Lecture Notes 3: Scale Mixture Distribution

Professor: Zhihua Zhang

3.1 Scale Mixture Distribution

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

$$X \sim F(\theta)$$
$$\theta \sim G(\lambda)$$

, 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.

3.1.1 Student's t-distribution

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

$$\begin{aligned}
&\int_0^\infty \frac{r^{-1/2}}{\sqrt{2\pi}\sigma} e^{-\frac{r(x-\mu)^2}{2\sigma^2}} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} r^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}r} dr \\
=&\frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})\sigma\sqrt{2\pi}} \int_0^\infty r^{\frac{\nu}{2}-\frac{3}{2}} e^{-\frac{r}{2}(\frac{(x-\mu)^2}{\sigma^2}+\frac{\nu}{2})} dr \\
=&\frac{\nu^{\frac{\nu}{2}}\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\pi}\Gamma(\frac{\nu}{2})} \left[\frac{(x-\mu)^2}{\sigma^2} + \frac{\nu}{2}\right]^{\frac{\nu+1}{2}}
\end{aligned}$$

Note that during the integral, we use a math trick. Since we know $\int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = 1$ 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.

3.1.2 Laplace Distribution

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

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

3.1.3 Negative Binomial Distribution

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

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

Homework 1. $\sum\limits_{k=0}^{\infty} Gamma(x|k, \beta) Poisson(k|\lambda)$.

3.2 Statistical Inference (I)

3.2.1 Jeffrey 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 θ upon which the probability of X depends.

Assume we have a model for random variable X, for example $\mathbb{P}(X|\theta)$. $\mathbb{P}(X|\theta)$ can be seen as a joint function of x and θ. Let $f(x, \theta) = \mathbb{P}(X|\theta)$. Then Fisher information of X about θ is given by:

$$
\begin{aligned}
I(\theta) =&\mathbb{E}[(\frac{\partial \log f(x,\theta)}{\partial \theta})^2] \\
=&\int \left(\frac{\partial \log f(x,\theta)}{\partial \theta}\right)^2 f(x,\theta) d\theta
\end{aligned}
$$

Lemma 3.1 *Under certain condition,*

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

Proof.

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

So,

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

□

Now let we go to see Jeffrey prior. When we do MAP(maximum a posteriori), we usually meet $\mathbb{P}(\theta|X) \propto \mathbb{P}(X|\theta)BP(\theta)$. Usually $\mathbb{P}(X|\theta)$ is easy to get, but $\mathbb{P}(\theta)$ (prior) needs our hypothesis. How to choose hypothesis? If we set a prior with hyper-parameter, the training process will be difficult. Jeffrey prior tells us how to choose hypothesis:

$$\mathbb{P}(\theta) \propto \sqrt{I(\theta)}$$

Remark: Jeffrey prior has a property called **invariant under reparameterization**, which means if we replace θ with φ, and there is a one to one rejection between θ and φ. Then we can get:

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

Example 3.1 $X \sim N(\mu, \sigma^2)$.

Case 1: Fix σ, the only parameter is μ. So we can get:

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

So we can get Jeffrey prior $\mathbb{P}(\mu) \propto \sqrt{I(\mu)} = \frac{1}{\sigma}$. As σ is fixed, then $\mathbb{P}(\mu) \propto 1$.

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

Case 2: Fix μ, the only parameter is σ. For convenience, let $\tau = \frac{1}{\sigma^2}$. So $f(x) = \frac{\tau^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{\tau(x-\mu)^2}{2}}$. Then we can get Fisher information:

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

So Jeffrey prior is $\mathbb{P}(\tau) \propto \sqrt{I(\tau)}$.

Homework 2: Compute the following integrals:

1. $m_0 = \int_{-\infty}^{\infty} \Phi(x) N(x|\mu, \sigma^2) dx$
2. $m_1 = \int_{-\infty}^{\infty} \Phi(x) N(x|\mu, \sigma^2) x dx$
3. $m_2 = \int_{-\infty}^{\infty} \Phi(x) N(x|\mu, \sigma^2)(x - m_1) dx$

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

Example 3.2 $X \sim Poisson(\lambda)$

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 Jeffrey prior is:

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

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

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

3.2.2 Problem: $X = \theta + \epsilon$

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

Case 1. Fix τ, or let it be hyper-parameter. The only parameter is θ. And we set the prior about θ is $N(\theta|0, \lambda)$. So

$$\begin{aligned}
\mathbb{P}(\theta|x) &\propto \mathbb{P}(x|\theta)\mathbb{P}(\theta) \\
&= \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{(x-\theta)^2}{2\tau}} \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{\theta^2}{2\lambda}} \\
&= \frac{1}{2\pi\sqrt{\tau\lambda}} e^{-\frac{1}{2}[(\frac{1}{\tau}+\frac{1}{\lambda})(\theta-\frac{\lambda x}{\tau+\lambda})^2+\frac{x^2}{\tau+\lambda}]}
\end{aligned}$$

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

Case 2. Let θ and τ both be parameters. In order to get MAP, we can make three hypothesis.

case 2.1. Assume θ and τ are independent, then $\mathbb{P}(\theta, \tau) = \mathbb{P}(\theta)\mathbb{P}(\tau)$. Let θ's prior be $\theta \sim N(0, \lambda)$, τ's prior be $\tau \sim Gamma(\alpha, \beta)$. So

$$\begin{aligned}
\mathbb{P}(\theta, \tau|X) &\propto \mathbb{P}(X|\theta, \tau)\mathbb{P}(\theta)\mathbb{P}(\tau) \\
&= \frac{\tau^{-\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2\tau}} \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{\theta^2}{2\lambda}} \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau}
\end{aligned}$$

In order to get the maximum, it's equivalent to compute the minimum log. Let $L = \log \mathbb{P}(X|\theta, \tau)\mathbb{P}(\theta)\mathbb{P}(\tau)$, remove the constant, we can get:

$$L = \beta\tau + \frac{1}{2}\left[\frac{(x-\theta)^2}{\tau} + \frac{\theta^2}{\lambda}\right] - (\alpha - \frac{3}{2}) \log \tau$$

To get the estimate of θ and τ, we need to solve:

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

Then we will get:

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

It is a difficult problem to solve, especially when θ is a vector.

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

case 2.2. Assume the conditional prior of $\theta|\tau$ is $\theta|\tau \sim N(0, \lambda\tau)$, the prior of τ is $\tau \sim \Gamma(\alpha, \beta)$ as case 2.1. So

$$\begin{aligned}
\mathbb{P}(\theta, \tau|X) &\propto \mathbb{P}(X|\theta, \tau)\mathbb{P}(\theta|\tau)\mathbb{P}(\tau) \\
&= \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{(x-\theta)^2}{2\tau}} \frac{1}{\sqrt{2\pi\lambda\tau}} e^{-\frac{\theta^2}{2\lambda\tau}} \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau}
\end{aligned}$$

Then the corresponding L is given by:

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

. To get the estimate of θ and τ, we need to solve:

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

Then we will get:

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

Form (1), we can easily get θ. It is called **decouple**.

case 2.3. From the two subcases 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 in grids. So if there are 2 hyper-parameters, the search space is 2-dimension. If there are 3 hyper-parameters, the search space is 3-dimension... It will cost much time when the search space is high dimension.

Simply, we can give an uninformative prior to τ, $\mathbb{P}(\tau) \propto 1$. Or we can consider Jeffrey prior for τ. According to $\theta|\tau \sim N(0, \lambda\tau)$. Then we can get Fisher information:

$$\begin{aligned} I(\tau) &= \mathbb{E}[(\frac{\partial \ln f}{\partial \tau})] \\ &= \frac{1}{2\tau^2} \end{aligned}$$

, where $f = \frac{1}{\sqrt{2\pi\lambda\tau}}e^{-\frac{\theta^2}{2\lambda\tau}}$. So we can get the prior for τ, $\mathbb{P}(\tau) \propto \frac{1}{\tau}$. Then we will get:

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

After $-\ln$ operation and remove constants, we will get:

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

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

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

. We can see it is easy to solve.