2d) contains many datapoints for both A_G and A_V, but we are still not able to infer a reliable distance due to the complicated distribution of the extinction (see Fig. 4.5). For that, we obtained 268^{+17}_{-20} pc, which is lower compared to the extinction breakpoint.

The results obtained from those small sub-regions also demonstrate the distance gradient across the Orion A. For both models, we obtained a lower distances to the sub-regions in head compared to those in the tail. All the figures obtained to those sub-regions are listed in Appendix A.

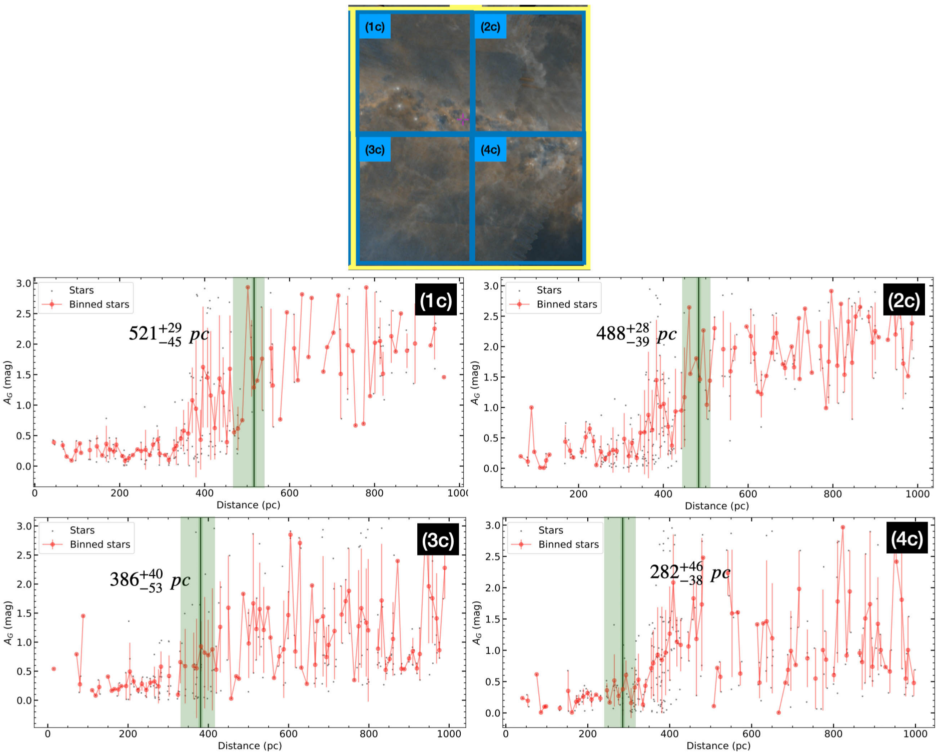

Figure 4.4: Distances towards the sub-regions of box(c) using A_G.

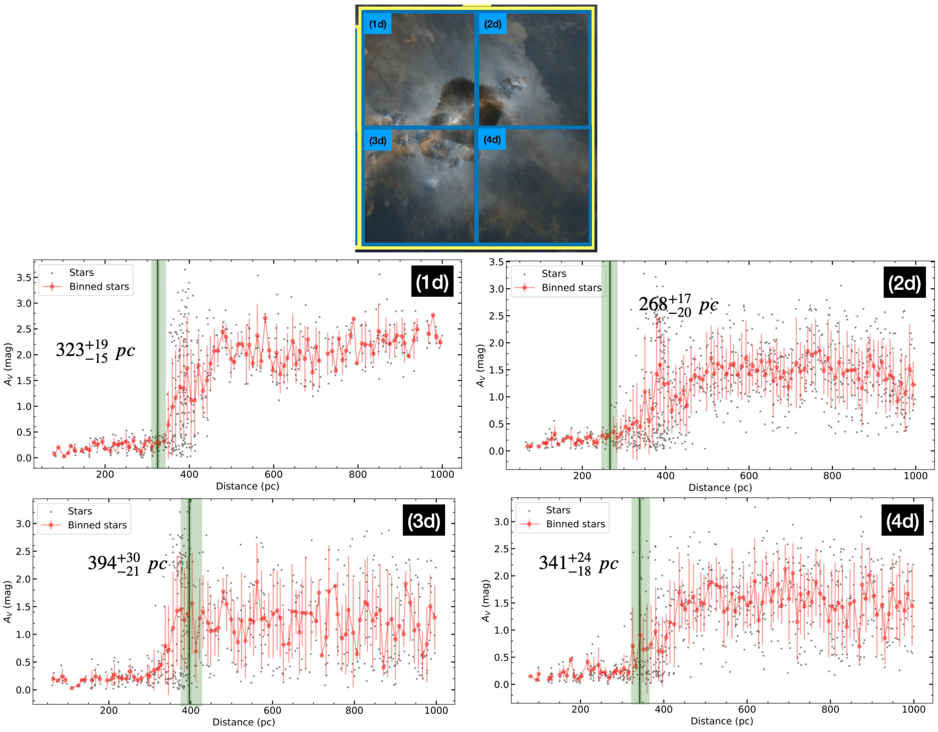

Figure 4.5: Distances towards the sub-regions of box(d) using A_V.

Extinction cutoff effects

The two models always detect distances even if a smaller jump in the extinction were used. By applying an extinction cutoff of 1.5 mag for A_G and 2 mag for A_V to the box (c), we inferred a distance of 438^{+21}_{-15} pc for A_G and 438^{+10}_{-12} pc for A_V. Compared to the distances obtained with no extinction cutoff, which are 410^{+19}_{-12} pc with A_G and 407^{+11}_{-10} pc with A_V, we observed an increase of about $\sim$30 pc in the distance for both models. The extinction cutoff removed background stars that suffer much extinction at the boundary of the cloud, and thus the observed distance naturally shifted with respect to the distribution stars along the sightline.

Distance cutoff effects

The larger distance cutoff 2 kpc applied has no considerable effect in the A_G distance, but it caused a slight decrease for about ∼60 pc for A_V. Compared to the measurements without distance cutoff, there is also an increase in the errorbar for the two models. I fact, increasing the number of stars into our model should lower the errorbar, however, we obtained higher errorbar since stars located at higher distance have large uncertainties.

The additional 50 % error in the distance sample has no significant effect in the results as it only caused a slight increase of about ∼10 pc in the distance and small increase in the errorbar (see Fig. 4.6).

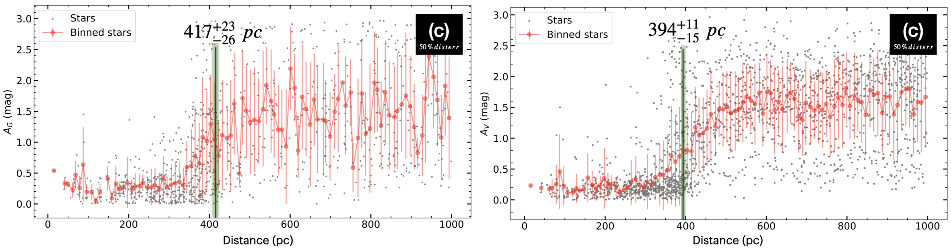

Figure 4.6: Test of the two models A_G (in the left panel) and A_V (in the right panel) by adding a 50 % error in the distance samples. See Caption 4.2 for more details.

4.1.4 Effective sample selection

According to the test, all changes applied to the datasets had effects in the distance measurement. As we used a simple dust screen model, good agreement of the extinction with this model is crucial and enough background stars that has extinction data is also required. In addition, applying a large distance cutoff did not yield to a considerable change in the distance, so it is not always necessary. One issue with a large distance cutoff was that it added more stars that are located very far away, which might not be part of the molecular cloud. Those many input stars also take too much time for the computation. According to the test, the distance cutoff for an object that has an estimated distance less than 500 pc in the literature should not exceed 1 kpc. This distance interval will be respected for the other distance measurements in this thesis.

4.1.5 Distance error and autocorrelation

Our distance measurements contain two categories of uncertainty: the statistical uncertainty and the systematic uncertainty. The systematic uncertainty is not included in the test but will be added in the next results. As mentioned by Luri *et al.* (2018), *Gaia* measurement may be affected by systematic errors due to the design of both the spacecraft and the implementation of the data processing software. This systematic error on the parallax is estimated to be around 0.04 mas which causes a systematic error about 5% in the distance. We adopted this 5% systematic error in our distance measurements in this work.

An alternative way to view the autocorrelation has been discussed by Foreman-Mackey *et al.* (2013) when using the *emcee* package. They pointed out that the acceptance fraction should be in the 0.2 to 0.5 range in order to see if *emcee* performs well. In this test, the acceptance fractions observed fell in this range of value, meaning that the way we used the MCMC algorithm was convenient.

The results obtained from the test are summarised in Table 4.1 for A_G, and in Table 4.2 for A_V. The rest of the figures are listed in Appendix.

Table 4.1: Summary of the result obtained for the test of the A_G model with the Orion A.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Box used	l	b	D_{cutoff}	A_G cutoff	N	D_{A_G}	$\mu_1^{A_G}$	$\sigma_1^{A_G}$	$\mu_2^{A_G}$	$\sigma_2^{A_G}$
	(deg)	(deg)	(kpc)	(mag)		(pc)	(mag)	(mag)	(mag)	(mag)
(a)	207 < l < 214	-21 < b < -18	1	3.6	6071	399^{+8}_{-8}	$0.09^{+0.01}_{-0.00}$	$0.07^{+0.01}_{-0.01}$	$0.5^{+0.09}_{-0.09}$	$1.27^{+0.05}_{-0.05}$
(b)	212 < l < 214	-20.5 < b < -18.5	1	3.6	869	455^{+21}_{-20}	$0.11^{+0.02}_{-0.02}$	$0.14^{+0.03}_{-0.03}$	$1.79^{+0.08}_{-0.08}$	$0.88^{+0.07}_{-0.06}$
(c)	210 < l < 212	-20.5 < b < -18.5	1	3.6	998	410^{+19}_{-12}	$0.09^{+0.01}_{-0.01}$	$0.09^{+0.02}_{-0.02}$	$1.31^{+0.08}_{-0.10}$	$1.03^{+0.08}_{-0.07}$
(d)	207.5 < l < 210	-20.5 < b < -18.25	1	3.6	1243	340^{+13}_{-10}	$0.07^{+0.01}_{-0.01}$	$0.50^{+0.01}_{-0.01}$	$1.14^{+0.07}_{-0.08}$	$1.07^{+0.06}_{-0.05}$
(c)	207 < l < 214	-21 < b < -18	1	1.5	749	438^{+21}_{-15}	$0.08^{+0.01}_{-0.01}$	$0.07^{+0.01}_{-0.00}$	$0.78^{+0.03}_{-0.03}$	$0.32^{+0.03}_{-0.03}$
(c)	207 < l < 214	-21 < b < -18	2	3.6	1529	406^{+25}_{-30}	$0.10^{+0.02}_{-0.02}$	$0.12^{+0.02}_{-0.02}$	$1.52^{+0.06}_{-0.06}$	$0.99^{+0.06}_{-0.05}$
(c) (50% dist err)	207 < l < 214	-21 < b < -18	1	3.6	998	417^{+23}_{-26}	$0.10^{+0.01}_{-0.01}$	$1.10^{+0.02}_{-0.02}$	$1.34^{+0.09}_{-0.01}$	$1.02^{+0.08}_{-0.07}$
(1b)	213 < l < 214	-19.5 < b < -18.5	1	3.6	199	469^{+38}_{-43}	$0.06^{+0.02}_{-0.02}$	$0.02^{+0.02}_{-0.01}$	$2.14^{+0.139}_{-0.11}$	$0.79^{+0.15}_{-0.09}$
(2b)	212 < l < 213	-19.5 < b < -18.5	1	3.6	218	477^{+38}_{-29}	$0.13^{+0.04}_{-0.05}$	$0.19^{+0.04}_{-0.04}$	$2.00^{+0.09}_{-0.10}$	$0.60^{+0.09}_{-0.07}$
(3b)	213 < l < 214	-20.5 < b < -19.5	1	3.6	198	437^{+34}_{-50}	$0.13^{+0.03}_{-0.03}$	$0.10^{+0.04}_{-0.04}$	$1.63^{+0.16}_{-0.22}$	$1.00^{+0.24}_{-0.15}$
(4b)	211 < l < 212	-20.5 < b < -19.5	1	3.6	254	452^{+66}_{-47}	$0.13^{+0.04}_{-0.04}$	$0.15^{+0.06}_{-0.06}$	$1.37^{+0.21}_{-0.27}$	$1.05^{+0.23}_{-0.15}$
(1c)	213 < l < 214	-19.5 < b < -18.5	1	3.6	209	521^{+29}_{-45}	$0.14^{+0.03}_{-0.03}$	$0.14^{+0.04}_{-0.03}$	$1.91^{+0.11}_{-0.12}$	$0.70^{+0.12}_{-0.08}$
(2c)	210 < l < 211	-19.5 < b < -18.5	1	3.6	264	488^{+28}_{-39}	$0.11^{+0.04}_{-0.05}$	$0.20^{+0.05}_{-0.05}$	$1.96^{+0.08}_{-0.08}$	$0.60^{+0.07}_{-0.06}$
(3c)	211 < l < 212	-20.5 < b < -19.5	1	3.6	242	386^{+40}_{-53}	$0.10^{+0.03}_{-0.2}$	$0.04^{+0.03}_{-0.03}$	$0.50^{+0.36}_{-0.32}$	$1.38^{+0.21}_{-0.22}$
(4c)	210 < l < 211	-20.5 < b < -19.5	1	3.6	283	282^{+46}_{-38}	$0.04^{+0.01}_{-0.01}$	$0.01^{+0.01}_{-0.00}$	$0.52^{+0.26}_{-0.31}$	$1.21^{+0.18}_{-0.16}$
(1d)	208.8 < l < 210	-19.6 < b < -18.4	1	3.6	365	386^{+27}_{-31}	$0.08^{+0.02}_{-0.02}$	$0.04^{+0.03}_{-0.02}$	$2.10^{+0.09}_{-0.09}$	$0.84^{+0.09}_{-0.07}$
(2d)	207.5 < l < 208.8	-19.6 < b < -18.4	1	3.6	482	253^{+27}_{-34}	$0.03^{+0.01}_{-0.01}$	$0.02^{+0.01}_{-0.01}$	$1.18^{+0.09}_{-0.09}$	$0.86^{+0.07}_{-0.07}$
(3d)	208.8 < l < 210	-20.5 < b < -19.5	1	3.6	365	283^{+46}_{-30}	$0.09^{+0.02}_{-0.02}$	$0.06^{+0.02}_{-0.02}$	$0.24^{+0.27}_{-0.17}$	$1.52^{+0.16}_{-0.14}$
(4d)	207.5 < l < 208.8	-20.5 < b < -19.5	1	3.6	345	337^{+37}_{-34}	$0.05^{+0.02}_{-0.02}$	$0.04^{+0.02}_{-0.02}$	$1.15^{+0.11}_{-0.14}$	$0.90^{+0.11}_{-0.09}$

NOTE: *Distances to the Orion A using A_G extinction. In (1) we list the region used for the test in Fig. 4.1. In (2) and (3) we list the Galactic coordinates range corresponding to each box. In (4) and (5) we list the distance cutoff and the extinction cutoff applied to each box. In (6) and (7) we list the number of stars used and the distances obtained from the A_G model. In (8), (9), (10), and (11) we list the inferred foreground extinction with its error, and background extinction with its error, respectively.*

Table 4.2: Summary of the result obtained for the test of the A_V model with the Orion A.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Box used	l	b	D_{cutoff}	A_V cutoff	N	D_{A_V}	$\mu_1^{A_V}$	$\sigma_1^{A_V}$	$\mu_2^{A_V}$	$\sigma_2^{A_V}$
	(deg)	(deg)	(kpc)	(mag)		(pc)	(mag)	(mag)	(mag)	(mag)
(a)	207 < l < 214	-21 < b < -18	1	3	11770	427^{+5}_{-4}	$0.21^{+0.00}_{-0.00}$	$0.09^{+0.00}_{-0.00}$	$1.41^{+0.01}_{-0.01}$	$0.65^{+0.01}_{-0.01}$
(b)	212 < l < 214	-20.5 < b < -18.5	1	3	1511	413^{+11}_{-10}	$0.20^{+0.01}_{-0.01}$	$0.08^{+0.01}_{-0.01}$	$1.66^{+0.02}_{-0.02}$	$0.56^{+0.02}_{-0.02}$
(c)	210 < l < 212	-20.5 < b < -18.5	1	3	1885	407^{+11}_{-10}	$0.20^{+0.01}_{-0.01}$	$0.08^{+0.01}_{-0.01}$	$1.58^{+0.02}_{-0.02}$	$0.57^{+0.01}_{-0.01}$
(d)	207.5 < l < 210	-20.5 < b < -18.25	1	3	3100	349^{+8}_{-7}	$0.23^{+0.01}_{-0.01}$	$0.10^{+0.00}_{-0.00}$	$1.57^{+0.02}_{-0.02}$	$0.62^{+0.01}_{-0.01}$
(c)	207 < l < 214	-21 < b < -18	1	2	1500	438^{+10}_{-12}	$0.19^{+0.01}_{-0.01}$	$0.08^{+0.01}_{-0.00}$	$1.30^{+0.02}_{-0.02}$	$0.44^{+0.01}_{-0.02}$
(c)	207 < l < 214	-21 < b < -18	2	3	2984	341^{+19}_{-16}	$0.20^{+0.01}_{-0.01}$	$0.09^{+0.01}_{-0.01}$	$1.59^{+0.01}_{-0.02}$	$1.60^{+0.01}_{-0.01}$
(c) (50% dist err)	207 < l < 214	-21 < b < -18	1	3	1885	394^{+11}_{-15}	$0.20^{+0.00}_{-0.00}$	$0.09^{+0.01}_{-0.01}$	$1.60^{+0.02}_{-0.02}$	$0.55^{+0.01}_{-0.01}$
(1b)	213 < l < 214	-19.5 < b < -18.5	1	3.5	316	381^{+24}_{-23}	$0.20^{+0.01}_{-0.01}$	$0.09^{+0.01}_{-0.01}$	$1.88^{+0.04}_{-0.04}$	$0.51^{+0.03}_{-0.03}$
(2b)	212 < l < 213	-19.5 < b < -18.5	1	3.5	319	448^{+29}_{-25}	$0.19^{+0.01}_{-0.01}$	$0.07^{+0.01}_{-0.01}$	$1.85^{+0.05}_{-0.05}$	$0.56^{+0.04}_{-0.04}$
(3b)	213 < l < 214	-20.5 < b < -19.5	1	3.5	354	370^{+30}_{-24}	$0.19^{+0.01}_{-0.01}$	$0.07^{+0.01}_{-0.01}$	$1.66^{+0.04}_{-0.04}$	$0.52^{+0.03}_{-0.03}$
(4b)	211 < l < 212	-20.5 < b < -19.5	1	3	511	426^{+27}_{-23}	$0.22^{+0.01}_{-0.01}$	$0.10^{+0.01}_{-0.01}$	$1.59^{+0.04}_{-0.04}$	$0.56^{+0.03}_{-0.03}$
(1c)	213 < l < 214	-19.5 < b < -18.5	1	3.5	386	425^{+25}_{-26}	$0.22^{+0.01}_{-0.01}$	$0.10^{+0.01}_{-0.01}$	$1.72^{+0.03}_{-0.03}$	$0.42^{+0.03}_{-0.02}$
(2c)	210 < l < 211	-19.5 < b < -18.5	1	3.5	504	350^{+22}_{-20}	$0.21^{+0.01}_{-0.01}$	$0.09^{+0.01}_{-0.01}$	$1.87^{+0.03}_{-0.03}$	$0.42^{+0.02}_{-0.02}$
(3c)	211 < l < 212	-20.5 < b < -19.5	1	3	478	328^{+33}_{-24}	$0.16^{+0.01}_{-0.01}$	$0.05^{+0.01}_{-0.01}$	$1.39^{+0.04}_{-0.04}$	$0.60^{+0.03}_{-0.03}$
(4c)	210 < l < 211	-20.5 < b < -19.5	1	3.5	484	402^{+31}_{-22}	$0.18^{+0.01}_{-0.01}$	$0.07^{+0.01}_{-0.01}$	$1.37^{+0.05}_{-0.05}$	$0.67^{+0.03}_{-0.03}$
(1d)	208.8 < l < 210	-19.6 < b < -18.4	1	3.5	577	323^{+19}_{-15}	$0.24^{+0.01}_{-0.01}$	$0.11^{+0.01}_{-0.01}$	$2.01^{+0.03}_{-0.03}$	$0.56^{+0.03}_{-0.02}$
(2d)	207.5 < l < 208.8	-19.6 < b < -18.4	1	3.5	961	268^{+17}_{-20}	$0.23^{+0.01}_{-0.01}$	$0.10^{+0.01}_{-0.01}$	$1.53^{+0.02}_{-0.03}$	$0.60^{+0.02}_{-0.02}$
(3d)	208.8 < l < 210	-20.5 < b < -19.5	1	3.5	628	394^{+30}_{-21}	$0.25^{+0.01}_{-0.01}$	$0.11^{+0.01}_{-0.01}$	$1.52{-}0.04^{+0.04}$	$0.72^{+0.03}_{-0.03}$
(4d)	207.5 < l < 208.8	-20.5 < b < -19.5	1	3.5	717	341^{+24}_{-18}	$0.22^{+0.01}_{-0.01}$	$0.10^{+0.01}_{-0.01}$	$1.57^{+0.03}_{-0.03}$	$0.58^{+0.02}_{-0.02}$

NOTE: *Distances to the Orion A using A_V extinction. For A_V extinction, we use use a 3 mag cutoff to remove stars with larger extinction error. For more details, see Caption 4.1.*

4.2 Distance catalogs

4.2.1 RMS distances

In this section, we present distances of YSOs selected from the Leeds Red MSX Source Survey (RMS, Lumsden *et al.* (2013)). Those stars are taken from different regions with different dust environments, and have distances ranging from 0.5 to 3 kpc which are mostly derived from kinematic method. The RMS selected candidates are displayed in Table 4.3, and our extinction distances are listed in Table 4.4 alongside the RMS distances.

Overall, we inferred distances to all selected target except the A_G distance of G034.4035+00.2282A. As shown in the top left panel of Fig. 4.12, this target is surrounded by a very few optical stars which is not enough for the distance measurement.

Table 4.3: List of RMS stars used in this work.

(1)	(2)	(3)	(4)	(5)
Name	l	b	Distance (kpc)	Distance type
G010.384+2.213	10.3844	2.2128	1.1	Kin
G108.929+2.595	108.9288	2.5954	0.7	Par
G126.714-0.822	126.714	-0.822	0.7	Kin
G014.4886+0.0219B	14.4886	0.0219	2.5	Kin
G065.3169-2.7141	65.3169	-2.7141	1.2	Kin
G034.4035+0.2282A	34.4035	0.228	1.6	Par

NOTE: *List of stars used in this work. The name, Galactic coordinates (l, b), and source type are shown in (1), (2), (3) and (4) respectively, while the estimated distances and the type of method used from the literature are displayed in (4) and (5), respectively.*

4.2.2 Cygnus X distances

Apart from that, we derived distances to the complex star formation region, Cygnus X. We are particularly interested in Cygnus X as it is a nearby massive star formation region located at $\sim$ 2 kpc (Reid *et al.*, 2011) that *Gaia* also can

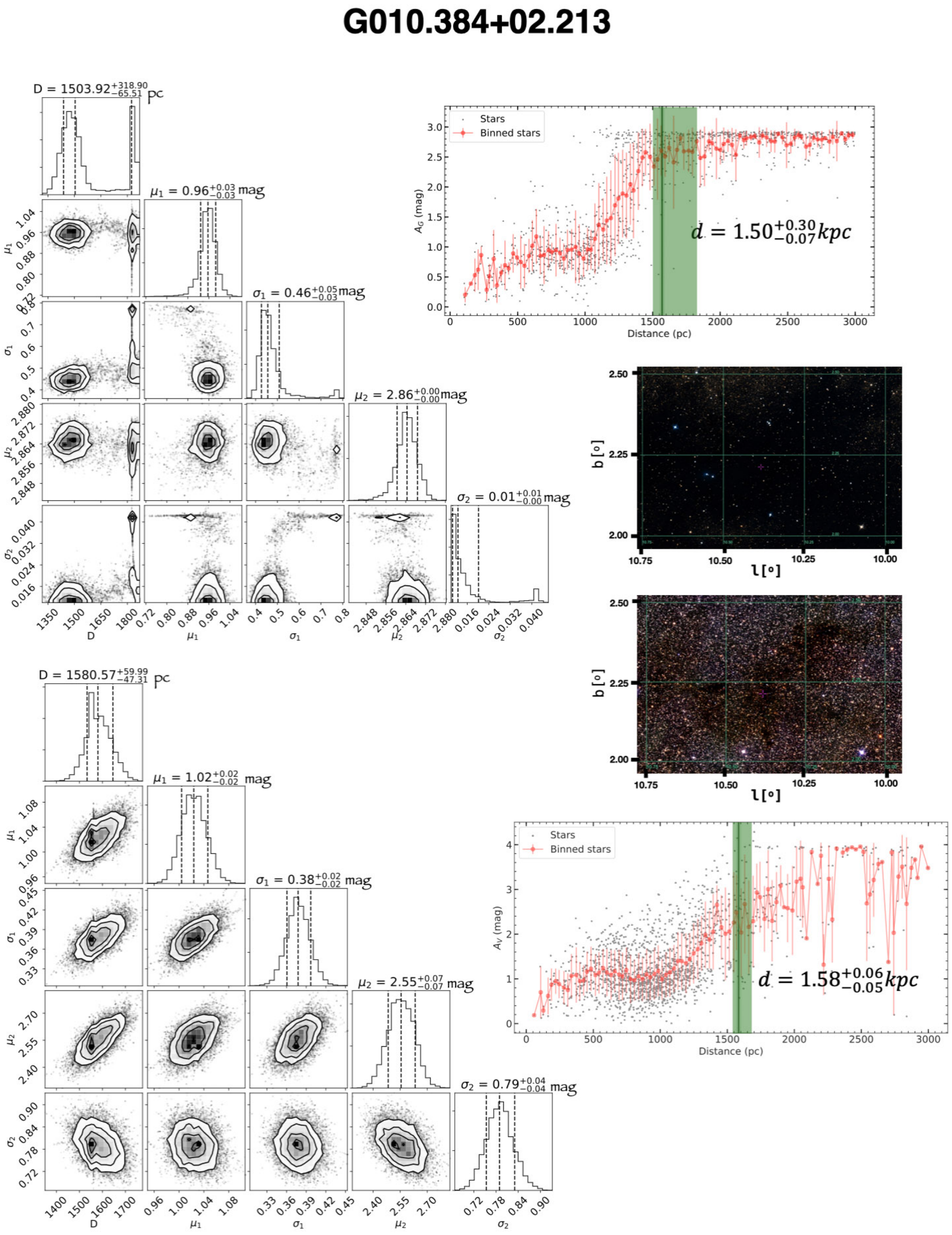

Figure 4.7: The distance of G010.384+02.213. The two images in the middle of right panels are taken from the ESASky website (Martí *et al.*, 2016), which is a great astronomical tool for viewing. The one in the top is an optical image (SDSS 2 York *et al.* (2000)), while the one in the bottom is a NIR image (2MASS, Skrutskie *et al.* (2006)). The purple plus in the centre of each image represents the location of our object of interest. In the top and bottom right panels, the grey points are the selected stars we used for the distance measurements, while the red points show their corresponding binned data (averaged in every 10 pc). The derived distances and the uncertainty are indicated by the green vertical line. The corner plots of the MCMC samples obtained from A_G (on the top) and A_V (on the bottom) are displayed on the left panels. They show the obtained distance D, the inferred extinction of *Off-cloud* stars μ_1 with its error σ_1, and the extinction for *On-cloud* stars μ_2 with its error σ_2 with the two models, respectively. The 16^{th}, 50^{th}, and the 86^{th} are shown by the dashed vertical lines, respectively.

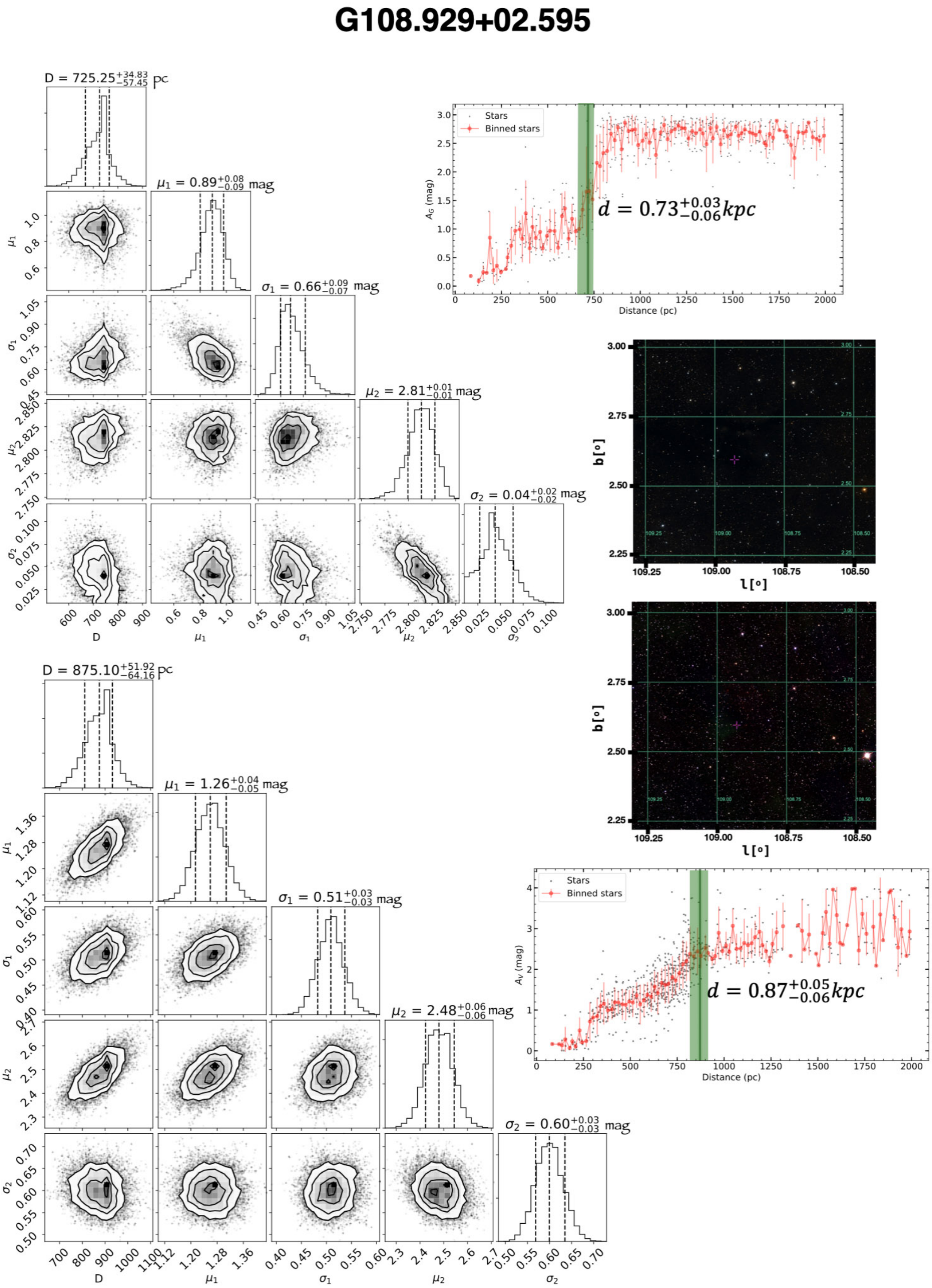

Figure 4.8: The distance of G108.929+02.595. Top left panel and bottom left panel show the corner plot of the MCMC samples obtained from A_G and A_G, respectively. In the right panels, the grey points and red points in the top and bottom show the stars used for the distance calculation and their corresponding binned data (averaged in every 10 pc). The two images in the middle of right panels are optical image and NIR image showing the object of interest and their surrounding stars. See the caption of Fig. 4.7 for more information.

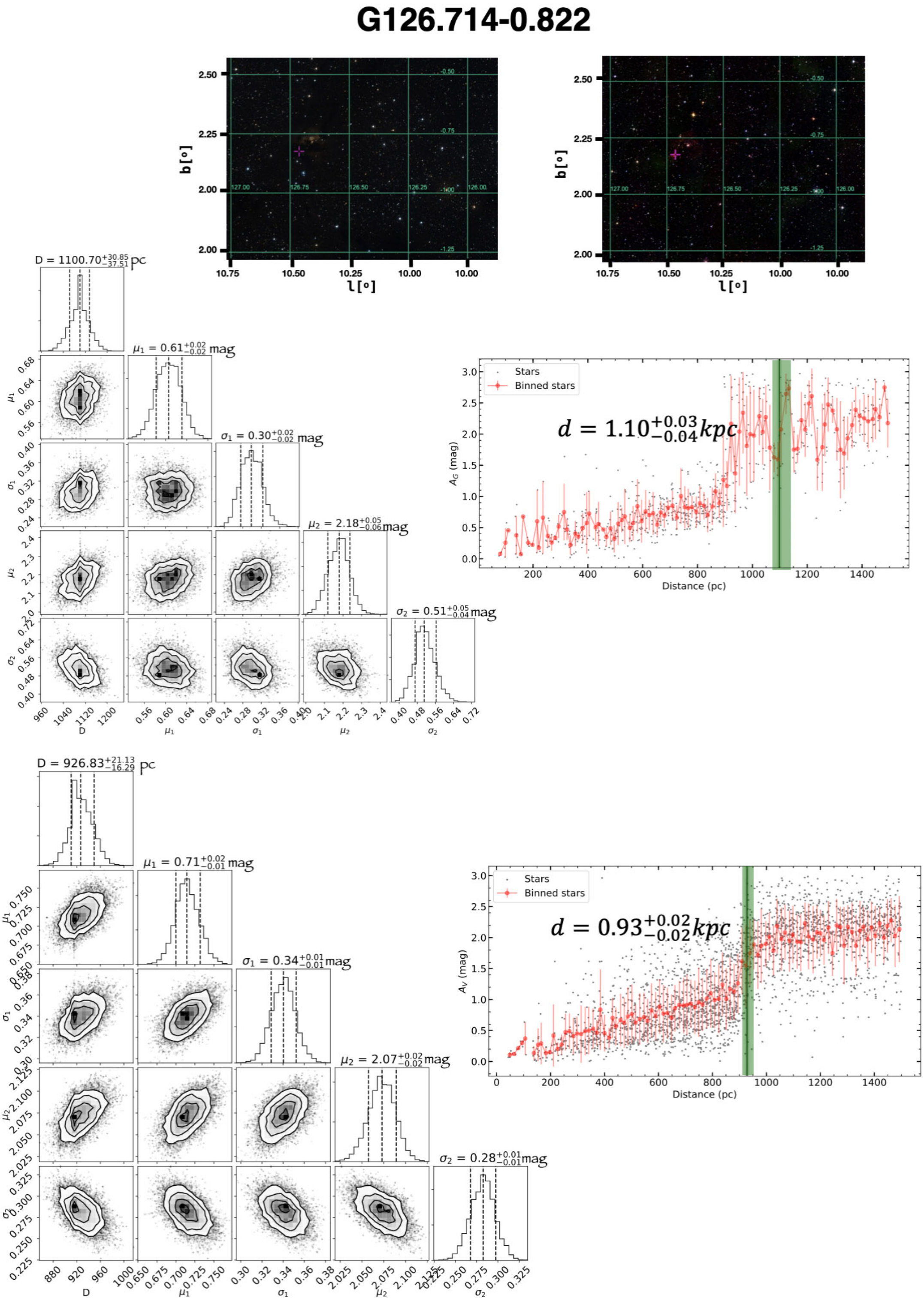

Figure 4.9: The distance of G126.714-0.822. Top panels show the optical image (on the left) and NIR image (on the right) of the region that contain our object of interest. Middle panels and bottom panels show the corner plot of the MCMC samples (on the left), and the stars used for the distance inference with their corresponding binned data (grey and red points on the right), which are obtained from A_G and A_V, respectively. See the caption of Fig. 4.7 for more information.

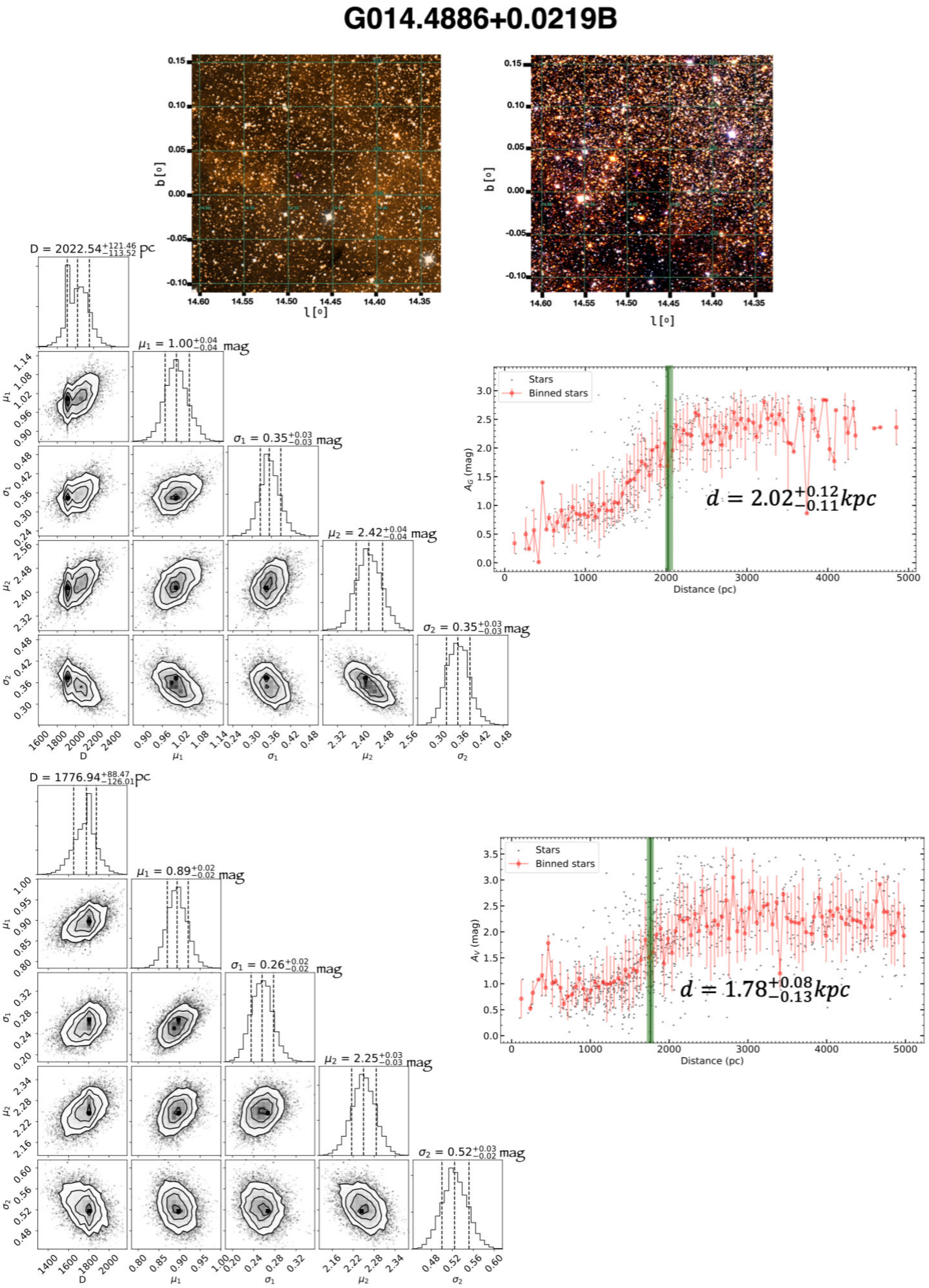

Figure 4.10: The distance of G014.4886+0.0219B. Top panels show the optical image (on the left) and NIR image (on the right) of the region that contains our object of interest. Middle panels and bottom panels show the corner plot of the MCMC samples (on the left), and the stars used for the distance inference with their corresponding binned data (grey and red points on the right), which are obtained from A_G and A_V, respectively. See the caption of Fig. 4.7 for more information.

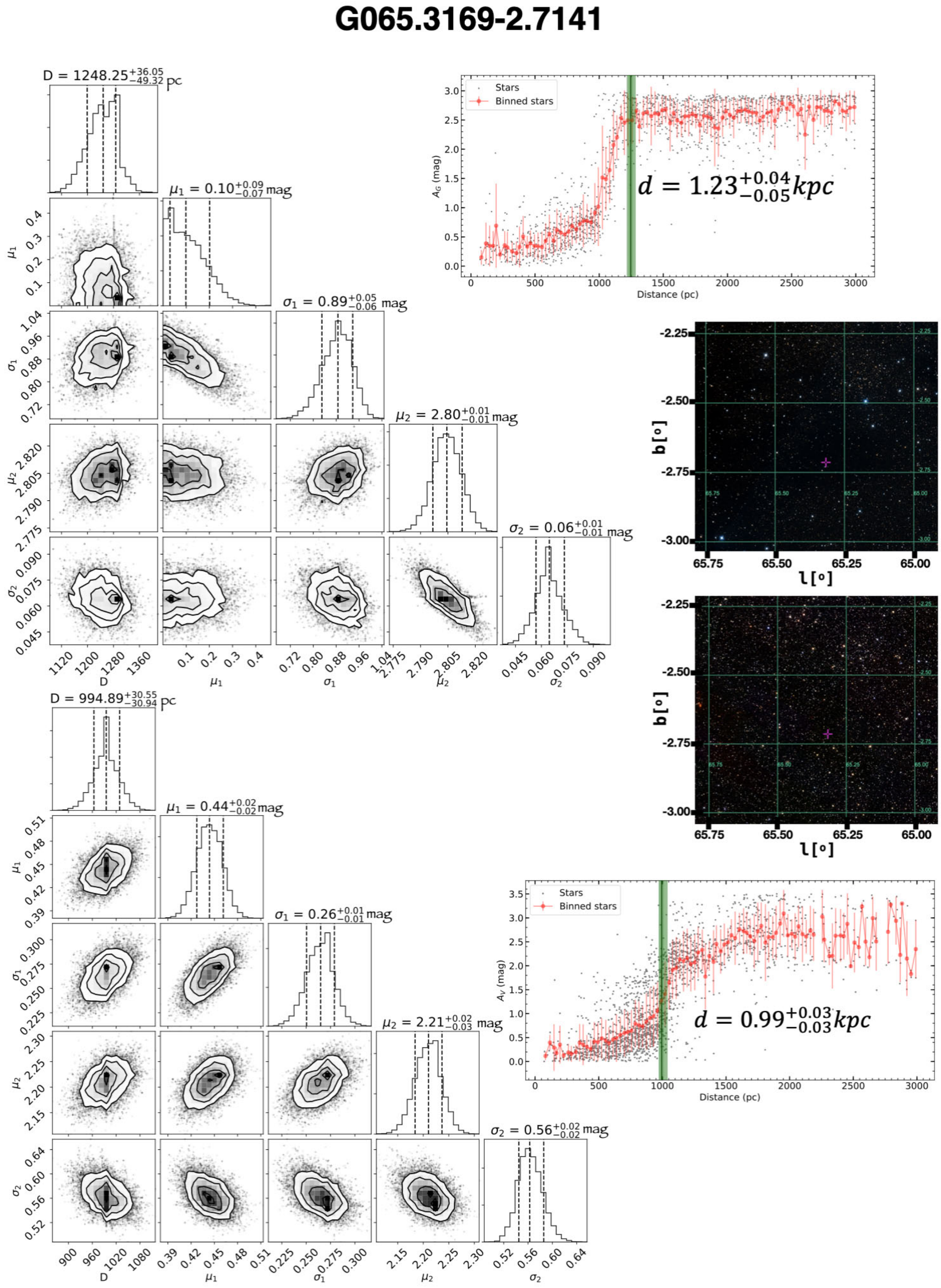

Figure 4.11: The distance of G065.3169-2.7141. Top left panel and bottom left panel show the corner plot of the MCMC samples obtained from A_G and A_G, respectively. In the right panels, the grey points and red points in the top and bottom show the stars used for the distance calculation and their corresponding binned data (averaged in every 10 pc). The two images in the middle of right panels are optical image and NIR image showing the object of interest and their surrounding stars. See the caption of Fig. 4.7 for more information.

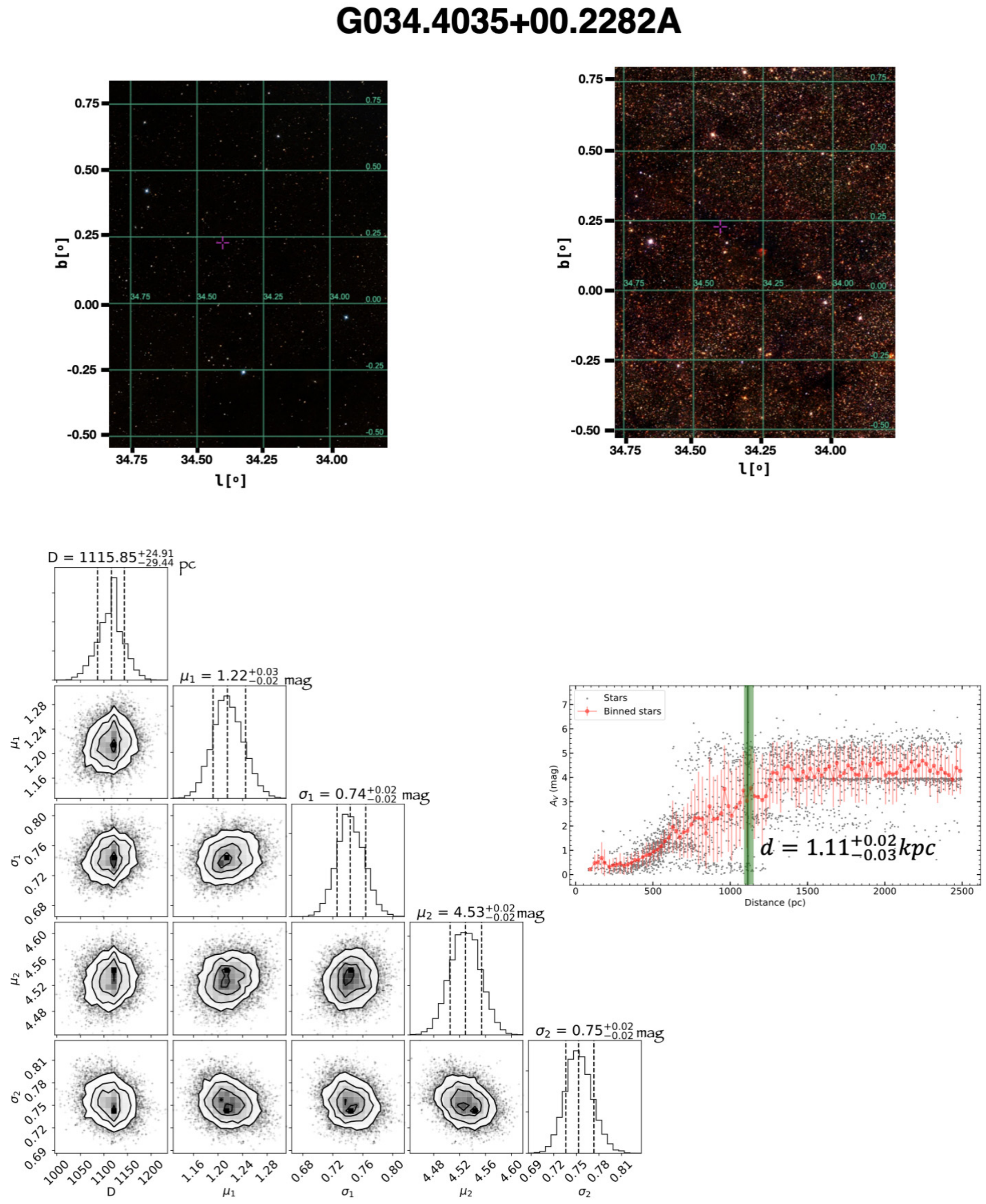

Figure 4.12: The distance of G034.4035+00.2282A. Top panels show the optical image (on the left) and NIR image (on the right) of the region that contains our object of interest. Bottom left panel shows the corner plot of the MCMC samples, while bottom right panel depicts the stars used for the distance inference (grey points) with their corresponding binned data (red points). See the caption of Fig. 4.7 for more details.

Table 4.4: RMS distances results.

(1)	(2)	(3)	(4)	(5)	(6)
Name	l	b	Distance	Distance using A_G	Distance using A_V
			(kpc)	(kpc)	(kpc)
G010.384+02.213	10.3844	02.2128	1.1	$1.50^{+0.3}_{-0.07} \pm 0.07$	$1.58^{+0.06}_{-0.05} \pm 0.08$
G108.929+02.595	108.9288	02.5954	0.7	$0.73^{+0.03}_{-0.06} \pm 0.03$	$0.87^{+0.05}_{-0.06} \pm 0.04$
G126.714-00.822	126.714	-00.822	0.7	$1.10^{+0.03}_{-0.04} \pm 0.05$	$0.93^{+0.02}_{-0.02} \pm 0.04$
G014.4886+00.0219B	14.4886	0.0219	2.5	$2.02^{+0.12}_{-0.11} \pm 0.10$	$1.78^{+0.08}_{-0.13} \pm 0.09$
G065.3169-02.7141	65.3169	-02.7141	1.2	$1.23^{+0.04}_{-0.05} \pm 0.06$	$0.99^{+0.03}_{-0.03} \pm 0.05$
G034.4035+00.2282A	34.4035	0.228	1.6	–	$1.11^{+0.02}_{-0.03} \pm 0.05$

Note: *The distances of YSOs selected from the RMS survey. The distance calculated with both A_G and A_V are shown in (5) and (6) respectively. The first error term in the distance estimates is the statistical uncertainty while the second error term is the 5% systematic uncertainty.*

deal with. In this part, provide distances to its sub-regions including DR 20, DR21, DR22, DR23, and W75N. This helps to know whether those sub-regions are located at the same distance or not.

Recent work such as Beerer *et al.* (2010) has proved the high number of YSOs in the Cygnus X region. In their work, they identified 670 Class I, 7,249 Class II, 112 transition disk, and 200 embedded protostellar sources.

The Cygnux X sub-regions we used are shown in Fig. 4.13. We failed to measure the distance towards DR23 using the A_V model due to the complex distribution of A_V extinction. The distances obtained for the other sub-regions are summarised in Table 4.5.

4.2.3 Distances to additional sources

We also derived distances to G5.89-0.39, G35.20-0.74, and G59.7+0.1. Those are a well-studied high mass star forming region, and also have distances measurement in the literature. For example, Motogi *et al.* (2011) derived distance $1.28^{+0.09}_{-0.08}$ kpc to G5.89-0.39 with the use of VERA (VLBI Exploration of Radio Astrometry), Zhang *et al.* (2009) found a distance of $2.19^{+0.24}_{-0.20}$ kpc to G35.20-0.74, and Xu

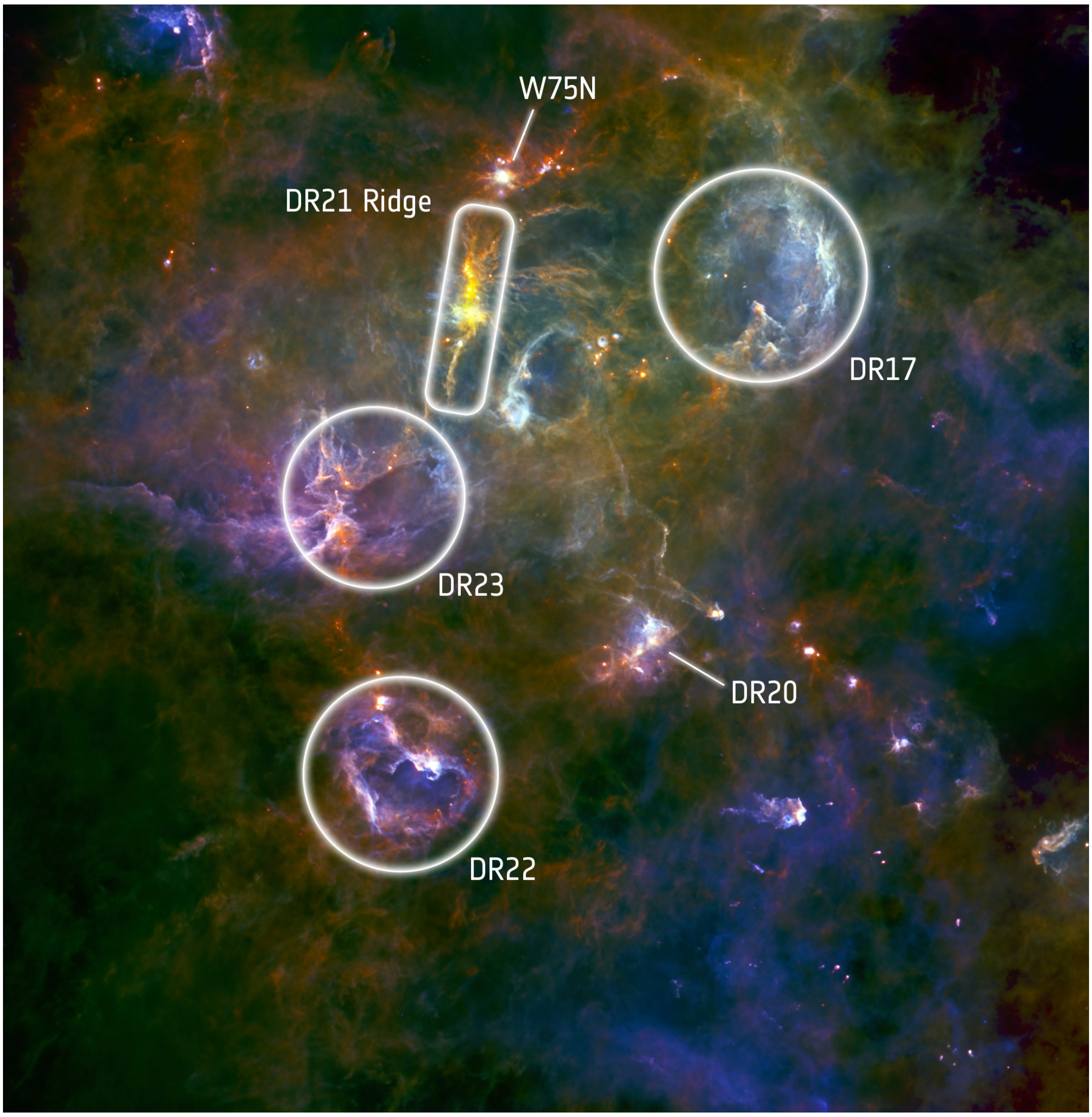

Figure 4.13: Infrared image of the extremely active star-forming region Cygnus X. The distance of the marked areas in the figure are calculated in this work. Credit: ESA.

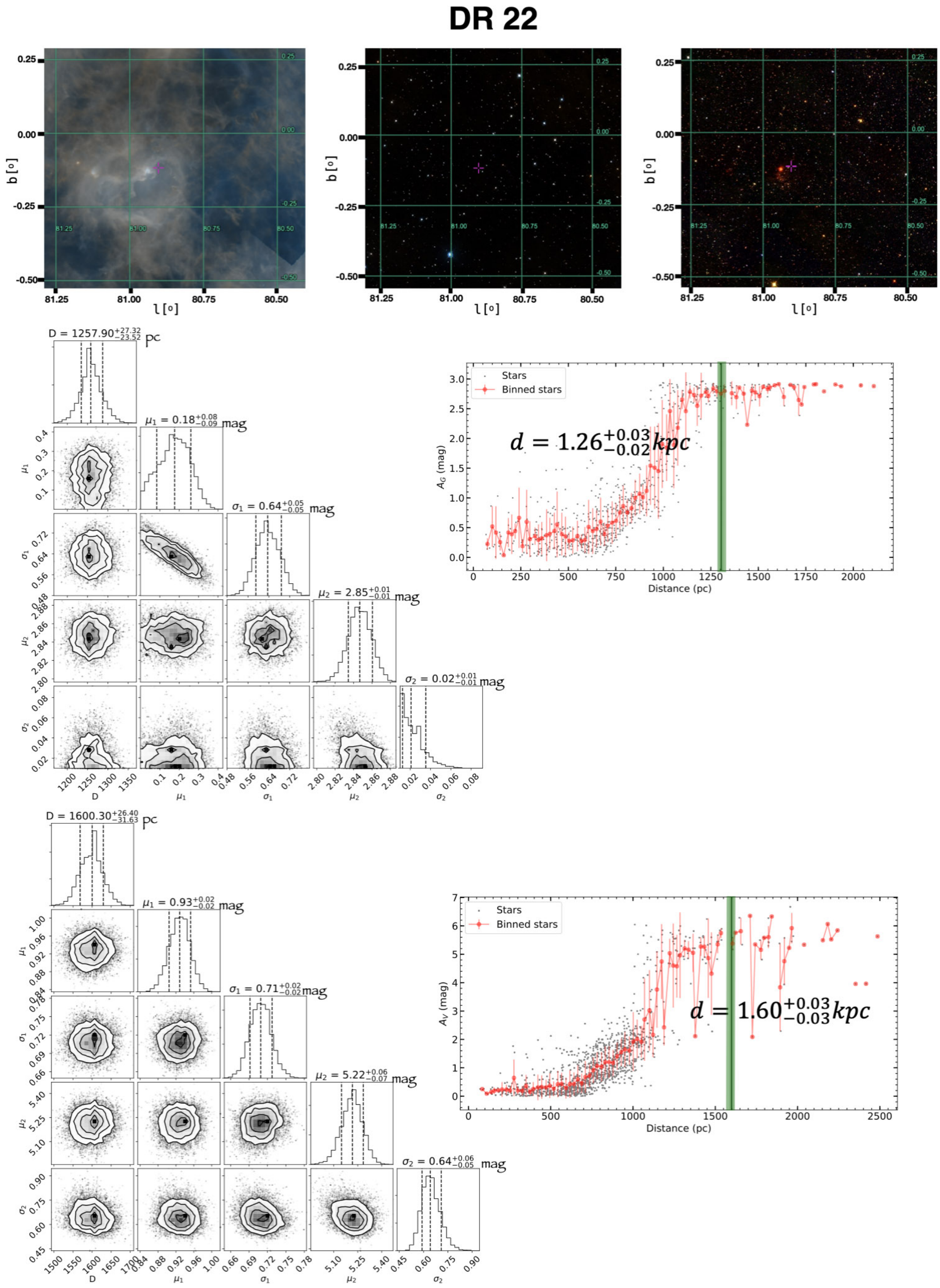

Figure 4.14: The distance of DR22. Top panels display, from left to right, a far-infrared (FIR) image (Herschel, Marton *et al.* (2015)), an optical image (SDSS 2, York *et al.* (2000)), and a near-infrared (NIR) image (Skrutskie *et al.* (2006)) of the source of interest. Middle and bottom left panel display the corner plot of the MCMC samples obtained from A_G and A_V, respectively. The corner plots show the obtained distance D, the inferred extinction of *Off-cloud* stars μ_1 with its error σ_1, and the extinction for *On-cloud* stars μ_2 with its error σ_2 with the two models. The 16^{th}, 50^{th}, and the 86^{th} are shown by the dashed vertical lines, respectively. The grey points in the middle and bottom right panel are the selected stars we used for the distance measurements. The the red points show their corresponding binned data (averaged in every 10 pc), while derived distances and the uncertainty are indicated by the green vertical line.

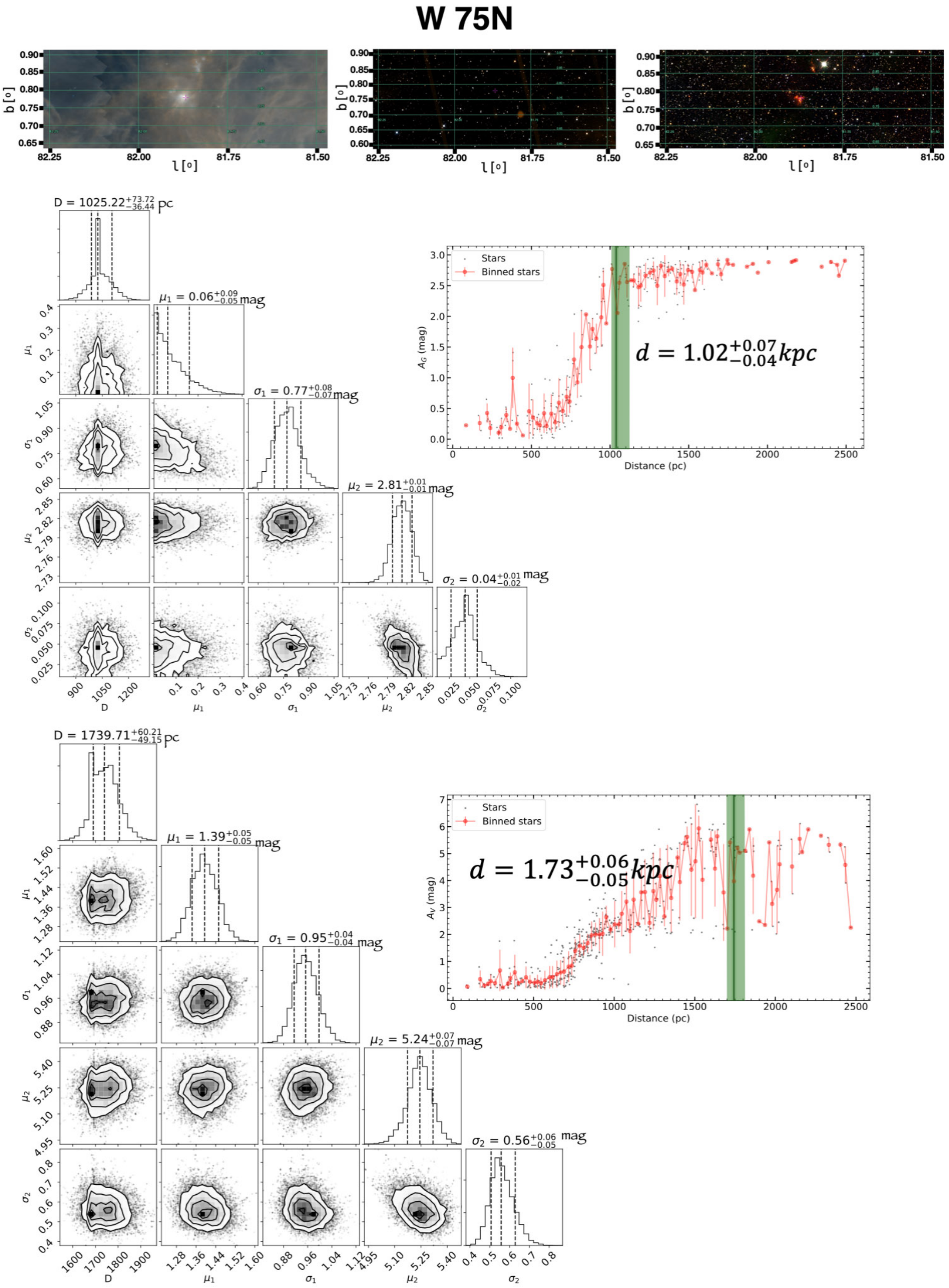

Figure 4.15: The distance of W75N. See the caption of Fig. 4.14 for more details.

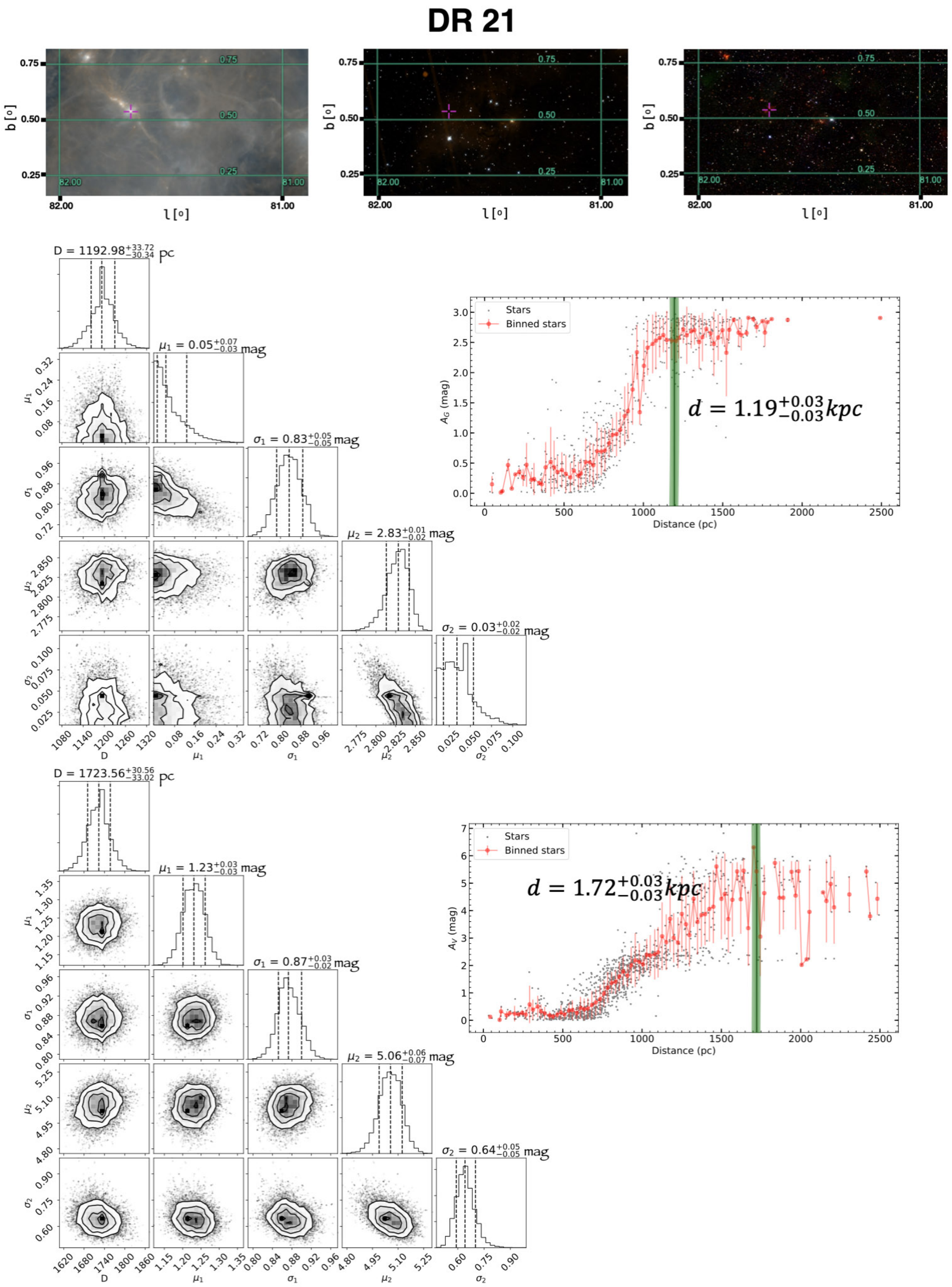

Figure 4.16: The distance of DR21. See the caption of Fig. 4.14 for more details.

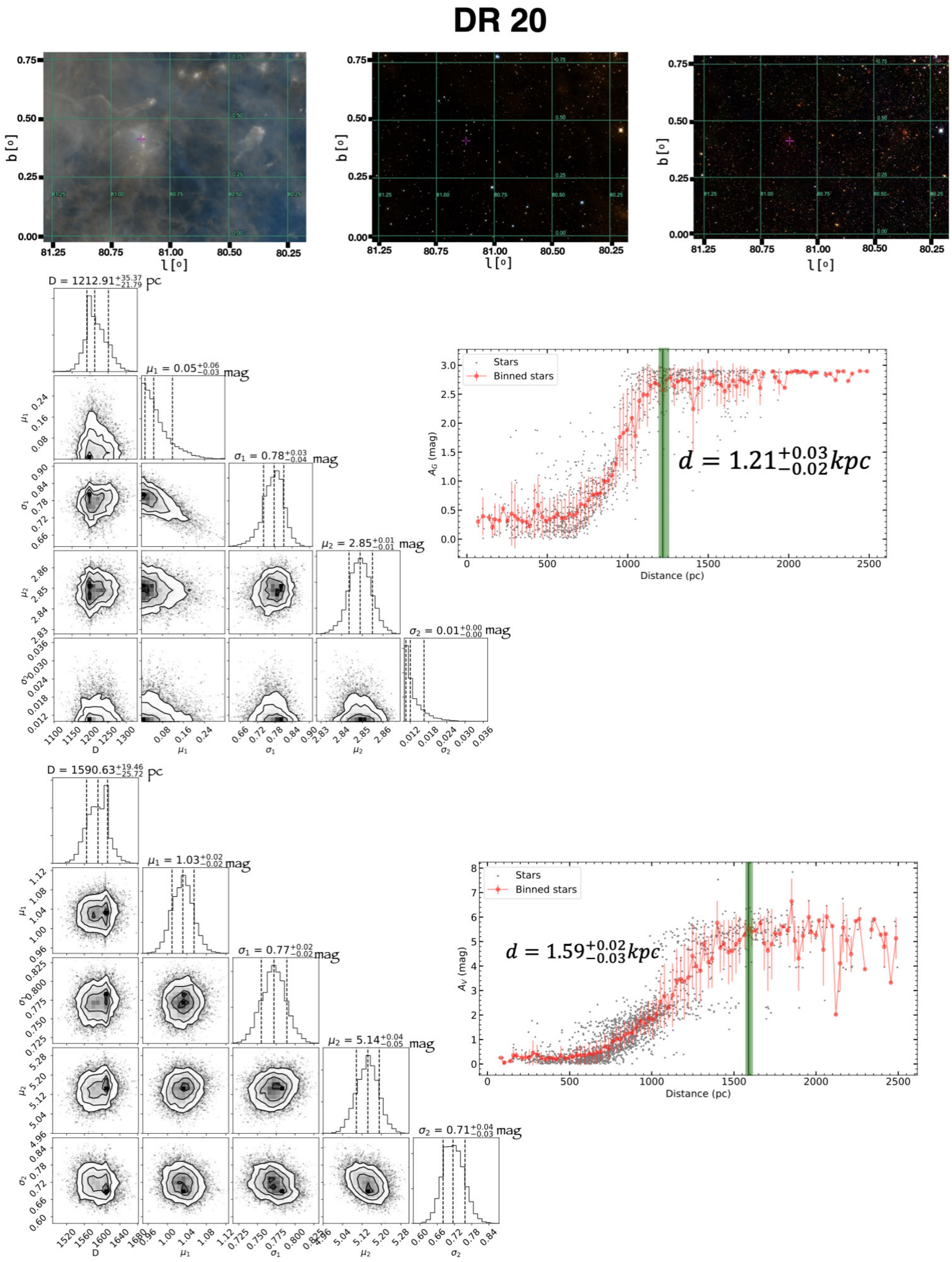

Figure 4.17: The distance of DR20. See the caption of Fig. 4.14 for more details.

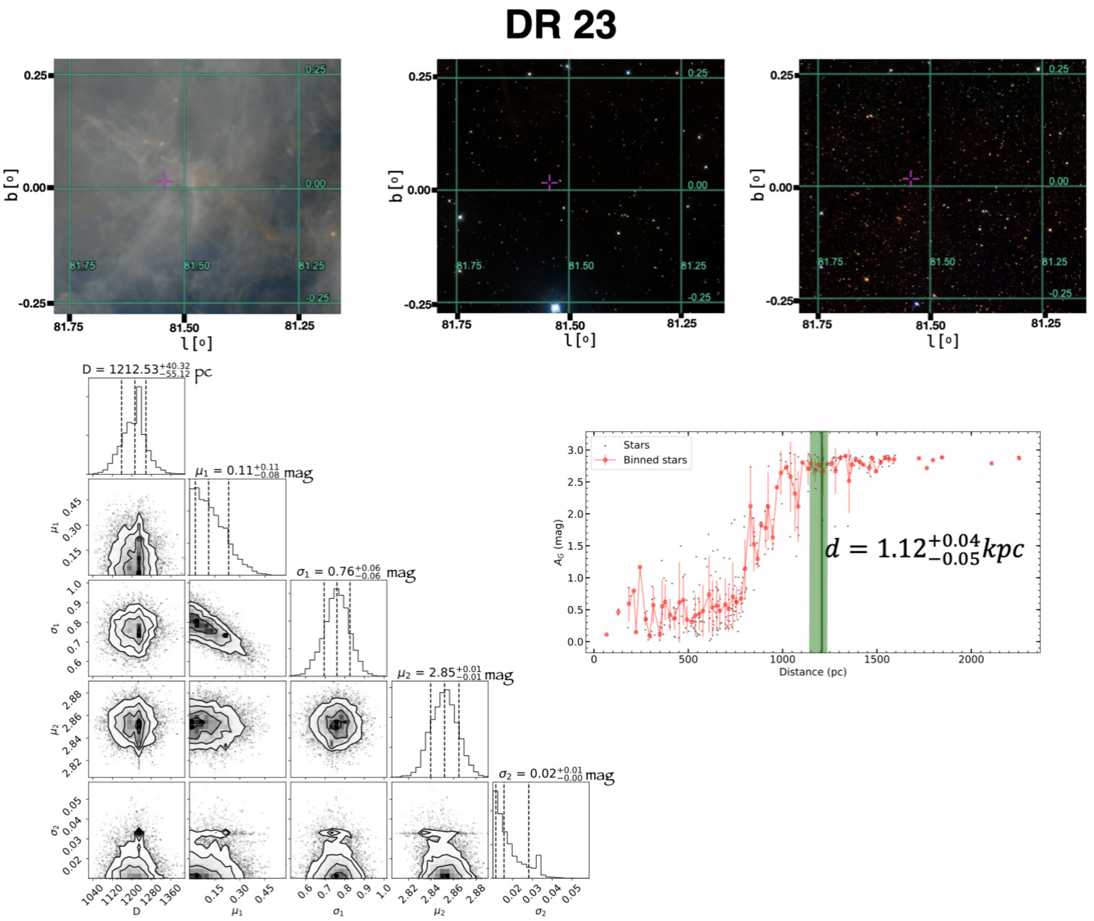

Figure 4.18: The distance of DR23. See the caption of Fig. 4.14 for more details.

Table 4.5: Distances to the sub-regions of Cygnus X. The statistical uncertainty and the 5% systematic uncertainty in the distances are shown.

Name	l	b	Distance using A_G (kpc)	Distance using A_V (kpc)
DR 22	80.903	-0.117	$1.26^{+0.03}_{-0.02} \pm 0.06$	$1.60^{+0.03}_{-0.03} \pm 0.08$
W 75N	81.867	+0.779	$1.02^{+0.07}_{-0.04} \pm 0.05$	$1.73^{+0.06}_{-0.05} \pm 0.08$
DR 21	81.680	+0.537	$1.19^{+0.03}_{-0.03} \pm 0.06$	$1.72^{+0.03}_{-0.03} \pm 0.08$
DR 20	80.872	+0.411	$1.21^{+0.3}_{-0.02} \pm 0.06$	$1.59^{+0.02}_{-0.02} \pm 0.08$
DR 23	81.543	+0.016	$1.12^{+0.04}_{-0.05} \pm 0.06$	–

et al. (2007) derived a distance of 2.20 ± 0.11 kpc to G59.7+0.1 using VLBA maser parallax.

By using our extinction models, the observed distances for G5.89-0.39 are $1.27^{+0.05}_{-0.08} \pm 0.06$ kpc with A_G and $1.25^{+0.06}_{-0.07} \pm 0.06$ kpc with A_V (see Fig. 4.19). For G35.20-0.74, we inferred $1.30^{+0.08}_{-0.09} \pm 0.06$ kpc with A_V (see Fig. 4.20), but we failed to derived A_G distance due to the lack of data. For G59.7+0.1, we obtained $2.24^{+0.10}_{-0.18} \pm 0.1$ kpc and $2.78^{+0.12}_{-0.09} \pm 0.1$ kpc with A_G and A_V, respectively (see Fig. 4.21).

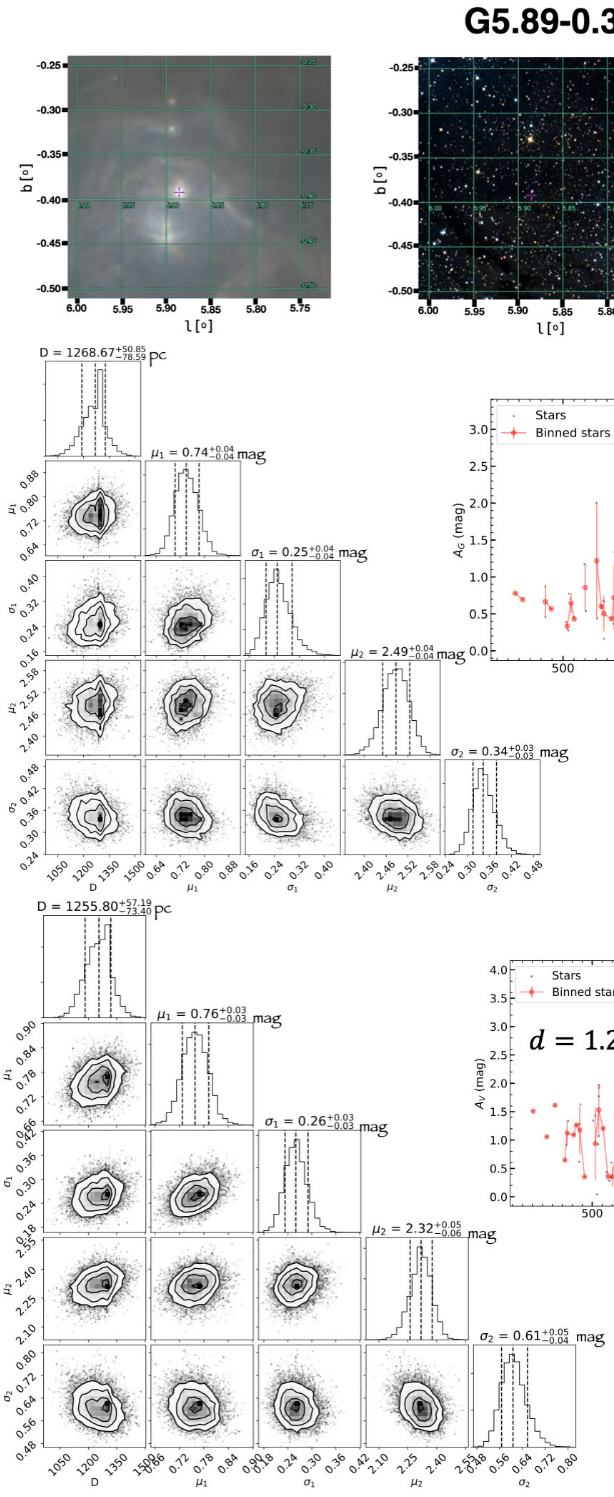

Figure 4.19: The distance of G5.89-0.39. Top panels display, from left to right, a far-infrared (FIR) image (Herschel, Marton *et al.* (2015)), an optical image (SDSS 2, York *et al.* (2000)), and a near-infrared (NIR) image (Skrutskie *et al.* (2006)) of the source of interest. Middle and bottom left panel display the corner plot of the MCMC samples obtained from A_G and A_V, respectively. The corner plots show the obtained distance D, the inferred extinction of *Off-cloud* stars μ_1 with its error σ_1, and the extinction for *On-cloud* stars μ_2 with its error σ_2 with the two models. The 16^{th}, 50^{th}, and the 86^{th} are shown by the dashed vertical lines, respectively. The grey points in the middle and bottom right panel are the selected stars we used for the distance measurements. The the red points show their corresponding binned data (averaged in every 10 pc), while derived distances and the uncertainty are indicated by the green vertical line.

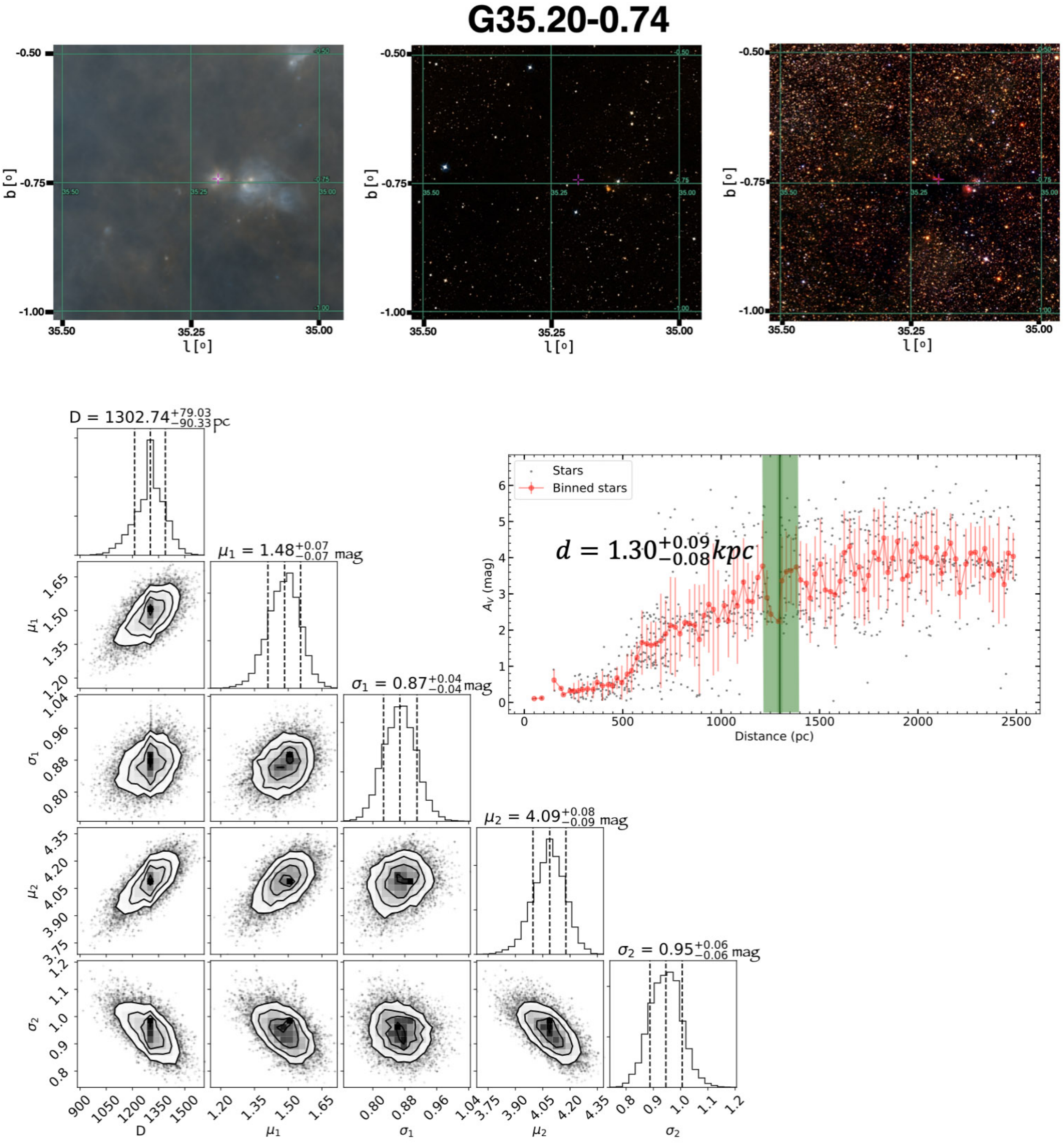

Figure 4.20: The distance of G35.20-0.74. See the caption of Fig. 4.19 for additional information.

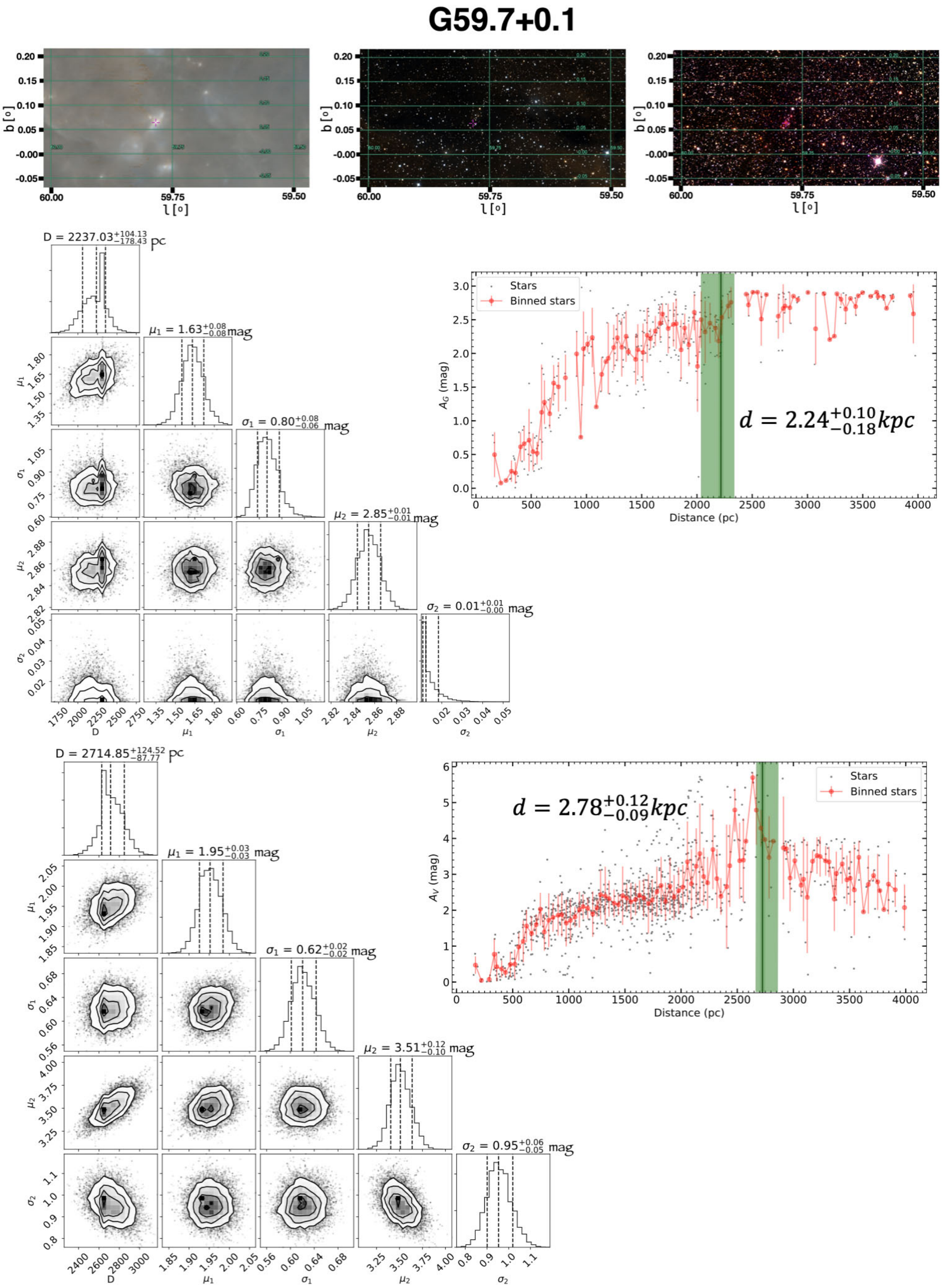

Figure 4.21: The distance of G59.7+0.1. See the caption of Fig. 4.19 for additional information.

4.3 Discussion

4.3.1 Comparison with RMS disances

We first compare our results to the RMS distances listed in Table 4.3. The distance to our selected RMS target are derived from kinematic method and VLBI maser parallax. Our estimated distances from the two models agree well with the literature distance of G108.929+02.595 and G065.3169-02.7141, showing only a ∼0.1 kpc difference in the distance. For the other targets G010.384+02.213, G126.714-00.822, G014.4886+00.0219B, G034.4035+00.2282A, we observed difference of ∼0.5 pc in the distance. The RMS distance are not accompanied with their uncertainty, but if we consider the typical 0.7 kpc error for kinematic distances, our findings are consistent with the literature.

4.3.2 Comparison to kinematic distances

In this section, we compare our results from Sect. 4.2.3 to their literature distances. Foster *et al.* (2012) derived kinematic distances to those high-mass star forming regions using two different rotational curves (Clemens (1985) and Reid *et al.* (2009)). As can be seen in Table 4.6, kinematic distances are systematically greater than our distance measurements. Our distances are about ∼1 kpc lower than those kinematic distances. A possible reason for our lower distance is that those are complex dark clouds that contain high-massive star formation rate (Hampton *et al.*, 2016; Xu *et al.*, 2007), where extinction is higher. Thus

Table 4.6: Comparison with the maser parallax distances from the literature.

Name	Distance using A_G	Distance using A_V	Kinematic distance (kpc)	
	(kpc)	(kpc)	Clemens (1985)	Reid *et al.* (2009)
G5.89-0.39	$1.27^{+0.05}_{-0.08} \pm 0.06$	$1.25^{+0.06}_{-0.07} \pm 0.06$	$2.0^{+0.7}_{-0.7}$	$1.9^{+0.6}_{-0.7}$
G35.20-0.74	–	$1.30^{+0.08}_{-0.09} \pm 0.06$	$2.3^{+0.2}_{-0.2}$	$2.4^{+0.2}_{-0.2}$
G59.7+0.1	$2.24^{+0.10}_{-0.18} \pm 0.1$	$2.78^{+0.12}_{-0.09} \pm 0.1$	$3.3^{+1.0}_{-0.5}$	$4.2^{+1.0}_{-1.0}$

Table 4.7: Maser used in this work.

(1)	(2)	(3)	(4)	(5)	(6)
Name	l	b	Distance [kpc]	Reference	Type
G5.89-0.39	5.8842	-0.3924	$1.28^{+0.09}_{-0.08}$	Motogi *et al.* (2011)	22 GHz H_2O
G35.20-0.74	35.1970	-0.7431	$2.19^{+0.24}_{-0.20}$	Zhang *et al.* (2009)	12 GHz methanol
G59.7+0.1	59.7828	0.0647	$2.16^{+0.10}_{-0.09}$	Xu *et al.* (2009)	12 GHz methanol
W 75N	81.867	+0.779	$1.3^{+0.7}_{-0.07}$	Rygl *et al.* (2012)	12 GHz methanol
DR 21	81.680	+0.537	$1.50^{+0.08}_{-0.07}$	Rygl *et al.* (2012)	12 GHz methanol
DR 20	80.872	+0.411	$1.46^{+0.10}_{-0.09}$	Rygl *et al.* (2012)	12 GHz methanol

NOTE: *Maser selected from the literature. The name, Galactic coordinates (l, b), and the distances are displayed in (1), (2), (3) and (4) respectively. The references and the maser type are shown in (5) and (6), respectively.*

Gaia suffers from the extinction and is not able to detect many stars in the cloud boundaries. Consequently, few stars are available and only lower distance can be produced in the same line of sight towards the dark cloud. Another possible reason is that the line of sight towards those cloud might contain several extinction breakpoint, and hence our models naturally infer the distance of the first jump as there are many datapoints in there.

4.3.3 Comparison with the VLBI maser parallax distances

Some of the sources we study in this thesis have maser parallax distances from the literature. They are listed in Table 4.7 with the literature reference and the maser type. As indicated by Reid & Honma (2014), the VLBI can measure parallaxes for sources across the Milky Way with ∼10 μas accuracy, which is a very accurate measurement. If we suppose that those maser parallax distances are the true distances for those regions, our results display a systematic error of 1-20% except G5.89-0.39, which shows a large 40% systematic error with A_V.

Table 4.8: Comparison of the VLBI maser parallax distance measurement from the literature with the results obtained from this work.

Name	Maser Parallax (kpc)	Distance using A_G (kpc)	Distance using A_V (kpc)
G5.89-0.39	$1.28^{+0.09}_{-0.08}$	$1.27^{+0.05}_{-0.08} \pm 0.06$	$1.25^{+0.06}_{-0.07} \pm 0.06$
G35.20-0.74	$2.19^{+0.24}_{-0.20}$	–	$1.30^{+0.08}_{-0.09} \pm 0.06$
G59.7+0.1	$2.16^{+0.10}_{-0.09}$	$2.24^{+0.10}_{-0.18} \pm 0.1$	$2.78^{+0.12}_{-0.09} \pm 0.1$
W 75N	$1.3^{+0.7}_{-0.07}$	$1.02^{+0.07}_{-0.04} \pm 0.05$	$1.73^{+0.06}_{-0.05} \pm 0.08$
DR 21	$1.50^{+0.08}_{-0.07}$	$1.19^{+0.03}_{-0.03} \pm 0.06$	$1.72^{+0.03}_{-0.03} \pm 0.08$
DR 20	$1.46^{+0.10}_{-0.09}$	$1.21^{+0.3}_{-0.02} \pm 0.06$	$1.59^{+0.02}_{-0.02} \pm 0.08$

4.3.4 Comparison to extinction distance

Here, we compare our results with the extinction distances of Foster *et al.* (2012). They performed two methods of extinction distances in 11 dark clouds that contain maser parallax measurement. The two extinction methods that they used are the Blue Number Count Extinction (BNCE) method and the Red Giant Extinction (RGE) method. Basically, the idea of their blue count extinction method is to count the blue number of stars within the region of interest and compare those number of blue stars to the Galactic model (Robin *et al.*, 2003) to estimate the distance. For the red giant method, they calculated the visual extinction A_V of giant stars along the same line of sight as the cloud, and compared the extinction to a Galactic model (Marshall *et al.*, 2006) to derive the distances. The extinction distances and our results are highlighted in Table 4.9.

Foster *et al.* (2012) used 2MASS and UKIDSS, which are deep infrared imaging surveys. They can observe stars that are heavily embedded in a dark cloud, and therefore they can rich larger distance. Here, their extinction distance are very large compared to our A_G and A_V distance. For G5.89-0.39, the extinction distances from the blue number count and the red giant are almost three times larger than our A_G and A_V distances, and it is two times larger for G35.200.74.

Table 4.9: Comparison with the extinction distances from Foster *et al.* (2012).

Name	Distance using A_G	using A_V	BNCE (kpc)		RGE (kpc)	
	(kpc)	(kpc)	2MASS	UKIDSS	2MASS	UKIDSS
G5.89-0.39	$1.27^{+0.05}_{-0.08} \pm 0.06$	$1.25^{+0.06}_{-0.07} \pm 0.06$	...	$4.5^{+0.6}_{-0.8}$	$3.9^{+1.4}_{-1.1}$	$3.7^{+0.3}_{-0.4}$
G35.200.74	–	$1.30^{+0.08}_{-0.09} \pm 0.06$	$2.4^{+0.8}_{-0.7}$	$2.7^{+0.3}_{-0.3}$	$2.4^{+0.3}_{-0.2}$	$2.6^{+0.7}_{-0.2}$
G59.7+0.1	$2.24^{+0.10}_{-0.18} \pm 0.1$	$2.78^{+0.12}_{-0.09} \pm 0.1$	...	$4.2^{+0.9}_{-1.2}$	$2.5^{+0.2}_{-0.2}$	...

For G59.7+0.1, however, our A_V distance agree with the 2MASS red giant extinction distance (only 10% systematic error).

4.3.5 Limitation of the method

As in Sect. 4.1, we used a simple dust screen model of the extinction along the line of sight. We modelled the extinction along the line of sight and infer the distance to one target according to a sudden jump in the extinction. Thus, the method rely on the distribution of star in the line of sight. With our method, we managed to derive reliable distances to several regions which agree with the literature. Unfortunately, however, certain regions have complicated dust distribution which violates our model assumption as *Gaia* cannot measure enough stars needed for our Bayesian method. Therefore, we are not able to report distance measurements to other sources with our A_G and A_V model. An example of a failure model we observed is shown in Fig. 4.22. As can be seen in the figure, there is no clear extinction breakpoint in the figure presented in the top left panel and bottom right panel. For the two other target, the number of on-cloud stars are not enough to produce the Bayesian inference.

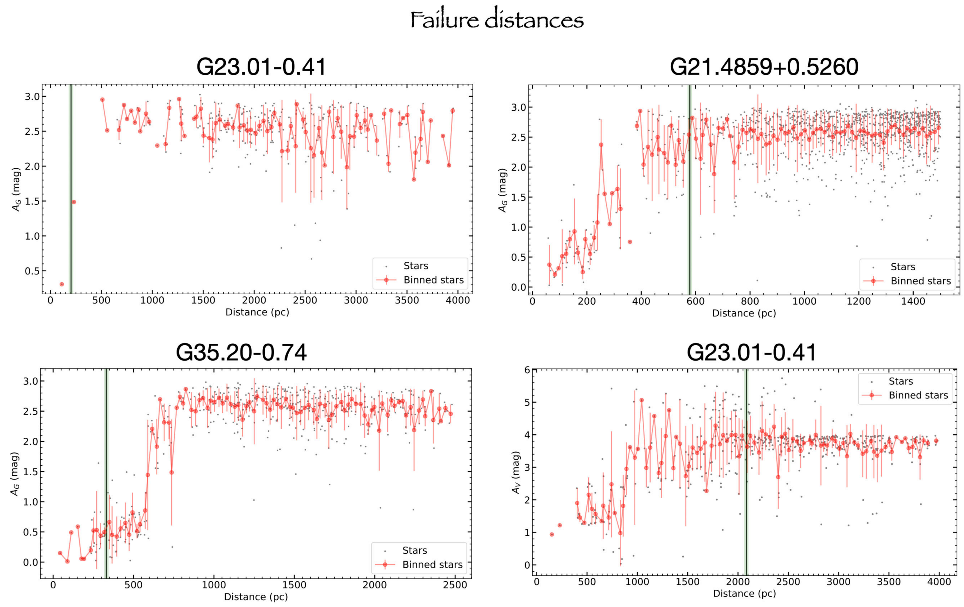

Figure 4.22: Examples of failure in our method. Top panels and bottom left panel represent the distance estimate towards G23.01-0.41, G21.4859+0.5260, and G35.20-0.74 using A_G model. Bottom right panel shows the distance of G23.01-0.41 using A_V model. The failed distance estimates is due to the weak distribution of off-cloud and on-cloud stars. There is no clear jump in the extinction along the line of sight, and the model failed to provide distance measurement of the target under study. For G35.20-0.74, we observed distance of the small component located in foreground of the target of interest.

Chapter 5

Conclusions and Future Work

The purpose of this thesis has been to derive distances to embedded stars using the second release of Gaia data (Gaia DR2). To do this, we started by drawing a box region centred on our object of interest and classified the stars along the line of sight into two categories: i) Off-cloud stars, which are stars located around the molecular cloud and have low extinction ii) On-cloud stars are stars that can be observed in the boundaries of the densest part of the molecular cloud and have higher extinction. Secondly, we built a Bayesian model of the A_G extinction and A_V extinction towards the region of interest to detect the jump point in the extinction from Off-cloud stars to On-cloud stars. As those extinction measurements have their corresponding distances, the distance to our region of interest is subsequently deduced by the Markov chain Monte Carlo (MCMC) technique with respect to the breakpoint in the extinction. The Bayesian model that we used in this work is fully described in Chapter 3.

Chapter 4 has presented and discussed our findings on the distances to different regions. We targeted Young Stellar Objects (YSOs) from the Leeds RMS survey, an additional three sources from the literature, and the sub-structures of Cygnus X. Overall, our distance measurements are consistent with the literature. If we suppose the maser parallax is the true distance, our distances show a small systematic uncertainty of less than 5% for objects that are associated with a molecular cloud with an average extinction (e.g . G108.929+02.59

and G065.3169-02.714). However, for a target associated with a complex dark cloud such as G5.89-0.39 and G59.7+0.1, the systematic uncertainty raised about ∼20%.

The results of this work suggest that the distance to an object associated with a dark cloud cannot be well reported with our method as Gaia stars are only sensitive to moderate-sized extinction. Consequently, our measurements with both the A_G and A_V model are lower compared to the extinction distances of Foster *et al.* (2012). Our models only determined the nearby distances along the line of sight either there is a presence of multiple clouds or several distances depending on the structure of the cloud. However, the use of A_V and set of distances provided by Anders *et al.* (2019) improved the distances to all regions selected. Compared to the kinematic distances of Foster *et al.* (2012), our distances are much lower. This large discrepancy may be the result of the presence of multiple clouds along the line of sight.

Another significant result of this thesis is the finding of the similarity between the distances to the components of Cygnus X. We found that DR20, DR21, DR22, DR23, and W75N are all located at ∼1.0 kpc according to A_G model and ∼1.6 kpc according to the A_V model. This is an excellent agreement with the maser parallax distances given by Rygl *et al.* (2012), who found an average individual distance of 1.4 kpc to the Cygnus X.

Interestingly, with the A_V extinction for only stars brighter than G=18 mag provided by Anders *et al.* (2019) that we used, we managed to detect reliable distances to several regions we arbitrary chose. The results using a similar methodology as we used in this work will absolutely bring more knowledge when more parallax data is available in the future. The case when our extinction model fails to infer distances towards a region with complex extinction distribution is also discussed in Sect. 4.3.5.

5.1 Future Work

As illustrated by previous studies such as Zucker *et al.* (2018) and Yan *et al.* (2019b), a precise classification of stars is essential for molecular cloud distance calculation. They used velocity slices of a CO spectral cube to trace the presence of the molecular cloud, and they classify the stars along the line of sight according to the CO emission. Although we have shown reliable distance measurements with our method (described in Chapter 4), it would be useful in the future to use CO velocity slices or a similar method before proceeding to the distance inference. This has to be taken into account for the study of large regions as it removes a large number of unnecessarily background stars.

As seen in Sect. 4, distances to certain regions cannot be reported with our models due to the insufficiency of Gaia stars within the region. In the future, as Gaia future data release is expected to provide many stars within those regions, suggest that combining Gaia DR3 astrometry with an ultra-deep infrared survey such as the UKIRT Infrared Deep Sky Survey (UKIDSS) will improve distance measurements to embedded stars. A similar work as which are done by Foster *et al.* (2012) using real astrometric data from the Gaia DR3 rather than using the Galactic model to derive the distances is also recommended.

Besides, we have seen in this study that only a few other approaches can be used to derive distances to embedded objects. Since they do not have visible data, it is impossible to carry out a single distance measurement to those sources. It is known that the kinematic method suffers from its large errors and the kinematic distance ambiguity (KDA) (Rice *et al.*, 2016), while the extinction distance methods are dependent on the Galactic model. In addition, we have introduced that the maser parallax distances is very accurate but cannot be applied to gauge distances to embedded objects in a region that lack maser. This accuracy on the parallax will be improved with the upcoming of the large project Square Kilometre Array (SKA) and can pin down distance measurements to a substantial part of the galactic plane including the central parts of the Milky Way. In the future, we recommend the use of maser parallax information to derive distances to complex dark clouds. The SKA is an international collaboration planned for full operation in 2030. When completed, it will be 50 times more sensitive than any current system and can measure astrometry with an accuracy better than ± 1 μas (Reid & Honma, 2014).

Appendix A

Appendix

A.1 Examples of ADQL query used in this work

```
*+*+*+*+*+*+*+*+*+*+*+*+*+*+
ADQL query for GAIA stars
*+*+*+*+*+*+*+*+*+*+*+*+*+*+
SELECT TOP 10000 dist.source_id, g.ra, g.dec, dist.r_est,
dist.r_len,g.parallax, g.parallax_error,g.a_g_val,
g.a_g_percentile_lower, g.a_g_percentile_upper, rw.ruwe
FROM external.gaiadr2_geometric_distance as dist,
gaiadr2.gaia_source AS g, gaiadr2.ruwe AS rw
WHERE dist.source_id = g.source_id
AND g.source_id = tbest.source_id
AND g.source_id = rw.source_id
AND g.parallax IS NOT NULL
AND 1./g.parallax <3
AND 33.4035 < g.l AND g.l < 35.4035
AND  -0.772 < g.b AND g.b < 1.228
AND g.a_g_val > 0
AND rw.ruwe < 1.4
*+*+*+*+*+*+*+*+*+*+*+*+*+*+*+*
ADQL query for STARHORSE stars
*+*+*+*+*+*+*+*+*+*+*+*+*+*+*+*
SELECT TOP 10000 g.ra, g.dec, s.dist16, s.dist50, s.dist84,
s.AV16, s.AV50, s.AV84, s.AG50
FROM gdr2.gaia_source AS g, gdr2_contrib.starhorse AS s
WHERE g.source_id = s.source_id
```

```
AND 33.4035 < g.l AND g.l < 35.4035
AND  -0.772 < g.b AND g.b < 1.228
AND s.AV50 > 0
AND s.dist50 > 0 AND s.dist50 < 3 AND s.ruwe < 1.4
```

A.2 Corner plot of the Orion A obtained in Sect. 4.1

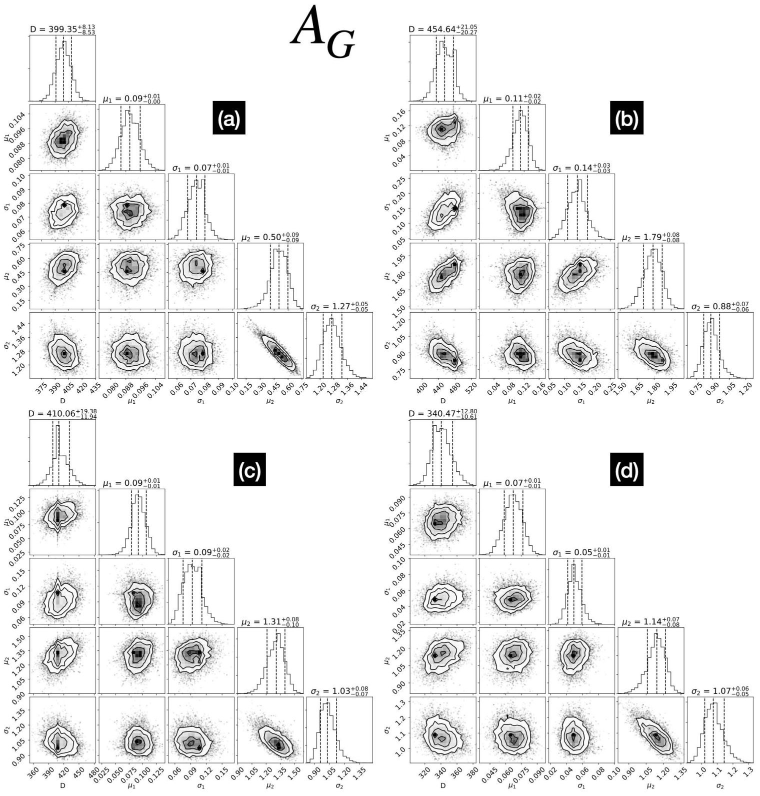

Figure A.1: Corner plot of the four boxes of Orion A using the A_G model.

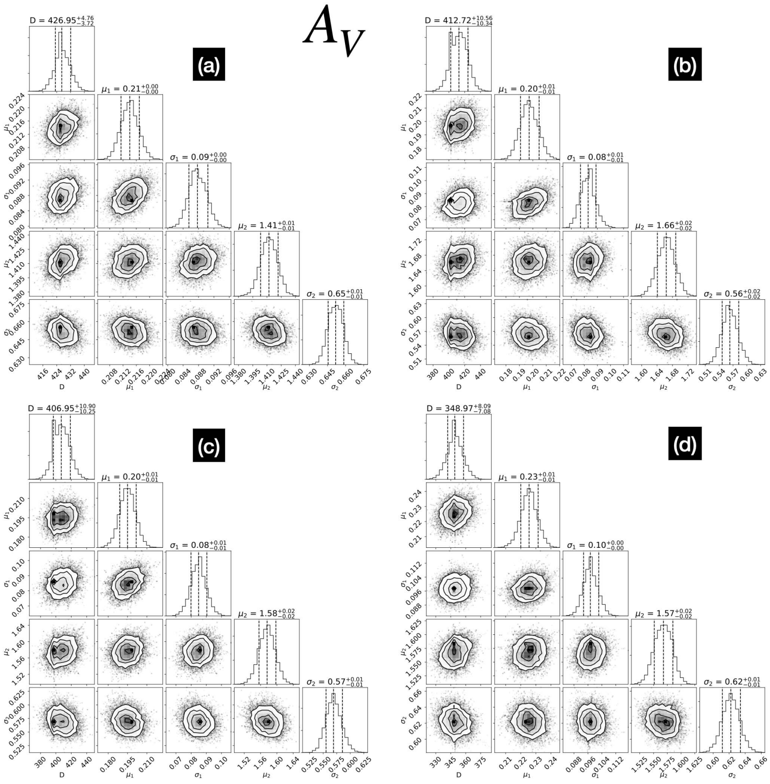

Figure A.2: Corner plot of the four boxes of Orion A using the A_V model.

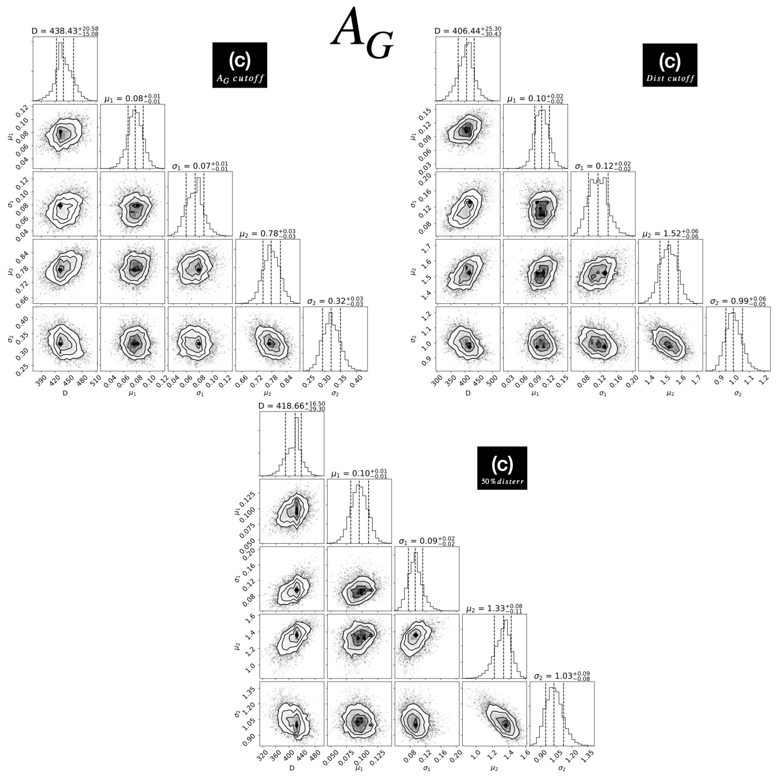

Figure A.3: Corner plot of the resulting cutoff applied to the box (c) of Orion A using the A_G model.

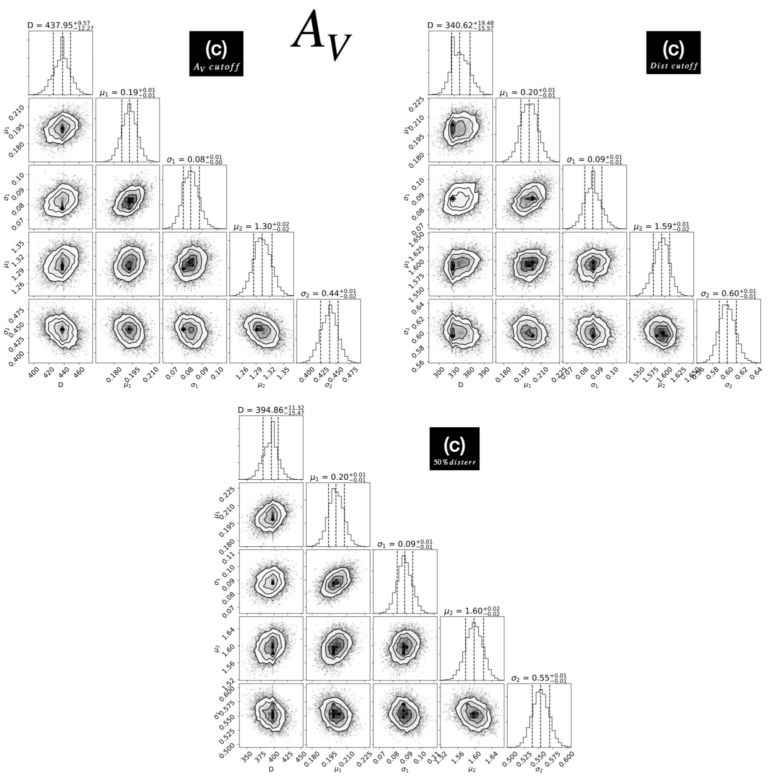

Figure A.4: Same figure as Fig. A.3 but using the A_V model.

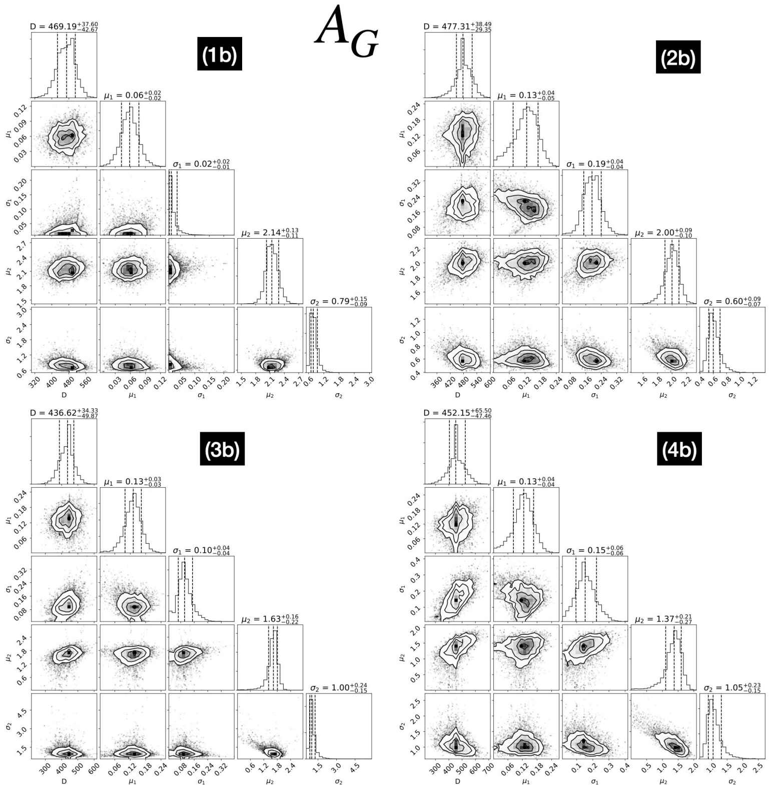

Figure A.5: Corner plot of the box (b) sub-regions of Orion A with A_G.

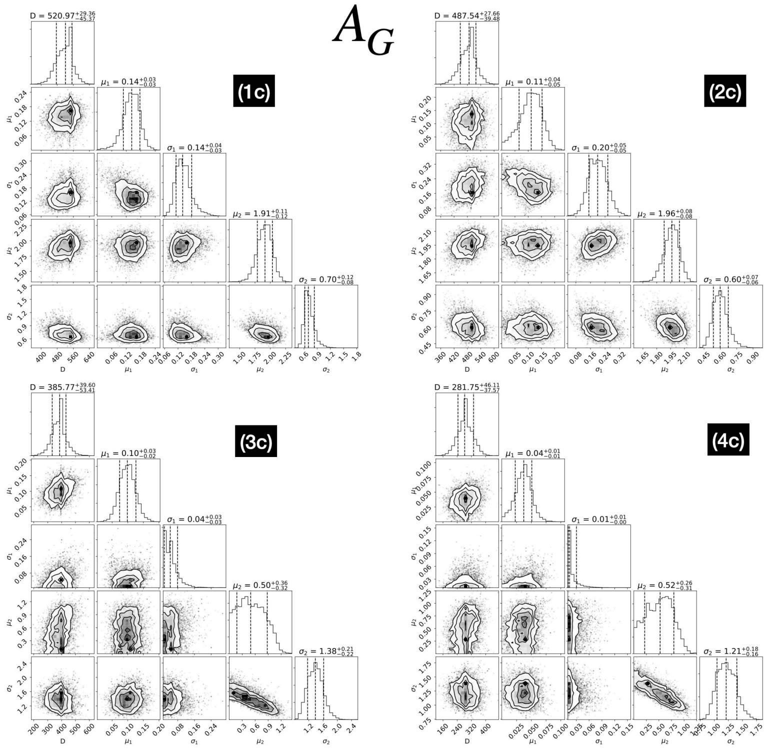

Figure A.6: Corner plot of the box (c) sub-regions of Orion A with A_G.

A.3 Summary of distances obtained from this work

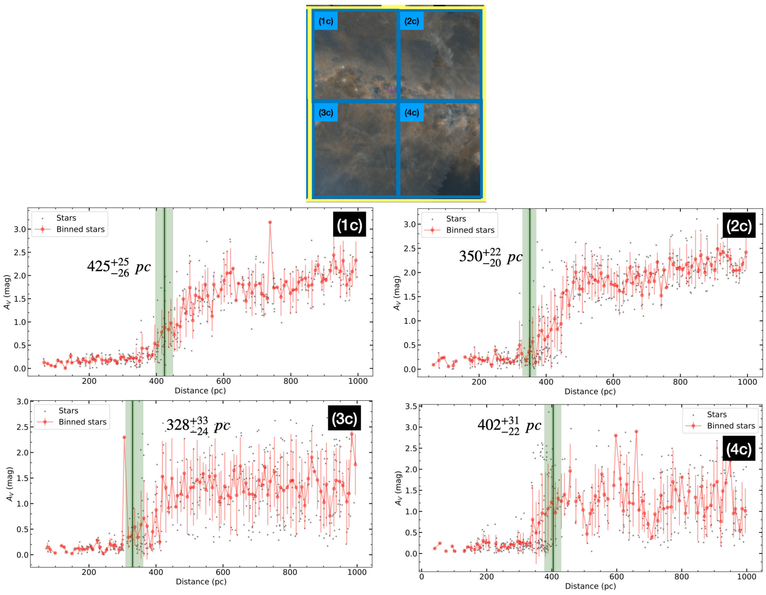

Figure A.7: Distance to the box (c) sub-regions of Orion A using A_V model.

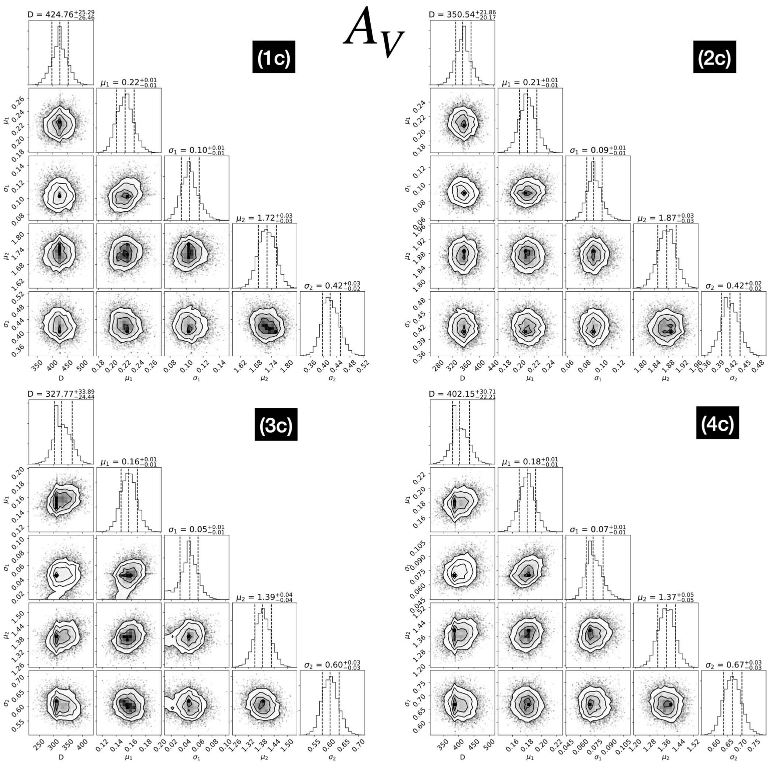

Figure A.8: Corner plot of the box (c) sub-regions of Orion A with A_V.

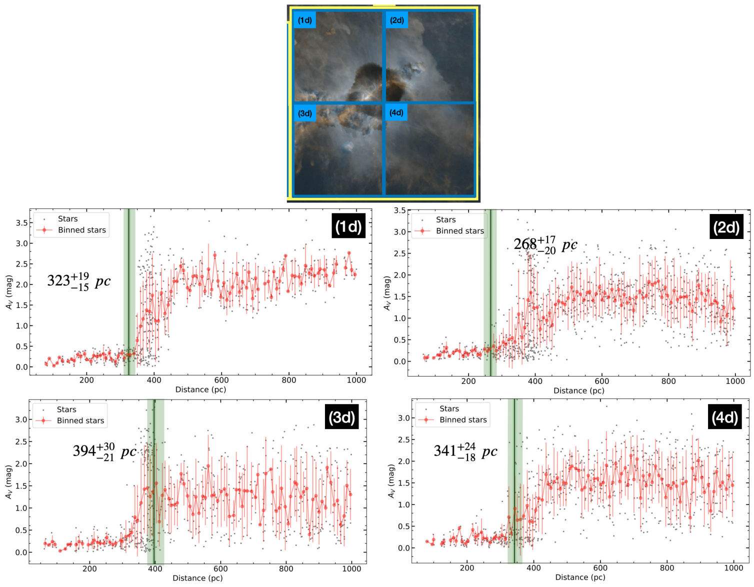

Figure A.9: Distance to the box (b) sub-regions of Orion A using A_V model.

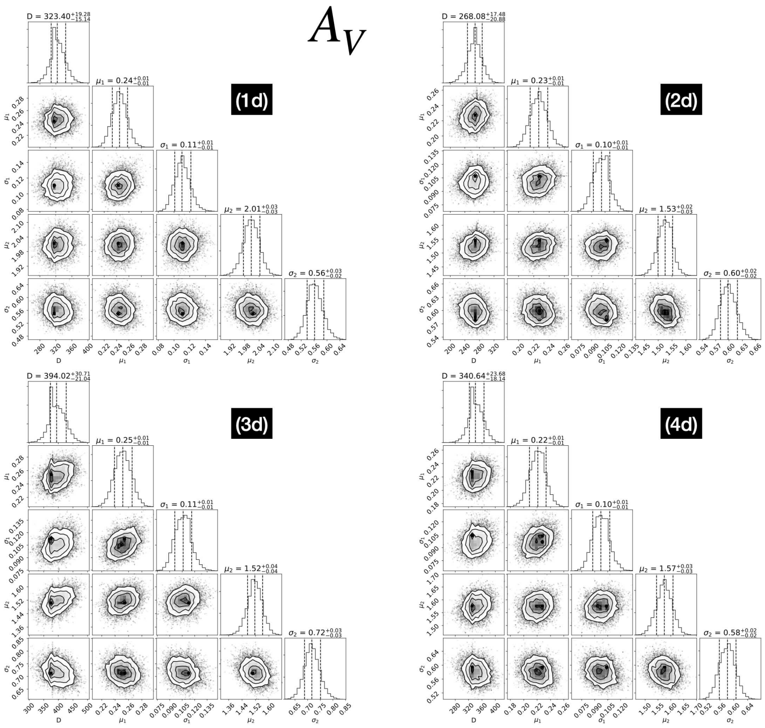

Figure A.10: Corner plot of the box (d) sub-regions of Orion A with A_V.

Table A.1: All distances obtained from this work.

(1)	(2)	(3)	(4)	(5)
Source	l	b	Distances with A_G	Distances with A_V
			(kpc)	(kpc)
G010.384+2.213	10.3844	2.2128	$1.50^{+0.3}_{-0.07} \pm 0.07$	$1.58^{+0.06}_{-0.05} \pm 0.08$
G108.929+02.595	108.9288	2.5954	$0.73^{+0.03}_{-0.06} \pm 0.03$	$0.87^{+0.05}_{-0.06} \pm 0.04$
G126.714-00.822	126.714	-00.822	$1.10^{+0.03}_{-0.04} \pm 0.05$	$0.93^{+0.02}_{-0.02} \pm 0.04$
G014.4886+00.0219B	14.4886	0.0219	$2.02^{+0.12}_{-0.11} \pm 0.10$	$1.78^{+0.08}_{-0.13} \pm 0.09$
G065.3169-02.7141	65.3169	-02.7141	$1.23^{+0.04}_{-0.05} \pm 0.06$	$0.99^{+0.03}_{-0.03} \pm 0.05$
G034.4035+00.2282A	34.4035	0.228	$0.48^{+0.02}_{-0.01} \pm 0.02$	$0.83^{+0.01}_{-0.01} \pm 0.04$
G5.89-0.39	5.8842	-0.3924	$1.27^{+0.05}_{-0.08} \pm 0.06$	$1.25^{+0.06}_{-0.07} \pm 0.06$
G35.20-0.74	35.1970	-0.7431	–	$1.30^{+0.08}_{-0.09} \pm 0.06$
G59.7+0.1	59.7828	0.0647	$2.24^{+0.10}_{-0.18} \pm 0.1$	$2.78^{+0.12}_{-0.09} \pm 0.1$
W 75N	81.867	+0.779	$1.02^{+0.07}_{-0.04} \pm 0.05$	$1.73^{+0.06}_{-0.05} \pm 0.08$
DR 21	81.680	+0.537	$1.19^{+0.03}_{-0.03} \pm 0.06$	$1.72^{+0.03}_{-0.03} \pm 0.08$
DR 20	80.872	+0.411	$1.21^{+0.3}_{-0.02} \pm 0.06$	$1.59^{+0.02}_{-0.02} \pm 0.08$
DR 22	80.904	-0.117	$1.26^{+0.03}_{-0.02} \pm 0.06$	$1.60^{+0.03}_{-0.03} \pm 0.08$
DR 23	81.543	+0.016	$1.12^{+0.04}_{-0.05} \pm 0.06$	–

NOTE: (1), (2), and (3) represent the source name and the Galactic coordinates (l,b), respectively. The distances D obtained from the two models A_G and A_V are given in (4) and (5), respectively. Our distances contain the two categories of uncertainties: the statistical uncertainty as 16th and 84th percentiles (in the first term error), and the 5% systematic uncertainty (shown in the second term error).

References

Alves, J., Lada, C.J., Lada, E.A., Kenyon, S.J. & Phelps, R. (1998). Dust extinction and molecular cloud structure: L977. *the Astrophysical Journal*, **506**, 292. 24

Anders, F., Khalatyan, A., Chiappini, C., Queiroz, A., Santiago, B., Jordi, C., Girardi, L., Brown, A., Matijevič, G., Monari, G. *et al.* (2019). Photo-astrometric distances, extinctions, and astrophysical parameters for gaia dr2 stars brighter than g= 18. *Astronomy & Astrophysics*, **628**, A94. iv, xi, xiv, 27, 29, 31, 38, 40, 41, 90

Anderson, L., Hogg, D.W., Leistedt, B., Price-Whelan, A.M. & Bovy, J. (2018). Improving gaia parallax precision with a data-driven model of stars. *The Astronomical Journal*, **156**, 145. 30

Andrae, R., Fouesneau, M., Creevey, O., Ordenovic, C., Mary, N., Burlacu, A., Chaoul, L., Jean-Antoine-Piccolo, A., Kordopatis, G., Korn, A. *et al.* (2018). Gaia data release 2-first stellar parameters from apsis. *Astronomy & Astrophysics*, **616**, A8. 25, 26, 37, 38, 41, 50, 51, 52

André, P. (1994). Observations of protostars and protostellar stages. *The Cold Universe*, 179. 20

Astraatmadja, T.L. & Bailer-Jones, C.A. (2016). Estimating distances from parallaxes. iii. distances of two million stars in the gaia dr1 catalogue. *The Astrophysical Journal*, **833**, 119. 4, 42

Bailer-Jones, C., Rybizki, J., Fouesneau, M., Mantelet, G. & Andrae, R. (2018). Estimating distance from parallaxes. iv. distances to 1.33 billion stars in gaia data release 2. *The Astronomical Journal*, **156**, 58. 4, 27, 31, 41, 42, 43, 44

Bailer-Jones, C.A. (2015). Estimating distances from parallaxes. *Publications of the Astronomical Society of the Pacific*, **127**, 994. 3, 42

Bailer-Jones, C.A.L., Andrae, R., Arcay, B., Astraatmadja, T., Bellas-Velidis, I., Berihuete, A., Bijaoui, A., Carrión, C., Dafonte, C., Damerdji, Y. & et al. (2013). Thegaiaastrophysical parameters inference system (apsis). *Astronomy Astrophysics*, **559**, A74. 24

Balser, D.S., Rood, R.T., Bania, T. & Anderson, L. (2011). H ii region metallicity distribution in the milky way disk. *The Astrophysical Journal*, **738**, 27. 13

Beerer, I., Koenig, X., Hora, J., Gutermuth, R., Bontemps, S., Megeath, S., Schneider, N., Motte, F., Carey, S., Simon, R. *et al.* (2010). A spitzer view of star formation in the cygnus x north complex. *The Astrophysical Journal*, **720**, 679. 73

Beltrán, M.T., Cesaroni, R., Codella, C., Testi, L., Furuya, R.S. & Olmi, L. (2006). Infall of gas as the formation mechanism of stars up to 20 times more massive than the sun. *Nature*, **443**, 427–429. 17

Beskin, V., Beskin, V., Henri, G., Menard, F., Dalibard, J. & Pelletier, G. (2003). *Accretion Disks, Jets and High-Energy Phenomena in Astrophysics: Les Houches Session LXXVIII, July 29-August 23, 2002*, vol. 78. Springer Science & Business Media. 21

Bessel, F.W. (1838). On the parallax of 61 cygni. *Monthly Notices of the Royal Astronomical Society*, **4**, 152–161. 3

Billington, S., Urquhart, J., Figura, C., Eden, D. & Moore, T. (2019). The rms survey: Ammonia mapping of the environment of young massive stellar objects–ii. *Monthly Notices of the Royal Astronomical Society*, **483**, 3146–3167. 16

Binney, J., Burnett, B., Kordopatis, G., McMillan, P.J., Sharma, S., Zwitter, T., Bienayme, O., Bland-Hawthorn, J., Steinmetz, M., Gilmore, G. et al. (2013). New distances to rave stars. *Monthly Notices of the Royal Astronomical Society*, **437**, 351–370. 30

Breddels, M.A., Smith, M.C., Helmi, A., Bienaymé, O., Binney, J., Bland-Hawthorn, J., Boeche, C., Burnett, B., Campbell, R., Freeman, K.C. et al. (2010). Distance determination for rave stars using stellar models. *Astronomy & Astrophysics*, **511**, A90. 30

Cardelli, J.A., Clayton, G.C. & Mathis, J.S. (1989). The relationship between infrared, optical, and ultraviolet extinction. *The Astrophysical Journal*, **345**, 245–256. 22, 25

Clemens, D.P. (1985). Massachusetts-stony brook galactic plane co survey-the galactic disk rotation curve. *The Astrophysical Journal*, **295**, 422–428. 84

Collaboration, G. et al. (2016). Description of the gaia mission (spacecraft, instruments, survey and measurement principles, and operations). *Gaia Collaboration et al.(2016a): Summary description of Gaia DR1*. 9

Cutri, R.e. et al. (2014). Vizier online data catalog: Allwise data release (cutri+ 2013). *VizieR Online Data Catalog*, **2328**. 40

De Wit, W., Testi, L., Palla, F. & Zinnecker, H. (2005). The origin of massive o-type field stars-ii. field o stars as runaways. *Astronomy & Astrophysics*, **437**, 247–255. 17

Dobashi, K., Uehara, H., Kandori, R., Sakurai, T., Kaiden, M., Umemoto, T. & Sato, F. (2005). Atlas and catalog of dark clouds based on digitized sky survey i. *Publications of the Astronomical Society of Japan*, **57**, S1–S386. 24

Draine, B.T. (2003). Interstellar dust grains. *Annual Review of Astronomy and Astrophysics*, **41**, 241–289. 22

Evans, D., Riello, M., De Angeli, F., Carrasco, J., Montegriffo, P., Fabricius, C., Jordi, C., Palaversa, L., Diener, C., Busso, G. *et al.* (2018). Gaia data release 2-photometric content and validation. *Astronomy & Astrophysics*, **616**, A4. 27

Ferriere, K.M. (2001). The interstellar environment of our galaxy. *Reviews of Modern Physics*, **73**, 1031. 12, 14

Foreman-Mackey, D. (2016). corner.py: Scatterplot matrices in python. *The Journal of Open Source Software*, **1**, 24. 35

Foreman-Mackey, D., Hogg, D.W., Lang, D. & Goodman, J. (2013). emcee: the mcmc hammer. *Publications of the Astronomical Society of the Pacific*, **125**, 306. 34, 63

Foster, J.B., Stead, J.J., Benjamin, R.A., Hoare, M.G. & Jackson, J.M. (2012). Distances to dark clouds: Comparing extinction distances to maser parallax distances. *The Astrophysical Journal*, **751**, 157. iv, xiv, 24, 84, 86, 87, 90, 91

Gaia Collaboration, A., Brown, Vallenari, A., Prusti, T., De Bruijne, J., Babusiaux, C., Bailer-Jones, C., Biermann, M., Evans, D.W., Eyer, L., Jansen, F. *et al.* (2018). Gaia data release 2-summary of the contents and survey properties. *Astronomy & astrophysics*, **616**, A1. 38

Gaia Collaboration, T., Prusti, De Bruijne, J., Brown, A.G., Vallenari, A., Babusiaux, C., Bailer-Jones, C., Bastian, U., Biermann, M., Evans, D.W., Eyer, L. et al. (2016b). The gaia mission. *Astronomy & Astrophysics*, **595**, A1. 4, 38

Gelman, A., Simpson, D. & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, **19**, 555. 33

Geurts, P., Ernst, D. & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, **63**, 3–42. 25

Gilmore, G. (2018). Gaia: 3-dimensional census of the milky way galaxy. *Contemporary Physics*, **59**, 155–173. xi, 8

Goldbaum, N.J., Krumholz, M.R. & Forbes, J.C. (2016). Mass transport and turbulence in gravitationally unstable disk galaxies. ii. the effects of star formation feedback. *The Astrophysical Journal*, **827**, 28. 19

Goodman, J. & Weare, J. (2010). Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science*, **5**, 65–80. 34

Grossschedl, J.E., Alves, J., Meingast, S., Ackerl, C., Ascenso, J., Bouy, H., Burkert, A., Forbrich, J., Fürnkranz, V., Goodman, A. et al. (2018). 3d shape of orion a from gaia dr2. *Astronomy & Astrophysics*, **619**, A106. 54, 56

Guzmán, A.E., Garay, G., Rodríguez, L.F., Contreras, Y., Dougados, C. & Cabrit, S. (2016). A protostellar jet emanating from a hypercompact h ii region. *The Astrophysical Journal*, **826**, 208. 16

Haffner, L., Dettmar, R.J., Beckman, J., Wood, K., Slavin, J., Giammanco, C., Madsen, G., Zurita, A. & Reynolds, R. (2009). The warm ionized medium in spiral galaxies. *Reviews of Modern Physics*, **81**, 969. 13

Hampton, E., Rowell, G., Hofmann, W., Horns, D., Uchiyama, Y. & Wagner, S. (2016). Chandra observations of the hii complex g5. 89-0.39 and tev gamma-ray source hessj1800-240b. *Journal of High Energy Astrophysics*, **11**, 1–19. 84

Hartquist, T.W. (2011). Star formation: a beginner's guide. *Astronomy Geophysics*, **52**, 5.37–5.37. 17

Hawkins, K., Leistedt, B., Bovy, J. & Hogg, D.W. (2017). Red clump stars and gaia: calibration of the standard candle using a hierarchical probabilistic model. *Monthly Notices of the Royal Astronomical Society*, **471**, 722–729. 30

Heiles, C. & Troland, T. (2003). The millennium arecibo 21 centimeter absorption-line survey. ii. properties of the warm and cold neutral media. *The Astrophysical Journal*, **586**, 1067. 13

Hogg, D.W. (2018). A likelihood function for the gaia data. *arXiv preprint arXiv:1804.07766*. 31

Isella, A. (2006). *Interferometric observations of pre-main sequence disks*. Ph.D. thesis, Ph. D. thesis, Universitá degli Studi di Milano, Facoltaa di Scienze .... xi, 21

Kass, R.E. & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370. 33

Kennicutt Jr, R.C. & Evans, N.J. (2012). Star formation in the milky way and nearby galaxies. *Annual Review of Astronomy and Astrophysics*, **50**, 531–608. 14

Krause, M.G., Offner, S.S., Charbonnel, C., Gieles, M., Klessen, R.S., Vazquez-Semadeni, E., Ballesteros-Paredes, J., Girichidis, P., Diederik Kruijssen, J., Ward, J.L. *et al.* (2020). The physics of star cluster formation and evolution. *Space Science Reviews*, **216**, 1–46. 18, 19

Krumholz, M.R. & Bonnell, I.A. (2007). Models for the formation of massive stars. 18

Krumholz, M.R., McKee, C.F. & Bland-Hawthorn, J. (2019). Star clusters across cosmic time. *Annual Review of Astronomy and Astrophysics*, **57**, 227–303. 18

Kurucz, R.L. (1993). Atlas9 stellar atmosphere programs and 2km/s grid. *Kurucz CD-Rom*, **13**. 29

Lada, C.J. (1987). Star formation: from ob associations to protostars. In *Symposium-International Astronomical Union*, vol. 115, 1–18, Cambridge University Press. 20

Lada, C.J. & Lada, E.A. (2003). Embedded clusters in molecular clouds. *Annual Review of Astronomy and Astrophysics*, **41**, 57–115. 19

Lada, C.J., Lada, E.A., Clemens, D.P. & Bally, J. (1994). Dust extinction and molecular gas in the dark cloud ic 5146. *The Astrophysical Journal*, **429**, 694–709. 24

Larson, R.B. (2003). The physics of star formation. *Reports on Progress in Physics*, **66**, 1651. 14

Leistedt, B. & Hogg, D.W. (2017). Hierarchical probabilistic inference of the color–magnitude diagram and shrinkage of stellar distance uncertainties. *The Astronomical Journal*, **154**, 222. 30

Li, A. & Mann, I. (2012). Nanodust in the interstellar medium in comparison to the solar system. In *Nanodust in the Solar System: Discoveries and Interpretations*, 5–30, Springer. 23

Lindegren, L., Lammers, U., Hobbs, D., O'Mullane, W., Bastian, U. & Hernández, J. (2012). The astrometric core solution for the gaia mission-overview of models, algorithms, and software implementation. *Astronomy & Astrophysics*, **538**, A78. 9

Lindegren, L., Hernández, J., Bombrun, A., Klioner, S., Bastian, U., Ramos-Lerate, M., De Torres, A., Steidelmüller, H., Stephenson, C., Hobbs, D. et al. (2018a). Gaia data release 2-the astrometric solution. *Astronomy & Astrophysics*, **616**, A2. 9, 11

Lindegren, L. et al. (2018b). Re-normalising the astrometric chi-square in gaia dr2. *Gaia Technical Note: GAIA-C3-TN-LU-LL-124-0*. 11, 40, 42

Lombardi, M. (2009). Nicest, a near-infrared color excess method tailored to small-scale structures. *Astronomy & Astrophysics*, **493**, 735–745. 24

Lombardi, M. & Alves, J. (2001). Mapping the interstellar dust with near-infrared observations: An optimized multi-band technique. *Astronomy & Astrophysics*, **377**, 1023–1034. 24, 46

Lombardi, M., Alves, J. & Lada, C.J. (2006). 2mass wide field extinction maps-i. the pipe nebula. *Astronomy & Astrophysics*, **454**, 781–796. 24

Lombardi, M., Lada, C.J. & Alves, J. (2008). Hipparcos distance estimates of the ophiuchus and the lupus cloud complexes. *Astronomy & Astrophysics*, **480**, 785–792. 46, 47

Loredo, T.J. (2013). Bayesian astrostatistics: a backward look to the future. In *Astrostatistical challenges for the new astronomy*, 15–40, Springer. 31

Lumsden, S.L., Hoare, M.G., Urquhart, J.S., Oudmaijer, R.D., Davies, B., Mottram, J.C., Cooper, H.D.B. & Moore, T.J.T. (2013). The red msx source survey: The massive young stellar population of our galaxy. *The Astrophysical Journal Supplement Series*, **208**, 11. 36, 54, 66

Luri, X., Brown, A., Sarro, L., Arenou, F., Bailer-Jones, C., Castro-Ginard, A., De Bruijne, J., Prusti, T., Babusiaux, C. & Delgado, H. (2018). Gaia data release 2-using gaia parallaxes. *Astronomy & Astrophysics*, **616**, A9. 42, 63

Machida, M.N. (2017). Protostellar jets and outflows in low-mass star formation. *arXiv preprint arXiv:1711.00384*. 17

Marshall, D.J., Robin, A., Reylé, C., Schultheis, M. & Picaud, S. (2006). Modelling the galactic interstellar extinction distribution in three dimensions. *Astronomy & Astrophysics*, **453**, 635–651. 86

Martí, B.L., Merín, B., Giordano, F., Baines, D., Racero, E., Salgado, J., Sarmiento, M.H., Gutiérrez, R., de Teodoro, P., González, J. *et al.* (2016). Esasky: The whole of space astronomy at your fingertips. *arXiv preprint arXiv:1610.09826*. 55, 67

Marton, G., Schulz, B., Altieri, B., Calzoletti, L., Kiss, C., Lim, T., Lu, N., Paladini, R., Papageorgiou, A., Pearson, C. *et al.* (2015). The herschel point source catalogue. *arXiv preprint arXiv:1510.08325*. 75, 81

Maund, J.R., Crowther, P.A., Janka, H.T. & Langer, N. (2017). Bridging the gap: from massive stars to supernovae. 17

McKee, C.F. & Ostriker, J.P. (1977). A theory of the interstellar medium-three components regulated by supernova explosions in an inhomogeneous substrate. *The Astrophysical Journal*, **218**, 148–169. xiv, 12

McMillan, P.J. (2018). Simple distance estimates for gaia dr2 stars with radial velocities. *arXiv preprint arXiv:1806.00426*. 41

Menten, K., Reid, M., Forbrich, J. & Brunthaler, A. (2007). The distance to the orion nebula. *Astronomy & Astrophysics*, **474**, 515–520. 56

Mignard, F. (2019). The gaia mission and significance. *arXiv preprint arXiv:1906.09022*. xiv, 39

Mints, A. & Hekker, S. (2017). A unified tool to estimate distances, ages, and masses (unidam) from spectrophotometric data. *Astronomy & Astrophysics*, **604**, A108. 41

Motogi, K., Sorai, K., Habe, A., Honma, M., Kobayashi, H. & Sato, K. (2011). New distance and revised natures of high-mass star formation in g5. 89–0.39. *Publications of the Astronomical Society of Japan*, **63**, 31–44. 73, 85

O'Malley, E.M., Gilligan, C. & Chaboyer, B. (2017). Absolute ages and distances of 22 gcs using monte carlo main-sequence fitting. *The Astrophysical Journal*, **838**, 162. 19

O'Mullane, W., Lammers, U., Lindegren, L., Hernandez, J. & Hobbs, D. (2011). Implementing the gaia astrometric global iterative solution (agis) in java. *Experimental Astronomy*, **31**, 215. xi, 9, 10

Perryman, M. *et al.* (1997). The hipparcos and tycho catalogues. *star*, **2**, 2π1. 4, 46

Queiroz, A.B.d.A., Anders, F., Santiago, B.X., Chiappini, C., Steinmetz, M., Dal Ponte, M., Stassun, K.G., da Costa, L.N., Maia, M.A.G., Crestani, J. *et al.* (2018). Starhorse: a bayesian tool for determining stellar masses, ages, distances, and extinctions for field stars. *Monthly Notices of the Royal Astronomical Society*, **476**, 2556–2583. 27, 28, 41

Reid, M., Menten, K., Zheng, X., Brunthaler, A., Moscadelli, L., Xu, Y., Zhang, B., Sato, M., Honma, M., Hirota, T. *et al.* (2009). Trigonometric parallaxes of massive star-forming regions. vi. galactic structure, fundamental parameters, and noncircular motions. *The Astrophysical Journal*, **700**, 137. 84

Reid, M.J. & Honma, M. (2014). Microarcsecond radio astrometry. *Annual Review of Astronomy and Astrophysics*, **52**, 339–372. 85, 92

Reid, M.J., McClintock, J.E., Narayan, R., Gou, L., Remillard, R.A. & Orosz, J.A. (2011). The trigonometric parallax of cygnus x-1. *The Astrophysical Journal*, **742**, 83. 66

Rezaei Kh., S., Bailer-Jones, C.A.L., Soler, J.D. & Zari, E. (2020). Detailed 3d structure of orion a in dust with gaia dr2. *Astronomy Astrophysics*, **643**, A151. 54

Rice, T.S., Goodman, A.A., Bergin, E.A., Beaumont, C. & Dame, T.M. (2016). A uniform catalog of molecular clouds in the milky way. *The Astrophysical Journal*, **822**, 52. 91

Robin, A.C., Reylé, C., Derriere, S. & Picaud, S. (2003). A synthetic view on structure and evolution of the milky way. *Astronomy & Astrophysics*, **409**, 523–540. 86

Roman-Duval, J., Heyer, M., Brunt, C.M., Clark, P., Klessen, R. & Shetty, R. (2016). Distribution and mass of diffuse and dense co gas in the milky way. *The Astrophysical Journal*, **818**, 144. 15

Roman-Lopes, A. (2013). An o2 if*/wn6 star caught in the act in a compact h ii region in the starburst cluster ngc 3603. *Monthly Notices of the Royal Astronomical Society*, **433**, 712–718. 17

Rygl, K.L., Brunthaler, A., Sanna, A., Menten, K.M., Reid, M.J., van Langevelde, H.J., Honma, M., Torstensson, K.J. & Fujisawa, K. (2012). Parallaxes and proper motions of interstellar masers toward the cygnus x star-forming complex-i. membership of the cygnus x region. *Astronomy & Astrophysics*, **539**, A79. 85, 90

Sanders, J.L. & Das, P. (2018). Isochrone ages for 3 million stars with the second gaia data release. *Monthly Notices of the Royal Astronomical Society*, **481**, 4093–4110. 41

Santiago, B.X., Brauer, D.E., Anders, F., Chiappini, C., Queiroz, A.B., Girardi, L., Rocha-Pinto, H.J., Balbinot, E., Da Costa, L.N., Maia, M.A. *et al.* (2016). Spectro-photometric distances to stars: A general purpose bayesian approach. *Astronomy & Astrophysics*, **585**, A42. 30

Schilke, P. (2015). High-mass star formation. *EAS Publications Series*, **75**, 227–235. 18, 56

Schilke, P. (2017). (mostly) observational aspects of high-mass star formation. *arXiv preprint arXiv:1712.05281*. 17

Schlafly, E., Green, G., Finkbeiner, D., Rix, H.W., Bell, E., Burgett, W., Chambers, K., Draper, P., Hodapp, K., Kaiser, N. *et al.* (2014). A large catalog of accurate distances to molecular clouds from ps1 photometry. *The Astrophysical Journal*, **786**, 29. 30

Schlafly, E., Green, G., Finkbeiner, D., Rix, H.W., Burgett, W., Chambers, K., Draper, P., Kaiser, N., Martin, N., Metcalfe, N. *et al.* (2015). Three-dimensional dust mapping reveals that orion forms part of a large ring of dust. *The Astrophysical Journal*, **799**, 116. 56, 59

Schlafly, E., Meisner, A., Stutz, A., Kainulainen, J., Peek, J., Tchernyshyov, K., Rix, H.W., Finkbeiner, D., Covey, K., Green, G.

ET AL. (2016). The optical–infrared extinction curve and its variation in the milky way. *The Astrophysical Journal*, **821**, 78. 29

Schönrich, R., McMillan, P. & Eyer, L. (2019). Distances and parallax bias in gaia dr2. *Monthly Notices of the Royal Astronomical Society*, **487**, 3568–3580. 40

Scolnic, D., Casertano, S., Riess, A., Rest, A., Schlafly, E., Foley, R.J., Finkbeiner, D., Tang, C., Burgett, W., Chambers, K. et al. (2015). Supercal: Cross-calibration of multiple photometric systems to improve cosmological measurements with type ia supernovae. *The Astrophysical Journal*, **815**, 117. 40

Sharma, S. (2017). Markov chain monte carlo methods for bayesian data analysis in astronomy. *Annual Review of Astronomy and Astrophysics*, **55**, 213–259. 30

Shu, F.H. (1977). Self-similar collapse of isothermal spheres and star formation. *The Astrophysical Journal*, **214**, 488–497. 15

Skrutskie, M., Cutri, R., Stiening, R., Weinberg, M., Schneider, S., Carpenter, J., Beichman, C., Capps, R., Chester, T., Elias, J. et al. (2006). The two micron all sky survey (2mass). *The Astronomical Journal*, **131**, 1163. 24, 40, 67, 75, 81

Stanway, E.R. (2016). What can distant galaxies teach us about massive stars? *Proceedings of the International Astronomical Union*, **12**, 305–312. 17

Stead, J. & Hoare, M. (2010). Molecular cloud distance determination from deep nir survey extinction measurements. *Monthly Notices of the Royal Astronomical Society*, **407**, 923–936. 24

Urquhart, J., Figura, C., Moore, T., Csengeri, T., Lumsden, S., Pillai, T., Thompson, M., Eden, D. & Morgan, L. (2015). The rms survey: ammonia mapping of the environment of massive young stellar objects. *Monthly Notices of the Royal Astronomical Society*, **452**, 4029–4053. 16

Van Dishoeck, E.F., Black, J.H. *et al.* (1988). The photodissociation and chemistry of interstellar co. *Astrophysical Journal*, **334**, 771–802. 12

Ward-Thompson, D. & Whitworth, A.P. (2011). *An introduction to star formation*. Cambridge University Press. 16

Wenger, T.V., Balser, D.S., Anderson, L. & Bania, T. (2018). Kinematic distances: A monte carlo method. *The Astrophysical Journal*, **856**, 52. 36

Xu, Y., Reid, M., Menten, K., Zheng, X., Brunthaler, A. & Moscadelli, L. (2007). The distance to g59. 7+ 0.1 and w3oh. *Proceedings of the International Astronomical Union*, **3**, 214–216. 73, 84

Xu, Y., Reid, M., Menten, K., Brunthaler, A., Zheng, X. & Moscadelli, L. (2009). Trigonometric parallaxes of massive star-forming regions: Iii. g59. 7+ 0.1 and w 51 irs2. *The Astrophysical Journal*, **693**, 413. 85

Yan, Q.Z., Yang, J., Sun, Y., Su, Y. & Xu, Y. (2019a). Molecular cloud distances based on the mwisp co survey and gaia dr2. *The Astrophysical Journal*, **885**, 19. 31, 50

Yan, Q.Z., Zhang, B., Xu, Y., Guo, S., Macquart, J.P., Tang, Z.H. & Walsh, A.J. (2019b). Distances to molecular clouds at high galactic latitudes based on gaia dr2. *Astronomy & Astrophysics*, **624**, A6. 31, 38, 46, 47, 49, 91

York, D.G., Adelman, J., Anderson Jr, J.E., Anderson, S.F., Annis, J., Bahcall, N.A., Bakken, J., Barkhouser, R., Bastian, S., Berman, E. *et al.* (2000). The sloan digital sky survey: Technical summary. *The Astronomical Journal*, **120**, 1579. 67, 75, 81

ZHANG, B., ZHENG, X., REID, M., MENTEN, K., XU, Y., MOSCADELLI, L. & BRUNTHALER, A. (2009). Trigonometric parallaxes of massive star-forming regions. iv. g35. 20–0.74 and g35. 20–1.74. *The Astrophysical Journal*, **693**, 419. 73, 85

ZINNECKER, H. & YORKE, H.W. (2007). Toward understanding massive star formation. *Annual Review of Astronomy and Astrophysics*, **45**. 18

ZUCKER, C., SCHLAFLY, E.F., SPEAGLE, J.S., GREEN, G.M., PORTILLO, S.K., FINKBEINER, D.P. & GOODMAN, A.A. (2018). Mapping distances across the perseus molecular cloud using co observations, stellar photometry, and gaia dr2 parallax measurements. *The Astrophysical Journal*, **869**, 83. 31, 91

ZUCKER, C., SPEAGLE, J.S., SCHLAFLY, E.F., GREEN, G.M., FINKBEINER, D.P., GOODMAN, A.A. & ALVES, J. (2019). A large catalog of accurate distances to local molecular clouds: The gaia dr2 edition. *arXiv preprint arXiv:1902.01425*. 31