

## Lecture Notes 3: Scale Mixture Distribution

Professor: Zhihua Zhang

**Notice:** In this lecture note,  $X \sim \mathcal{N}(a, b)$  means that  $\mu = a, \sigma^2 = b$ . Prof. Zhang sometimes means  $\mu = a, \sigma = b$  in the class. So the notation or results may be a little different with your notes.

## 2.4 Scale Mixture Distribution

We will show several distributions can be seen as the scale mixture of distributions, which is defined as follows,

$$\begin{aligned} X &\sim F(\theta) \\ \theta &\sim G(\lambda) \end{aligned}$$

, So,  $T(x) = \int_{\theta} F(\theta)G(\lambda)d\theta$  can be seen as a scale mixture of  $F$ , where the scale has distribution  $G$ .

### 2.4.1 Student's t-distribution

The Student's t-distribution is a scale mixture of Gaussian distribution, where the scale has a Gamma distribution. Let  $X \sim \mathcal{N}(\mu, \frac{\sigma^2}{r})$ ,  $r \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$ , then the integral will be:

$$\begin{aligned} &\int_0^{\infty} \frac{r^{1/2}}{\sqrt{2\pi}\sigma} e^{-\frac{r(x-\mu)^2}{2\sigma^2}} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} r^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}r} dr \\ &= \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})\sigma\sqrt{2\pi}} \int_0^{\infty} r^{\frac{\nu+1}{2}-1} e^{-\frac{r}{2}(\frac{(x-\mu)^2}{\sigma^2} + \nu)} dr \\ &= \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{2\pi}\Gamma(\frac{\nu}{2})} \left[ \frac{(x-\mu)^2}{\sigma^2} + \nu \right]^{-\frac{\nu+1}{2}} \\ &= \frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left[ \frac{(x-\mu)^2}{\nu\sigma^2} + 1 \right]^{-\frac{\nu+1}{2}} \\ &= t_{\nu}(\mu, \sigma^2) \end{aligned}$$

Note that during the integral, we use a mathematical trick. Since we have

$$\int_0^{\infty} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = 1$$

from Gamma distribution, so we can get  $\int_0^{\infty} x^{\alpha-1} e^{-\beta x} dx = \frac{\Gamma(\alpha)}{\beta^{\alpha}}$ . This trick will be often used in the follows.

### 2.4.2 Laplace Distribution

Laplace distribution is:

$$f(x) = \frac{1}{4\sigma} \exp\left(-\frac{|x - \mu|}{2\sigma}\right)$$

Let we see  $2\sigma$  as  $\sigma$  for convenience, that is:

$$f(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x - \mu|}{\sigma}\right)$$

The Laplace distribution is a scale mixture of Gaussian distribution, where the scale has a exponential distribution. Let  $X \sim \mathcal{N}(\mu, r)$ ,  $r \sim \text{Exponential}(\frac{1}{2\sigma^2})$ , then we can get the mixture distribution:

$$\begin{aligned} & \int_0^\infty \frac{1}{\sqrt{2\pi r}} e^{-\frac{(x-\mu)^2}{2r}} \frac{1}{2\sigma^2} e^{-\frac{r}{2\sigma^2}} dr \\ &= \frac{1}{2\sigma^2 \sqrt{2\pi}} \int_0^\infty r^{\frac{1}{2}-1} e^{-\frac{1}{2}\left(\frac{(x-\mu)^2}{r} + \frac{r}{\sigma^2}\right)} dr \\ &= \frac{1}{2\sigma^2 \sqrt{2\pi}} \frac{2K_{1/2}\left(\sqrt{\frac{1}{\sigma^2}(x-\mu)^2}\right)}{\left(\frac{1}{\sigma^2(x-\mu)^2}\right)^{\frac{1}{4}}} \\ &= \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}} \end{aligned}$$

The integral term is a integral of generalized inverse Gaussian distribution,  $\text{GIG}(\frac{1}{2}, \frac{1}{\sigma^2}, (x - \mu)^2)$ .

### 2.4.3 Negative Binomial Distribution

Negative Binomial Distribution is a scale of Poisson distribution, where the scale has a Gamma distribution. Let  $K \sim \text{Poisson}(\lambda)$ ,  $\lambda \sim \text{Gamma}(r, \frac{1-p}{p})$ , then we can get the mixture distribution:

$$\begin{aligned} & \int_0^\infty \frac{\lambda^k}{k!} e^{-\lambda} \frac{\lambda^{r-1} e^{-\frac{1-p}{p}\lambda}}{\Gamma(r) \left(\frac{p}{1-p}\right)^r} d\lambda \\ &= \frac{1}{k! \Gamma(r) \left(\frac{p}{1-p}\right)^r} \int_0^\infty \lambda^{k+r-1} e^{-\frac{\lambda}{p}} d\lambda \\ &= \frac{\Gamma(r+k) p^k (1-p)^r}{\Gamma(k) \Gamma(r)} \\ &= \binom{k+r-1}{k} p^k (1-p)^r \end{aligned}$$

**Homework 1:**  $\sum_{k=0}^{\infty} \text{Gamma}(x|k + \rho + 1, \beta) \text{Poisson}(k|\lambda)$ ,  $\rho$  is a constant,  $\rho > -1$ .

### 3 Statistical Inference (I)

#### 3.1 Jeffreys Prior

In order to show Jeffrey prior, we first introduce **Fisher information**. In mathematical statistics, the Fisher information is a way of measuring the amount of information that an observable random variable  $X$  carries about an unknown parameter  $\theta$  upon which the probability of  $X$  depends.

The probability function for  $X$ , which is also the likelihood function for  $\theta$ , is a function  $f(X; \theta)$ ; it is the probability mass (or probability density) of the random variable  $X$  conditional on the value of  $\theta$ . Then we define Fisher information:

**Definition 3.1.** *Fisher Information:*

$$I(\theta) = \mathbb{E}\left(\left(\frac{\partial \log f(x; \theta)}{\partial \theta}\right)^2\right)$$

**Lemma 3.1.** *If  $\log f(x; \theta)$  is twice differentiable with respect to  $\theta$  and under certain regularity conditions, then*

$$I(\theta) = -\mathbb{E}\left(\frac{\partial^2 \log f}{\partial \theta^2}\right)$$

*Proof.* first

$$\begin{aligned}\frac{\partial^2 \log f}{\partial \theta^2} &= \frac{\partial}{\partial \theta} \left( \frac{\frac{\partial f}{\partial \theta}}{f} \right) \\ &= \frac{\frac{\partial^2 f}{\partial \theta^2}}{f} - \frac{(\frac{\partial f}{\partial \theta})^2}{f^2} \\ &= \frac{\frac{\partial^2 f}{\partial \theta^2}}{f} - \left( \frac{\partial \log f}{\partial \theta} \right)^2\end{aligned}$$

then

$$\begin{aligned}\mathbb{E}\left(\frac{\partial^2 \log f}{\partial \theta^2}\right) &= \int \frac{\partial^2 \log f}{\partial \theta^2} f dx \\ &= \int \frac{\partial^2 f}{\partial \theta^2} dx - I(\theta)\end{aligned}$$

if  $\int \frac{\partial^2 f}{\partial \theta^2} dx = \frac{\partial^2}{\partial \theta^2} \int f dx = \frac{\partial^2}{\partial \theta^2} 1 = 0$  (the certain condition), then we had proved the lemma.  $\square$

**Definition 3.2.** *Jeffreys prior is defined in terms of Fisher information*

$$p(\theta) \propto \sqrt{I(\theta)}$$

**Remark:** It has the key feature that it is **invariant under reparametrization** of parameter  $\theta$ . For an alternate parametrization  $\varphi$  we can derive

$$p(\varphi) \propto \sqrt{I(\varphi)}$$

from

$$p(\theta) \propto \sqrt{I(\theta)}$$

where  $\theta$  and  $\varphi$  exist a one-to-one mapping.

*Proof.*

$$\begin{aligned} p(\varphi) &= p(\theta) \left| \frac{d\theta}{d\varphi} \right| \propto \sqrt{I(\theta) \left( \frac{d\theta}{d\varphi} \right)^2} \propto \sqrt{\mathbb{E} \left( \left( \frac{d \log f}{d\theta} \right)^2 \right) \left( \frac{d\theta}{d\varphi} \right)^2} \\ &\propto \sqrt{\mathbb{E} \left( \left( \frac{d \log f}{d\theta} \frac{d\theta}{d\varphi} \right)^2 \right)} = \sqrt{\mathbb{E} \left( \left( \frac{d \log f}{d\varphi} \right)^2 \right)} \propto \sqrt{I(\varphi)} \end{aligned}$$

□

**Example 3.1.**  $X \sim \mathcal{N}(\mu, \sigma^2)$ .

**Case 1:** Fix  $\sigma$ , the only parameter is  $\mu$ . The likelihood is:

$$f(X|\mu) \propto \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

so

$$\log f \propto -\left(\frac{x - \mu}{\sigma}\right)^2$$

So we can get:

$$\begin{aligned} I(\mu) &= \mathbb{E} \left[ \left( \frac{(x - \mu)^2}{\sigma^2} \right)^2 \right] \\ &= \frac{\mathbb{E}(x - \mu)^2}{\sigma^4} \\ &= \frac{1}{\sigma^2} \end{aligned}$$

Thus the Jeffreys prior  $p(\mu) \propto \sqrt{I(\mu)} = \frac{1}{\sigma}$ . As  $\sigma$  is fixed, so  $p(\mu) \propto 1$ .

**Remark:** Although  $p(\mu) = 1$  is a improper prior, as  $\int_{-\infty}^{\infty} 1dx = \infty$ , the posterior is proper. The prior is also called **uninformative prior**.

**Case 2:** Fix  $\mu$ , the only parameter is  $\sigma$ . For convenience, let  $\tau = \frac{1}{\sigma^2}$ . So  $f(x) = \frac{\tau^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{\tau(x-\mu)^2}{2}}$ . The likelihood is denoted by  $f(\tau)$ :

$$\begin{aligned} f(\tau) &= \frac{\tau^{\frac{1}{2}}}{\sqrt{2\pi}} \exp\left(-\frac{\tau(x-\mu)^2}{2}\right) \\ \implies \log f &\propto \frac{1}{2} \log \tau - \frac{\tau}{2} (x - \mu)^2 \\ \implies \frac{\partial \log f}{\partial \tau} &\propto \frac{1}{2\tau} - \frac{(x-\mu)^2}{2} \end{aligned}$$

Hence,

$$\begin{aligned}
I(\tau) &= \mathbb{E} \left[ \left( \frac{\partial \log f}{\partial \tau} \right)^2 \right] \\
&= \mathbb{E} \left[ \frac{1}{4} \left( \frac{1}{\tau} - (x - \mu)^2 \right)^2 \right] \\
&= \mathbb{E} \left[ \frac{1}{4\tau^2} - \frac{(x - \mu)^2}{2\tau} + \frac{(x - \mu)^4}{4} \right] \\
&= \frac{1}{4\tau^2} - \frac{1}{2\tau^2} + \frac{1}{4} \mathbb{E}(x - \mu)^4 \\
&= \frac{1}{4\tau^2} - \frac{1}{2\tau^2} + \int_{-\infty}^{\infty} \frac{1}{4} (x - \mu)^4 \mathcal{N}(x|\mu, \tau^{-1}) dx \\
&= \frac{1}{2\tau^2}
\end{aligned}$$

where the integral can be computed from variance :

$$\begin{aligned}
&\int (x - \mu)^2 \frac{\tau^{\frac{1}{2}}}{\sqrt{2\pi}} \exp\left(-\frac{\tau(x - \mu)^2}{2}\right) dx = \tau^{-1} \\
\Rightarrow &\int (x - \mu)^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\tau(x - \mu)^2}{2}\right) dx = \tau^{-\frac{3}{2}} \\
&\text{(taking the derivate of both side)} \\
\Rightarrow &\int (x - \mu)^4 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\tau(x - \mu)^2}{2}\right) dx = 3\tau^{-\frac{5}{2}} \\
\Rightarrow &\int (x - \mu)^4 \frac{\tau^{\frac{1}{2}}}{\sqrt{2\pi}} \exp\left(-\frac{\tau(x - \mu)^2}{2}\right) dx = 3\tau^{-2} \\
\Rightarrow &\mathbb{E}((x - \mu)^4) = \frac{3}{4\tau^2}
\end{aligned}$$

So Jeffreys prior is  $\pi(\tau) \propto \frac{1}{\tau}$ . Note that  $p(\sigma) = \pi(\tau)|d\tau/d\theta|$ , hence,

$$\begin{aligned}
p(\sigma) &\propto \left| \frac{1}{\tau} - 2\sigma^{-3} \right| \\
&\propto \left| \sigma^2 - 2\sigma^{-3} \right| \\
&\propto \frac{1}{\sigma}
\end{aligned}$$

**Homework 2:** Compute the following integrals:

1.  $u_0 = \int_{-\infty}^{\infty} \Phi(x) \mathcal{N}(x|\mu, \sigma^2) dx$
2.  $u_1 = \int_{-\infty}^{\infty} \Phi(x) \mathcal{N}(x|\mu, \sigma^2) x dx$
3.  $u_2 = \int_{-\infty}^{\infty} \Phi(x) \mathcal{N}(x|\mu, \sigma^2) (x - m_1) dx$

where  $\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$

**Example 3.2.**  $X \sim \text{Poisson}(n; \lambda)$

$$\log f = -\lambda + n \log \lambda$$

Fisher information is:

$$\begin{aligned} I(\lambda) &= \mathbb{E} \left[ \left( \frac{n}{\lambda} - 1 \right)^2 \right] \\ &= 1 + \frac{\mathbb{E}(n^2)}{\lambda^2} - 2 \\ &= \frac{\lambda + 1}{\lambda} - 1 \\ &= \frac{1}{\lambda} \end{aligned}$$

So Jeffreys prior is:

$$p(\lambda) \propto \sqrt{\frac{1}{\lambda}}.$$

**Homework 3:**  $f(x; \theta) = \theta^x(1 - \theta)^{1-x}$ ,  $0 < \theta < 1$ .

1. Compute Jeffreys prior about  $\theta$ .
2. If  $\theta = \sin^2 \alpha$ , compute Jeffreys prior about  $\alpha$ .

### 3.2 Compute Posterior Probability

Assume we have a model  $x = \theta + \epsilon$ , where  $x$  is data which we observed or predict,  $\theta$  is the parameters,  $\epsilon \sim \mathcal{N}(0, \tau)$  is the error term. So given  $\theta$ ,  $X \sim \mathcal{N}(\theta, \tau)$ . When we use MAP(maximum a posteriori) to estimate parameter  $\theta$ , we will get  $p(\theta|x) \propto p(x|\theta)p(\theta)$ . We will discuss this problem under several different conditions in the following.

**Case 1** Fix  $\tau$ . The only parameter is  $\theta$ . And we set the prior about  $\theta$  is  $\mathcal{N}(\theta|0, \lambda)$ . So

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &\propto \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{(x-\theta)^2}{2\tau}} \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{\theta^2}{2\lambda}} \\ &\propto \frac{1}{2\pi\sqrt{\tau\lambda}} e^{-\frac{1}{2}[(\frac{1}{\tau} + \frac{1}{\lambda})(\theta - \frac{\lambda x}{\tau + \lambda})^2 + \frac{x^2}{\tau + \lambda}]} \\ &\propto \mathcal{N}\left(\frac{\lambda x}{\lambda + \tau}, \frac{\lambda\tau}{\lambda + \tau}\right) \end{aligned}$$

Then we can get the estimation about  $\theta$  from MAP,  $\hat{\theta} = \frac{\lambda x}{\lambda + \tau}$ .

**Case 2** Let  $\theta$  and  $\tau$  both be parameters. In order to get MAP, we may make three solutions.

**Solution 1** Assume that  $\tau, \lambda$  are independent.  $p(\theta, \tau) = p(\theta)p(\tau)$  and we use a new  $\tau$  here,  $\tau = \tau_{old}^{-1}$  for convenience of computing (It is different with the  $\tau$  used in case 1). Suppose  $\tau \sim \text{Gamma}(\alpha/2, \beta/2)$

$$\begin{aligned} p(\theta, \tau|x) &\propto p(x|\theta, \tau)p(\theta, \tau) = p(x|\theta, \tau)p(\theta)p(\tau) \\ &\propto \frac{1}{\sqrt{2\pi}}\tau^{\frac{1}{2}} \exp\left(-\frac{\tau(x-\theta)^2}{2}\right) \frac{\lambda^{\frac{1}{2}}}{\sqrt{2\pi}} \exp\left(-\frac{\lambda\theta^2}{2}\right) \left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}} \exp\left(-\frac{\beta\tau}{2}\right) \frac{\tau^{\frac{\alpha}{2}-1}}{\Gamma(\alpha/2)} \end{aligned}$$

To get maximum posterior, let

$$L = \tau^{\frac{\alpha+1}{2}-1} \lambda^{\frac{1}{2}} \exp\left(-\frac{\tau}{2}((x-\theta)^2 + \beta) - \frac{\lambda\theta^2}{2}\right)$$

then

$$\ln L = \left(\frac{\alpha+1}{2} - 1\right) \ln \tau - \frac{\tau}{2}((x-\theta)^2 + \beta) + \frac{1}{2} \ln \lambda - \frac{\lambda\theta^2}{2}$$

Let

$$Q = -2 \ln L = -(\alpha-1) \ln \tau + \tau((x-\theta)^2 + \beta) - \ln \lambda + \lambda\theta^2$$

We need to solve equations:

$$\begin{cases} \frac{\partial Q}{\partial \theta} = -2\tau(x-\theta) + 2\lambda\theta = 0 \\ \frac{\partial Q}{\partial \tau} = \frac{1-\alpha}{\tau} + (x-\theta)^2 + \beta = 0 \end{cases}$$

It is a difficult problem to solve, especially when  $\theta$  is a vector.

**Remark :** One way to solve the problem above is to compute one parameter, for example  $\theta$ , while fixing the other parameter, i.e.  $\tau$ . Then fix  $\theta$ , compute  $\theta$ . Hold on until they get convergent. Well, then we need to think about the convergence problem.

**Solution 2** Assume  $p(\theta, \tau) = p(\theta|\tau)p(\tau)$ , and  $p(\theta|\tau) \sim \mathcal{N}(0, (\lambda\tau)^{-1})$ , suppose  $\tau \sim \text{Gamma}(\alpha/2, \beta/2)$ , follow the same steps in solution 1, we get:

$$\begin{aligned} p(\theta, \tau|x) &\propto p(x|\theta, \tau)p(\theta|\tau)p(\tau) \\ &\propto \frac{\tau^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{\tau(x-\theta)^2}{2}} \frac{(\lambda\tau)^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{\lambda\tau\theta^2}{2}} \left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}} e^{-\frac{\beta\tau}{2}} \frac{\tau^{\frac{\alpha}{2}-1}}{\Gamma(\alpha/2)} \end{aligned}$$

Then the corresponding  $L$  is given by:

$$L = \tau^{\frac{\alpha+1}{2}-1} (\lambda\tau)^{\frac{1}{2}} \exp\left(-\frac{\tau}{2}((x-\theta)^2 + \beta) - \frac{\lambda\tau\theta^2}{2}\right).$$

Then the  $Q$  is

$$Q = -2 \ln L = -(\alpha-1) \ln \tau + \tau((x-\theta)^2 + \beta) - \ln(\lambda\tau) + \lambda\tau\theta^2$$

To get the estimation of  $\theta$  and  $\tau$ , we need to solve:

$$\begin{cases} \frac{\partial Q}{\partial \theta} = 0 \\ \frac{\partial Q}{\partial \tau} = 0 \end{cases}$$

Then we will get:

$$\begin{cases} -\tau(x - \theta) + \tau\lambda = 0 \\ \beta - \frac{\alpha-1}{\tau} - \frac{1}{\tau} + (x - \theta)^2 + \lambda\theta^2 = 0 \end{cases}$$

We can easily solve  $\theta$  and  $\tau$ . It is called **decouple**.

**Solution 3** From the two sub-cases above, we can find the major problem is computing complexity. Another problem will occurs if there are too many hyper-parameters. As we need to search the best hyper-parameters by grid search. So if there are 2 hyper-parameters, the search space is 2-dimensional. If there are 3 hyper-parameters, the search space is 3-dimensional... It will cost too much time when the search space is high dimensional.

Simply, we can give an uninformative prior to  $\tau$ ,  $p(\tau) \propto 1$ . Or we can consider Jeffreys prior for  $\tau$ . According to Example 3.1, case 2, we get  $p(\tau) \propto \frac{1}{\tau}$ . So we will have:

$$\begin{aligned} p(\theta, \tau|x) &\propto p(x|\theta, \tau)p(\theta|\tau)p(\tau) \\ &\propto \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{(x-\theta)^2}{2\tau}} \frac{1}{\sqrt{2\pi\lambda\tau}} e^{-\frac{\theta^2}{2\lambda\tau}} \frac{1}{\tau} = L \end{aligned}$$

Removing constants, we will get:

$$Q = -2 \ln L = 3 \ln \tau + \ln(\lambda\tau) + \frac{(x - \theta)^2}{\tau} + \frac{\theta^2}{\lambda\tau}$$

Then according to  $\frac{\partial Q}{\partial \theta} = 0$  and  $\frac{\partial Q}{\partial \tau} = 0$ , we will get the followings:

$$\begin{cases} -2(x - \theta) + \frac{2\theta}{\lambda} = 0 \\ \frac{3}{\tau} + \frac{1}{\tau} - \frac{1}{\tau^2} \left( \frac{(x-\theta)^2}{2} + \frac{\theta^2}{\lambda} \right) = 0 \end{cases}$$

We can see it is easy to solve and there is no extra parameter. (Solution 2 has two new parameters,  $\alpha$  and  $\beta$ )