BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2016

4 Model selection

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,

$$P(\mathbf{a} \mid D, I) = \frac{P(D \mid \mathbf{a}, I)\, P(\mathbf{a} \mid I)}{P(D \mid I)}, \tag{4.1}$$

where $\mathbf{a}$ 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.

4.1 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. We really should be explicit and include which model we're using in our probabilities.[13] Let M_i be model i. Then, for parameter estimation, we have

$$P(\mathbf{a}_i \mid D, M_i, I) = \frac{P(D \mid \mathbf{a}_i, M_i, I)\, P(\mathbf{a}_i \mid M_i, I)}{P(D \mid M_i, I)}. \tag{4.2}$$

Notice that we have also assigned the subscript i to the set of parameters we are determining to specify that they are associated with model M_i. So this is a more explicit description of the probabilities associated with the parameter estimation problem.

4.2 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.

$$P(M_i \mid D, I) = \frac{P(D \mid M_i, I)\, P(M_i \mid I)}{P(D \mid I)}. \tag{4.3}$$

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

[13] We haven't been this explicit so we don't get over burdened with notation.

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

$$\sum_i P(M_i \mid D, I) = 1 \;\Rightarrow\; P(D \mid I) = \sum_i P(D \mid M_i, I)\, P(M_i \mid I). \tag{4.4}$$

- $P(M_i \mid I)$ 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. I.e., we have uninformative priors for the models.
- $P(D \mid M_i, I)$ 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 (4.2), we see that the likelihood for the model selection problem is the evidence for the parameter estimation problem! Because the posterior in the parameter estimation problem, $P(\mathbf{a}_i \mid D, M_i, I)$, must be normalized, the evidence in the parameter estimation problem, and therefore also the likelihood in the model selection problem, is given by

$$P(D \mid M_i, I) = \int \mathrm{d}\mathbf{a}_i\, P(D \mid \mathbf{a}_i, M_i, I)\, P(\mathbf{a}_i \mid M_i, I). \tag{4.5}$$

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 selection problem.

4.3 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, $P(D \mid I)$. We typically therefore make pairwise comparisons between models. This comparison is called an **odds ratio**. It is the ratio of the probabilities of two models being true.

$$O_{ij} = \frac{P(M_i \mid I)}{P(M_j \mid I)} \left[\frac{P(D \mid M_i, I)}{P(D \mid M_j, I)}\right]. \tag{4.6}$$

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

$$P(M_i \mid D, I) = \frac{O_{ik}}{\sum_j O_{jk}}. \tag{4.7}$$

4.4 Approximate computation of the Bayes factor

Evaluating the integral in equation (4.5) 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

$$\begin{aligned} P(D \mid M_i, I) &= \int \mathrm{d}a_i \, P(D \mid a_i, M_i, I) \, P(a_i \mid M_i, I) \\ &\approx P(D \mid a_i^*, M_i, I) \, P(a_i^* \mid M_i, I) \sqrt{2\pi\sigma_i^2}, \end{aligned} \tag{4.8}$$

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

$$P(D \mid M_i, I) = \int \mathrm{d}a_i \, P(D \mid \mathbf{a}_i, M_i, I) \, P(\mathbf{a}_i \mid M_i, I) \tag{4.9}$$

$$\approx P(D \mid \mathbf{a}_i^*, M_i, I) \, P(\mathbf{a}_i^* \mid M_i, I) \; (2\pi)^{|\mathbf{a}_i|/2} \sqrt{\det \boldsymbol{\sigma}_i^2}, \tag{4.10}$$

where $\boldsymbol{\sigma}_i^2$ is now the covariance matrix of the Gaussian approximation of the posterior under model M_i. Note that we have already computed all of factors in the above product in the parameter estimation problem. Therefore, we already have what we need to compute the (approximate) odds ratio.

4.5 The factors in the odds ratio

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

$$O_{ij} \approx \left(\frac{P(M_i \mid I)}{P(M_j \mid I)}\right)\left(\frac{P(D \mid \mathbf{a}_i^*, M_i, I)}{P(D \mid \mathbf{a}_j^*, M_j, I)}\right)\left(\frac{P(\mathbf{a}_i^* \mid M_i, I)\ (2\pi)^{|\mathbf{a}_i|/2}\sqrt{\det \boldsymbol{\sigma}_i^2}}{P(\mathbf{a}_j^* \mid M_j, I)\ (2\pi)^{|\mathbf{a}_j|/2}\sqrt{\det \boldsymbol{\sigma}_j^2}}\right). \tag{4.11}$$

- 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 single parameter model where the parameter has a uniform prior. Then,

$$\text{Occam factor} \propto P(a \mid M_i, I)\sigma_i = \frac{\sigma_i}{a_{\max} - a_{\min}}. \tag{4.12}$$

  Now, compare a model, with one parameter (a) with uniform prior to one with two (a and b). In this case, we have

$$P(a^* \mid M_i, I) = \frac{1}{a_{\max} - a_{\min}}, \tag{4.13}$$

$$P(a^*, b^* \mid M_j, I) = \frac{1}{a_{\max} - a_{\min}}\,\frac{1}{b_{\max} - b_{\min}}. \tag{4.14}$$

  So, the volume of the parameter space model M_j is larger than M_i, so the this part of the odds ratio is greater than one, favoring the model with fewer parameters. The ratio of Occam factors is then

$$\frac{\sigma_i}{\sqrt{2\pi\ \det \boldsymbol{\sigma}_j^2}}\,(b_{\max} - b_{\min}). \tag{4.15}$$

  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 it penalizes more complicated models.[14]

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

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

4.6 Example: Are two data sets from the same 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. Model 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

$$O_{12} = \frac{P(M_1 \mid I)}{P(M_2 \mid I)} \frac{P(D_c, D_e \mid M_1, I)}{P(D_c, D_e \mid M_2, I)}, \tag{4.16}$$

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 $P(M_i \mid I) = P(M_j \mid I)$. Then, we are left to compute $P(D_c, D_e \mid M_1, I)$ and $P(D_c, D_e \mid M_2, I)$. 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

$$O_{12} \approx \frac{\sigma_{\max}\,(\mu_{\max} - \mu_{\min})}{\pi\sqrt{2}} \; \frac{n_1\, n_2\, s^{2-n1-n2}}{(n_1 + n_2)\, s_1^{2-n_1}\, s_2^{2-n_2}}, \tag{4.17}$$

where

$$s^2 = \frac{1}{n_1 + n_2} \sum_{i \in D_1 \cup D_2} (x_i - \bar{x})^2, \tag{4.18}$$

$$s_1^2 = \frac{1}{n_1} \sum_{i \in D_1} (x_i - \bar{x}_1)^2, \tag{4.19}$$

$$s_2^2 = \frac{1}{n_2} \sum_{i \in D_2} (x_i - \bar{x}_2)^2, \tag{4.20}$$

with

$$\bar{x} = \frac{1}{n_1 + n_2} \sum_{i \in D_1 \cup D_2} x_i, \tag{4.21}$$

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i \in D_1} x_i, \tag{4.22}$$

$$\bar{x}_2 = \frac{1}{n_2} \sum_{i \in D_2} x_i. \tag{4.23}$$

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 it.

4.7 Computing odds ratios without the Laplace approximation

We can use a technique called parallel-tempering Markov chain Monta Carlo (PTMCMC) to compute odds ratios without making the Laplace approximation. As you likely have guessed, this is computationally intensive, but effective. We will learn about this in an upcoming lecture.