BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2017

5 Model comparison

We have spent a lot of time in the past couple of weeks looking at the problem of parameter estimation. Really, we have been stepping through the process of bringing our thinking about a biological system into a concrete statistical model that defines a likelihood for the data and the parametrization thereof. Writing down Bayes's theorem then gives the posterior,

$$g(\theta \mid D) = \frac{f(D \mid \theta)\, g(\theta)}{P(D)}, \tag{5.1}$$

where θ is the set of parameters. Solving the parameter estimation problem involves computing the posterior, which usually involves summarizing the posterior into a form that can be processed intuitively.

5.1 Adding models to the probabilities

When we write Bayes's theorem for the parameter estimation problem, implicit in the definition of the likelihood is the fact that we are using a specific statistical model. To be complete, especially in the context of model comparison, we should include which model we're using in the conditions of the probabilities. Let M_i denote a model i, and θ_i be the set of parameters associated with M_i.[10] Then, we have

$$g(\theta_i \mid D, M_i) = \frac{f(D \mid \theta_i, M_i)\, g(\theta_i \mid M_i)}{f(D \mid M_i)}. \tag{5.2}$$

This is a more explicit description of the probabilities associated with the parameter estimation problem.

5.2 Probabilities of models

Remember that Bayesian probability is a measure of the plausibility of any logical conjecture. So, we can talk about the probability of models being true. So, what is the probability that a model is true, given the observed data? Again, this is given by Bayes's theorem.

$$g(M_i \mid D) = \frac{f(D \mid M_i)\, g(M_i)}{f(D)}. \tag{5.3}$$

This is Bayes's theorem stated for the model comparison problem. Let's look at each term in turn.

[10] Do not be confused by the subscript here. The i does not signify the ith parameter of a set of parameters for a given mode. Here, it means that θ_i describes the set of parameters for model i.

- $g(M_i \mid D)$, as we said before, is the probability that model M_i is true given the measured data.
- $f(D)$ is a normalization constant for the posterior that is computed by marginalizing over all possible models

$$\sum_i g(M_i \mid D) = 1 \;\Rightarrow\; f(D) = \sum_i f(D \mid M_i)\, g(M_i). \tag{5.4}$$

- $g(M_i)$ is a measure of how plausible we thought model M_i is a priori, the prior probability for model M_i. For example, if a proposed model violates a physical conservation law, we know it is unlikely to be true even before we see the data. In practice, we typically assign equal probability to all models we have not ruled out prior to seeing the data.
- $f(D \mid M_i)$ is the likelihood of observing the data, given that model M_i is true.

As usual, we need to specify the likelihood and prior to assess the posterior probability of any given model. We already discussed how to specify the prior. We usually assume all models are equally likely. How about the likelihood? Well, glancing at equation (5.2), we see that the likelihood for the model comparison problem is the evidence for the parameter estimation problem! Because the posterior in the parameter estimation problem, $g(\theta_i \mid D, M_i)$, must be normalized, the evidence in the parameter estimation problem, and therefore also the likelihood in the model comparison problem, is given by

$$f(D \mid M_i) = \int \mathrm{d}\theta_i\, f(D \mid \theta_i, M_i)\, g(\theta_i \mid M_i). \tag{5.5}$$

So, if we can compute the likelihood and priors from the parameter estimation problem and can integrate their product, we have the likelihood for the model comparison problem.

5.3 Bayes factors and odds ratios

Computing the absolute probability of a model is difficult, since it would require considering all possible models, as is required to compute the normalization constant, $f(D)$. We therefore typically make pairwise comparisons between models. This comparison is called an **odds ratio**. It is the ratio of the probabilities of two models being true.

$$O_{ij} = \frac{g(M_i)}{g(M_j)} \left[\frac{f(D \mid M_i)}{f(D \mid M_j)}\right]. \tag{5.6}$$

The first factor in the product is the ratio of our prior knowledge of the truth of the models. If they are equally likely, this ratio is unity. The bracketed ratio is called the **Bayes factor**, which is the ratio of the evidences of the respective models.

Note that if we compute all of the odds ratios comparing a given model k to all others (and somehow did manage to consider all models that have nonzero probability), we can compute the posterior probability of model M_i as

$$g(M_i \mid D) = \frac{O_{ik}}{\sum_j O_{jk}}. \tag{5.7}$$

5.4 Approximate computation of the Bayes factor

Evaluating the integral in equation (5.5) to compute the Bayes factor is in general difficult. If the posterior is sharply peaked, we may compute this integral using the **Laplace approximation** in which we approximate the integral by the height of the peak times its width. In one dimension, this is

$$\begin{aligned} f(D \mid M_i) &= \int \mathrm{d}\theta_i f(D \mid \theta_i, M_i)\, g(\theta_i \mid M_i) \\ &\approx f(D \mid \theta_i^*, M_i)\, g(\theta_i^* \mid M_i) \sqrt{2\pi\,\sigma_i^2}, \end{aligned} \tag{5.8}$$

where θ_i^* is the MAP estimate, and σ_i^2 is the variance of the Gaussian approximation of the posterior. In n-dimensions, this is

$$g(D \mid M_i) = \int \mathrm{d}\theta_i f(D \mid \theta_i, M_i)\, g(\theta_i \mid M_i) \tag{5.9}$$

$$\approx f(D \mid \theta_i^*, M_i)\, g(\theta_i^* \mid M_i)\ (2\pi)^{|\theta_i|/2} \sqrt{\det \Sigma_i}, \tag{5.10}$$

where Σ_i is now the covariance matrix of the Gaussian approximation of the posterior under M_i. We have also denoted the number of parameters in M_i to be $|\theta_i|$. Note that we have already computed all of factors in the above product in the parameter estimation problem if we solved it by optimization. Therefore, we already have what we need to compute the (approximate) odds ratio.

5.5 The factors in the odds ratio

We can now write the approximate odds ratio as the product of three factors.

$$O_{ij} \approx \left(\frac{g(M_i)}{g(M_j)}\right) \left(\frac{f(D \mid \theta_i^*, M_i)}{f(D \mid \theta_j^*, M_j)}\right) \left(\frac{g(\theta_i^* \mid M_i)\ (2\pi)^{|\theta_i|/2} \sqrt{\det \Sigma_i}}{g(\theta_j^* \mid M_j)\ (2\pi)^{|\theta_j|/2} \sqrt{\det \Sigma_j}}\right). \tag{5.11}$$

- The first term represents the prior probability of the models. This is how plausible we thought the models were before the experiment.

- The second term is a measure of the goodness of fit. In other words, it comments on how probable the data are given the model and the MAP estimate.
- The third term is a ratio of **Occam factors**. An Occam factor is the ratio of the volume of parameter space accessible to the posterior to that of the prior. This is best seen by example. Consider a model M_1 with a single parameter where the parameter a that has a Uniform prior. Then,

$$\text{Occam factor} = \sqrt{2\pi} g(a^* \mid M_1)\sigma_1 = \frac{\sqrt{2\pi}\ \sigma_1}{a_{\max} - a_{\min}}. \tag{5.12}$$

  Remember, σ_1^2 is the variance of the Gaussian approximation of the posterior. So, the numerator here is the width of the posterior and the denominator is the width of the prior.

  Now, consider a model, M_2 with two parameters, b and c, each with Uniform priors. In this case, we have

$$g(b^*, c^* \mid M_j) = \frac{1}{b_{\max} - b_{\min}} \frac{1}{c_{\max} - c_{\min}}, \tag{5.13}$$

  and the Occam factor is

$$\text{Occam factor} = \frac{2\pi\sqrt{\det \Sigma_2}}{(b_{\max} - b_{\min})(c_{\max} - c_{\min})}. \tag{5.14}$$

  So, the volume of the parameter space accessible to the prior for model M_2 is larger than for M_1, so the part of the odds ratio is greater than one, favoring the model with fewer parameters. The ratio of Occam factors is then

$$\frac{\sigma_i}{\sqrt{2\pi\ \det \sigma_j^2}} (b_{\max} - b_{\min}). \tag{5.15}$$

  Comparing the Occam factors of the two models, we see that the more parameters you have, the bigger the denominator of the Occam factor is, making the Occam factor smaller. Furthermore, it is also often the case that complicated models with lots of parameters also have smaller determinants of the covariance because the multitude of parameters are "locked in" around the MAP estimate. Thus, we see where the Occam factor gets its name, since it penalizes more complicated models.[11]

This approximate calculation shows us everything that goes into the odds ratio. Any one factor can overwhelm the others:

- What we knew before
- How well the model fits the data
- How simple the model is

[11] Remember that Occam's razor states that among competing hypotheses, the one with fewest assumptions is preferred.

5.6 Example: Are two data sets from the same Gaussian distribution?

We will now look at an example. Say I do two sets of measurements of property x, a control and an experiment. We make n_c control measurements and n_e experiment measurements. We consider two models. M_1 says that both the control and the experiment are chosen from the same underlying Gaussian distribution with mean μ and variance σ. Model M_2 says that control and experiment come from different Gaussian distributions with means μ_c and μ_e. We wish to compare models M_1 and M_2. The odds ratio is

$$O_{12} = \frac{g(M_1)\, f(D_c, D_e \mid M_1)}{g(M_2)\, f(D_c, D_e \mid M_2)}, \tag{5.16}$$

where D_c denotes the data from the control experiment and D_e denotes the data from the experiment.

We will assume a prior that $g(M_i) = g(M_j)$. Then, we are left to compute $f(D_c, D_e \mid M_1)$ and $f(D_c, D_e \mid M_2)$. We can do this by approximate integration (see section 4.3.1 of Sivia). Note that we assume a uniform prior on σ, with $0 < \sigma < \sigma_{\max}$. We could also try the problem with a Jeffreys prior on σ, but I do not feel like doing the nasty integration. The result for the odds ratio is

$$O_{12} \approx \frac{\sigma_{\max}\,(\mu_{\max} - \mu_{\min})}{\pi\sqrt{2}}\; \frac{n_1\, n_2\, s^{2-n1-n2}}{(n_1 + n_2)\, s_1^{2-n_1}\, s_2^{2-n_2}}, \tag{5.17}$$

where

$$s^2 = \frac{1}{n_1 + n_2} \sum_{i \in D_1 \cup D_2} (x_i - \bar{x})^2, \tag{5.18}$$

$$s_1^2 = \frac{1}{n_1} \sum_{i \in D_1} (x_i - \bar{x}_1)^2, \tag{5.19}$$

$$s_2^2 = \frac{1}{n_2} \sum_{i \in D_2} (x_i - \bar{x}_2)^2, \tag{5.20}$$

with

$$\bar{x} = \frac{1}{n_1 + n_2} \sum_{i \in D_1 \cup D_2} x_i, \tag{5.21}$$

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i \in D_1} x_i, \tag{5.22}$$

$$\bar{x}_2 = \frac{1}{n_2} \sum_{i \in D_2} x_i. \tag{5.23}$$

It seems that this question is often asked: does the experiment come from a different process than the control? My opinion is that in most situations, the answer is an obvious yes, and the more pertinent question is by how much they differ. Nonetheless, if we are asking the "if they are different" question, we can plug our data in and easily compute the odds ratio. Be careful, though. This, too, should not be a yes-or-no question. We should not really be asking *if* they come from different distributions, what are the odds that they do.

5.7 Caveats and motivation for information criteria

Gelman, et al., in their book *Bayesian Data Analysis* (3rd Ed, page 182), express some concern about the approach we have taken here. I quote them, emphasis theirs, bracketed comment mine.

> This fully Bayesian approach has some appeal but we generally do *not* recommend it because, in practice, the marginal likelihood [which we have been calling the evidence from the parameter estimation prior] is highly sensitive to aspects of the model that are typically assigned arbitrarily and are untestable from data.

These arbitrary and untestable aspects are typically the priors. We try to be uninformative, but they *must* be proper (meaning normalized) in order to do model comparison as we have done here. The Bayes factor is highly sensitive to the width of the priors.

In my opinion, this method of model comparison is often perfectly legitimate because the prior is part of the model and, while untestable from data, is not arbitrary. If constructed properly, the prior represents our knowledge before data acquisition and should therefore naturally be included in model comparison.

Whatever your position on this matter, it is still useful to have other metrics for assessing models.

5.8 Watanabe-Akaike Information Criterion (WAIC)

A good model is a predictive model. If we were to acquire more data under identical conditions, the parameters we derived from the posterior should be able to accurately predict what those new data would look like. It makes sense to assess a model on how well it can predict new data. Furthermore, the connection between predictive capabilities and the Bayes factor is clear if you think about what must be true of a predictive model. First, it must describe the data we have actually acquired well. The goodness-of-fit term in the Bayes factor covers this. Second, it must describe

new data well. If the parameters are such that it describes the data already collection very well, but cannot predict, the model is not good. This usually happens when the model has many parameters tailor-made for the data (such as fitting data with a higher-order polynomial). This is captured in the Occam factor. So, models with good Bayes factors are often predictive.

With that in mind, I will introduce a good metric for comparing models, the **Watanabe-Akaike Information Criterion**, also known as the **Widely Applicable Information Criterion** (WAIC). I will discuss it intuitively and not provide much rigor. For detailed descriptions of what follows, I recommend reading chapter 7 of Gelman, et al., *Bayesian Data Analysis, 3rd Ed.* and chapter 6 of McElreath, *Statistical Rethinking*.

In what follows, for notational convenience, I will drop explicit dependence of M_i, and also drop the subscripts from the parameter set θ_i. We define the predictive density of a single data point $x \in D$ as

$$\text{single point predictive density} = \int \mathrm{d}\theta \; f(x \mid \theta)\, g(\theta \mid D). \tag{5.24}$$

This is the likelihood for observing data point x, averaged over the posterior probability distribution of parameter values θ. We are therefore taking into account posterior information and using the likelihood to assess goodness-of-fit. We can take the product of each of the single point predictive densities in the data set and take the logarithm to get the **log pointwise predictive density**, or **lppd**,

$$\begin{aligned} \text{lppd} &= \ln \left(\prod_{x \in D} \int \mathrm{d}\theta \; f(x \mid \theta)\, g(\theta \mid D) \right) \\ &= \sum_{x \in D} \ln \left(\int \mathrm{d}\theta \; f(x \mid \theta)\, g(\theta \mid D) \right). \end{aligned} \tag{5.25}$$

This gives a metric of how well the model manages to predict the observed data. Put succinctly, the lppd is the sum of the logarithm of the average likelihood of each observation in a data set.

This metric is biased toward complicated models, so we add a correction. We compute the **effective number of parameters**, p_{WAIC} as

$$p_{\text{WAIC}} = \sum_{i \in D} \text{variance}(\ln f(x \mid D)), \tag{5.26}$$

where the variance is computed over the posterior. Written out, this is

$$\text{variance}(\ln f(x \mid D)) = \int \mathrm{d}\theta \; g(\theta \mid D)\, (\ln f(x \mid D))^2$$

$$- \left(\int \mathrm{d}\theta \, g(\theta \mid D) \, \ln f(x \mid D) \right)^2 . \tag{5.27}$$

This parameter p_{WAIC}, can be thought of as the number of unconstrained parameters in a model. Parameters that are influences only by the prior contribute little to p_{WAIC}, while those that are determined mostly by the data contribute more.

The WAIC is then

$$\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}}). \tag{5.28}$$

The factor of -2 is there for historical reasons to enable comparisons to the Akaike Information Criterion (AIC) and the Deviance Information Criterion (DIC). These two information criteria are also widely used, but have assumptions about Gaussianity, and in the case of the AIC, also flat priors. The WAIC is a better choice.

Computing the WAIC is difficult, unless, of course, you managed to get MCMC samples! Given a set of S MCMC samples of the parameters θ (where $\theta^{(s)}$ is the sth sample), the lppd may be calculated as

$$\text{lppd} = \sum_{x \in D} \ln \left(\frac{1}{S} \sum_{s=1}^{S} f(x \mid \theta^{(s)}) \right) . \tag{5.29}$$

Another beautiful example of how sampling converts integrals into sums. Similarly we can compute p_{WAIC} from samples.

$$p_{\text{WAIC}} = \sum_{x \in D} \frac{1}{S-1} \sum_{s=1}^{S} \left(\log f(x \mid \theta^{(s)}) - q(x) \right)^2 , \tag{5.30}$$

where

$$q(x) = \frac{1}{S} \sum_{s=1}^{S} \ln f(x \mid \theta^{(s)}). \tag{5.31}$$

While you can compute the WAIC from your MCMC samples, PyMC3 has a built-in function to do it.

For an intuitive description of the WAIC, you may think of it as an estimate of the negative log likelihood of new data.[12] That is, it is an estimate of how badly the model would perform with new data. So, the lower the WAIC, the better the model.

[12] Stated precisely, the WAIC is an estimate of the out-of-sample deviance. "Out-of-sample" just means data that is yet to come. I did not want to go through the trouble of defining deviance.

5.9 The Akaike weights

The value of a WAIC by itself does not tell us anything. Only comparison of two or more WAICs makes sense. Recalling that the WAIC is a measure of a log likelihood, if we exponentiate it, we get something proportional to a probability. If we have two models, M_i and M_j, the **Akaike weight** of model i is

$$w_i = \frac{\exp\left[-\frac{1}{2}\,\text{WAIC}_i\right]}{\exp\left[-\frac{1}{2}\,\text{WAIC}_i\right] + \exp\left[-\frac{1}{2}\,\text{WAIC}_j\right]}. \tag{5.32}$$

This weight may be interpreted as an estimate of the probability that M_i will make the best predictions of new data.[13] We can generalize this to multiple models.

$$w_i = \frac{\exp\left[-\frac{1}{2}\,\text{WAIC}_i\right]}{\sum_j \exp\left[-\frac{1}{2}\,\text{WAIC}_j\right]}. \tag{5.33}$$

We can compute a quantity analogous to the Bayesian odds ratio,

$$\frac{w_i}{w_j} = \exp\left[-\frac{1}{2}(\text{WAIC}_i - \text{WAIC}_j)\right]. \tag{5.34}$$

5.10 Computing odds ratios and information criteria

You may have noticed that computing the WAIC almost always required performing an MCMC calculation In the approximate calculation of the odds ratio, I only used MAP information that could be found by optimization. This, however, is approximate, and has all the perils associated with posteriors that are strongly non-Gaussian. There are information criteria that can be computed from MAP estimates as well. These also have dangers associated with them.

So, how do you compute the odds ratio (via Bayes factor) from MCMC? We can use a technique called parallel-tempering Markov chain Monte Carlo (PTMCMC) to exactly compute the odds ratio. As you likely have guessed, this is computationally intensive, but effective. We will learn about this in an auxiliary lesson.

[13] This interpretation is common, but not entirely agreed upon.