Bayesian Statistics: Model Comparison

Suppose we have a data set, and two models to compare. How do we determine which model is favored by the data? At first glance this may seem easy: just figure out which model matches the data better. But think about models with different numbers of parameters; intuitively, we should give the benefit of the doubt to the model with fewer parameters, based on Ockham's principle. In addition, one could imagine a situation in which the parameters of two models are qualitatively different. Some of the parameters could be continuous (e.g., temperature), and some could be discrete (e.g., the quantum spin of a particle). For example, suppose we have two models. The first one has one parameter, which can take on any real number between 0 and 1. The second has three parameters, but each of them can only take on the value 0 *or* the value 1, but nothing in between. Which model is simpler, and how would we take that into account?

This, in my opinion, is where Bayesian statistics shines. It provides a simple procedure that *automatically* takes into account different numbers of parameters in an intuitively satisfying way. As before we'll give the general principles, then try some examples.

Say we have two models, 1 and 2. Model 1 has parameters $a_1,\ a_2, \ldots,\ a_n$, and a normalized prior probability distribution $p_1(a_1,\ a_2, \ldots,\ a_n)$. Model 2 has parameters $b_1,\ b_2, \ldots,\ b_m$ and a normalized prior probability distribution $p_2(b_1,\ b_2, \ldots,\ b_m)$. For a given set of values $a_1,\ a_2, \ldots,\ a_n$, let the likelihood of the data given the model (defined in previous lectures) for model 1 be $\mathcal{L}_1(a_1,\ a_2, \ldots,\ a_n)$, and similarly for model 2. Then the "Bayes factor" for model 1 in favor of model 2 is

$$\mathcal{B}_{12} = \frac{\int \mathcal{L}_1(a_1,\ a_2, \ldots,\ a_n) p_1(a_1,\ a_2, \ldots,\ a_n) da_1\, da_2 \ldots da_n}{\int \mathcal{L}_2(b_1,\ b_2, \ldots,\ b_m) p_2(b_1,\ b_2, \ldots,\ b_m) db_1\, db_2 \ldots db_m} \tag{1}$$

where the integration in each case is over the entire model parameter space. Therefore, it's just a ratio of the integrals of the likelihoods times the priors for each model. When you multiply the Bayes factor by the prior probability ratio you had for model 1 in favor of model 2 (you could set that ratio to unity if you had no reason to prefer one model over another), you get the odds ratio of model 1 in favor of model 2. Each integral (e.g., the numerator of this equation, or the denominator of this equation) is sometimes called the "evidence" for the model in question, so the Bayes factor is the ratio of the evidences.

What does this mean? Don't tell a real Bayesian I explained it this way, but consider the following. Suppose you and a friend place a series of bets. In each bet, one has two possible models. You compute the odds ratio as above, and get $\mathcal{O}_{12}$ in each case. Ultimately, it will be determined (by future data, say) which of the two models is correct (we're assuming these are the only two possible models). If your friend puts down \$1 on model 2 in each case, how much money should you place on model 1 in each bet so that you expect to break even

after many bets? You put down $\$\mathcal{O}_{12}$. That is, it really does act like an odds ratio. The reason a hard-core Bayesian might get agitated about this analogy is that Bayesian statistics emphasizes considering only the data you have before you, rather than imagining an infinite space of data (as happens in more familiar frequentist statistics). Still, I think this is a good description.

Why does this automatically take simplicity into account? Think of it like this. If your data are informative, then for a given set of data it is likely that only a small portion of the parameter space will give a reasonably large likelihood. For example, if you are modeling the interstellar medium in some region, you might have temperature and density as parameters; with good enough data, only temperatures and densities close to the right ones will produce a likelihood close to the maximum. Now, think about the priors. For a complicated model with many parameters, the probability density is "spread out" over the many dimensions of parameter space. Thus, the probability density is comparatively small in the region where the likelihood is significant. If instead you have few parameters, the prior probability density is less spread out, so it's larger where the likelihood is significant and therefore the integral is larger.

If the parameters have discrete instead of continuous values, you do a sum instead of an integral but otherwise it's the same. Note that (assuming that Poisson statistics apply) we have to use more of the full Poisson likelihood here. When we did parameter estimation we could cancel out lots of things, but here we have an integral or sum of likelihoods so we can't do the cancellation as easily. The product $\prod(1/d_i!)$ will be the same for every likelihood, and if your model is normalized so that the total number of expected counts is set to the number of observed counts (which as we've said before is common, but not universal) then $\prod \exp(-m_i)$ is the same for every likelihood. Thus those factors can be cancelled, but one still has a sum of likelihoods and so taking the log requires some interesting finesses that you might discover during your coding exercise.

Let's try an example. Consider a six-sided die. We want to know the probabilities of each of the six faces. Model 1 is that the probability is the same (1/6) for each face. Model 2 is that the probability is proportional to the number on the face. Normalized, this means a probability of 1/21 for 1; 2/21 for 2; and so on. We roll the die ten times and get 5, 2, 6, 2, 2, 3, 4, 3, 1, 4. What is the Bayes factor between the two models?

We're starting with an easy one, in which there are no parameters, so we don't even have to do an integral, just a likelihood ratio. For model 1 the normalized model expectations per bin are $m_1 = 10/6$, $m_2 = 10/6$, and so on. For model 2 we have $n_1 = 10/21$, $n_2 = 20/21$, $n_3 = 30/21$, and so on. Therefore,

$$\mathcal{L}_1 = \left(\frac{10}{6}\right)^1 \cdot \left(\frac{10}{6}\right)^3 \cdot \left(\frac{10}{6}\right)^2 \cdot \left(\frac{10}{6}\right)^2 \cdot \left(\frac{10}{6}\right)^1 \cdot \left(\frac{10}{6}\right)^1 = 165.4 \tag{2}$$

and

$$\mathcal{L}_2 = \left(\frac{10}{21}\right)^1 \cdot \left(\frac{20}{21}\right)^3 \cdot \left(\frac{30}{21}\right)^2 \cdot \left(\frac{40}{21}\right)^2 \cdot \left(\frac{50}{21}\right)^1 \cdot \left(\frac{60}{21}\right)^1 = 20.7 \; . \tag{3}$$

Thus, from this data,

$$\mathcal{B}_{12} = \mathcal{L}_1/\mathcal{L}_2 = 7.98 \; . \tag{4}$$

Model 1 is favored, assuming that we didn't have strong prior favoritism toward model 2.

Now try another example, with the same data. Model 1 is the same as before, but now model 2 has a parameter. In model 2, the probability of a 1 is $1 - p$, and the probability of a 2, 3, 4, 5, or 6 is $p/5$. Therefore, model 2 encompasses model 1, so by maximum likelihood alone it will do better. But will it do enough better to be favored? Let's assume as a prior that p is equally probable from 0 through 1. The numerator is the same as before, but for the denominator we need to do an integral. For probability p and our given data, the Poisson likelihood of the data given the model is

$$\mathcal{L}_2(p) = [10(1-p)] \cdot (2p)^3 \cdot (2p)^2 \ldots = 10(1-p)(2p)^9 \; . \tag{5}$$

Therefore the denominator is

$$\int_0^1 5120(1-p)p^9 \, dp = 46.5 \tag{6}$$

and the Bayes factor is

$$\mathcal{B}_{12} = 165.4/46.5 = 3.55 \; , \tag{7}$$

so the first model is still preferred. Note that the maximum likelihood for model 2 occurs for $p = 0.9$ (it should! We have one 1 in ten rolls) and gives 198.4, so as expected the more complicated model has a higher *maximum* likelihood; it's just not enough to make up for the extra complication.

Now it's your turn. Apply this last pair of models to the data sets on the web (which are the same as they were for Class 3, but are reprinted for convenience). For each data set, what are the Bayes factors between the two models? Note that in the coding section for this class we discuss how to do model comparison using χ^2 in the special case that one model is nested inside the other (which is true for the example we're exploring). What conclusions do you draw?

Model comparison in Bayesian statistics is always between precisely defined models. There is no analogue to the idea of a null hypothesis. Hard-core Bayesians consider this to be a strength of the approach. For example, suppose that you try to define a null hypothesis and do a standard frequentist analysis, finding that the null hypothesis can be rejected at the 99% confidence level. Should you, in fact, reject the null hypothesis? Not necessarily, according to Bayesians. Unless you know the full space of possible hypotheses, it could be that there are

10,000 competing hypotheses and of those your null hypothesis did the best. For example, suppose I think that gamma-ray bursts should come from isotropically distributed positions in the sky; that's my null hypothesis. A hundred positions are measured, and they are all found to cluster within 1° of each other. Surely I can reject my null hypothesis? Well, if I compare it with another hypothesis that says that all bursts should come from within 1" of each other, my null hypothesis does much better!

I'm not fully satisfied with this line of argument. To me, the testing of a null hypothesis as it's done in frequentist statistics is important because it gives you a way to tell if your model is reasonably close or not. That is, a standard chi squared per degree of freedom can give you an idea of whether you need to work a lot harder to get a good model, or if you're nearly there. In my opinion, it's important to have that kind of information, but there is reasoned disagreement on this issue.

Where does all of this leave us? I feel that when evaluating a model or derivation or whatever you should use quick, easy methods first (e.g., order of magnitude estimation) before settling in for more detailed treatments. The same goes for statistics. I recommend doing a quick analysis (chi squared, Kolmogorov-Smirnov test, or whatever) first, to see if your data are informative. If they are, then you may be justified in spending time with a more rigorous method to get the most out of your data. In all cases, however, you have to know the limitations of your method! If your model is terrible, getting detailed confidence or credible regions around your best fit isn't meaningful. If you have 40 degrees of freedom (defined as the number of data points minus the number of parameters in your model) and your total chi squared is 4000, you can't say much. On the other hand, if you get a reduced chi squared much *less* than one, doing delta chi squareds is also not really meaningful; actually, what it means is that you overestimated your error bars. So, the lesson as always is that you need to understand your method. That's why I've found it helpful to think in the Bayesian way. In many circumstances it means I can figure out what *should* be done, then I have a better sense of how good an approximation a simpler method is.

Now that we have our all-too-short introduction to Bayesian methods, we will spend the rest of the course discussing specific tasks, which we will perform on real astronomical data sets.