A Bayesian Interlude: Marginalization and Priors

Marginalization

Suppose that your model has multiple parameters, but you're really only interested in the posterior probability distribution of one of the parameters. For example, maybe you are doing a Gaussian fit to a line, and of the three parameters involved in the Gaussian (the centroid wavelength, the width, and the amplitude) you only care about the centroid wavelength. What should you do?

The way this task is performed in a number of analysis packages is that (1) you find the single best fit to the data, then (2) you freeze all of the values of the *other* parameters at their best-fit values, and finally (3) you vary the parameter of interest, with the other parameter values frozen, and use some criterion to figure out the uncertainty in the parameter of interest (e.g., by effectively doing a Bayesian fit with one parameter, or by doing a $\Delta\chi^2$).

But this procedure is incorrect; it yields the right value only in special circumstances. To understand why, let's think about what we mean by the posterior probability distribution for a single parameter when we have additional parameters.

First, if we go back to a single-parameter model (call that parameter a), suppose that we take the posterior probability distribution $P(a)$ and sample from it. That means that the probability that we choose a given a is proportional to $P(a)$. With lots of samples, we will simply recreate $P(a)$. That's simple enough to be tautological.

So what should we do when there are multiple parameters? The logical and correct generalization is that once we have the full multidimensional posterior $P(a, b, c, \ldots)$ such that the total probability contained in the volume spanned by a and $a + da$, b and $b + db$, etc., is $P(a, b, c, \ldots)da\,db\,dc\,d\ldots$, we imagine picking parameter combinations $(a, b, c, \ldots)$ with probability proportional to $P(a, b, c, \ldots)$ and then just storing the value of a. That makes sense: pick according to the (now multidimensional) posterior probability distribution, and determine the distribution of the values of a that you get as a result.

But it might not be clear why this definition is different from the oft-used definition in the second paragraph above.

To make the difference clear, let's think about an extreme case with two parameters. We'll set this up so that the overall peak in the two-dimensional posterior has only a narrow range in the y parameter with a large amplitude, but at other values of x (the parameter we care about), there is a broad range of values of y that give a significant posterior probability density $P(x, y)$. The posterior, which we assume to apply in the range $x = 0.05$ to ∞ and $y = -\infty$ to $+\infty$, is

$$P(x, y) = 0.23008e^{-(x-1.5)^2/2}e^{-x^2y^2} \ . \tag{1}$$

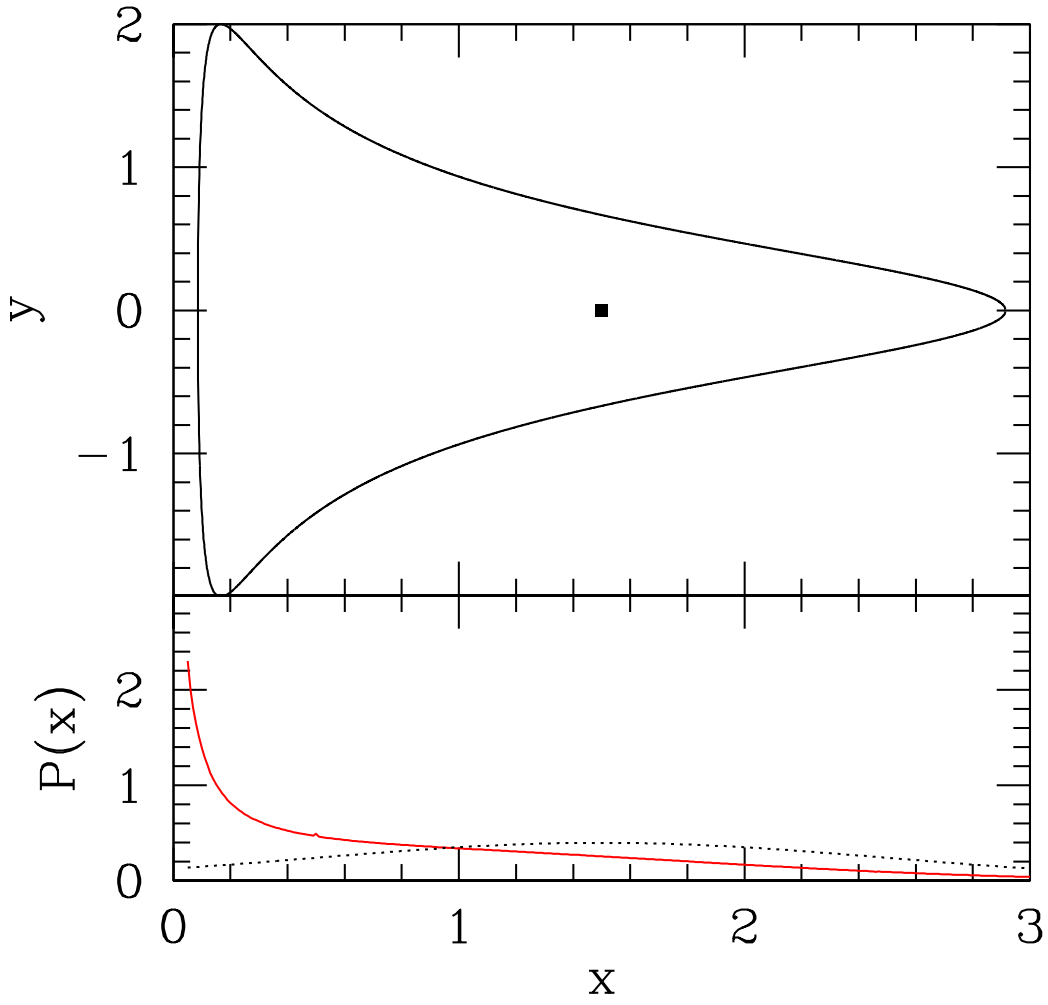

Fig. 1.— Demonstration that in general, you cannot find the one-dimensional posterior probability distribution from a multiparameter fit by fixing the other parameters to their best values. The posterior here is $P(x,y) = 0.23008e^{-(x-1.5)^2/2}e^{-x^2y^2}$, where the prefactor is chosen so that $\int_{0.05}^{\infty} P(x,y)dxdy = 1$. The solid square in the upper panel shows the location of the maximum of $P(x,y)$, where $x = 1.5$ and $y = 0$. That panel also shows the contour lines where $P(x,y)$ is e^{-1} times its maximum value; the asymmetry is obvious. The lower panel shows the one-dimensional posterior for x computed in two different ways. The black dotted line shows the result from the standard procedure of fixing the other parameters to their best values (here, $y = 0$) and then determining the probability distribution for x. The solid red line shows the correct answer we get by marginalization, i.e., by integrating $P(x,y)$ over all y at a given x. In this case, and in general, the procedure of fixing the other parameters to their best values gives an incorrect answer.

where the weird factor of 0.23008 is so that $\int_{-\infty}^{\infty}\int_{0.05}^{\infty} P(x,y)dxdy = 1$. The overall maximum of $P(x,y)$ is at $x = 1.5$, $y = 0$; that's the only place where the exponent is nonnegative. Thus if we used the common approach described above, we would expect that the one-dimensional posterior for x, $P(x)$, should peak at $x = 1.5$. Indeed, if we go to the overall maximum in the two-dimensional posterior ($x = 1.5$, $y = 0$) and fix y at its best value ($y = 0$), then what we have is just a standard Gaussian with variance 1, centered on $x = 1.5$.

But as you see in the upper panel of Figure 1, the contour lines are asymmetric; here we show the locus of points in the (x, y) plane that give $P(x, y) = 0.23008e^{-1}$. For $x < 1.5$, there is a much wider range in y that gives $P(x, y) > 0.23008e^{-1}$ than there is for $x > 1.5$.

You may say that this is an extreme example. And it is: I selected the function to make a point. But if you think about it, it is virtually *never* the case that a posterior has the property that fixing all the parameters but one and looking at that single-parameter cut through the posterior gives you exactly the same answer as the correct marginalization procedure. Multidimensional Gaussians, oriented in any direction, do have that property, but posteriors will never have exactly that form. So if your black box code fixes parameters to get one-dimensional posteriors, beware!

To be a bit more formal, suppose that we have parameters $a_1,\ a_2,\ \ldots, a_n$ in our model. Our posterior probability distribution is $P(a_1,\ a_2,\ \ldots, a_n)$, normalized so that $\int P(a_1,\ a_2,\ \ldots, a_n)da_1da_2\ldots da_n = 1$. 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 (this integration is called *marginalization*):

$$p(a_1) = \int P(a_1,\ a_2,\ \ldots, a_n)da_2\ldots da_n\ . \tag{2}$$

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.

Priors

I must say that priors were a sticking point for me when I first encountered Bayesian statistics. The normal statistics you use give the impression that they'll just tell you the answer, with no subjective priors. So why do we need them and, if we do need them, how should we pick our priors?

To address that let's begin with a hypothetical example. Suppose that you have a coin. The coin appears and feels completely ordinary and we assume that you acquired it in an ordinary way, e.g., maybe you got it in change from a store. You idly flip the coin ten times, and get eight heads and two tails. Your friend watches this and then, being a betting person,

puts down \$1 to bet that in the next 100 flips there will be at least 50 tails. You don't have anything else to do, and \$1 isn't too bad, so you agree to the bet and put down \$1 of your own. But your friend objects. Because the first ten flips got eight heads, it is clear that this coin is biased toward heads. Indeed, when the two of you do a careful calculation, you find that based on the data the probability that flips with this coin give tails at least half the time is only 0.0238. Thus your friend demands that you put down $(1-0.0238)/0.0238 \approx \$41$ to make the bet fair.

Do you take the bet?

For a more astronomical example, we can think about measurements of the Hubble constant H_0. This seems like a relatively straightforward calculation. The Hubble constant can be pictured as an average of the ratio between the apparent recession speed of a galaxy (or other cosmological object) and the distance to that object. So all we need to do is to measure the recession speeds to our objects, and the distances to our objects, and take ratios and some sort of suitable average and we're set. Right?

No, it's more complicated than that. To get a specific understanding of why, we can think about the latest method to enter the discussion, which involves gravitational waves.

On August 17, 2017, the gravitational wave detectors LIGO and Virgo detected the gravitational waves from the coalescence of two neutron stars; this event was called GW170817. A bit less than two seconds later, gamma rays from the event were detected using the *Fermi* telescope. This touched off an amazingly intense multimessenger observing campaign that had lots of implications, including for H_0.

The idea is that gravitational waves from a coalescing binary that is close enough that the redshift is nearly zero (GW170817 qualifies in this respect) serve as a self-calibrating source. What we mean by that is that from the frequency and frequency derivative of the gravitational waves, we know how luminous the source should be in gravitational waves, and thus by measuring the flux of energy in gravitational waves we can figure out the distance. Then, if the host galaxy can be identified (as happened for GW170817), its redshift and thus its recession speed can be determined and thus H_0 can be computed based on this event. That was done. But there are some hidden prior choices we need to make:

1. To get H_0 we want to think about the "Hubble flow", which is the net, average, motion of spacetime. Individual galaxies can move relative to the Hubble flow; for example, we appear to be moving at about 370 km s^{-1} relative to the flow (which we say because that's our speed relative to a frame in which the cosmic microwave background would appear isotropic). This movement can be in any direction, potentially, and it is usually not possible to determine the direction or speed for any individual galaxy (it wasn't possible for the host galaxy of GW170817, for example). Thus we need to make

assumptions about the probability distribution of the speed of the host galaxy relative to the Hubble flow, and fold those into our calculations. Such assumptions are made based on our prior observations of galaxies. We really can't "just let the data speak", because if we did this single analysis with no prior information, then we would have no idea how fast the galaxy could be moving relative to the Hubble flow. 10,000 km s^{-1} toward us? Away from us? Sideways? Our measurement would be useless.

2. Even the distance requires priors. The emission of gravitational waves from a binary is not isotropic. For example, the flux is four times as large along the binary axis as it is edge-on to the binary. The gravitational wave measurements themselves constrain the angle only very poorly. So we need to assume something about the probability distribution of the angle the binary axis makes with our line of sight. Should it be uniform in that angle, from (say) 0 to 90 degrees? That wouldn't be a good choice, because solid geometry tells you that from (for example) 0 to 1 degrees from an axis there is a much smaller solid angle than from 89 to 90 degrees, so your uniform-angle approximation could be far off. So maybe we should assume something that is uniform in the solid angle; as it turns out, uniformity in the cosine of the angle from the axis does that. Is that a good choice? Not necessarily, because given that binary gravitational wave events are brighter along the axis than edge-on, even if binaries are *intrinsically* isotropically oriented relative to us, we have an extra chance to see them when they're bright, i.e., close to face-on. In addition, models of gamma-ray bursts suggest that we see bright bursts when we are close to face-on; but this particular event had a weak gamma-ray burst, which might mean that it wasn't very close to face-on. And so on.
3. There are additional, hidden priors. There have been various papers projecting how well we can measure H_0 from future gravitational wave events, with optimistic answers. It has been pointed out that those analyses typically assume that all *other* aspects of cosmology are extremely well known a priori, such as the energy density of dark energy. If these assumptions are not made, then the precision of H_0 measurements can be degraded significantly.

It is important to emphasize that the probability distribution for H_0 that is derived from the gravitational wave data *depends* on the choices for the priors. If the new data had fantastically high signal to noise it would be different, because the likelihood would overwhelm the prior, but in this case, as in many cases, the data aren't overwhelming.

Most astronomical problems are like this to some degree. Even if you think you are just analyzing data, you have to make some prior assumptions to make progress.

So this is why you need to make prior assumptions, and why it is a *good* thing that in Bayesian statistics you are required to specify your priors explicitly. If you don't, then people can't reproduce your work.

But you might now be settling into despair. How are you supposed to *choose* your priors? After all, if you choose your priors narrowly enough, you'll get any answer you like. For example, in the coin-flipping example that started us off, if you have an absolutely unshakable faith that the probability of tails is 0.5 to an arbitrary number of significant figures, no number of flips would suffice to change your mind. Someone could flip a coin 1000 times, getting all heads, and it wouldn't matter to you. Come to think of it, this is actually how many people approach issues of religion or politics...

Despair, however, is not necessary. I would divide the choice of priors into two broad categories:

- Cases in which you have specific prior information. For example, other observations give us a reasonable idea for the distribution of galaxy speeds relative to the Hubble flow, so we can incorporate that into our new analysis. Somewhat trickier is the incorporation of prior measurements of the quantity you care about, e.g., H_0. For example, in the case of the gravitational wave measurement of H_0, we can ask whether we should present the results from that analysis only (with assumptions as described above) or present the results of that analysis with other H_0 estimates incorporated.
- Cases in which we don't have a lot of prior information. For this case, it is appropriate to use *uninformative* priors, which we'll now discuss.

Clearly, unless we really do have a lot of prior information, we shouldn't use priors so narrow that they give us the final answer. Thus we'd like to use a broad prior, so that if we end up with a narrow posterior it's because the data demand it, rather than because we used a narrow prior.

What the right broad prior is can depend a bit on the problem. If you're interested in the fraction of times a weirdly-shaped object lands on "H" rather than "T", then maybe a uniform prior from 0 to 1 probability would be appropriate. If you don't know the *scale* of the problem, then a logarithmic prior could be appropriate. For example, if we are interested in the distance to an object and have no idea how far it is, you might think that you should have a uniform prior over the distance, e.g., from 0 distance to 10 billion light years. But in doing that you would be unintentionally setting the prior so that large distances are preferred; after all, with that prior, there is a 99.99% prior probability that the distance is more than a million light years. To be scale-independent, you would want a logarithmic prior, which would mean that there is an equal prior probability that the distance is between 1 and 10, or 10 and 100, or 100 and 1000, etc., light years.

One can say with some justification that if you try several reasonable priors and these give you wildly different answers, your data didn't contain enough information to judge between them, so you can't say much. It is appropriate to try to select priors that are as

uninformative as possible so that the data speak for themselves. But you need to specify your prior and indicate what it is, so that others know how you performed your analysis!