Determining stellar density laws with *TGAS*

Jo Bovy[1,2,3]

ABSTRACT

This note discusses how to observationally determine stellar density laws from the *TGAS* data, taking into account the *TGAS* selection function.

1. Introduction

The general problem under consideration here is how to determine stellar density laws $\nu_*(X,Y,Z)$ of stellar populations in the Milky Way from *TGAS* observations of stars providing precise (α,δ,ϖ,G). These notes follow the discussion in Bovy et al. (2016), which the reader should familiarize themselves with, although the discussion below is largely self-contained.

2. Likelihood-based density modeling

We are interested in $\nu_*(X,Y,Z)$, where (X,Y,Z) is a set of cartesian coordinates related by a rotation to the spherical coordinates $(\alpha,\delta,D=1/\varpi)$ in which observations are made[1]. Following Bovy et al. (2016), we determine $\nu_*(X,Y,Z)$ from a population of stars that is not complete in any (simple) geometric sense and thus need to take the selection function into account. We assume that the selection function is a function of (a) position on the sky (α,δ), (b) an apparent magnitude (here *Gaia* G), and potentially (c) a color c. We further assume that the absolute magnitude M_G of stars in the population under investigation can be (closely enough) determined from G and c, such that we can connect the underlying stellar density in (X,Y,Z) to the distribution in apparent magnitude.

---

[1] Department of Astronomy and Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON, M5S 3H4, Canada; bovy@astro.utoronto.ca

[2] Center for Computational Astrophysics, Flatiron Institute, 162 5th Ave, New York, NY 10010, USA

[3] Alfred P. Sloan Fellow

[1]We write distance $D=1/\varpi$ as a shorthand here; if σ_ϖ is large, care and prior information must be used to infer D from ϖ.

To determine $\nu_*(X,Y,Z)$ we use a likelihood approach that models the full rate function $\lambda(O|\theta)$ that gives the number of stars as a function of all observables O of interest for a set of model parameters θ. These observables O are in this case (α,δ,D,G,c)—we will use (X,Y,Z) and (α,δ,D) interchangeably because they are related by coordinate transformation, but will keep track of the Jacobian—and we can write

$$\begin{aligned}\lambda(O|\theta) &= \lambda(\alpha,\delta,D,G,c)\,,\\ &= \nu_*(X,Y,Z|\theta)\,D^2\,\cos\delta\,\rho(M_G,c|X,Y,Z)\,S(\alpha,\delta,G,c)\,,\end{aligned} \tag{1}$$

where we have assumed that the model parameters only affect ν_*. In this decomposition, the factor $D^2\cos\delta$ comes from the Jacobian of the transformation between (α,δ,D) and (X,Y,Z), $S(\alpha,\delta,G,c)$ is the *TGAS* selection function, and $\rho(M_G,c|X,Y,Z)$ gives the density distribution of (M_G,c), which may be a function of position (X,Y,Z), of the stellar population.

An observed set of stars indexed by i forms a Poisson process with the likelihood $\mathcal{L}(\theta)$ of the parameters θ describing the density law given by

$$\begin{aligned}\ln\mathcal{L}(\theta) &= \sum_i \ln\lambda(O_i|\theta) - \int \mathrm{d}O\lambda(O|\theta)\,,\\ &= \sum_i \ln\nu_*(X_i,Y_i,Z_i|\theta)\\ &\quad - \int \mathrm{d}D\,D^2\mathrm{d}\alpha\mathrm{d}\delta\,\cos\delta\,\nu_*(X,Y,Z|\theta)\int \mathrm{d}G\mathrm{d}c\,\rho(M_G,c|X,Y,Z)\,S(\alpha,\delta,G,c)\,,\end{aligned} \tag{2}$$

where in the second line we have dropped terms that do not depend on θ. As in Bovy et al. (2016) we simplify this expression by defining the *effective selection function* $\mathfrak{S}(\alpha,\delta,D)$ defined by

$$\mathfrak{S}(\alpha,\delta,D) \equiv \int \mathrm{d}G\mathrm{d}c\,\rho(M_G,c|X,Y,Z)\,S(\alpha,\delta,G,c)\,, \tag{3}$$

where we use that $5\log_{10}(D/10\,\mathrm{pc}) = G - M_G$[2]. The ln likelihood then becomes

$$\ln\mathcal{L}(\theta) = \sum_i \ln\nu_*(X_i,Y_i,Z_i|\theta) - \int \mathrm{d}D\,D^2\mathrm{d}\alpha\mathrm{d}\delta\,\cos\delta\,\nu_*(X,Y,Z|\theta)\,\mathfrak{S}(\alpha,\delta,D)\,. \tag{4}$$

Unlike the selection function $S(\cdot)$ which is a function of the survey's operations only (which parts of the sky were observed, for how long, ...), the effective selection function $\mathfrak{S}(\cdot)$ is a function of both the survey operations *and* the stellar population under investigation. Its

[2]This assumes that there is no extinction due to dust. If there is extinction $A_G(\alpha,\delta,D)$, this should be taken into account in this equation.

usefulness derives from the fact that it encapsulates all observational effects due to selection and dust obscuration and turns the inference problem into a purely geometric problem.

To determine the best-fit parameters $\hat{\theta}$ of a parameterized density law $\nu_*(X,Y,Z|\theta)$ one has to optimize the ln likelihood given above. This ln likelihood can be marginalized analytically over the overall amplitude of the density (the local normalization if you will); this is discussed in Bovy et al. (2016) and similar expressions would apply here.

3. Non-parametric binned density laws

Now suppose that one wants to determine the density $\nu_*(X,Y,Z)$ of a stellar population in a set of bins in (X,Y,Z). The bins are given by a set $\{\Pi_k\}_k$ of rectangular functions that are equal to one within the domain of the bin and zero outside of it. The domain can have an arbitrary shape, but typically this would be an interval in each of X, Y, and Z or perhaps in R and Z, where R is the Galactocentric radius. We can then write the density as

$$\nu_*(X,Y,Z|\theta) = \sum_k n_k\,\Pi_k(X,Y,Z)\,, \tag{5}$$

where $\theta \equiv \{n_k\}_k$ is a set of numbers that give the density in each bin and that therefore parameterizes the density law.

The ln likelihood then becomes

$$\begin{aligned}\ln\mathcal{L}(\{n_k\}_k) &= \sum_i \ln\sum_k n_k\,\Pi_k(X_i,Y_i,Z_i) - \int \mathrm{d}D\,D^2\mathrm{d}\alpha\mathrm{d}\delta\,\cos\delta\,\sum_k n_k\,\Pi_k(X,Y,Z)\,\mathfrak{S}(\alpha,\delta,D)\,,\\ &= \sum_k\left[N_k\ln n_k - n_k\int \mathrm{d}D\,D^2\mathrm{d}\alpha\mathrm{d}\delta\,\cos\delta\,\Pi_k(X,Y,Z)\,\mathfrak{S}(\alpha,\delta,D)\right]\,,\end{aligned} \tag{6}$$

where N_k is the number of points i in the observed set that fall within bin k. We can maximize this likelihood for each n_k analytically and find best-fit $\hat{n}_k$

$$\hat{n}_k = \frac{N_k}{\int \mathrm{d}D\,D^2\mathrm{d}\alpha\mathrm{d}\delta\,\cos\delta\,\Pi_k(X,Y,Z)\,\mathfrak{S}(\alpha,\delta,D)}\,. \tag{7}$$

The denominator in this expression is known as the *effective volume*. Using the same symbol Π_k to denote the integration volume and using $x = (X,Y,Z)$ and (α,δ,D) interchangeably because they are related through coordinate transformation (as we've been doing all along), this can be written as the following simple expression

$$\hat{n}_k = \frac{N_k}{\int_{\Pi_k}\mathrm{d}^3x\,\mathfrak{S}(\alpha,\delta,D)}\,. \tag{8}$$

Thus, the effective volume is the spatial integral of the effective selection function. This expression makes sense, because for a complete sample $\mathfrak{S}(\alpha, \delta, D) = 1$ and this expression simplifies to the number divided by the volume of the bin, the standard way to compute a number density.

From the second derivative of the ln likelihood, we find the uncertainty on the $\hat{n}_k$

$$\sigma_{\hat{n}_k} = \frac{\hat{n}_k}{\sqrt{N_k}}\,, \tag{9}$$

which again makes sense.

4. Application to *TGAS*

To apply this formalism to the *TGAS* data, one requires the following

- $S(\alpha, \delta, G, c)$: This can be obtained by comparing the *TGAS* catalog to the 2MASS catalog. In principle this is best done in a pixelized representation of the sky (α, δ), e.g., through use of HEALPix. The best form of the dependence on G may be a smooth function or a binned representation. Whether the selection function has any significant dependence on a color c for samples of interest is unknown to me.

  If a cut on parallax uncertainty is made, then the selection function needs to be determined by comparing the sample resulting from the cut with the underlying 2MASS sample. Because parallax uncertainty depends on c, this will almost certainly introduce a dependence on c, although this may be minimized by restricting the color range. To allow for different parallax-uncertainty cuts (and other *TGAS* cuts), we would therefore ideally be able to determine the selection function on the fly. However, the denominator (the 2MASS number counts) could be cached on a fine HEALPix grid (HEALPix is best, because TGAS has a HEALPix index that can then be used to quickly determine the selection function).

  A complication in practice is that the *Gaia* catalog itself is not complete (otherwise we could just use it!), so not all 2MASS sources will have a G magnitude. There are various options to deal with this: (a) use J instead of G, (b) determine a transformation $(J, H, K) \rightarrow G$ and assign G based on this to 2MASS sources missing a G measurement.

  Another option may be to use the Tycho-2 catalog as the underlying catalog instead of 2MASS. Tycho-2 is 99 % complete down to $V = 11$, so could function instead of 2MASS down to that limit.

- $\mathbf{\mathfrak{S}}(\alpha,\delta,D)$: The calculation of $\mathbf{\mathfrak{S}}(\alpha,\delta,D)$ requires the density $\rho(M_G,c|X,Y,Z)$ [in addition to $S(\alpha,\delta,G,c)$]. This can be obtained most easily from a Monte Carlo sampling of the *TGAS* sample of interest. The *TGAS* sample provides a sample $(M_{G,j},c_j)$ and the integral over $\rho(M_G,c)$ in the effective selection function can thus be performed through Monte Carlo integration using this sample:

$$\mathbf{\mathfrak{S}}(\alpha,\delta,D)\approx\sum_j S(\alpha,\delta,5\log_{10}(D/10\,\mathrm{pc})+M_{G,j},c_j)\,. \tag{10}$$

  The sample $(M_{G,j},c_j)$ can be restricted to a certain spatial region (e.g., Π_k) if there is any concern that the density $\rho(M_G,c|X,Y,Z)$ depends significantly on (X,Y,Z), but this is unlikely.

REFERENCES

Bovy, J., Rix, H.-W., Green, G. M., Schlafly, E. F., & Finkbeiner, D. P. 2016, ApJ, 818, 130

This preprint was prepared with the AAS LaTeX macros v5.2.