# Bayesian Statistics Notes

Haolin Li

September 18, 2024

**Abstract**

The general purpose of Bayesian Statistics is to find/estimate the joint distribution $\mathbf{P}(\mathbf{y_1}, \ldots, \mathbf{y_n}, \theta_\mathbf{1}, \ldots, \theta_\mathbf{m})$, from which we explore applications with the help of posterior and predictive distributions, and of course, the Bayes' Rule. It includes paramatric, from one-parameter distributions like Bernoulli and Poission to multiple-parameter like Normal, and unparametric methods. This serves as an introductory course to the world of Bayesian.

# Contents

# Chapter 1

# Introduction and Notation

The core difference between a Bayesian and a frenquencist is the belief on whether the latent or the parameter $\theta$ is a random variable or a constant. One of the major impacts is that iid assumptions changes to conditional independence instead of mutual independence due to the connection between $\theta$ and $Y$ through the joint distribution $\mathbf{P}(\mathbf{Y_i}, \theta_\mathbf{j})$, where the marginal distribution of $Y_i$ are parametrized by $\theta$. Before everything, we should introduce some simple notations.

## 1.1 Notation

- $\mathcal{Y}$ denote the set of all possible observation values
- $Y$ denote the random variable
- $y$ denote the value of a single observation
- $\Theta$ is the space of parameters

**Note.** Let $\theta \in \Theta$, define $\pi(\theta)$ or $p(\theta)$ as the prior distribution.

**Note.** For any $\theta \in \Theta,\ y \in Y$, $\mathbf{P}(\mathbf{y}|\theta)$ describes the sampling model

**Note.** Let $\theta \in \Theta$, the posterior distribution $\mathbf{P}(\theta|\mathbf{y})$ describes our belief about the parameters based on samples.

**Theorem 1.1.1** (Bayes' Rule).

$$\mathbf{P}(\theta|\mathbf{y}) = \frac{\mathbf{P}(\mathbf{y}|\theta)\pi(\theta)}{\mathbf{P}(\mathbf{y})} \tag{1.1}$$

**Example.** Suppose $\theta \in [0,1]$, and $Y_i|\theta \sim Bernoulli(\theta)$ with sample size 20. Then let $y|\theta = \sum Y_i \sim Binomial(20, \theta)$. We will see how the choice of the prior has on the posterior with $\theta \sim Beta(a, b)$, we have

$$\mathbb{E}[\theta] = \frac{a}{a+b}, \qquad mode[\theta] = \frac{a-1}{a-1+b-1}$$

Usually, from a prior sampling, we denote $a$ be the number of events counted, and $b$ as the total sample size. In this case, say the event happens twice, we have

$$\theta \sim Beta(2, 20)$$

We know that the pdf of $Beta(a, b)$ is

$$pdf(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1},\ \theta \in [0,1]$$

Using Bayes' Rule, we have

$$\begin{aligned}\mathbf{P}(\theta|\mathbf{Y_1},\ldots,\mathbf{Y_{20}}) &\propto \mathbf{P}(\mathbf{Y_1},\ldots,\mathbf{Y_{20}}|\theta)\cdot\pi(\theta)\\ &\propto \theta^{\Sigma Y_i}(1-\theta)^{n-\Sigma Y_i}\cdot\theta^{a-1}(1-\theta)^{b-1}\\ &= \theta^{a+y-1}(1-\theta)^{b+n-y-1}\\ &\sim Beta(a+y, b+n-y)\end{aligned}$$

This means, amazingly, the posterior falls in the same family of distribution like the prior. For priors like this, we give them a special name: Conjugate Prior.

From the example, we can derive

$$\mathbb{E}[\theta] = \frac{a}{a+b}, \qquad mean(sample) = \frac{y}{n}, \qquad \mathbb{E}[\theta|y] = \frac{a+y}{a+b+n}$$

With a little massage, we have

$$\mathbb{E}[\theta|y] = \frac{a+b}{a+b+n}\cdot\frac{a}{a+b} + \frac{n}{a+b+n}\cdot\frac{\sum y_i}{n}$$

This break down is intriguing because the posterior mean is in fact the weighted sum of the prior mean and sample mean, implying insensitivity to the prior as $n \to \infty$ since the weight dominates. We can further conclude the above example with the following proposition:

**Proposition 1.1.1.** With $Y_i|\theta \sim Bernoulli(\theta)$ and $\theta \sim Beta(a,b)$ as the conjugate prior, we have the posterior $\theta|Y_1,\ldots,Y_n \sim Beta(a+\sum Y_i, b+n-\sum Y_i)$

**Remark.**

$$Uniform[0,1] = Beta(1,1)$$

## 1.2 One-parameter Models

### 1.2.1 Sufficient Statistics and Conjugate Prior

In this section, we talk about single-parameter models, where the example in **??** about Bernoulli with Beta priors was a perfect example. Let's start with a closer look at the posterior with uniform prior:

$$\mathbf{P}(\theta|\mathbf{Y_1},\ldots,\mathbf{Y_n}) \ \propto\ \mathbf{P}(\mathbf{Y_1},\ldots,\mathbf{Y_n}|\theta)\cdot\pi(\theta) \ =\ \theta^{\mathbf{\Sigma Y_i}}(\mathbf{1}-\theta)^{\mathbf{n-\Sigma Y_i}}$$

By observing the last term, the posterior distribution is determined by the statistic $\sum Y_i$(we assume sample size known at all time). This means we don't need to examine the exact values of $Y_i$, but the sum would be enough/sufficient to find out the parameters of the posterior. As a result, we say the sum $\sum Y_i$ is the sufficient statistic of the posterior distribution.

**Definition 1.2.1** (Sufficient Statistics). Given any subject $\mathcal{S}$ we are trying to estimate, a distribution, a parameter, or even another statistic, a statistic $T(Y_i)$ is a sufficient statistic of $\mathcal{S}$ if $T(Y_i)$ contains enough information for us to determine that subject.

In the previous section, we also mentioned the rough idea of a conjugate prior, now we give it a formal definition:

**Definition 1.2.2** (Conjugate Prior). A class of prior distributions $\mathcal{P}$ for $\theta$ is called conjugate for a sampling model $\mathbf{P}(\mathbf{Y}|\theta)$ if

$$\pi(\theta) \in \mathcal{P} \Rightarrow \mathbf{P}(\theta|\mathbf{Y}) \in \mathcal{P}$$

Now we see how these two concepts play together with the following example.

**Example.** Previously we have talked about the posterior conditioned over the entire sequence, $\theta|Y_1, \ldots, Y_n$. What would happen if we instead condition the parameter with posterior's sufficient statistic? i.e. $\theta|y = \sum_{i=1}^{n} Y_i$

$$\begin{aligned}
\mathbf{P}(\theta|\mathbf{y}) &\propto \mathbf{P}(\mathbf{y}|\theta) \cdot \pi(\theta) \\
&= \binom{n}{y} \theta^y (1-\theta)^{n-y} \theta^{a-1} (1-\theta)^{b-1} \\
&\propto \theta^{a+y-1} (1-\theta)^{b+n-y-1} \\
&\sim Beta(a+y, b+n-y)
\end{aligned}$$

Surprisingly, it looks the same as $\theta|Y_1, \ldots, Y_n$! This is because $y$, as the sufficient statistic for the posterior, is enough to determine the distribution.

### 1.2.2 Predictive Distribution

There are two main reasons why we model things with mathematics in general: to explain, and to predict. In this section, we explore how can we make predictions within the Bayesian framework.

**Definition 1.2.3** (Predictive Distribution). Given data points $Y_1, \ldots, Y_n$, the predictive distribution refers to

$$Y_{n+1}|Y_1, \ldots, Y_n$$

**Example.** Say we have $Y_i|\theta \sim Bernoulli(\theta)$ and prior $\pi(\theta) \sim Beta(a, b)$, then the predictive distribution $Y_{n+1}|Y_1, \ldots, Y_n \sim Bernoulli(\hat{\theta})$, and we will try to find the exact value of $\hat{\theta}$ in order to determine the predictive distribution. Using marginalization, we have

$$\begin{aligned}
\hat{\theta} &= \mathbf{P}(\mathbf{Y_{n+1}} = \mathbf{1}|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) \\
&= \int_0^1 \mathbf{P}(\mathbf{Y_{n+1}} = \mathbf{1}, \theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) \mathbf{d}\mathbb{P}(\theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) \\
&= \int_0^1 \theta \cdot pdf(Beta(a + \sum Y_i, b + n - \sum Y_i)) d\theta \\
&= \mathbb{E}[\theta|Y_1, \ldots, Y_n] \\
&= \frac{a + \sum Y_i}{a + b + n}
\end{aligned}$$

As a result, we have our predictive distribution $Y_{n+1}|Y_1, \ldots, Y_n \sim Bernoulli(\frac{a+\Sigma Y_i}{a+b+n})$

### 1.2.3 Confidence Regions

After we have known how to estimate the next observation with predictive distribution, now let's find out how to estimate the parameters.

If we were frenquencists, we first assume the sample distribution, with which then determine the confidence interval(also random variables), after that we sample and see if the result lies in the interval, which determines whether we reject hypothesis $H_0$.

However, as a Bayesian, the confidence of our estimation on $\theta$ originates from the posterior $\mathbf{P}(\theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n})$, which means we have to sample before determining the "interval" or in essence, the distribution itself. Also, we are able to tell the odds very clearly because right now we know exactly how $\theta|Y_1, \ldots, Y_n$ distributes, at least that's what we hope for.

Moreover, in the Bayesian way, there are two major ways to determine the interval with the posterior distribution.

1. Highest Posterior Density(HPD): points with the highest of the posterior pdf, also $\mathbb{P}(HDP_\alpha) = 1 - \alpha$.

2. Quantile based interval: old-fashion $(\theta_{\alpha/2}, \theta_{1-\alpha/2})$

### 1.2.4 Case Study: Possion + Gamma Prior

Now we are in business, in this section, we introduce another pair of conjugate prior. A little recap, if $Y \sim Poission(\theta)$, then $\mathbf{P}(\mathbf{Y} = \mathbf{y}|\theta) = \frac{\mathbf{e}^{-\theta}\theta^{\mathbf{y}}}{\mathbf{y}!}$, $\mathbf{y} \in \underline{\mathbf{N}}$. We will leave the pdf of Gamma for now because in this case study, we present a way to actually find the distribution. Firstly, a little analysis on the joint sampling distribution:

$$\begin{aligned}\mathbf{P}(\mathbf{Y_1}, \ldots, \mathbf{Y_n}|\theta) &= \prod_{i=1}^{n} \mathbf{P}(\mathbf{Y_i}|\theta) \\ &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{Y_i}}{Y_i!} \\ &= \frac{e^{-n\theta}\theta^{\Sigma Y_i}}{\prod_{i=1}^{n} Y_i!}\end{aligned}$$

**Note.** $Y = Y_1 + \cdots + Y_n \sim Possion(n\theta)$.

Right now we can work on the posterior.

$$\mathbf{P}(\theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) \propto \pi(\theta) \cdot \mathbf{P}(\mathbf{Y_1}, \ldots, \mathbf{Y_n}|\theta) \tag{1.2}$$

$$\propto \pi(\theta) \cdot e^{c_1\theta}\theta^{c_2} \tag{1.3}$$

Let's guess what the conjugate prior should look like. If $\pi(\theta) \propto e^{d_1\theta}\theta^{d_2}$, then the prior and the posterior will be proportionate to the same pattern, i.e. $\mathbf{P}(\theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) \propto \mathbf{e}^{(\mathbf{c_1}+\mathbf{d_1})\theta}\theta^{(\mathbf{c_2}+\mathbf{d_2})}$. Then all we need to do now is to determine the constant depending on $c_i, d_i$, the result is

$$\pi(\theta) \sim Gamma(a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}$$

Therefore, we can push further with 1.3 so we have

$$\mathbf{P}(\theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) \propto \mathbf{Gamma}(\mathbf{a} + \mathbf{\Sigma Y_i}, \mathbf{b} + \mathbf{n})$$

And we conclude this case study with the following proposition

**Proposition 1.2.1.** With $Y_i|\theta \sim Poission(\theta)$ and $\theta \sim Gamma(a, b)$ as the conjugate prior, we have the posterior $\theta|Y_1, \ldots, Y_n \sim Gamma(a + \Sigma Y_i, b + n)$.

## 1.3 Exponential Family and Monte Carlo Method

In previous examples, we have seen two conjugate distribution pairs that saves our lives from tedious computation. A natural question one may ask is, does there exist a pattern for us to find conjugate priors given any sampling distribution? The answer is almost yes, there's indeed a pattern for huge family of distributions: the exponential family.

**Definition 1.3.1** (One-parameter Exponential Family)**.** A sampling model $\mathbf{P}(\mathbf{Y}|\theta)$ is an one-parameter exponential family model if we have the following decomposition:

$$\mathbf{P}(\mathbf{Y}|\theta) = \mathbf{h}(\mathbf{Y})\mathbf{c}(\theta)\exp(\theta\mathbf{t}(\mathbf{Y})) \tag{1.4}$$

where $\theta$ is a parameter, and $t(Y)$ is the sufficient statistic for the posterior.

In imitation to (1.4), with $n_0, t_0$ being the sample size and sufficient statistic of the prior sample, we suppose our prior takes the form:

$$\pi(\theta) = k(n_0, t_0)c(\theta)^{n_0} \exp(n_0 t_0 \theta)$$

We then can derive the posterior distribution using the Bayes' Rule:

$$\begin{aligned}\mathbf{P}(\theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) &\propto \pi(\theta) \cdot \mathbf{P}(\mathbf{Y_1}, \ldots, \mathbf{Y_n}|\theta) \\ &\propto c(\theta)^{n_0+n} \exp\{(n_0+n)\frac{n_0 t_0 + n\bar{t(y)}}{n_0+n}\theta\}\end{aligned}$$

With the color, we can easily see that the prior and the posterior fall into the same family of distribution, which by definition is conjugate.

As for Monte Carlo, it's just a fancy name for simulation/experiment. Specifically, it applies to the field of Bayesian by allowing us to simulate the sampling process. Given a joint distribution $\mathbf{P}(\mathbf{Y}, \theta)$, we can rewrite it as $\pi(\theta)\mathbf{P}(\mathbf{Y}|\theta)$. There are 3 steps for us to follow, given sample $Y_i$:

1. Sample a list of parameter $\vec{\theta} = \{\theta_1, \ldots, \theta_n\}$ with $\theta_i \sim \pi(\theta|Y_1, \ldots, Y_{n_0})$
2. For each $\theta_i$, sample $\tilde{Y_i} \sim \mathbf{P}(\mathbf{Y}|\theta, \mathbf{Y_1}, \ldots, \mathbf{Y_{n_0}})$
3. return $\{(\theta_i, \tilde{Y_i})\}_{i=1}^n$

Sometimes we may find our simulation does not align with our empirical data. Firstly, it's possible we were unlucky, so we can run this entire experiment $k$ times, meaning a total of $k \cdot n$ pairs of $(\theta, Y)$ will be generated. It's also possible that we need to change our model, both the sampling model and the prior have room for adjustment.

# Chapter 2

# Multi-Parameter Models

In the previous chapter, we introduced the idea of Bayesian Statistics, two one-parameter models, and one-parameter exponential family. In this chapter, we will go explore models with multiple parameters, together with some advanced ideas.

## 2.1 The most objective: Jeffreys' Prior

When we are trying to select a prior, we often turn to the uniform distribution if we have no idea about how the parameters are distributed. Uniform is believed to be objective because all possible parameters are assigned with the same weight, meaning we are not favoring any particular options. But it's far from being perfect, because if so, this section won't exist. In this section, we will first explain why uniform can be problematic sometimes, and then introduce the Jeffreys' prior, which is believed to be more objective to some.

One of the good sides of using uniform as prior is that, as long as you are selecting the parameter(s) within a finite region, you don't need to adjust the weight since it will be offset in the normalization factor afterall. However, if $|\Omega| = \infty$ for example $[0, +\infty), (-\infty, 0]$, then the integration of the prior is bound to be infinite. The problem is that, not only the decomposition $\mathbf{P}(Y = y) = \int_\Omega \mathbf{P}(Y = y|\theta)d\theta$ will very likely to be invalid because the integration might not converge. In addition, even if it converges, uniform prior with $\pi(\theta) = 1$ becomes less objective because now it tends to favor larger $\theta$ since for all finite interval/region $I$, $\int_I \pi(\theta)d\theta < \infty$ but $\int_{\Omega\setminus I} \pi(\theta)d\theta = \infty$.

Apart from infinite regions, uniform prior also doesn't work well with change of variables. Take the following as an example:

**Example.** Let $\theta \sim Unif(0, 1)$ and odds $\tau = \frac{\theta}{1-\theta}$. Then we have $\theta = \frac{\tau}{1+\tau}$ and $\frac{d\theta}{d\tau} = \frac{1}{(1+\tau)^2}$. Using the formula for change of variable, we have $pdf(\tau) = \pi(\theta(\tau))\frac{d\theta}{d\tau} = \frac{1}{(1+\tau)^2}$, $\tau \in (0, +\infty)$.

To address these issues, mostly on the second one, we introduce Jeffreys' Prior

$$\pi(\theta) = \sqrt{I(\theta)}$$

where $I(\theta)$ is the *Fisher Information* with $I(\theta) = -\mathbb{E}[\frac{\partial^2 \log p(y|\theta)}{\partial\theta^2}|\theta] = -\mathbb{E}[\frac{\partial^2 \log L}{\partial\theta^2}|\theta]$. It's not intuitive at the first glance, but the following property enbales Fisher Information to be a part of the prior.

**Proposition 2.1.1.**

$$I(\theta) = -\mathbb{E}[\frac{\partial^2 \log L}{\partial\theta^2}] = \mathbb{E}[(\frac{\partial \log L}{\partial\theta})^2]$$

With this property, we can prove that Jeffreys' prior is invariable under change of variables. In other words, we have:

**Theorem 2.1.1.** With $\phi = \phi(\theta)$ and $p(\theta) \propto \sqrt{I(\theta)}$, we have $p(\phi) \propto \sqrt{I(\phi)}$.

One question one may have right now is why is Jeffreys' prior objective? Clearly it doesn't assign the same weight on all possible parameter candidates since it's not a uniform. The answer is, Jeffreys'

Prior is in fact the maximizer of the KL-divergence bewteen the prior and posterior distribution. In this sense, Jeffreys' prior is objective because it "allows" the data to speak the most of it. Apparently, this idea can be philosophical and up to individual's personal perspective, but at least mathematically it's a handy wrench in the toolbox by introducing invariability to reparametrization.

**Note** (The maximizer of KL-divergence). TBD

## 2.2 The Normal Model

Normal distribution will be the first multi-parameter model we are going to study. The methodology is intuitive, instead of modeling two parameters at a time, we first fix $\sigma^2$ as a constant, which now equals to a one-parameter model. Then, we introduce $\sigma^2$ as a random variable and repeat the procedure.

### 2.2.1 Condition Posterior

We have joint sampling density as the product of $n$ normal due to conditional independence

$$\begin{aligned}
\mathbf{P}(y_1,\dots,y_n|\theta,\sigma^2) &\propto \prod_{i=1}^{n}\exp\left(-\frac{1}{2\sigma^2}(y_i-\theta)^2\right) && (2.1)\\
&= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right) && (2.2)\\
&= \exp\left(-\frac{1}{2}(\frac{n}{\sigma^2}\theta^2 - 2\frac{\sum y_i}{\sigma^2} + \frac{\sum y_i^2}{\sigma^2})\right) && (2.3)\\
&= \exp\big(c_1(\theta-c_2)^2 + c_3\big) && (2.4)
\end{aligned}$$

By fixing $\sigma^2$ as a constant, we have conditional posterior

$$\mathbf{P}(\theta|y_1,\dots,y_n,\sigma^2) \propto \mathbf{P}(y_1,\dots,y_n|\theta,\sigma^2)\cdot\mathbf{P}(\theta|\sigma^2) \tag{2.5}$$

To make it conjugate, one straightforward choice for prior is also normal, specifically $N(\mu_0,\tau_0^2)$. Then we have the following decomposition of the prior pdf:

$$\begin{aligned}
\mathbf{P}(\theta|\sigma^2) &\propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)\right) && (2.6)\\
&= \exp\left(-\frac{1}{2}(\frac{1}{\tau_0^2}\theta^2 - 2\frac{\mu_0}{\tau_0^2}\theta + \frac{\mu_0^2}{\tau_0^2})\right) && (2.7)\\
&= \exp\left(-\frac{1}{2}(a\theta^2 - 2b\theta^2 + c)\right) && (2.8)
\end{aligned}$$

This expression offers us a quick way to compute the mean and gradient of the normal distribution:

$$\tau_0^2 = \frac{1}{a}, \quad \mu_0 = b\cdot\tau_0^2 = \frac{b}{a} \tag{2.9}$$

Now we have collected all of the ingredients we need, by combining (2.3) and (2.7) into (2.5), we obtain the following decomposition:

$$\mathbf{P}(\theta|y_1,\dots,y_n,\sigma^2) \propto \exp\left(-\frac{1}{2}\left[(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2})\theta^2 - 2(\frac{\sum y_i}{\sigma^2}+\frac{\mu_0}{\tau_0^2})\theta + (\frac{\sum y_i^2}{\sigma^2}+\frac{\mu_0^2}{\tau_0^2})\right]\right) \tag{2.10}$$

By using the formula (2.9) and substituting the variance terms with precision, we obtain conditional posterior mean and variance:

$$\tau_n^2 = \frac{1}{a} = \frac{1}{n\tilde{\sigma}^2+\tilde{\tau_0}^2}, \quad \mu_n = \frac{b}{a} = \frac{\sum y_i\tilde{\sigma}^2 + \mu_0\tilde{\tau_0}^2}{n\tilde{\sigma}^2+\tilde{\tau_0}^2}$$

which means we have the conditional postertior

$$\theta|y_1,\dots,y_n,\sigma^2 \sim N(\mu_n,\tau_n^2)$$

### 2.2.2 Law of Total Expectation/Variance

Conditional expectation $\mathbb{E}[U|V = v]$ is a random variable in $V$, which means its value changes according to $V$. The low of total expectation states that

$$\mathbb{E}[\mathbb{E}[U|V]] = \mathbb{E}[U]$$

The proof is quite simple, by writing the left hand side explicitly:

$$\begin{aligned}
&\int_V \int_{U|V=v} u \cdot p(U = u|V = v)du \cdot p(V = v)dv \\
&= \int_V \int_{U|V=v} u \cdot p(U = u)dudv \\
&= \int_V \int_U u \cdot p(U = u)dudv \\
&= \int_V \mathbb{E}[U]dv = \mathbb{E}[U]
\end{aligned}$$

By doing something similar, we derive the law of total variance:

$$Var(U) = \mathbb{E}[Var(U|V)] + Var(\mathbb{E}[U|V])$$

# Appendix

# Appendix A

# Normalization Tricks

This trick is the direct application of $\int pdf(x)dx = 1$, so we only need to remember what the pdfs look like for those frequently used distributions. The integration with respect to the parameters is the inverse of the constant.

- $Beta(a, b) \sim \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$
- $Gamma(a, b) \sim \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$