Assorted Thoughts on Bayesian Techniques

Ilya Mandel* and Will M. Farr†

Northwestern University

(Dated: October 29, 2009)

I. INTRODUCTION

Bayesian techniques can be used for parameter estimation, including the estimation of uncertainties in parameter determination, as well as model selection.

Bayes' theorem is fundamental to Bayesian techniques. It relates the probabilities of observational data, d, and parameters describing the target of the observations, $\vec{\theta}$, within the framework of a model, M. Consider the joint probability of a particular observation and parameter value given the model, $p(\vec{\theta}, d|M)$. By the axioms of probability, we can write this in two ways:

$$p(\vec{\theta}, d|M) = p(\vec{\theta}|d, M)p(d|M) = p(d|\vec{\theta}, M)p(\vec{\theta}|M). \tag{1}$$

We are particularly interested in the quantity $p(\vec{\theta}|d, M)$, the probability of a particular parameter value given the observations we have made (in the context of the model we have chosen). This quantity is called the *posterior probability*. Isolating the posterior probability in equation (1) gives

$$p(\vec{\theta}|d, M) = \frac{p(d|\vec{\theta}, M)p(\vec{\theta}|M)}{p(d|M)}. \tag{2}$$

The quantities on the right hand side of this relation are either computable or under our control. The probability of the data given parameter values, $p(d|\vec{\theta}, M)$, known as the *likelihood*, is computable in the context of the model. The probability of the parameters given the model, $p(\vec{\theta}|M)$, called the *prior probability*, is something we choose; it reflects all our knowledge about parameter values before we made the observation that resulted in the data, d. The probability of the data given the model, $p(d|M)$, called the *evidence*, can be computed

*Electronic address: ilyamandel@chgk.info

†Electronic address: w-farr@northwestern.edu

using

$$p(d|M) = \int p(d|\vec{\theta}, M)p(\vec{\theta}|M)d\vec{\theta}. \tag{3}$$

(The evidence is required for normalization of the posterior probability. If we only care about relative posterior probabilities, we can ignore the evidence.)

Example 1 *A disease occurs throughout the population at a rate of one per ten thousand. A test exists for the disease that returns a positive result 99.9% of the time when the test subject has the disease, and a negative result 99% of the time when the test subject does not have the disease. Given a positive test result, what is the probability that the test subject has the disease?*

Here the data, d, is the positive result, the model is the given rates of test results and disease incidence, and the parameter we wish to estimate is the disease status of the test subject. Denote the diseased state by $\vec{\theta} = +$ *and the un-diseased state by* $\vec{\theta} = -$*. We have*

$$p(d|+, M) = 0.999 \tag{4}$$

$$p(d|-, M) = 0.01 \tag{5}$$

$$p(+|M) = 0.0001 \tag{6}$$

$$p(-|M) = 0.9999 \tag{7}$$

From the first two probabilities, we conclude that the evidence is

$$p(d|M) = p(d|+, M)p(+|M) + p(d|-, M)p(-|M) = 0.0100989. \tag{8}$$

Bayes' theorem tells us that

$$p(+|d, M) = \frac{p(d|+, M)p(+|M)}{p(d|M)} = 0.00989\ldots, \tag{9}$$

from which we see that it is still much more likely that the unfortunate test subject is the victim of a false positive than that the test subject actually has the disease.

Example 2 *We are interested in knowing the maximum spin frequency of pulsars in the galaxy. Unfortunately, we have not actually observed any pulsars whose spin we can determine (of course, this isn't true, but play along). All we know is that the pulsar spin frequency must be greater than zero, and must be less than a maximum value,* Ω*, that results in the surface of the pulsar rotating with a velocity exceeding the speed of light. To parametrize our*

ignorance, we will assume a prior that is uniform in frequency over the range $[0, \Omega]$ *for the maximum spin frequency of pulsars,* ω_m*:*

$$p(\omega_m|M) = \begin{cases} \frac{1}{\Omega} & 0 \leq \omega_m \leq \Omega \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

Furthermore, we will assume that the distribution of pulsar frequencies in the galaxy is uniform between 0 and ω_m *(this is also not actually true for pulsars in the galaxy):*

$$p(\omega|\omega_m, M) = \begin{cases} \frac{1}{\omega_m} & 0 \leq \omega \leq \omega_m \\ 0 & \text{otherwise} \end{cases}. \tag{11}$$

Now we observe a single pulsar spinning at frequency ω*. We want to know what the probability distribution for* ω_m *is given this observation. We begin by computing the evidence:*

$$p(\omega|M) = \int d\omega_m \, p(\omega|\omega_m, M) p(\omega_m|M) = \frac{1}{\Omega} \int_\omega^\Omega \frac{1}{\omega_m} d\omega_m = \frac{1}{\Omega} \log\left(\Omega/\omega\right). \tag{12}$$

By Bayes' theorem, we have

$$p(\omega_m|\omega, M) = \frac{p(\omega|\omega_m, M) p(\omega_m|M)}{p(\omega|M)} = \begin{cases} \frac{1}{\log(\Omega/\omega)} \frac{1}{\omega_m} & \omega \leq \omega_m \leq \Omega \\ 0 & \text{otherwise} \end{cases}. \tag{13}$$

Note that the posterior distribution for the maximum spin frequency is zero for $\omega_m < \omega$ *(since we have definitely observed a pulsar spinning with frequency* ω*), obtains a maximum at* $\omega_m = \omega$*, and drops (discontinuously) to 0 for* $\omega_m > \Omega$*. Bayes' theorem has blended the information we had in the prior (*$0 \leq \omega_m \leq \Omega$*) with the information from our new observation to obtain the posterior distribution.*

A warning: be careful with priors. In this example, if $\omega > \Omega$*, Bayes' theorem naively tells us that* $p(\omega_m|\omega, M) = 0/0$*. This breakdown occurs because our prior tells us that the probability of the measurement is zero!*

Example 3 *A second pulsar is observed with spin frequency* $\omega' > \omega$*. What is the distribution for the maximum pulsar spin frequency incorporating this new data?*

We use the posterior distribution from the last example for our prior in this example:

$$p(\omega_m|M) = \begin{cases} \frac{1}{\log(\Omega/\omega)} \frac{1}{\omega_m} & \omega \leq \omega_m \leq \Omega \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

Computing the evidence gives

$$p(\omega'|M) = \int d\omega_m p(\omega'|\omega_m, M) p(\omega_m|M)$$
$$= \int_{\omega'}^{\Omega} d\omega_m \frac{1}{\omega_m} \frac{1}{\log(\Omega/\omega)} \frac{1}{\omega_m} = \frac{1}{\log(\Omega/\omega)} \left[\frac{1}{\omega'} - \frac{1}{\Omega}\right]. \quad (15)$$

Plugging into Bayes' theorem, we have

$$p(\omega_m|\omega', M) = \begin{cases} \left[\frac{1}{\omega'} - \frac{1}{\Omega}\right]^{-1} \frac{1}{\omega_m^2} & \omega' \leq \omega_m \leq \Omega \\ 0 & \text{otherwise} \end{cases}. \quad (16)$$

Note that the posterior probability is independent of ω (as it should be, since ω is not the maximum observed pulsar spin anymore). But, the probability falls faster for $\omega_m > \omega'$ than it would had we only observed one pulsar spinning at frequency ω': Bayes' theorem has incorporated the fact that we have made two *observations instead of one, and are therefore more certain about the maximum spin. It is an interesting exercise to verify that the result is the same if one computes the posterior probability assuming that both observations are made simultaneously instead of sequentially (hint: consider carefully the evidence in this case).*

Finally, note that we can take $\Omega \to \infty$ in equation (16) *(unlike in equation* (13)*), to obtain*

$$p_{\Omega\to\infty}(\omega_m|\omega', M) = \begin{cases} \frac{\omega'}{\omega_m^2} & \omega_m > \omega' \\ 0 & \text{otherwise} \end{cases}. \quad (17)$$

Apparently, two observations is sufficient to constrain the posterior distribution for ω_m even in the presence of "infinite" uncertainty on the prior, while one observation is insufficient.

Example: the elephant and the soda can.

The likelihood depends on the particular problem you are dealing with. Some likelihoods are physically motivated. For example, maybe one is making a measurement of a quantity using a device which is known to have Gaussian errors with a particular standard deviation; in this case, the likelihood should be a Gaussian about the parameters with the given standard deviation. Or, perhaps the device making the measurement is sufficiently well-understood that it can be simulated. Then the likelihood could come from simulations of the device with the given parameters and random errors representing measurement noise included in the simulation. Other times there is no obvious physical reason for preferring

one likelihood over another. In these cases it is common to choose

$$p(d|\vec{\theta}, M) \propto \exp\left[-\sum_i \left(\frac{d_i - \hat{d}_i(\vec{\theta})}{\sigma_i(\vec{\theta})}\right)^2\right], \tag{18}$$

where $\hat{d}(\vec{\theta})$ is the data that would be predicted by the model with parameters $\vec{\theta}$, and $\sigma(\vec{\theta})$ is the model of the standard deviation of the corresponding data generated by model parameters $\vec{\theta}$. For this choice, the maximum likelihood parameters are the same as would be found via χ-squared minimization. Other choices for the likelihood may work better for some problems. In any case, the following methods for determining the posterior PDF are independent of the form of the likelihood.

II. MARKOV CHAIN MONTE CARLO MOTIVATION

How do we determine the PDFs over a large, multi-dimensional parameter space?

Could try to lay out a grid in the parameter space. But this quickly becomes impractical for large numbers of parameters (grid size grows exponentially).

Could try to use some variant on gradient-descent (e.g., simplex, amoeba). But this can get stuck on a local maximum. Could start may chains - but still no guarantee of finding the true maximum.

And even if the true posterior maximum is found, it's hard to estimate the accuracy. Can try to locally approximate the parameter-space likelihood contours quadratically - this yield the Fisher information matrix. But it's local, doesn't take other maxima into account, could overstate accuracy.

Idea: sample stochastically, but sample more in parameter-space regions with high posterior. In fact, can sample according to posterior PDF: Markov Chain Monte Carlo (MCMC).

III. METROPOLIS-HASTINGS MCMC

We want to sample the posterior distribution via a sequence of parameter values, $\{\vec{\theta}_i \,|\, i = 0, 1, \ldots\}$, where each $\vec{\theta}_i$ is generated by a Markov chain with transition probability $P(\vec{\theta}_{i-1} \to \vec{\theta}_i)$ that depends only on the previous state. Let $p(\vec{\theta}_i)$ be the posterior PDF. Then, the transition probability must satisfy a condition called *detailed balance*:

$$p(\vec{\theta})P(\vec{\theta} \to \vec{\theta}') = p(\vec{\theta}')P(\vec{\theta}' \to \vec{\theta}). \tag{19}$$

Detailed balance says that the probability of transitions between states *in the equilibrium distribution* is the same in both directions.

To see that detailed balance is a necessary condition for sampling the posterior PDF with a Markov chain, suppose that at some point, i, in the sequence the distribution for $\vec{\theta}_i$ is the posterior PDF, $p(\vec{\theta}_i)$. Then the PDF for the next element of the sequence, $\tilde{p}(\vec{\theta}_{i+1})$, is given by

$$\tilde{p}(\vec{\theta}_{i+1}) = \int d\vec{\theta}_i \, P(\vec{\theta}_i \to \vec{\theta}_{i+1}) p(\vec{\theta}_i). \tag{20}$$

We are using $\tilde{p}$ for the probability distribution *implied* by our jump probability; p denotes the posterior PDF we want to sample. When detailed balance is satisfied, equation (19) lets us rewrite this as

$$\tilde{p}(\vec{\theta}_{i+1}) = \int d\vec{\theta}_i \, \frac{P(\vec{\theta}_{i+1} \to \vec{\theta}_i) p(\vec{\theta}_{i+1})}{p(\vec{\theta}_i)} p(\vec{\theta}_i) = p(\vec{\theta}_{i+1}) \int d\vec{\theta}_i \, P(\vec{\theta}_{i+1} \to \vec{\theta}_i) = p(\vec{\theta}_{i+1}). \tag{21}$$

Thus we see that detailed balance is required for the probability distribution on subsequent points to be the posterior PDF if the initial probability distribution is the posterior PDF. It turns out that detailed balance is also a sufficient condition for the probability distribution on the sequence of values $\{\vec{\theta}_i \,|\, i = 0, 1, \ldots\}$ to approach p *asymptotically.*

Now the problem of sampling from the posterior PDF has been reduced to the problem of computing a jump probability for a Markov chain that satisfies detailed balance, equation (19). The Metropolis-Hastings algorithm is one algorithm that allows us to construct such a jump probability.

The Metropolis-Hastings algorithm works as follows. When in state $\vec{\theta}_i$, we choose a trial state $\vec{\theta}'$ based on some "trial jump" probability distribution $Q(\vec{\theta}_i \to \vec{\theta}')$. We are free to choose the trial jump probability distribution, so long as it is possible to completely explore parameter space with trial jumps[1]. We accept the trial state $\vec{\theta}'$ with a probability given by

$$R(\vec{\theta}_i \to \vec{\theta}') \equiv \min \left\{ \frac{p(\vec{\theta}') Q(\vec{\theta}' \to \vec{\theta}_i)}{p(\vec{\theta}_i) Q(\vec{\theta}_i \to \vec{\theta}')}, \; 1 \right\} \tag{22}$$

If the proposed trial state is accepted, $\vec{\theta}_{i+1} = \vec{\theta}'$; otherwise, we set $\vec{\theta}_{i+1} = \vec{\theta}_i$.

The Metropolis-Hastings procedure generates a jump probability $P(\vec{\theta}_i \to \vec{\theta}_{i+1})$ that satisfies detailed balance. Here is the proof. First, note that if $R(\vec{\theta}_i \to \vec{\theta}') < 1$ then $R(\vec{\theta}' \to \vec{\theta}_i) = 1$, and vice versa. Assume that $R(\vec{\theta}_i \to \vec{\theta}') < 1$. Then

$$P(\vec{\theta}_i \to \vec{\theta}') p(\vec{\theta}_i) = R(\vec{\theta}_i \to \vec{\theta}') Q(\vec{\theta}_i \to \vec{\theta}') p(\vec{\theta}_i) = p(\vec{\theta}') Q(\vec{\theta}' - \vec{\theta}_i), \tag{23}$$

while

$$P(\vec{\theta}' \to \vec{\theta}_i)p(\vec{\theta}') = R(\vec{\theta}' \to \vec{\theta}_i)Q(\vec{\theta}' \to \vec{\theta}_i)p(\vec{\theta}') = Q(\vec{\theta}' \to \vec{\theta}_i)p(\vec{\theta}') = P(\vec{\theta}_i \to \vec{\theta}')p(\vec{\theta}_i), \quad (24)$$

from which we see that P satisfies detailed balance.

IV. CHALLENGES AND VARIATIONS ON MCMC

Biggest problem: the difficulty of sampling a complicated, highly structured parameter space, which may have a multimodal structure. Islands in an ocean.

Jump proposal distribution Q is key to good sampling. Random jumps. Jumps that follow correlations between parameters - bigger jumps in directions to which the likelihood is less sensitive; estimated with correlation matrix or Fisher matrix. Adaptive jump size refinement.

Two broad classes of techniques that can improve sampling efficiency:

1. Quickly finding local maxima.

General smoothing of structure: simulated/thermostated annealing; frequency annealing; parallel tempering.

Using techniques that violate detailed balance to first get a picture of the structure (e.g., sometimes you can cheaply maximize over some of the parameters), then using true MCMC to obtain PDFs.

2. Better jumps between maxima.

Directed jump proposal distributions. Case-specific, based on likelihood space knowledge. Island hopping, harmonic identification, reconnaissance, delayed rejection.

How to set the initial burn-in to avoid sensitivity to initial conditions?

When can you confidently stop?

V. MODEL SELECTION

We can also use Bayesian techniques to carry out model selection when we have several different models available. The odds ratio for two models M_1 and M_2 is:

$$O = \frac{p(M_1)}{p(M_2)} \frac{\int d\vec{\theta} p(\vec{\theta}|d, M_1)}{\int d\vec{\theta} p(\vec{\theta}|d, M_2)}. \quad (25)$$

Here, the first factor is the prior ratio of the odds for the two models, which is either set by a general theoretical understanding, or fixed to unity. The second factor is the ratio of the marginal likelihoods, i.e., the integrals of the posteriors for the two models

$$B = \frac{\int d\vec{\theta} p(\vec{\theta}|d, M_1)}{\int d\vec{\theta} p(\vec{\theta}|d, M_2)} = \frac{\int d\vec{\theta} p(\vec{\theta}|M_1) p(d|\vec{\theta}, M_1)}{\int d\vec{\theta} p(\vec{\theta}|M_2) p(d|\vec{\theta}, M_2)}. \tag{26}$$

This is known as the Bayes factor.

Another way to think about model selection is to consider a "meta-model", $\tilde{M}$, that encompasses both models M_1 and M_2. Then the specific model being applied becomes just another parameter; the priors now include the best estimate we have for the probability of M_1 versus M_2 before taking any data. Applying Bayes' theorem gives

$$p(\vec{\theta}_i, M_i|d, \tilde{M}) = \frac{p(d|\vec{\theta}_i, M_i, \tilde{M}) p(\vec{\theta}_i|M_i, \tilde{M}) p(M_i|\tilde{M})}{p(d|\tilde{M})}, \tag{27}$$

where we write $\vec{\theta}_i$ for the parameters of model M_i (since, for example, the models may have entirely different numbers of parameters we must distinguish which parameter space we are talking about) and the evidence is now calculated by summing over the two models:

$$p(d|\tilde{M}) = \int d\vec{\theta}_1 \, p(d|\vec{\theta}_1, M_1, \tilde{M}) p(\vec{\theta}_1|M_1, \tilde{M}) p(M_1|\tilde{M}) + \int d\vec{\theta}_2 \, p(d|\vec{\theta}_2, M_2, \tilde{M}) p(\vec{\theta}_2|M_2, \tilde{M}) p(M_2|\tilde{M}). \tag{28}$$

We get both the probability distribution of the parameters *within* each model, $p(\vec{\theta}_i, M_i|d, \tilde{M})$, and the total probability of each model for any value of its parameters,

$$p(M_i|d, \tilde{M}) = \int d\vec{\theta}_i \, p(\vec{\theta}_i, M_i|d, \tilde{M}). \tag{29}$$

(Compare to equation (25).)

VI. ALTERNATIVES

Genetic algorithms.

Nested Sampling – to be discussed next week. MultiNest.

Model-independent exploration.

Join us for a reading group if you want to learn more!

---

[1] We will see later that jump distributions that make jumping to regions of high posterior PDF probable are more efficient. Apparently, the optimal distribution for trial steps when the posterior is a high-dimensional Gaussian results in an "acceptance rate" of about 25%; this is a good target in general.