Bayesian stellar distances from parameter inference

Author: Billy Harrison

Abstract

Estimating stellar distances from parallax measurements becomes increasingly challenging in the low signal-to-noise/high uncertainty regime, where naive inversion leads to biased and unstable results. In this work, we investigate Bayesian distance inference using a sample of 10,000 Gaia DR2 stars within 1 kpc, comparing progressively more informative posterior models. We first implement a parallax-only Bayesian framework with a physical distance prior and assess its performance relative to simple parallax inversion. We then extend the model to a full posterior that jointly infers distance and absolute magnitude by incorporating photometric, spectroscopic, and astrometric information. Posterior distance estimates are compared against Gaia GSP-PHOT distances using both visual diagnostics and quantitative performance metrics. We find that the full posterior produces more constrained and symmetric distance distributions, significantly reducing fractional distance errors and log-distance scatter while improving rank correlation. These results demonstrate how incorporating observational constraints within a Bayesian framework improves stellar distance inference.

1. Introduction

Trigonometric parallax provides a direct geometric measurement of stellar distances. They are defined as an objects observed angular displacement with respect to a stationary frame of reference given the expected movement of the observer over a distance baseline. Stellar parallaxes are inferred using twice the distance from the Earth to the Sun (AU) as said baseline. Geometric construction and the choice of standard distance units define the distance of an object (star) from the sun d pc as $1/\varpi$ arcsec. This is a very good approximation for close, well-resolved stars (up tp $\sim$ 300pc or so). However, in the presence of measurement uncertainty, the naive estimator $d = 1/\varpi$ becomes biased and ill-defined.

The goal of this short report is to estimate distances and the associated uncertainties of a set of 10000 GAIA DR2 (Lindegren et al., 2018) stars from parallax alone. But, as the above inversion relation only holds strictly true when there are no measurement uncertainties, it is in our best interest to figure out how to infer the distances instead given parallax (and eventually other spectroscopic and astrometric) measurements, uncertainties, and any involved assumptions about the problem or the parameter space it occupies. This sort of problem is well-suited for probabilistic analysis, namely Bayesian methodology, adopted for instance by (Bailer-Jones, 2015). This involves building a likelihood and prior in order to combine into a posterior probability distribution function in distance space - from which we can infer relevant statistics.

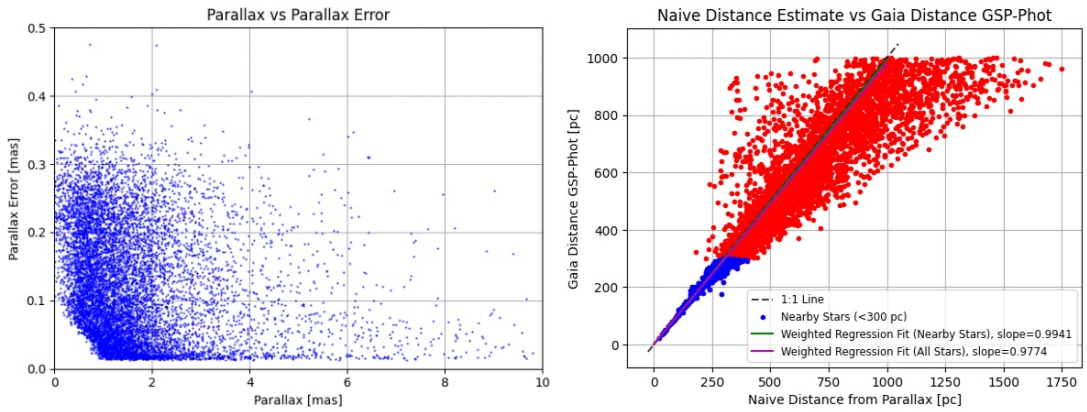

Figure 1: *Left*: Parallax error σ_{ϖ}/ϖ vs parallax ϖ for the full GAIA DR2 10000 star sample. *Right*: GAIA distances vs naive distances inferred from parallax. Nearby (blue) stars converge between both distance estimates, whereas further stars tend to diverge, with the naive estimates overestimating compared to the GAIA data. This is reflected in the weighted regression fits for both populations.

Previous studies in the literature have looked at estimating physical quantities from parallax in tandem with additional data (e.g. colour, magnitude). Example studies include (Palmer et al., 2014)'s ML approach to cluster distance estimates, and (Verbiest et al., 2012)'s pulsar distance study. The objective here is to delve into issues surrounding maybe the simplest application of these problems - inferring distance from parallax - and comparing our varying results via different methodology with the (Lindegren et al., 2018) GSP-PHOT distance estimates provided by the data set. Hence we will project the distance posterior PDFs for each star into the relevant data space as a function of uncertainty, and cross reference with the so-called "expected" values.

2. GAIA DR2

The GAIA Data Release 2 (DR2) catalogue provides astrometric measurements for $\sim$1.3 billion sources, with typical parallax uncertainty on the order of milliarcseconds for nearby bright stars. Here we initially sample parallax ϖ, parallax uncertainty σ_{ϖ}, as the primary observables for distance inference. It must be emphasised that the GSP-PHOT distance estimates d used for model comparison throughout the work are inferred quantities, whereas parallaxes are direct stellar observables - hence the need for a statistical treatment.

We select a random sample of 10000 stars with an imposed distance (GSP-PHOT) of d ¡ 1kpc - we want to focus on regions where parallaxes have high Signal-to-Noise (SNR) and where distance priors do not extremely skew our Bayesian estimates compared to naive ones. Negative parallaxes are also discarded for ease of simplicity within the models used. Random sampling avoids spatial or magnitude-driven biases on distance compared to a sky-specific sample. For example, targeting cluster stars or stellar streams would introduce a selection bias and would hence over-represent specific sources relative to the grand array available. Imposing such a selection function would in combination with potential more advanced priors (absolute CMD distribution, isochrones) cause the sample to be a function of sky position and distance (Sloan Digital Sky Survey (SDSS) Collaboration, 2025). Additionally, the 1kpc cut reduces the impact of very small ϖ and their associated large fractional uncertainties σ_{ϖ}/ϖ.

Figure 1 (left) describes the trend of parallax uncertainty - stars cluster at relatively low par-

allax ($\lesssim$ 2mas) and increasing uncertainty ($\lesssim$ 0.3mas). This seems to reflect GAIA's astrometric precision and the selection effect of an increasing number of stars at large distance (low parallax) within the sampled volume. The absolute uncertainties remain small when the corresponding fractional uncertainties increase as parallax decreases (σ_ϖ/ϖ increases as ϖ, the denominator, decreases). Hence simple naive distance estimates (d = 1 / ϖ) become increasingly biased and noisy in this regime. This stands as a motivating argument to employ Bayesian distance estimators that explicitly account for parallax uncertainty and prior physical information. On the inverse, nearby stars ($\sim$ 300pc or $\sim$ 3.3mas) have large parallaxes and smaller fractional errors, we can see that the naive distance estimate works well in this regime. The weighted regression statistics (Figure 1, right) agrees with this interpretation, showing near agreement (r $\simeq$ 0.994) for close stars - parallax measurement is precise - and the degradation of the fit (r $\simeq$ 0.977) over the full sample reflects increasing parallax bias and scatter at larger distances. These empirical trends further motivate the use of Bayesian methods, which should remain well-behaved in the low SNR regime.

3. Bayesian Framework

In order to estimate parameter distributions, we need to formulate a Bayesian framework. This allows for proper propagation of measurement uncertainties, the incorporation of physically motivated background prior information, and produces well-defined distance estimates even for small parallaxes - compared to naive estimates which do not encode uncertainty.

3.1. Prior

We adopt an exponentially decreasing space density prior,

$$P(d) \propto d^2 \exp\left(-\frac{d}{L}\right), \tag{1}$$

where $L > 0$ is a scale length, as described by (Bailer-Jones, 2015). Here, it is important that distance remains strictly positive, and it is worth noting that stellar density is not isotropic, i.e. not uniform along the line of sight. Hence, L varies as a function of galactic position $L(l, b)$ according to a predetermined simplified piecewise model defined as:

$$L_{lb}(l,b) = \begin{cases} 800, & |b| < 30 \\ 550, & 30 \leq |b| < 35 \\ 400, & 35 \leq |b| < 40 \\ 300, & 40 \leq |b| < 45 \\ 250, & |b| \geq 45. \end{cases} \tag{2}$$

This reflects the vertical structure of the disk and the expected variation of stellar distances observed by GAIA within. Although this does not fully explore the distribution along the line of sight given by various Galactic models, it is more descriptive than an isotropic prior (fixed L). As for the rest of the prior, we can explain each term with physical motivation. d^2 encodes a volume element in spherical coordinates and alone implies more stars at larger distances. However the exponential decay factor $e^{-d/L}$ drops asymptotically towards zero as $d \to \infty$ thanks to the finite scale length of the Galactic stellar distribution. Figure 2 shows the shape of the physically motivated distance prior P(d). It is clear that the prior behaves differently for nearby

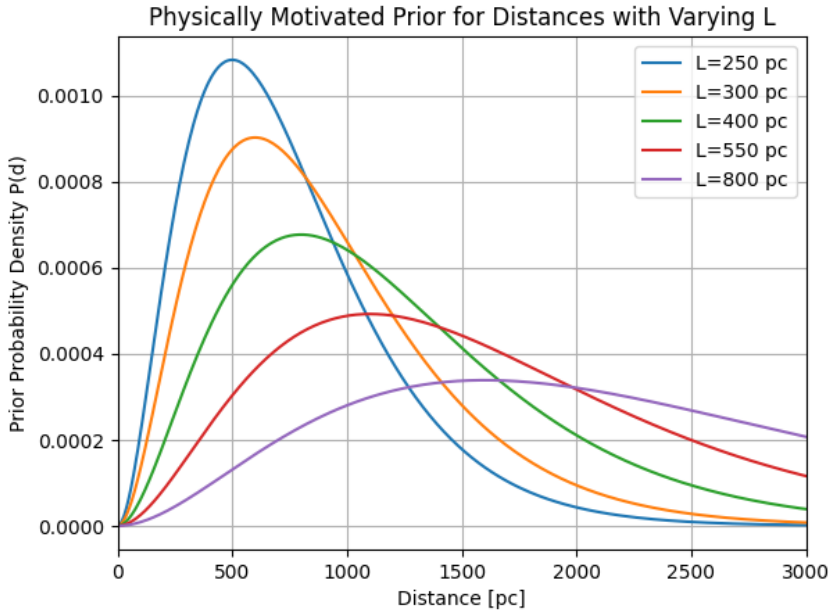

Figure 2: Fit of our physically motivated distance prior for varying scale lengths L = [250, 300, 400, 550, 800]pc

vs distant stars, and implies a preferred distance before even observing the data. It peaks at finite distance (as opposed to at 0 or ∞), with the peak shifting depending on the scale length.

3.2. Likelihood

Given GAIA parallax measurements are well approximated by Gaussian uncertainties, we can model the parallax ϖ with standard deviation σ_ϖ as:

$$L(\varpi \mid d) = \frac{1}{\sqrt{2\pi}\,\sigma_\varpi} \exp\left[-\frac{1}{2}\left(\frac{\varpi - \varpi_{\text{zp}} - 1000/d}{\sigma_\varpi}\right)^2\right] \tag{3}$$

where ϖ_{zp} is the global parallax zeropoint, determined as -0.029 Lindegren et al. (2018). This likelihood models the probability of observing a parallax ϖ given an unknown true distance d - converted to a noisy measurement of 1000/d (factor of 10^3 due to conversion of arcsec to milliarcsec). It effectively gives a probability density function (PDF); probability per unit parallax for any given measurement of d and σ_ϖ to infer a distribution in ϖ. The likelihood is Gaussian in ϖ but not in d - when transformed into distance parameter space, it becomes skewed and asymmetric. This asymmetry increases as a function of parallax SNR or parallax uncertainty. Strictly speaking, this likelihood only applies at larger distances (at extremely small distances, the inverse parallax relationship breaks down).

3.3. Posterior Inference

We now have a PDF distribution of parallax given distance and prior information on the stellar distance distribution observed by GAIA for our $d < 1000$pc cut. But the quantity we want to infer is in fact distance d. As measurements of such quantities are inherently noisy, it is difficult to infer d exactly - it is easier, as with the case with the likelihood, to instead infer the probability distribution over the possible values of distance. In other words we want the posterior PDF

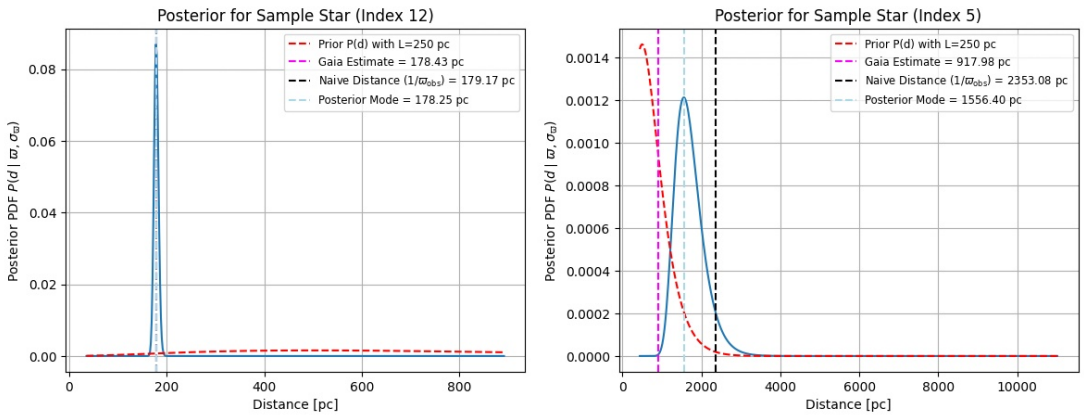

Figure 3: *Left*: Distance posterior for a close star, and *Right*: posterior for a more distant star, Both are taken from the GAIA DR2 dataset and the exponentially decreasing prior is shown for both regimes of scale length.

P(d | ϖ, σ_ϖ) - the probability per unit distance for any given measurement of parallax ϖ, with parallax uncertainty σ_ϖ.

The posterior follows from Bayes' theorem, where we can express

$$P(d \mid \varpi, \sigma_\varpi) = \frac{1}{Z} P(\varpi \mid d, \sigma_\varpi)\, P(d) \tag{4}$$

where Z is a constant of normalisation

$$Z = \int_0^\infty P(\varpi \mid d, \sigma_\varpi)\, P(d)\, \mathrm{d}d. \tag{5}$$

The posterior is obtained by essentially multiplying the likelihood by the prior, transforming a relation for the probability distribution of the known data ϖ given an unknown parameter d, to an interpretation of the desired quantity for the probability of said quantity given the data (parallax). The properties of this posterior include that it is always proper, or normalisable, and is either unimodal or bimodal - unimodal for the vast majority of cases.

4. Initial Results

Here we observe the resulting posteriors for two example stars from our 10000 star GAIA sample, one with a "close" estimate on GSP-PHOT distance, and one with a "distant" estimate. The posteriors for both stars, along with associated priors, naive distance estimates, GAIA estimates, and mode statistics, are shown in Figure 3

We can immediately see the relative success of the posteriors predictive power in both regimes. For nearby stars, the likelihood is narrow and informative, whereas the prior holds little influence. The posterior peak aligns with the naive inverse-parallax estimate and the GAIA GSP-PHOT distance, implying that all estimators agree within uncertainties. However, in the distant star regime, the likelihood is broader and slightly asymmetric, and the prior shapes the posterior to a significant extent. The naive distance estimate is a gross overestimation compared to the posterior mode, and the GSP-PHOT distance is lower than both. These cross-methodology incompatibilities for distant stars can be explained by the compromise proposed by the posterior between the asymmetric long-distance tail of the parallax likelihood and the exponentially

decreasing space-density prior. The prior suppresses un-physically large distances while also retaining a volume-weighted (d^2) component which biases against very small distances. The distinct direction-dependent prior is not as vigorously implemented as those imposed on the GAIA catalogue estimates, which are population-level and are optimised for all-sky performance, leading to weaker constraints on small distances along particular lines of sight. The posteriors disagreement for distant stars implies prior, measurement uncertainty, and Galactic structure sensitivity within this regime, which is unaccounted for with our basic inverse parallax distance estimates.

Particularly at smaller parallaxes, posteriors tend to be non-gaussian for GAIA distances, and are often skewed. Hence, mean, median, and mode do not often coincide. As briefly mentioned above, this is due to the inverse transformation between d and ϖ which produces sharp cutoffs at small distances and long tails towards large ones (partially suppressed by the prior but does not completely remove it). As a result, posterior distance PDFs are typically unimodal and positively skewed, see Fig 3, right. We can evaluate the posteriors with these different statistics.

The mode $d_{mode} = argmax P(d \mid \varpi, \sigma_\varpi)$ corresponds to the maximum likelihood of the prior. This is closely aligned with the most physical distance and less sensitive to the distance tails, however less stable if multimodal (rare in these cases), and can be sensitive to small-scale posterior structure. The mean, or the expectation value $\langle d \rangle = \int_0^\infty d\, P(d \mid \varpi, \sigma_\varpi)\, \mathrm{d}d$ is, however, strongly influenced by the tail and can lie in regions of low PDF density, especially for skewed posteriors that are weakly constrained by the parallax data. The median satisfies $\int_0^{d_{\mathrm{med}}} P(d \mid \varpi, \sigma_\varpi)\, \mathrm{d}d = 0.5$ such that half the posterior PDF lies on either side. This is useful as it is more sensitive to longer tails and is well-defined for normalised posteriors.

Throughout the rest of this work, we report both posterior modes and medians. The mode will cover the most "probable" distance estimate for individual sources, and the median provides a comparison statistic across the full GAIA DR2 sample due to its insensitivity to the asymmetric tails. For nearby stars, these statistics are roughly invariant, but for distant stars with skewed posteriors, the differences may report valuable inference on the Bayesian framework's ability to provide analysis.

Figure 4 shows some of these ensemble statistics across the data set. The top panels plot distance estimates against GAIA GSP-PHOT. The mode sees tight correlation at small distances, with scatter increasing vertically, i.e. there is a tendency for the posterior mode to overestimate relative to GAIA. For high parallax SNR (low σ_ϖ/ϖ and distance), the likelihood expectedly dominates and the estimators converge, in comparison to further afield where the posterior is more prior-sensitive. The median shows a slightly tighter correlation at low d and reduced upward scatter above. It is more catalogue-consistent with GAIA GSP-PHOT which are typically median-based Bayesian estimates (Sloan Digital Sky Survey (SDSS) Collaboration, gai) and is less sensitive to tradeoffs between regions of likelihood-prior incompatibility. The plot of naive distance ($1000/\varpi$) acts as the control experiment which we weigh the other estimators against - it showcases similar systematic breakdown of the approximation at large distances, leading to overestimation.

The bottom right panels define a distance estimator for the posterior mode, $d_{mode}\varpi$, which should be equal to unity for an unbiased and resolved source. This is calculated for each star and plotted against σ_ϖ/ϖ, the right panel is a zoomed in representation. We observe that at high SNR ($\sigma_\varpi/\varpi << 1$), the values cluster around the unity line - as SNR decreases, the estimator drops systematically below 1 with a continuous deviation. As uncertainty increases, the Bayesian estimator moves distance inward, suppressing unphysical large-distance solutions. Even at moderate uncertainties (< 0.2), the posterior estimator departs significantly from inversion;

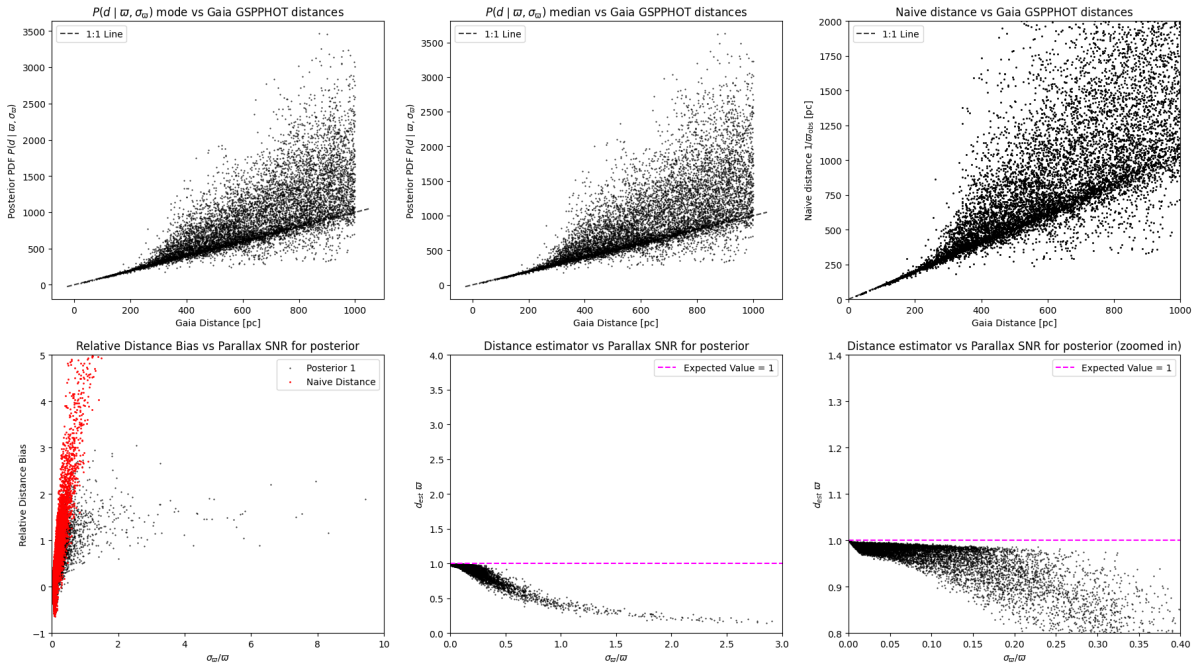

Figure 4: Initial posterior statistics. Clockwise from top left: posterior mode, posterior mean, naive inversion estimates vs GAIA, distance estimators and relative distance bias vs parallax SNR σ_ϖ/ϖ

.

this highlights again the non-linearity of the distance-parallax transformation. The bottom left panel shows distance bias $(d_{est} - d\ )\ /\ d$ against σ_ϖ/ϖ. For naive distances, we observe large positive bias with rapid divergence/extreme scatter at low SNR. Compared to the this, the Bayesian estimates show much smaller bias, and the scatter is consistent and controlled with less divergence. It trades variance for bias in a controlled way, underscoring the need for a Bayesian framework for distance inference from parallax measurements.

5. Posterior Extension

While parallax measurements alone provide direct and immediate geometric constraints on distance, their constraining power diminishes rapidly at high σ_ϖ/ϖ and larger distances. We have seen that the inverse parallax-distance relation transforms un-linearly, and that it breaks down as an estimator at intermediate distance (>300pc or so). Luckily, GAIA also provides photometric and colour information that encode individual distance tracers via the intrinsic luminosity associated with stars. In the following section, we will extend the parallax-only posterior into a multi parameter-space joint Bayesian framework that simultaneously infers distance and absolute GAIA-band magnitude M_G. Absolute magnitude, while not directly observed, can be inferred empirically, and directly links distance to colour and apparent magnitude. After the joint posterior is evaluated, it can also be marginalised out to provide distance-only probability distributions, and hence posterior mode and median statistics.

Assuming conditional dependence between astrometric, photometric, colour, and extinction measurements, the full extended posterior can be proposed as the product of several linked

likelihood and prior terms, again linked via Bayes' theorem:

$$P(d, M_G \mid \varpi, m_G, c, A) \propto L_{\varpi}\, L_{\text{phot}}\, L_{\text{colour}}\, P(A \mid d)\, P(d) \tag{6}$$

where m_G is apparent GAIA-band magnitude, c is BP-RP colour, and A is extinction magnitude.

L_{ϖ} is the unchanged parallax likelihood - Gaussian with included global zero-point correction. It is identical to the one used in parallax-only inference and measures ϖ as a Gaussian-distributed observation of the "true" parallax.

L_{phot} is the photometric likelihood which links distance and luminosity. We use the standard distance modulus with extinction: m_G = M_G + 5$\log_{10}$(d) - 5 + A. Given distance, absolute magnitude, and extinction, we can fully specify apparent magnitude up to measurement noise. For a given apparent magnitude, larger distances will require intrinsically brighter stars, coupling distance and M_G in the posterior. The likelihood manifests as a Gaussian with the residuals equalling the difference between the apparent magnitude inferred by distance and M_G, and the provided GAIA apparent magnitudes.

L_{colour} imposes a probabilistic relation between colour and absolute magnitude. This relation is fitted empirically using the data set itself - modelled as a polynomial fit which captures the intrinsic well-defined loci in colour-magnitude space which most stars reside on. The parameter transformation from magnitude to colour is then weighed against the actual data in the Gaussian residuals defined by the likelihood. This term reflects the astrophysical relation that stellar colour strongly constrains intrinsic luminosity, while allowing for intrinsic scatter.

$P(A \mid d)$ is the extinction prior as a function of distance. This encodes the expectation that dust column density decreases with distance, while retaining sufficient flexibility to accommodate local variations. We model extinction as an exponentially decreasing function of distance with scale height h = 150pc. This allows us to constrain estimates on distance given the GAIA extinction data at a given scale-height ratio. The distance prior itself, P(d), also remains unchanged in order to be consistent with the parallax-only posterior analysis.

Once we have the joint posterior in distance and absolute magnitude, we can regard M_G as a nuisance parameter and marginalise it out - ultimately we are only interested in distance estimates. this can be done numerically with the shorthand equation $P(d \mid \text{data}) = \int P(d, M_G \mid \text{data})\, \mathrm{d}M_G$. This improves distance estimates as photometry and colour constrain M_G, indirectly sharpening the posterior in distance space - they inform the distance estimate without requiring a single fixed luminosity assumption.

In the following section, we compare distance estimates (mode, median) obtained from the marginalised distance posterior with those derived above from the parallax-only inference framework. This should demonstrate the improved precision and reduced bias achieved by incorporating photometric and spectroscopic information into our previously astrometric-bound methodology.

6. Full Results

In this section, we compare distance estimates derived from the parallax-only posterior with the extended posterior and the naive distance estimates, using both individual stellar examples and population-level diagnostics. We start with the same example stars as before, one "close" and one "distant" according to their GAIA GSP-PHOT values.

Figure 5 shows the distance and absolute magnitude PDFs for both stars depicted in a corner-plot. The nearby star (left) has a distance marginal which is narrow and approximately symmetric, with the absolute magnitude also being similarly narrow. The joint distribution

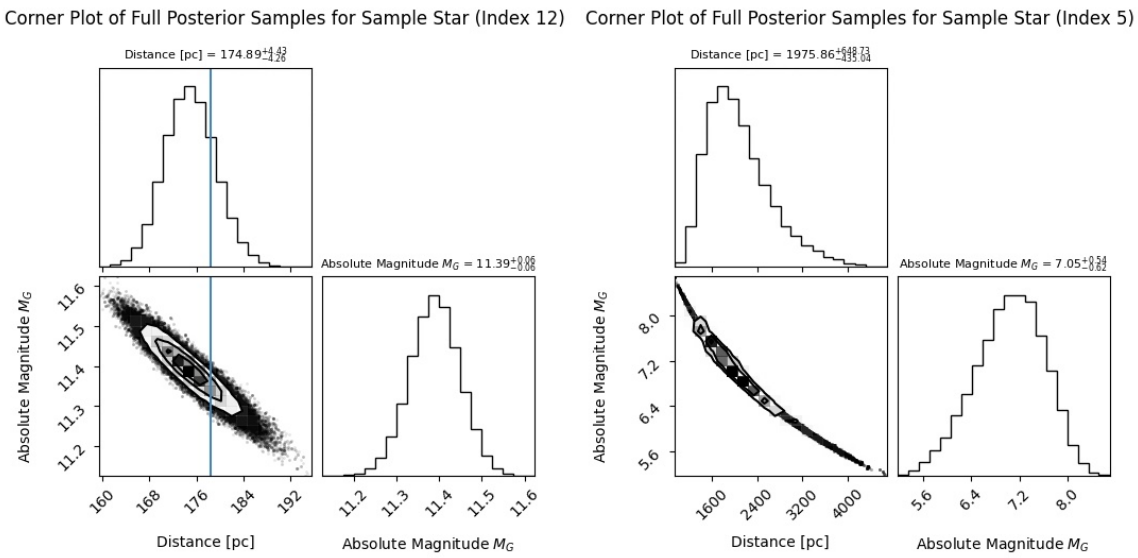

Figure 5: Corner plots displaying the MCMC sampled posteriors for both the close (*left*) and distant (*right*) example stars examined above. The blue line depicts the "true" distance value encoded by GAIA.

shows a moderate anti-correlation, forming a tight tilted ellipse. It is clear that at small distances, and in the high SNR regime, the parallax likelihood is already highly informative - the photometric/spectroscopic additions only serve to tighten the constraints to make the distribution more nearly Gaussian, and fully suppress the tail. Conversely, the distance marginal for distant star (right) is still asymmetric, as in the parallax-only case, but significantly more compact, the posterior occupying a thin curved loci in (d, M_G) space. Photometry constrains the combination of these parameters tightly, and the colour-magnitude relation in the likelihood further restricts the allowed wedge, so that the posterior is concentrated along a physically plausible distance-luminosity relation. This indicates the inference is driven by meaningful astrophysical constraints rather than arbitrary priors. The suppressed long-distance tail demonstrates the effects of photometry and colour at breaking degeneracies present in the previous parallax-only inference.

We expand the full posterior across the entire data set of 10000 GAIA sources and compute both the mean and median distances. Figure 6 show the modes (top panel) and medians (bottom) across the whole sample, with 68% and 95% confidence intervals shaded to depict spread. They are also compared to the mode and median distance estimates from the parallax-only posterior, and the naive inversion distances. Comparing said distance estimates reveals systematic differences between the central correlation with GAIA distances, and the structure of the uncertainty. As seen above, naive distances exhibit rapidly increasing vertical scatter above the regime of "close" stars, and strong positive bias at large distances, reflecting the non-linearity propagation of σ_{ϖ} under transformation. The explicit distance prior introduced by the parallax posterior regularises this behaviour producing stable but symmetric posteriors at low distances, and less stable posteriors with long tails at large distances - leading to distinct mode-median separation. This can also be seen in Figure 7. In contrast, the full extended posterior reveals tighter, more symmetric distance distributions with reduced confidence intervals and improved agreement with GAIA catalogue distances. The converging agreement between median and mode points to a substantial suppression of the long-distance tails and an agreement between prior-dominated and likelihood/data informed inference.

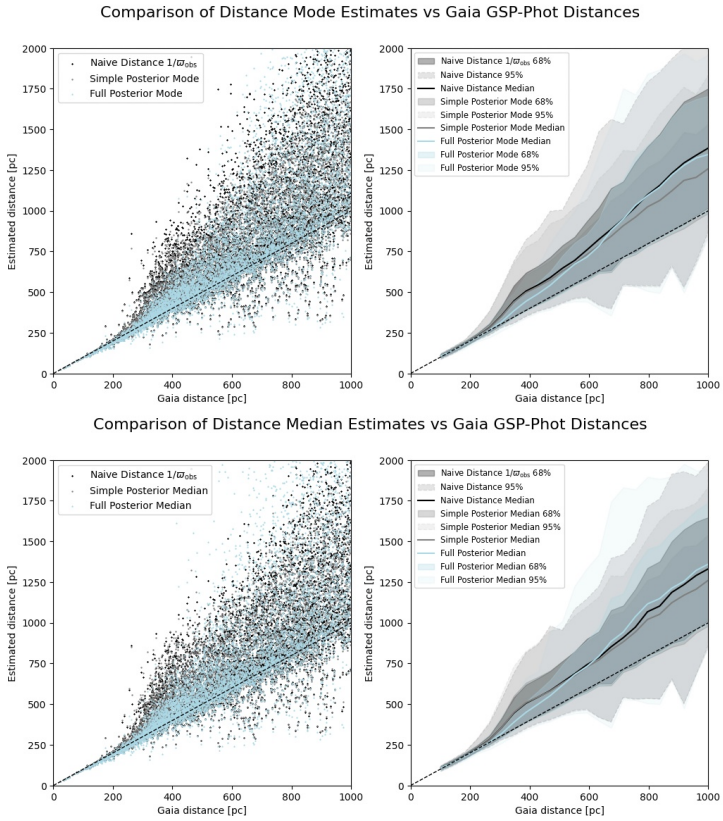

Figure 6: Full posterior modes (*top*) and medians (*bottom* across the full sample. Left plots show the raw data, right plots show the 68% and 95% shaded confidence intervals, as well as median best fits.

Figure 7 further illustrates the internal behaviour of both distance posteriors. The left panel shows distance bias $(d_{est} - d\,) \,/\, d$ against the GAIA catalogue distances as a baseline, showcasing that while the parallax-only posterior observes a scatter and mild shrinkage bias starting at around 300pc. Again, this is evidence that the prior is working as intended - suppressing implausibly large distances. The full posterior remains,to a certain extent at least, tightly centred around 0 bias for all of the regime - reduction of shrinking bias indicated the posterior is data informed rather than prior-dominated. There is no catastrophic divergence for either formulation, unlike the naive distance estimates. The right panel shows the separation between mode and median estimates, a crucial diagnostic of determining asymmetry in the posterior. While both posteriors exhibit increasing separation with distance, the initial parallax posterior tends to be more extreme in this regard, beginning the separation at a similar point to the shrinking bias. This is a sign of the presence of long distance tails induced by parallaxes with high σ_{ϖ}, compared to the full posterior which maintains more uniform coupled modes and medians up to the 1kpc cutoff. This indicates the posterior shape is well behaved, even in the low-SNR regime. Together, these diagnostics confirm that the full posterior yields more symmetric, data-driven (likelihood dominated) distance distributions and substantially reduces prior-induced bias.

Distance inference errors are inherently non-Gaussian (sometimes multi-modal), asymmetric (long tails), and heteroscedastic, particularly in the low-parallax signal-to-noise regime (high σ_{ϖ}). We therefore adopt a complementary set of rank-based and scale-sensitive statistics that together probe correlation, sensitivity, and dispersion across several distance regimes and posterior formulation methods.

The Spearman rank correlation coefficient, ρ, measures the degree of monotonic association

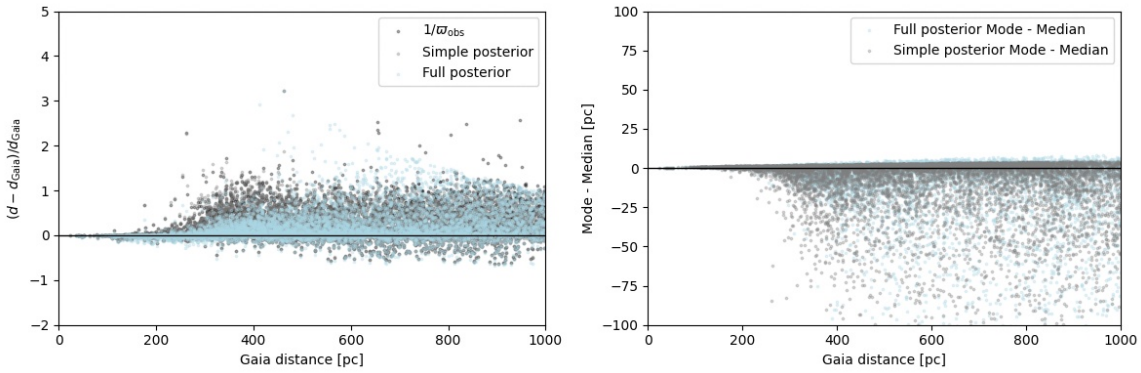

Figure 7: *Left*: distance bias for naive distances + both posteriors. *Right*: mode-median asymmetry as a function of GAIA distances. Both show similar breakdown of uniformity at the $\sim$300pc mark.

Table 1: Performance comparison of distance estimation methods. Spearman correlation coefficients are reported with corresponding p-values in parentheses.

Metric	**Naive Distance**	**Simple Posterior**	**Full Posterior**
Spearman ρ	0.778 (0.0)	0.804 (0.0)	0.890 (0.0)
Median Absolute Fractional Deviation	0.266	0.249	0.113
Log-Distance Scatter	0.222	0.153	0.114

between two variables, independent of their absolute scaling or linearity. It is non-parametric, insensitive to trends, and captures trends above linearity, i.e. those that are monotonic. Distances may be biased, but preserving the ordering (distinguishing nearer and further stars) is important for population studies. Rank preservation provides meaning to this assessment when we have asymmetric posteriors, compared to linear combination alone. The increase in ρ from naive $\rightarrow$ parallax $\rightarrow$ full posterior reflects the success of the photometric/colour constraints to ensure correct relative ordering even when parallax alone is weak and the inverse relationship breaks down.

The median absolute fractional deviation is just the median of our bias estimate as explored in Figure 7. Fractional deviations provide physically meaningful performance quantification across a wide distance range, the median again protecting against heavy-tailed error distributions. The highest value for naive estimates show they are dominated by low-SNR parallaxes, and the reduction in fractional deviation for the extended posterior indicated the breaking of the distance-parallax degeneracy through photometric data.

The log-distance scatter is the standard deviation of $\Delta \log d = \log_{10}(d_{\rm est}) - \log_{10}(d_{\rm ref})$, which measures multiplicative errors rather than additive ones. These are relevant as distance errors are approx. multiplicative, and is sensitive to posterior width/tail behaviour, making it well suited for evaluating prior impact and additional constraints from the likelihood. The reduction in log-distance scatter closely mirrors the suppression of long-distance tails observed in the posterior geometry diagnostics, confirming that the full posterior yields more compact and symmetric distance distributions.

Taken together, these metrics quantitatively confirm the trends observed in the posterior comparison plots and discussion. The full posterior simultaneously improves rank preservation, reduces fractional error, and suppresses multiplicative scatter, consistent with its more symmetric geometry, reduced skewness, and weaker prior dominance. The consistent improvement across all three metrics demonstrates that incorporating photometric and colour information into the

likelihood yields distance estimates that are not only more precise, but better ordered, and less sensitive to low-SNR/high σ_{ϖ} parallax measurements.

7. Conclusion

In this work, we investigated to a surface level degree the problem of stellar distance inference from GAIA DR2 parallaxes using a Bayesian framework. Beginning with parallax-only likelihood and distance prior, we observed how the naive inversion of parallaxes without reference to uncertainty leads to bias distance estimates, and how introducing a probabilistic treatment of inference deals with the fractional errors growing at large distance.

The parallax-only posterior provided effective regularisation and yielded distance estimates that were more stable than simple inversion, however, they included residual biases and broad tails for distant stars. This, along with increasing mode-median separation and asymmetric posteriors highlighted the limitations of parallax-only inference. The posterior was therefore extended to incorporate spectroscopic, photometric, and astrometric information across parameter space, jointly inferring distance and absolute magnitude, eventually marginalising over the nuisance parameter. This full posterior encoded additional astrophysical constraints via the colour-magnitude relation, and the extinction prior, allowing for the breakdown of degeneracies. The resulting posteriors exhibit tighter, more symmetric distance PDFs, and reduced prior sensitivity.

This was confirmed by quantitative comparisons across all three approaches (naive inversion, parallax-only, full posterior), demonstrating systematic improvement with increasing model complexity. We achieved higher rank correlation, reduced fractional deviations, and lower log-distance scatter. These improvements were consistent across the full distance range of the GAIA dataset and are particularly pronounced for low-SNR sources where parallax-only inference breaks down.

Overall, this study highlights the importance of Bayesian distance inference for modern astrometric surveys and demonstrates how incorporating additional observational constraints leads to more accurate and physically interpretable distance estimates. The framework presented here is readily extensible to larger samples and other/future Gaia data releases, and illustrates the broader principle that astrophysical inference benefits most when observational uncertainties and prior knowledge are treated in a unified Bayesian manner.

8. Bibliography

Bailer-Jones C. A. L., 2015, , 127, 994

Lindegren L., et al., 2018, , 616, A2

Palmer M., Arenou F., Luri X., Masana E., 2014, Astronomy & Astrophysics, 564, A49

Sloan Digital Sky Survey (SDSS) Collaboration 2025, Selection Biases, https://www.sdss4.org/dr17/irspec/targets/selection-biases/

Verbiest J. P. W., Weisberg J. M., Chael A. A., Lee K. J., Lorimer D. R., 2012, The Astrophysical Journal, 755, 39

Gaia Data Release 3: distance_gspphot description, https://gea.esac.esa.int/archive/documentation/GDR3/Gaia_archive/chap_datamodel/sec_dm_main_source_catalogue/ssec_dm_gaia_source.html