6. Bayesian Phylogenetics
EEOB563
Spring 2021

1 Preliminary considerations

In ML we choose the hypothesis that maximizes the probability of the data [$e.g.$, $L(H) = P(D|H)$]. This makes sense, but the L number itself is rather meaningless to us. What we really want to know is the probability of the hypothesis given the data or $P(H|D)$.

To evaluate this probability we use an equation named after Thomas Bayes (1702-1761), an English mathematician and Presbyterian minister, which shows us how opinions about a hypothesis held prior to the experiment should be modified by the evidence of the experiment:

$$P(H_i|E) = \frac{P(E|H_i)P(H_i)}{P(E)} \tag{1}$$

In this equation $P(H_i|E)$ is the posterior probability of the hypothesis i after experiment E; $P(E|H_i)$ is the likelihood of the hypothesis i; and $P(H_i)$ is its prior probability.

In essence, Bayesian data analysis has two foundational ideas.

1. Bayesian inference is reallocation of credibility across possibilities;

2. the possibilities, over which we allocate credibility, are parameter values in meaningful mathematical models.

2 Some examples

2.1 Example with marbles

To understand the equation above, we have to understand the concepts of joint, marginal, and conditional probabilities, which can be explained using an example with marbles reviewed in class.

Here the joint probability ($e.g.$,$P(W, D)$) can be calculated as the product of a certain condition for one of the traits ($e.g.$, being white ($P(W)$) and a conditional probability of the second trait given the state of the first ($e.g.$, $P(D|W)$). Notice that we can't simply multiply probabilities of the two traits because we don't know if they are independent.
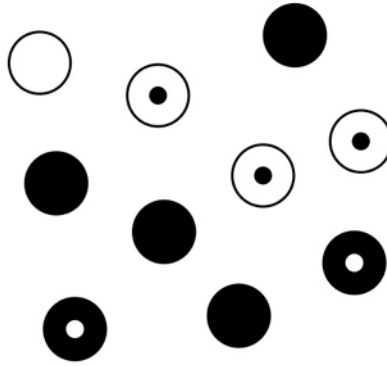
2.2 Paternity testing

Our second (hypothetical) example is about paternity testing. Let's say that we have a mother, who is homozygous recessive for a trait and a child who is heterozygous. What is the genotype of the father?

Genotype of a mother: aa
Genotype of a child: Aa
Genotype of the father?

If we know genetics, and the father is biological, we have to consider two possibilities: AA and Aa, hence two hypotheses:

	H1	H2	Row Sum
Genotype	AA	Aa	-
Prior probability	0.5	0.5	1
Likelihood	1	0.5	-
Likelihood * Prior	0.5	0.25	0.75
Posterior probability	$\frac{2}{3}$	$\frac{1}{3}$	1

So, even if we don't know anything about the frequencies of two alleles in a population (flat prior), but we know that the child is Aa, the probability that the father is AA is 2/3.

Homework question:
• Suppose that we have a rare genetic disease in the human population with the occurrence rate of 1 in 1000.
• Suppose that you bought one of those fancy genomic screening kits, which included a test for the disease and the test came back positive (although you have no symptoms yet!)
• You frantically read the instructions and find out that the test has a 99% hit rate, meaning that if a person has the disease, then the test result is positive 99% of the time.
• You also read that the test has a false alarm rate of 5%
• What is the probability that you have a disease?

2.3 More complex examples

So far, we considered discrete characters/parameters. In case of continuous parameters (e.g., rates, branch lengths, frequencies of nucleotides, etc), we need to modify and use probability densities rather than probabilities per se:

$$f(\Theta|D) = \frac{f(D|\Theta)f(\Theta}{\int_{\Theta} f(D|\Theta)f(\Theta)d\Theta} \tag{2}$$

where $f(\Theta|D)$ is the posterior probability density; $f(D|\Theta)$ – the likelihood, $f(\Theta)$ – the prior probability density, and the whole denominator – the marginal probability of the data.

It is usually impossible to do Bayesian calculations analytically because denominator involves summing and integrating over all possible hypotheses. But it is relatively easy to calculate the ratio of posterior probabilities, as two denominators cancel out. Random sampling of parameter space is also impossible (Huge space, most of it with very low probability). Thus posterior probability density is approximated by the Markov Chain Monte Carlo (MCMC) method.

3 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a general computational technique for evaluating sums and integrals, especially those that arise as probabilities or expectations under complex probability distributions. **Monte Carlo** implies that the method is based on using chance. **Markov Chain** indicates a *dependent sampling scheme* with the probability distribution of each sampled point depending on the value of the previous one.

MCMC works by taking a series of steps that form a conceptual chain. The algorithm used to make these steps is known as **Metropolis-Hastings** algorithm. At each iteration, the algorithm picks a candidate for the next sample value based on the current sample value. Then, with some probability, the candidate is either accepted (in which case the candidate value is used in the next iteration) or rejected (in which case the candidate value is discarded, and current value is reused in the next iteration). The probability of acceptance is determined by comparing the likelihoods of the current and candidate sample values.

Metropolis-coupled Markov chain Monte Carlo (MCMCMC) involves running several chains simultaneously. One is the cold chain, the rest are heated chains. Chain is heated by raising densities to a power less than 1.0. These heated chains increase the acceptance probabilities making it easier to cross the likelihood valleys. Chains are allowed to swap with the cold chain from time to time. Only the cold chain is sampled.

4   Comparison between Bayesian and ML inference (p. 175-176 of the textbook)

1. Both are model-based methods in which a probabilistic description of the data generation process is assumed;

2. Both involve computing P(data — p) for parameter values p. In Maximum Likelihood, this is the Likelihood function we seek to maximize; in a Bayesian analysis it appears in the process of computing the posterior distribution.

3. A Bayesian analysis requires specifying a prior distribution of the parameters we seek to infer, which expresses our beliefs in what parameter values are likely before the data are considered.

4. Maximum Likelihood produces a single estimate, called a point estimate, of the parameters we infer. A Bayesian analysis produces a distribution of estimates, the posterior distribution, which indicates differing levels of support for different parameter values.