7.3 Conjugate Priors

Previously, we observed that using a Beta$(3,1)$ prior with Geometric sampling produced a posterior distribution that was also Beta distributed. This was not an accident, and in fact, if we had chosen a different Beta prior, we also would have ended up with a beta posterior.

Suppose instead of performing a single experiment where $X \sim \text{Geom}(\theta)$, that we performed a different experiment, sampling n times where each $X_1, \ldots, X_n \sim \text{Bern}(\theta)$, conditionally independent given θ. Note that one way to write the PMF of X_i is

$$f(x_i) = \theta^{x_i}(1-\theta)^{1-x_i} \qquad \text{for } x_i \in \{0,1\}$$

Then the conditional distribution for $\mathbf{X}|\theta$ is

$$f(\mathbf{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta) = \theta^{x_1+\cdots+x_n}(1-\theta)^{n-(x_1+x_2+\cdots+x_n)}$$

Letting $x = x_1 + \cdots + x_n$, then the posterior distribution of $\theta|\mathbf{X} = \mathbf{x}$ is

$$\xi(\theta \,|\, \mathbf{x}) \propto \theta^x(1-\theta)^n\xi(\theta)$$

So if the prior is of the form

$$\xi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

then the posterior will be of the form

$$\xi(\theta \,|\, \mathbf{x}) \propto \theta^{x+a-1}(1-\theta)^{n-x+b-1}$$

and so $\theta|X = x \sim \text{Beta}(x+a, n-x+b)$.

Def: Let $X_1, X_2, \ldots$ be conditionally iid given θ, with common distribution $f(x|\theta)$ and let Ψ be a family of distributions indexed by parameter space Ω. Suppose that, for any prior distribution $\xi \in \Psi$, for any number of observations $\mathbf{X} = (X_1, \ldots, X_n)$, and for any observed values of these observations $\mathbf{x} = (x_1, \ldots, x_n)$, the posterior distribution $\xi(\theta|\mathbf{x})$ is also a member of Ψ. Then Ψ is called a **conjugate family of prior distributions** for samples from the distribution $f(x|\theta)$.

Thm: The family of beta distributions is a conjugate family of prior distributions for samples from a Geometric distribution.

Before we present the next example, we should discuss an algebraic trick that will be useful in that example, as well as many more to come.

Thm: Let $x_1, \ldots, x_n$ be a list of numbers with mean $\bar{x}$ and let a be an arbitrary number. Then

$$\sum_{i=1}^{n}(x_i - a)^2 = n(\bar{x} - a) + \sum_{i=1}^{n}(x_i - \bar{x})^2$$

That is, the sum of squared distances between the x_i and a fixed number a can be calculated as a sum of squared distances between the x_i and their mean, and and the distance from the mean to the fixed number a.

Note: It is possible to restate this equation using the notion of distances in vector spaces!

Proof. We proceed by creatively adding 0 to an expression, and then expanding the square:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-a)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-a)^2 \\
&= \sum_{i=1}^{n}\Big((x_i-\bar{x})+(\bar{x}-a)\Big)^2 \\
&= \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-a) + \sum_{i=1}^{n}(\bar{x}-a)^2 \\
&= \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2(\bar{x}-a)\sum_{i=1}^{n}(x_i-\bar{x}) + n(\bar{x}-a)^2 \\
&= n(\bar{x}-a)^2 + \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2(\bar{x}-a)\left(\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x}\right) \\
&= n(\bar{x}-a)^2 + \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2(\bar{x}-a)\left(\sum_{i=1}^{n}x_i - n\bar{x}\right) \\
&= n(\bar{x}-a)^2 + \sum_{i=1}^{n}(x_i-\bar{x})^2 + 0
\end{aligned}$$

□

Ex 1: Suppose we want to predict the high temperature in Grinnell on the 1st of February. We have n historical recorded temperatures for this day, which we'll treat as a random sample. Based on natural weather trends and domain knowledge, we have good reason to believe that temperature values on a given day are Normally distributed with some mean θ and variance σ^2.

What is parameter space Ω?

Suppose we know (based on national weather data) that the variance is $\sigma^2 = 9$, so only the mean θ is unknown.

What are some priors we could put on θ?

One possible prior is $N(\mu, \nu^2)$. Suppose in particular, we take $\mu = 30$ and $\nu^2 = 4$, obtained by estimating θ to be 30, but with some uncertainty in our belief.

Note the difference between the parameter(s) we are building priors for (θ, in this case), and the parameters of the prior distribution, called hyperparameters (μ and ν^2, in this case).

Returning to the general case with prior $N(\mu, \nu^2)$, let's compute the posterior distribution. Let $\mathbf{x} = (x_1, \ldots, x_n)$ denote the n observations. Then

$$\begin{aligned}
f(\mathbf{x}|\theta) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left(n(\theta-\bar{x})^2 + \sum_{i=1}^{n}(x_i-\bar{x})^2\right)\right\} \\
&\propto \exp\left\{-\frac{n}{2\sigma^2}(\theta-\bar{x})^2\right\}
\end{aligned}$$

By assumption, we have a Normal prior:

$$\xi(\theta) \propto \exp\left\{-\frac{1}{2\nu^2}(\theta-\mu)^2\right\}$$

Therefore, the posterior distribution is

$$\begin{aligned}\xi(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta)\xi(\theta)\\ &\propto \exp\left\{-\frac{n}{2\sigma^2}(\theta-\bar{x})^2-\frac{1}{2\nu^2}(\theta-\mu)^2\right\}\end{aligned}$$

$$\text{note:}\quad \frac{n}{\sigma^2}(\theta-\bar{x})^2+\frac{1}{\nu^2}(\theta-\mu)^2=\frac{1}{\nu_1^2}(\theta-\mu_1)^2+\frac{n}{\sigma^2+n\nu^2}(\bar{x}-\mu)^2$$

$$\text{where}\quad \mu_1=\frac{\sigma^2\mu+n\nu^2\bar{x}}{\sigma^2+n\nu^2}\quad \nu_1^2=\frac{\sigma^2\nu^2}{\sigma^2+n\nu^2}$$

$$\xi(\theta|\mathbf{x}) \propto \exp\left\{-\frac{1}{2\nu_1^2}(\theta-\mu_1)^2\right\}$$

$$\theta|\mathbf{x} \sim N(\mu_1,\nu_1^2)$$

Ex 2: Suppose in the previous example we observe that $\bar{x}=25$ with $n=16$. What is our posterior distribution?

$$\mu_1=\frac{9\cdot 30+16\cdot 4\cdot 25}{9+16\cdot 4}\approx 25.62 \qquad \nu_1^2=\frac{9\cdot 4}{9+16\cdot 4}\approx 0.49\approx (0.7)^2$$

What effect do ν^2, σ^2, n and $\bar{x}$ have on the posterior?

1. μ_1 is the weighted average of the prior mean and the sample mean.
2. The only way the data enters is the calculation is via the sample mean $\bar{x}$.
3. For fixed ν_0^2 and σ^2, larger sample sizes will weight $\bar{x}$ more heavily.
4. For fixed values of ν_0^2 and n, the larger the variance of each observation σ^2, the smaller the relative weight given to $\bar{x}$.
5. For fixed σ^2 and n, the larger the variance ν_0^2 of the prior distribution, the larger the relative weight given to $\bar{x}$.
6. The variance of the posterior doesn't depend on the observed values; only on their number.
7. What happens if we take $\nu^2\to\infty$?

   $$\mu_1=\frac{\sigma^2}{\sigma^2+n\nu^2}\mu+\frac{n\nu^2}{\sigma^2+n\nu^2}\bar{x} \qquad \nu_1^2=\frac{\sigma^2}{n}\frac{\nu^2}{\frac{\sigma^2}{n}+\nu^2}$$

   So

   $$\mu_1\to\bar{x} \qquad \nu_1^2\to\frac{\sigma^2}{n}$$

   That is, with $\nu^2=\infty$, our posterior is determined completely by the sample mean and sample variance. This is an example of an **improper prior**—a prior "distribution" which is not a valid probability distribution, but which nonetheless gives a valid posterior distribution.

Additional Conjugate Priors

Thm: The family of Beta distributions is a conjugate family of prior distributions for samples from a Bernoulli-θ distribution.

Thm: The family of Gamma distributions is a conjugate family of prior distributions for samples from a Poisson-θ distribution.

Thm: The family of Beta distributions is a conjugate family of prior distributions for samples from Geometric-θ distribution.