Gaia parallax and SED fitting challenge

1 Astrophysical context

1.1 Motivation

With the recent Gaia first public data release, there are now 2 million stars which have parallax information. However, the parallax measurement alone is often not good enough to reliably infer the distance of a star. Other information is required.
Fortunately, many of these stars also have a lot of photometric information. Therefore, if we combine the parallax and the available photometry, we can improve the distance estimation. To achieve this, we have to invoke models for the spectral energy distribution (SED) of stars. The model parameters will be:

- distance d in units of parsec (pc)
- absolute magnitudes in magnitudes (mag)

This challenge is focussing on distance estimation, comparing the results for different methods that use different amounts of information.

1.2 TGAS parallaxes

The Tycho-Gaia astrometric solution (TGAS) provides parallax angles ϖ, which we assume to have Gaussian errors σ_ϖ. From a model distance d (in pc) we can predict a model parallax $\frac{1000}{d}$ (in milli-arcsec). The parallax data (values and errors) are given in the data file `data-for-challenge.csv` for 224 real stars.

1.3 Photometry data

Photometry is measured in terms of apparent magnitudes m_X in some filter X. There are numerous different filters, but we will restrict ourselves to the following filters:

- Johnson filters U, B, V
- Cousins filters R, I
- 2MASS filters J, H, K

These cover the wavelength ranges from 450nm to 2 000nm, which is the range where most stars emit most of their light. The photometric data (apparent magnitudes and errors) are given in the data file `data-for-challenge.csv` for 224 real stars.

The apparent magnitude is obviously dependent on distance d. The further a star away, the fainter it gets and the apparent magnitude *increases*.[1] The crucial relation is the distance

[1] Yes, magnitude is an inverse scale ... it gets higher if the star gets fainter. Astronomers love to hold on to historical burdens that have far outlived their usefulness.

modulus:

$$m_X - M_X = A_X + 5\log_{10} d - 5 \tag{1}$$

Here M_X is the *absolute magnitude*, which is the magnitude the star would have at fixed distance of 10pc. A_X is the dust extinction in filter X. For the sake of simplicity, we will ignore dust and set $A_X = 0$ throughout this challenge! Thus we simplify:

$$m_X - M_X = 5\log_{10} d - 5 \tag{2}$$

It is now obvious that if we insert $d = 10pc$, we obtain $m_X = M_X$.

Our SED model will predict M_X for all filters X. We compare it to the observed apparent magnitudes m_X, assuming these have Gaussian measurement errors σ_X.

1.4 SED models

We recruit the SED models from the PARSEC evolutionary tracks.[2] As the name suggests, these are models that follow a star through its life and monitor how its parameters change with age. These evolutionary tracks do not only monitor how the star's parameters evolve, but also how its SED evolves. They provide our desired SED models M_X and are given in the file `PARSEC-tracks-Z0.019-Av0.dat`. It also contains lots of other parameters, such as effective temperature, which we will ignore.

1.5 Limitations

In this challenge, we are neglecting several effects: First, we are ignoring extinction by interstellar dust (A_X is set to 0). Second, we assume that all stars have solar metallicity, which is not true. Finally, we are not going to interpolate over the models, i.e., we may lack the resolution in model space to find the best-fit parameters.

[2] `http://stev.oapd.inaf.it/cgi-bin/cmd`

2 Acquaint with the data

2.1 Plot distribution of relative parallax errors

(a) From the data file `data-for-challenge.csv`, plot the distribution of relative parallax errors σ_ϖ/ϖ as a histogram in the range from -1 to 3.
(b) There are a few stars with negative parallax measurements. Why can you not discard them?

2.2 Plot colour-magnitude diagram

(a) From the data file `data-for-challenge.csv`, plot $J-K$ colour on the horizontal axis vs. absolute magnitude $M_V = V + 5\log_{10}\varpi + 5$ on the vertical axis (careful with the units of ϖ). Do not forget to invert the y-axes such that bright stars (negative M_V) are on the top and faint stars (positive M_V) are at the bottom.
(b) What types of stars (very roughly) do you recognise in the given data?
(c) From the data file `PARSEC-tracks-Z0.019-Av0.dat`, produce the same plot $M_J - M_K$ vs. M_V. Overplot the real data from `data-for-challenge.csv` in a different colour.
(d) Why do the models not cover the whole data space? What effects do we have to expect from this shortcoming of our models?

3 Mathematical background

3.1 Definition of likelihood function

(a) Write down the likelihood functions of the parallax measurement, $P(\varpi|d)$, and the photometric measurement in filter X, $P(m_X|M_X, d)$.
(b) Explain why the joint likelihood factorises such that:

$$P(\varpi, m_U, m_B, m_V, m_R, m_I, m_J, m_H, m_K|\theta) = P(\varpi|d) \prod_{X\in U,B,V,R,I,J,H,K} P(m_X|M_X, d) \quad (3)$$

(c) If there is *no parallax measurement*, i.e., you only have photometry, show that $\prod_{X\in U,B,V,R,I,J,H,K} P(m_X|M_X, d)$ implies

$$\chi^2 = \sum_{X\in U,B,V,R,I,J,H,K} \left(\frac{m_X - M_X - 5\log_{10} d + 5}{\sigma_X}\right)^2 \quad (4)$$

(d) Show that for *given SED* M_X, minimising this χ^2 has an *analytic* solution for $\log_{10} d$ that is:

$$\log_{10} d = \frac{1}{5} \frac{\sum_X \frac{m_X - M_X + 5}{\sigma_X^2}}{\sum_X \frac{1}{\sigma_X^2}} \quad (5)$$

(e) If we also have a parallax measurement, do we still get an analytic result?

3.2 Definition of distance prior $P(d)$

(a) If we are using a prior, we will use the uniform space density prior with exponential cutoff:

$$P(d) = \frac{1}{2L^3} d^2 e^{-d/L} \quad (6)$$

Overplot this prior for various exponential scale lengths $L = 100$pc, $L = 200$pc, and $L = 1350$pc, in the range from 0pc to 10 000pc.[3]
(b) Show that this prior assigns maximum a-priori probability to $d = 2L$.
(c) If we adopt this prior, do we still have an analytic solution for (logarithmic) distance from the photometry alone?

[3] For your orientation: The centre of the Milky Way is ≈8 000pc away.

4 Distance estimation

4.1 Parallaxes only without prior

Ignoring the photometry, take only the given parallax measurements and infer the distance through "plain" parallax inversion, i.e., $d = \frac{1000}{\varpi}$. Stars with negative parallax values will get negative distances. Store these distance estimates in some file (we plot them in comparison later).

4.2 Photometry only without prior

(a) Ignoring the parallax, take only the given photometric measurements and the SED models for M_X from `PARSEC-tracks-Z0.019-Av0.dat`, which contains ≈17 000 SED models. For every such model, use the analytic solution from Eq. (5) to get an estimate of the distance d. Also compute its likelihood. For the first five stars given in the data file `data-for-challenge.csv`, plot the resulting likelihood as a function of d (from 0pc to 1 000pc). Does it look Gaussian?
(b) For every star you have ≈17 000 distances and their likelihoods. For every star, take the distance which has the highest likelihood and plot it in comparison to the distance estimate you obtained from the parallax only.

4.3 Photometry only with prior

(a) Still ignoring the parallax, take only the given photometry, the SED models for M_X from `PARSEC-tracks-Z0.019-Av0.dat`, but now also include the prior from Eq. (6) for $L = 200$pc. There are ≈17 000 SED models in that data file. For *every* such model, take its SED M_X as given and optimise numerically for the distance by maximising the product of photometric likelihood and prior. You may use Newton's method here, with numerical gradient and numerical Hessian. For the first five stars given in the data file `data-for-challenge.csv`, plot the resulting posterior (not likelihood) as a function of d. Does it look Gaussian?
(b) For every star, take the distance that has highest posterior probability and make a plot to compare it to the distance inferred from the photometry without prior.

4.4 Parallax *and* photometry with prior

(a) Now using parallax and photometry, the SED models for M_X from `PARSEC-tracks-Z0.019-Av0.dat`, and the prior from Eq. (6) for $L = 200$pc. There are ≈17 000 SED models in that data file. For *every* such model, take its SED M_X as given and optimise numerically for the distance by maximising the product of parallax likelihood, photometric likelihood and prior. You may use Newton's method here, with numerical gradient and numerical Hessian. Again, for the first five stars given in the data file `data-for-challenge.csv`, plot the resulting posterior as a function of d. Does it look Gaussian?
(b) For every star, take the distance that has highest posterior probability and make a plot to compare it to the distance inferred from the photometry with prior.
(c) Do the parallaxes actually help or are the distance estimates from photometry alone good enough?
(d) Let us assume that, in principle, the photometry is good enough such that a parallax is not strictly necessary to get good distance estimates. Explain why parallaxes are still important to obtain. (Check your plots from Sect. 2.2 to get an inspiration.)