Statistics Mini-course Problem Set

Due on Mar. 9

The exercises in this problem set must be solved on a computer and the best way to hand in the problem set is as an `jupyter notebook`. Rather than sending me the notebook, you can upload it to `GitHub`, which will automatically render the notebook. Rather than starting a repository for a single notebook, you can upload your notebook as a `gist`, which are version-controlled snippets of code that can optionally be made private.

If you want to upload your notebook as a gist from the command-line, you can use the package at this http URL and use it as follows. Log into your `GitHub` account:

```
gist --login
```

and then upload your notebook `statmini_2018_PS1_YOURNAME.ipynb` as

```
gist -p statmini_2018_PS1_YOURNAME.ipynb
```

(the `-p` option will make the gist private). If you want to make further changes, you can clone your gist in a separate directory and use it as you would any other git repository. Please make sure that the input and output are fully consistent (if you are re-running cells etc.).

If you are unfamiliar with notebooks, you can also hand in a traditional write-up, but you also need to send in well-commented code for how you solved the problems. Thus, notebooks are strongly preferred :-)

Determining the Hubble constant through the distance ladder

As you probably know, contemporary cosmology is roiled by a discrepancy between the Hubble constant H_0 as determined from the early Universe—the CMB—and large-scale structure on the one hand and H_0 determined in the local Universe through direct measurements of the Hubble law on the other hand. To explore model building, fitting, and MCMC sampling in the context of a real astrophysical problem we will determine H_0 using the latest measurements of the distance ladder.

We will mostly use powerful, open-source software for the numerics, so first a warm-up exercise:

Problem 1: Write a Metropolis-Hastings sampler for a general one-dimensional probability distribution $p(x)$ with a Gaussian proposal distribution (characterized by a width parameter that should be passed to the code) that returns a sampling and the acceptance fraction. Test it with a Gaussian with zero mean and unit variance: plot a normalized histogram of the samples and compare it to the analytical PDF. Then apply it to sample a probability distribution consisting of the sum of two Gaussians with equal weights, unit variance for each, and means 0 and 10 (again plot a histogram of the samples and the analytical PDF).

Try to find a relatively high acceptance fraction.

Okay, back to the Hubble constant. We will use the Cepheid and SN Ia data from Riess et al. (2016), which you can find `at this URL`. We will determine H_0 in four (coupled) steps:

1. Obtain the distance to a galaxy that contains Cepheids. We will use the galaxy NGC 4258, the nucleus of which contains a disk of masers rotating in a Keplerian manner around the central black hole. Modeling of this system provides a (quasi-)geometric distance to NGC 4258: distance modulus $\mu_{\rm N4258} = 29.387 \pm 0.0568\,{\rm mag}$. We will use this as a measurement, assuming that the given uncertainty is Gaussian.

2. NGC 4258 contains a bunch of Cepheid variable stars. These stars follow a linear relation between their log periods, log metallicity, and apparent magnitude (the Leavitt law)
$$m_i^W = {\rm zp}_{W,\rm N4258} + b_W\ \log P_i + Z_w \Delta \log({\rm O/H})_i\,, \tag{1}$$
where m_i^W is the "Wesenheit" magnitude
$$m^W = m_H - 0.39\,(V-I) \tag{2}$$
of Cepheid i, P_i is the period in days, $\Delta \log({\rm O/H})_i$ the metallicity[1]; the model parameters are ${\rm zp}_{W,\rm N4258}$, b_W, and Z_w. Measurements of periods, metallicities, and apparent magnitudes of these Cepheids allow us to determine all parameters of the Leavitt law for NGC 4258 (without using its distance).

3. To obtain H_0 we need to determine distances to galaxies that are further into the Hubble flow than the galaxies in which we can find Cepheids. For this we use type Ia supernovae (SN Ia) which can be seen out to cosmological distances ($z > 1$ even) and which are so-called standardizable candles: from the properties of their lightcurves we can normalize these supernovae such that they all have the same absolute magnitude. The missing link in the distance ladder is given by galaxies which host both Cepheids and an SN Ia, which can be used to calibrate the SN Ia absolute magnitude. The Leavitt law applied to Cepheids in these galaxies allows use to determine their distances relative to NGC 4258: $(\mu - \mu_{\rm N4258})_j$ for galaxy j. The general Leavitt law for Cepheids i in galaxies j is
$$m_{i,j}^W = (\mu - \mu_{\rm N4258})_j + {\rm zp}_{W,\rm N4258} + b_W\ \log_{10} P_{i,j} + Z_w \Delta \log_{10}({\rm O/H})_{i,j}\,. \tag{3}$$
where the additional term $(\mu - \mu_{\rm N4258})_j$ accounts for the distance between galaxy j and NGC 4258.

4. After standardization, all SN Ia absolute magnitudes are the same and the difference between their apparent magnitudes in different galaxies is therefore only due to their different distance. We will use magnitudes measured in the 'B' band. Then
$$m_{B,j} = (\mu - \mu_{\rm N4258})_j + m_{B,\rm N4258}\,. \tag{4}$$

[1] Your instructor is not sure how the 'Δ' is defined here, but presumably it is with respect to the Solar oxygen abundance, 8.66 in the units of Table 4.

Using galaxy distances relative to NGC 4258 determined in the previous step, we can determine the apparent magnitude $m_{B,\mathrm{N4258}}$ of a SN Ia *if it occurred in NGC 4258* and, using the distance from 1), its absolute magnitude. By comparing this absolute magnitude with the redshift zero intercept of the redshift-magnitude relation of higher redshift SN Ia the Hubble constant can be determined. This intercept can be written as a_B and then

$$\log_{10} H_0 = \frac{m_{B,\mathrm{N4258}} - \mu_{\mathrm{N4258}} + 5a_B + 25}{5} \tag{5}$$

in units of $\mathrm{km\,s^{-1}\,Mpc^{-1}}$. We will not determine the intercept from higher redshift SN Ia ourselves, but simply use the measurement $a_B = 0.71273 \pm 0.00176$.

At this point it may be helpful to sketch out the relation between the different parameters and data points, distinguishing between Cepheids in different galaxies and SN Ia. Ready? Okay, let's implement the model and get the Hubble constant!

Problem 2: Grab the data for Cepheids in NGC 4258 from Table 4 of Riess et al. (2016; see link above). Assuming that there is no intrinsic scatter in the Leavitt law, the parameters of Equation (1) can be determined through *ordinary least squares*. Determine the best-fit parameters $\mathrm{zp}_{W,\mathrm{N4258}}$, b_W, and Z_w and their Gaussian covariance matrix (e.g., see Sec. 1 and Exercise 1 in `https://arxiv.org/abs/1008.4686`). Plot the data m_i^W vs. $\log_{10} P_i$ and overlay the best fit model (evaluated at the mean metallicity of the sample).

Problem 3: Now assume that the relation in (1) has intrinsinc scatter that can be assumed to be Gaussian with a constant variance V. Write down the likelihood and the posterior PDF for this model (use a reasonable prior for V and the other parameters). Write a function that returns the (natural) logarithm of the PDF and use `emcee` to sample the parameters of the model. Convince yourself that the MCMC chain is well mixed. Make plots of the posterior samples with `corner.py`. Again plot the data m_i^W vs. $\log_{10} P_i$ and overlay the best fit model (evaluated at the mean metallicity of the sample) and the 1σ band corresponding to the intrinsic scatter.

Problem 4: Using the data for Cepheids in the other galaxies we will fit Equation (3). We assume that the parameters of the Leavitt law ($\mathrm{zp}_{W,\mathrm{N4258}}, b_W, Z_w, V$) are the same for all galaxies (we will again assume intrinsic scatter and assume that it is the same in all galaxies). What are the model parameters? Given what we have done so far, what are good priors for ($\mathrm{zp}_{W,\mathrm{N4258}}, b_W, Z_w, V$)? Given that we have many parameters, `emcee` might not work too well, so we will use `stan` (e.g., through its Python interface `PyStan`) to fit and sample this model. Make diagnostic plots of the MCMC chain returned by `stan`. The main output that we are looking for are the relative distance moduli $(\mu - \mu_{\mathrm{N4258}})_j$. Using the distance modulus of NGC 4258 from point 1) above, compare your distance moduli to those from Table 5 of Riess et al. (2016).

Problem 5: Combine your measurements of $(\mu - \mu_{\mathrm{N4258}})_j$ from Problem 4 with the data in Table 5 of Riess et al. (2016) to determine the SN Ia magnitude in NGC 4258 $m_{B,\mathrm{N4258}}$ and the Hubble constant H_0 using Equations (4) and (5), respectively. Note that the SN Ia

magnitudes have uncertainties of their own.

Bonus Problem 1: Because steps 2) and 3) above involve the same Leavitt law, we should really treat the NGC 4258 data at the same time as the other Cepheid data and combine Problems 3 and 4. And to properly propagate the uncertainties in $(\mu - \mu_{\rm N4258})_j$ to the Hubble constant, we should just include Problem 5 as well, creating one big model connecting all of the data and all of the parameters. Write down this model in `stan` and sample it. How does your final value of H_0 compare to that in Problem 5?

Bonus Problem 2: The Leavitt law has outliers which we have not removed. One option for modeling these is with a mixture model (e.g., see Hogg, Bovy, & Lang 2010), but that can be somewhat difficult to sample. An alternative is to assume that the intrinsic scatter is not Gaussian, but instead has a distribution with wider tails. Explore this by modeling the Leavitt law using a Student t distribution of two degrees of freedom instead in the analysis above. This distribution is available in `stan`, so is straightforward to implement in a `stan` analysis.