Week 2 Derivation of Conjugacy (Optional)

Duke University

In this file, we will explore further why the conjugate families we mentioned in the course work. And we will also discuss briefly the updating the prior distribution to the posterior distribution of the parameter affects the effective sample size.

To derive conjugacy, we will apply the Bayes' Rule

$$X \text{ discrete:} \qquad \pi^*(\theta \mid X) = \frac{\mathbb{P}(X \mid \theta) \times \pi(\theta)}{\int \mathbb{P}(X \mid \theta) \times \pi(\theta)\, d\theta} \propto \mathbb{P}(X \mid \theta) \times \pi(\theta)$$

$$X \text{ continuous:} \qquad \pi^*(\theta \mid X) = \frac{f(X \mid \theta) \times \pi(\theta)}{\int f(X \mid \theta) \times \pi(\theta)\, d\theta} \propto f(X \mid \theta) \times \pi(\theta)$$

1 Beta-Binomial Conjugacy

1.1 Definitions

- Beta distribution: $\pi(p) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$. Here α and β are the hyperparameters.
- Binomial distribution: $\mathbb{P}(X = k \mid p) = \binom{n}{k} p^k (1-p)^{n-k}$

Using Bayes' Rule, we get

$$\pi^*(p \mid X = k) = \frac{\left[\binom{n}{k} p^k (1-p)^{n-k}\right] \times \left[\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right]}{\displaystyle\int_0^1 \left[\binom{n}{k} p^k (1-p)^{n-k}\right] \times \left[\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right] dp}$$

The right-hand side is extremely complicated, especially the calculation of the denominator. Are we going to actually calculate this?

Of course the answer is "No". What we will use instead is the "proportional to" part of Bayes' Rule. Observing that the denominator of the right-hand side, $\binom{n}{k}$, and $\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ are just some numbers which will scale $\pi^*(p \mid X = k)$ well so that the total area under the curve of $\pi^*(p \mid X = k)$ is always 1, we can first ignore these constants and see what form of $\pi^*(p \mid X = k)$ we will get.

Note: As a probability distribution, one property is to guarantee that the total area under the curve is always 1. This corresponds to the fact that probability will never exceed 1.

Using "proportional to", we rewrite the calculation into

$$\pi^*(p \mid X = k) \propto \left[p^k (1-p)^{n-k}\right] \times \left[p^{\alpha-1}(1-p)^{\beta-1}\right] = p^{(\alpha+k)-1}(1-p)^{(\beta+n-k)-1}$$

Since $\pi^*(p \mid X = k)$ is still a continuous function, the **only** continuous probability distribution that follows the form of $p^{\text{some power}-1}(1-p)^{\text{another power}-1}$ is the Beta distribution. So we argue that $\pi^*(p \mid X = k)$ has to be another Beta distribution.

Mirroring the form of Beta distribution, we can see that the new hyperparameters for $\pi^*(p \mid X)$ is

$$\alpha^* = \alpha + k, \qquad \beta^* = \beta + n - k$$

And the constants that we have ignored from our calculation must give us

$$\frac{\binom{n}{k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}{\int_0^1 \left[\binom{n}{k}p^k(1-p)^{n-k}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta}\right] dp} = \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)} = \frac{\Gamma((\alpha+k)+(\beta+n-k))}{\Gamma(\alpha+k)\Gamma(\beta+n-k)}$$

This looks like magic, but it is indeed how it works.

1.2 Summary

The whole process tells us

$$\text{Beta}(\alpha,\beta) \xrightarrow{\text{Binomial distribution, } k \text{ successes out of } n \text{ trials}} \text{Beta}(\alpha+k, \beta+n-k)$$

1.3 Effective Sample Size

We can also compare the change of the mean between the two Beta distributions (the prior, and the posterior):

$$\text{posterior mean} = \frac{\alpha^*}{\alpha^*+\beta^*} = \frac{\alpha+k}{(\alpha+k)+(\beta+n-k)} = \frac{\alpha+k}{\alpha+\beta+n} = \frac{\alpha}{\alpha+\beta+n} + \frac{k}{\alpha+\beta+n}$$

$$= \frac{\alpha+\beta}{\alpha+\beta+n} \cdot \frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n} \cdot \frac{k}{n}$$

We see that we can split the mean of the posterior distribution into a sum of two products. We can recognize that

$$\frac{\alpha}{\alpha+\beta}: \quad \text{mean of the prior,} \qquad \frac{k}{n}: \quad \text{mean of the observed data}$$

We would like to interpret this sum as a weighted sum of the two means, with weights $\frac{\alpha+\beta}{\alpha+\beta+n}$ and $\frac{n}{\alpha+\beta+n}$. Intuitively, we define

- Effective sample size of $\text{Beta}(\alpha,\beta)$ as $\alpha+\beta$.

The we can interpret the mean of the posterior as follow:

$$\underbrace{\frac{\alpha^*}{\alpha^*+\beta^*}}_{\text{posterior mean}} = \underbrace{\left(\frac{\overbrace{\alpha+\beta}^{\text{prior effective sample size}}}{\underbrace{\alpha+\beta+n}_{\text{posterior effective sample size}}}\right)}_{\text{prior weight}} \cdot \underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{prior mean}} + \underbrace{\left(\frac{\overbrace{n}^{\text{data size}}}{\underbrace{\alpha+\beta+n}_{\text{posterior effective sample size}}}\right)}_{\text{data weight}} \cdot \underbrace{\frac{k}{n}}_{\text{data mean}}$$

1.4 Effect of Stronger Prior

Consider when we impose 2 different Beta priors, $\text{Beta}(1,1)$ and $\text{Beta}(100,100)$. They share the same prior mean $\frac{1}{2} = \frac{100}{200}$. With the same amount of data (size n), compare that

$$\text{Beta}(1,1): \quad \text{prior weight } = \frac{2}{2+n}, \qquad \text{vs} \qquad \text{Beta}(100,100): \quad \text{prior weight } = \frac{200}{200+n}$$

Since $\frac{200}{200+n}$ is closer to 1, the mean of the posterior $\frac{\alpha^*}{\alpha^*+\beta^*}$ is closer to the mean of the prior $\frac{1}{2}$. This means **stronger** the prior we have, the less we will shift away from our original beliefs.

2 Gamma-Poisson Conjugacy

With the detailed discussion in the previous Beta-Binomial Conjugacy, we can derive the Gamma-Poisson Conjugacy in a similar way.

2.1 Definitions

Note: Gamma distribution can have 2 different definitions.

- Gamma distribution:
  - Definition 1 (used in video): $\pi(\lambda) = \frac{1}{\Gamma(k)\theta^k}\lambda^{k-1}e^{-\lambda/\theta}$. Here hyperparameters are k and θ.
  - Definition 2 (used in lab): $\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$. Here hyperparameters are α and β.
  - Relationship between the 2 definitions: $\alpha = k$ while $\beta = \frac{1}{\theta}$.
  - Statistics: mean $= k\theta = \frac{\alpha}{\beta}$, and variance $= k\theta^2 = \frac{\alpha}{\beta^2}$.
- Poisson distribution: $\mathbb{P}(X = x_i \mid \lambda) = \frac{\lambda^{x_i}}{(x_i)!}e^{-\lambda}$. Here, x_i denotes the count **within 1 time period**.

Using Bayes' Rule, we get:

Version 1 with k and θ, over 1 time period:

$$\pi^*(\lambda \mid X = x_i) \propto \mathbb{P}(X = x_i \mid \lambda) \times \pi(\lambda) \propto \left[\lambda^{x_i}e^{-\lambda}\right] \times \left[\lambda^{k-1}e^{-\lambda/\theta}\right] = \lambda^{x_i+k-1}e^{-(1+1/\theta)\lambda}$$

Version 2 with α and β, over 1 time period:

$$\pi^*(\lambda \mid X = x_i) \propto \mathbb{P}(X = x_i \mid \lambda) \times \pi(\lambda) \propto \left[\lambda^{x_i}e^{-\lambda}\right] \times \left[\lambda^{\alpha-1}e^{-\beta\lambda}\right] = \lambda^{x_i+\alpha-1}e^{-(1+\beta)\lambda}$$

Since the only continuous probability distribution sharing the form $\lambda^{\text{some power}-1}e^{-(someconstant)\times\lambda}$ is the Gamma distribution. We conclude that $\pi^*(p)$ must also be the Gamma distribution.

2.2 Summary

Version 1 notation:

$$\text{Gamma}(k, \theta) \xrightarrow{\text{Poisson distribution, } x_i \text{ count within 1 period}} \text{Gamma}\left(x_i + k,\ 1/(1 + \frac{1}{\theta}) = \frac{\theta}{\theta + 1}\right)$$

or Version 2 notation:

$$\text{Gamma}(\alpha, \beta) \xrightarrow{\text{Poisson distribution, } x_i \text{ count within 1 period}} \text{Gamma}(x_i + \alpha,\ 1 + \beta)$$

After sequentially updating over n time periods

Version 1 notation:

$$\text{Gamma}(k, \theta) \xrightarrow{\text{Poisson distribution, } \sum x_i \text{ total counts within } n \text{ period}} \text{Gamma}\left(\sum x_i + k,\ 1/(n + \frac{1}{\theta}) = \frac{\theta}{n\theta + 1}\right)$$

Version 2 notation:

$$\text{Gamma}(\alpha, \beta) \xrightarrow{\text{Poisson distribution, } \sum x_i \text{ total counts within } n \text{ period}} \text{Gamma}\left(\sum x_i + \alpha,\ n + \beta\right)$$

2.3 Effective Sample Size and Analysis of Posterior Mean

Here we use the Gamma(α, β) version to illustrate the idea.

- Effective sample size of Gamma(α, β) is β.

After updating the distribution, we have

$$\text{posterior mean} = \frac{\alpha^*}{\beta^*} = \frac{\alpha + \sum x_i}{\beta + n}$$

$$= \underbrace{\left(\frac{\overbrace{\beta}^{\text{prior effective sample size}}}{\underbrace{\beta + n}_{\text{posterior effective sample size}}} \right)}_{\text{prior weight}} \cdot \underbrace{\frac{\alpha}{\beta}}_{\text{prior mean}} + \underbrace{\left(\frac{\overbrace{n}^{\text{data size}}}{\underbrace{\beta + n}_{\text{posterior effective sample size}}} \right)}_{\text{data weight}} \cdot \underbrace{\frac{\sum x_i}{n}}_{\text{data mean}}$$

When the prior effective sample size β is larger, the posterior mean will be closer to the prior mean, and it will take more time period n to change our prior beliefs.

3 Normal-Normal Conjugacy

IMPORTANT: This conjugacy only works under the assumption that the data variance σ^2 is **known**. So the only parameter we are interested is the **data mean** μ. In Week 3, we will encounter another conjugate family, the **Inverse Gamma-Normal Conjugacy**, which deals with unknown data variance.

3.1 Definitions

- Normal distribution for **prior**: $\pi(\mu) = \mathcal{N}(\nu, \tau^2) = \dfrac{1}{\sqrt{2\pi\tau^2}} e^{-(\mu-\nu)^2/(2\tau^2)}$. Hyperparameters are ν, τ^2.
- Normal distribution for 1 piece of **data** with **known** σ^2: $f(x_i \mid \mu) = \mathcal{N}(\mu, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/(2\sigma^2)}$

Using Bayes' Rule, we have

$$\pi^*(\mu \mid x_i) \propto f(x \mid \mu) \times \pi(\mu) \propto \left[e^{-\dfrac{(x_i-\mu)^2}{2\sigma^2}} \right] \times \left[e^{-\dfrac{(\mu-\nu)^2}{2\tau^2}} \right]$$

$$= \exp\left\{ -\frac{1}{2} \frac{(\mu - \frac{\nu\sigma^2 + x_i\tau^2}{\tau^2+\sigma^2})^2}{\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}} + (1 - \frac{1}{\sigma^2+\tau^2})(\sigma^2\nu^2 + (x_i)^2\tau^2) \right\} \propto \exp\left\{ -\frac{1}{2} \frac{(\mu-\nu^*)^2}{(\tau^*)^2} \right\}$$

where the new hyperparameters ν^* and $(\tau^*)^2$ are

$$\nu^* = \frac{\nu\sigma^2 + x_i\tau^2}{\sigma^2 + \tau^2}, \qquad (\tau^*)^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}$$

(In the above derivation, we use the fact that x_i, ν, σ^2, and τ^2 are known once we have set our prior. Therefore, $\exp\left\{ (1 - \dfrac{1}{\sigma^2+\tau^2})(\nu^2\sigma^2 + (x_i)^2\tau^2) \right\}$ is also a constant.)

If we have observed a set of data $\{x_1, x_2, \cdots, x_n\}$, using **sequential update**, we have

$$\nu^* = \frac{\nu\sigma^2 + (\sum x_i)\,\tau^2}{\sigma^2 + n\tau^2}, \qquad (\tau^*)^2 = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}$$

3.2 Summary

After observing a set of n data points $\{x_1, x_2, \cdots, x_n\}$ with the same **known** variance σ^2, we can update the distribution of the mean μ:

$$\mathcal{N}(\nu, \tau) \xrightarrow{\text{Normal distribution, } n \text{ data points } \{x_1, \cdots, x_n\}} \mathcal{N}\left(\frac{\nu\sigma^2 + (\sum x_i)\,\tau^2}{\sigma^2 + n\tau^2},\ \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$$

3.3 Effective Sample Size

- Effective sample size of $\mathcal{N}(\nu, \tau^2)$ under the **Normal-Normal Conjugacy** is $\dfrac{\sigma^2}{\tau^2}$.

To see this, we need to manipulate the formula of the posterior mean

$$\text{posterior mean } = \nu^* = \frac{\nu\sigma^2 + (\sum x_i)\tau^2}{\sigma^2 + n\tau^2} = \frac{\nu\sigma^2}{\sigma^2 + n\tau^2} + \frac{(\sum x_i)\,\tau^2}{\sigma^2 + n\tau^2}$$

$$= \frac{\sigma^2}{\sigma^2 + n\tau^2} \cdot \nu + \frac{n\tau^2}{\sigma^2 + n\tau^2} \cdot \frac{\sum x_i}{n}$$

$$= \underbrace{\left(\frac{\overbrace{\frac{\sigma^2}{\tau^2}}^{\text{prior effective sample size}}}{\underbrace{\frac{\sigma^2}{\tau^2} + n}_{\text{posterior effective sample size}}}\right)}_{\text{prior weight}} \cdot \underbrace{\nu}_{\text{prior mean}} + \underbrace{\left(\frac{\overbrace{n}^{\text{data size}}}{\underbrace{\frac{\sigma^2}{\tau^2} + n}_{\text{posterior effective sample size}}}\right)}_{\text{data weight}} \cdot \underbrace{\frac{\sum x_i}{n}}_{\text{data mean}}$$

In the Normal-Normal Conjugacy, it turns out that for a fixed data variance σ^2, the **smaller** τ^2 is, the **larger** the prior effective sample size is. This makes sense because the variance τ^2 represents our prior belief of the **spread** of the prior data. The less variance of our belief in the prior data, the more centered the prior will be, then the stronger we trust our prior.