

Chapter 15

Bayesian Methods

We have already described several applications of Bayes' Theorem to problems in statistical inference. Shortly after introducing Bayes' Theorem in Chapter ?? we used it in Example ?? to evaluate the posterior probability of dementia following a screening test. In Section ?? we showed that Bayes classifiers are optimal, in the sense of producing the smallest possible error rate, and we described how they could be used to recover which of four possible saccade directions had been used to stimulate firing-rate responses among a set of 55 neurons. In Section ?? we said that Bayes' Theorem could be used to quantify uncertainty in parameter estimation using the posterior distribution, whose pdf has the form

$$f_{\theta|x}(\theta|x) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)d\theta}$$

where $L(\theta) \propto f_{X|\theta}(x|\theta)$ is the likelihood function and $\pi(\theta) = f_{\theta}(\theta)$ is the prior pdf. We illustrated the approach, on page ??, by applying it to Example ?? (concerning blindsight in patient P.S.), and we also discussed the highly intuitive interpretation of credible intervals, which are confidence intervals produced by Bayes' Theorem. In Section ?? we noted that large-sample confidence intervals produced by ML estimation are the same as large-sample credible intervals obtained with Bayes' Theorem. Taken together, these results were intended to show that Bayesian formulation of statistical problems can be helpful conceptually, and Bayesian methods can be useful for scien-

tific inference. In this chapter we extend the discussion by introducing a few additional Bayesian techniques.

15.1 Posterior Distributions

15.1.1 Conjugate priors are convenient.

Let us return to the Binomial setting discussed in Sections ?? and ?. There we used a uniform prior $\pi(\theta) = 1$ and obtained the posterior pdf

$$f(\theta|x) = \frac{\theta^x(1-\theta)^{n-x}}{\int \theta^x(1-\theta)^{n-x}d\theta}$$

which matched the beta pdf form

$$f(w) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} w^{\alpha-1} (1-w)^{\beta-1},$$

giving the posterior distribution $\theta|X = x \sim \text{Beta}(x+1, n-x+1)$. A similar form is obtained when we instead use a beta prior: if $\theta \sim \text{Beta}(\alpha_\pi, \beta_\pi)$ then the posterior pdf becomes

$$f(\theta|x) = \frac{\theta^x(1-\theta)^{n-x}\theta^{\alpha_\pi}(1-\theta)^{\beta_\pi}}{\int \theta^x(1-\theta)^{n-x}\theta^{\alpha_\pi}(1-\theta)^{\beta_\pi}d\theta} \quad (15.1)$$

and this may be recognized as a $\text{Beta}(\alpha_{post}, \beta_{post})$ pdf where

$$\begin{aligned} \alpha_{post} &= x + 1 + \alpha_\pi \\ \beta_{post} &= n - x + 1 + \beta_\pi. \end{aligned}$$

Thus, when a beta prior is used in conjunction with the binomial likelihood, the posterior is also a beta distribution. This is advantageous computationally because algorithms and software are readily available for evaluating beta pdfs and probabilities. In such situations, where a prior distribution leads to a posterior within the same parametric family of distributions, the prior is called *conjugate*.

Here is another example. Let $X_1, X_2, \dots, X_n$ be i.i.d. $N(\theta, \sigma^2)$ random variables, write $X = (X_1, \dots, X_n)$, take $\bar{X}$ to be the usual sample mean of the X_i variables so that $\bar{X} \sim N(\theta, \sigma^2/n)$, and assume σ is known. If we let the prior distribution be normal with $\theta \sim N(\mu_\pi, \tau^2)$ then the posterior distribution is also normal: we have $\theta|X = x \sim N(\mu_{post}, \sigma_{post}^2)$ where

$$\begin{aligned}\mu_{post} &= \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \tau^2} \mu + \frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2} \bar{x} \\ \sigma_{post}^2 &= \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1}.\end{aligned}\tag{15.2}$$

More generally, exponential families have conjugate priors. For instance, for Poisson likelihood functions gamma distributions become conjugate priors.

15.1.2 The posterior mean is often a weighted combination of the MLE and the prior mean.

In the case of the normal likelihood and conjugate prior, above, the posterior mean is a weighted combination of the MLE and the prior mean, with the weights determined by the relative precision of data and prior. As the precision in the data increases relative to the prior, i.e., as $n\tau^2/\sigma^2$ increases, more weight is placed on $\bar{x}$ and the posterior mean becomes nearly the same as $\bar{x}$. When the data are imprecise relative to the prior (so $n\tau^2/\sigma^2$ gets small), more weight is placed on the prior mean, so that the posterior mean is “pulled” away from $\bar{x}$ and toward the prior mean.

Similar statements may be made in the binomial case. Let us reparameterize the $Beta(\alpha, \beta)$ distribution by defining

$$\begin{aligned}\mu &= \frac{\alpha}{\alpha + \beta} \\ \nu &= \alpha + \beta.\end{aligned}$$

The beta distribution may then be written $Beta(\mu\nu, (1-\mu)\nu)$. We may then write the beta prior $\theta \sim Beta(\alpha_\pi, \beta_\pi)$ instead as $\theta \sim Beta(\mu_\pi\nu_\pi, (1-\mu_\pi)\nu_\pi)$.

The posterior becomes $\theta|X = x \sim \text{Beta}(\mu_{post}\nu_{post}, (1 - \mu_{post})\nu_{post})$ and we have

$$\mu_{post} = \frac{\nu_{\pi}}{n + \nu_{\pi}} \mu_{\pi} + \frac{n}{n + \nu_{\pi}} \frac{x}{n}. \quad (15.3)$$

Here, the data precision is not exactly the reciprocal of the variance but is instead represented by n and the prior precision is represented by μ_{π} . With these definitions of precision it is again true that as the precision in the data increases relative to the prior more weight is placed on the observed proportion $\frac{x}{n}$ and the posterior mean becomes nearly the same as $\frac{x}{n}$, while when the data precision gets relatively smaller the posterior mean is pulled away from $\frac{x}{n}$ toward the prior mean.

The binomial posterior mean may be interpreted as equivalent to the MLE that would be obtained from the original data x together with some *pseudo-data* represented by the prior. For example, the posterior mean based on a uniform prior (so that $\alpha = \beta = 1$) is equal to the MLE based on $x+1$ successes and $n - x + 1$ failures. That is, we imagine first supplementing the actual data with 1 success and 1 failure, and then finding the observed proportion of successes; this is the posterior mean. A similar statement remains true whenever α and β are integers. The non-integer case is sometimes interpreted by analogy. For example, if we use the conjugate prior with $\alpha = \beta = \frac{1}{2}$ the posterior mean is equal to the MLE we would get by “adding half a success and half a failure” to the data before finding the proportion of successes.

The normal case may be interpreted similarly. Let us suppose, first, that $\tau = \sigma$. Then the posterior mean is the same as the sample mean from the original n observations supplemented by 1 observation having the value μ_{π} . Then, if $\tau^2 = \sigma^2/k$, the posterior mean is the same as the sample mean from the original n observations supplemented by k observations having mean μ_{π} . When the ratio σ^2/τ^2 is not an integer the interpretation is by analogy: the prior again injects some additional information, beyond the data, represented as if based on other data having sample mean μ_{π} and variance of that mean equal to τ^2 .

15.1.3 There is no compelling choice of prior distribution.

Numerous methods have been proposed and discussed in an attempt to define “noninformative” prior distributions. While particular choices seem reasonable, there is a degree of arbitrariness in all and no consensus has emerged. See Kass and Wasserman (1996) for an extensive review.

In the case of a binomial $B(n, \theta)$ distribution, it is quite common to take the prior for θ to be uniform on $(0, 1)$, so that the prior pdf is $\pi(\theta) = 1$. This seems to capture the notion that the prior is “non-informative” about the parameter value. Working by analogy, in the case of estimating a normal mean, where the data distribution follows $N(\theta, \sigma^2)$, the prior on θ is often taken to be uniform on $(-\infty, \infty)$ with $\pi(\theta) = 1$. This, however, is not a probability density because its integral (over its domain $(-\infty, \infty)$) is not 1 but rather is infinite. Nonetheless, the posterior turns out to be a well-defined probability distribution, so inferences may be made. Such formal priors that are not actually probability distributions are called *improper*.

15.2 Hierarchical Models

The term hierarchical model refers to a model in which not only is the family of data densities $\{p(x | \theta) : \theta \in \Theta\}$ indexed by a parameter θ , but θ itself is assumed to be distributed according to some parametric family of densities $\{p(\theta | \lambda) : \lambda \in \Lambda\}$. This process can continue, with λ distributed according to a family of densities, and so on, but in practice models with unknown parameters for a distribution of λ are rare. Thus, we will concentrate on two-stage hierarchical models, in which X , θ , and λ are vectors with θ and λ usually unknown.

The most common applications of hierarchical models are those in which there is an obvious source of variability among values of the parameter θ , as when θ could vary from subject-to-subject, or neuron-to-neuron, etc. For generality, we will refer to individual subjects or individual neurons, etc., as *units*. In other words, we will say that we are interested in the variation of

some quantity across units. In neuroimaging, for example, we might have task-related effects at particular voxels whose magnitude varies across subjects, and these could be assumed to follow some probability distribution. In analyzing neural responses, the way a particular measure of neural activity varies across neurons may be of interest, and might be assumed to follow a given probability distribution. Example ?? on page ?? provides an example. As described there, Behseta *et al.* considered spike counts from 54 neurons during performance of a serial-order eye-movement task, and the authors computed a rank order selectivity index

$$I_{\text{rank}} = \frac{(f_3 - f_1)}{(f_3 + f_1)}$$

where f_1 and f_3 were the mean firing rates measured at the times of the first and third saccades respectively, the mean being taken across trials. As part of the analysis, the rank selectivity indices across neurons were considered to follow a normal distribution.

To treat these situations we introduce a random vector X_i to represent measurements made on unit i . For instance, X_i could be the rank order selectivity index for neuron i . We then assume the observations X_i (and the parameters θ_i) are *conditionally independent* across units, with variation being described by a two-stage hierarchical model:

Stage one: Conditionally on $(\theta_1, \dots, \theta_k)$ and λ , the vectors X_i are independent with densities $p(x_i | \theta_i, \lambda)$, $i = 1, \dots, k$, belonging to a family $\{p(x | \theta, \lambda) : \theta \in \Theta, \lambda \in \Lambda\}$;

Stage two: Conditionally on λ , the vectors θ_i are i.i.d. with density belonging to a family $\{p(\theta | \lambda) : \lambda \in \Lambda\}$.

In general, θ and λ are multidimensional. In the case of the rank order selectivity index, X_i is a random variable representing I_{rank} for neuron i and Behseta *et al.* assumed a model of the form

$$\begin{aligned} X_i &\sim N(\theta_i, \sigma_i^2) \\ \theta_i &\sim N(\mu, \tau^2). \end{aligned}$$

Here, θ_i is the theoretical mean of the the rank order selectivity index for neuron i and σ_i^2 is its variance. The value of θ_i becomes a quantity to be estimated, but is here considered to follow a distribution across the population of neurons, with population mean μ and variance τ^2 . Actually, Behseta *et al.* also considered a second index, but we are ignoring that for the time being.

It would be possible to estimate θ_i as x_i , but the model suggests something different: it assumes that the values of θ_i are related to each other (according to the second stage of the model) and the posterior therefore uses data *from the other neurons* in estimating θ_i . This would be especially valuable if σ_i^2 happened to be large, possibly due to a very small number of trials for that neuron.

Let us assume $X_i \sim N(\theta_i, \sigma_i^2)$, independently and $\theta_i \sim N(\mu, \tau^2)$ i.i.d. for $i = 1, \dots, k$ with the σ_i 's and τ known but μ unknown, and let μ have the improper uniform prior. (In most practical cases, including that in Behseta *et al.*, τ is unknown and must be estimated, but we are ignoring that complication here.) Calculations show that this results in normal posteriors for θ_i , $i = 1, \dots, k$ given $x = (x_1, \dots, x_k)$ with

$$E(\theta_i | x) = \frac{\sigma_i^2}{\sigma_i^2 + \tau^2} \bar{x}_\alpha + \frac{\tau^2}{\sigma_i^2 + \tau^2} x_i$$

$$V(\theta_i | x) = \left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2} \right)^{-1} + \left(\sum_i \frac{1}{\sigma_i^2 + \tau^2} \right)^{-1} \left(\frac{\sigma_i^2}{\sigma_i^2 + \tau^2} \right)^2$$

where $\bar{x}_\alpha = (\sum_i \alpha_i x_i) / (\sum_i \alpha_i)$ and $\alpha_i = (\sigma_i^2 + \tau^2)^{-1}$.

The expression for the posterior mean is beautifully simple. In the case of rank order selectivity index, each value of σ_i could be estimated directly from the data and was therefore taken to be known. But these could vary across neurons. Some neurons could have highly variable X_i , and thus poorly-determined values of θ_i , while other neurons could have less variable X_i and better-determined values of θ_i . To estimate the θ_i 's it would make sense to use x_i if σ_i were small. But if we assume it is appropriate to take $\theta_i \sim N(\mu, \tau^2)$, then what happens with one neuron provides at least some information about what is happening with another. The posterior mean incorporates this information in a simple way: when σ_i is very small, the posterior mean is roughly

equal to that neuron's measured value x_i , whereas as σ_i gets larger, the value $\bar{x}_\alpha$ plays a role according to the combination $w_i\bar{x}_\alpha + (1 - w_i)x_i$ where $w_i = \sigma_i^2(\sigma_i^2 + \tau^2)^{-1}$. A common way to describe this is that the estimate x_i is *shrunk* toward $\bar{x}_\alpha$ with the amount of *shrinkage* determined by w_i .

A NON-NEURAL EXAMPLE. In a microbiology experiment, 13 strains of E. coli were tested for association of two traits. The raw data for each strain were two pairs of sample sizes and corresponding proportions $(n_{i1}, \hat{p}_{i1})$ and $(n_{i2}, \hat{p}_{i2})$. From preliminary analysis, it appeared that p_{i1} was greater than p_{i2} in most strains. In two strains, however, $\hat{p}_{i1}$ was less than $\hat{p}_{i2}$ and the issue was whether this was due to sampling fluctuation or a genuinely different phenomenon for either or both of the two strains in question. We assume here that the data are distributed as binomial proportions, we transform to the logit scale according to $X_i = \log[\hat{p}_{i1}(1 - \hat{p}_{i2})/(\hat{p}_{i2}(1 - \hat{p}_{i1}))]$, and we take σ_i^2 to be known and equal to the large-sample variance formula (based on binomial sampling) $\sigma_i^2 = (n_{i1}\hat{p}_{i1})^{-1} + (n_{i1}(1 - \hat{p}_{i1}))^{-1} + (n_{i2}\hat{p}_{i2})^{-1} + (n_{i2}(1 - \hat{p}_{i2}))^{-1}$.

The transformed data are shown below. Although τ^2 is not known in this case, it is assumed known in the formula for $E(\theta_i | x)$. We used $\tau^2 = .39$ to obtain the tabled values. (The reason for this choice is that it is the MLE of τ^2 .) The weighted mean is $\bar{x}_\alpha = 1.30$. Note that the “shrinkage” behavior is as described in the previous paragraph. $\square$

Strain	x_i	σ_i	$E(\theta_i y)$
1	1.36	.28	1.35
2	2.26	1.04	1.56
3	2.23	.75	1.68
4	1.32	.36	1.31
5	1.21	.38	1.24
6	1.27	.49	1.28
7	1.43	.57	1.37
8	1.85	.54	1.62
9	1.34	.56	1.30
10	3.44	.73	2.20
11	-0.42	.69	.53
12	-0.10	.31	.17
13	1.25	.39	1.27

When parameters such as σ_i and τ are unknown, the Bayesian approach is to put priors on them and then to proceed as before with the calculation of the relevant marginal posteriors, such as that of θ_i given y . This requires possibly intractable integrations over these other parameters. Thus, a major part of practical Bayesian analysis of these problems (and others, too) involves evaluation of integrals. Sometimes the integrals can be evaluated analytically but often they must be evaluated by some numerical approximation or, most commonly, by Monte Carlo simulation.

An alternative is to estimate unknown parameters, such as τ by ML or some variant of it. This is often called *empirical Bayes*. Once we have an estimate $\hat{\tau}$, it may be inserted in place of τ in the expression of the posterior mean. The resulting quantity

$$\tilde{\theta}_i = \hat{w}_i \bar{x}_\alpha + (1 - \hat{w}_i) x_i$$

where

$$\hat{w}_i = \sigma_i^2 / (\sigma_i^2 + \hat{\tau}^2)$$

is usually called an *empirical Bayes estimator*. Similarly, σ_i could be estimated, if it were unknown, and then the estimate could again replace σ_i in the formula.

Details: Derivation of basic result

Suppose $X_i \sim N(\theta_i, \sigma_i^2)$, independently, $i = 1, \dots, k$, $\theta_i \sim N(\mu, \tau^2)$, i.i.d., $i = 1, \dots, k$, σ_i and τ are known, and we put an improper uniform prior on μ . Then

$$\begin{aligned} p(\theta \mid x) &\propto \int L(\theta) p(\theta \mid \mu, \tau^2) d\mu \\ &= L(\theta) \int p(\theta \mid \mu, \tau^2) d\mu \\ &\propto \exp\{-(1/2) \Sigma \sigma_i^{-2} (x_i - \theta_i)^2\} \int \exp\{-(1/2) \tau^{-2} \Sigma (\theta_i - \mu)^2\} d\mu. \end{aligned}$$

Writing

$$\Sigma (\theta_i - \mu)^2 = k(\mu^2 - 2\mu\bar{\theta} + \bar{\theta}^2) + \Sigma \theta_i^2 - k\bar{\theta}^2$$

where $\bar{\theta} = k^{-1} \Sigma \theta_i$, we get

$$\int \exp\{-(1/2) \tau^{-2} \Sigma (\theta_i - \mu)^2\} d\mu \propto \exp\{-(1/2) \tau^{-2} \Sigma (\theta_i - \bar{\theta})^2\}$$

so that

$$p(\theta | x) \propto \exp\{-(1/2)[\sum \sigma_i^{-2}(x_i - \theta_i)^2 + \tau^{-2}\sum(\theta_i - \bar{\theta})^2]\}.$$

Expanding the sums of squares and collecting terms we find

$$\begin{aligned} & \sum \sigma_i^{-2}(x_i - \theta_i)^2 + \tau^{-2}\sum(\theta_i - \bar{\theta})^2 \\ = & \sum(\sigma_i^{-2} + \tau^{-2})\theta_i^2 - 2\sum(\sigma_i^{-2}x_i + \tau^{-2}\bar{\theta})\theta_i + \tau^{-2}k\bar{\theta}^2 + \sum \sigma_i^{-2}x_i^2 \\ = & \sum(\sigma_i^{-2} + \tau^2 + k^{-1}\tau^{-2})\theta_i^2 - k^{-1}\tau^{-2}\sum\sum\theta_i\theta_j - 2\sum\sigma_i^{-2}x_i\theta_i + \text{constant} \end{aligned}$$

which is quadratic in θ . In general, for a matrix V and vector z we have

$$\theta^T V^{-1} \theta - 2z^T \theta = (\theta - m)^T V^{-1} (\theta - m) - z^T m$$

where $m = Vz$. Writing $v^{ij} = (V^{-1})_{ij}$ and defining V^{-1} and z according to

$$v^{ij} = -k^{-1}\tau^{-2} + (\sigma_i^{-2} + \tau^{-2})\delta_{ij}$$

$$z_i = \sigma_i^{-2}x_i$$

where δ_{ij} is 1 if $i = j$ and 0 otherwise, we thus get

$$p(\theta | x) \propto \exp[-(1/2)(\theta - m)^T V^{-1} (\theta - m)]$$

where $m = Vz$ and V and z are defined by the components above. Thus, the posterior distribution of θ is multivariate normal with expectation vector m and covariance matrix V . All that remains is to write down the explicit formulae for m and V .

For this we use the following matrix identity: writing $c^{ij} = (C^{-1})_{ij}$, if

$$c_{ij} = -k^{-1}b + (a_i + b)\delta_{ij}$$

then

$$c^{ij} = \left(\sum \frac{a_i b}{a_i + b}\right)^{-1} \left(\frac{b}{a_i + b}\right) \left(\frac{b}{a_j + b}\right) + (a_i + b)^{-1}\delta_{ij}.$$

(This may be verified by direct calculation.) The matrix V becomes specified by

$$v_{ij} = (\sum(\sigma_i^2 + \tau^2)^{-1})^{-1} \left(\frac{\sigma_i^2}{\sigma_i^2 + \tau^2}\right) \left(\frac{\sigma_j^2}{\sigma_j^2 + \tau^2}\right) + \left(\frac{\sigma_i^2 \tau^2}{\sigma_i^2 + \tau^2}\right) \delta_{ij}$$

and the vector m then becomes

$$\begin{aligned} m_i &= \sum v_{ij} \sigma_j^{-2} x_j \\ &= w_i \bar{x}_\alpha + (1 - w_i) x_i \end{aligned}$$

where

$$\begin{aligned} w_i &= \sigma_i^2 / (\sigma_i^2 + \tau^2) \\ \alpha_i &= (\sigma_i^2 + \tau^2)^{-1} \\ \bar{x}_\alpha &= (\sum \alpha_i x_i) / (\sum \alpha_i). \end{aligned}$$