9.07 Introduction to Statistics for Brain and Cognitive Sciences
Emery N. Brown
Updated 11/25/09

Lecture 10 Bayesian Methods

I. Objectives

1. Understand the concept of conditional probabilities and Bayes' Rule

2. Understand how to apply the Bayesian inference paradigm

3. Understand the relation among method-of-moments, likelihood and Bayesian methods

4. Understand how to compute Bayesian estimates by Gaussian approximation

5. Understand how to compute Bayesian estimates by Monte Carlo

6. Understand the properties of Bayesian estimates

II. Bayesian Methods

A. Conditional Probabilities and Bayes' Rule

We recall from **Lecture 1** that conditional probability allows us to assess how likely one event is given that another has happened. If A and $B \in \Omega,$ then

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}. \tag{10.1}$$

We read this as the probability of A given $B.$ Heuristically, we can think of this as coming from

$$\Pr(A \mid B) = \frac{(\text{Area of } A \cap B)}{\left(\text{Area } B\right)} = \frac{(\text{Area of } A \cap B) / \text{Area of } (\Omega)}{(\text{Area of } B) / \text{Area of } (\Omega)}$$
$$= \frac{\Pr(A \cap B)}{\Pr(B)} \tag{10.2}$$

By a similar argument we have

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)} \tag{10.3}$$

or we have

$$\Pr(A \mid B) = \frac{\Pr(A)\Pr(B \mid A)}{\Pr(B)} \tag{10.4}$$

If we write $\Pr(A)\Pr(B \mid A) = \Pr(B)\Pr(A \mid B)$ we have the **Multiplication Rule of Probability**. Given an event A and a set $B = \bigcup_{i=1}^{n} B_i = \Omega$ of disjoint events such that

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B_i) = 1 \qquad (10.5)$$

then

$$\Pr(A) = \Pr(A \cap B) = \sum_{i=1}^{n} \Pr(B_i \cap A) = \sum_{i=1}^{n} \Pr(B_i)\Pr(A \mid B_i). \qquad (10.6)$$

The above result is sometimes referred to as the **Law of Total Probability**. Now for $j = 1, \ldots, n$ we may write

$$\Pr(B_j \mid A) = \frac{\Pr(A \cap B_j)}{\Pr(A)} = \frac{\Pr(B_j)\Pr(A \mid B_j)}{\sum_{i=1}^{n} \Pr(B_i)\Pr(A \mid B_i)}. \qquad (10.7)$$

This last expression is **Bayes' Rule**. In its simplest form, it is merely a re-statement of the **Multiplication Rule of Probability.**

As a first application of Eq 10.7, let's consider the problem of decoding MI neural spiking activity.

Example 1.6. Reach Direction Given an Observed Neural Firing Pattern (Simplest Decoding Problem). Suppose that a monkey is making reaching movements with a manipulandum in 8 directions while spiking activity is being recorded from a set of single neurons in primary motor cortex MI. If A is an observed ensemble firing pattern, and B_j is the j^{th} direction, then $\Pr(B_j \mid A)$ above represents the probability that the observed firing pattern A encodes direction B_j. This is the simple model for neural spike train decoding that appeared in Sanger (1996) using a Poisson model.

Example 5.3 (continued). Dynamics of Dendritic Spines of Adult Cortical Neurons (Lee et al. 2006). To examine the extent of neuronal remodeling that occurs in the brain on a day-to-day basis, Elly Nedivi and colleagues used a multiphoton-based microscopy system for chronic in vivo imaging and reconstruction of entire neurons in the superficial layers of the rodent cerebral cortex. Over a period of months, they observed neurons extending and retracting existing branches and, in rare cases, budding new tips. 35 of 259 non-pyramidal interneuron dendritic tips showed rearrangement and 0 of 124 pyramidal cell dendritic tips showed rearrangement. We would like to construct a Bayesian analysis of these data to decide whether the results of these two groups are different.

B. Bayes' Rule for Probability Mass Functions and Probability Densities

Let $x = x_1, \ldots, x_n$ be a random sample from a probability distribution $f(x \mid \theta)$. We can make a statement of Bayes' rule for probability mass functions and probability density functions. Let $f(x \mid \theta)$ be the sampling probability density or likelihood and $f(\theta)$ be the prior probability density for the model parameter. The posterior probability density is

$$f(\theta \mid x) = \frac{f(\theta) f(x \mid \theta)}{f(x)} \qquad (10.8)$$

where $f(x) = \int f(\theta) f(x \mid \theta) d\theta$ is the normalizing constant.

Remark 10.1. Bayesian estimation combines the sample probability density $f(x \mid \theta)$ with a prior probability density on the parameter θ to compute a posterior probability density for $\theta.$ In this formulation of the estimation problem, θ is a random variable unlike in the case of maximum likelihood or methods-of-moments estimation in which θ is viewed as a fixed unknown parameter. The parameter θ has all its uncertainty characterized by its posterior density.

Remark 10.2. The Bayesian approach represents an important paradigm for combining information from different sources. In the simplest case one source is what is known about a parameter prior to an experiment and the second source is the likelihood for the experiment. Equation 10.8 therefore shows how the likelihood can be converted directly into a probability statement.

Remark 10.3. We can take a summary statistic (function) from $f(\theta \mid x)$ to be a point estimate of $\theta.$ The most typical is the **posterior mean**, which is defined as

$$E(\theta \mid x) = \int \theta f(\theta \mid x) d\theta. \tag{10.9}$$

Other estimators to be used are the **posterior median** and the **posterior mode**. We will illustrate computation of the posterior mode below. The posterior variance is defined as

$$Var(\theta \mid x) = \int (\theta - E(\theta \mid x))^2 f(\theta \mid x) d\theta \tag{10.10}$$

Remark 10.4. A **Bayesian credibility interval** (confidence interval) evaluates the probable values of the parameter relative to the posterior probability density. The parameter is a random variable and not a fixed quantity. The credible interval is usually chosen to be the smallest one with the highest posterior probability. These confidence statements have the more intuitive interpretation that one might guess a confidence statement should have.

Remark 10.5. The two basic questions in Bayesian inference are: 1) how to specify the prior probability density in a principled way for a given problem; and 2) how to compute the posterior probability density in Eq.10.8. To address the first question we will present several examples illustrating prior probability densities for commonly used **sampling densities or likelihood functions**. To address the second question we will consider four approaches: analytic methods, numerical methods, approximation methods and Monte Carlo methods.

Example 2.1 (continued). Let $x = x_1, \ldots, x_n$ be the sample of responses from the learning experiment. We assume a binomial probability mass function with parameters $n = 40$ and unknown propensity for a correct response $p.$ Assume that the prior probability density for p is the beta probability density with parameters α and $\beta.$ Find the posterior density of $p.$ Take $y = \sum_{k=1}^{n} x_k.$ What does this say about how well y summarizes the information in the sample x? Notice that y is the total number of successes in the n trials. We have that

$$\begin{aligned} f(p\,|\,x) &\propto f(p)f(x\,|\,p) \\ &\propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\binom{n}{y}p^{y}(1-p)^{n-y} \\ &\propto p^{\alpha-1}(1-p)^{\beta-1}p^{y}(1-p)^{n-y} \\ &\propto p^{\alpha+y-1}(1-p)^{n-y+\beta-1}. \end{aligned} \tag{10.11}$$

Hence by the definition of the pdf of a beta random variable, we have that

$$f(p\,|\,x) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)\Gamma(n-y+\beta)} p^{\alpha+y-1}(1-p)^{n-y+\beta-1}. \tag{10.12}$$

is a beta distribution with parameters $(\alpha+y)$ and $n-y+\beta.$ It follows from Eq. 3.24 and Eq. 3.25 that $E(p\,|\,x) = \frac{\alpha+y}{n+\alpha+\beta}$ and $Var(p\,|\,x) = \frac{(\alpha+y)(n-y+\beta)}{(n+\alpha+\beta)(n+\alpha+\beta+1)}.$ The beta distribution is a **conjugate prior distribution** for the binomial probability mass function.

A **conjugate prior probability distribution** for a parameter is a prior distribution that is proportional to the likelihood function of that parameter. This formulation makes the computation of the posterior distribution more tractable mathematically.

Example 5.3 (continued). To conduct a Bayesian analysis of the data we apply Eq. 10.8. Let us denote the binomial probability mass function for the interneurons as

$$f_i(k_i\,|\,n_i, p_i) = \binom{n_i}{k_i} p_i^{k_i}(1-p_i)^{n_i-k_i} \tag{10.13}$$

and the binomial probability mass function for the pyramidal neurons as

$$f_p(k_p\,|\,n_p, p_p) = \binom{n_p}{k_p} p_p^{k_p}(1-p_p)^{n_p-k_p}. \tag{10.14}$$

What important assumptions do these models make about the individual branchtips in both the interneuron and pyramidal neuron groups? Because we want to minimize the set of *apriori* assumptions we make about the data, let us assume that for both the interneurons and the pyramidal neurons that there is no prior knowledge about what the propensity to change is. Hence, we assume that the prior distributions for both p_i and p_p are both uniform. We recall that we can represent the uniform probability density as a beta probability distribution with $\alpha = \beta = 1$ (see Lecture 3). We write this as

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}. \tag{10.15}$$

Therefore, using Bayes' rule in Eq. 10.8 we can write for the interneurons the posterior probability density

$$f(p_i \mid k_i) = \frac{f(p_i) f(k_i \mid p_i)}{f(k_i)}. \tag{10.16}$$

where $f(k_i) = \int f(p_i) f(k_i \mid p_i) dp_i$ is the normalizing constant and for the pyramidal neurons the posterior probability density

$$f(p_p \mid k_p) = \frac{f(p_p) f(k_p \mid p_p)}{f(k_p)}. \tag{10.17}$$

Where $f(k_p) = \int f(p_p) f(k_p \mid p_p) dp_p$. Equations 10.16 and 10.17 provide explicit probability densities for the probability of change for the interneuron and pyramidal neuron branchtips. We can now make a comparison by computing the posterior probability densities for the two distributions and comparing the extent to which they overlap. We can specifically compute

$$\begin{aligned} \Pr(p_p > p_i) &= \int_{p_p > p_i} f(p_p, p_i \mid k_p, k_i) dp_p dp_i \\ &= \int_{p_p > p_i} f(p_p \mid k_p) f(p_i \mid k_i) dp_p dp_i \end{aligned} \tag{10.18}$$

This quantity can be computed explicitly numerically. We devise a simple Monte Carlo scheme for computing the probability in Eq. 10.18.

Algorithm 10.1 (Bayesian Comparison)

```
    $\text{Sum} = 0$
    For $j = 1, ..., 10,000$
1.  Draw $p_{p,j}$ from $f_p(p_p \mid k_p)$ and $p_{i,j}$ from $f_i(p_i \mid k_i)$
2.  If $p_{p,j} > p_{i,j}$  $\text{Sum} \leftarrow \text{Sum} + 1$
3.  If $j = 10,000$ then compute $\Pr(p_p > p_i) \doteq 10,000^{-1} \text{Sum}$
```

We will use Algorithm 10.1 to analyze these data in **Homework Assignments 7** and **8**.

Example 2.2 (continued). We consider $x = x_1, ..., x_n$ observations from the quantal release experiment. If we assume the x_i's a Poisson distribution with unknown parameter λ, then the joint distribution of the data is

$$f(x \mid \lambda) = \frac{\lambda^y e^{-n\lambda}}{\prod_{i=1}^{n} (x_i)!} \tag{10.19}$$

where $y = \sum_{i=1}^{n} x_i$. Assume λ has a prior probability density given by the gamma probability model

$$f(\lambda \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \tag{10.20}$$

The posterior probability density

$$f(\lambda \mid x) = \frac{f(\lambda) f(x \mid \lambda)}{f(x)} \tag{10.21}$$

$$\propto \lambda^{y} e^{-n\lambda} \lambda^{\alpha-1} e^{-\beta\lambda} \tag{10.22}$$

$$\propto \lambda^{\alpha+y-1} e^{-(n+\beta)\lambda}. \tag{10.23}$$

We recognize that the posterior density is a gamma probability density with parameter $\alpha + y$ and $n + \beta$. Could we have predicted this based on our analysis of the Poisson likelihood in **Lecture 9**? It follows that the posterior mean is $E(\lambda \mid x) = (\alpha + y)/(n + \beta)$ and the posterior variance is $Var(\lambda \mid x) = (\alpha + y)/(n + \beta)^2$.

Example 10.3. **A Gaussian random variable with a Gaussian prior probability density for the mean parameter** θ. Assume that x is a single observation from the Gaussian probability density

$$x \sim N(\theta, \sigma^2) \tag{10.24}$$

Assume further that the mean parameter θ has a prior probability density

$$\theta \sim N(\mu, \tau^2). \tag{10.25}$$

It suffices to find the posterior distribution of θ. It is straight forward, albeit tedious, to show directly using Eq. 10.8 and completing the square that this posterior distribution is Gaussian. Because of this, it suffices to report the mean and variance of the posterior distribution which are

$$E[\theta \mid x] = \frac{\tau^2}{\tau^2 + \sigma^2} x + \frac{\sigma^2}{\sigma^2 + \tau^2} \mu \tag{10.26}$$

$$Var[\theta \mid x] = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}. \tag{10.27}$$

The Gaussian distribution is a **conjugate prior distribution** for the Gaussian probability density when the unknown parameter is the mean. Notice that the posterior mean is a weighted average between the prior mean and the data. Here, we see explicitly how the prior assumptions affect the resulting estimate of the mean. How do the relative sizes of the observation and prior variances affect the estimate of the prior mean? Now let $x_t = x$, $\theta_t = E[\theta \mid x_t]$ and $\mu = \theta_{t-1}$

$$\theta_t = \theta_{t-1} + \frac{\tau^2}{\tau^2 + \sigma^2}(x_t - \theta_{t-1}), \tag{10.28}$$

we obtain the simplest version of the well-known Kalman filter. This is a point of departure for our dynamic estimation algorithms for **state-space models**. We will expand on these points a little further in the next section when we study Gaussian approximations to posterior probability densities.

Question 10.1. What would be a possible conjugate prior model for the variance of a Gaussian distribution? To answer this remember that the joint likelihood for μ and σ^2 is

$$L(\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \tag{10.29}$$

Now let $\eta = 1/\sigma^2$ and let $S_n^2 = \sum_{i=1}^{n}(x_i - \mu)^2$. We can reexpress Eq. 10.29 as

$$L(\eta) \propto \eta^{n/2} \exp\left\{-\frac{\eta}{2} S_n^2\right\} \tag{10.30}$$

and see that the likelihood expressed in terms of η is proportional to a gamma probability density (Eq. 3.23) with $\alpha = n/2 + 1$ and $\beta = \frac{S_n^2}{2}$. This suggests that a gamma or chi-squared distribution would be the choice of conjugate prior model for $1/\sigma^2$ (Why?) It is sometimes stated that the inverse chi-squared is the conjugate prior for the variance of a Gaussian distribution.

D. Approximations to Posterior Probability Densities

In high-dimensional problems, with realistic prior densities and realistic likelihoods, it is difficult to compute the posterior probability density in closed form or with exact numerical integration beyond a dimension of say 10. Therefore, it is useful to have approaches to approximate the posterior densities. Useful approximations for posterior probability densities include the Gaussian and Laplace's approximations (Tanner, 1996; Pawitan, 2001). We illustrate the Gaussian approximation.

1. Gaussian Approximation: The Essentials of Maximum-a-Posteriori Estimation

The posterior probability density is proportional to the product of a prior probability density and a likelihood function. We saw in **Lecture 9** that likelihood functions become more Gaussian in shape as the sample size gets large. Moreover, we motivated the fact that the maximum likelihood estimates have Gaussian distributions as the sample size gets large. Together these facts suggest that the Gaussian distribution should provide a reasonable approximation to posterior densities particularly those for which the likelihood is regular and the prior is not informative.

Let $f(x|\theta)$ be the sampling probability density or likelihood and $f(\theta)$ be the prior probability density. The posterior probability density is

$$\begin{aligned} f(\theta|x) &= \frac{f(\theta)f(x|\theta)}{f(x)} \\ &\propto f(\theta)f(x|\theta), \end{aligned} \tag{10.31}$$

where $f(x) = \int f(\theta)f(x|\theta)d\theta.$

Recall that the maximum likelihood (ML) estimate of θ is defined as

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} f(x|\theta). \tag{10.32}$$

Given $f(\theta)$ we define the **maximum-a-posteriori (MAP)** estimate of θ as

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} f(\theta|x). \tag{10.33}$$

Simply stated, the MAP estimate is the mode of the posterior density.

Example 10.3 (continued). Gaussian Prior and Gaussian Sampling Density. Before deriving a general Gaussian approximation to a posterior density, let's gain some intuition by considering the Gaussian sampling probability density and prior probability density defined as

$$f(x|\theta) = \left(\frac{1}{2\pi\sigma_x^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{(x-\theta)^2}{\sigma_x^2}\right\} \tag{10.34}$$

$$f(\theta|\theta_0) = \left(\frac{1}{2\pi\sigma_\theta^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{(\theta-\theta_0)^2}{\sigma_\theta^2}\right\}. \tag{10.35}$$

To find the ML estimate of θ we consider the log likelihood

$$\log f(x|\theta) = -\frac{1}{2}\log(2\pi\sigma_x^2) - \frac{1}{2}\frac{(x-\theta)^2}{\sigma_x^2}. \tag{10.36}$$

Differentiating with respect to θ gives

$$\frac{\partial \log f(x|\theta)}{\partial\theta} = \frac{(x-\theta)}{\sigma_x^2}, \tag{10.37}$$

and setting the derivative of the log likelihood equal to zero and solving for θ gives

$$\hat{\theta}_{ML} = x. \tag{10.38}$$

To compute the MAP estimate of θ we consider the log posterior probability density

$$\log f(\theta \mid x) \propto \log f(\theta) + \log f(x \mid \theta) \tag{10.39}$$

$$\propto -\frac{1}{2}\left[\frac{(\theta-\theta_0)^2}{\sigma_\theta^2} + \frac{(x-\theta)^2}{\sigma_x^2}\right]. \tag{10.40}$$

Differentiating with respect to θ gives

$$\frac{\partial \log f(\theta \mid x)}{\partial \theta} = -\frac{(\theta-\theta_0)}{\sigma_\theta^2} + \frac{(x-\theta)}{\sigma_x^2}, \tag{10.41}$$

and setting the derivative of log posterior to zero yields

$$\hat{\theta}_{MAP} = \left[\frac{1}{\sigma_\theta^2} + \frac{1}{\sigma_x^2}\right]^{-1}\left[\frac{\theta_0}{\sigma_\theta^2} + \frac{x}{\sigma_x^2}\right] = \frac{\sigma_x^2}{\sigma_\theta^2 + \sigma_x^2}\theta_0 + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_x^2}x \text{ (statistician's view)} \tag{10.42}$$

$$= \theta_0 + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_x^2}(x-\theta_0) \text{ (control engineer's view)} \tag{10.43}$$

If $\sigma_x^2 \to \infty$, then $\hat{\theta}_{MAP} \to \theta_0$ and there is no information from the likelihood. If $\sigma_\theta^2 \to \infty$, then $\hat{\theta}_{MAP} \to x = \hat{\theta}_{ML}$ and there is no information from the prior. The last expression above is the essential result for deriving the well-known Kalman filter. When the sampling probability density is a point process observation model, then the corresponding form of Eq. 10.43 may be used to derive neural spike train decoding algorithms (Brown et al. 1998).

Gaussian Approximation to a Posterior Probability Density. Let us now analyze the posterior probability density the same way we analyzed the likelihood in **Lecture 9**. Given a posterior probability density $f(\theta \mid x)$ and $\hat{\theta}_{MAP}$, a MAP estimate of θ, expand the log posterior probability density in a Taylor series about $\hat{\theta}_{MAP}$ to obtain

$$\begin{aligned}\log f(\theta \mid x) = \log f(\hat{\theta}_{MAP} \mid x) &+ \frac{\partial \log f(\hat{\theta}_{MAP} \mid x)}{\partial \theta}(\theta - \hat{\theta}_{MAP}) \\ &+ \frac{1}{2}\frac{\partial^2 \log f(\hat{\theta}_{MAP} \mid x)}{\partial \theta^2}(\theta - \hat{\theta}_{MAP})^2 + \ldots\end{aligned} \tag{10.44}$$

$$\approx \log f(\hat{\theta}_{MAP} \mid x) + \frac{1}{2}\frac{\partial^2 \log f(\hat{\theta}_{MAP} \mid x)}{\partial \theta^2}(\theta - \hat{\theta}_{MAP})^2, \tag{10.45}$$

because $\frac{\partial \log f(\hat{\theta}_{MAP} \mid x)}{\partial \theta} = 0$ by definition of the MAP estimate. Hence,

$$f(\theta \mid x) \approx f(\hat{\theta}_{MAP} \mid x)\exp\{\frac{1}{2}\frac{\partial^2 \log f(\hat{\theta}_{MAP} \mid x)}{\partial \theta^2}(\theta - \hat{\theta}_{MAP})^2\}. \tag{10.46}$$

gives a general Gaussian approximation to a posterior probability density.

Example 10.3 (continued). Notice that for all θ

$$\frac{\partial^2 \log f(\theta \mid x)}{\partial \theta^2} = -\left[\frac{1}{\sigma_\theta^2} + \frac{1}{\sigma_x^2}\right]. \tag{10.47}$$

and thus

$$f(\theta \mid x) = f(\hat{\theta}_{MAP} \mid x)\exp\{-\frac{1}{2}[\frac{1}{\sigma_\theta^2} + \frac{1}{\sigma_x^2}](\theta - \hat{\theta}_{MAP})^2\} \tag{10.48}$$

$$= \left(\frac{1}{2\pi}\left[\frac{\sigma_\theta^2 \sigma_x^2}{\sigma_\theta^2 + \sigma_x^2}\right]\right)^{\frac{1}{2}} \exp\{-\frac{1}{2}\left[\frac{\sigma_\theta^2 + \sigma_x^2}{\sigma_\theta^2 \sigma_x^2}\right]^{-1}(\theta - \hat{\theta}_{MAP})^2\}. \tag{10.49}$$

In this case the approximation is exact.

Example 2.2 (continued). Poisson probability mass function and a gamma prior probability density. We computed above the exact posterior probability density for this problem in Eq.10.23. Here we derive the Gaussian approximation. The Gaussian approximation to $f(\lambda \mid x)$ is defined by

$$f(\lambda \mid x) \doteq f(\hat{\lambda}_{MAP} \mid x)\exp\{\frac{1}{2}\frac{\partial^2 \log f(\hat{\lambda}_{MAP} \mid x)}{\partial \lambda^2}(\lambda - \hat{\lambda}_{MAP})^2\}. \tag{10.50}$$

We have the log posterior and the first derivative are

$$\log f(\lambda \mid x) \propto (\alpha + y - 1)\log \lambda - (n + \beta)\lambda \tag{10.51}$$

$$\frac{\partial \log f(\lambda \mid x)}{\partial \lambda} = \frac{\alpha + y - 1}{\lambda} - (n + \beta), \tag{10.52}$$

where $y = \sum_{i=1}^{n} x_i$. Solving for $\hat{\lambda}_{\text{MAP}}$ gives

$$\hat{\lambda}_{MAP} = \frac{\alpha + y - 1}{n + \beta}. \tag{10.53}$$

The second derivative of the log posterior is

$$\frac{\partial^2 \log f(\lambda \mid x)}{\partial \lambda^2} = -\frac{\alpha + y - 1}{\lambda^2}. \tag{10.54}$$

Evaluating this at $\hat{\lambda}_{\mathrm{MAP}}$ gives

$$\frac{\partial^2 \log f(\hat{\lambda}_{MAP} \mid x)}{\partial \lambda^2} = -\frac{(n+\beta)^2 \alpha + y - 1}{(\alpha + y - 1)^2} = -\frac{(n+\beta)^2}{(\alpha + y - 1)} = -\frac{n+\beta}{\hat{\lambda}_{MAP}}. \qquad (10.55)$$

The normalizing constant is $f(\lambda \mid x) = \frac{(n+\beta)^{\alpha+y}}{\Gamma(\alpha+y)} \lambda^{(\alpha+y-1)} e^{-(n+\beta)\lambda}$ evaluated at $\hat{\lambda}_{MAP}$ which is

$$f(\hat{\lambda}_{MAP} \mid x) \doteq \frac{(n+\beta)^{\alpha+y} (\alpha+y-1)^{\alpha+y-1} e^{-(n+\beta)\left(\frac{\alpha+y-1}{n+\beta}\right)}}{\Gamma(\alpha+y)(n+\beta)^{\alpha+y-1}}$$

$$\doteq \frac{(n+\beta)(\alpha+y-1)^{\alpha+y-1}}{\Gamma(\alpha+y)} e^{-(\alpha+y-1)} \qquad (10.56)$$

$$\approx \frac{(n+\beta)(\alpha+y-1)^{\alpha+y-1} e^{-(\alpha+y-1)}}{(2\pi)^{\frac{1}{2}} (\alpha+y-1)^{\alpha+y-1+\frac{1}{2}} e^{-(\alpha+y-1)}} = \frac{(n+\beta)}{[2\pi(\alpha+y-1)]^{\frac{1}{2}}},$$

using Stirling's formula $z! \approx (2\pi)^{\frac{1}{2}} z^{z+\frac{1}{2}} e^{-z}$ for z large and taking $\Gamma(\alpha+y) = (\alpha+y-1)!$ Hence,

$$f(\lambda \mid y) \approx \frac{n+\beta}{[2\pi(\alpha+y-1)]^{\frac{1}{2}}} \exp\left\{-\frac{1(\alpha+y-1)}{2(n+\beta)^2}[\lambda - \frac{\alpha+y-1}{(n+\beta)}]^2\right\}. \qquad (10.57)$$

The Gaussian approximation is in good agreement with the true posterior probability density provided that $\alpha + y$ is large relative to 1. Notice that the true posterior mean and variance are respectively $E(\lambda \mid x) = (\alpha+y)/(n+\beta)$ and $Var(\lambda \mid x) = (\alpha+y)/(n+\beta)^2$.

2. Monte Carlo Approximations to Posterior Densities

One of the things which has made wide use of Bayesian methods possible in the last 17 years is the increase in computational power and as a consequence, the development of Monte Carlo algorithms to compute high-dimensional posterior probability densities. In some applications such as climatology, astronomy and imaging, the dimension of the posterior probability densities can be several thousands if not millions!!! We illustrate two very basic Monte Carlo methods for computing posterior probabilities: importance sampling and the rejection method.

Importance Sampling

Posterior distributions are all of the form

$$f(\theta \mid x) \propto f(\theta) f(x \mid \theta). \qquad (10.58)$$

It may be the case that the prior and sampling distribution are easy to specify but the resulting form of the posterior does not suggest any easy analytic approach to summarizing it. We can generate a summary by Monte Carlo. Suppose we wish to compute for some function $g(\theta)$ the quantity

$$E(g(\theta)) = \int g(\theta) f(\theta \mid x) d\theta. \tag{10.59}$$

Suppose we have probability density $h(\theta)$ which has the same support as $f(\theta \mid x)$ and is easy to sample. We can then write

$$\begin{aligned} E(g(\theta)) &= \int \frac{g(\theta) f(\theta \mid x)}{h(\theta)} h(\theta) d\theta \\ &= \int g(\theta) w(\theta \mid x) h(\theta) d\theta \end{aligned} \tag{10.60}$$

where $w(\theta \mid x) = f(\theta) f(x \mid \theta) h(\theta)^{-1}$ Equation 10.60 defines the process of **importance sampling** and the distribution $h(\theta)$ is the **importance distribution**. Effectively, importance sampling converts the original problem into a problem of drawing a weighted sample from a distribution that is easy to manipulate. We can simulate Eq. 10.60 with the following algorithm.

Algorithm 10.2 (Importance Sampling)

```
    $\text{Sum} = 0$
    For $j = 1, ..., 10,000$
1.  Draw $\theta_j$ from $h(\theta)$
2.  Compute $w(\theta_j \mid x) = f(\theta_j) f(x \mid \theta_j) h(\theta_j)^{-1}$ and $g(\theta_j)$
3.  $\text{Sum} \leftarrow \text{Sum} + w(\theta_j \mid x) g(\theta_j)$
4.  Compute $E(g(\theta)) \doteq 10,000^{-1} \text{Sum}$
```

A common choice for $h(\theta)$ is a probability density which approximates $f(\theta \mid x)$ such as the Gaussian approximation to $f(\theta \mid x)$ derived above. A second choice of $h(\theta)$ is $f(\theta)$ the prior distribution. In this case, we have simply $w(\theta \mid x) = f(x \mid \theta)$. The better the extent to which $h(\theta)$ approximates $f(\theta \mid x)$, the more efficient the importance sampling, where efficiency here means the number of Monte Carlo samples required to evaluate Eq. 10.60 accurately. Ideally, we would like to have $w(\theta \mid x)$ close to 1 for as many values of θ as possible.

Rejection Method

Suppose that $f(\theta \mid x)$ can be approximated by $h(\theta)$, a probability density which has the same support or whose support contains the support of $f(\theta \mid x)$. Suppose there is a constant c such that

$$\frac{f(\theta \mid x)}{h(\theta)} \leq c \tag{10.61}$$

for all θ. We then have the following algorithm for simulating from $f(\theta \mid x)$. It is called the rejection method.

Algorithm 10.3 (Rejection Method).

Pick n large $j = 1.$

1. Draw θ_i^* from $h(\theta)$
2. Draw U_i from $U(0,1)$
3. If $U_i \le \frac{f(\theta_i^* \mid x)}{ch(\theta_i^*)}$ take $\theta_j = \theta_i^*; j \leftarrow j+1$
4. If $j = n$ stop else $i \leftarrow i+1$ Return to Step 1

This algorithm was originally due to John Von Neumann. We accept the value θ_i^* with probability $f(\theta_i^* \mid x) / ch(\theta_i^*)$ by generating a uniform $(0,1)$ random variable U and accepting if $U_i \le f(\theta_i^* \mid x) / ch(\theta_i^*)$.

When $j = n$; the quantity n / i measures the efficiency of the algorithm.

E. Summary of Properties of Bayesian Procedures.

Remark 10.6. Bayes' estimates are generally biased.

Remark 10.7. Bayes' estimates are consistent and they are hence asymptotically unbiased.

Remark 10.8. Bayes' estimates are asymptotically efficient. This is because like the maximum likelihood estimates Bayes' estimates are functions of the sufficient statistics of the sample.

Remark 10.9. Remarks 10.7 and **10.8** imply that Bayes' estimates have optimal mean squared error properties.

Remark 10.10.Given the posterior distribution of $\theta,$ the posterior distribution of some function of θ, say $h(\theta)$ can be computed directly by change-of-variable, analytically in many cases, or by Monte Carlo.

III. Summary

Bayesian methods provide a powerful framework for statistical analysis that enjoys the optimality properties of the likelihood approach. They also provide a principled approach for addressing the important problem of combining information from different sources.

Acknowledgments

I am grateful to Julie Hartin for technical assistance and to Jim Mutch for careful proofreading and comments.

Textbook References

DeGroot MH, Schervish MJ. *Probability and Statistics*, 3rd edition. Boston, MA: Addison Wesley, 2002.

Pawitan Y. *In All Likelihood: Statistical Modeling and Inference Using Likelihood.* London: Oxford, 2001.

Rice JA. *Mathematical Statistics and Data Analysis*, 3rd edition. Boston, MA, 2007.

Tanner MA. *Tools for Statistical Inference*, 3rd edition. New York, NY: Springer, 1996.

Literature References

Brown EN, Frank LM, Tang D, Quirk MC, Wilson MA. A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience* 1998, 18: 7411-7425.

Lee WCA, Huang H, Feng G, Sanes JR, Brown EN, So PT, Nedivi E. Dynamic remodeling of dendritic arbors in GABAergic interneurons of adult visual cortex. *Public Library of Science – Biology* 2005, Dec 27; 4(2): e29.

Sanger TD. Probability density estimation for the interpretation of neural population codes. *J Neurophys* 1996, 76: 2790-2793.