Statistical Methods 2024: Assignment 3 – modelling star density profiles and estimating cluster masses

This is the third of three assignments, for which extensive help is available during the tutorials. It is worth 30% of the final course grade.

The deadline for this assignment is Wednesday 31 Jan at 23:59. Due to the exam, late work is not permitted unless due to exceptional circumstances. The rubric for the assessment of all the assignments, listing the categories assessed and the requirements for each of them, will be provided separately on Canvas.

What you should submit
You should submit your work via Canvas. ***It must be in the form of a Jupyter notebook.*** Make sure that you upload the correct file, and check that all the cells run successfully (***and in the correct order***, from start to finish) before you submit!

Before you start it is essential that you read through the assignment grading explanation document on Canvas, since this explains what we expect from you in your answers. When answering each question, use markdown cells for explanations, assumptions and comments on your results: do not include these as comments in code cells, which are reserved only for comments about the code itself!
Remember that the usual plagiarism rules apply to your work: if you cut-and-paste code from somewhere/someone else (including code generated by an AI) you must cite the source (simply replacing variable names is not sufficient to make it your own!). See the grading explanation for more details. We also expect you to help each other, at least early on, and/or be inspired by methods you see online, so programming ***your own version*** (i.e. not cut and pasted) of someone else's method is fine and does not require citation.

The Assignment

For this final assignment, you will focus on studying the stellar density profiles (that is, star counts per unit area vs. radius from the centre of the cluster) of three open clusters in the Gaia DR3 dataset, which have been uniquely assigned to each of you. You will use maximum-likelihood methods to model the data with an analytical function, estimate errors on function parameters and compare subsets of the data to see if they have the same profiles. Finally, you will use MCMC to infer the cluster masses from your model fits.

Initial setup

First, load in the stars data as shown in Assignment 1, and create a dataframe containing only the stars from the three open clusters which have been assigned to you, which you will use for the remainder of this assignment. You can include all the stars in your analysis, not just the high-probability cluster members.

Assignment tasks

Each task contributes an equal weight to the assignment total grade:

1. First, make stellar density profiles for each of your three clusters. To do this, you first need to create a new column in your dataframe which contains the radial separation of each star (in arcseconds) from the centre of its cluster, i.e. $r = 3600.\sqrt{(x-\bar{x})^2 + (y-\bar{y})^2}$ where x and y are the RA and DE positions of the star (in degrees) and $\bar{x}$ and $\bar{y}$ are the means of RA and DE for all the stars in the cluster (i.e. the estimated location of the centre of the cluster). Then, use the radial locations of the stars to make a histogram of stellar number density ρ_* vs. radius, that is, the number of stars in a radial bin, normalized by the area of the radial bin, which will correspond to an annulus on the sky (use arcsec^2 as the units of area). Choose appropriate radial binning and plot the stellar density profiles (and error bars, if appropriate) for your 3 clusters on separate plots.

2. Now you will fit a model to your cluster stellar density profiles, using lmfit, which is demonstrated in the course online material. The model you will use is a variant of the King profile, an empirical function which is usually able to provide good fits the stellar density profiles of open and globular clusters:

$$\rho_{\text{King}}(r) = \rho_0\left[\frac{1}{\sqrt{1+(r/r_c)^2}} - \frac{1}{\sqrt{1+(r_t/r_c)^2}}\right] + c\ ,\ \text{where } r \le r_t$$

$$\rho_{\text{King}}(r) = c\ ,\ \text{where } r > r_t.$$

Here, ρ_0 is a normalization factor, r_c is known as the core radius of the cluster, r_t is the tidal radius (where the cluster is truncated), and c is a constant which corresponds to the number density of unassociated foreground/background stars (which may be very small for these data, since the cluster stars have already been pre-selected based on association in astrometric parameter space).

Fit the King model to the stellar density profiles of your three clusters to obtain the MLEs for the parameters and a goodness of fit of the model where appropriate. Plot your data again with the model fits and residuals or ratios as appropriate.

Obtain 1-σ confidence intervals on the model parameters, and if the intervals are not closed (i.e. they include the bounds of the fitted parameters, or zero or infinity), obtain 3-σ upper or lower (as appropriate) limits on the parameters.

3. An interesting question is whether or not the stars in a cluster follow the same stellar density profiles when they are selected according to different properties. In particular, we would like to see if the populations are well mixed according to their G band magnitude and their absolute proper motion (combining RA and DE proper motion directions). If they are well-mixed, subsamples selected on the given quantity will show similar density profiles. For each of these two quantities, split the stars from each cluster into two subsamples corresponding to different percentile ranges (your choice) of the given observable quantity. Then for each cluster, fit the resulting two density profiles together and determine whether the stellar density profile model parameters depend on the chosen quantity.

4. Making some plausible assumptions about the mass distribution in the cluster, we can use the virial theorem[1] to estimate the mass M of the cluster as follows:

$$M = \frac{3\, r_c \langle \sigma_{\text{pm}}^2 \rangle}{2G}$$

where G is the gravitational constant, r_c is the cluster core radius and $\langle \sigma_{\text{pm}}^2 \rangle$ is the population mean of the squared-velocity dispersion of the cluster stars which can be measured from the sample variance of the cluster stars proper motions[2]. To calculate the mass, these need to be turned into the correct physical units, using the distance to the cluster d.

Use Bayes' theorem to write out the equation relating the posterior distribution of mass M to the likelihood of the data given the parameters r_c, d and $\langle \sigma_{\text{pm}}^2 \rangle$. Then use MCMC to fit the complete dataset for one of your clusters (your choice) with appropriate priors on M, r_c, d and $\langle \sigma_{\text{pm}}^2 \rangle$ and use the results to obtain the MLE and 1-σ confidence interval on the mass. Be sure to calculate the mass in physical units (kg, g or solar masses are all fine) and be careful to explain your reasoning and state your assumptions clearly!

Hints: you can use your prior data to determine the priors on d and $\langle \sigma_{\text{pm}}^2 \rangle$. For d you can either use your own Bayesian estimate of the posterior distribution from the parallax measurements, or simply invert the weighted mean parallax (and propagate its error) which should provide a reliable estimate of the posterior, given sufficient data. For the prior on $\langle \sigma_{\text{pm}}^2 \rangle$ you can assume the MLE is given by the observed σ_{pm}^2 for the stars and estimate the standard deviation on the observed σ_{pm}^2 and assume the CLT to give the appropriate distribution.

---

[1] which states that the total potential energy U and kinetic energy K of the cluster stars is related by $2K + U = 0$

[2] $\sigma_{pm}^2 = \sigma_{pm,RA}^2 + \sigma_{pm,DE}^2$