Project #5: Galaxy Velocities

Our model of the Universe is that it is an infinite space filled with a uniform "gas" of galaxies. To a first approximation we can think of these galaxies as moving in random directions with each component of the velocity being drawn from a Gaussian distribution (surprise!) centered on zero (so each component has an equal chance of being positive or negative) with some standard deviation σ_v. Our goal in this project is to determine σ_v.

We measure galaxy velocities by measuring the shift in the spectral lines of a galaxy from the wavelengths we expect, called the redshift z, a dimensionless quantity that can be measured very accurately. The speed of light times this redshift, cz, gives us the velocity at which the galaxy appears to be moving away or toward us, usually expressed in km/sec. However, the redshift is actually a combination of the doppler shift, due to the galaxy's motion, which is what we want, and a cosmological redshift due to the light having travelled through an expanding of the Universe. Hubble's law tells us that the part of the velocity due to the cosmological redshift is given by $H_0 r$, where r is the distance in megaparsecs (Mpc) (an Mpc is about 3 million light years) and $H_0 = 70$km/sec/Mpc is Hubble's constant. Thus if we can measure the redshift z and the distance r we can obtain the radial velocity via

$$v = cz - H_0 r \tag{1}$$

Now, it turns out that measuring distance is difficult, and the uncertainty in distance measurements is 5-15%. For distant galaxies, where $H_0 r >> v$, this introduces crazy large uncertainties in the radial velocity. For example, suppose we measure the distance of a galaxy to be 100Mpc with an uncertainty of 10Mpc. Since r is multiplied by H_0 in the formula, this translates into an uncertainty of 700km/sec in the velocity. Given that velocities of galaxies are typically around a few hundred km/sec, this means that the uncertainty in velocity can be greater than the velocity itself! Thus galaxy velocity data is typically only useful when you have many measurements that you can combine to reduce the uncertainty.

Note that we can only measure the radial velocity of a galaxy, i.e. it's motion toward or away from us. For galaxies in different directions this will be a different component of the velocity vector $\vec{v}$, but since we have good evidence that the Universe is *isotropic* (the same in all directions), then we should be able to treat the radial component of velocity the same as if it were the x, y, or z component. In other words, we would expect the radial component of velocity to be drawn from a Gaussian distribution centered on zero with standard deviation σ_v.

On the course Canvas site you will find the files galvel.dat and grpvel.dat, which contain galaxy velocities and galaxy group velocities respectively. You will be using the 1st and 2nd columns of these files, which are measured radial velocity of the ith galaxy, v_i, and its measurement uncertainty σ_i, both of which are in units of km/sec. Galaxies with positive velocities are moving toward us and those with negative velocities are moving away from us. We can model each measured radial velocity as the sum of the actual velocity u_i of the galaxy plus some noise, $v_i = u_i + \epsilon_i$, where the noise ϵ_i is drawn from a Gaussian distributions with standard deviation σ_i and the actual velocity u_i is drawn from a Gaussian distribution with standard deviation σ_v. Thus the measured velocity v_i is drawn from a Gaussian distribution centered on zero with standard deviation given by $\sqrt{\sigma_v^2 + \sigma_i^2}$.

The problem of estimating σ_v is thus very similar to the problem of estimating the standard deviation of a set of numbers, with the differences that 1) we already know that the "true" average is zero, and 2) we need to account for the measurement uncertainty for each velocity. You should be able to convince yourself that the likelihood function for σ_v is

$$L(\sigma_v|v_i) = \prod_i \frac{A}{\sqrt{\sigma_v^2 + \sigma_i^2}} e^{-v_i^2/2(\sigma_v^2+\sigma_i^2)} \tag{2}$$

where we absorbed all the constants into A. Unfortunately, it's too messy to find the maximum likelihood by taking a derivative, so we will instead just plot the likelihood vs. σ_v and find the maximum likelihood value that way.

One complication for a large data set is that the product in the formula above can get so small that the computer will round it to zero. One solution is to calculate the log likelihood instead,

$$\ln L(\sigma_v|v_i) = C + \sum_i -\frac{1}{2}\ln(\sigma_v^2 + \sigma_i^2) - \frac{v_i^2}{2(\sigma_v^2 + \sigma_i^2)} \tag{3}$$

where C is a constant. First calculate ln$L$ with $C = 0$ and find where the maximum occurs. This is your maximum likelihood value for σ_v. To plot the likelihood, choose C to be the negative of the maximum value of the ln$L$, so that the maximum value of ln$L$ becomes zero. Then you can plot the Likelihood using this value of C and the likelihood at the peak will be 1. You can find the uncertainty in σ_v by fitting a Gaussian to the Likelihood peak and determining the width of the peak from the standard deviation of the Gaussian.

Now, in principle groups of galaxies should have a smaller σ_v than individual galaxies since their velocities are essentially averages of the velocities of their constituent galaxies. From your estimates of σ_v for the two datasets, calculate a confidence level that the two datasets have a different σ_v.

Finally, we can ask the question of how good our model is at describing the data. From the discussion above, the quantities $v_i/\sqrt{\sigma_v^2 + \sigma_i^2}$ should be drawn from a Gaussian distribution centered on zero with a standard deviation of 1. Using your maximum likelihood value of σ_v^2, make a histogram of these values for both datasets. Plot on top of your histogram a Gaussian centered on zero with standard deviation of 1.

Give your result for σ_v for each dataset together with its uncertainties. A good name for σ_v is the velocity dispersion, which here means the spread of velocity values. Discuss how confidently you can say that the σ_v of the two data sets is different. To calculate this, find the standard deviation of the difference of the two values and determine how many standard deviations your value is from zero. Discuss how well the Gaussian model describes the data, i.e. how well does your histogram match a Gaussian centered on zero with a standard deviation of 1. If it doesn't match, discuss possible reasons for the failure of the model. For example, galaxies falling into clusters of galaxies will have unexpectedly large velocities. Do you see evidence for these nonGaussian tails in your histogram?