# Distance Estimators' performance on Gaia DR2

Author: Ariadna Ribes Metidieri
*Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.**

Advisor: Xavier Luri Carrascoso
(Dated: January 14, 2018)

**Abstract:** An improvement in the estimation of distance and distance modulus cannot be achieved by only an enhancement of the precision of the trigonometric parallax, but with the correct statistical treatment of the parallaxes to derive these parameters. We aim to provide a recommendation regarding the distance estimators to be used for Gaia DR2 and onwards, as well as to rise awareness about the practice of unquestioningly inverting the parallax. We test the performance of two Bayesian and a frequentist methods over a simulated sample of $10^7$ Gaia DR2 stars, using a specifically developed Python software. We conclude that the use of the Bayesian method with the Exponentially Decreasing Space Density Prior improves the estimation of distance, since it has a good behavior for high relative error parallaxes, with a much smaller bias and dispersion than the rest of estimates.

## I. INTRODUCTION

The Gaia mission, launched on December $19^{th}$ 2013, aims to create a three dimensional map of our Galaxy, with full astrometric (position, distance and motion) and photometric (brightness and color) parameters of 1% of its whole population, which amounts to about 100 billion objects. Last year, the first data release (DR1) was delivered with more than 1.1 billion cataloged sources [1]. Today, Gaia continues observing the sky and will produce more data releases in the coming years, continuing with the second Gaia Data Release (DR2), that is expected to be published on April 2018.

Although Gaia measures the trigonometric parallax, that is, the apparent angular displacement of a stellar object with respect to two opposite points of Earth's orbit, up to a precision of microarcseconds, it cannot directly measure quantities such as the distance or the distance modulus, thus they have to be computed afterwards using the measured parallaxes. The correct way to estimate distances from parallaxes has been addressed by several authors, since the method traditionally used to compute the distance, i.e., inverting the parallax and computing its formal error as the first order Taylor expansion $\sigma_r = \frac{\sigma}{\varpi^2}$ (with $\sigma$ the parallax uncertainty and $\varpi$ the observed parallax), has been demonstrated not to be appropriate enough. The main complication arises from the fact that the measured parallax is an stochastic variable with an associated error, a noisy measurement of the true parallax. In fact, these observational errors can lead to undesired effects when estimating the distance to the stars as $\frac{1}{\varpi}$. On the one hand, the error can cause a measured negative parallax, and this a physically meaningless negative value of $r = \frac{1}{\varpi}$. Removing these values from the sample in order to leave only positive parallaxes will make it biased, as shown in [2] and [3]. More in general, the statistical behavior of $r = \frac{1}{\varpi}$ has undesired properties, including an asymmetric error distribution and a bias. The need of improvement is thus mandatory, so alternative methods, both frequentist and Bayesian, have been suggested by several authors.

The main goal of the present study is twofold. On the one hand, we aim to explore the performance of three distance estimators (the Bayesian method with Uniform Distance and Exponentially Decreasing Space Density Priors Eqs. (3) and (4), and the Transformation Method Eq. (6)), following the recommendation of the members the Gaia DPAC (Data Processing & Analysis Consortium), after the Sitges meeting of 01/2017 [4]. Consequently, our intention is to provide a recommendation about a better method of estimating distances for the next Gaia's data release. On the other hand, we also have an educational purpose, so through the development of several tools we aim to rise awareness about the practice of unquestioningly inverting the parallax [5].

In the second section, the theoretical basis of the two Bayesian Methods suggested by Coryn A. L. Bailer-Jones and the frequentist one described by Haywood Smith are discussed. In the third one, the results of the performance of the different distance estimators are detailed. Although the full study has been developed for both distance and distance modulus, we will only present the whole development and results for the estimation of distances for the sake of briefness. The whole development for the case of the distance modulus can be found in the tutorial *Distance and Distance Modulus Estimator Tool* which will be included in the documentation of the Gaia Archive DR2[1].

*Electronic address: aribesme8@alumnes.ub.edu

[1] Gaia Archive: https://gea.esac.esa.int/archive/

## II. THEORETICAL BASIS

The trigonometric parallax $\varpi$ is a measured variable, a noisy measurement of the true parallax $\varpi_{true}$, that consequently, has an associated formal error $\sigma$. The measured parallax does not need to coincide with the true one, which satisfies $\varpi_{true} = \frac{1}{r}$, being $r$ the true distance of the object.

In this section we will present the recommended methods [4] used in order to estimate the distance $r$ and the distance modulus $\mu$ using the trigonometric parallax $\varpi$.

The distance modulus is defined as:

$$\mu = m - M = 5(\log_{10} r - 1) \tag{1}$$

with $m$ the apparent magnitude, $M$ the absolute magnitude and $r = 1/\varpi_{true}$ the true distance in parsecs. The distance modulus is widely used for the calculation of absolute magnitudes from parallaxes. In neither the case of the distance nor the distance modulus, the relation $r = \frac{1}{\varpi}$ (with $\varpi$ the observed parallax) can be used unquestioningly.

The theoretical development in the distance modulus case can be found in the Jupyter Notebook tutorial *Distance and Distance Modulus Estimator Tool* which is included in the documentation of the Gaia archive DR2[1].

### A. Bayesian methods

Bayesian methods allow us to infer the distance of an object through a model error (consequence of the distribution of photons in the detectors, calibration errors, processing errors, etc.) and *a priori* assumption, the prior. In the case of Gaia parallaxes the model error behaves approximately as a Gaussian distribution of mean $1/r$ and a formal error $\sigma$ as a standard deviation of the distribution [6]. Thus, we can write the pdf of the error distribution as

$$P(\varpi|r,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(\varpi - \frac{1}{r}\right)^2\right) \tag{2}$$

where $P(\varpi|r,\sigma)$ expresses the probability of observing a parallax $\varpi$ with an associated formal error $\sigma$ given a true parallax $1/r$.

Since we dispose of the observed parallax but not of the true distance, we infer the probability distribution function (pdf) of the real distance through the application of Bayes Theorem $P(r|\varpi,\sigma) = P(r)P(\varpi|r,\sigma)$, where $P(r|\varpi,\sigma)$ expresses the probability of finding a true distance $r$ given the observed parallax $\varpi$ and its associated error $\sigma$, and $P(r)$ is an *a priori* assumption about the true distance distribution of the sample. We implement two simple, uninformative and unnormalized priors (Eq. (3) and Eq. (4)), described by Coryn A. L. Bailer-Jones in [6] for stellar objects in our galaxy.

The Uniform Distance Prior (UDP)

$$P_{UD}(r) = \begin{cases} \frac{1}{r_{lim}} & \text{for } 0 < r \leq r_{lim} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

assumes that the probability of finding a stellar object is constant until a certain distance limit and drops to 0 for more distant objects. It also imposes that there are not negative distances.

The Exponentially Decreasing Space Density Prior (EDSDP)

$$P_{EDSD}(r) = \begin{cases} \frac{1}{2L^3} r^2 e^{-\frac{r}{L}} & \text{for } r > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

assumes a spherical star distribution where the probability of finding a star decreases exponentially with the distance.

Based on the resulting pdfs (obtained from applying the Bayes theorem using these priors) we will define two estimators of the distance: The mode and the median of the pdfs. The mode of the distribution is the maximum probability value, derived at [6] for the two resulting pdfs. The median is the 50% quantile, i.e., the distance value separating the higher and the lower half of the probability distribution. We also compute it as a distance estimator, although in Bayesian methods we are usually interested in the performance of the mode. In addition to the estimation of distance, we also need to provide an associated finite uncertainty interval that does not include negative distances. In order to do this, we have chosen a 90% uncertainty interval following the lead of Coryn A. L. Bailer-Jones [6], which is given computing the 5% and the 95% percentiles. The median, and the 5% and 95% quantiles are computed solving the implicit equation on $x$

$$\frac{1}{N}\int_{r_0}^{x} dr P(r) P(\varpi|r,\sigma) - p = 0 \tag{5}$$

where $N$ is the normalization factor, $P(r)$ refers to both the described priors (Eqs. (3) and (4)), $P(\varpi|r,\sigma)$ stands for the error distribution Eq. (2) and $p = 0.05, 0.5, 0.95$ to the normalized percentiles. The lower integration limit $r_0 = 0.001$ kpc has been chosen to be superior to 0 to avoid numerical divergences.

With respect to the distance modulus $\mu$ Eq. (1), we have applied the translation to the distance modulus space of the priors Eqs. (3) and (4) defined in distance space, using the Jacobian of the transformation in order to derive the pdf of the distance modulus and be able to extract the expression of the mode, the median and the 90% uncertainty interval.

### B. Transformation Method

The Transformation Method is a frequentist method described by Haywood Smith in [7] and [8], which tries to find a transformation of the observed parallax that, once inverted, approaches better the true distance and corrects the problematic issues of the definition $r = \frac{1}{\varpi}$, i.e., the divergence of the distance for small $\varpi$ (the observed parallax) and the existence of negative distances. It is a completely frequentist method, since it tries to optimize several parameters in order to obtain a transformation that provides better results over a given sample of stars.

We have used the estimate $r^*$ described in [8]

$$r^* = \frac{1}{\varpi^*} \tag{6}$$

with $\varpi^* = \beta\sigma\phi g_\phi$, $\phi = \frac{1}{0.8}\ln(1 + e^{\frac{0.8\varpi}{\sigma}})$ and $g_\phi = 1$ if $\varpi > 0$ and $g_\phi = e^{-0.605\frac{\varpi^2}{\sigma^2}}$ for $\varpi \leq 0$, with $\beta = 1.01$.

In order to estimate an associated uncertainty interval to this quantity we compute $\varpi^*(\varpi + 2\sigma)$ as the inferior bound and $\varpi^*(\varpi - 2\sigma)$ as the superior bound. The performance of this uncertainty interval has also been tested.

We also apply the Transformation Method in the case of the distance modulus $\mu^*$,

$$\mu^* = m - \hat{M} = -(5\log(\hat{\varpi}) + 5) \tag{7}$$

that we define using the expression of the transformed absolute magnitude $\hat{M}$ and of the transformed parallax $\hat{\varpi}$, provided by Haywood in [7] and being $m$ the apparent magnitude.

## III. RESULTS AND DISCUSSION

In order to estimate the distance and the distance modulus with their uncertainty intervals, using the methods described in section II, we have developed a Python module named *pyrallaxes*, a Tkinter GUI called *DistanceEstimatorApplication* which estimates the distance and distance modulus for a single star and provides the plots of the pdfs, a tool that computes the distance given an input file *DistanceEstimatorTool* and an interactive Jupyter Notebook tutorial, *Distance and Distance Modulus Estimator Tool*. We validated the produced tools using the Gaia mock catalog of Red Clump stars [9], comparing the results on the sample with the Java implementation produced independently by Tri L. Astraatmadja and Coryn A.L. Bailer-Jones [6] and adapted for our usage by E. Utrilla. Both implementations match up to the thousandth of parsecs.

We have estimated the mode, the median and the 90% uncertainty interval of the distance and the distance modulus of the UD and EDSD Bayesian methods, as well as the estimates $r^*$ and $\mu^*$ with their associated uncertainty intervals, using the described tools over a sample of about $10^7$ randomly chosen objects over a simulation of Gaia DR2, that will be available in the article [10]. We have also computed the distance and distance modulus resulting of inverting the parallax $r_{inv} = \frac{1}{\varpi}$, with $\varpi$ the observed parallax and uncertainty interval $[r_{inv} - 2\sigma, r_{inv} + 2\sigma]$. The true fractional parallax error $f_{true} = \frac{\sigma}{\varpi}$ of the sample ranges between $16 \cdot 10^{-4}$ and 94.

We compare the performance of the different estimators computing the dimensionless quantity

$$x_i = \frac{r_i - r_{true,i}}{r_{true,i}} \tag{8}$$

that provides the bias ratio of the estimate $r_i$ with respect to the true distance $r_{true,i}$ for every element of the sample. We divide the whole sample in $M$ bins of $f_{true}$ in which we compute the mean bias per bin $\bar{x}_j$ as

$$\bar{x}_j = \frac{1}{n_j}\sum_{\forall i \in j} x_i \tag{9}$$

the root mean square (r.m.s.) per bin $\bar{x^2}_j^{1/2}$ as

$$\bar{x^2}_j^{1/2} = \sqrt{\frac{1}{n_j}\sum_{\forall i \in j} x_i^2} \tag{10}$$

and the standard deviation per bin $\sigma_{x,j}$

$$\sigma_{x,j} = \sqrt{\frac{1}{n_j - 1}\sum_{\forall i \in j}(x_i - \bar{x}_j)^2} \tag{11}$$

that indicates the dispersion of the quantities $x_i$ around the mean bias of the bin $j$, $\bar{x}_j$ . In all three expressions, $n_j$ is the number of elements in bin $j$ and $j$ ranges from 0 to $M$, with $M$ the total number of bins.

The mode and the median of the Uniform Distance (UD) and of the Exponentially Decreasing Space Density (EDSD) Bayesian methods, as well as the inverse of the observed parallax (INV) $1/\varpi$ and the Transformation method (TM) estimate $r^*$ are shown in Fig. 1. In Figs. 1.a and 1.b the bias ratio with $M = 500$ bins in the range of $f_{true} \in [0, 100]$ and $f_{true} \in [0, 1]$ are represented, in Fig. 1.c the bias ratio with $M = 300$ bins for $f_{true} \in [0, 0.3]$ is plotted.

Computing the bias ratio Eq. (9), the root mean square Eq. (10) and the standard deviation (Std. Dev.) Eq. (11) of the whole sample Table I (for $i = j = N$, with $N$ the total number of samples) or a given subsample of $f_{true}$ Tables II and III for $f_{true} \leq 1$ and $f_{true} \leq 0.3$ respectively, allow us to compare the global goodness of the estimators.

Globally, we can observe in Fig. 1.a and Table I, that the best performance over our simulated sample of Gaia DR2 corresponds to the EDSD method, since

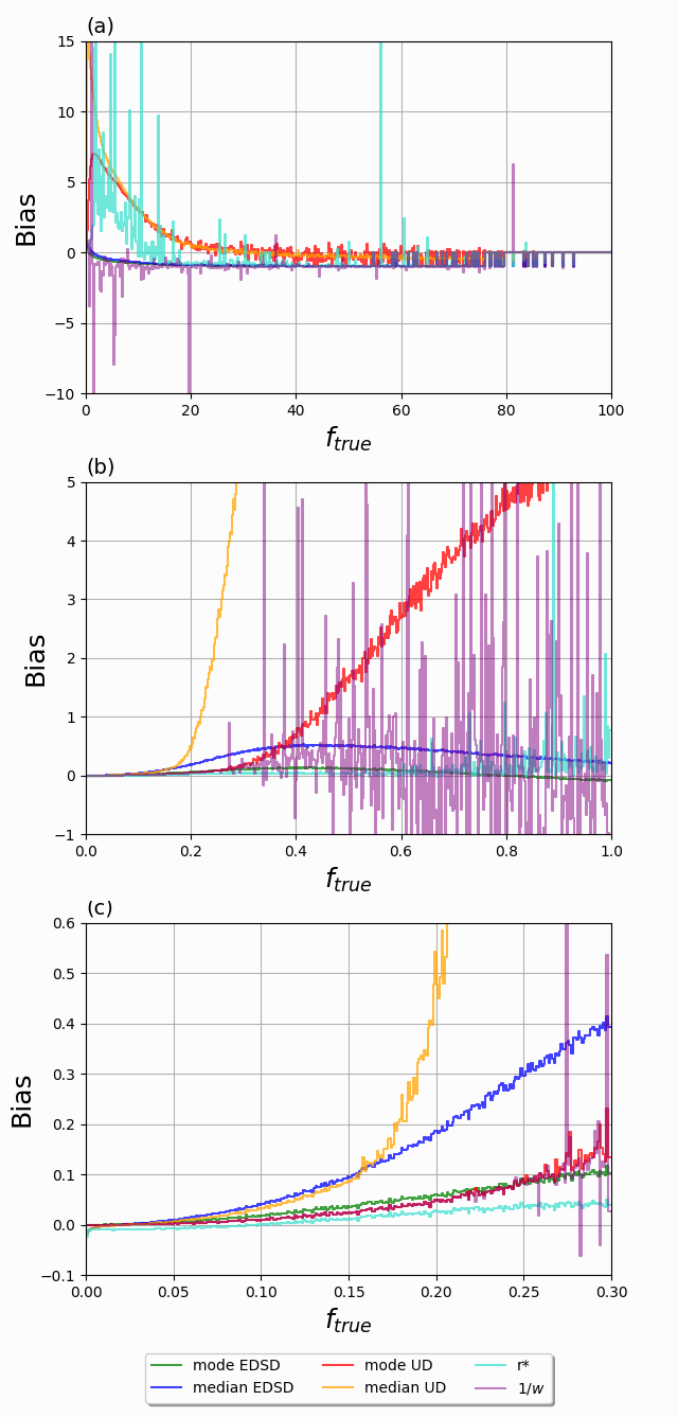


FIG. 1: The bias ratio as a function of the true fractional parallax error $f_{true}$. In (a) the whole simulated sample of Gaia DR2 with $10^7$ stars have been arranged in 500 bins of $f_{true}$, in (b) around $4.37 \cdot 10^6$ samples with $f_{true} \leq 1$ have been arranged in 500 bins and in (c) around $1.9 \cdot 10^6$ samples with $f_{true} \leq 0.3$ have been arranged in 300 bins. The given bias per bin corresponds to the mean of all the samples in the same bin of $f_{true}$. The bias range has been cut in all three cases in order to show more relevant information, the bias ratio goes from -40 to 1540 in (a) case, from -370 to 220 in the (b) case and from -0.1 to 6 in the last case.

| | | Bias | r.m.s | Std. Dev |
|---|---|---|---|---|
| EDSD | mode | -0.2144435 | 0.4676504 | 0.4155850 |
| | median | -0.0012707 | 0.5645264 | 0.5645250 |
| UD | mode | 4.1649385 | 10.4278158 | 9.5599497 |
| | median | 8.3483038 | 12.1676157 | 8.8519321 |
| TM | $r^*$ | 3.2090449 | 3233.2025951 | 3233.2011634 |
| INV | $1/\varpi$ | 1.4505190 | 9630.4302163 | 9630.4305861 |

TABLE I: Around $10^7$ Gaia DR2 simulated samples have been used.

| | | Bias | r.m.s | Std. Dev. |
|---|---|---|---|---|
| EDSD | mode | 0.0523262 | 0.4067274 | 0.4033475 |
| | median | 0.2884500 | 0.6994112 | 0.6371599 |
| UD | mode | 1.5656870 | 8.2809021 | 8.1315422 |
| | median | 7.7665353 | 14.4385161 | 12.1717587 |
| TM | $r^*$ | 0.1061284 | 88.5215926 | 88.5215391 |
| INV | $1/\varpi$ | -0.1867738 | 1287.9354960 | 1287.9356297 |

TABLE II: We have selected a DR2 Gaia subsample of around $4.37 \cdot 10^6$ samples with $f_{true} \leq 1$.

both the mode and the median present the smallest bias, standard deviation and root mean square. The EDSD method tends to systematically underestimate the predicted distances (negative bias) with small dispersion. The median EDSD is the less biased estimator, although it presents higher dispersion than the mode EDSD. We can also observe that the inverted parallax and the estimate $r^*$ present the next smaller positive bias, meaning that these methods tend to overestimate the distance, but with a huge dispersion. These two methods tend to strongly overestimate and underestimate the distance for different ranges of $f_{true}$. Even though the positive and negative bias can compensate globally in the sample, the estimated distance can differ with the real one in more than a 5000%. Finally, both the mode and the median of the UD method present the highest bias but with smaller dispersion than both the inverted parallax and Transformation methods.

In the case of the subsamples with small fractional parallax error $f_{true}$, presented in Tables II and III and Figs. 1.b and 1.c, we can observe that in the range $f_{true} \leq 1$ the mode EDSD presents the smallest global bias and dispersion. It seems really interesting to highlight the change of behavior of the Transformation Method estimate $r^*$ in the range $f_{true} \in [0.1, 0.7]$, in which it presents the smallest bias and dispersion, although it increases outside of this range, even surpassing the bias of the mode UD. In Fig. 1.c we can observe that for small $f_{true}$, that is, for very precise parallaxes, all the estimators have a similar good behavior, as one could expect. For instance, the global bias of the mode for the UD, EDSD and inverted parallax methods are around the 4%, for $f_{true} \leq 0.15$ all the estimators present a bias inferior to the 10%.

| | | Bias | r.m.s | Std. Dev. |
|---|---|---|---|---|
| EDSD | mode | 0.0378425 | 0.2082295 | 0.2047621 |
| | median | 0.1191844 | 0.3705107 | 0.3508181 |
| UD | mode | 0.0350620 | 0.4728545 | 0.4715529 |
| | median | 0.7533940 | 4.9125000 | 4.8543863 |
| TM | $r^*$ | 0.0113476 | 0.1838725 | 0.1835220 |
| INV | $1/\varpi$ | 0.037587 | 6.1906233 | 6.1905108 |

TABLE III: We have selected a DR2 Gaia subsample of around $1.9 \cdot 10^6$ samples with $f_{true} \leq 0.3$.

As it has been explained in section II.B, the Transformation method is a frequentist method, meaning that the numerical parameters used for adjusting the estimator $r^*$ have been chosen according to a specific sample. The change in the behavior of the Transformation method indicates that the sample used for defining $r^*$ can be similar to the Gaia DR2 simulation sample in the range $f_{true} \in [0.1, 0.7]$, but different outside it, so the adjusted parameters are not longer suitable.

## IV. CONCLUSIONS

In this work we have compared the performance of three different methods, two Bayesian methods and a frequentist method with the result of inverting the parallax. We have used the developed tools to estimate the mode and the median of the Uniform Distance and of the Exponentially Decreasing Space Density for the distance and the distance modulus, as well as the 90% uncertainty interval associated to these quantities and the Transformation method estimates $r^*$ and $\mu^*$ with their associated uncertainty intervals. We have compared the bias, the root mean square and the standard deviation of these estimates with the inverted parallax as a function of the true fractional parallax error $f_{true}$ for the whole range of $f_{true} \in [0, 100]$, for $f_{true} \leq 1$ and $f_{true} \leq 0.3$. In the tables above, the mean bias, standard deviation and root mean square over all Gaia DR2 subsamples limited by $f_{true}$ are summarized.

The developed software, as well as the tutorial on its use, that will be available from April 2018 in the Gaia Archive [1] at DR2 section, allows to use arbitrary priors for the calculation of the estimation of distances and distance modulus. As an improvement, more specific and complex priors can be developed in the framework of *pyrallaxes*, for instance, priors including the photometric data. In order to use these tools we would strongly recommend to analyze the sample beforehand and create the most suitable prior for each sample. Nevertheless, if a faster method is sought, we can recommend to use the mode or the median of the Bayesian method with the Exponentially Decreasing Space Density Prior, which improves the estimation of distance from the trigonometric parallax for the Gaia Data Release 2. In any case, we would not recommend the usage of the Transformation method, since it is only well behaved for a given range of $f_{true}$ and the parameters used for defining the transformation have been adjusted empirically with a specific sample that does not need to match with Gaia DR2.

### Acknowledgments

Special thanks to my supervisor Xavier Luri for his patient guidance and advice, as well as to Cesca Figueras and Carme Jordi for giving me this great opportunity, to all the people who has aided me in this work, Alfred Castro, Enrique Utrilla, Mercè Romero, Cecilia, Carine Babusiaux, Frédéric Arenou, Coryn Bayler-Jones, Raul Borrachero and Sergio Soria. Thank you very much to my family and friends, and specially to my parents, my sister and David, I wouldn't have managed so far without you.

[1] ESA Science & Technology: Gaia. Gaia/Data Release 1.
[2] Haywood Smith. Is there really a Lutz-Kelker bias? Reconsidering calibration with trigonometric parallaxes. *Royal Astronomical Society*, January 2003.
[3] P.Teerikorpi A.G. Butkevich, A.V. Berdyuin. Statistical biases in stellar astronomy: the Malmquist bias revisited. *Royal Astronomical Society*, July 2005.
[4] E. Utrilla. Parallax estimation tools. High level description. Technical report, DPAC, January 2017.
[5] X. Luri. Distance estimation from parallax. Results and recommendations for GST, May 2017.
[6] Coryn A. L. Bailer-Jones. Estimating distances from parallaxes. *The Astronomical Society of the Pacific*, 127:994–1009, 2015.
[7] Jr Haywood Smith and Heinrich Eichhorn. On the estimation of distances from trigonometric parallaxes. *Royal Astronomical Society*, 281:211–218, 1996.
[8] Jr. Haywood Smith. Transformation methods for trigonometric parallaxes. *Kluwer Academic publishers*, pages 1–12, 2001.
[9] M.Romero-Gómez, F.Figueras, H.Abedi T.Antoja, and L.Aguilar. The analysis of realistic Stellar Gaia mock catalogs.i. Red Clump stars as tracers of the central bar. *Monthly Notices of the Royal Astronomical Society*, 2014.
[10] X. Luri, A. Brown, L. Sarro, F. Arenou, C. Babusiaux, C.A.L. Bailer-Jones, J.de Bruijne, A. Castro-Ginard, and T. Prusti. On the proper use of Gaia parallaxes, distances, distance modulus and tangential velocities from Gaia astrometry. Expected publication at Astronomy and Astrophysics on April 2018.