Bayesian Statistics

Instructor: Malay Ghosh
Taken by Yu Zheng

Conjugate Priors

For a given likelihood, $L(\theta), \theta \in \Theta$ if f_θ constitutes a family of priors such that the resulting posterior also belongs to the family $\mathcal{F}_\theta$, then the family of priors f_θ is to be a conjugate family.

Examples:

- Binomial-Beta
- Poisson-Gamma
- Normal-Normal
- Uniform-Pareto:

$$f(x \mid \theta) = \frac{1}{\theta} I_{[0<x\leq\theta]}$$
$$\pi(\theta) \propto \theta^{-\alpha}$$

Bayesian Inference

Point estimation: Posterior mean/variance/mode/...

Improper Prior

$X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} \mathcal{N}(\theta, 1), \pi(\theta) = 1, -\infty < \theta < \infty.$

$$
\begin{aligned}
\pi(\theta \mid x_1, \dots, x_n) \propto & \exp[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2] \\
= & \exp[-\frac{n}{2}(\bar{x} - \theta)^2 - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2] \\
\propto & \exp[-\frac{n}{2}(\theta - \bar{x})^2] \\
\Rightarrow \theta \mid x_1, \dots, x_n \sim \mathcal{N}(\bar{x}, 1/n).
\end{aligned}
$$

Non-conjugate Priors

$X_1, \dots, X_n \mid \theta \overset{iid}{\sim} Bin(1, \theta), 0 < \theta < 1.$

$$
\phi = \text{logit}(\theta) = \log(\frac{\theta}{1-\theta})
$$
$$
\phi \sim \mathcal{N}(\mu, \sigma^2)
$$

Conjugacy is lost.

Credible Intervals and Sets

$$
X_1, \dots, X_n \mid \theta \sim f
$$
$$
\text{Prior } \pi(\theta), \ \theta \in \Theta
$$

A set $C \subset \Theta$ is said to be a credible set of size $1 - \alpha$ if $P(\theta \in C \mid X) = 1 - \alpha$.

Remarks:

- In a frequentist framework, it is not possible to make a probabilistic statement related to a confidence interval. In a Bayesian framework, this is possible, since we are looking at the conditional distribution of a parameter given the data.
- Often the confidence set is an interval. E.g., $[\theta_1(x), \theta_2(x)]$ such that $P(\theta_1(x) \le \theta \le \theta_2(x) \mid x) = 1 - \alpha$.

  Example:

$$X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} \mathcal{N}(\theta, \underbrace{\sigma^2}_{\text{known}})$$

$$\theta \sim \mathcal{N}(\mu, \tau^2)$$

$$\theta \mid x_1, \ldots, x_n \sim \mathcal{N}\left[(1-B)\bar{x} + B\mu, \frac{\sigma^2}{n}(1-B)\right], \text{ where } B = \frac{\sigma^2/n}{\sigma^2/n + \tau^2}$$

An equal-tailed CI for θ of size $1-\alpha$ is given by

$$(1-B)\bar{x} + B\mu \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{1-B}.$$

- Equal-tailed intervals does not always provide the shortest interval.

  E.g. Binomial with beta prior, Poisson with gamma prior, etc.

  Bayesians have recommended what they call **Highest Posterior Density (HPD)** interval or more generally HPD set.

  Suppose the posterior density of θ is unimodal. Then the HPD interval for θ is given by $C = \{\theta : \pi(\theta \mid x) \geq k\}$ where k is chosen so that $P(\theta \in C \mid x) = 1 - \alpha$.

  For symmetric unnimodal distributions, HPD intervals are equal-tailed.

Credible sets

$$X \mid \theta \sim \mathcal{N}(\theta, \Sigma)$$
$$\theta \sim \mathcal{N}(\mu, \Gamma)$$

$$X = \theta + e, \quad e \sim \mathcal{N}(0, \Sigma), \ e \text{ independent of } \theta$$

$$\begin{pmatrix} X \\ \theta \end{pmatrix} \sim \mathcal{N}_{2p}\left[\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \Sigma + \Gamma & \Gamma \\ \Gamma & \Gamma \end{pmatrix}\right]$$

$$\Theta \mid X \sim \mathcal{N}[\mu + \Gamma(\Sigma+\Gamma)^{-1}(X-\mu), \Gamma - \Gamma(\Sigma+\Gamma)^{-1}\Gamma]$$

$$B \overset{\Delta}{=} \Sigma(\Sigma+\Gamma)^{-1},$$
$$I - B = I - (\Sigma + \Gamma - \Gamma)(\Sigma+\Gamma)^{-1} = \Gamma(\Sigma+\Gamma)^{-1}$$

$$
\begin{aligned}
\Gamma-\Gamma(\Sigma+\Gamma)^{-1}\Gamma =&\Gamma-(\Gamma+\Sigma-\Sigma)(\Sigma+\Gamma)^{-1}\Gamma \\
=&\Gamma-\Gamma+\Sigma(\Sigma+\Gamma)^{-1}\Gamma \\
=&B\Gamma=(I-B)\Sigma
\end{aligned}
$$

$$
\begin{aligned}
(I-B)\Sigma =&\Sigma-\Sigma(\Sigma+\Gamma)^{-1}\Sigma \\
=&\Sigma-(\Sigma+\Gamma-\Gamma)(\Sigma+\Gamma)^{-1}\Sigma \\
=&\Gamma B=(I-B)\Sigma
\end{aligned}
$$

HPD Confidence set is an ellipsoid

$$
\left\{\theta:(\theta-((I-B)X+B\mu))^T((I-B)\Sigma)^{-1}(\theta-((I-B)X+B\mu))\leq K\right\}
$$

K is the chisquare percentile point.

Bayesian Hypothesis Testing

$$
\begin{aligned}
H_0:\theta\in\Theta_0 \quad & H_a:\theta\in\Theta_1 \\
\text{prior} \quad & \pi(\theta)
\end{aligned}
$$

Posteriror odds in favor of H_0: $\frac{P(H_0|x)}{P(H_1|x)}$.

Prior odds in favor of H_0: $\frac{\Pi(\Theta_0)}{\Pi(\Theta_1)}=\frac{\int_{\Theta_0}\pi(\theta)d\theta}{\int_{\Theta_1}\pi(\theta)d\theta}$.

$\Pi_0=P(\theta\in\Theta_0)$

Neukel Prior: $\Pi_0=1/2$

$g_0(\theta)$: density of θ when $\theta\in\Theta_0$; $g_1(\theta)$: density of θ when $\theta\in\Theta_1$. We have $\int_{\Theta_0}g_0(\theta)d\theta=\int_{\Theta_0}g_1(\theta)d\theta=1$.

$$
\pi(\theta)=\pi_0g_0(\theta)+(1-\pi_0)g_1(\theta)
$$

$$
m_\pi(x)=\int f(x\mid\theta)\pi(\theta)d\theta=\pi_0\int_{\Theta_0}f(x\mid\theta)g_0(\theta)d\theta+(1-\pi_0)\int_{\Theta_1}f(x\mid\theta)g_1(\theta)d\theta
$$

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{m_\pi(x)} = \begin{cases} \frac{f(x|\theta)g_0(\theta)}{m_\pi(x)}, & \theta \in \Theta_0 \\ \frac{f(x|\theta)g_1(\theta)}{m_\pi(x)}, & \theta \in \Theta_1 \end{cases}$$

$$p^\pi(H_0 \mid x) = \frac{\pi_0 \int_{\Theta_0} f(x \mid \theta) g_0(\theta) d\theta}{\pi_0 \int_{\Theta_0} f(x \mid \theta) g_0(\theta) d\theta + (1-\pi_0) \int_{\Theta_1} f(x \mid \theta) g_1(\theta) d\theta}$$

Posterior odds in favor of H_0:

$$\frac{P(H_0 \mid x)}{P(H_1 \mid x)} = \frac{\pi_0 \int_{\Theta_0} f(x \mid \theta) g_0(\theta) d\theta}{(1-\pi_0) \int_{\Theta_1} f(x \mid \theta) g_1(\theta) d\theta}$$

Prior odds in favor of H_0:

$$\frac{\pi_0}{1-\pi_0}$$

$$\begin{aligned} \mathrm{BF}_{01} &= \text{Bayes Factor in favor of } H_0 \\ &= \frac{\text{Posterior odds in favor of } H_0}{\text{Prior odds in favor of } H_0} \\ &= \frac{\int_{\Theta_0} f(x \mid \theta) g_0(\theta) d\theta}{\int_{\Theta_1} f(x \mid \theta) g_1(\theta) d\theta} \end{aligned}$$

Remarks:

- While BF_{01} does not depend on the choice of π_0, it does depend on the choice of g_0 and g_1.
- $\mathrm{BF}_{10} = \frac{1}{\mathrm{BF}_{01}}$.
- $\log(\mathrm{BF}_{01}) = \log(\text{Posterior odds in favor of } H_0) - \log(\text{Prior odds in favor of } H_0)$, change from prior odds to posterior odds in the log scale.

Example:

$X_1, \ldots, X_n \mid \mu, \sigma^2 \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$

$H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$

$$\begin{cases} g_0(\sigma^2) = (\sigma^2)^{-1} \\ g_1(\mu, \sigma^2) = \mathcal{N}(\mu \mid 0, \sigma^2)(\sigma^2)^{-1} \end{cases}$$

Simple Null vs Simple Alternative

$$H_0 : \theta = \theta_0 \quad H_1 : \theta = \theta_1$$

Posterior odds in favor of $H_0 = \frac{\pi_0 f(x|\theta_0)}{(1-\pi_0) f(x|\theta_1)}$

$\text{BF}_{01} = \frac{f(x|\theta_0)}{f(x|\theta_1)}$ Neyman-Pearson Likelihood Ratio

Example:

$X_1, \dots, X_n \mid \theta, \underbrace{\sigma^2}_{\text{known}} \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$

$$H_0 : \theta \leq \theta_0, \quad H_1 : \theta > \theta_0$$
$$\theta \sim \mathcal{N}(\mu, \tau^2)$$

$$\theta \mid x \sim \mathcal{N}((1-B)\bar{x} + B\mu, \frac{\sigma^2}{n}(1-B))$$

Posterior odds in favor of H_0:

$$\frac{P(\theta \leq \theta_0 \mid x)}{P(\theta > \theta_0 \mid x)} = \frac{\Phi(\frac{\theta_0 - (1-B)\bar{x} + B\mu}{\frac{\sigma}{n}\sqrt{1-B}})}{1 - \Phi(\frac{\theta_0 - (1-B)\bar{x} + B\mu}{\frac{\sigma}{n}\sqrt{1-B}})}$$

Examples:

- Sugar level for a person two hours after his breakfast.

  Let $X \mid \theta \sim \mathcal{N}(\theta, 100)$, θ being he true level, and $\theta \sim \mathcal{N}(100, 900)$.

  $\theta \mid X = x \sim \mathcal{N}((1-B)x + 100B, 100(1-B)) \xlongequal{B=100/(100+900)=0.1} \mathcal{N}(0.9x + 10, 90)$.

  Observe $x = 130$. Then $P(\theta \leq 130 \mid x) = 0.624$.

- IQ Test:

$$X \mid \theta \sim \mathcal{N}(\theta, 100)$$
$$\theta \sim \mathcal{N}(100, 15^2)$$

Posterior:

$$\theta \mid X = x \sim \mathcal{N}((1-B)x + 100B, 100(1-B)) \xlongequal{B=100/(100+225)=4/13} \mathcal{N}(\frac{9}{13}x + \frac{400}{13}, \frac{900}{13})$$

$P(\theta \leq 130 \mid x = 130) = ?$.

Point Null Hypothesis

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$$
$$\pi_0 = P(\theta = \theta_0)$$
$$\pi(\theta) = \pi_0 I_{[\theta=\theta_0]} + (1-\pi_0) g_1(\theta) I_{[\theta \neq \theta_0]}$$
$$m_\pi(x) = \pi_0 f(x \mid \theta_0) + (1-\pi_0) \int_{\theta \neq \theta_0} f(x \mid \theta) g_1(\theta) d\theta$$

Spike and Slab Priors

$$\pi(\theta) = \pi_0 I_{[\theta=0]} + (1-\pi_0) g_1(\theta)$$
$$m_\pi(x) = \pi_0 f(x \mid 0) + (1-\pi_0) \int_{\theta \neq 0} f(x \mid \theta) g_1(\theta) d\theta$$

$g_1(\theta)$ is typically $\mathcal{N}(\theta \mid \mu, \tau^2)$.

$$P(\theta = \theta_0 \mid x) = \frac{\pi_0 f(x \mid \theta_0)}{m_\pi(x)} = \frac{\pi_0 f(x \mid \theta_0)}{\pi_0 f(x \mid \theta_0) + (1-\pi_0) \int_{\theta \neq \theta_0} f(x \mid \theta) g_1(\theta) d\theta}$$
$$= \left[1 + \frac{1-\pi_0}{\pi_0} \cdot \frac{\int_{\theta \neq \theta_0} f(x \mid \theta) g_1(\theta) d\theta}{f(x \mid \theta_0)} \right]^{-1}$$

$$\mathrm{BF}_{01} = \frac{f(x \mid \theta_0)}{\int_{\theta \neq \theta_0} f(x \mid \theta) g_1(\theta) d\theta}$$

Interpretation of Bayes Factors (A. Raftry, Bren 1996 251-266)

BF_{10}	$2\log_e(\mathrm{BF}_{10})$	**Evidence for H_0**
<1	<0	Negative (Supports H_0)
1 - 3	0 - 2.2	Not worth more than a mention
3 - 20	2.2 - 6	Positive
20 - 150	6 - 10	Strong
>150	>10	Very strong

(Paper) Bayes Factors: What they are what they are not

Parameter space Ω; $\Omega_H \subset \Omega, \Omega_A = \Omega - \Omega_H$; $f(x \mid \theta)$: pdf of x given $\theta \in \Omega$.

$$H_0 : \theta \in \Omega_H, \quad H_1 : \theta \in \Omega_A = \Omega - \Omega_H$$

$$f_H(x) = \frac{\int_{\Omega_H} f(x \mid \theta) d\mu(\theta)}{\underbrace{\mu(\Omega_H)}_{:=p}}, \quad f_A(x) = \frac{\int_{\Omega_A} f(x \mid \theta) d\mu(\theta)}{\mu(\Omega_A)}$$

$$\text{Posterior odds} = \frac{p f_H(x)}{(1-p) f_A(x)}$$
$$\text{Prior odss} = \frac{p}{1-p}$$
$$\mathrm{BF}_{01} = \frac{f_H(x)}{f_A(x)}$$

$\Omega = \{0, 1/2, 1\}$

6 hypotheses:

- $H_1 : \theta = 1$
- $H_2 : \theta = 1/2$
- $H_3 : \theta = 0$
- $H_4 : \theta \neq 1$

- $H_5 : \theta \neq 1/2$
- $H_6 : \theta \neq 0$

A coint is tossed 4 times, all ending in heads.

$$f_{H_2}(x) = \frac{(1/2)^4 \mu(\{1/2\})}{\mu(\{1/2\})} = \frac{1}{16}$$
$$f_{H_5}(x) = \frac{(1)^4 \mu(\{1\}) + 0^4 \mu(\{0\})}{\mu(\{1\}) + \mu(\{0\})} = \frac{\mu(\{1\})}{\mu(\{1\}) + \mu(\{0\})}$$

$$\begin{aligned}
&\text{Posterior expected cost of rejection } H = clP(H \text{ is true} \mid x) \\
&\text{Posterior expected cost of accepting } H = lP(H \text{ is not true} \mid x) \\
&\text{Bayesian reject } H \text{ if } \frac{clP(H \text{ is true} \mid x)}{lP(H \text{ is not true} \mid x)} < 1 \\
\text{i.e.} \quad &\frac{P(H \text{ is true} \mid x)}{1 - P(H \text{ is true} \mid x)} < \frac{1}{c} \\
&\Leftrightarrow P(H \text{ is true} \mid x) < \frac{1}{c+1}, \\
&\text{reject } H \text{ if Bayes Factor in favor of } H \text{ is less than } k
\end{aligned}$$

$$f_{H_1}(x) = \frac{(1)^4 \mu(\{1\})}{\mu(\{1\})} = 1$$
$$f_{H_2}(x) = \frac{1}{16}$$
$$f_{H_3}(x) = \frac{0^4 \mu(\{0\})}{\mu(\{0\})} = 0$$

$$\text{Assume } \mu(\{1\}) = 0.01, \mu(\{1/2\}) = 0.98, \mu(\{0\}) = 0.01$$

$$
\begin{aligned}
f_{H_4}(x) &= \frac{\int_{H_2 \cup H_3} f(x \mid \theta) d\mu(\theta)}{\mu(H_2 \cup H_3)} \\
&= \frac{(1/2)^4 \mu(\{1/2\}) + 0^4 \mu(\{0\})}{\mu(\{1/2\}) + \mu(\{0\})} \\
&= \frac{0.98/16}{0.98 + 0.01} = 0.0619 \\
f_{H_5}(x) &= \frac{0^4 \mu(\{0\}) + 1^4 \mu(\{1\})}{\mu(\{0\}) + \mu(\{1\})} = \frac{0.01}{0.01 + 0.01} = 0.5 \\
f_{H_6}(x) &= \frac{1^4 \mu(\{1\}) + (1/2)^4 \mu(\{1/2\})}{\mu(\{1\}) + \mu(\{1/2\})} = \frac{0.01 + 0.98/16}{0.99} = 0.72
\end{aligned}
$$

$$
\frac{f_{H_4}(x)}{f_{H_1}(x)} = 0.0619, \quad \frac{f_{H_2}(x)}{f_{H_5}(x)} = \frac{0.0625}{0.5} = 0.125
$$

If we now choose $k \in (0.0619, 0.125)$, we have contradictory results.

$$
\text{Assume } f_{H_3}(x) < \min(f_{H_1}(x), f_{H_2}(x))
$$

$$
\begin{aligned}
f_{H_4}(x) =& \frac{\int_{H_2 \cup H_3} f(x \mid \theta) d\mu(\theta)}{\mu(H_2 \cup H_3)} \\
=& \frac{\int_{H_2} f(x \mid \theta) d\mu(\theta) + \int_{H_3} f(x \mid \theta) d\mu(\theta)}{\mu(H_2) + \mu(H_3)} \\
=& \frac{\mu(H_2) \int_{H_2} \frac{f(x|\theta)}{\mu(H_2)} d\mu(\theta) + \mu(H_3) \int_{H_3} \frac{f(x|\theta)}{\mu(H_3)} d\mu(\theta)}{\mu(H_2) + \mu(H_3)} \\
=& \frac{\mu(H_2)}{\mu(H_2) + \mu(H_3)} f_{H_2}(x) + \frac{\mu(H_3)}{\mu(H_2) + \mu(H_3)} f_{H_3}(x) \\
<& f_{H_2}(x)
\end{aligned}
$$

Therefore,

$$
\frac{f_{H_4}(x)}{f_{H_1}(x)} < \frac{f_{H_2}(x)}{f_{H_1}(x)}.
$$

On the other hand,

$$
\begin{aligned}
f_{H_5}(x) =& \frac{\int_{H_1 \cup H_3} f(x \mid \theta) d\mu(\theta)}{\mu(H_1 \cup H_3)} \\
=& \frac{\mu(H_1)}{\mu(H_1) + \mu(H_3)} f_{H_1}(x) + \frac{\mu(H_3)}{\mu(H_1) + \mu(H_3)} f_{H_3}(x) \\
<& f_{H_1}(x)
\end{aligned}
$$

Therefore,

$$
\frac{f_{H_2}(x)}{f_{H_5}(x)} > \frac{f_{H_2}(x)}{f_{H_1}(x)}.
$$

Default Priors for Binomial(p)

- Laplace: $\pi(p) = 1, 0 \leq p \leq 1$
- Jeffreys: $\pi(p) = p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}, 0 < p < 1$
- Haldane: $\pi(p) = p^{-1}(1-p)^{-1}, 0 < p < 1$

For the prior $\pi(\sigma^2) = 1$, consider $z = \sigma$. We have $z^2 = \sigma^2 \Rightarrow \frac{d\sigma^2}{dz} = 2z \Rightarrow \pi(z) \propto z$ -- Not invariant under one-to-one transformation.

Jeffreys's Prior

Invariant under 1:1 transformation

$|I(\theta)|^{1/2}$

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

Suppose $\phi \overset{1:1}{\leftrightarrow} \theta$.

$$
\begin{aligned}
I(\phi) =& \mathbb{E}\left[\frac{d\log f}{d\theta}\right]^2 = \mathbb{E}\left[\frac{d\log f}{d\theta}\frac{d\theta}{d\phi}\right]^2 \\
=& \mathbb{E}\left[\frac{d\log f}{d\theta}\right]^2\left(\frac{d\theta}{d\phi}\right)^2 \\
\Rightarrow |I(\phi)|^{1/2} =& |I(\theta)|^{1/2}\left|\frac{d\theta}{d\phi}\right|
\end{aligned}
$$

Now, for Binomial(p),

$$
\begin{gathered}
f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x} \\
\frac{d\log f}{dp} = \frac{x}{p} - \frac{n-x}{1-p} \\
\frac{d^2\log f}{dp^2} = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2} \\
I(p) = \frac{np}{p^2} + \frac{n(1-p)}{(1-p)^2} = \frac{n}{p(1-p)} \\
\pi_J(p) = p^{-1/2}(1-p)^{-1/2}
\end{gathered}
$$

Multiparameter

$\theta = (\theta_1, \ldots, \theta_p)$

$\pi(\theta) = |I(\theta)|^{1/2}$

$\phi = (\phi_1, \ldots, \phi_p) \overset{1:1}{\Leftrightarrow} (\theta_1, \ldots, \theta_p)$

$$
I(\phi) = \mathbb{E}\left[\frac{\partial\log f}{\partial\phi}\left(\frac{\partial\log f}{\partial\phi}\right)^T\right] = \mathbb{E}\left[\begin{pmatrix} \frac{\partial\log f}{\partial\phi_1} \\ \vdots \\ \frac{\partial\log f}{\partial\phi_p} \end{pmatrix}\begin{pmatrix} \frac{\partial\log f}{\partial\phi_1} & \cdots & \frac{\partial\log f}{\partial\phi_p} \end{pmatrix}\right]
$$

$$
\begin{aligned}
\frac{\partial \log f}{\partial \phi_j} &= \sum_{k=1}^{p} \frac{\partial \log f}{\partial \theta_k} \frac{\partial \theta_k}{\partial \phi_j} \\
&= \begin{pmatrix} \frac{\partial \theta_1}{\partial \phi_j} & \cdots & \frac{\partial \theta_p}{\partial \phi_j} \end{pmatrix} \begin{pmatrix} \frac{\partial \log f}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log f}{\partial \theta_p} \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
I(\phi) =& \mathbb{E}\left(\begin{pmatrix} \frac{\partial \theta_1}{\partial \phi_1} & \cdots & \frac{\partial \theta_p}{\partial \phi_1} \\ \vdots & & \vdots \\ \frac{\partial \theta_1}{\partial \phi_p} & \cdots & \frac{\partial \theta_p}{\partial \phi_p} \end{pmatrix} \begin{pmatrix} \frac{\partial \log f}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log f}{\partial \theta_p} \end{pmatrix} \begin{pmatrix} \frac{\partial \log f}{\partial \theta_1} & \cdots & \frac{\partial \log f}{\partial \theta_p} \end{pmatrix} \begin{pmatrix} \frac{\partial \theta_1}{\partial \phi_1} & \cdots & \frac{\partial \theta_1}{\partial \phi_p} \\ \vdots & & \vdots \\ \frac{\partial \theta_p}{\partial \phi_1} & \cdots & \frac{\partial \theta_p}{\partial \phi_p} \end{pmatrix} \right) \\
=& J I(\theta) J^T
\end{aligned}
$$

So,

$$
\begin{aligned}
|I(\phi)| = |J||I(\theta)||J^T| = |J|^2 |I(\theta)| \\
\Rightarrow |I(\phi)|^{1/2} = |J||I(\theta)|^{1/2}
\end{aligned}
$$

<u>Example:</u>

$X_1, \ldots, X_n \mid \mu, \sigma^2 \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$

$$
L(\mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right]
$$

$$
\begin{cases}
\frac{\partial \log L}{\partial \mu} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^2} = \frac{n(\bar{x} - \mu)}{\sigma^2} \\
\frac{\partial \log L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2
\end{cases}
$$

$$
\Rightarrow \begin{cases}
\frac{\partial^2 \log L}{\partial \mu^2} = -\frac{n}{\sigma^2}, \\
\frac{\partial^2 \log L}{\partial \mu \partial \sigma} = -\frac{2n(\bar{x} - \mu)}{\sigma^3}, \\
\frac{\partial^2 \log L}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^n (x_i - \mu)^2
\end{cases}
$$

So,

$$
\begin{aligned}
I(\mu, \sigma) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix} \\
\Rightarrow |I(\mu, \sigma)|^{1/2} \propto \sigma^{-2}
\end{aligned}
$$

- **Jeffreys' general rule prior** $|I(\theta)|^{1/2}$; here $|I(\mu,\sigma)|^{1/2} \propto \sigma^{-2}$
- **Jeffreys' independence prior** $\pi(\mu,\sigma) \propto \sigma^{-1}$ -- Recommended for point estimation and construction of credible sets

If $\pi(\mu,\sigma) \propto \sigma^{-1}$, consider $z = \sigma^2$. We have $\pi(\mu,z) = z^{-1/2}\frac{1}{2\sqrt{z}} \propto z^{-1}$, i.e., $\pi(\mu,\sigma^2) \propto (\sigma^2)^{-1}$.

$$L(\mu,\sigma) \propto (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$
$$\pi(\mu,\sigma^2) \propto (\sigma^2)^{-1}$$

Then,

$$\pi(\mu,\sigma^2 \mid X_1,\ldots,X_n) \propto (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

Integrating out σ^2, we get

$$\begin{aligned}\pi(\mu) &\propto \left[\sum_{i=1}^{n}(X_i-\mu)^2\right]^{-n/2}\\ &= \left[n(\bar{X}-\mu)^2 + \sum_{i=1}^{n}(X_i-\bar{X})^2\right]^{-n/2}\\ &\propto \left[1+\frac{n(\bar{X}-\mu)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]^{-n/2}\\ &= \left[1+\frac{(\bar{X}-\mu)^2}{(n-1)(S^2/n)}\right]^{-(n-1)/2-1/2}\end{aligned}$$

Posterior for μ is Student's t with location μ, scale $S/\sqrt{n}$, df $n-1$.

A Student's t-distribution is a Scale-Mixed Normal Dsitribution

$$X \mid \theta, \sigma^2, r \sim \mathcal{N}(\theta, \sigma^2/r)$$
$$r \sim \text{Gamma}(\nu/2, \nu/2)$$

Then

$$X \mid \theta, \sigma^2 \sim \text{Student's t}(\nu(\text{df}), \theta(\text{location}), \sigma(\text{scale}))$$

- If $\nu = 1$, the student's t distribution becomes a Cauchy distribution.

Proof:

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

A Double Exponential Distribution is a Scale-Mixed Normal Distribution

$$X \mid \theta, \sigma^2, r \sim \mathcal{N}(\theta, \sigma^2/r)$$
$$f(r) = \exp(-\frac{1}{2r})\frac{1}{2r^2} = \exp(-\frac{1}{2r})\frac{1}{2} r^{-1-1} \sim \text{InvGamma(1,1/2)}$$

Then

$$f(x) = \frac{1}{2\sigma}\exp\left[-\frac{|x-\theta|}{\sigma^2}\right]$$

Proof:

$$
\begin{aligned}
f(x) = & \int_0^\infty \left(\frac{r}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{r(x-\theta)^2}{2\sigma^2}\right] \exp(-\frac{1}{2r}) \frac{dr}{r^2} \\
= & \frac{1}{2\sigma} \int_0^\infty \frac{1}{(2\pi r^3)^{1/2}} \exp\left[-\frac{1}{2r}\left[\frac{r(x-\theta)^2}{\sigma^2}+1\right]\right] dr \\
= & \cdots \\
= & \frac{1}{2\sigma} \exp[-\frac{1}{\sigma}|x-\theta|]
\end{aligned}
$$

Inverse Gaussian

$$
f(z) = \frac{1}{(2\pi z^3)^{1/2}} \exp\left[-\frac{1}{2z}\left(\frac{z}{\mu}-1\right)^2\right]
$$

$$
\mathbb{E}Z = \mu
$$

Logistic pdf

$$
\begin{aligned}
f(x) = & \frac{\exp(x)}{(1+\exp(x))^2} = \exp(-x)[1+\exp(-x)]^{-2} \\
= & \exp(-x)[1 - 2\exp(-x) + 3\exp(-2x) - 4\exp(-3x) + \cdots] \\
= & \sum_{r=1}^\infty (-1)^{r-1} r \exp(-rx) \\
\xlongequal{f(x)=f(-x)} & 2\sum_{r=1}^\infty (-1)^{r-1} \frac{r}{2} \exp(-r|x|) \text{ -- Mixture of DE}
\end{aligned}
$$

[More on the lecture notes in paper]

Asymptotic Normality of the Posterior

Heuristics

$$f(x \mid \theta) \quad \pi(\theta)$$
$$\pi(\theta \mid x)$$

$$\begin{aligned}
\log \pi(\theta \mid x_n) =& \log \pi(\hat{\theta}_n \mid x_n) + (\theta - \hat{\theta}_n) \frac{d \log \pi(\theta \mid x_n)}{d\theta} \mid_{\theta=\hat{\theta}_n} + \frac{1}{2}(\theta - \hat{\theta}_n)^2 \frac{d^2 \log \pi(\theta \mid x_n)}{d\theta^2} \mid_{\theta=\hat{\theta}_n} \\
=& \log \pi(\tilde{\theta}_n \mid x_n) + \frac{1}{2}(\theta - \tilde{\theta}_n)^2 (-\tilde{I}_n(\tilde{\theta}_n)) \\
\Rightarrow \pi(\theta \mid x) =& \pi(\tilde{\theta}_n \mid x_n) \exp[-\frac{1}{2}(\theta - \tilde{\theta}_n)^2 \tilde{I}_n(\tilde{\theta}_n)] \\
& \propto \exp[-\frac{1}{2}(\theta - \tilde{\theta}_n)^2 \tilde{I}_n(\tilde{\theta}_n)] \\
& \sim \mathcal{N}(\tilde{\theta}_n, \tilde{I}_n^{-1}(\tilde{\theta}_n))
\end{aligned}$$

where $\hat{\theta}_n$ is the MLE of θ, $\tilde{\theta}_n$ is very close to $\hat{\theta}_n$.

Berstein von Mises proves asymptotic $\mathcal{N}(\hat{\theta}_n, \hat{I}_n^{-1}(\hat{\theta}_n))$, where $\hat{I}_n^{-1}(\hat{\theta}_n) = -\frac{d^2 \log f(x_n|\theta)}{d\theta^2}|_{\theta=\hat{\theta}_n}$ for normalized posterior.

Total Variation Between Two Densities f_1 and f_2

$$TV(f_1, f_2) = \sup_A \left| \int_A (f_1 - f_2) d\mu \right|$$

Recall

$$TV(f_1, f_2) = \frac{1}{2} \int |f_1 - f_2| d\mu.$$

Renyi Divergence

$$D_\alpha(f_1, f_2) = \frac{1 - \int f_1^\alpha f_2^{1-\alpha} d\mu}{\alpha(1-\alpha)}, \quad 0 < \alpha < 1$$

$$
\begin{aligned}
\lim_{\alpha\to 0} D_\alpha(f_1,f_2) =& \lim_{\alpha\to 0} \frac{1-\int\left(\frac{f_1}{f_2}\right)^\alpha f_2 d\mu}{\alpha(1-\alpha)} \\
=& \lim_{\alpha\to 0} \frac{-\int\left(\frac{f_1}{f_2}\right)^\alpha (\log\frac{f_1}{f_2}) f_2 d\mu}{1-2\alpha} \\
=& \int (\log\frac{f_2}{f_1}) f_2 d\mu \\
=& KL(f_2,f_1),
\end{aligned}
$$

$$
\begin{aligned}
\lim_{\alpha\to 1} D(f_1,f_2) =& \lim_{\alpha\to 0} \frac{-\int\left(\frac{f_1}{f_2}\right)^\alpha (\log\frac{f_1}{f_2}) f_2 d\mu}{1-2\alpha} \\
=& \int \log(\frac{f_1}{f_2}) f_1 d\mu \\
=& KL(f_1,f_2).
\end{aligned}
$$

Bhattachwyga-Hellinger

$$
H(f_1,f_2) = \left[\int \left(f_1^{1/2} - f_2^{1/2}\right)^2 d\mu\right]^{1/2}
$$

$$
\begin{aligned}
D_{1/2}(f_1,f_2) =& 4\left[1 - \int f_1^{1/2} f_2^{1/2} d\mu\right] \\
=& 2\left[2 - 2\int f_1^{1/2} f_2^{1/2} d\mu\right] \\
=& 2\left[\int f_1 d\mu + \int f_2 d\mu - \int f_1^{1/2} f_2^{1/2} d\mu\right] \\
=& 2\int \left(f_1^{1/2} - f_2^{1/2}\right)^2 d\mu \\
=& 2H^2(f_1,f_2)
\end{aligned}
$$

$$
\begin{aligned}
D_{-1}(f_1,f_2) =& \frac{1-\int f_1^{-1} f_2^2 d\mu}{(-1)(2)} \\
=& \frac{1}{2}\left[\int \frac{f_2^2}{f_1} d\mu - 1\right] \\
=& \frac{1}{2}\int \frac{(f_2-f_1)^2}{f_1} d\mu \quad \text{-- Chi-square divergence}
\end{aligned}
$$

$$
\begin{aligned}
H^2(f_1, f_2) =& 2\left[1 - \int f_1^{1/2} f_2^{1/2} d\mu\right] \\
=& 2\left[1 - \int (\frac{f_1}{f_2})^{1/2} f_2 d\mu\right] \\
\leq& 2\left[1 - \int [1 + \log(\frac{f_1}{f_2})^{1/2}] f_2 d\mu\right] \\
=& 2\left[1 - \int f_2 d\mu + \frac{1}{2}\int \log(\frac{f_2}{f_1}) f_2 d\mu\right] \\
=& \int \log(\frac{f_2}{f_1}) f_2 d\mu \\
=& KL(f_2, f_1)
\end{aligned}
$$

$$
\begin{aligned}
KL(f_1, f_2) &= \int (\log \frac{f_1}{f_2}) f_1 d\mu \\
&\leq \int \left(\frac{f_1}{f_2} - 1\right) f_2 d\mu \\
&= \int \frac{f_1^2}{f_2} d\mu - 1 \\
&= \int \frac{(f_1 - f_2)^2}{f_2} d\mu
\end{aligned}
$$

Scheffe's Theorem

Let $\{p_n, n \geq 1\}$ be a sequence of pdf's such that $p_n(x) \to p(x)$ pointwise. Then $\int |p_n - p| d\mu \to 0$.

proof:

$$
\begin{aligned}
\int |p_n - p| d\mu =& \int [p_n + p - 2\min(p_n, p)] d\mu \\
=& 2[1 - \int \min(p_n, p) d\mu]
\end{aligned}
$$

Because $\min(p_n, p) \to p$ and $\min(p_n, p) \leq p$ and $\int p d\mu = 1$, apply DCT to get

$$
\int |p_n - p| d\mu \to 2(1 - 1) = 0.
$$

Result: $TV(p_n,p)\to 0 \Rightarrow D_\alpha(p_n,p)\to 0$ **for all** $0<\alpha<1$.

proof:

$$\begin{aligned}D_\alpha(p_n,p) &= \frac{1-\int p_n^\alpha p^{1-\alpha}d\mu}{\alpha(1-\alpha)}\\ &\leq \frac{1-\int \min(p_n,p)d\mu}{\alpha(1-\alpha)}\\ &\leq \frac{\int |p_n-p|d\mu}{2\alpha(1-\alpha)}\to 0\end{aligned}$$

Result: $D_{1/2}(p_n,p)\to 0 \Rightarrow TV(p_n,p)\to 0$.

proof:

$$\begin{aligned}\int |p_n-p|d\mu &= \int \left|(p_n^{1/2}+p^{1/2})(p_n^{1/2}-p^{1/2})\right|d\mu\\ &\leq \left[\int (p_n^{1/2}+p^{1/2})^2 d\mu\right]^{1/2}\left[\int (p_n^{1/2}-p^{1/2})^2 d\mu\right]^{1/2}\\ &= \left[\int (p_n+p+2p_n^{1/2}p^{1/2})d\mu\right]^{1/2} H(p_n,p)\\ &= \left[2\left(1+\int p_n^{1/2}p^{1/2}d\mu\right)\right]^{1/2} H(p_n,p)\\ &= \left[4-2(1-\int p_n^{1/2}p^{1/2}d\mu)\right]^{1/2} H(p_n,p)\\ &= \left[4-H^2(p_n,p)\right]^{1/2} H(p_n,p)\to 0\end{aligned}$$

To put everything in a nutshell:

Final Result: $D_{1/2}(p_n,p)\to 0 \Rightarrow TV(p_n,p)\to 0 \Rightarrow D_\alpha(p_n,p)\to 0$ **for all** $0<\alpha<1$.

Bernstein-von Mises Theorem

See the lecture notes in paper