Statistical Methods 2024: Assignment 2 – correlations and Bayesian distance estimation with the Gaia cluster stars sample

This is the second of three assignments, for which extensive help is available during the tutorials. It is worth 15% of the final course grade.

The deadline for this assignment is Monday 22 Jan at 12:00 noon. Late work which is submitted up to 24 hours after the deadline will receive a penalty of -20% of the awarded grade, while work submitted 24-48 hours after the deadline will receive a penalty of -40% of the awarded grade. Work submitted later than this will not be assessed. 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 assignment, you will continue to use the Gaia cluster stars data set from Hunt & Reffert (2023), which is described in more detail in the Assignment 1 explanation. In Assignment 1, you examined the properties of individual clusters to search for spatial variations of observed quantities such as proper motion, parallax and optical emission and colour, across specific clusters and over the general population. In Assignment 2, you will extend your study to look at the correlation between astrometric quantities calculated for the clusters as a whole, such as mean parallax, mean proper motion, cluster angular size and parallax and proper motion velocity dispersion.

Initial setup

First, load in the stars data as shown in Assignment 1, and create a dataframe containing only stars with `Prob` > 0.8, which you will use for the remainder of this assignment.

Assignment tasks

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

1. Use your stars dataframe to calculate the following 6 sample quantities per cluster: the number of stars in the cluster n_*; the mean parallax, $\overline{\omega}$, of the stars in the cluster; the standard deviation σ_ω of the parallax of the stars in the cluster; the 'size' of the cluster $\sigma_{\rm pos}$ calculated using the standard deviation in RA and Dec position of the stars[1]; the mean proper motion of the cluster stars $\bar{\delta}$; the standard deviation of the proper motion of the cluster stars σ_δ (which can be calculated by adding in

[1] First calculate the standard deviations of the star positions for each coordinate σ_{RA}, σ_{DE} and then use: $\sigma_{\rm pos} = \sqrt{\sigma_{RA}^2 + \sigma_{DE}^2}$.

quadrature the RA and dec proper motion standard deviations, i.e. the same as for $\sigma_{\rm pos}$). To help you do this in Pandas, you can adapt the code you used for Assignment 1:

```
clusters_hiprob = stars_hiprob.groupby(['Name']).size().reset_index(name='n_star')
clusters_sd_hiprob = stars_hiprob.groupby(['Name']).std(numeric_only=True).reset_index()
```

where the `.std` method calculates the sample standard deviations for all numerical quantities in the stars data frame grouped according to cluster name, and a similar method exists for the sample mean. Those can then be used to add the required columns to the dataframe with n_*.

Now, select only the clusters which satisfy $n_* > 200$, $\sigma_{\rm pos} < 1^\circ$and for these clusters plot the 5 astrometric quantities only (i.e. excluding n_*) in a scatter-matrix plot, to compare each pair of quantities and show their histograms. Comment on whether there are any clear correlations revealed by the scatter plots.

2. Correlation tests:
    a. Now search for correlations by calculating the Pearson and Spearman correlation coefficients and *p*-values for each of the combinations shown in the scatter-matrix plots. Comment on whether the tests are appropriate given the observed distributions of the measured quantities.
    b. You can improve the situation by performing the tests on log-transformed data (i.e. on the logarithm of the quantities instead of the original values). Why should this make the tests more reliable? Finally, comment on your results and their implications.

3. Now randomly select a cluster from the sample used in Task 2. Using the parallaxes of the individual stars in the cluster, use Bayes' theorem to calculate the posterior pdf for the distance d (in kpc) to the cluster, using the formula $d = 1/p$ where p is the parallax in milliarcsec (mas). ***Gaia has a known 'zero-point' offset - a systematic error – in the parallax, so before you do your calculation you should first add a correction of 0.029 mas to the parallax measurements.*** You should calculate the posterior pdf for two different prior pdfs:
    a. A uniform prior pdf.
    b. A more realistic pdf corresponding to constant volume density modified by an exponential decrease with distance:
    $$p(d) \propto d^2 e^{-d/L}$$
    where the length-scale $L = 1$ kpc.

    You may assume that the corrected parallax measurements are normally distributed about the true parallax, with standard deviation given by the errors on the parallax measurements. For each prior, calculate the distance corresponding to the mode (maximum) of the posterior pdf, plot your posterior pdf (on the same plot for both priors) and determine the 1-σ confidence interval on the distance and indicate the mode and interval on your pdf. Also on the same plot, compare the Bayesian posterior pdfs with the pdf you obtain just by assuming the normal distribution with mean obtained by inverting the mean parallax for the cluster, $\overline{\omega}$ (to obtain d) and standard deviation obtained by propagating the standard error on $\overline{\omega}$, to obtain the standard deviation on d.

    Finally, repeat the analysis above using just 10 stars from the cluster (you should pick a random sub-sample), to see the effect of smaller numbers.

    Hint for fast numerical calculation: to output a pandas data column to a numpy array which you can reshape as needed, use the `.values` method.

4. Now repeat the Bayesian distance calculation for all clusters used in the correlation sample in Task 2, using only the more realistic, constant density exponentially decreasing prior. You should obtain the distance estimate for each cluster from the maximum posterior probability of d. Then, use your distances to correct these 4 observed quantities to the values they would have at a fixed distance of 1 kpc: σ_ω , $\sigma_{\rm pos}$, $\bar{\delta}$, σ_δ. For these four corrected quantities, show the scatter-matrix plot and perform Pearson and Spearman correlation tests. Comment on your results and how they differ compared to what you obtained in Task 1 and 2 for the same quantities.