01 /13/ 22

$\mathcal{F} = \{ N(\mu, \tau^2) ; \mu \in \mathbb{R}, \tau^2 > 0 \}$

† Conjugate Priors (CR Sec 3.3)

- **Def 3.3.1:** A family $\mathcal{F}$ of probability distributions on Θ is said to be *conjugate* (or closed under sampling) for a likelihood function $f(x \mid \theta)$ if, for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta \mid x)$ also belong to $\mathcal{F}$.

e.g1 A beta prior distribution and a binomial sampling model lead to a beta posterior distribution. We say "The class of beta priors is conjugate for the binomial sampling distribution."

e.g2 Similarly, normal priors are a conjugate family for normal sampling distributions.

† Examples: Conjugate Priors

e.g1 Assume $x \mid \theta \sim \mathrm{N}(\theta, \sigma^2)$ and $\theta \sim \mathrm{N}(\mu, \tau^2)$.

$$\Rightarrow \quad \theta \mid x \sim \mathrm{N}\left(\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{x}{\sigma^2} + \frac{\mu}{\tau^2}\right), \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right).$$

⋆⋆ Normal priors are a conjugate family for normal sampling distributions.

e.g2 Assume $X \mid \theta \sim \mathrm{Bin}(n, \theta)$ and $\theta \sim \mathrm{Be}(\alpha, \beta)$.

$$\Rightarrow \theta \mid x \sim \mathrm{Be}(\alpha + x, \beta + n - x).$$

⋆⋆ Beta priors are a conjugate family for binomial sampling distributions.

- If $\mathcal{F}$ is a conjugate family,

  obtaining the posterior $\Leftrightarrow$ updating the corresponding parameters

  i.e, data does not modify the whole structure of the distribution of θ, but simply updates its parameters.

- A classical parametric approach to build up prior distributions based on limited prior input

- main motivation: tractability

- A conjugate family can frequently be determined by examining the likelihood functions $\ell(\theta \mid x)$ and choosing, as a conjugate family, the class of distributions with the same functional form as these likelihood functions.

  $\Rightarrow$ often called natural conjugate priors.

  $\Rightarrow$ can find a conjugate family for the sampling distribution in the exponential family.

Find a conjugate prior for a Poisson sampling distribution.

- ~~Show a Poisson distribution, $X \sim \text{Poi}(\theta)$ with $\theta > 0$ is an exponential family.~~

$\theta > 0$

$X \mid \theta \sim \text{Poi}(\theta)$, $x = 0, 1, 2, \ldots$

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

$f(x \mid \theta) = \frac{e^{-\theta} \theta^x}{x!}$

$\theta \sim Ga(a, b)$

$l(\theta \mid x) \propto \theta^x e^{-\theta}$ $\Rightarrow$ conjugate prior is Ga

Assume $\theta \sim Ga(a, b)$ and lets find $\pi(\theta \mid x)$

$$\pi(\theta \mid x) \propto f(x \mid \theta)\, \pi(\theta)$$

$$= \frac{e^{-\theta} \theta^x}{x!} \cdot \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}$$

$$\propto \underline{\theta^{a+x-1} e^{-(1+b)\theta}}$$

$\Rightarrow$ a kernel for $Ga(a+x, b+1)$

$\Rightarrow$ $\theta \mid x \sim Ga(a+x, b+1)$

† Exponential Families (CR §3.3.3, Casella & Berger §3.4)

- A family of pdfs or pmfs is called an *exponential family* if it can be expressed as

$$f(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(R(\boldsymbol{\theta})T(x)\right).$$

≥0

⋆⋆ $h(x) \geq 0$

⋆⋆ $T(x) = [t_1(x), \ldots, t_k(x)]$ are real-valued functions of the observations x (cannot depend on $\boldsymbol{\theta}$)

∗∗ natural sufficient statistic.

∗∗ all the information about $\boldsymbol{\theta}$ in the sample is summarized in $T(x)$.

⋆⋆ $c(\boldsymbol{\theta}) \geq 0$

⋆⋆ $R(\boldsymbol{\theta}) = (r_1(\boldsymbol{\theta}), \ldots, r_k(\boldsymbol{\theta}))$ are real-valued functions of the possibly vector-valued parameter $\boldsymbol{\theta}$ (cannot depend on x)

† Exponential Families (contd)

- The sufficient statistic and the parameter vectors are usually of equal length.

- These include the continuous families- normal, gamma, and beta, and the discrete families- binomial, Poisson, and negative binomial.

⋆⋆ consider a change of variables $\boldsymbol{z} = T(x)$ and a reparameterization $\boldsymbol{\eta} = R(\boldsymbol{\theta})$ (natural parameter) and rewrite

$$f(z \mid \boldsymbol{\eta}) = C^{\star}(\boldsymbol{\eta}) h^{\star}(z) \exp\left(\boldsymbol{\eta}\boldsymbol{z}\right)$$

$\Rightarrow$ the canonical form

- Show a Poisson distribution, $X \sim \mathrm{Poi}(\theta)$ with $\theta > 0$ is an exponential family.

$$f(x|\theta) = \frac{e^{-\theta}\theta^{x}}{x!} = \frac{1}{x!}\, e^{-\theta}\, e^{x\log\theta}$$

$$\begin{cases} h(x) = \frac{1}{x!} \times 2 \\ C(\theta) = e^{-\theta} \times \frac{1}{2} \\ R(\theta) = \log\theta \\ T(x) = x \end{cases} \Rightarrow \quad \mathrm{Poi}(\theta) \text{ is an exponential family.}$$

- Show a normal distribution, $X \sim \mathrm{N}(\mu, \sigma^2)$ with $\underline{\boldsymbol{\theta} = (\mu, \sigma)}$, where $\mu \in \mathbb{R}$ and $\sigma > 0$, is an exponential family.

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2}\right)$$

$$h(x) = \frac{1}{\sqrt{2\pi}}$$

$$C(\theta) = C(\mu, \sigma) = \frac{1}{\sigma} \cdot \exp\left(-\frac{\mu^2}{2\sigma^2}\right)$$

$$T(x) = \left(-\frac{x^2}{2}, x\right), \quad R(\theta) = \left(\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}\right)$$

$\Rightarrow$ $N(\mu, \sigma^2)$ is an exp. family

- CR §3.3.4 Conjugate distributions for exponential families: See Propositions 3.3.13 and 3.3.14.

♣ Table 3.3.1 Natural conjugate priors for some common exponential families

Table 3.3.1. *Natural conjugate priors for some common exponential families*

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta,\sigma^2)$	Normal $\mathcal{N}(\mu,\tau^2)$	$\mathcal{N}(\varrho(\sigma^2\mu+\tau^2x),\varrho\sigma^2\tau^2)$ $\varrho^{-1}=\sigma^2+\tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+x,\beta+1)$
Gamma $\mathcal{G}(\nu,\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+\nu,\beta+x)$
Binomial $\mathcal{B}(n,\theta)$	Beta $\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+x,\beta+n-x)$
Negative Binomial $\mathcal{N}eg(m,\theta)$	Beta $\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+m,\beta+x)$
Multinomial $\mathcal{M}_k(\theta_1,\ldots,\theta_k)$	Dirichlet $\mathcal{D}(\alpha_1,\ldots,\alpha_k)$	$\mathcal{D}(\alpha_1+x_1,\ldots,\alpha_k+x_k)$
Normal $\mathcal{N}(\mu,1/\theta)$	Gamma $\mathcal{G}a(\alpha,\beta)$	$\mathcal{G}(\alpha+0.5,\beta+(\mu-x)^2/2)$

$N(3,\theta)$

$f(x|\theta)$

$\cancel{f(x|\theta)}$

$\frac{1}{\sqrt{2\pi\theta}}e$

$N(\mu,\sigma^2)$

$\sigma^2 \sim IG(\alpha,\beta)$

$\sigma^2|x \sim IG$

$\Rightarrow 1/\sigma^2 \sim Ga(\alpha,\beta)$

$\Rightarrow 1/\sigma^2|X \sim Ga$

$\theta \sim Ga \Leftrightarrow 1/\theta \sim IG$

$X \mid \theta \sim N(\mu, \theta)$, μ fixed $\quad \theta$: variance

$\theta \sim IG(\alpha, \beta)$

$$\pi(\theta \mid x) \propto f(x \mid \theta)\, \pi(\theta)$$

$$\propto \theta^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\theta}\right) \theta^{-\alpha-1} \exp\left(-\frac{\beta}{\theta}\right)$$

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

$$\Rightarrow \theta \mid x \sim IG\left(\alpha + \frac{1}{2}, \beta + \frac{(x-\mu)^2}{2}\right)$$

$X \mid \eta \sim N(\mu, 1/\eta)$ i.e. $\eta = \frac{1}{\theta}$

let $\eta \sim Ga(\alpha, \beta)$

$$\pi(\eta \mid x) \propto f(x \mid \eta)\, \pi(\eta)$$

$$\propto (\eta)^{1/2} \cdot \exp\left(-\frac{\eta(x-\mu)^2}{2}\right) \cdot \eta^{\alpha-1} \exp(-\beta\eta)$$

$$= \eta^{\alpha+\frac{1}{2}-1} \exp\left(-\eta\left(\frac{(x-\mu)^2}{2} + \beta\right)\right)$$

$$\Rightarrow \eta \mid x \sim Ga\left(\alpha + \frac{1}{2}, \beta + \frac{(x-\mu)^2}{2}\right)$$

From this

let $\theta = \frac{1}{\eta}$ and find the distribution of θ by a change-variable technique

$\Rightarrow$ we can find $\theta \mid x \sim IG\left(\alpha + \frac{1}{2}, \beta + \frac{(x-\mu)^2}{2}\right)$,

which is the same as the above.

In other words, placing a IG prior for the variance is the same as placing a Ga prior for the precision

because variance $= \frac{1}{\text{precision}}$

† Improper Prior Distributions (CR 1.4)

- Recall that the parameter is a random variable following a probability distribution $\pi(\theta)$.

- We say the prior distribution is *improper* (or *generalized*) if

$$\int_{\Theta} \pi(\theta) d\theta = +\infty.$$

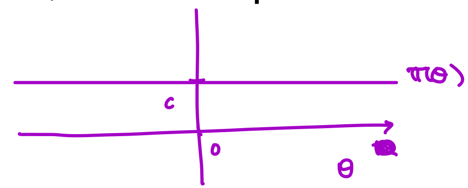
- Bayesian methods apply as long as the posterior distribution is defined.

- The posterior exists when the pseudo marginal distribution $\int_{\Theta} \pi(\theta) f(x \mid \theta) d\theta$ is well defined.

$< \infty$

$\theta \in \mathbb{R}$

♣ Example 3: Assume that an observation, x is normally distributed with mean θ and known variance σ^2. The parameter of interest, θ has an improper prior distribution, $\pi(\theta) = \underset{>0}{c}$. Check it produces a proper posterior distribution. If so, find the posterior distribution.



$$\int_{-\infty}^{\infty} \pi(\theta)\, d\theta = \int_{-\infty}^{\infty} c \, d\theta = \infty$$

$\Rightarrow \quad m(x) < \infty$??

$$m(x) = \int_{-\infty}^{\infty} f(x|\theta)\, \pi(\theta)\, d\theta$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\theta)^2\right) \cdot c \, d\theta$$

$E(\theta|x) = x$

$$= c < \infty$$

$\Rightarrow \boxed{\theta | x \sim N(x, \sigma^2)}$

$$\pi(\theta|x) \propto f(x|\theta)\, \pi(\theta)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2\sigma^2}(x-\theta)^2\right) \cdot c$$

† Two fundamental principles for the Bayesian paradigm

- Sufficiency principle
- Likelihood principle

† Sufficient Statistics

- **Def 5.2.1 (Casella & Berger)** Let $x_1, \ldots, x_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(x_1, \ldots, x_n)$. Then the random variable or random vector $T(x_1, \ldots, x_n)$ is called a *statistic*. The probability distribution of $T(x_1, \ldots, x_n)$ is called the *sampling distribution* of T.

  e.g. If an independent sample $x_1, \ldots, x_n$ is taken, the sample mean $\bar{x} = \sum_{i=1}^{n} x_i / n$, the sample variance $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)$ and the sample standard deviation $s = \sqrt{s^2}$ are statistics that are often used and provide good summaries of the sample.

† Sufficient Statistics–contd

- **Def 1.3.1** When $x \sim f(x \mid \theta)$, a function T of x (also called a statistic) is said to be *sufficient* if the distribution of x conditional upon $T(x)$ does not depend on θ. h(x | T(x))

- *How to show that a certain statistic $T(x)$ is or is not a sufficient statistic?* Use the **Fisher–Neyman factorization lemma**.

  Under some measure theoretic regularity conditions, the likelihood can be represented as

$$f(x \mid \theta) = g(T(x) \mid \theta) \boxed{h(x \mid T(x))}$$

  $\Rightarrow$ $T(x)$: a function of data which summarizes all the available *sample* information concerning θ

  $\Rightarrow$ Any additional information in the sample, besides the value of the sufficient statistic, does not contain any more information about θ.

- **Casella & Berger: Example 6.2.3** Consider $x_1, \ldots, x_n$ be iid Bernoulli random variables with unknown parameter θ, $0 < \theta < 1$. Show $T(x) = x_1 + \ldots + x_n$ is a sufficient statistic for θ.

$x_i \mid \theta \overset{iid}{\sim} \text{Ber}(\theta)$, $x_i \in \{0, 1\}$ $\quad 0 < \theta < 1 \quad \frac{1}{\binom{5}{3}}$

$x = (x_1, \ldots, x_5) \quad n = 5$

$= (1, 1, 1, 0, 0)$

or $(1, 0, 1, 0, 1)$

$$f(x \mid \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\overbrace{\sum x_i}^{=T}} (1-\theta)^{n - \overbrace{\sum x_i}^{=T}}$$

$$= \underbrace{\frac{1}{\binom{n}{t}}}_{h(x \mid t)} \cdot \underbrace{\binom{n}{t} \theta^t (1-\theta)^{n-t}}_{g(t \mid \theta)}$$

$T = \sum x_i \sim \text{Bin}(n, \theta)$

$\Rightarrow$ t is sufficient by the factorization lemma.

- **Example 1.3.2** Consider $x_1, \ldots, x_n$ independent observations from a normal distribution $\mathrm{N}(\mu, \sigma^2)$ where μ and σ^2 are unknown.
  - By the factorization theorem, the pair $T(x) = (\bar{x}, s^2)$ where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$ is a sufficient statistic for the parameter (μ, σ).

$$f(x \mid \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i \pm \bar{x} - \mu)^2\right)$$

$$= \underbrace{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\underbrace{\sum_{i=1}^{n}(x_i-\bar{x})^2}_{=s^2} - \frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right)}_{= g(t \mid \theta)}$$

$$h(x \mid \mu, \sigma^2) = 1$$

$\Rightarrow$ $(\bar{x}, s^2)$ are sufficient.

$x = (1, 1, 0, 0, 1) \rightarrow t(x) = 3$

$y = (1, 1, 1, 0, 0) \rightarrow t(y) = 3$

† Sufficiency Principle

- **Sufficiency Principle** Two observations x and y factorizing through the same value of a sufficient statistic T, that is, such that $T(x) = T(y)$, must lead to the same inference.

- If principle is adopted, all inference about θ should depend on sufficient statistics since $\ell(\theta) \propto g(T(x), \theta)$.

- Sometimes criticized since it assumes that the statistical model is the one underlying the data generation.

$$\ell(\theta) = f(x \mid \theta) = g(T(x) \mid \theta) \cdot h(x \mid T(x)) \propto g(T(x) \mid \theta)$$

$(1, 1, 1, 0, 0)$

$(0, 0, 1, 1, 1)$

$x_i \mid \theta \overset{iid}{\sim} \text{Ber}(\theta)$