Statistical Methods 2024: Assignment 1 – Exploratory analysis of Gaia cluster candidate stars

This is the first of two assignments, for which extensive help is available during the tutorials. It is worth 25% of the final course grade.

The deadline for this assignment is Wednesday 14 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! Advice: before you submit, restart your kernel and run through all the cells in order, to make sure they run and produce the expected output. Questions that ask you to write text should be submitted as text cells within that same notebook.

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!

Some questions ask you to make a prediction about what you will see in the data before you look at it: **there will be no points deducted for the prediction being incorrect**! You can achieve full points with an incorrect prediction, and task completion and quality will be judged based on whether you applied the statistical reasoning you are learning in the class to motivate your prediction.

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.

You may use AI for this work in the following ways:

- To help you **implement code**: You may use AI to figure out the precise syntax for rote programming tasks, such as plotting and database manipulation, but **do not ask for and copy full solutions to assignment questions**
- to help you **debug code**: you may ask the AI for hints, explanations, or checks, to help you think through error messages, but **do not ask for or simply copy full solutions**
- to **help you understand concepts (e.g. for astrophysical quantities and processes you may not know) and think through mistakes**, but AI **cannot and must not decide for you** which assumptions are reasonable, which tests are appropriate, or how to interpret results.

You may **not** ask AI to generate full functions for you, or to provide full answers to the questions. **To do so is considered academic fraud and will be addressed as such**. See the course guide on Canvas for more details.

Note that you are responsible for the correctness of all work you submit, no matter its source. You must be able to explain your solution in detail if asked. You must flag any support by AI clearly and specifically (e.g. if it helped you figure out code, make sure you specify how in a docstring or text cell).

Any written answer must explicitly reference at least one figure, table, or numerical value produced by *your own code*. Answers that could plausibly have been written *before running the notebook* will receive no credit.

The Assignment:

For this assignment you will be using a dataset of astrometric and photometric data for nearly 1.3 million stars observed by the Gaia mission, which have been identified as belonging to more than 7000 star clusters and looser associations in our galaxy. Gaia, launched in 2013, is ESA's successor to the Hipparcos astrometry satellite, data which is used in the 'Extras' episode to introduce Pandas (which will also be useful for this assignment). Gaia continuously scans the sky using two telescopes set at a precisely known (via an on-board laser interferometer) angle to each other. The relative positions of many stars, focused on to two CCDs allows a precise astrometric solution that can measure the angles between the stars down to a precision of tens of microarcseconds – note that one microarcsec is $\approx 5 \times 10^{-12}$ radian! Besides measuring extremely accurate positions on the sky, Gaia provides measurement of the star's proper motion – its movement on the sky – and parallax – the annual angular motion against distant background objects as the Earth (and satellite) moves in its orbit. The parallax can be used to directly estimate the distance to the star.

The data you will be using is taken from a recent paper by Emily L. Hunt & Sabine Reffert[1] which identifies and analyses star clusters found in the data from Gaia's Data Release 3 (DR3), a database of 1.46 billion sources (c.f. 2.5 million for Hipparcos!) which includes 5D astrometric data (RA and Dec positions, proper motion on the sky and parallaxes to measure the distances), as well as photometry in 3 optical bands[2]. Hunt & Reffert use a sophisticated clustering analysis of the astrometric information to identify which objects are associated with one another, e.g. in globular and open clusters or looser associations of stars.

Initial setup

The data for each star which is a candidate cluster member is contained in the FITS file gaiadr3_cluster_stars.fits. Besides the descriptions in the FITS header, the data columns are described in the additional file data_description.txt.

To load in the FITS file you need to have installed the astropy package. If you do not have it yet, you can install it using conda or pip depending on whether you use Anaconda python distribution or not (see here). Then in your first cell add:

```
from astropy.io import fits
```

and to load in the data, you can use e.g.:

```
dr3stars = fits.open('gaiadr3_cluster_stars.fits')
dr3stars.info()
print(dr3stars[1].columns)

stars = pd.DataFrame(dr3stars[1].data)
stars['Name'] = stars['Name'].str.strip()
```

which will open the FITS file, provide information about the file structure and columns in the data table, and then (if you use Pandas, which we recommend) assigns the table to the Pandas dataframe stars (you can choose your own dataframe name if you wish). The last line of code strips the trailing white spaces from the original cluster names given in the Name column (which are 20 characters long including trailing white spaces). This fix makes it easier to obtain data for a given cluster name, without having to pad the string with white spaces to obtain a match.

This should be all you need to start working with the data using Pandas, scipy, numpy etc., e.g. using the commands described in the Extras episode on Working with and plotting large multivariate data sets.

---

[1] https://ui.adsabs.harvard.edu/abs/2023A%26A...673A.114H/abstract - it is not necessary for you to read this paper in order to complete the assignment!

[2] https://www.cosmos.esa.int/web/gaia/iow_20180316

However, for more discussion on working with FITS files in Python, see this Episode from the Programming for A&A course.

For the following tasks it will be useful to have information on the number of stars associated with a given cluster name in a pandas dataframe, and information on the extent of the cluster in the RA and Dec directions. You can obtain this information quickly using the following code:

```
clcounts = stars.groupby(['Name']).size().reset_index(name='count')
```

which creates a dataframe `clcounts`, where each column shows the number of counts which is identical to the number of stars associated with that cluster name in the stars dataframe. You can use this information to find clusters with a certain minimum number of stars.

Assignment tasks

There are 25 points available total. The points awarded to each question are printed at the start of each question.

1. **Exploratory Data Analysis (6 pts):** The parameter `Prob` gives a conservative estimate of the probability that the star is associated with the cluster, by doing a 'clustering'[3] analysis of the stars in the 5-dimensional astrometric parameter space, i.e. using `RAdeg, DEdeg, Plx, pmRA` and `pmDE`.
    - **a.** Use the Pandas sample function on your cluster star counts dataframe, to randomly select 4 clusters, only from clusters with >1000 candidate stars. Please make sure to **include the random seed** you used in the notebook for reproducibility. Be sure to include the names of the clusters in your notebook!
    - **b.** Split each cluster into two subsamples corresponding to stars with `Prob ≤ 0.8` and stars with `Prob > 0.8` and for each cluster, make a scatter plot matrix (see the extras episode, or use libraries that have this functionality like `corner` or `seaborn`) to show **on the same figure** the data points for both subsamples on this 5-D parameter space. Use the same axis limits for `both Prob ≤ 0.8` and `Prob > 0.8` and use transparency (`alpha < 0.3`). For **each cluster**, answer the following questions, and explain the reasoning for your answer. For each answer, include one quoted observation (e.g. "the clouds overlap almost entirely in pmDE"), and one Figure number where that observation is apparent:
        - **i. Which pair of dimensions most strongly separates high- and low-Prob stars?** Choose exactly one: RA–Dec, Parallax–pmRA, pmRA–pmDE
        - **ii. Which dimension contributes least to separation?** Choose exactly one: RA, Dec, Parallax, pmRA, pmDE
        - **iii. Which statement matches best the evidence in your plots?**
            1. The algorithm primarily exploits spatial clustering (i.e. in position)
            2. The algorithm primarily exploits kinematic coherence (i.e. in stellar motion)
            3. **The separation is weak and likely noise-dominated.**
    - **c.** **Self-check:** Briefly, answer the following two questions: Could this explanation have been written without looking at any Gaia data? If yes, go back and make sure your answers are backed up with observations from your figures. Which observation changes the most between different clusters?

2. **Exploratory Data Analysis and Hypothesis Testing (6pts):** We would like to select clusters that are tightly clustered together, to rule out looser associations which will have a wide spatial variance in their parameters due to the range of distances from us. To select such clusters, you can use the following lines of code to create a new stars dataframe containing only the higher-probability stars based on a threshold and to create a new clusters dataframe with the star counts and standard deviations in RA and Dec for each cluster:

    ```
    p_thres = 0.8 # !! select your threshold here !!
    ```

[3] Of data points in the parameter space, not stars!

```
stars_hiprob = stars[stars.Prob > p_thres]
clusters_hiprob =
stars_hiprob.groupby(['Name']).size().reset_index(name='count')
clusters_sd_hiprob =
stars_hiprob.groupby(['Name']).std(numeric_only=True).reset_index()
clusters_hiprob['sd_RAdeg'] = clusters_sd_hiprob['RAdeg']
clusters_hiprob['sd_DEdeg'] = clusters_sd_hiprob['DEdeg']
```

Now use the new `clusters_hiprob` dataframe to make a new cluster sample containing only clusters with > 200 stars and standard deviations of RA and Dec < 0.1° (to constrain the cluster size). An interesting question is whether there is any spatial (RA and Dec) variation of the other astrometric and photometric parameters in each cluster.

a. Randomly select one filtered cluster, note down its name and the random seed you used in the notebook. Before making any new plots, based on your previous analysis, answer:
    i. Will the parallax distribution vary with RA?
    ii. Will proper motion dispersion increase, decrease or stay constant with RA and Dec?

   Justify each choice in one sentence, noting down which previous figure/result you base your prediction on. To be clear: **incorrect predictions will not be penalized**, you will be graded on the **coherence of your argument**, not on the correctness ofE the prediction.
b. Now split it into 2 subsamples in `RA`, corresponding to stars with `RA`: i) greater than the mean `RA`, ii) less than the mean `RA`. Do the same for Dec, to create 2 subsamples selected on mean `Dec`. Then for the `RA` -selected subsamples, plot a figure with 5 separate subplots (e.g. side-by-side) which show the histograms of the following parameters for each of the 2 subsamples: `Plx, pmRA, pmDE, Gmag` and `BP-RP`. Each subplot will show two histograms, one for each subsample, so you can compare the distributions for stars on one side of the cluster vs the other. Repeat this for the 2 subsamples selected on `Dec`. Answer the following question: Did your predictions agree with the data? Motivate your answer based on the figures you've made. If not, state one possible reason for the observed discrepancy.
c. Use hypothesis testing via t-tests to compare the 2 subsamples in `RA` and then the 2 subsamples in `Dec` for the following parameters: `Plx, pmRA, pmDE, Gmag` and `BP-RP`. For the t-test you can assume populations with the same variance. You will do 5 t-tests for the subsamples selected on `RA` and 5 for the subsamples selected on `Dec`, to see if there is any evidence that the populations of stars which each subsample is drawn from is different from the other subsample, i.e. does it change with position in the cluster.
    i. Identify one test result that you believe may not reflect reality. Fill in the blanks in this sentence: "This test is misleading because the histograms in Figure ___ show ___, which violates the assumption of ___." If you don't believe any test results are in conflict with what you see in the Figures, fill in the blanks in the following sentence. "No tests are misleading. For example, the histograms in Figure ___ show ___, which is in line with the assumption of ___ required by the t-test."
d. Based on the parameter distributions you plot, is the t-test an appropriate test in all cases?

3. **Hypothesis testing: (7pts)** Now you will analyse all the clusters in your new sample (i.e. star counts > 200, constrained in RA and Dec standard deviation).
    a. Before computing any p-values, choose exactly one expectation about the distribution of p-values across your sample of clusters, based on the table below (include your expectations in your Jupyter notebook). Justify your choices, each in one sentence. Reference one figure or observation from a previous exercise to justify your expectation.

Test	Expected p-value distribution
Plx (RA split)	☐ Uniform ☐ Skewed low ☐ Skewed high
pmRA (RA split)	☐ Uniform ☐ Skewed low ☐ Skewed high
BP–RP (Dec split)	☐ Uniform ☐ Skewed low ☐ Skewed high

b. Control test: Before looking for spatial variation, validate your method. For every cluster, randomly shuffle the stars into two groups (ignoring RA/Dec) and perform the t-test on `pmRA`. Plot the histogram of these p-values. *Self-Check:* If this histogram is not roughly uniform, debug your code before proceeding.
c. For all clusters in your sample, perform the t-tests from exercise 2c on Plx, pmRA and BP-RP, with both RA and Dec splits. For each parameter and coordinate (RA or Dec) being tested, plot the resulting p-values in a histogram (so you have 6 histograms in total, each calculated from N_c p-values where N_c is the number of clusters analysed). Use identical binning, include a vertical line at `p=0.05` and produce figures that span [0, 1] in `p`. Include the number of clusters used in the Figure title.
d. For at least one of the three expectations you justified above, answer the following questions (make sure to include references to at least one specific figure and to features within that figure):
    i. Which description best matches the histogram? (1) consistent with uniform, (2) excess of small p-values, (3) depleted at small p-values
    ii. Which explanation is the most plausible, and which is least plausible: (1) Genuine spatial substructure, (2) violation of test assumptions, (3) selection effects from earlier cuts, (4) pure chance
e. You performed N_c×3 tests. How many p-values <0.05 did you find? How many would you expect by pure chance if there were no spatial differences? Based on this, estimate the **False Discovery Rate** (Observed Positives / Expected False Positives). Is `p<0.05` a good threshold? What would you choose? Justify your choice based on one specific figure (include the Figure number you are referring to in your answer) or on one specific number you have calculated.
f. Should you base your scientific conclusion about the spatial distribution of stellar properties on the tests you've performed above? Why or why not? Motivate your answer using the at least one of the quantities and figures you've generated above.

4. **Hypothesis testing significance versus sample size: (6 pts)** In very large datasets, even tiny, physically irrelevant deviations can produce extremely small p-values because the standard error of the mean shrinks as $1/\sqrt{N}$. We need to determine if our "significant" results are physically meaningful or just a byproduct of having a lot of data. For this exercise, **use the original `clusters_hiprob` DataFrame** (i.e. without filtering for RA and Dec) in order to have a sufficient number of clusters.
    a. Select one parameter where you previously found an excess of small p-values and create a scatter plot of the number of stars in the cluster N_c against the p-value from your t-test for each cluster. Plot every cluster in your sample as a single point. Make one plot with the y-axis in linear space and one with the y-axis in log-space. Keep the x-axis in log-space both times.
    b. Calculate the correlation coefficient (e.g., Pearson or Spearman) between N_c and the p-value (hint: this may work better in the logarithm of N_c and p). Does the correlation suggest that larger clusters are "more significant" (lower p-values)?
    c. Identify the cluster with the lowest p-value in your sample. For this specific cluster, plot the distributions and calculate the actual difference in means between the two subsamples: $\Delta\mu = |\mu_1 - \mu_2|$
    d. Compare this $\Delta\mu$ to the dispersion σ of the cluster.
    e. Based on the scatter plot and the effect size $\Delta\mu$, do you believe the spatial variations in these clusters are **astrophysically important** (indicating real structure like rotation/tides) or merely **statistically detectable** due to large N_c?