Bayesian Statistics: Parameter Estimation 2, continuous distributions

In this class, we'll do another parameter estimation task. Last time we had a finite number of bins (two or six, in fact). In many circumstances we instead have to deal with a continuous variable, which therefore has a potentially unlimited number of bins. We can estimate parameters in the same way as before, but it's useful to give it a try because there are some apparent difficulties to overcome.

Suppose that we have measured radial velocities, and we fit a zero-centered Gaussian to the data. The single parameter of interest is the standard deviation. We have ten data points, which (after sorting) are velocities of -1.84623, -1.76493, -0.926109, -0.909967, 0.110899, 0.296846, 1.80029, 2.55558, 2.78944, and 3.57825, in units of km s^{-1}. How do we estimate the standard deviation? It might seem that you would *have* to bin the measurements, because otherwise it doesn't look like a distribution at all (if the bins are narrow, there are either zero points or one point in a bin, so there are no peaks). This is, however, not the case, so let's see how it works.

First, we realize that our Gaussian has the form

$$N(v)dv = A\exp(-v^2/2\sigma^2)dv \ , \tag{1}$$

where σ is the standard deviation and A is a normalization factor. Therefore, it might appear that there are two parameters. However, if we normalize the distribution so that the total number expected in the model equals the number of data points, then in our case $\int_{-\infty}^{\infty} N(v)\,dv = 10$. This implies that $A = 10/(\sqrt{2\pi\sigma^2})$.

Now we construct the log likelihood. As we showed earlier, when we normalize so that $\sum_i m_i$ is a constant, the only important term in the log likelihood is $\sum d_i \ln(m_i)$, where the sum is over all bins, d_i is the number of counts in bin i, and m_i is the predicted number of counts in bin i from the model. If we imagine dividing the data space into an enormous number of narrow bins, though, we realize that the ones without counts don't contribute, because $d_i = 0$. Therefore, the sum really only needs to go over the bins that contain counts. Next, what is m_i? It is the expected number of counts in a bin. Suppose that a bin has width dv at velocity v. Then the expected number is $N(v)dv$. This appears to depend crucially on the bin width, but remember that we're just comparing *differences* of log likelihoods. Therefore, if we use the same bin widths for every value of σ (which we obviously will), the $\ln dv$ values will be in common between all models, and hence will cancel. If we make the further assumption that we've done the smart thing and chosen small enough bins that the ones with data all have $d_i = 1$, then we get finally

$$\ln\mathcal{L} = \sum_i \ln[N(v_i)] + \text{const} \ , \tag{2}$$

where the v_i are the measured velocities. The depends only on the values of the distribution function at the measured velocities.

This is actually a general result for continuous distributions, in any number of dimensions. After you've normalized, the log likelihood is just the sum of the log of the distribution function at the measured locations if you have enough precision that there is at most one count per bin.

For a general model, one would now calculate the log likelihood numerically for a set of parameter values, then maximize to get the best fit. In our particular case, we can do it analytically. Dropping the constant,

$$\ln \mathcal{L} = \sum_i \left[\ln(10/\sqrt{2\pi}) - \ln \sigma - v_i^2/2\sigma^2 \right] \ . \tag{3}$$

This sum is over the ten measured velocities. We note that the first term is in common between all models, so we drop it. We then have

$$\ln \mathcal{L} = -10 \ln \sigma - (1/2\sigma^2) \sum_i v_i^2 \ . \tag{4}$$

The sum of the squares of our velocities is 38.7. Taking the derivative with respect to σ and setting to zero (to maximize) gives

$$\begin{aligned} -10/\sigma_{\rm best} + 38.7/\sigma_{\rm best}^3 &= 0 \\ \sigma_{\rm best} = (3.87)^{1/2} &= 1.967 \ . \end{aligned} \tag{5}$$

When we compute the 68.3% credible region, we need to (1) select a prior on the standard deviation, and (2) decide on how we want to define the credible region. When we think about a prior, there are apparently many choices. For example, unlike with our previous discrete case, where a was limited to being between 0 and 1, σ could in principle range from 0 to ∞. What should we choose? We can get an answer to that by looking at the likelihood:

$$\mathcal{L} \propto \sigma^{-10} e^{-38.7/(2\sigma^2)} \ . \tag{6}$$

We see that at very small σ the likelihood drops off sharply (because of the $e^{-38.7/(2\sigma^2)}$ factor) and similarly at very large σ (because of the σ^{-10} factor). Thus we actually can take σ to be equally probable in a large range, say 0 to 20, and then have it end abruptly above 20. There is virtually no likelihood above 20, so having it be constant from 0 to 50, or 0 to 1000, will lead to almost identical conclusions. This is an example in which the data are informative enough that reasonable priors will lead to the same conclusion. There are also times when we might not know the *scale* of σ, in which case perhaps a reasonable prior might be that there is equal probability in equal ranges of $\ln \sigma$ (so that, for example, the prior probability would be the same from 0.1 km s^{-1} to 1 km s^{-1} as it is from 1 km s^{-1} to 10 km s^{-1}). Then

the prior would be proportional to $1/\sigma$ from some minimum to maximum (can you see why?) and the posterior would thus be proportional to σ^{-11} instead of σ^{-10}; not a big difference.

So if we choose a flat prior $p(\sigma) = 1/\sigma_{\max}$ from $\sigma = 0$ to $\sigma = \sigma_{\max}$, with $\sigma_{\max} > 20$, and define the credible region as before (minimum contiguous region in σ that includes 68.3% of the probability), we find that the credible region runs from $\sigma = 1.56$ to $\sigma = 2.565$. In fact, $\sigma = 2.5$ was used to generate the data.

As with our discrete-data example, let's now think about how we would do this using χ^2. First of all, we wouldn't use χ^2; if we have only 10 total points, which we then have to bin, that's just ridiculous. But let's do it anyway to see how it would be done.

How should we bin the data? If, for example, we put all of the negative radial velocities in one bin, and all of the positive radial velocities in the other, then we have no discriminatory power at all! Why? Because the integral from $-\infty$ to 0 (or 0 to $+\infty$) of a normalized zero-centered Gaussian is the same regardless of the standard deviation, which is what we want to calculate. Thus the chi squared will be completely independent of the standard deviation. This is an extreme example of the loss of information due to binning! It also points out the arbitrary nature of binning; what should you group together?

But let's forge on anyway. Suppose that we look at the data before making our bins, and decide to make the bins $-\infty$ to -0.91 (in km s^{-1}), -0.91 to 1.0, and 1.0 to $+\infty$, so that there are respectively 3, 3, and 4 counts in our bins. Call those bins 1, 2, and 3. Then the expected number in each bin, for a standard deviation σ, are

$$m_1 = \int_{-\infty}^{-0.91} \frac{10}{\sqrt{2\pi\sigma^2}} e^{-v^2/2\sigma^2} dv \ , \tag{7}$$

$$m_2 = \int_{-0.91}^{1.0} \frac{10}{\sqrt{2\pi\sigma^2}} e^{-v^2/2\sigma^2} dv \ , \tag{8}$$

and

$$m_3 = \int_{1.0}^{\infty} \frac{10}{\sqrt{2\pi\sigma^2}} e^{-v^2/2\sigma^2} dv \ , \tag{9}$$

and $d_1 = 3$, $d_2 = 3$, and $d_3 = 4$. If we do this then we find that the data-variance χ^2 is minimized at $\sigma = 2.45$ and using $\Delta\chi^2 = 1$ gives a range from $\sigma = 1.603$ to $\sigma = 4.767$. Feel free to calculate the best value and the range for the correct model-variance χ^2.

This is obviously worse and less representative than what we got from the Bayesian analysis, but it's also clearly unfair to χ^2; 3, 3, and 4 counts in the bins is hardly in the $n \to \infty$ asymptotic limit! So now it's your turn, using the data file for the blackbody spectrum. Using the Bayesian approach with likelihoods and a flat prior on the temperature from 0 to 1 keV, determine the highest-probability temperature and the smallest contiguous 68.3% credible region. Then group them into 3 bins of 20 counts each (with the idea that

we need at least 20 counts to be in the Gaussian regime), and do the same analysis using χ^2. How much information do you lose due to the binning?

Now for some closing remarks regarding parameter estimation and Bayesian analysis more generally. First, there are times when you simply can't have infinitely fine data. For example, in X-ray astronomy, the spectral data you receive is in the form of counts in discrete energy channels. You can't break the data apart more finely than that, so you will have to deal with channels that have more than one count. In addition, if you are interested in the parameters of the source, you'll need to use a model of the detector (for X-ray observations, this comes in the form of a response matrix and effective energy curve) and propagate your intrinsic source model (in the form of photons per area per time at each photon energy) through the detector response to get a prediction in energy channel space. Always compare your model with the data in data space, which is called forward folding! Don't "backward fold", where you try to use the counts in the energy channels to infer the photon properties of the source. That leads to ambiguity and pain.

Second, so far we have focused for simplicity on cases in which there is only one model parameter. If there are several parameters, we can do similar things in the sense that we can still compute the likelihood of the data given the model for a particular combination of its parameters, and the prior should be specified for all the parameters (jointly in general, rather than as a product of individual priors). In such cases, it is common that one is only interested in the distribution of a subset of the parameters (say, only one of them!). In that case, one *marginalizes* the posterior probability distribution. That is, suppose now we have lots of parameters $a_1,\ a_2,\ \ldots, a_n$. Our posterior probability distribution is $P(a_1,\ a_2,\ \ldots, a_n)$, normalized so that $\int P(a_1,\ a_2,\ \ldots, a_n) da_1 da_2 \ldots da_n = 1$. If we only want to know the probability distribution for parameter a_1, independent of the values of the other parameters, we simply integrate over those other parameters:

$$p(a_1) = \int P(a_1,\ a_2,\ \ldots, a_n) da_2 \ldots da_n \ . \tag{10}$$

We then have $\int p(a_1) da_1 = 1$. Similarly, one could find the distribution for the two parameters a_1 and a_2 by integrating P over a_3 through a_n. The parameters you integrate over are called *nuisance* parameters.

Marginalization is *not* the same thing as (1) finding the parameter combination that maximizes the likelihood and then (2) finding the distribution of a single parameter by fixing all but that parameter at the maximum-likelihood values and then treating the problem like a single-parameter model! This is another approach that is sometimes taken by black-box analysis codes, and it will only give you a reasonable answer is special circumstances.

Finally, one cautionary point: because the *value* of the likelihood never enters, one can happily calculate maximum likelihoods and credible regions for models that are awful!

It's an automatic procedure. That's why Bayesians draw a distinction between parameter estimation and model comparison, which we will treat in the next class.