Astronomía
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

AS4501-1 2023A
Astroinformática

HOMEWORK 1

Due date: 23/04/2023 22:59 hrs

> Homeworks in this class are turned in as Jupyter Notebooks. Use them to show your code, figures, explanations.

Problem 1

The *Fundamental Plane* is an empirical relation for Elliptical galaxies. It relates two distance-independent galaxy observables, and another observable that depends on distance. Because of this, it has been used in the past to estimate distances to galaxies. Perhaps more importantly, the fact that this relation exists at all, is indicative of something profound about how galaxies form and evolve.

The file `fundamental_plane.csv` contains data compiled by Djorgovski & Davis (1987) that can be used to infer the fundamental plane. In specific, we will be working with the following columns:
`log_re_pc`: logarithm of the effective radius in units of pc.
`log_sigma`: logarithm of the central velocity dispersion in units of $\mathrm{km\,s^{-1}}$.
`mu`: mean surface brightness inside de effective radius in units of magnitude per square arcsecond.

We will ignore the measurement uncertainties in this problem (you can use them if you want to be extra thorough but we will not require it for grading). In classes we have seen linear regressions applied to problems with a single variable but here you will work with a two parameter linear model. Your task is to model $log(r_e)$ as a linear function of $log(\sigma)$ and μ.

1. Use an ordinary linear regression to model $log(r_e)$ as a function of $log(\sigma)$ and μ.

   - Separate the sample into a training and testing sample. Take a look at:

   `from sklearn.model_selection import train_test_split`

   - Attempt models of increasing degree (go at least up to 5^{th} degree). For this, take a look at:

   `from sklearn.preprocessing import PolynomialFeatures`

   and PolynomialFeatures

   - Use what you learned about cross-validation to decide what degree linear function is best. In particular, show that the MLE for the training sample always decreases with increasing degree of the function being fit but the testing sample does not behave the same way.

2. Now stick to a first degree linear model and perform a Lasso regression and a Ridge regression. Take a look at the coefficients of the linear fits. Can you explain the differences?

3. The data set may contain one or two outliers. Perform a Hubber loss regression (Why?). How did you choose the hyper-parameter? How does it compare to the regressions done before?

It is difficult to visualize these 3 dimensional relations. One possible strategy is to plot $log(r_e)$ in the y-axis, and in the x-axis place the linear function of the other two parameters, $log(\sigma)$ and μ (e.g., $log(r_e)$ vs. $(A \times log(\sigma) + B \times \mu)$, where A and B are the best fit constants). That way, you expect to obtain a 1:1 relation and you can judge visually how good or bad is your fit or if there are obvious outliers. We do not request figures explicitly but you should include some that you think are relevant.

Problem 2

The file `M_sigma.csv`, contains the data used by Harris et al. (2013). The data corresponds to estimates of $M_\bullet$: the mass of the supermassive black hole in the center of a galaxy, in units of $M_\odot = 2 \times 10^{33}$ g; and σ: the velocity dispersion of the stars in the bulge of those same galaxies, in units of $\sigma_0 = 200\ \mathrm{km\,s^{-1}}$. The data was used by Harris et al. (2013) to explore the so called $M_\bullet$–σ relation, a very important observational constraint that relates the properties of the supermassive black hole in the center of a galaxy (a parsec sized region) with the global potential of the galaxy (on kiloparsec scales...–by the way, in case you are looking for an interesting scientific project, it is not understood how such a correlation arises). Here you will use the same data to derive the $M_\bullet$–σ relation yourself. The linear relation to fit has the following shape:

$$\log\left(\frac{M_\bullet}{M_\odot}\right) = \alpha + \beta \log\left(\frac{\sigma}{\sigma_0}\right) + \mathcal{N}(0, w^2)$$

In the above expression, $\mathcal{N}(0, w^2)$ represents the intrinsic scatter of the relation, assumed to be gaussian and constant, parametrized by w.

Your task is to perform a full Bayesian analysis to find the joint probability distribution for the three parameters of this model: (α, β, w). The analysis must including uncertainties in both variables ($M_\bullet$, σ), possibility of outliers, and the intrinsic scatter. The following article may prove useful: https://ui.adsabs.harvard.edu/abs/2010arXiv1008.4686H/abstract. It is also discussed in Ch. 8.8 and 8.9 of the book "Statistics, Data Mining and Machine Learning in Astronomy" by Ivezic et al.

Problem 3

It is important to develop an idea of the intricacies associated with the information provided by large surveys in their databases. Here we will estimate the depth of an SDSS image to get a sense for the uncertainties associated with the photometry that the SDSS database provides. To that end:

1. Download the r-band SDSS dr7 (data release 7) image centered at coordinates: `RA 13:29:52.7, DEC +47:11:43s`. Download a region of 20 × 20 arcmin with a resolution of 0.6 arcseconds per pixel. The `python` package `astroquery.skyview` is recommended for this. Beware that the file size will be fairly large, it may take a while to download. If your computer has trouble handling it, download a smaller image and specify in your solution.

2. We will estimate the depth of the image in 3" diameter apertures. For that, you will have to place a large number of apertures in random positions of the sky and study the distribution of the fluxes within those apertures. Avoid the beautiful galaxy in the center (this is the Whirlpool galaxy). The `python` package `photutils` can be useful. Note that the image is not

background-subtracted, so, for every aperture you must estimate the local background using an annulus (see here).

Note: the transformation from the image units into magnitudes is not straightforward for these images. The zeropoint should be 28.2576 but there are several other complications that need to be taken into account to obtain accurate photometry (airmass at which the image was taken, for example). If you are interested, you can find all the information necessary here but to keep things simple, we will only care about the S/N, which should be independent of the units of the image. Why?

3. Report the 5σ limiting *flux* (in whatever units the image is in) for 3 arcsec-diameter apertures.

4. Now pick a few *compact* sources (our circular apertures work best with stars) in the field and compare your estimates of their S/N with those reported by SDSS. To do this, you have to identify the position of the sources you are interested in and query the SDSS tables for their magnitudes and uncertainties. You can do this manually in the SkyServer webpage but you should try to do it programmatically.

5. Do you see differences? What explains the differences?