Bayesian Inference of Galactic Rotation Using Gaia DR2 Data

Colombo Gabriele, Maszkowski Michal, Seghezzi Daniele

April 8, 2025

Abstract

In this report, we present a Bayesian model for the rotational motion of stars in the Milky Way galaxy, using observational data from the Gaia Data Release 2 (DR2) [1]. We begin by introducing a simplified physical model in which both the stars and the Local Standard of Rest (LSR) follow circular orbits around the Galactic Center (GC) with a constant velocity $V_{\rm rot}$. The Sun is modeled as moving relative to the LSR with components $U_\odot$ (in the x-direction) and $V_\odot$ (in the y-direction), all defined within the GC frame. We then extend this model by introducing a random velocity component $v'_{\rm rand} \sim \mathcal{N}(0, \sigma)$ for each star, to account for intrinsic stellar motion not captured by pure circular rotation.

Using Bayesian inference and applying Markov Chain Monte Carlo (MCMC) sampling via the `emcee` [2] package, we estimate the posterior distributions of the model parameters. The inference is performed using stellar longitude, parallax, and radial velocity measurements, along with their associated uncertainties, all expressed in the Sun-centered reference frame.

For the simpler model, we estimate the parameters (with 95% confidence intervals) as: $V_{\rm rot} = 211.45^{+0.04}_{-0.04}$ km s^{-1}; $U_\odot = 11.638^{+0.005}_{-0.005}$ km s^{-1}; $V_\odot = 21.604^{+0.005}_{-0.005}$ km s^{-1}.

In the extended model, which includes the velocity dispersion of all stars, we find: $V_{\rm rot} = 204^{+2}_{-2}$ km s^{-1}; $U_\odot = 11.7^{+0.3}_{-0.3}$ km s^{-1}; $V_\odot = 21.7^{+0.3}_{-0.3}$ km s^{-1}; $\sigma = 30.6^{+0.2}_{-0.2}$ km s^{-1}.

1 Data

In this work, we utilize stellar data from Gaia Data Release 2 (DR2) [1]. From the extensive Gaia catalog, we select only those stars for which radial velocities $v_{\rm rad}$ were measured relative to the Sun via the Doppler effect. To manage the dataset size efficiently, we apply a random subsampling procedure, restricting the dataset to entries with a random index less than 10^8.

For each selected star, we extract key parameters from Gaia DR2: parallax p and its associated uncertainty σ_p, radial velocity $v_{\rm rad}$ with measurement uncertainty σ_v, and Galactic coordinates — specifically, latitude b and longitude l.

To focus our analysis on stars approximately lying in the Galactic plane, we apply a cut on the latitude: $|b| < 5°$. To ensure data quality, we retain only stars with a relative parallax uncertainty below 20%, and a radial velocity uncertainty less than 5 km s^{-1}.

After applying these criteria, we obtain a sample of $\mathrm{N_{stars}} = 75{,}659$ stars. The upper panel of Figure 1 shows the distribution of their radial velocities with respect to the Sun, plotted as a function of Galactic longitude. For clarity, only values within the range $[-200, 200]$ km s^{-1} are displayed, although the full dataset includes velocities extending up to approximately 500 km s^{-1}. These extreme values are likely outliers, and a more refined data-cleaning procedure could further improve the quality of the analysis.

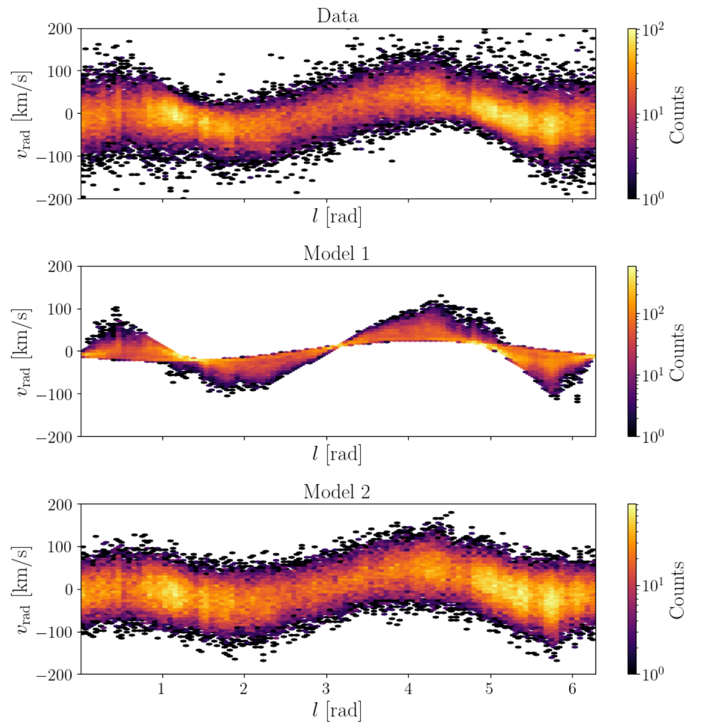

Figure 1: Upper: observed radial velocities in the range $[-200, 200]$ km s^{-1}. Middle: predicted radial velocities from the first model. Lower: simulated radial velocity distribution under the second model.

2 Model

In this section we present the physical and the statistical models we developed for the study of the Galactic kinematics of the Milky Way. Our physical assumptions, presented in the paragraph 2.1, lead to a prediction for the radial motion of stars with respect to the Sun as a function of their longitude and parallax. The prediction, for each star, can then be compared to the direct measurement. In the paragraph 2.2 we formally present our statistical assumptions on the data and the parameters of the two models with their priors and likelihoods.

2.1 Physical Model

In our model, we restrict attention to stellar motion confined to the Galactic plane ($b \approx 0$). We assume that each star moves in a circular orbit around the Galactic Center (GC) with a uniform speed V_{rot}. Fixing the coordinate system as shown in Figure 2, the velocity $\boldsymbol{v}_{\mathrm{s}}$ of a generic star s is given by:

$$\boldsymbol{v}_{\mathrm{s}} = V_{\mathrm{rot}} \begin{pmatrix} \sin\varphi \\ -\cos\varphi \end{pmatrix}, \tag{1}$$

where φ is the angle between the star's position vector and the x-axis, measured from the GC. In the following, primed quantities refer to the Sun's frame of reference, while unprimed quantities are defined in the GC frame. All angles are expressed in radians.

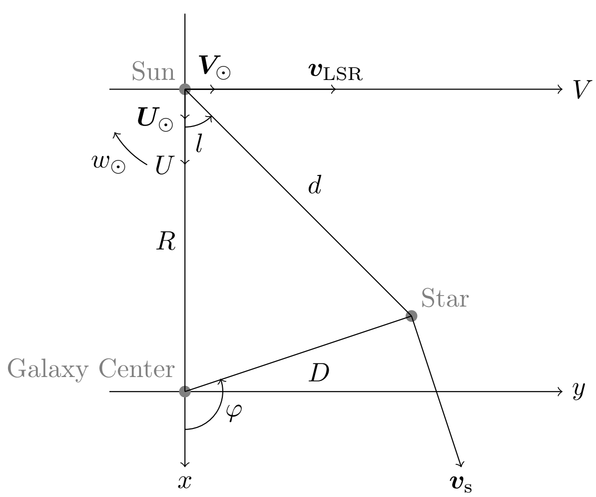

Figure 2: Frames of reference. The angular frequency $w_\odot$ only contributes to the tangential component of the velocities of the stars in the Sun's frame of reference, and is therefore omitted in equation 3.

In this reference frame, the Sun is located at $\varphi = \pi$, at a fixed distance of $R = 8300\,\mathrm{pc}$ from the GC [3]. The Local Standard of Rest (LSR) shares the same rotational velocity as described by Eq. 1. Additionally, the Sun has a peculiar motion with respect to the LSR, with components $U_\odot$ along the x-axis and $V_\odot$ along the y-axis. The total velocity of the Sun is therefore:

$$\boldsymbol{v}_\odot = \begin{pmatrix} 0 \\ V_{\mathrm{rot}} \end{pmatrix} + \begin{pmatrix} U_\odot \\ V_\odot \end{pmatrix}. \tag{2}$$

The velocity of a star in the Sun's frame is then given by:

$$\boldsymbol{v}'_s = \boldsymbol{v}_s - \boldsymbol{v}_\odot. \tag{3}$$

We now project $\boldsymbol{v}'_s$ onto the radial direction relative to the Sun, denoted $\hat{e}'_r$, obtaining the observed radial velocity v'_{rad} of a star with Galactic longitude l:

$$\begin{aligned} v'_{\mathrm{rad}} &= \boldsymbol{v}'_s \cdot \hat{e}'_r \\ &= V_{\mathrm{rot}}\left[\sin\varphi\cos l - (1+\cos\varphi)\sin l\right] - U_\odot \cos l - V_\odot \sin l. \end{aligned} \tag{4}$$

To relate this model to Gaia observations, we must express $\sin\varphi$ and $\cos\varphi$ in terms of l and the parallax p (in milliarcseconds). The distance d (in parsecs) from the Sun to a star is calculated as:

$$d\ [\mathrm{pc}] = \frac{1000}{p\ [\mathrm{mas}]}. \tag{5}$$

Using simple geometric relations (see Figure 2), we find the distance D from the star to the GC, and express $\sin\varphi$ and $\cos\varphi$ as:

$$\begin{aligned} D &= \sqrt{d^2 + R^2 - 2dR\cos l}, \\ \sin\varphi &= \frac{d\sin l}{D}, \\ \cos\varphi &= \frac{d\cos l - R}{D}. \end{aligned} \tag{6}$$

Substituting Eqs. 5–6 into Eq. 4, we obtain the prediction for the radial velocity of star i under the first model, $\hat{v}^{(1)}_{\mathrm{rad},i}$, as a function of its longitude l_i and parallax p_i:

$$\begin{aligned} \hat{v}^{(1)}_{\mathrm{rad},i}(l_i, p_i) = V_{\mathrm{rot}} \sin l_i & \left(\frac{R}{\sqrt{\left(\frac{1000}{p_i}\right)^2 + R^2 - 2\left(\frac{1000}{p_i}\right) R\cos l_i}} - 1 \right) \\ & - U_\odot \cos l_i - V_\odot \sin l_i. \end{aligned} \tag{7}$$

In a more refined second model, we account for the random peculiar motion of stars. We model each star's radial velocity component due to random motion as a Gaussian variable: $v_{\mathrm{rand},i} \sim \mathcal{N}(0, \sigma)$, where σ is the velocity dispersion. The corresponding prediction becomes:

$$\begin{aligned} \hat{v}^{(2)}_{\mathrm{rad},i}(l_i, p_i) = V_{\mathrm{rot}} \sin l_i & \left(\frac{R}{\sqrt{\left(\frac{1000}{p_i}\right)^2 + R^2 - 2\left(\frac{1000}{p_i}\right) R\cos l_i}} - 1 \right) \\ & - U_\odot \cos l_i - V_\odot \sin l_i + v_{\mathrm{rand},i}. \end{aligned} \tag{8}$$

2.2 Statistical Model

In this section, we describe the Bayesian inference framework employed to estimate the parameter distributions of the two models. The unnormalized posterior probability distributions were sampled using a Markov Chain Monte Carlo (MCMC) method implemented via the `emcee` [2] Python package. Specifically, we supplied `emcee` with the unnormalized posterior distribution, i.e., the numerator of Bayes' theorem:

$$\mathbb{P}^*(\theta|\mathcal{D}) = \mathcal{L}(\mathcal{D}|\theta)\mathbb{P}(\theta) \ , \tag{9}$$

where $\mathcal{L}(\mathcal{D}|\theta)$ is the likelihood and $\mathbb{P}(\theta)$ the prior. In practice, their logarithms were used to enhance numerical stability. Constant factors independent of the model parameters were omitted, as they do not affect the sampling of the posterior distribution and are thus irrelevant for our purposes.

In both models, we assume all measurements and parameters to be statistically independent, given no a priori reason to consider correlations among them. This allows the likelihood and prior to be factorized into products over individual terms, and their logarithms expressed as corresponding sums. We model each measured quantity q_{measured} as the sum of its true value q_{true} and a random error ϵ drawn from a normal distribution centered at 0, with standard deviation σ_{q}, given by the reported statistical uncertainty provided by the GAIA database: $\epsilon \sim N(0, \sigma_{\text{q}})$. Thus, $q_{\text{measured}} = q_{\text{true}} + \epsilon$.

For the first model, the parameter vector is $\boldsymbol{\theta_1} = (V_{\text{rot}}, U_\odot, V_\odot)$. Neglecting uncertainties on parallax measurements, the difference between the observed radial velocity $v_{\text{rad},i}$ and its model prediction $\hat{v}^{(1)}_{\text{rad},i}$ is treated as a Gaussian random variable centered at 0, with standard deviation given solely by the reported radial velocity uncertainty $\sigma_{\text{v},i}$. The log-likelihood for this model is therefore:

$$\log \mathcal{L}^{(1)}(\mathcal{D}|\theta_1) = -\frac{1}{2} \sum_{i=1}^{\text{N}_{\text{stars}}} \left[\log(2\pi\sigma^2_{\text{v},i}) + \frac{(v_{\text{rad},i} - \hat{v}^{(1)}_{\text{rad},i})^2}{\sigma^2_{\text{v},i}} \right] \ . \tag{10}$$

We assume the three parameters to be mutually independent. For V_{rot}, we adopt a flat prior over the interval $[0, 500\,\text{km}\,\text{s}^{-1}]$, which covers the typical rotational velocities in Sb galaxies [5], found in the range [144, 330] $\text{km}\,\text{s}^{-1}$[5]. For $U_\odot$ and $V_\odot$, we use Gaussian priors centered at 0, based on the assumption that the Sun's peculiar motion is akin to a stochastic thermal motion. Therefore, we model their log-prior to be proportional to:

$$\log \mathbb{P}^{(1)}(U_\odot) + \log \mathbb{P}^{(1)}(V_\odot) \sim -\frac{U_\odot^2 + V_\odot^2}{v_{\text{gal}}^2} \ ,$$

with $v_{\text{gal}} = 200\,\text{km}\,\text{s}^{-1}$, reflecting the typical velocity scale in spiral galaxies [5], as previously mentioned.

In the second model, the parameter vector is $\boldsymbol{\theta_2} = (V_{\text{rot}}, U_\odot, V_\odot, \sigma)$. In this case, uncertainties in both radial velocity and parallax are taken into account. The model prediction $\hat{v}^{(2)}_{\text{rad},i}$ (eq.8) is treated as a random variable drawn from a normal distribution whose variance includes two contributions: the intrinsic dispersion σ^2, and the uncertainty due to parallax measurement errors. Assuming the parallax error is small, its contribution to the model variance can be approximated by $\left(\frac{\partial \hat{v}^{(2)}_{\text{rad},i}}{\partial p_i}\right)^2 \sigma^2_{p_i}$, where the derivative is analytically derived from eq.8. The difference between the measured radial velocity $v_{\text{rad},i}$ and the model prediction $\hat{v}^{(2)}_{\text{rad},i}$ is then modeled as a Gaussian random variable centered at zero, with total variance equal to the sum of the variances of the radial velocity measurement, the parallax-induced component, and the intrinsic dispersion:

$$\sigma^2_{\text{tot},i} = \sigma^2_{\text{v},i} + \left(\frac{\partial \hat{v}^{(2)}_{\text{rad},i}}{\partial p_i}\right)^2 \sigma^2_{p_i} + \sigma^2 \ . \tag{11}$$

The log-likelihood is therefore:

$$\log \mathcal{L}^{(2)}(\mathcal{D}|\theta_2) = -\frac{1}{2} \sum_i \left\{ \log[2\pi(\sigma_i^2 + \sigma^2)] + \frac{(v_{\text{rad},i} - \hat{v}^{(2)}_{\text{rad},i})^2}{\sigma^2_{\text{tot},i}} \right\} \ . \tag{12}$$

The prior distributions for V_{rot}, $U_\odot$, and $V_\odot$ are retained from the first model. For the additional parameter σ, we assume it is uncorrelated with the others and assign it the non-informative prior appropriate for the standard deviation in Gaussian models, since it has a similar interpretation in the model we implemented: $\log \mathbb{P}(\sigma) = -\log(\sigma)$ [4].

3 Conclusions

The median values and 95% confidence intervals of the posterior distributions for both models are reported in Table 1. The estimates for $U_\odot$ and $V_\odot$ are consistent across the two models, while the estimates for V_{rot} differ significantly — their respective 95% confidence intervals do not overlap.

The inferred values of $U_\odot$ from both models are in good agreement with results from the literature [6], while the estimates of $V_\odot$ are notably higher and incompatible with those reported in the same source. Regarding V_{rot}, our values are lower than the literature estimate [3], but both models yield confidence intervals that partially overlap with the reference value, indicating general consistency.

The posterior distributions for all parameters are presented in Appendix A. Additionally, the middle and lower panels of Figure 1 show the model predictions based on the median posterior values, using the longitude and parallax of the observed stars. For the second model, the effect of stellar random motion was simulated by adding a Gaussian noise term $v_{\text{rand},i} \sim \mathcal{N}(0, \sigma)$, using the inferred value of σ.

Visually, it is evident that the second model provides a better fit to the data, highlighting the importance of incorporating the random motion of stars to accurately model their observed radial velocities.

Parameter	Model 1	Model 2	Literature
V_{rot} [$\text{km}\,\text{s}^{-1}$]	$211.45^{+0.04}_{-0.04}$	204^{+2}_{-2}	225 ± 10 [3]
$U_\odot$ [$\text{km}\,\text{s}^{-1}$]	$11.638^{+0.004}_{-0.005}$	$11.7^{+0.3}_{-0.3}$	$11.1^{+0.7}_{-0.8}$ [6]
$V_\odot$ [$\text{km}\,\text{s}^{-1}$]	$21.604^{+0.005}_{-0.005}$	$21.7^{+0.3}_{-0.3}$	$12.2^{+0.5}_{-0.5}$ [6]
σ [$\text{km}\,\text{s}^{-1}$]	—	$30.6^{+0.2}_{-0.2}$	—

Table 1: Estimated parameters and their 95% confidence intervals for the two models. The literature value of V_{rot} includes total uncertainty (1σ), combining statistical and systematic contributions as reported in [3].

References

[1] A. G. A. Brown et al. "Gaia Data Release 2: Summary of the contents and survey properties". In: *Astronomy & Astrophysics* 616 (Aug. 2018), A1. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201833051. URL: http://dx.doi.org/10.1051/0004-6361/201833051.

[2] Daniel Foreman-Mackey et al. "emcee: The MCMC Hammer". In: 125.925 (Mar. 2013), p. 306. DOI: 10.1086/670067. arXiv: 1202.3665 `[astro-ph.IM]`.

[3] Richard de Grijs and Giuseppe Bono. "Clustering of Local Group Distances: Publication Bias or Correlated Measurements? V. Galactic Rotation Constants". In: *The Astrophysical Journal Supplement Series* 232.2 (Oct. 2017), p. 22. DOI: 10.3847/1538-4365/aa8b71. URL: https://dx.doi.org/10.3847/1538-4365/aa8b71.

[4] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press, 2003. ISBN: 9780521642989. URL: http://www.inference.phy.cam.ac.uk/mackay/itila/.

[5] Peter Schneider. *Extragalactic Astronomy and Cosmology: An Introduction*. 2nd. See table 3.2 for the typical rotational speed of spiral galaxies. See appendix B for the estimation of the mass from measurements of luminosity. Heidelberg, Germany: Springer-Verlag Berlin Heidelberg, 2015. ISBN: 978-3-642-54082-0. DOI: 10.1007/978-3-642-54083-7.

[6] Ralph Schönrich, James Binney, and Walter Dehnen. "Local kinematics and the local standard of rest". In: *Monthly Notices of the Royal Astronomical Society* 403.4 (Apr. 2010), pp. 1829–1833. ISSN: 1365-2966. DOI: 10.1111/j.1365-2966.2010.16253.x. URL: http://dx.doi.org/10.1111/j.1365-2966.2010.16253.x.

A Posterior distributions

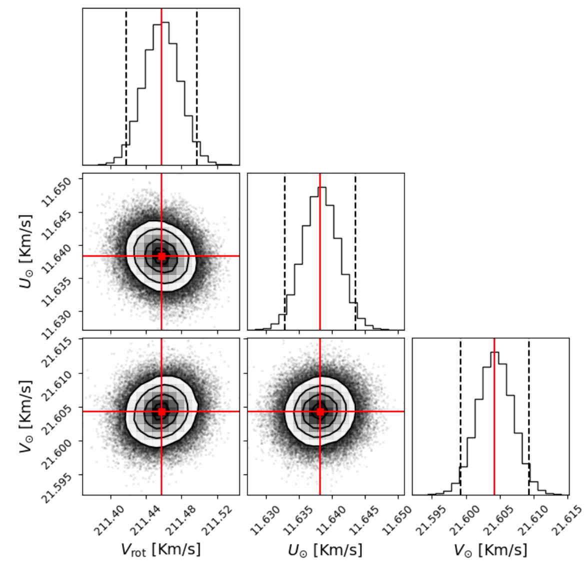

Figure 3: Posterior distribution of the parameters for the first model. Red continuous lines indicate the median values found for the respective parameters, while the black dashed ones delimit their 95% confidence interval.

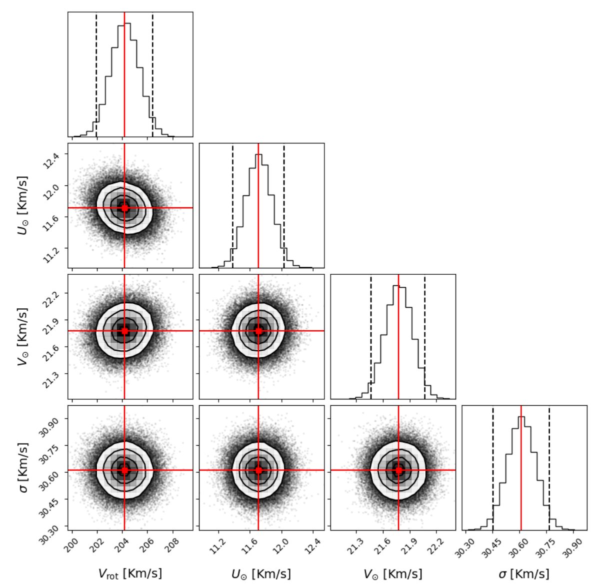

Figure 4: Posterior distribution of the parameters for the second model. Red continuous lines indicate the median values found for the respective parameters, while the black dashed ones delimit their 95% confidence interval.