Homework # 4

PHY-905-005
Computational Astrophysics and Astrostatistics
Spring 2023

This assignment is due by 11:59 p.m. on Friday May 5, 2023.

Instructions: Turn in all materials via GitHub. Include your code, plots, and answers to any questions asked in your assignment. Your code must adhere to the class coding standards, use text files rather than Jupyter notebooks, and pass a PyLint test with an acceptably high score. The solutions to individual parts of the assignment should be saved in separate, clearly-named subdirectories named `part_1`, `part_2`, etc. Plots should have easily readable and logical axis labels and titles, and the source code and data used to generate the plots should be included. Questions should be answered in a file in each subdirectory named `ANSWERS.md` or in a LaTeX-created PDF document of a similar name (e.g., `ANSWERS.pdf`).

Part 1: The galaxy luminosity function is a measure of the number of galaxies per luminosity (or magnitude) interval, and its behavior over time and in different galactic environments is an important probe of the astrophysics of galaxy evolution.

The galaxy luminosity function is typically described as a Schechter luminosity function, which takes the form

$$n(L)dL = \phi^* \left(\frac{L}{L^*}\right)^{\alpha} e^{-L/L^*} \frac{dL}{L^*} \tag{1}$$

where n(L) is the number density of galaxies at a given lumosity L, α is the slope of the power law, L^* (typically pronounced “L star”) is a characteristic luminosity where the power-law component of the function cuts off, and ϕ^* is a normalization parameter with units of (typically comoving) number density. Alternately, this equation can be written in terms of magnitudes[1]:

$$n(M)dM = (0.4\ \ln(10))\phi^* (10^{0.4(M^*-M)})^{\alpha+1} e^{-10^{0.4(M^*-M)}} dM \tag{2}$$

Typical measurements for the constants in this equation in low-redshift field galaxies[2] are $\alpha \simeq -1.25$, $\phi^* \simeq 1.2 \times 10^{-2}$ h^3 Mpc^{-3} (with h being the Hubble parameter in units of 100 km/s/Mpc), and $M^* \simeq -23$.

In this problem we are going to experiment with the significance of assumptions about Bayesian priors, the amount of data, the magnitude of errors in data, and the impact of these things on the resulting model fits. The file `luminosity_schechter.py` contains two importable functions that have detailed doc strings explaining their inputs and outputs. One, `luminosity_function()`, creates a synthetic “observed” luminosity function where you can control the Schechter function model parameters (α, ϕ^*, and M^*), number of bins, and magnitude of the errors for each bin, and get back numpy arrays of absolute magnitude bins, comoving number density, and errors in number density. The included figure `galaxy_luminosity_function.png` has an example of the output from this function. The second function, `schechter_function()`, takes in a numpy array of magnitude bins and values for the Schechter function model parameters and returns number densities for each bin.

Using the Bayesian Markov Chain Monte Carlo code we wrote in class as your starting point, conduct the following experiments and report the outcome in an appropriately-named file. Use a reasonable likelihood function for comparing your model and data (reduced chi-square is fine) and create a function describing the Bayesian prior distribution for your input model parameters α, ϕ^*, and M^*, assuming a lognormal distribution for each model parameter with a user-specified mean and variance.

[1] A unitless, logarithmic measurement of brightness of objects that is used in astronomy. Its invention predates telescopes and its continued usage is traditional but difficult to justify in the 21st century, not unlike the CGS unit system.

[2] “Field galaxies” mean galaxies that are in a typical cosmic environment; galaxies in high-density environments (such as clusters) or low-density environments (such as in void walls) typically have quantitatively different luminosity functions, though the functional form is generally the same.

The experiments you will conduct are:

1. Generate an "observed" luminosity function using the default number of magnitude bins ($n = 10$), model parameters, and error levels. For your priors, assume that the mean is the "correct" model parameter (i.e., the ones that generated the input dataset) and choose a large variance for each model parameter (a value of roughly 1 for α; a factor of a few for ϕ^*; several magnitudes for M^*). Use your MCMC code to find the "best-fit" model parameters and their probability distribution functions, and create a figure or figures showing (1) the posterior distributions for the model parameters and the "triangle plot" showing their relationships and (2) a plot of the "observed" luminosity function overplotted with the model your MCMC code declares is the most likely. Briefly comment on the results – do the things that you observe make sense to you? This is your "fiducial" result and model parameters – i.e., the result we're going to use as a basis of comparison for most of the remaining experiments.

2. Repeat experiment #1, but increase the number of model bins from 10 to 20, and then try decreasing it to 5 (keeping all other parameters the same as the fiducial model). As you increase the number of model bins, how does it affect the posterior distributions for your model parameters compared to the fiducial result? Why do you think that is?

3. Repeat experiment #1 with the fiducial number of bins, but now try decreasing and increasing the errors in the "observed" luminosity function by a factor of a few. What happens to the posterior distribution, and why do you think that is?

4. Repeat experiment #1 with the fiducial number of bins and error level, but now modify your prior distributions. Set the mean values of your priors to be far away from the "correct" model and significantly reduce the magnitude of your variance – the goal is to ensure that the correct model is now several standard deviations away from the mean of your prior. How does this affect the posterior distributions, and why do you think that is? And, if you modify your prior distribution so that it is closer to the correct model by adjusting the mean value, how does the posterior distribution change?

5. Repeat the previous experiment (with modified priors), but now try changing both the number of model bins and the error level. Qualitatively, how do your posterior distributions behave as you change these properties of your "observed" dataset?

6. In the document describing your answers, reflect on the sum total of what you've observed as you experiment with the various components of Bayesian MCMC model fitting. What intuition have you developed about the use of Bayesian priors in Markov Chain Monte Carlo? In particular, comment on the relationships between the amount of observed data, the magnitude of the errors in observed data, your assumptions about the prior distribution, and the measured posterior distribution.

Part 2: We're going to examine exoplanet data obtained using two different methods – by the transit method (using, e.g., the Kepler mission), and by using radial velocity measurements of the central star. For the transit method we will use the star KIC-8554498 (in the data file `KIC-8554498_lightcurve.dat`, which contains the relative flux of the star as a function of time since the start of the Kepler mission) and for radial velocity measurements we will use 51-Pegasi (in the data file `51_Pegasi_RadVel.dat`, which contains the Julian date, the star's radial velocity, and the error in the radial velocity).

For each of the two files, do the following:

1. Plot the light curve or radial velocity data for each of the two planetary systems, zooming in as necessary to show interesting features. Can you see by eye the evidence of exoplanets in the data? If so, what do you see?

2. Use the SciPy or Astropy Lomb-Scargle periodogram routines to identify the orbital periods of the exoplanets orbiting these stars. Include your plots, zooming in on regions as necessary. Note any interesting features in the periodogram. Are the results consistent with the expected periods of the planets orbiting 51 Peg ($P \simeq 4.23$ days) and KIC-8554498 ($P \simeq 4.78$ days)? What features do you see in these graphs, and what do you hypothesize causes them?

Part 3: We're going to continue using the data for 51-Pegasi, which was the first confirmed discovery of a planet outside of our own solar system. Assuming a period of $P = 4.230785$ days, created a stacked radial velocity curve that folds all of the orbital data you have been provided into a single period. Then, create a model for the observed radial velocity curve of a single planet in a circular orbit around a star (you may wish to start here), and use a Bayesian Markov Chain Monte Carlo algorithm to determine the planet's mass (really $M \sin(i)$), the semi-major axis of the orbit, and the orbital period. Take the mass of 51 Pegasi as a given ($\mathrm{M}_* = 1.06\ \mathrm{M}_\odot$), and assume reasonably broad priors for the other quantities. Use a reduced sum of squares to estimate the agreement between the model and the data. Using the data and the errors provided, can you reproduce the values for the period, semi-major axis of the orbit, and mass of the planet? Show 2D histograms for various combinations of the unknowns, and identify any degeneracies between model parameters that may exist.