# 3D Modelling of the Large Magellanic Cloud using Markov Chain Monte Carlo

Author: Jan Strunk.

*Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.**

Advisors: Xavier Luri and Óscar Jiménez-Arranz

**Abstract:** The Large Magellanic Cloud (LMC) is one of the closest galaxies to the Milky Way, at a distance of about 50 kpc. With the latest *Gaia* Data Release, *Gaia* DR3, and its unprecedented precision and number of stars, the study of the galaxy can reach new depths. However, the uncertainty in the parallax for LMC stars is of the same order of magnitude of the values themselves, making the use of individual measurements to determine distance and structure unfeasible. In this work we will study the possibility of exploiting DR3 through the use of a Markov-Chain Monte Carlo (MCMC) algorithm. By studying the algorithm's ability to return parameters used to generate synthetic models of the LMC with varying uncertainties, we will assess the feasibility of the use of *Gaia* parallaxes in the 3D modelling of the LMC structure.

## I. INTRODUCTION

The *Gaia* Mission, launched by the European Space Agency (ESA) in 2013, is an astrometric survey aiming to generate a three-dimensional map of the Milky Way with unprecedented precision [1]. *Gaia* is able to measure the positions of over a billion stars, and the distance through their parallax [2]. *Gaia* Data Release 2, published in 2018, provided an extensive and high-quality dataset of parallax measurements, which greatly enhanced our understanding of our galaxy's composition, formation and evolution [3]. While parallax is a good distance estimator for stars in close proximity, its uncertainty quickly increases with distance. With *Gaia* Data Release 3, DR3, parallax precisions were increased by 30%, and systematic errors in astrometry reduced by 30-40% [4], making it more feasible to study stars in galaxies of the Local Group.

Because of its proximity to us, the Large Magellanic Cloud (LMC) provides an extraordinary sample of stellar populations to study galactic dynamics and star formation [5]. Moreover, its proximity and its Cepheid star population makes the LMC an excellent tool to calibrate the cosmic distance ladder [6].

As much as the LMC has proved to be a useful laboratory for astrophysics and cosmology, its 3D modelling has not been a straightforward enterprise. Distance measurements have been done through several standard candle methods, based for instance on Cepheid variables or eclipsing binaries, with the latter providing the most precise measurements of $D = 49.59 \pm 0.55$ kpc [7]. However, where more information is lacking is in terms of its geometric parametritzation. While it was initially assumed that the LMC was a planar galaxy with all its stars at roughly the same distance to Earth, in 1986, distance measurements from Cepheid variables found that the galaxy had a certain inclination with respect to our line of sight [8]. This has been latter confirmed through core helium burning red clump stars [9] and near IR observations of red giants [10], with all finding inclination values of around $i = 35°$.

The LMC is at the very limit of usefulness of *Gaia* parallaxes for determining distances. The average parallax for the LMC is of 0.02 mas, corresponding to a distance of about 50 kpc, with the instrumental uncertainty of *Gaia* being typically an order of magnitude greater. Additionally to this high relative random uncertainty, we also have systematic errors due mainly to the parallax zero-point offset, which for the LMC is of -0.0242 mas [11]. Due to these uncertainties, *Gaia*'s parallax data cannot be used through direct sampling to produce a parametrization of the LMC. The following work will instead attempt to parametrize the 3D LMC structure by using a Markov Chain Monte Carlo (MCMC) algorithm, a stochastic method for sampling from complex, high-dimensional probability distributions [12]. In order to study whether such an algorithm will be able to infer the LMC 3D structure, we have generated synthetic samples of *Gaia* with different uncertainty values.

The work is organised as follows. In Section II, we describe the model generation, the initial parametrization and the algorithm. In Section III, we discuss the obtained results and computational times for the fitting of the generated synthetic models. Finally, in section IV, we present the conclusions of the work.

## II. METHODOLOGY

### A. MCMC Algorithm

The astrometric observables that *Gaia* provides are the positional parameters of individual stars, those being the right ascension $\alpha$, declination $\delta$ and parallax $\varpi$, as well as the proper motions, which were not used in our investigation but can be a subject of further study. The objective of our investigation is to explore whether these measurements can be used to infer the 3D LMC structure, since

*Electronic address: jan@strunk.cat

the uncertainty of the parallax measurements means that we cannot do so by individually studying each *Gaia* DR3 sample.

*Gaia* measures the parallax by measuring changes of position of individual stars, which are smaller for larger distances, producing a higher relative uncertainty for a given error of measurement. This leads to the distance uncertainty for LMC measurements being significantly larger than with standard candle methods, such as Cepheid variables, eclipsing binaries or top of the giant branch IR astrometry. However, those methods rely on stars and phenomena that are rare in comparison to the standard star population of the LMC, and are thus insufficient to determine the overall structure of the galaxy. Therefore, although the *Gaia* parallax measurements incorporate a higher uncertainty, this paper will seek to exploit the much larger sample of stars from DR3 in order to get more information on the geometric parameters of the LMC. As such, we will explore whether the overall instrumental uncertainty and the systematic uncertainties of *Gaia* DR3 mask the signature of the LMC 3D structure to the point that its parameters cannot be inferred, or if alternatively some information can still be extracted.

In 2021, *Luri et al.* attempted to infer the structural properties of the LMC through two processes based on Bayesian inference: MCMC and Approximate Bayesian Computation (ABC) [11], but both approaches were unsuccessful. The latter failed because eDR3 did not provide enough information to both infer the zero-point parallax variations as well as the LMC 3D structure. The former, however, was unsuccessful because of high computational times when scaling the MCMC algorithm to the full size of the *Gaia* DR3 sample population, thus leaving the door open for further study. Through the use of Markov chains [13], MCMC algorithms generate sample distributions that iteratively narrow down on an approximation of the parameters that lead to the distribution of observations [14]. This makes them particularly effective for high-dimensional analysis with large data samples such as ours. This work studies the computational scaling and uncertainty constraints of such an approach, in order to determine the feasibility of inferring the LMC 3D structure.

With the purpose of reducing computational times, the MCMC algorithm computations have been run using a distributed computing framework. This has been done with the TensorFlow library [15], more specifically TensorFlow Probability [16], and the algorithm has been run at the CTE-Power cluster of the Barcelona Supercomputer Centre (BSC).

Written by *A. Berihuete*, our particular MCMC algorithm requires a probabilistic generative model of the *Gaia* measurements, i.e. a model that informs the algorithm how the observations will be distributed. While the stellar distribution of the LMC has a dense off-center stellar bar and spiral arms [17], we have simplified the modelling of the galaxy using the usual assumptions of an elliptic disc, with a stellar population given by a gamma radial distribution and a vertical Laplacian distribution. For this initial stellar population generation, the algorithm requires three global parameters: the disc scale height, $h_0$, the disc scale length, $R_0$, and the disc ellipticity, $\epsilon$. Afterwards, the algorithm transforms the galactocentric Cartesian $(x, y, z)$ coordinates of the generated stars to heliocentric coordinates, and those are transformed into the observables measured by *Gaia*, namely the position and parallax. For a more detailed description of the coordinate transformations applied by the algorithm, see the Appendix. These coordinate transformations use a set of six more global parameters: the disc minor axis position angle, $\theta_{ma}$, the disc inclination, $i$, the LMC line of nodes angle position, $\theta_{LON}$, and the position of the galactic centre, $\alpha_0$, $\theta_0$ and $D$. The combination of these six global parameters with the three mentioned previously lead to the algorithm using nine global parameters for its probabilistic generative model.

The algorithm will then propose random changes to these parameters within an interval of values that are plausible based on the scientific literature, and use the generative model to infer what the resulting simulated observations would be. It will then proceed to evaluate whether those changes lead to simulated *Gaia* observables more similar to the *Gaia* measurements than before the proposed change, accepting the change if it increases the likelihood functions and rejecting it if it does not. The sufficiently repeated iteration of this process will make the generated distribution converge to the distribution of stars leading to the *Gaia* measurements.

While the method is guaranteed to converge to the desired distribution as long as the number of iterations is large enough, MCMC algorithms can be particularly sensitive to the choice of tuning parameters and initial steps, with convergence being computationally very slow [18]. In order to reduce the number of iterations necessary, and thus lower the computational time required to reach a solution, we have aided the algorithm reach convergence earlier through two methods.

First, we can artificially set a starting point for the algorithm that is already close to the desired distribution. Our MCMC algorithm will maximize the likelihood function by comparing the observables for each individual model generated star to a corresponding individual DR3 source. The number of iterations required for each generated star to correspond one-to-one to a data point is very large, and thus the computation time would be considerable. In order to help the algorithm compare one-to-one, we can ensure that the model generated stars are generated in the vicinity of a DR3 data point. We can do that by applying an inverse transformation of the data coordinates (that is to say, we use *Gaia*'s position and parallax to obtain the corresponding disc-centered Cartesian coordinates, and then generate the model's star close to those coordinates). This solution can present problems when the uncertainty in the parallax is large, as it will thus generate stars that will be within the predefined disk, but can be relatively far away from the true positions of the stars whose positions we want to infer. An alternative solution can be to estimate the (x,y,z) position of individual stars with *Gaia*'s $\alpha$ and $\delta$ measurements, and then projecting in a straight line by using measurements of the distance to the galactic center, inclination and line of nodes angle position from other studies. This method should reduce the dependency with the uncertainty of the parallax measurements, but it would also incorporate a further source of confirmation bias, given that the D, $i$ and $\theta_{LON}$ are parameters that we are trying to estimate.

Secondly, we can restrict the interval of values of the parameters that the algorithm will explore. If we know that the parameter's value has to be at a specific interval, we can establish prior limits so that the algorithm does not spend computational time exploring values that do not make physical sense for our problem. In a generic example, we could establish a prior limit that the distance to the galactic centre, $D$, has to be positive. However, if we want to further reduce computational time we can extend this process, by not only establishing prior limits that correspond to physical possibility but also constraints that come from measurements done using other methods. In the case of the distance to the galactic centre, to continue the example, eclipsing binary measurements have lead to an uncertainty of less than 2%, and thus we can decide to force the algorithm to only explore that interval. This procedure presents some problems, given that again we are introducing a certain confirmation bias: if the prior interval gives a restriction bigger than the algorithm is able to yield based on *Gaia* DR3, our results will not be based on the algorithm but rather on the prior constraints given. It can be argued however that given the multidimensionality of the parametrization, restricting one parameter more than the measurements allow might lead to narrower and thus more precise distributions for the other parameters for which we lack more precise prior knowledge. A more detailed discussion of what prior restrictions have been used is found in the following Section II B.

### B. Prior parametrization for the MCMC fitting

As mentioned before, we have established a set of simplifications and restrictions to facilitate the convergence of the algorithm. First of all, in order to reduce the number of calculations that the algorithm has to produce at each iteration, we have taken the scale height and scale length as constants that the algorithm does not change, with $h_0 = 0.35$ and $R_0 = 1.6$ taken from *Weinberg & Nikolaev* [19]. Additionally, we have only generated synthetic models with an ellipticity of $\epsilon = 1$, thus simplyfying the algorithm's calculations by not having to consider either $\epsilon$ or the minor axis angle, $\theta_{ma}$. All these parameters are still part of the generative model, however, and can easily be introduced as modifiable parameters to the algorithm in further studies, at the cost of higher computation times. Our algorithm thus will focus on generating distributions for the 5 remaining parameters: $D$, $\alpha_0$,$\delta_0$, $i$ and $\theta_{LON}$.

For the prior limits of the distance to the galactic center, we have established the interval $D = 50.0 \pm 1.0$ kpc in accordance with the measured by *Pietrzyński et al.* using eclipsing binaries [7]. For the inclination and position angle of nodes, $i$ and $\theta_{LON}$, there exists a certain dispute depending on the method used to measure them, with ranges between $i = 25°$ and $i = 38°$ for the inclination angle and between $\theta_{LON} = 212.5°$ and $\theta_{LON} = 235°$ for the line of nodes angle. A detailed discussion of the different stellar populations in the LMC and the differences that it generates between methods to measure these angles, such as near-IR color magnitud diagrams[10], red clump magnitudes [9] or cepheids[19] is found in *Subramanian & Subramanian, 2010*[20]. Given that the *Gaia* DR3 is a general star catalogue, we elected to use the near-IR measurements by *van der Marel & Cioni* as the central values, but with large enough intervals to consider the different possibilities when appropriate, leading to prior limits of the inclination angle, $i = 34.7° \pm 5.0°$ and the position angle of nodes, $\theta_{LON} = 212.5° \pm 10.0°$. Lastly, the celestial coordinates of the galactic center, $\alpha_0$ and $\delta_0$, were taken from the Third Reference Catalogue of Bright Galaxies, with $\alpha_0 = 80.89° \pm 2°$ and $\delta_0 = -69.75° \pm 2°$ [21].

Finally, it is worth to mention that if we had ended up being able to use *Gaia* DR3 as the algorithm's input sample, we would have required to establish LMC membership, as done in *Jiménez-Arranz et al.* [22]. For further study, we could attempt to reduce the uncertainties of our *Gaia* data sample by only considering a subset of all the observed samples with the brightest stars, given that those will have lower parallax uncertainties on average.

## III. RESULTS

We have studied the accuracy and precision of the results of the algorithm when it has been given synthetic models that simulate *Gaia* observables. The algorithm returns the values at each iteration of the parameters that have been accepted, generating thus a density distribution of accepted values, of which we have studied its mean and standard deviation, $\sigma$. We considered the algorithm's results to be accurate if the mean of the algorithm generated distributions was within $3\sigma$ of the parameters corresponding to the simulated data. Precision has been determined more quantitatively, by studying the order of magnitude of the standard deviation, which is different for each parameter studied. The algorithm was deemed to be successful in convergence if it was both accurate and precise.

First, in order to validate the formal written code, a synthetic sample with $D = 50.0$ kpc, $\alpha_0 = 80.89°$ $\delta_0 = -69.75°$, $i = 34.7°$ and $\theta_{LON} = 212.5°$ is generated without any uncertainty. With that sample as the algorithm's input, we confirmed that the programming worked and that it was able to recover the original parameters. Additionally, it allowed us to see the dependency of the computational time with the computational parameters of our algorithm, i.e. the number of iteration steps done and the number of stars of the sample, which is shown in Figure 1.

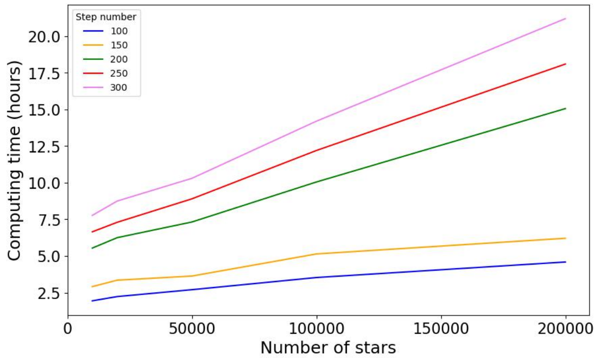


FIG. 1: Comparison of the computational time taken by the MCMC algorithm as a function of the number of stars of the synthetic sample, for different numbers of iteration steps.

We observe that the behaviour of the computation time is roughly lineal both with the number of steps and the number of samples studied, the latter one being an exceptionally good sign for using this algorithm for further study using the full *Gaia* LMC dataset of the order of $10^7$ samples[11]. We additionally observed that beyond 250 steps and $2 \cdot 10^4$ samples, any further increase in these parameters did not lead to more precise nor accurate results for the range of uncertainties studied. Thus, for the rest of the work, we used 250 iteration steps and $2 \cdot 10^4$ stars as our data sample.

With the purpose of validating the concept of our investigation and the MCMC method, we introduced random uncertainties into our synthetic sample. The uncertainties were introduced by individually sampling from a Gaussian distribution with mean equal to the true value of the corresponding data sample and with the desired simulated uncertainty as its standard deviation. Progressively increasing the uncertainties, we found that starting at uncertainties in parallax of the order of $\Delta\omega \approx 0.001$ mas, the algorithm had to be aided in finding an appropriate first step, as mentioned in II A. The algorithm struggled to converge when given samples with parallax uncertainties of around $\Delta\omega \gtrsim 0.005$ mas, and was unable to reach the combined random and systematic uncertainties of $\Delta\omega \approx 0.2$ mas that *Gaia* DR3 presents. In Figure 2, we show the distributions generated by the MCMC algorithm for the 5 global parameters studied, and for the various synthetic catalogues with predefined uncertainties. In Table I the mean values and the standard deviation of the generated distributions are presented.

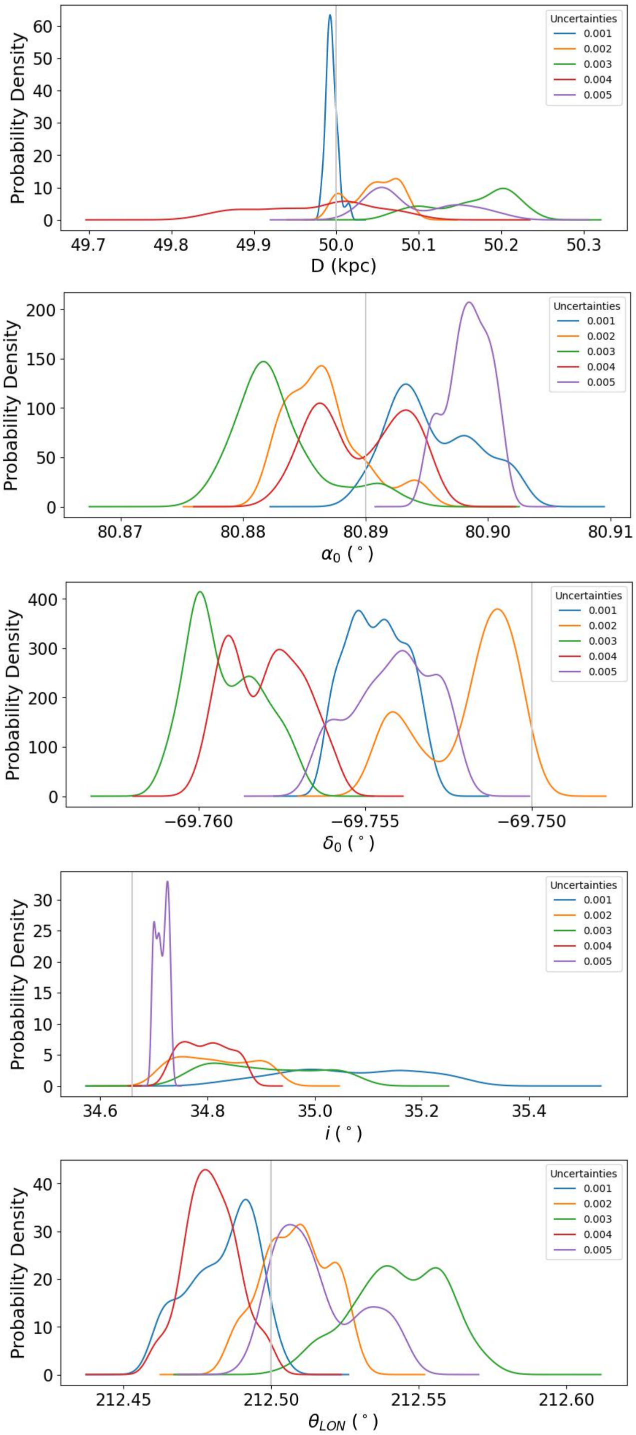


FIG. 2: Histograms generated by 250 steps of the MCMC algorithm from a synthetic catalogue of $2 \cdot 10^4$ stars with different parallax uncertainties. The probability density is calculated through a Kernel Density Estimation, and the vertical line indicates the value of the parameter used to generate the sample.

| $\Delta\omega$ (mas) | 0.001 | 0.002 | 0.003 | 0.004 | 0.005 |
|---|---|---|---|---|---|
| D | 49.994 | 50.046 | 50.166 | 49.974 | 50.095 |
| (kpc) | ±0.007 | ±0.029 | ±0.048 | ±0.071 | ±0.054 |
| $\alpha_0$ | 80.896 | 80.887 | 80.883 | 80.89 | 80.898 |
| (°) | ±0.004 | ±0.003 | ±0.004 | ±0.004 | ±0.002 |
| $\delta_0$ | -69.755 | -69.752 | -69.759 | -69.758 | -69.754 |
| (°) | ±0.001 | ±0.001 | ±0.001 | ±0.001 | ±0.001 |
| $i$ | 35.07 | 34.82 | 34.906 | 34.798 | 34.715 |
| (°) | ±0.13 | ±0.07 | ±0.098 | ±0.043 | ±0.011 |
| $\theta_{LON}$ | 212.483 | 212.508 | 212.543 | 212.48 | 212.516 |
| (°) | ±0.011 | ±0.011 | ±0.015 | ±0.009 | ±0.014 |

TABLE I: Table with the mean and standard deviation values of the distributions shown in Fig. 2

As we can see from Fig. 2 and Table I, the probability densities that present higher dispersion are those that have a higher dependency on the parallax measurement ($D$,$i$ and $\theta_{LON}$), while those that are projections of stellar coordinates ($\alpha_0$ and $\delta_0$) have much narrower distributions. That is to be expected, given that in accordance to *Gaia*'s precision, we introduced relative uncertainties into our synthetic model for parallax much greater than the ones for the right ascension and declination measurements for each star. From the histograms shown in Fig. 2 and the standard deviations, we see that the algorithm struggled to return the original distance of $D = 50.0$ kpc, but even then for all uncertainties its value was within $3\sigma$. Particularly remarkable was the accuracy of the algorithm to infer the true values of inclination, $i$, and line of nodes angles, $\theta_{LON}$, with the means of the generated distributions all having discrepancies smaller than $2\sigma$. The standard deviations for all generated probability densities were relatively low, indicating that, at least for the ranges of uncertainty where the algorithm was able to run, it was considerably precise. Additionally, while it would be expected for the precision to decrease when the parallax uncertainty was increased, it remained constant. Both these observations indicate that the algorithm should be able to infer parameters even when the uncertainties are larger than the ones here presented. However, for these higher uncertainties, the algorithm does not propose any successful iterations apart from its first step, thus effectively yielding Dirac $\delta$ distributions centered at the randomly chosen initial values.

## IV. CONCLUSIONS

In this work, we have studied the ability of our MCMC algorithm to infer the LMC 3D structure when given synthetic samples with predefined parallax uncertainties.

The successful parametrization of the simulated LMC samples for parallax uncertainties of up to $\Delta\omega = 0.005$ mas, validates the formal and conceptual aspect of the method. However, the algorithm's failure to converge for greater uncertainties, and consequently the lack of an improvement in precision for an increase in the sample size, means that we were not able to validate that the method can be applied to *Gaia* DR3. The relatively high and constant precision of the algorithm for the ranges of parallax uncertainty studied seems to indicate that our algorithm should conceptually be able to succeed at higher uncertainties, specially with larger sample sizes. The failure to do so likely lies in either not finding a suitable first step, or on the configuration for the implementation of the MCMC not being optimal. Further study can thus be undertaken through additional tuning of the algorithm both to the initial parameters and larger sample sizes.

**Acknowledgments**

This work would not have been possible without the help from its advisors, Xavier Luri and Óscar Jiménez-Arranz, who guided me when was needed but also gave me the freedom to learn and develop independently. The code from Ángel Berihuete, as well as the helpful correspondence to understand it, were also pivotal, for which I am grateful. Finally, I am most thankful for the unconditional support from my partner and my parents.

[1] Perryman et al., A&A 369 (2001)
[2] Gaia Collaboration (Prusti et al.), A&A 595, A1 (2016)
[3] Gaia Collaboration (Brown et al.), A&A 616, A1 (2018)
[4] Gaia Collaboration (Brown et al.), A&A 649, A1 (2021)
[5] Koerwer, AJ, 138, 1 (2009)
[6] Freedman et al., AJ, 553, 1 (2001)
[7] Pietrzynski et al., AJ, 697, 1 (2009)
[8] Caldwell and Coulson, MNRAS, 218, 2 (1986)
[9] Olsen and Salyk, AJ 124 (2002)
[10] Cioni and van der Marel AJ 122 (2001)
[11] Gaia Collaboration (Luri et al.) A&A 649, A7 (2021)
[12] Gelman et al., "Bayesian Data Analysis" Chapman and Hall/CRC, 3rd Edition (2013)
[13] Markov, Bulletin of the Society of Physics, 2 (1906)
[14] Robert and Casella, "Monte Carlo Statistical Methods" Springer, 2nd Edition (2004)
[15] Abadi et al., CCS'16: 2016 ACM (2016)
[16] Dillon et al., arXiv e-prints:1711.10604 (2017)
[17] El Youssoufi et al. MNRAS, 490, 1, (2019)
[18] Rosenthal. "Markov Chain Monte Carlo Algorithms: Theory and Practice", Springer (2009)
[19] Nikolaev et al., AJ (2004)
[20] Subramanian and Subramanian, A&A 520, A24 (2010)
[21] de Vaucouleurs et al., "Third Reference Catalogue of Bright Galaxies", Springer (1991)
[22] Jiménez-Arranz et, al, A&A 669, A91 (2023)

# V. APPENDIX

## A. Coordinate transformations

The first coordinate transformation from the galactocentric Cartesian coordinates is a rotation along the third axis, by an angle equal to $\theta_{ma}$. If the generated disc has no ellipticity, this rotation does not affect our coordinate changes.

$$R_{rot1} = R_{[3]}(\theta_{ma}) = \begin{pmatrix} \cos\theta_{ma} & -\sin\theta_{ma} & 0 \\ \sin\theta_{ma} & \cos\theta_{ma} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Afterwards, we introduce the inclination and the line of nodes angles. We achieve that by a rotation of an angle $-i$ along the first axis and then an angle $\theta_{LON}$ along the third axis.

$$R_{rot2} = R_{[3]}(\theta_{LON})R_{[1]}(-i) =$$

$$= \begin{pmatrix} \cos\theta_{LON} & -\sin\theta_{LON} & 0 \\ \sin\theta_{LON} & \cos\theta_{LON} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos i & \sin i \\ 0 & -\sin i & \cos i \end{pmatrix} =$$

$$= \begin{pmatrix} \cos\theta_{LON} & -\sin\theta_{LON}\cos i & -\sin\theta_{LON}\sin i \\ \sin\theta_{LON} & \cos\theta_{LON}\cos i & \cos\theta_{LON}\sin i \\ 0 & -\sin i & \cos i \end{pmatrix}$$

Subsequently, we transform the coordinates into heliocentric ones. We achieve that by applying a rotation related to the angular position from Earth (or rather, the Sun), as well as applying a shift in regards to the position of the center of the LMC in relation to the position of the Sun.

$$\begin{pmatrix} x_H \\ y_H \\ z_H \end{pmatrix} = \begin{pmatrix} \sin\alpha_0 & -\cos\alpha_0\sin\delta_0 & -\cos\alpha_0\cos\delta_0 \\ -\cos\alpha_0 & -\sin\alpha_0\sin\delta_0 & -\sin\alpha_0\cos\delta_0 \\ 0 & \cos\delta_0 & -\sin\delta_0 \end{pmatrix} \times$$

$$\begin{pmatrix} \xi_{rot2} \\ \eta_{rot2} \\ \zeta_{rot2} \end{pmatrix} + \begin{pmatrix} r_0\cos\delta_0\cos\alpha_0 \\ r_0\cos\delta_0\sin\alpha_0 \\ r_0\sin\delta_0 \end{pmatrix}$$

Finally, we want to transform the heliocentric coordinates to the observables that *Gaia* will measure, the parallax $\omega$, the right ascension $\alpha$ and the declination $\delta$:

$$\omega = \frac{1}{\sqrt{x_H^2 + y_H^2 + z_H^2}}$$

$$\alpha = \tan^{-1}\left(\frac{y_H}{x_H}\right)$$

$$\delta = \sin^{-1}\left(\frac{z_H}{1/\omega}\right)$$