Bayesian Statistics Notes

Haolin Li

September 28, 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 θ 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 θ and Y through the joint distribution $\mathbf{P}(\mathbf{Y_i}, \theta_\mathbf{j})$, where the marginal distribution of Y_i are parametrized by θ. 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
- Θ 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 θ 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 θ 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 θ 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 θ_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 (θ, 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 θ 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 σ^2 as a constant, which now equals to a one-parameter model. Then, we introduce σ^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 σ^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 law 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:

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

2.2.3 Predictive Conditional Posterior Distribution

In this part we will model the predictive model. Conditioned on σ^2, we have

$$\begin{aligned}
\mathbf{P}(\tilde{Y}|y_1, \ldots, y_n, \sigma^2) &= \int \mathbf{P}(\tilde{Y}, \theta|y_1, \ldots, y_n, \sigma^2) \cdot \mathbf{P}(\theta|y_1, \ldots, y_n, \sigma^2)d\theta \\
&= \int \underbrace{\mathbf{P}(\tilde{Y}, \theta|\sigma^2)}_{\text{sampling model: normal}} \cdot \underbrace{\mathbf{P}(\theta|y_1, \ldots, y_n, \sigma^2)}_{\text{conditional prior: normal}} d\theta
\end{aligned}$$

This tells us that, the conditional predictive distribution is also normal with some mean and variance. In the following we are going to find out what they are using the laws we mentioned above.

$$\begin{aligned}
\text{mean} &= \mathbb{E}[\tilde{Y}|\vec{Y}, \sigma^2] \\
&= \mathbb{E}[\mathbb{E}[\tilde{Y}|\theta, \vec{Y}, \sigma^2]|\vec{Y}, \sigma^2] \\
&= \mathbb{E}[\theta|\vec{Y}, \sigma^2] \\
&= \mu_n
\end{aligned}$$

$$\begin{aligned}
\text{Var} &= Var(\tilde{Y}|\vec{Y}, \sigma^2) \\
&= \mathbb{E}[Var(\tilde{Y}|\vec{Y}, \theta, \sigma^2)|\vec{Y}, \sigma^2] + Var(\mathbb{E}[\tilde{Y}|\vec{Y}, \theta, \sigma^2]|\vec{Y}, \sigma^2) \\
&= \mathbb{E}[\sigma^2|\vec{Y}, \sigma^2] + Var(\theta|\vec{Y}, \sigma^2) \\
&= \sigma^2 + \tau_n^2
\end{aligned}$$

This implied that $\tilde{Y}|y_1, \ldots, y_n, \sigma^2 \sim N(\mu_n, \sigma^2 + \tau_n^2)$.

2.2.4 Joint Inference for Mean and Variance

Previously, we derived our work by fixing σ^2. In this section, we will find out what's going to happen if we bring σ^2 back to life. Firstly, it is natural to write the joint posterior as

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

where the first two terms are already known to be $N(\theta, \sigma^2)$ and $N(\mu_0, \tau_0^2)$. It is obvious that if σ^2 also follows normal, then the posterior will also be a normal. However, this is impossible because σ^2 must be non-negative even in the loosest case, where normal distribution is not legitimate.

To address this issue, we decomopse the posterior in a different way and assume a specific dependence of θ on σ^2.

$$\frac{1}{\sigma^2} \sim Gamma(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$$
$$\sigma^2 \sim InvGamma(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$$

We have decomposition:

$$\begin{aligned}\mathbf{P}(\theta, \sigma^2|y_1, \ldots, y_n) &\propto \mathbf{P}(\theta|\vec{Y}, \sigma^2) \cdot \mathbf{P}(\sigma^2|\vec{Y}) \\ &= N(\mu_n, \tau_n^2) \cdot \mathbf{P}(\sigma^2|\vec{Y})\end{aligned}$$

where we can further write the second term in the following way:

$$\begin{aligned}\mathbf{P}(\sigma^2|\vec{Y}) &\propto \mathbf{P}(\vec{Y}, \sigma^2) \cdot \mathbf{P}(\sigma^2) \\ &= \mathbf{P}(\sigma^2) \cdot \int \mathbf{P}(\vec{Y}\theta|\sigma^2) d\theta \\ &= \mathbf{P}(\sigma^2) \cdot \int \mathbf{P}(\vec{Y}|\theta, \sigma^2) \cdot \mathbf{P}(\theta|\sigma^2) d\theta \\ &= InvGamma \int Normal \cdot Normal \; d\theta \\ &= InvGamma(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n}{2})\end{aligned}$$

where $\nu_n = \nu_0 + n$, $\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - y_0)^2$, $\kappa_0 = \frac{\sigma^2}{\tau_0^2}$. Here κ_0 can be interpreted as the prior sample size.

Semi-Conjugate Case

In the previous example, by modeling the normal model to be *normal-inverse-gamma*, we obtained a full-conjugate configuration. We have $\theta|\sigma^2, \vec{Y}$ samely distributed as $\theta|\sigma^2$ as normal, and $\sigma^2|\vec{Y}$ samely distributed as σ^2 as inverse gamma. What's more important is that, this kind of structure allows the parameters to be implicitly connected, i.e. having θ dependent on σ^2.

However, this might not always be the case in real life. In some cases, a priori we may believe that θ and σ^2 are in fact independent to each other, i.e. $\theta = \theta|\sigma^2$. We refer to this situation or this kind of modeling structure as semi-conjugate. While it still allows conjugate behaviors, the core difference compared with full-conjugate is the way we model the distribution of parameters.

2.2.5 Clarification on Some Terms

Many different notations popped in our previous discussion over normal models which can be confusing for first time readers. This section specifically addresses this issue. We will discuss and categorize the notations mentioned above for clarification purposes.

It's almost a mess when we are looking at $\theta, \sigma^2, \mu_0, \mu_n, \tau_0^2, \tau_n^2, \nu_0, \nu_n$ altogether, even with the context. As a matter of fact, they can be taken and categorized in the following ways:

Firstly, θ, σ^2 are the parameters of our true interest. It is the posterior of these two parameters that we truly care about. These two determines the sampling/predictive model.

$$\tilde{Y}|\vec{Y} = \tilde{Y}|\theta, \sigma^2 \sim N(\theta, \sigma^2)$$

Secondly, the (μ_n, τ_n) pair. This notation was introduced when we were studying the behavior when conditioned on fixed σ^2. In other words, they are the parameters of the conditional posterior of θ

$$\theta|\vec{Y}, \sigma^2 \sim N(\mu_n, \tau_n^2)$$

And as for the initial pair (μ_0, τ_0), they were the parameters for the prior

$$\theta|\sigma^2 \sim N(\mu_0, \tau_0^2)$$

Similarly, ν_0 and ν_n are the parameters for the prior and posterior of σ^2 after we reintroduce σ^2 as a random variable.

From my current understanding and the lecture, κ_0 is understood as the prior sample size, which can be confusing since by definition this can be a decimal, and act as a relation between the params of two distribution without any necassary implicit relations.

So Where is the dependency in prior?

This is a natural question to ask if we go back to how we defined the conditional prior $\theta|\sigma^2$. Because in the distribution

$$\theta|\sigma^2 \sim \mathcal{N}(\mu_0, \tau_0^2)$$

both μ_0, τ_0^2 came from expert opinion which *should* be independent from the parameter σ^2.

However, later on when we were computing the posterior, we defined $\kappa_0 = \frac{\sigma^2}{\tau_0^2}$ and reparametrized the conditional prior to be

$$\theta|\sigma^2 \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{\kappa_0})$$

In other words, instead of thinking the experts are giving direct conjecture over the prior's variance, we can interpret it as giving relative relation to the latent σ^2, and this is where the dependency kicks in.

2.2.6 Monte Carlo Simulation and Marginal Posterior

To simulate the joint posterior distribution, since we assumed θ's dependency on σ^2, we first sample σ^2 from an InvGamma distribution, and then sample $\theta|\sigma^2$ from $\mathcal{N}(\mu_n, \tau_n^2)$ which depends on σ^2. This means when we are sampling the parameter pair (θ, σ^2), it follows

$$\mathbf{P}(\theta, \sigma^2) = \mathbf{P}(\sigma^2) \cdot \mathbf{P}(\theta|\sigma^2)$$

This looks nice, but what if we don't care about σ^2 and only want to sample a sequence of $\{\theta_i\}$? Then the above simulation can be expensive since we are running an additional sampling for n times. A faster approach is to directly sample θ from its marginal posterior distribution

$$t(\theta) = \frac{\theta - \mu_n}{\sigma_n/\sqrt{\kappa_n}}|\vec{Y} \sim t_{\nu_0+n}$$

where $\mu_n = \frac{\kappa_0\mu_0 + n\langle y|}{\kappa_0+n}$ and $\sigma_n^2 = \frac{1}{\nu_0+n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu)^2\right]$. In other words, we have the marginal posterior of θ as a t-distribution.

2.2.7 Weak but Improper Priors

To find out how a weak prior affects the posterior, we set both $\kappa_0, \nu_0 \to 0$. But this results in improper priors:

$$\begin{aligned}\mathbf{P}(\theta, \sigma^2) &= \mathbf{P}(\theta|\sigma^2) \cdot \mathbf{P}(\sigma^2) \\ &= \mathcal{N}(\mu_0, \frac{\sigma^2}{\kappa_0}) \cdot IG(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}) \\ &\to \mathcal{N}(\mu_0, \infty) \cdot IG(0, 0)\end{aligned}$$

However, the posterior is still proper since we have

$$\mu_n \to \bar{y} \qquad \sigma_n^2 \to \frac{1}{n}\sum(Y_i - \bar{Y})^2 \qquad \nu_n \to n$$

Let's Get Practical

To solve a real-world problem, we are usually given a belif on what should the parameters be like. One way to encode this information is to ensure that for the joint prior $p(\theta, \sigma^2)$ satisfies the following:

$$\mathbb{E}[\sigma^2] = \sigma_0^2 \qquad \mathbb{E}[\theta] = \mu_0$$

One possible parametrization is :

$$\theta|\sigma^2 \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{n}) \qquad \sigma^2 \sim InvGamma(\frac{n_0+3}{2}, \frac{(n_0+1)\sigma_0^2}{2})$$

It can be verified that this joint distribution indeed satisfies the above requirements.

Chapter 3

Gibbs Sampling

While not much to include, a new chapter is opened for Gibbs sampling due to both of its efficiency especially when the situation is complicated, and its effectiveness which is obtained via Markov Chain. First we will introduce how it's done, and then explain why it's good.

What is going on?

Gibbs sampling was introduced in class as a better replacement for plain Monte Carlo method in high dimensional cases. Say we have a prior distribution of parameters $\{\theta_i\}_{i=1}^{p}$ as $\mathbf{P}(\theta_1, \ldots, \theta_p)$. While we don't know what the joint distribution is, we assume we know the conditional distribution

$$\mathbf{P}(\theta_j|\theta_1, \ldots, \hat{\theta_j}, \ldots, \theta_p), \quad \forall\, 1 \leq j \leq p$$

where $\hat{\theta_j}$ means the j-th parameter is omitted. Then we can start by selecting **ANY** initial values $(\theta_1^{(0)}, \ldots, \theta_p^{(0)})$ as long as they are in the domain. After that, we update each parameter/entry at a time while fixing the resting of the values fixed. In other words, we would firstly sample from:

$$\theta_1|\hat{\theta_1} \sim \mathbf{P}(\theta_1|\hat{\theta_1})$$

We denote the this observation as $\theta_1^{(1)}$ and right now the entire tuple of parameters has experienced the following change

$$(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)}) \to (\theta_1^{(1)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)})$$

After that, we focus on θ_2 and sample one observation from the corresponding distribution

$$\theta_2|\hat{\theta_2} \sim \mathbf{P}(\theta_2|\hat{\theta_2})$$

And now we plug the new sampled point $\theta_2^{(1)}$ in the tuple and have

$$(\theta_1^{(1)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)}) \to (\theta_1^{(1)}, \theta_2^{(1)}, \theta_3^{(0)}, \ldots, \theta_p^{(0)})$$

We continue the scan until the p-th parameter is updated and the result looks like

$$(\theta_1^{(1)}, \ldots, \theta_p^{(1)})$$

Till this point, we can say that we have finished an entire scan. Then we can continue this pattern until we have reached the diresed number of iterations. We can interpret each step of Gibbs sampling as moving along one of the axes in the parameter space while the other parameters are fixed.

Why is it good?

From my point of view, Gibbs sampling is beneficial mostly because it's efficiency in generating samples in the parameter space. This can be broken down to two main reasons.

Complicated even Intractable Joint Prior Distribution

Sometimes, especially when the number of parameters goes to a crazy level, the joint prior $\mathbf{P}(\theta_1, \ldots, \theta_p)$ can be hard to obtain. Even if we can model this joint distribution in a hierarchical way(conditioned on the other parameters), for each data points obtained through Monte Carlo, we have to sampling at least p times to get all of the coordinates $(\theta_1, \ldots, \theta_p)$. In other words, we have to firstly sample from θ_p, and then $\theta_{p-1}|\theta_p$, all the way to $\theta_1|\theta_2, \ldots, \theta_p$ in order to get a data point from the parameter space Θ.

In contrast, when we are doing Gibbs sampling, although it seems like we are doing things that are quite similar, we actually finds a data point each time we update an **ENTRY**. In other words, $(\theta_1^{(1)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)})$ is a new data point. So we only need to sample from one distribution to get a parameter data point instead of p times just like Monte Carlo.

Curse of Dimension

Another reason is derived from the curse of dimensionality. The volume of an object grows exponentially with the dimension. This significantly increases the parameter space, which means we need a lot more points so as to get an plausible interpretation/simulation on the real joint prior distribution. And again, from the arguement above, the efficiency of Gibbs sampling makes it suitable for high-dimensional filthy simulations.

Any Downside?

Nothing is perfect, Gibbs sampling's problem I think mainly embeds in it's sensitivity to initial configurations. Although as a Markov Chain it converges to the joint distribution as the number of iteration goes to infinity, still, just like the case of CLT, there's no way a priori we can know what is the sufficient number of iterations to obtain a satisfactory simulation. On the contrary, although expensive, sampling directly from the joint distribution is guaranteed to be *objective*.

Chapter 4

Multivariate Normal Model

Now we have seen a little bit about how Bayesian models work when the sampling random variable maps to a 1-dimensional real number. In this chapter, through the perspective of multivariate normal, we take a tour around the world when

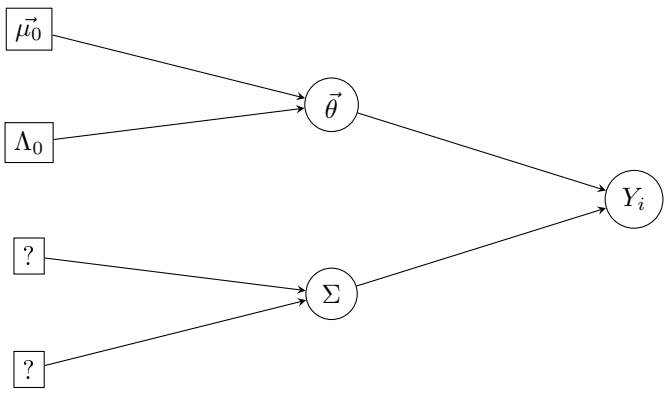
$$X : \mathcal{F} \to \mathbb{R}^p$$

4.1 Semi-Conjugate Prior

Semi-conjugate modeling requires parameters to be independent to each other, which is easier to begin with. In high dimension cases, each data point/sample looks like this

$$Y_i = (Y_{i1}, \ldots, Y_{ip})^T \in \mathbb{R}^p$$

As a semi-conjugate case, the following plot illustrates how will we structure this model. While both $\vec{\theta}, \Sigma$ depends on some latent variables, but they are not implicitely connected in any way. Firstly, we will talk about how we are going to model the prior of $\vec{\theta}$ and leave Σ's latent for now.



4.1.1 θ as Multivariate Normal

In the 1-dimension case, we modeled θ as a normal. As a high-dimensional generalization, we choose to model $\vec{\theta}$ as multivariate normal:

$$\vec{\theta} \sim MVN(\mu_0, \Lambda_0)$$

If we dive deeper into its pdf, we have the following:

$$\begin{aligned}
p(\vec{\theta}) &= (2\pi)^{-p/2}|\Lambda_0|^{-1/2} \exp\big(-\frac{1}{2}(\theta-\mu_0)^T\Lambda_0^{-1}(\theta-\mu_0)\big) \\
&\propto \exp\big(-\frac{1}{2}\theta^T\Lambda_0^{-1}\theta + \frac{1}{2}\theta^T\Lambda_0^{-1}\mu_0 + \frac{1}{2}\mu_0^T\Lambda_0^{-1}\theta - \frac{1}{2}\mu_0^T\Lambda_0^{-1}\mu_0\big) \\
&\propto \exp\big(-\frac{1}{2}\theta^T\Lambda_0^{-1}\theta + \theta^T\Lambda_0^{-1}\mu_0\big) \\
&= \exp\big(-\frac{1}{2}\theta^T A_0\theta + \theta^T b_0\big)
\end{aligned}$$

where we reparametized $A_0 = \Lambda_0^{-1}$ and $b_0 = A_0\mu_0 = \Lambda_0^{-1}\mu_0$.

As a direct consequence of the reparametrization, we have

$$\vec{\theta} \sim MVN(A_0^{-1}b_0, A_0^{-1})$$

Joint Sampling Model

Now, assuming we have known Σ, then our sampling model can be written as

$$\begin{aligned} p(\vec{\theta}|\vec{Y}, \theta, \Sigma) &= \Pi_{i=1}^{n}(2\pi)^{-p/2}|\Sigma|^{-\frac{1}{2}} \exp\big(-\frac{1}{2}(Y_i - \vec{\theta})^T\Sigma^{-1}(Y_i - \vec{\theta})\big) \\ &\propto \exp\big(-\frac{n}{2}\theta^T\Sigma^{-1}\theta + \theta^T\Sigma^{-1}\sum_{i=1}^{n}Y_i - \frac{1}{2}\sum_{i=1}^{n}(Y_i^T\Sigma^{-1}Y_i)\big) \\ &\propto \exp\big(-\frac{1}{2}\theta^T A_1\theta + \theta^T b_1\big) \end{aligned}$$

where we reparametized $A_1 = n\Sigma^{-1}$ and $b_1 = \Sigma^{-1}\sum_{i=1}^{n}Y_i = n\Sigma^{-1}\bar{Y}$

Posterior

With the sampling and prior, we now formulate the posterior:

$$\begin{aligned} p(\vec{\theta}|\vec{Y}, \Sigma) &\propto p(\vec{Y}|\theta, \Sigma) \cdot p(\theta) \\ &\propto \exp\big(-\frac{1}{2}\theta^T A_1\theta + \theta^T b_1\big) \cdot \exp\big(-\frac{1}{2}\theta^T A_0\theta + \theta^T b_0\big) \\ &= \exp\big(-\frac{1}{2}\theta^T(A_0 + A_1)\theta + \theta^T(b_0 + b_1)\big) \end{aligned}$$

Notice the first line where we used Bayes' rule, we have $p(\theta|\Sigma) = p(\theta)$ because we are modeling the semi-conjugate case. In conclusion, we have obtained the posterior as

$$\vec{\theta}|\vec{Y}, \Sigma \sim MVN(A_n^{-1}b_n, A_n^{-1})$$

where $A_n = A_0 + A_1 = \Lambda_0^{-1} + n\Sigma^{-1}$ and $b_n = b_0 + b_1 = \Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{Y}$

4.1.2 Σ as Inverse-Wishart

In 1-dimension case, we would model the variance term of normals as χ_n^2. Its high dimensional resemblence is the Wishart distribution.

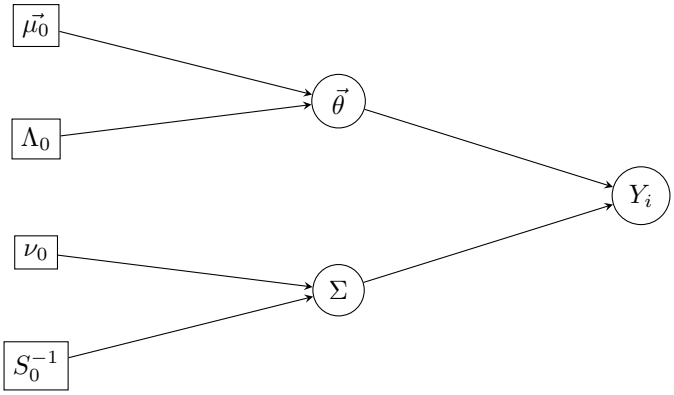
1. Firstly we sample $z_i \sim MVN(\vec{0}, \Phi_0)$, $i = 1, \ldots, \nu_0$
2. Then we have $\sum_{i=1}^{\nu_0} z_i z_i^T \sim Wishart(\nu_0, \Phi_0)$

One thing worth noticing is that we actually can sample PSD fro Wishart, for example the all-zero matrix. However, we are not very concerned with this setback because firstly it's a zero-measure set, secondly we usually tolerates the covariance to be PSD in the sense that there exists colliniarity between variables. Generally speaking, it's sensible to model something that is bound to be PD as Wishart distributed. In conclusion, we have

$$\sum_{i=1}^{n} z_i z_i^T \sim Wishart(\nu_0, S_0^{-1})$$

$$\Sigma = (\sum_{i=1}^{n} z_i z_i^T)^{-1} \sim InvWishart(\nu_0, S_0^{-1})$$

In other words, we have completed the dependency plot


$\vec{\mu_0}$
Λ_0
ν_0
S_0^{-1}
$\vec{\theta}$
Σ
Y_i

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.

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