

L2.1: SOME BAYESIAN EXAMPLES

1. BAYESIAN CLASSICS: CONJUGATE PRIORS

Bernoulli (Binomial) / Beta Prior Model:

$$p \sim \text{Beta}(\alpha, \beta)$$

$$X_1, \dots, X_n | p = p \stackrel{iid}{\sim} \text{Bernoulli}(p)$$

Posterior:

$$p | X_1 = x_1, \dots, X_n = x_n \sim \text{Beta}(\alpha + \sum_{i=1}^n x_i, \beta + \sum_{i=1}^n (1 - x_i))$$

Generalization: Dirichlet / Multinomial

Normal / Normal location model (σ known):

$$\theta \sim \text{Normal}(\mu, \tau^2)$$

$$X_1, \dots, X_n | \theta = \theta \stackrel{iid}{\sim} \text{Normal}(\theta, \sigma^2)$$

Posterior: (page 12 in G&H)

$$\theta | X_1 = x_1, \dots, X_n = x_n \sim \text{Normal}(\mu_n, \tau_n^2)$$

Where the posterior variance is $\tau_n^2 = [(\tau^2)^{-1} + (\frac{\sigma^2}{n})^{-1}]^{-1}$. Notice that this means that the posterior precision (inverse posterior variance) is the sum of the prior precision and the “data precision.” The posterior mean μ_n is a convex combination of the prior mean μ and the data mean \bar{x} . We compute $\mu_n = (\frac{\mu}{\tau^2} + \frac{n\bar{x}}{\sigma^2})\tau_n^2$ and reexpress this as:

$$\mu_n = \frac{(\frac{\sigma^2}{n})^{-1}}{(\tau^2)^{-1} + (\frac{\sigma^2}{n})^{-1}} \bar{x} + \frac{(\tau^2)^{-1}}{(\tau^2)^{-1} + (\frac{\sigma^2}{n})^{-1}} \mu$$

Notice that the denominators are the same, here, and that the multipliers sum to 1. We interpret this as the posterior mean is pulled toward the data mean and the prior mean respectively by linear factors proportional to the data-precision and prior precision.

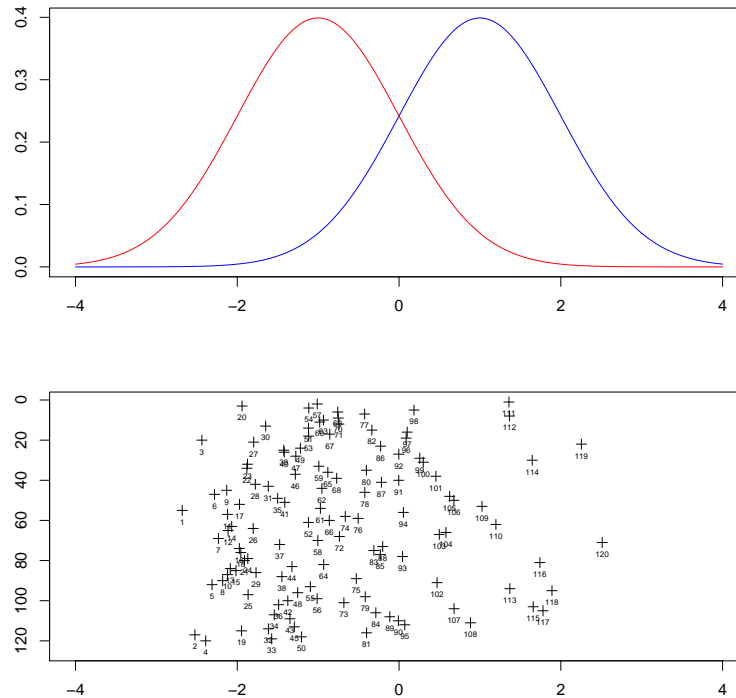
Generalization: Multivariate Normal (widely used in e.g. spatial statistics, smoothing splines)

We will see other classic cases. E.g. a Normal with a Gamma prior on the precision and a Poisson/Gamma pairing (on HW 1).

All of these are examples of *conjugate families*. When the prior is of the conjugate form to the likelihood, the posterior is of the same form and, furthermore, the updates to the parameters are easy [properly parametrized they become linear updating rules]. Conjugate priors are, therefore, mathematically convenient and are ubiquitous in Gibbs sampling and hierarchical-Bayesian work. They are not always, however, the right prior to use. From the subjective viewpoint, your prior should express your (personal) uncertainty about the value of the parameter. Only occasionally can this be well approximated by a conjugate form. Sometimes a carefully chosen mixture of conjugate priors will make a suitable approximation but this may not be obvious.

2. COMPOUND DECISION MAKING EXAMPLE

Suppose θ is a vector of length n with each $\theta_i \in \{-1, 1\}$. Suppose you get a dollar for every coordinate of θ that you estimate correctly. For data, suppose we observe $X_i \sim \text{Normal}(\theta_i, 1)$ for each i independently.



In the figure, I've displayed the two possible sampling densities from which each X_i is drawn and I've show a data set; each point in the scatterplot is of the form (x_i, i) . I labeled the points by their order statistics so that the smallest is 1 up to the largest of 120.

A reasonable approach is to use the maximum likelihood estimate of θ . This simplifies down to be: choose $\hat{\theta}_i = 1$ if $x_i \geq 0$ and -1 otherwise. This is because 0 is where the two normal densities cross in the figure.

But there's money at stake here, can we do better? If we looked at the data and saw a clear pattern, like if the first 50 x_i 's were roughly centered around 1 and the remainder centered around -1, then it wouldn't be smart to just use the MLE. I don't see that here, but there is rather more mass near -1 than near 1 overall. This suggests that there are more θ_i 's equal to -1 than 1. Can we estimate that, and show should that affect our decision?

What would a Bayesian do? Let's start in the simple case that there's only one observation. Suppose he believes that there is a prior probability of p that $\theta = 1$. More formally, his joint density is the product of these two terms:

$$(1) \quad f_{\Theta}(\theta) = p\mathbf{1}(\theta = 1) + (1 - p)\mathbf{1}(\theta = -1)$$

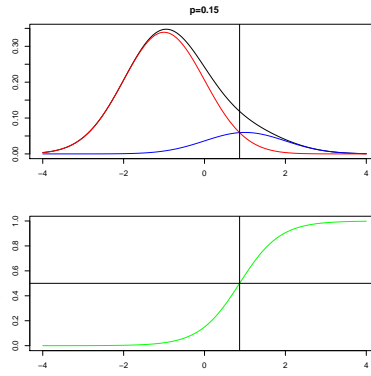
$$(2) \quad f_{X|\Theta}(x|\theta) = \phi(x - \theta)$$

where $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$ is the standard normal density. So that by our usual calculation, the conditional density of θ given $X = x$:

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)}{\int f_{\Theta}(\tilde{\theta})f_{X|\Theta}(x|\tilde{\theta})d\tilde{\theta}}$$

Converting the "integral" into the sum over the two actual possibilities for $\tilde{\theta}$ and evaluating at $\theta = 1$ we get that the posterior probability that $\theta = 1$ given that $X = x$:

$$\frac{p\phi(x - 1)}{p\phi(x - 1) + (1 - p)\phi(x - (-1))}$$



The figure depicts the case that $p = 0.15$ and shows the mixture of two normals that he *believes* will produce the data; the lower figure shows the conditional probability of $\theta = 1$ given $X = x$. As we can see, this Bayesian would switch his guess near 1 instead of near 0.

We don't have any special prior belief about θ , certainly we don't know a good value for p a priori, but looking at the data we have made the guess that $p < 0.5$. Let's be more systematic. Let's invent a new Bayesian whose beliefs are still simple, but closer to ours. This Bayesian doesn't know p , but has a uniform prior on p . Then, conditional on p he believes that the θ_i 's result from flipping independent coins with probability p . His joint probability density on everything is the product of the following terms with n copies of the latter two densities, one for each value of i :

$$(3) \quad f_P(p) = \mathbf{1}(0 \leq p \leq 1)$$

$$(4) \quad f_{\Theta_i|P}(\theta_i|p) = p\mathbf{1}(\theta_i = 1) + (1-p)\mathbf{1}(\theta_i = -1)$$

$$(5) \quad f_{X_i|\Theta_i}(x_i|\theta_i) = \phi(x_i - \theta_i)$$

We understand here that the Θ_i 's are discrete random variables and the others are continuous.

The point of considering this Bayesian is that he can infer the value of p and will act accordingly. Our main interest at the moment, is figuring out what his posterior belief about p is. We want to work out the posterior density on p given $\mathbf{X} = \mathbf{x}$. (I use bold to denote the vector of all of the x 's.)

$$f_{P|\mathbf{X}}(p|\mathbf{x}) = \frac{f_{P,\mathbf{X}}(p, \mathbf{x})}{\int f_{P,\mathbf{X}}(\tilde{p}, \mathbf{x})d\tilde{p}}$$

Where:

$$(6) \quad f_{P,\mathbf{X}}(p, \mathbf{x}) = \int f_{P,\Theta,\mathbf{X}}(p, \theta, \mathbf{x})d\theta$$

$$(7) \quad = \int f_P(p) \prod_{i=1}^n \{f_{\Theta_i|P}(\theta_i|p)f_{X_i|\Theta_i}(x_i|\theta_i)\}d\theta$$

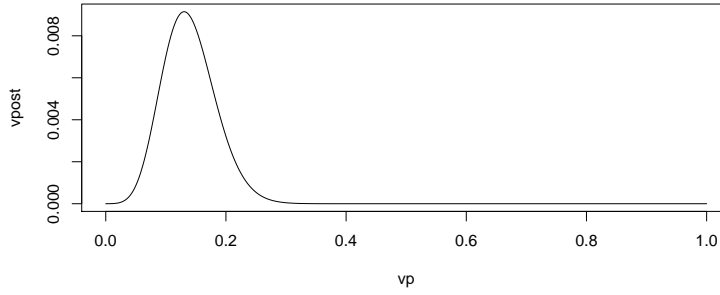
$$(8) \quad = f_P(p) \prod_{i=1}^n (p\phi(x_i - 1) + (1-p)\phi(x_i + 1))$$

This latter expression (aside from the prior density term on p) can be interpreted as the likelihood function for p . It's just the product of all of the mixed normal densities. Since his prior is flat, the posterior is just be the likelihood renormalized to have area 1. We can compute these on our data, just by evaluating it on a fine grid.

Finally, how should we decide on each θ_i ? We should decide $\theta_i = 1$ if the posterior probability of that is larger than $\frac{1}{2}$. The easiest way to compute this is to observe that it is just the average with respect to the posterior on p of the conditional density given p . We have already worked out the formulas for these parts. [* actually I might be slightly off here, but it simplifies things and I'll correct the

error next time if necessary]

$$(9) \quad f_{\Theta_i|\mathbf{X}}(\theta_i|\mathbf{x}) = \int f_{\Theta_i|P, X_i}(\theta_i|p, x_i) f_{P|\mathbf{X}}(p|\mathbf{x}) dp$$



Best cut is 0.935727933322809 resulting in a score of 111

