Draft version March 11, 2023
Typeset using LaTeX **twocolumn** style in AASTeX631

Gaia Parallax Distances: What Can Go Wrong and How to Fix It

Michael Tauraso

ABSTRACT

With the unique and revolutionary measurements of the Gaia era affecting many areas of the study of the structure and history of the Milky Way, this paper looks at the parallax distances and explores the high level causes of spurious parallaxes in Gaia DR3, as well as exploring some approaches for using these parallaxes and the astrometric solution as a whole to measure distance.

1. INTRODUCTION

ESA's Gaia mission has been a boon to the study of the Milky Way. At time of writing Data Release 3 (DR3) has a publically searchable catalog of some 1.8 billion sources with 88% of those having a high quality astrometric solution, and some $Y\%$ containing a published astrometric parallax with. In comparison with Gaia's predecessor sattellite, Hipparcos, this is an increase of $10x$??? in the number of sources, and a $x\%$?? increase in precision.

The detailed study of the structure of the milky way benefits greatly from the enhanced precision of these parallax measures; however Gaia's astrometric parallaxes contain an estimated ($y\%$???) spurious astrometric solutions including ($z\%$???) negative valued parallaxes. The interpretation of these parallax values as distances is the subject of some study, and their presence is in some ways a commitment to a certain predictability in the structure of astrometric data processing, which has independent scientific value outside of the accuracy of the data themselves.

Section 2 will sketch the process by which Gaia parallaxes are determined, and in Section 3 I will discuss some of the ways this process can lead to a erronious (or even negative) parallax. Section 4 will sketch and compare efforts to derive distances from DR3 parallaxes. Section 5 will discuss improvements expected in Gaia's upcoming data release (DR4), as well as comparing DR3 with the prior major data release (DR2). Finally, Section 6 will conclude by summarizing an unorthodox use of the astrometric solution which are enabled by Gaia's data processing discipline.

2. HOW GAIA DOES PARALLAX

The standard introductory treatment of parallax as the apex angle of a giant triangle subtended by two ends of the earth's orbit and a distant star, while pedagogically interesting and trigonometrically correct, pales in comparison to the complexity measuring nearly 2 billion parallaxes on a modern space telescope. The Gaia spacecraft sits in a Lissajous orbit at the Earth-Sun L2 Lagrange point. Gaia's two telescopes, separated by a basic angle shown in figure 1, scan the sky as the satellite rotates. As the telescope rotates, sources of light across the universe shine on to CCD detectors for each telescope as shown by the yellow star in figure 1. The times of these transits and the response of the CCDs are recorded.

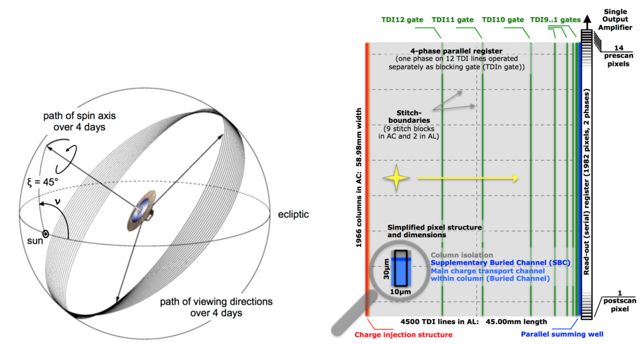

Figure 1. Gaia Scan illustration (left) and Gaia CCD schematic (right) (Collaboration 2016)

The first steps of this analysis are focused on constructing a frame of reference for the astronomic observations and accurately placing the spacecraft and its instruments within that reference frame. Well known astronomic sources with predictable behavior are identified algorithmically and used as points of reference to calibrate detailed mathematial models of the spacecraft's location over time. Once this model of the spacecraft is calibrated, it is used to convert timed transit events to sky locations and times where a source was observed. The time that a source passes by each telescope is critical, because the time that each source appears and leaves the field of view of the spacecraft is an additional piece of information that allows its sky location to be inferred with greater accuracy

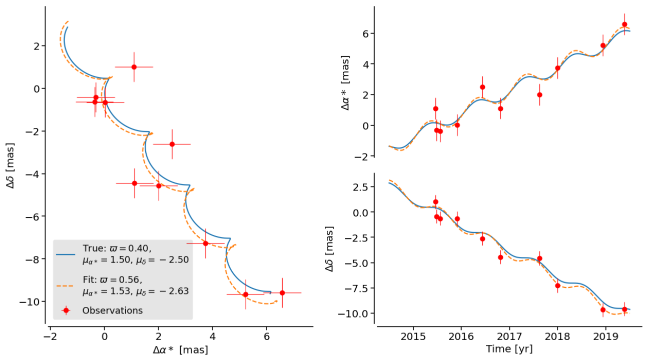

Figure 2. A simplified but representative astrometric curve fit. Simulated star location is in blue, simulated measurements are in red, and the fit curve is orange. The error bars are of representative size to Gaia's data. Actual and fit parameters for proper motion and Parallax are reported on the plot. Simulation was generated by author with code from Luri et al. (2018)

These observations must then be grouped algorithmically, such that each group of observations is identified as being from the same source in the night sky. Parallaxes then flow from a least-squares fit of a linear model to each group of observations identified with a particular source. A simplified and simulated view of such a curve fit is shown in Figure 2.

Each gaia source includes around 200 observations, and the published astrometric parameters, errors and correlations flow directly from this curve fit. The ideal case for this pipeline is a high quality 5-parameter fit, so called because the 5 astrometric parameters Right Ascension (α), Declination (δ), Right Ascension Proper Motion (μ_α), Declination Proper Motion (μ_δ), and Parallax ($\bar{\omega}$) are determined[1] by the least squares fit(Lindegren et al. 2021).

There can be issues with the veracity of this fit. Gaia's CCD instrument and telescope have non-linear response to different frequences of light, and it is well known that objects in the sky are not monochromatic. These two effects together affect the calibration of the instrument, because there are chromatic effects that affect where, when, and how much a given source is recorded on the CCD. Gaia corrects for these effects using a model that takes as input a single color, reported in the published data as `nu_eff_used_in_astrometry`. In the case where this color can be determined from spectral observations, and the resulting 5 parameter fit is of sufficiently high quality, the 5-parameter fit is reported.

In the case that the 5-parameter fit is not high enough quality to be reported, Gaia Astrometry falls back to either a 6 parameter fit or a 2 parameter fit. The 6 parameter fit treats `nu_eff_used_in_astrometry` as an additional unknown taking part of the least squares fit. The value found by the curve fit is reported in the database as `pseudocolor`. Typically very bright and very dim sources require this treatement for different reasons. Dim sources because there is not enough light to determine their true color. Bright sources end up having a 6-parameter fit because their 5-parameter fit is uncertain as their light is washing out much of the positional precision that Gaia would otherwise have[2]. 2-parameter fits (only α and δ) are reported when neither 5-parameter nor 6-parameter fits reach the desired level of quality.

3. WHAT CAN GO WRONG?

The process of source identification is possibly the most error-prone step in the entire astrometric pipeline, involving both components on-spacecraft that initially identify sources and a large amount of earthbound data analysis. Dim sources in crowded fields are particularly suceptible to source misidentification; however, there is a long list of astrophysical phenomena, mostly affecting dim objects, which can cause a source misidentification error. Tracking down and improving the source identification algorithm is an area of nontrivial current work.

In addition to source identification issues, there are astrophysical phenomena that can cause a correctly identified source to not fit the linear 5 parameter model. Some of these have a somewhat mean-reverting property to linear motion such as binary systems, gravitationally lensed sources, and sources with dark companions. Others, such as stellar close encounters, and exceptionally fast sources have motion that simply diverges from the underlying model. These types of systems can cause a low quality or even a spurious curve fit depending on the magnitude of the effect and the timing of observations. The distribution of these different sorts of fits, as well as the fraction of them that are spurious are illustrated in Figure 3.

[1] Radial Velocity (v_r) is unmentioned here, because radial velocity is determined spectroscopically, and is used as input data to the curve fit in DR3 where it is available. While in theory v_r could be an output of an astrometric curve fit, in practice this method yields useful results only for the closest and brightest sources(Lindegren et al. 2021).

[2] Researchers familiar with the DR2 data may note it contains several close, high proper motion, bright sources. These reportedly arose from falling back to a 6-parameter fit for a bright source with high error in its 5-parameter fit. The acceptance parameters for the 5-parameter sources were incorrectly tuned for bright sources in DR2; however, this error has been corrected in DR3 with a G magnitude dependent criteria (Lindegren et al. 2021).

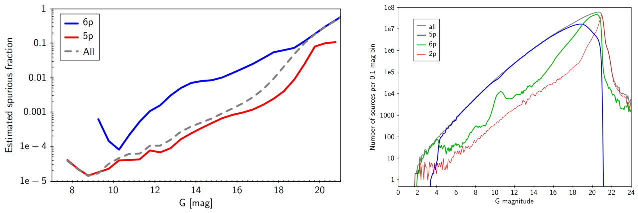

Figure 3. Left: Estimated fraction of spurious astrometric solutions by type in EDR3 (Fabricius et al. 2021). Right: Total number of astrometric solutions by type in EDR3 by G magnitude (Lindegren et al. 2021).

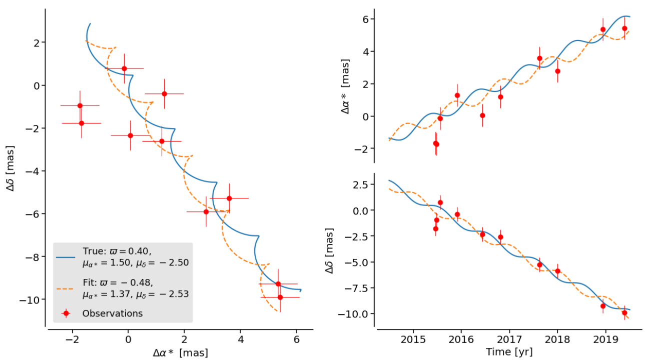

Figure 4. Same as figure 2, except the fit is to a negative parallax value. Note the orange curve (fit) differs from the blue (actual) by a phase of π. Figure generated from code in Luri et al. (2018)

Even if source identification works and the source moves linearly over the observation timescale, you can still have a curve fit that is simply wrong due to the timing and uncertainty in the observations. Many negative parallaxes fall into this category. Figure 4 shows a bad fit for a simulated star, yielding a negative parallax. In this figure, the "measured" data points were derived from a simulation of a star with linear motion, and normally distributed measurement uncertainties.

Spurious astrometric solutions can vary greatly in terms of quality. While every negative parallax is definitely spurious, there are also spurious astrometric solutions that generate slightly-wrong values, many of which are within the formal errors for the reported astrometric parameters. There is ultimately still information about distance in some of these spurious solutions. For example, one step in verifying the calibration of the astrometric pipeline is computing the parallaxes for several known quasar sources(Luri et al. 2018). These parallaxes ought all be zero, but they are in-fact normally distributed around zero and half of them are negative! Determining whether an astrometric parallax has information is primarily an issue of looking at both the parallax value and the error estimate. When the error is small

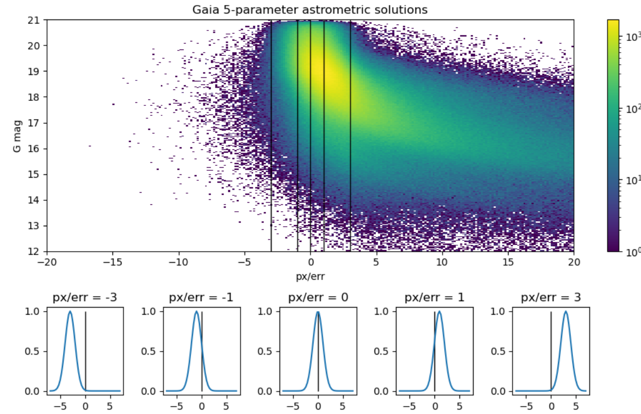

Figure 5. Histogram of Gaia DR3 5-parameter solutions by `parallax_over_error` measure and G magnitude. Lower plots show example normal distributions for parallax measurements implied by selected `parallax_over_error` values. These same selected values are marked on the main plot by the black vertical lines. Data from Collaboration et al. (2022)

relative to the value, there is a much greater chance that even a negative parallax has information.

Figure shows a histogram of a randomly selected subset of Gaia DR3 sources. The lines show schematically how as one samples the diagram further and further to the left, the negative parallax is greater and relatively more certain, as illustrated by the normal distributions at the bottom of the figure. The brightness dependence of spurious parallaxes can clearly be seen in this diagram, as well as the prominence of negative parallaxes with high relative uncertainty.

4. REASONING DISTANCES FROM PARALLAX

Researchers working on Gaia recommend that the issue of finding distances from negative and spurious parallax be treated as a full Baysean inference problem(Luri et al. 2018). The most well known of these is Bailer-Jones method[3]. The most recent Bailer-Jones distances for EDR3 are calculated two ways, once using only a geometric prior, and also using a photogeometric prior. In comparison to other Bayesian methods, Bailer-Jones attempts to keep the priors simple and focused on the geometry of the sky, avoiding more complex prior assumptions that model stellar systems.

The most recent set of Bailer-Jones distances are derived from Markov Chain Monte Carlo (MCMC) based sampling of an un-normalized postieror probability dis-

[3] Bailer-Jones geometric and photogeometric distances for 1.4 billion EDR3 sources are accessible by adding ADQL resembling "`...JOIN external.gaiaedr3_distance as d USING (source_id)...`" to your Gaia archive query (Bailer-Jones et al. 2021)

tribution of distance. For both methods the likelihood is that of a particular parallax method given a distance and parallax uncertainty. The priors are each derived from the GeDR3 Mock Galaxy, and some simplifying assumptions allow the same likelihood to be used in both methods (Bailer-Jones et al. 2021). The photogeometric method achieves slightly greater accuracy than the geometric method by incorporating the G magnitude and $BP - RP$ color (c) into a photometric prior.

With the stars representing un-normalized probability density, the two methods can be summarized in equation form as follows, where p is a sky location, and r is the distance. The first P term on the right hand side of each equation is the shared likelihood, and $Q_g = G - 5log_{10}(r) + 5$ is a measure of absolute magnitude with extinction added in[4].

$$P_g^*(r|\bar{\omega}, \sigma_{\bar{\omega}}, p) = P(\bar{\omega}|r, \sigma_{\bar{\omega}})P(r|p)$$

$$P_g^*(r|\bar{\omega}, \sigma_{\bar{\omega}}, p, G, c) = P(\bar{\omega}|r, \sigma_{\bar{\omega}})P(r|p)P(Q_g|c, p)$$

The output of the MCMC algorithm is samples of the posterior $P_g^*(r)$ function. The maximum of $P_g^*(r)$ for positive r is reported as the most probable distance. Because negative distances are excluded, when this method is applied to an uninformative parallax (like those on the far left in figure 3), the result is that the likelihood term does not contribute to the published distance. The positive distance reported therefore only has information from the priors. This is a common failure mode with Bayesian methods, and a main reason why it is desireable to have a known and well-scoped set of priors, to avoid the method hallucinating data that is not present.

Figure 6 shows a comparison of Bailer-Jones distances derived from Gaia EDR3 both to well characterized Red Cluster (RC) distance measurements, and to StarHorse, a bayesian distance measurement based primarily on stellar modelsQueiroz et al. (2018). The StarHorse sample shown considers only sources for which there is a Bailer-Jones distance and a StarHorse distance; however, many of these sources have larger parallax uncertanties than are present in the Red Cluster sample. While the Bailer-Jones method applied to Gaia EDR3 does relatively well with the Red Clusters, where parallaxes have small error, it begins to deviate from the StarHorse measurements around 6 kpc Bailer-Jones et al. (2021).

Ultimately Bayesian methods to improve parallax can only be as accurate as their priors and the information content of their likelihood. As a comparison to Bailer-Jones, Gaia Collaboration publishes distances derived through their General Stellar Parameterizer from Photometry (GSP-Phot)[5]. GSP-Phot uses several priors derived from stellar models to predict stellar parameters, and then uses those stellar parameters to derive a distance. The distance itself is constrained by a geometric prior similar to the Bailer-Jones geometric prior above. This combination approach systematically underestimates distances, differing significantly from other methods past 3kpc. GSP-Phot distances do not have enough accuracy to map the Milky Way's spiral arms(Andrae et al. 2022).

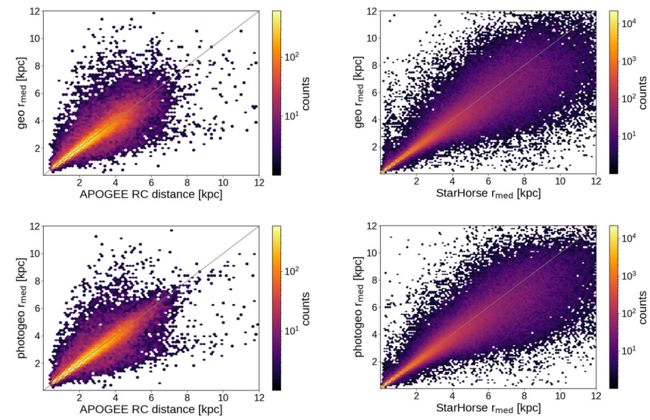

Figure 6. Comparison of Bailer Jones distances to other methods. Left: Bailer-Jones on Gaia EDR3 vs APOGEE Red Cluster (RC) measurements. Right: Bailer-Jones on Gaia EDR3 vs StarHorse Bayesean method. Top: Bailer-jones geometric method. Bottom: Bailer-Jones photogeometric method (Bailer-Jones et al. 2021).

5. IMPROVEMENT EFFORTS

Given the state of the art of Gaia parallax distances, What should we expect in the future? Source identification is the root cause of most spurious astrometric measurements, and therefore the most likely root cause of negative and spurious parallax distances. Given that the gaia catalog has 1.8 billion sources, this is also an extremely time consuming problem. Gaia DR4 is expected to contain more than twice the observation time as Gaia DR3, which will help reduce uncertainty driven by source selection and fitting(Lindegren et al. 2021). Comparing Gaia DR3 to the prior release, DR2, yields some striking improvements in the quality of the astrometric solutions overall.

[4] This construction of the photometric prior enables the construction of the photogeometric posterior listed here from a more formally bayesian expression. See Bailer-Jones et al. (2021)

[5] GSP-Phot distances are available in the `distance_gspphot` column of `gaiadr3.gaia_source` and `gaiadr3.gaia_source_lite`. Additional GSP-Phot data can be found in the `gaiadr3.astrophysical_parameters` table of the Gaia Archive

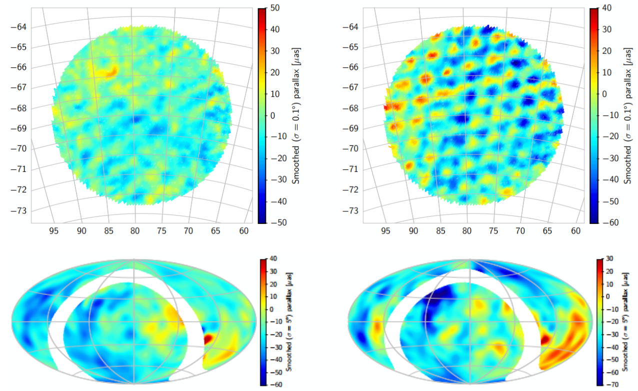

Figure 7. Comparison of parallaxes from Gaia DR2 (left) and Gaia EDR3 (right). Top fields show the Large Magellenic Cloud sky area. Bottom fields show distant quasars. Figures from Lindegren et al. (2021)

Figure 7 shows a selection of parallaxes from the Large Magellenic Cloud (LMC) and of distant quasars, with negative parallaxes appearing in red. The reduction in the waffle pattern of systemic errors in the LMC view, as well as the overall reduction in red in both diagrams shows the striking difference that additional data collection can make. Researchers estimate that the sensitivity limits of the Gaia mission as a whole are still quite far off in Gaia DR3, and that in general uncertanties will be reduced by a factor of 0.7 for positions and parallaxes, and a factor of 0.35 for proper motions(Lindegren et al. 2021).

6. UNORTHODOX USE OF THE ASTROMETRIC SOLUTION

I'd also like to thank Jim Davenport for a truly excellent and informative Galaxies class. The class reminded me of all the messy curiosity, drama, and general insanity that I appreciate about the sciences. I'd also like to thank Jake Kurlander for the first citation of my academic career, given shortly after the precursor presentation to this paper.

REFERENCES

Andrae, R., Fouesneau, M., Sordo, R., et al. 2022, A&A, doi: 10.1051/0004-6361/202243462

Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Demleitner, M., & Andrae, R. 2021, AJ, 161, 147, doi: 10.3847/1538-3881/abd806

Collaboration, G. 2016, doi: 10.1051/0004-6361/201629272

Collaboration, G., Vallenari, A., Brown, A. G. A., et al. 2022, doi: 10.48550/arXiv.2208.00211

Fabricius, C., Luri, X., Arenou, F., et al. 2021, A&A, 649, A5, doi: 10.1051/0004-6361/202039834

Jabłońska, M., Wyrzykowski, Ł., Rybicki, K. A., et al. 2022, A&A, 666, L16, doi: 10.1051/0004-6361/202244656

Lindegren, L., Klioner, S. A., Hernández, J., et al. 2021, A&A, 649, A2, doi: 10.1051/0004-6361/202039709

Luri, X., Brown, A. G. A., Sarro, L. M., et al. 2018, A&A, 616, A9, doi: 10.1051/0004-6361/201832964

Queiroz, A. B. A., Anders, F., Santiago, B. X., et al. 2018, Monthly Notices of the Royal Astronomical Society, 476, 2556, doi: 10.1093/mnras/sty330

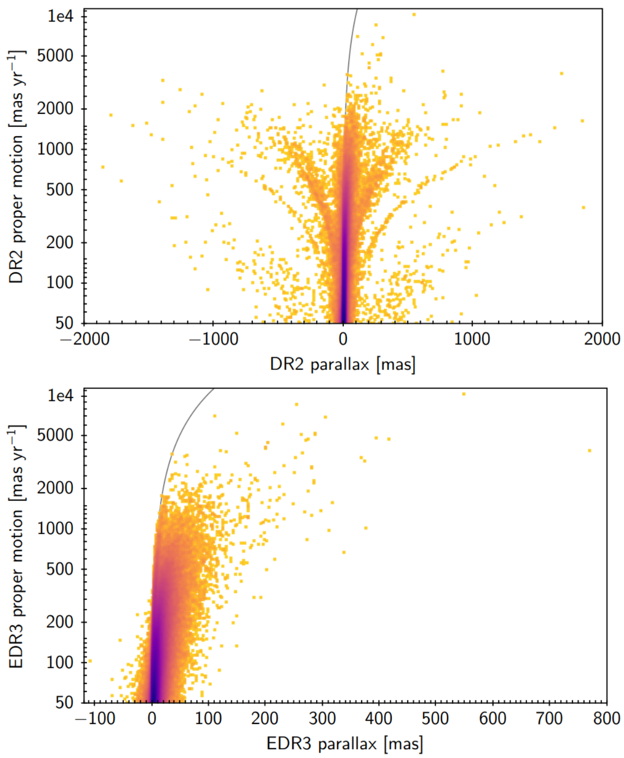

Figure 8. Figure from Fabricius et al. (2021)

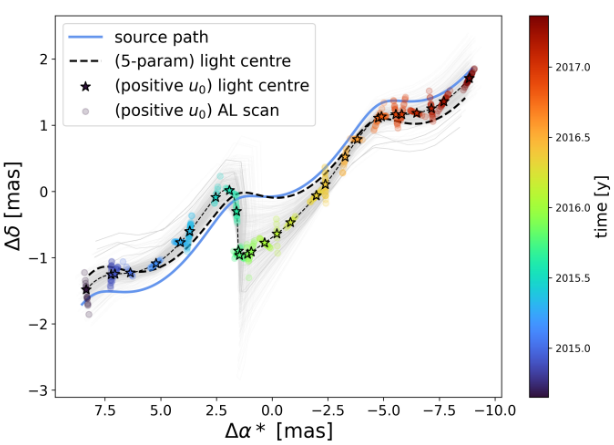

Figure 9. Figure from Jabłońska et al. (2022)

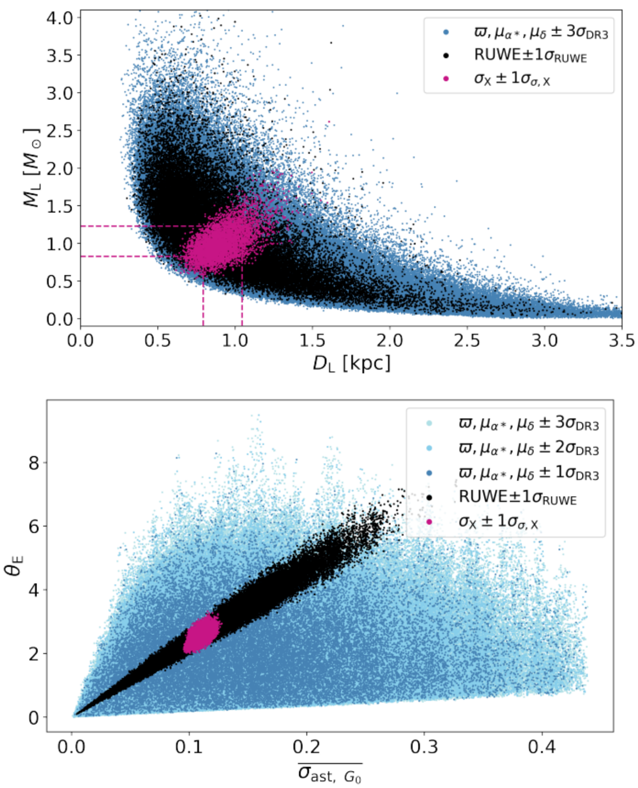

Figure 10. Figure from Jabłońska et al. (2022)