Bayesian Statistics Review

Jiahui Shui

October 10, 2025

1 Introduction

1.1 Basic Concepts in Bayesian Statistics

There are two basic inferences:

- MLE:
$$\hat{\theta}^{MLE} = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta) = \arg\max_{\theta\in\Theta} \log \mathcal{L}(\theta)$$
where $\hat{\theta}^{MLE}$ is a statistics (i.e., a random variable), while the true parameter θ_0 is a constant.

- Bayesian approach:
$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$
where $p(\theta|y)$ is a (conditional) distribution.

Then let's recall the Bayesian rules:

- For events A and B:
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

- For random variables X and Y,
$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)}$$

In Bayesian statistics, parameters are also considered as random variables. Suppose that the econometrician observes data $\boldsymbol{y}$ from some sample $\boldsymbol{Y} \in \mathbb{R}^n$. The purpose of Bayesian

analysis is to use the data $\boldsymbol{y}$ to update the prior belief of θ.

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{p(\boldsymbol{y})}$$

Here $p(\theta|\boldsymbol{y})$ is called *posterior* distribution. And $p(\boldsymbol{y})$ is called *marginal likelihood* or *normalizing constant*. $p(\theta)$ is called *prior distribution*. In addition, we define the hyper parameters as coefficients that parameterize the prior and posterior distributions but do not directly affect the likelihood. We denote λ_0 as the prior hyper-parameters and $\lambda_1 = \lambda_1(y, \lambda_0)$ as the posterior hyper-parameters. Then the prior is $p(\theta; \lambda_0)$, the posterior is

$$p(\theta|\boldsymbol{y}; \lambda_1) = p(\theta|\boldsymbol{y}; \lambda_0)$$

and the marginal likelihood is

$$p(\boldsymbol{y}; \lambda_0, \lambda_1) = p(\boldsymbol{y}; \lambda_0)$$

Definition 1. *In Bayesian statistics, the kernel of a PDF is the form of the PDF in which factors that are not functions of any of the model parameters are omitted,* $p \propto K$

Take normal distribution as an example. Say, $Y|\mu, \sigma^2 \sim N(\mu, \sigma^2)$. Then the (exact) likelihood of the sample would be

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

If σ^2 is known, then

$$p(y|\mu) \propto e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{\frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}}$$

If μ is known, then

$$p(y|\sigma^2) \propto \frac{1}{\sqrt{\sigma^2}} e^{-\frac{(y-\mu)^2}{\sigma^2}}$$

Note that posterior can be also written as

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{\int p(\boldsymbol{y}|\theta)p(\theta)d\theta}$$

Here actually we can ignore $p(\boldsymbol{y})$ since it does not contain any information about true parameter θ. Thus

$$\underbrace{p(\theta|\boldsymbol{y})}_{\text{posterior}} \propto \underbrace{p(\boldsymbol{y}|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$

1.2 Some Examples

- Binomial with uniform prior: Suppose that $Y \sim B(n, \theta)$. Then

$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad 0 \leq y \leq n, y \in \mathbb{N}$$

If we have uniform prior on θ, i.e. $\theta \sim U([0,1])$, then

$$p(\theta|y) \propto p(y|\theta)p(\theta) = \theta^y(1-\theta)^{n-y}$$

This is the posterior kernel. And it corresponds to Beta$(y+1, n-y+1)$ distribution. Then how can we find marginal likelihood $p(y)$? The first way is through integration:

$$\begin{aligned} p(y) &= \int p(y|\theta)p(\theta)\mathrm{d}\theta \\ &= \int_0^1 \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y}\mathrm{d}\theta \\ &= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\mathcal{B}(y+1, n-y+1) \\ &= \frac{\Gamma(n+1)}{\Gamma(n+2)} \\ &= \frac{1}{n+1} \end{aligned}$$

where $\mathcal{B}(p,q)$ is Beta function, and

$$\mathcal{B}(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

The other way is to use normalizing constant method. Omitted (you have to recognize the kernel is Beta distribution).

- Binomial with Beta prior: Similar setting, but this time, the prior of θ is $\mathcal{B}(\alpha_0, \beta_0)$. Then

$$p(\theta|y) \propto p(y|\theta)p(\theta) \propto \theta^{y+\alpha_0-1}(1-\theta)^{n-y+\beta_0-1}$$

Hence the posterior should be $\mathcal{B}(y+\alpha_0, n-y+\beta_0)$

1.3 Model Comparison

A model is defined by a likelihood function and a prior. Suppose that we have m models, $\mathcal{M}_i$ for $i = 1, \cdots, m$. They are all going to explain $\boldsymbol{y}$. And model $\mathcal{M}_i$ depends

upon parameters θ^i. The posterior for the parameters θ^i calculated using model $\mathcal{M}_i$ is

$$p(\theta^i|\boldsymbol{y}, \mathcal{M}_i) = \frac{p(\boldsymbol{y}|\theta^i, \mathcal{M}_i)p(\theta^i|\mathcal{M}_i)}{p(\boldsymbol{y}|\mathcal{M}_i)}$$

The posterior model probability is

$$p(\mathcal{M}_i|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathcal{M}_i)p(\mathcal{M}_i)}{p(\boldsymbol{y})}$$

where $p(\mathcal{M}_i)$ is the prior model probability, which measures how likely we believe $\mathcal{M}_i$ to be correct model before seeing the data. $p(\boldsymbol{y}|\mathcal{M}_i)$ is the marginal likelihood, which is calculated using

$$p(\boldsymbol{y}|\mathcal{M}_i) = \int p(\boldsymbol{y}|\theta^i, \mathcal{M}_i)p(\theta^i|\mathcal{M}_i)\mathrm{d}\theta^i$$

The posterior odds ratio is defined as:

$$\mathrm{PO}_{ij} = \frac{p(\mathcal{M}_i|\boldsymbol{y})}{p(\mathcal{M}_j|\boldsymbol{y})} = \underbrace{\frac{p(\boldsymbol{y}|\mathcal{M}_i)}{p(\boldsymbol{y}|\mathcal{M}_j)}}_{\text{Bayes Factor}} \underbrace{\frac{p(\mathcal{M}_i)}{p(\mathcal{M}_j)}}_{\text{prior odd ratio}}$$

How to present the posterior distribution if there is no closed form solution? Consider the following example: Binomial likelihood with truncated normal prior: $\theta \sim p_{TN(0,1,0,1)}(\theta), 0 \leq \theta \leq 1$. That is

$$p(\theta) = \frac{\sqrt{2\pi}}{\Phi(1) - \Phi(0)} \exp\left(-\frac{\theta^2}{2}\right) \propto \exp\left(-\frac{\theta^2}{2}\right)$$

The posterior kernel is

$$p(\theta|y) \propto \theta^y(1-\theta)^{n-y} \exp\left(-\frac{\theta^2}{2}\right)$$

According to the Bayes' rule, one can derive that

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)\mathrm{d}\theta} = \frac{\theta^y(1-\theta)^{n-y}\exp(-\frac{\theta^2}{2})}{\int \theta^y(1-\theta)^{n-y}\exp(-\frac{\theta^2}{2})\mathrm{d}\theta}$$

Thus, the posterior mean is

$$\mathbb{E}[\tilde{\theta}|y] = \int_0^1 \theta p(\theta|y)\mathrm{d}\theta = \int \frac{\theta^y(1-\theta)^{n-y}\exp(-\frac{\theta^2}{2})}{\int \theta^{y+1}(1-\theta)^{n-y}\exp(-\frac{\theta^2}{2})\mathrm{d}\theta}\mathrm{d}\theta$$

1.4 Monte Carlo Integration

We would like to evaluate the following integral:

$$\mathbb{E}[g(\theta)|\boldsymbol{Y}] = \int g(\theta)p(\theta|\boldsymbol{y})\mathrm{d}\theta$$

Law of Large number can help us to achieve that.

1. Generate S i.i.d random draws $\{y^{(s)}\}_{s=1}^{S}$ from $p_Y(y)$, where each $y^{(s)} = Y_s(\omega)$ is a realization of $Y_s \sim p_Y(y)$ i.i.d
2. Calculate
$$\frac{1}{S}\sum_{s=1}^{S} g(y^{(s)})$$

2 Single-parameter models

2.1 Conjugate Prior

Definition 2. *A prior distribution $p(\theta) \in \mathcal{F}$ is said to be conjugate for a likelihood function $p(y|\theta)$ if the posterior distribution $p(\theta|y) \in \mathcal{F}$.*

Definition 3. *A conjugate prior that has the same functional form as the likelihood function regarded as a function of θ.*

2.2 Some Examples

- Exponential-Gamma system: Suppose that $Y_i|\theta \sim \mathrm{Exp}(\theta)$, then the likelihood is given by
$$p(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} \theta \exp(-\theta y_i) = \theta^n \exp(-n\bar{y}_n\theta)$$
The conjugate prior is Gamma distribution: $\theta \sim \mathrm{Gamma}(\alpha_0, 1/\beta_0)$, i.e.
$$p(\theta) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\theta^{\alpha_0-1}\exp(-\beta_0\theta) \propto \theta^{\alpha_0-1}\exp(-\beta_0\theta)$$
The kernel of the posterior is given by
$$p(\theta|\boldsymbol{y}) \propto \theta^{n+\alpha_0-1}\exp(-(\beta_0 + n\bar{y}_n)\theta)$$
which is $\mathrm{Gamma}(\alpha_1, 1/\beta_1)$ distribution.

2.3 Exponential Family

Definition 4 (Exponential Family)**.** *A PDF $p(y|\theta)$ where $\theta \in \mathbb{R}$ is said to belong to the one-parameter exponential family if it has form*

$$p(y|\theta) = c(\theta)h(y)\exp\left(\phi(\theta)t(y)\right) = h(y)\exp\left(\phi(\theta)t(y) - \kappa(\theta)\right)$$

for some functions $h(y), \phi(\theta), \kappa(\theta) = -\log c(\theta)$ and $t(y)$. If the support of Y is independent of θ, then the family is said to be regular and otherwise it is irregular.

Here are some examples

- Exponential distribution:

$$p(y|\theta) = \theta\exp(-\theta y)\mathbb{1}_{\{y>0\}}$$

  Then $\phi(\theta) = -\theta$, $t(y) = y$, $\kappa(\theta) = -\log\theta$, $h(y) = \mathbb{1}_{\{y>0\}}$.

- Poisson distribution:

$$p(y|\theta) = \frac{\theta^y}{y!}e^{-\theta}\mathbb{1}_{\{y\in\mathbb{Z}_+\}}$$

  Then $\phi(\theta) = \log\theta$, $t(y) = y$, $\kappa(\theta) = \theta$, $h(y) = \frac{1}{y!}\mathbb{1}_{\{y\in\mathbb{Z}_+\}}$

- Uniform distribution
- Cauchy distribution
- Normal distribution with unknown mean
- Normal distribution with unknown variance
- What about normal distribution with unknown mean and variance? See next section.

Theorem 1. *Any exponential family population has a conjugate prior, with kernel*

$$p(\theta) \propto \exp(b_0\phi(\theta) - a_0\kappa(\theta)) \tag{1}$$

Proof. The posterior is given by

$$\begin{aligned} p(\theta|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\theta)p(\theta) \propto \prod_{i=1}^{n} h(y_i)\exp\left(\phi(\theta)t(y_i) - \kappa(\theta)\right)\exp(b_0\phi(\theta) - a_0\kappa(\theta)) \\ &\propto \exp\left\{\left(b_0 + \sum_{i=1}^{n} t(y_i)\right)\phi(\theta) - (a_0+n)\kappa(\theta)\right\} = \exp(b_1\phi(\theta) - a_1\kappa(\theta)) \end{aligned}$$

where $b_1 = b_0 + \sum_{i=1}^{n} t(y_i)$ and $a_1 = a_0 + n$. □

2.4 Normal Distribution

Normal Mean with Known Variance

First, let us consider normal distribution with known variance, and the mean μ, is the parameter that we are interested in. Let $\theta = \mu$, then

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} e^{\frac{y\theta}{\sigma^2} - \frac{\theta^2}{2\sigma^2}}$$

Then

$$\phi(\theta) = \frac{\theta}{\sigma^2}, \quad t(y) = y, \quad \kappa(\theta) = \frac{\theta^2}{2\sigma^2}, \quad h(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}$$

By the previous theorem, we know that the conjugate prior would be

$$p(\theta) \propto \exp(b_0\phi(\theta) - a_0\kappa(\theta)) = \exp\left(b_0\frac{\theta}{\sigma^2} - a_0\frac{\theta^2}{2\sigma^2}\right)$$

One can parameterize this family as

$$\theta \sim N(\mu_0, \tau_0^2) \Rightarrow p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right)$$

Suppose that $Y_i|\theta \sim N(\theta, \sigma^2)$ with σ^2 being known. Then the likelihood is given by

$$\begin{aligned}
p(\boldsymbol{y}|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta)^2}{2\sigma^2}} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\theta + \theta^2)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(n\theta^2 - 2n\theta\bar{y} + \sum_{i=1}^{n} y_i^2 + \bar{y}_n^2 - \bar{y}_n^2\right)\right) \\
&\propto \exp\left(-\frac{n}{2\sigma^2}(\theta - \bar{y}_n)^2\right)
\end{aligned}$$

where $\bar{y}_n := \frac{1}{n}\sum_{i=1}^n y_i$. Then the MLE is given by

$$\hat{\theta}^{ML} = \bar{y}_n \sim N(\theta, \sigma^2/n)$$

The posterior is given by

$$p(\theta|\boldsymbol{y}) \propto p(\theta)p(\boldsymbol{y}|\theta) \propto \exp\left(-\frac{1}{2}\left(\frac{(\theta-\mu_0)^2}{\tau_0^2} + \frac{(\theta-\bar{y})^2}{\sigma^2/n}\right)\right)$$

What is this distribution? We need some algebra. (Check the completing the squares method). It can be written as

$$p(\theta|\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\left(\frac{(\theta-\mu_1)^2}{\tau_1^2} + \frac{(\mu_0-\bar{y})^2}{\tau_0^2+\sigma^2/n}\right)\right) \propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)$$

where

$$\tau_1^2 = \frac{1}{1/\tau_0^2 + n/\sigma^2}, \quad \mu_1 = \tau_1^2\left(\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}_n}{\sigma^2}\right) = \gamma\mu_0 + (1-\gamma)\bar{y}_n$$

if we define

$$\gamma = \frac{1/\tau_0^2}{1/\tau_0^2 + n/\sigma^2}$$

Normal Variance with Known Mean

Let $\theta = \sigma^2$, then

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\theta}}\exp\left(-\frac{(y-\mu)^2}{2\theta}\right) \propto \exp\left(-\frac{(y-\mu)^2}{2\theta} - \frac{1}{2}\log\theta\right)$$

Hence

$$\phi(\theta) = -\frac{1}{2\theta}, \quad t(y) = (y-\mu)^2, \quad \kappa(\theta) = \frac{1}{2}\log\theta$$

By previous theorem, the kernel of conjugate prior is given by

$$p(\theta) \propto \exp(b_0\phi(\theta) - a_0\kappa(\theta)) = \exp\left(-\frac{b_0}{2\theta} - \frac{a_0}{2}\log\theta\right) = \theta^{-\frac{a_0}{2}}\exp\left(-\frac{b_0}{2\theta}\right)$$

This is Inverse Gamma distribution. We can parameterize this family as

$$\theta \sim IG(\nu_0, s_0^2)$$

Then

$$p(\theta) \propto \theta^{-(\frac{\nu_0}{2}+1)} \exp\left(-\frac{\nu_0}{2}\frac{s_0^2}{\theta}\right)$$