Introduction to Bayesian Statistics

Harvey Thornburg
Center for Computer Research in Music and Acoustics (CCRMA)
Department of Music, Stanford University
Stanford, California 94305

August 9, 2005

Statistical approaches to parameter estimation and hypothesis testing which use *prior distributions* over parameters are known as *Bayesian* methods. The following notes briefly summarize some important facts.

Outline

- Bayesian Parameter Estimation
- Bayesian Hypothesis Testing
- Bayesian Sequential Hypothesis Testing

Bayesian Parameter Estimation

- Let y be distributed according to a *parametric family*: $y \sim f_\theta(y)$. The goal is, given iid observations $\{y_i\}$, to estimate θ. For instance, let $\{y_i\}$ be a series of coin flips where $y_i = 1$ denotes "heads" and $y_i = 0$ denotes "tails". The coin is weighted, so $P(y_i = 1)$ can be other than $1/2$. Let us define $\theta = P(y_i = 1)$; our goal is to estimate θ. This simple distribution is given the name "Bernoulli".
- Without prior information, we use the *maximum likelihood* approach. Let the observations be $y_1 \ldots y_{H+T}$. Let H be the number of heads observed and T be the number of tails.

$$\begin{aligned} \hat{\theta} &= \text{argmax} f_\theta(y_{1:H+T}) \\ &= \text{argmax} \theta^H (1-\theta)^T \\ &= H/(H+T) \end{aligned}$$

- Not surprisingly, the probability of heads is estimated as the empirical frequency of heads in the data sample.
- Suppose we remember that yesterday, using the same coin, we recorded 10 heads and 20 tails. This is one

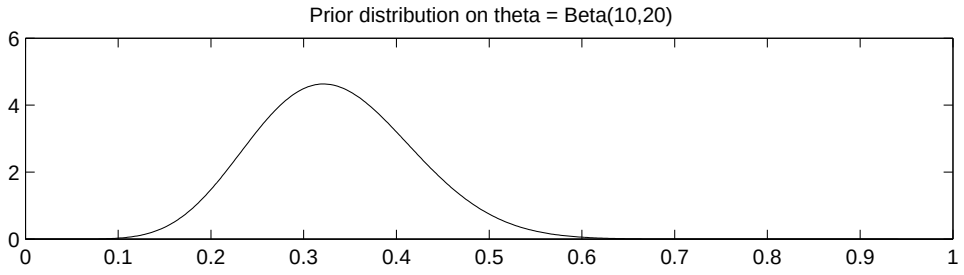
way to indicate "prior information" about θ. We simply include these past trials in our estimate:

$$\hat{\theta} = (10 + H)/(10 + H + 20 + T)$$

- As (H+T) goes to infinity, the effect of the past trials will wash out.

- Suppose, due to computer crash, we had lost the details of the experiment, and our memory has also failed (due to lack of sleep), that we forget even the number of heads and tails (which are the *sufficient statistics* for the Bernoulli distribution). However, we believe the probability of heads is about $1/3$, but this probability itself is somewhat uncertain, since we only performed 30 trials.
- In short, we claim to have a *prior distribution* over the probability θ, which represents our prior belief. Suppose this distribution is $P(\theta)$ and $P(\theta) \sim \mathrm{Beta}(10, 20)$:

$$g(\theta) \;=\; \frac{\theta^9(1-\theta)^{19}}{\int \theta^9(1-\theta)^{19} d\theta}$$



- Now we observe a new sequence of tosses: $y_{1:H+T}$. We may calculate the *posterior* distribution

$P(\theta|y_{1:H+T})$ according to *Bayes' Rule*:

$$\begin{aligned} P(\theta|y) &= \frac{P(y|\theta)P(\theta)}{P(y)} \\ &= \frac{P(y|\theta)P(\theta)}{\int P(y|\theta)P(\theta)d\theta} \end{aligned}$$

The term $P(y|\theta)$ is, as before, the *likelihood* function of θ. The *marginal* $P(y)$ comes by integrating out θ:

$$P(y) = \int P(y|\theta)P(\theta)d\theta$$

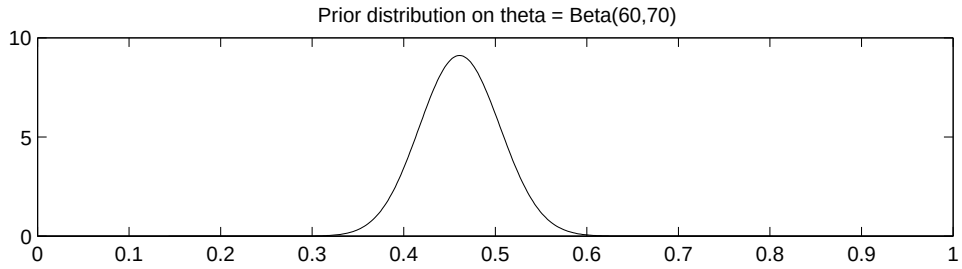
- To continue our example, suppose we observe in the new data $y(1:H+T)$ a sequence of 50 heads and 50 tails. The likelihood becomes:

$$P(y|\theta) = \theta^{50}(1-\theta)^{50}$$

- Plugging this likelihood and the prior into the Bayes Rule expression, and doing he math, obtains the posterior distribution as a $\mathrm{Beta}(10+50, 20+50)$:

$$P(\theta|y) = \frac{\theta^{59}(1-\theta)^{69}}{\int \theta^{59}(1-\theta)^{69}d\theta}$$

- Note that the posterior and prior distribution have the same form. We call such a distribution a *conjugate prior*. The Beta distribution is conjugate to the



binomial distribution which gives the likelihood of iid Bernoulli trials. As we will see, a conjugate prior perfectly captures the results of past experiments. Or, it allows us to express prior belief in terms of "invented" data. More importantly, conjugacy allows for efficient *sequential updating* of the posterior distribution, where the posterior at one stage is used as prior for the next.

- **Key Point** The "output" of the Bayesian analysis is not a *single estimate* of θ, but rather *the entire posterior distribution*. The posterior distribution summarizes all our "information" about θ. As we get more data, if the samples are truly iid, the posterior distribution will become more sharply peaked about a single value.
- Of course, we can use this distribution to make *inference* about θ. Suppose an "oracle" was to tell us the true value of θ used to generate the samples. We want to guess θ that minimizes the mean squared error between our guess and the true value. This is the same criterion as in maximum likelihood estimation. We would choose the *mean* of the posterior distribution, because we know conditional mean minimizes mean square error.
- Let our prior be $\mathrm{Beta}(H_0, T_0)$ and

$$\begin{aligned} \hat{\theta} &= E(\theta | y_{1:N}) \\ &= \frac{H_0 + H}{H_0 + H + T_0 + T} \end{aligned}$$

- The same way, we can do *prediction*. What is

$P(y_{N+1} = 1|y_{1:N})$?

$$\begin{aligned} P(y_{N+1} = 1|y_{1:N}) &= \int P(y_{N+1} = 1|\theta, y_{1:N}) P(\theta|y_{1:N}) d\theta \\ &= \int P(y_{N+1} = 1|\theta) P(\theta|y_{1:N}) d\theta \\ &= \int \theta P(\theta|y_{1:N}) d\theta \\ &= E(\theta|y_{1:N}) \\ &= \frac{H_0 + H}{H_0 + H + T_0 + T} \end{aligned}$$

Bayesian Hypothesis Testing

- Suppose we have a fixed iid data sample $y_{1:N} \sim f_\theta(y)$. We have two choices: $\theta = \theta_0$ or $\theta = \theta_1$. That is, the data $y_{1:N}$ is generated by *either* θ_0 or θ_1. Call θ_0 the "null" hypothesis and θ_1 the "alternative". The alternative hypothesis indicates a disturbance is present. If we decide $\theta = \theta_1$, we signal an "alarm" for the disturbance.
- We process the data by a decision function $g(y_{1:N})$

$$\begin{aligned} D(y_{1:N}) &= 0, \theta = \theta_0 \\ &= 1, \theta = \theta_1 \end{aligned}$$

- We have two possible errors:
  - **False Alarm**: $\theta = \theta_0$, but $D(y_{1:N}) = 1$
  - **Miss**: $\theta = \theta_1$, but $D(y_{1:N}) = 0$
- In the non-Bayesian setting, we wish to choose a family of $D(\cdot)$, which navigate the optimal tradeoff between the probabilities of miss and false alarm.
- The probability of miss, P_M, is $P(D(y_{1:N})) = 0 | \theta = \theta_1)$ and the probability of false alarm, P_{FA}, is $P(D(y_{1:N})) = 1 | \theta = \theta_0)$.

- We optimize the tradeoff by comparing the likelihood ratio to a nonnegative threshold, say $\exp(T) > 0$:

$$D_*(y_{1:N}) \;=\; 1_{\frac{f_{\theta_1}(y)}{f_{\theta_0}(y)} > \exp(T)}$$

- Equivalently, compare the log likelihood ratio to an arbitrary real threshold T:

$$D_*(y_{1:N}) \;=\; 1_{\log \frac{f_{\theta_1}(y)}{f_{\theta_0}(y)} > T}$$

- Increasing T makes the test less "sensitive" for the disturbance: we accept a higher probability of miss in return for a lower probability of false alarm. Because of the tradeoff, there is a limit as to how well we can do, which improves exponentially as we collect more data. This limit relation is given by *Stein's lemma*. Fix $P_M = \epsilon$. Then, as $\epsilon \to 0$, and for large N, we get:

$$\frac{1}{N} \log P_{FA} \;\to\; -K(f_{\theta_0}, f_{\theta_1})$$

- The quantity $K(f_{\theta_0}, f_{\theta_1})$ is the *Kullback-Leibler* distance, or the expected value of the log likelihood ratio. We define, where f and g are densities:

$$K(f, g) \;=\; E_f\left[\log(g/f)\right]$$

  The following facts about Kullback-Leibler distance hold:

- $K(f, g) \geq 0$. Equality holds when $f \equiv g$ except on a set of $(f + g)/2$-measure zero. I.E. for a continuous sample space you can allow difference on sets of Lebesgue measures zero, for a discrete space you cannot allow any difference.
- $K(f, g) \neq K(g, f)$, in general. So the K-L distance is not a metric. The triangle inequality also fails.

- When f, g belong to the same parametric family, we adopt the shorthand: $K(\theta_0, \theta_1)$ rather than $K(f_{\theta_0}, f_{\theta_1})$. Then we have an additional fact. When hypotheses are "close", K-L distance behaves approximately like the square of the Euclidean metric in parameter (θ)-space. Specifically:

$$2K(\theta_0, \theta_1) \approx (\theta_1 - \theta_0)' J(\theta_0)(\theta_1 - \theta_0).$$

  where $J(\theta_0)$ is the Fisher information. The right hand side is sometimes called the square of the *Mahalanobis distance*.

- Furthermore, we may assume the hypotheses are "close" enough that $J(\theta_0) \approx J(\theta_1)$. Then, K-L information appears also symmetric.

- Practically there is still the problem to choose T, or to choose "desirable" probabilities of miss and false alarm which obey Stein's lemma, which gives also the data size. We can solve for T given the error probabilities. However, it is often "unnatural" to specify these probabilities; instead, we are concerned about other, observable effects on the system. Hence, the usual scenario results in a lot of lost sleep, as we are continually varying T, running simulations, and then observing some distant outcome.

- Fortunately, the Bayesian approach comes to the

rescue. Instead of optimizing a probability tradeoff, we assign *costs*: $C_M > 0$ to a miss event and $C_{FA} > 0$ to a false alarm event. Additionally, we have a *prior distribution* on θ

$$P(\theta = \theta_1) \; = \; \pi_1$$

- Let $D(y_{1:N})$ be the decision function as before. The *Bayes risk*, or expected cost, is as follows.

  $$R(D) \;=\; \pi_1 E\left[D(y_{1:N}) = 0 | \theta = \theta_1\right] + (1-\pi_1) E\left[D(y_{1:N}) = 1\right.$$

- It follows, the optimum-Bayes risk decision *also* involves comparing the likelihood ratio to a threshold:

  $$\begin{aligned} D(y_{1:N}) \;&=\; 1_{\frac{P(y|\theta_1)}{P(y|\theta_0)} > \frac{C_{FA} P(\theta_0)}{C_M P(\theta_1)}} \\ &=\; 1_{\frac{f_{\theta_1}(y)}{f_{\theta_0}(y)} > \frac{C_{FA}(1-\pi_1)}{C_M \pi_1}} \end{aligned}$$

  We see the threshold is available in closed form, as a function of costs and priors.

Bayesian Sequential Analysis

- In *sequential analysis* we don't have a fixed number of observations. Instead, observations come in sequence, and we'd like to decide in favor of θ_0 or θ_1 as soon as possible. For each n we perform a test: $D(y_{1:n})$. There are three outcomes of D:
  - Decide $\theta = \theta_1$
  - Decide $\theta = \theta_0$
  - Keep testing
- **NonBayesian Case** Let T be the stopping time of this test. We wish to find an optimal tradeoff between:
  - P_{FA}, the probability:: [$D(y_{1:T}) = 1$, but $\theta = \theta_0$]
  - P_M, the probability: [$D(y_{1:T}) = 0$, but $\theta = \theta_1$]
  - $E_\theta(T)$, where $\theta = \theta_0$ or θ_1
- It turns out, the optimal test again involves monitoring the likelihood ratio. This test is called *SPRT* for "Sequential Probability Ratio Test". It is more insightful to examine this test in the "log" domain. The test involves comparing the log

likelihood ratio:

$$S_n = \frac{f_{\theta_1}(y_{1:n})}{f_{\theta_0}(y_{1:n})}$$

to positive and negative thresholds $-a < 0$, $b > 0$. The first time $S_n < -a$, we stop the test and decide θ_0 The first time $S_n > b$, we stop and declare θ_1. Otherwise we keep on testing.

- There is one "catch"; in the analysis, we ignore *overshoots* concerning the threshold boundary. Hence $S_T = -a$ or b.
- **Properties of SPRT** The change (first difference) of S_n is

$$\begin{aligned} s_n &= S_n - S_{n-1} \\ &= \frac{f_{\theta_1}(y_n|y_{1:n-1})}{f_{\theta_0}(y_n|y_{1:n-1})} \end{aligned}$$

For an iid process, we drop the conditioning:

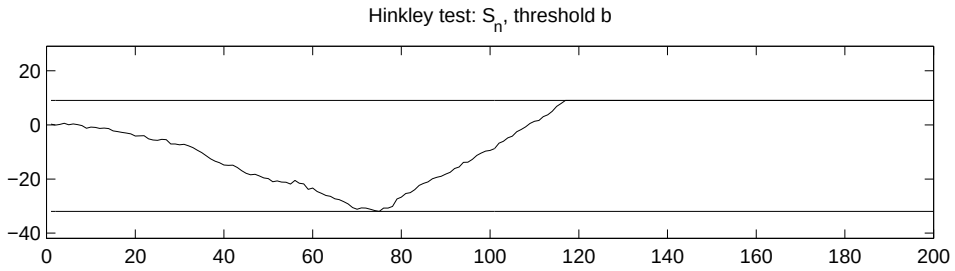
$$s_n = \frac{f_{\theta_1}(y_n)}{f_{\theta_0}(y_n)}$$

The *drift* of S_n is defined as $E(s_n|\theta)$. From definitions, it follows that the drifts under $\theta = \theta_0$ or

θ_1 are given by the K-L informations:

$$E(s_n|\theta_0) = -K(\theta_0, \theta_1)$$
$$E(s_n|\theta_1) = -K(\theta_1, \theta_0)$$

- We can visualize the behavior of S_n, when in fact θ undergoes a step transition from θ_0 to θ_1:



- Again, we have practical issues concerning how we choose thresholds a, b. By invoking Wald's equation, or some results from martingale theory, these are easily related to the probabilities of error at the stopping time of the test. However, the problem arises how to choose both probabilities of error, since we have a three-way tradeoff with the *average run lengths* $E_{\theta_0}(T)$, $E_{\theta_1}(T)$!!
- Fortunately, the Bayesian formulation comes to our rescue. We can again assign costs to the probabilities of false alarm and miss C_{FA}, C_M. We also include a cost proportional to the number of observations prior to stopping. Let this cost equal the number of observations, which is T. The goal is to minimize expected cost, or *sequential Bayes risk*. What is our prior information? Again, we must know $P(\theta = \theta_1) = \pi_1$.
- It turns out that the optimal Bayesian strategy is again a SPRT. This follows from the theory of *optimal stopping*. Suppose at time n, our we have yet to make a decision concerning θ. We must decide among the following alternatives:
  - Stop, and declare θ_0 or θ_1.
  - Take one more observation.

- We choose to stop only when the minimum additional cost of stopping is less than the minimum expected additional cost of taking one more observation.
- We compute these costs using the *posterior* distribution of θ, i.e:

$$\pi_1(n) \;=\; P(\theta = \theta_1 | y_{1:n})$$

which comes by recursively applying Bayes' rule.

$$\begin{aligned} \pi_1(n+1) \;&=\; \frac{\pi_1(n) P(y_{n+1}|\theta_1)}{(1-\pi_1(n))P(y_{n+1}|\theta_0) + \pi_1(n)P(y_{n+1}|\theta_1)} \\ \pi_1(0) \;&=\; \pi_1 \end{aligned}$$

- If we stopped after observing y_n and declared $\theta = \theta_0$, the expected cost due to "miss" would be $\pi_1(n)C_M$. Therefore if we make the decision to stop, the (minimum) additional cost is

$$\rho_0(\pi_1(n)) \;=\; \min\left\{\pi_1(n)C_M, (1-\pi_1(n))C_{FA}\right\}$$

- The overall minimum cost is:

$$\rho(\pi_1(n)) \;=\; \min\left\{\rho_0(\pi_1(n)), 1 + E_{\pi_1(n)}[\rho(\pi_1(n+1))]\right\}$$

- In the two-hypothesis case, the implied recursion for the minimum cost can be solved, and the result is a SPRT(!)

- Unfortunately, one cannot get a close form expression for the thresholds in terms of the costs, but the "Bayes" formulation allows at least to involve prior information about the hypotheses.
- We will see a much richer extension to the problem of Bayesian change detection.