

- In practice, the first step in the analysis is the same as in frequentist framework
 - we have to decide/specify how we are best going to describe the data that we have
 - once we have this, we have to choose a prior
 - can do conjugate, but only one that goes w/ the data model we chose
 - ⇒ This gives a posterior w/ a nice form, but it is not very flexible (limited # of shapes for a given prior)

→ We could also use a noninformative prior, but this will likely lead to an ugly posterior
 ⇒ then the question becomes: how we are going to sample from this ugly form??
 ⇒ can use strategy in Lab - Single-parameter-inference

Single parameter inference

STA 427/527, Fall 2019, Xin Wang

2.4 Choices of priors

2.4.1 Conjugate priors

- **Conjugacy of priors:** If \mathcal{F} is a class of sampling distributions $p(y|\theta)$, and \mathcal{P} is a class of prior distributions $\pi(\theta)$, then the class \mathcal{P} is conjugate for \mathcal{F} if $p(\theta|y) \in \mathcal{P}$.

we want to choose a prior such that the prior & posterior distributions are in the same class of distributions
 ⇒ Keeps derivation of the posterior easier (this is the only way to guarantee a nice posterior dist)

- **Example 1:** Beta is conjugate for Binomial

– $Y_1, Y_2, \dots, Y_n \stackrel{iid}{\sim} \text{Bin}(m, \theta)$, where m is known, $\theta \sim \text{Beta}(\alpha, \beta)$ $\underline{y} = (y_1, \dots, y_n)$

end result → – Posterior: $\theta|y \sim \text{Beta}(\alpha + \sum_{i=1}^n y_i, \beta + mn - \sum_{i=1}^n y_i)$

$$\begin{aligned}
 \hookrightarrow \text{show this} \rightarrow \text{Data dist} \rightarrow p(\underline{y}|\theta) &= p(y_1, \dots, y_n|\theta) = \prod_{i=1}^n p(y_i|\theta) \propto \prod_{i=1}^n (y_i)^\alpha \theta^{y_i} (1-\theta)^{m-y_i} \\
 &\downarrow \\
 &= \left(\prod_{i=1}^n (y_i) \right) \theta^{\sum_{i=1}^n y_i} (1-\theta)^{mn - \sum_{i=1}^n y_i} \\
 \Rightarrow \text{prior} \rightarrow \pi(\theta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\
 \rightarrow \text{Posterior} \rightarrow p(\theta|\underline{y}) &\propto p(\underline{y}|\theta) \pi(\theta) \\
 &\propto \theta^{\sum_{i=1}^n y_i} (1-\theta)^{mn - \sum_{i=1}^n y_i} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\
 &= \theta^{\left(\sum_{i=1}^n y_i + \alpha\right) - 1} (1-\theta)^{\left(mn - \sum_{i=1}^n y_i + \beta\right) - 1} \\
 &\downarrow \\
 &\sim \text{Beta}\left(\sum_{i=1}^n y_i + \alpha, mn - \sum_{i=1}^n y_i + \beta\right) \Rightarrow E(\theta|\underline{y}) = \frac{\sum_{i=1}^n y_i + \alpha}{mn + \beta + \alpha} \quad \text{Posterior mean}
 \end{aligned}$$

- **Example 2:** Gamma is conjugate for Poisson

– $Y_1, Y_2, \dots, Y_n \stackrel{iid}{\sim} \text{Poisson}(\lambda)$, prior of λ is $\lambda \sim \text{Gamma}(\alpha, \beta)$

$$\begin{aligned}
 \rightarrow \text{Data dist} \rightarrow p(\underline{y}|\lambda) &= p(y_1, \dots, y_n|\lambda) = \prod_{i=1}^n p(y_i|\lambda) \propto \prod_{i=1}^n \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \\
 &\downarrow \\
 &= \left(\prod_{i=1}^n y_i! \right) \lambda^{\sum_{i=1}^n y_i} e^{-n\lambda}
 \end{aligned}$$

$$\Rightarrow \text{prior} \rightarrow \pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\begin{aligned}
 \rightarrow \text{Posterior} \rightarrow p(\lambda|\underline{y}) &\propto p(\underline{y}|\lambda) \pi(\lambda) \\
 &\propto \lambda^{\sum_{i=1}^n y_i} e^{-n\lambda} \lambda^{\alpha-1} e^{-\beta\lambda} \\
 &= \lambda^{\left(\sum_{i=1}^n y_i + \alpha\right) - 1} e^{-\left(n + \beta\right)\lambda} \\
 &\sim \text{Gamma}\left(\sum_{i=1}^n y_i + \alpha, n + \beta\right) \Rightarrow E(\lambda|\underline{y}) = \frac{\sum_{i=1}^n y_i + \alpha}{n + \beta} \quad \text{Posterior mean}
 \end{aligned}$$

Now for new observations

- Question: What is prior predictive distribution and posterior predictive distribution? \rightarrow Poisson example

\rightarrow we need to have an actual estimate of our model (bc we don't have data yet)

\Rightarrow first we make the prior predictive dist

$$\textcircled{1} \rightarrow \text{Prior predictive dist} \rightarrow p(\tilde{y}) = \int p(\tilde{y}|\lambda) \pi(\lambda) d\lambda = \int p(\tilde{y}|\lambda) \frac{\lambda^{\tilde{y}} e^{-\lambda}}{\Gamma(\tilde{y})} \frac{\alpha^{\alpha} e^{-\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda} d\lambda$$

$$= \frac{\alpha^{\alpha} e^{-\alpha}}{\Gamma(\alpha)} \frac{1}{\tilde{y}!} \int \lambda^{\tilde{y}+\alpha-1} e^{-(\alpha+1)\lambda} d\lambda$$

$$= \frac{(\tilde{y}+\alpha)!}{(\tilde{y}+\alpha-\alpha)!} \frac{(\tilde{y}+\alpha-1)!}{\tilde{y}!}$$

$$= \frac{(\tilde{y}+\alpha-1)!}{(\tilde{y}-1)! \tilde{y}!} \left(\frac{\alpha}{\alpha+1} \right)^{\tilde{y}} \left(\frac{1}{\alpha+1} \right)^{\tilde{y}}$$

$$\sim \text{Negative Binomial } (\alpha, \frac{\alpha}{\alpha+1})$$

$\tilde{y} = 0, 1, 2, \dots$

if α is an integer

$\Rightarrow \tilde{y} = \# \text{ failures until } \alpha \text{ successes w/ prob } \frac{\alpha}{\alpha+1}$

\rightarrow If $X \sim \text{Gamma}(\alpha, \beta)$

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

\rightarrow If $X \sim NB(r, p)$

 $f(x) = \binom{x-1}{r-1} p^r q^{x-r}, x = \# \text{ trials until } r \text{ successes}$

OR

 $f(y) = \binom{y+r-1}{r} p^r q^y, y = \# \text{ failures} \dots$

\rightarrow after data collection, we can then use the additional information we have

- \rightarrow The process of finding the posterior predictive is the same as finding the prior predictive, except we replace the prior w/ the posterior of θ bc it takes into account the data we collected

$$\textcircled{2} \rightarrow \text{Posterior Predictive dist} \rightarrow p(\tilde{y}|y) = \int p(\tilde{y}|\lambda) \frac{p(\lambda|y)}{\text{posterior}} d\lambda = \int \frac{\tilde{y}^{\tilde{y}-1}}{\tilde{y}!} \left(\frac{(n+\beta) \sum y_i + \alpha}{\Gamma(\sum y_i + \alpha)} \right)^{\sum y_i + \alpha - 1} \lambda^{\sum y_i + \alpha - 1} e^{-(n+\beta)\lambda} d\lambda$$

$$= \frac{(n+\beta) \sum y_i + \alpha}{\Gamma(\sum y_i + \alpha)} \frac{1}{\tilde{y}!} \int \lambda^{\sum y_i + \alpha - 1} e^{-(n+\beta+1)\lambda} d\lambda$$

$$= \frac{\Gamma(\sum y_i + \alpha + \tilde{y})}{\Gamma(\sum y_i + \alpha)} \frac{\tilde{y}^{\tilde{y}-1}}{\Gamma(\sum y_i + \alpha + \tilde{y})}, \tilde{y} = 0, 1, 2, \dots$$

$$\underset{\alpha > 0 \text{ integer}}{=} \frac{(\tilde{y} + \sum_{i=1}^n y_i + \alpha - 1)!}{(\sum_{i=1}^n y_i + \alpha - 1)! \tilde{y}!} \left(\frac{n+\beta}{n+\beta+1} \right)^{\sum y_i + \alpha} \left(\frac{1}{n+\beta+1} \right)^{\tilde{y}}$$

$$\sim \text{Negative Binomial } (\sum y_i + \alpha, \frac{n+\beta}{n+\beta+1})$$

$$\Rightarrow \text{Posterior predictive mean}$$

$$E(\tilde{y}|y) = \frac{(\sum y_i + \alpha)(\frac{1}{n+\beta+1})}{\frac{n+\beta}{n+\beta+1}} = \frac{\sum y_i + \alpha}{n+\beta}$$

\hookrightarrow This is a combo of info from the data ($\sum y_i + \alpha$) & the prior ($\alpha + \beta$)

- \rightarrow In simulation \rightarrow often we cannot derive this posterior predictive dist (or it would take lots of effort)
- \Rightarrow So instead we can use simulation

\rightarrow 2 steps

- \rightarrow 1) Sample $\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(M)}$ from posterior dist $\sim \text{Gamma}(\sum y_i + \alpha, n+\beta)$
- \rightarrow 2) Sample $\tilde{y}^{(m)}, m=1, 2, \dots, M$ from $\text{Poisson}(\lambda^{(m)})$

→ Posterior means as weighted average of data likelihood + prior mean

⇒ Beta - Binomial Example

$$E(\theta | \mathcal{Y}) = \frac{\sum y_i + \alpha}{n\mu + \alpha + \beta}$$

↳ sample size

$$= \frac{\frac{\sum y_i}{n\mu} (\underbrace{n\mu}_{\text{frequency}}) + \frac{\alpha}{\alpha + \beta} (\underbrace{\alpha + \beta}_{\text{prior mean}})}{n\mu + \alpha + \beta}$$

↓

$$= \left\{ \begin{array}{l} \frac{\sum y_i}{n\mu} \left(\frac{n\mu}{n\mu + \alpha + \beta} \right) + \frac{\alpha}{\alpha + \beta} \left(\frac{\alpha + \beta}{n\mu + \alpha + \beta} \right) \\ \text{frequentist point estimate} \quad \text{weight} \quad \text{prior mean} \quad 1 - \text{weight} \\ = \hat{P}_{MLE} \text{ for } \mu \quad \lim_{n \rightarrow \infty} = 1 \quad \lim_{n \rightarrow \infty} = 0 \end{array} \right.$$

as the sample size increases, the first weight increases

the first term dominates the posterior mean
(data likelihood)

specifically $\lim_{n \rightarrow \infty} E(\theta | \mathcal{Y}) = \hat{P}_{MLE}$

e.g. $y_1, y_2, y_3 \stackrel{iid}{\sim} \text{Bin}(n=5, p)$

↳ $y = (3, 4, 3) \Rightarrow \hat{P}_{MLE} = \frac{(\sum y_i)}{n} = \frac{10}{15} = \frac{2}{3}$

↳ $\bar{y} = \frac{\sum y_i}{n}$

→ Poisson - Gamma

$$E(\lambda | \mathcal{Y}) = \frac{\sum y_i + \alpha}{n + \beta}$$

↓

$$= \frac{\sum y_i \cdot \underbrace{\frac{1}{\lambda}}_{\text{data MLE weight}} + \frac{\alpha}{\beta} \cdot \underbrace{\frac{\lambda}{\alpha + \beta}}_{\text{prior mean}}}{n + \beta}$$

$$= \left\{ \begin{array}{l} \frac{\sum y_i}{n} \left(\frac{n}{n + \beta} \right) + \frac{\alpha}{\beta} \left(\frac{\beta}{n + \beta} \right) \\ \text{data MLE weight} \quad \text{prior mean} \quad 1 - \text{weight} \end{array} \right.$$

- Example 3: Conjugate priors for exponential family

A single-parameter exponential family: The family consists of any distribution whose pmf/pdf can be written as

$$f(y|\theta) = \exp \{a(y)b(\theta) + c(\theta) + d(y)\} \quad a(\cdot), b(\cdot), c(\cdot) + d(\cdot) \text{ are known functions}$$

where $a(y)$ and $d(y)$ do not depend on θ , $b(\theta)$ and $c(\theta)$ do not depend on y .

- If we have n iid observations from the distribution above. The joint distribution of $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ is

$$\begin{aligned} \rightarrow \text{Data dist} \rightarrow p(\mathbf{y}|\theta) &= \prod_{i=1}^n \exp \left\{ a(y_i) b(\theta) + c(\theta) + d(y_i) \right\} \\ &\downarrow \\ &= \exp \left\{ b(\theta) \sum_{i=1}^n a(y_i) + n c(\theta) + \sum_{i=1}^n d(y_i) \right\} \end{aligned}$$

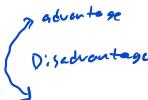
- A conjugate prior for θ is

$$\text{general form} \rightarrow \pi(\theta) = g(k, \gamma) \exp [k \cdot \gamma \cdot b(\theta) + kc(\theta)]$$

- Posterior distribution

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto p(\mathbf{y}|\theta) \pi(\theta) \\ &= \exp \left\{ b(\theta) \sum_{i=1}^n a(y_i) + n c(\theta) + \sum_{i=1}^n d(y_i) \right\} * \left[g(k, \gamma) \exp \{ k \gamma b(\theta) + kc(\theta) \} \right] \\ &\propto \exp \left\{ b(\theta) \left(\sum_{i=1}^n a(y_i) + k\gamma \right) + (n+k)c(\theta) \right\} \end{aligned}$$

- Comments



- Easy to implement
- Not flexible enough to allow other shapes of priors \rightarrow not flexible enough to allow other shapes of priors
e.g. if have gamma prior \Rightarrow then can only have prior shapes like 
- Still possible to incorporate prior knowledge about θ

- Some conjugate priors

Data likelihood	Prior
Bernoulli	Beta
Binomial	Beta
Poisson	Gamma
Normal(σ^2 known)	Normal for μ
Normal(μ known)	Inverse-gamma for σ^2
Uniform($0, \theta$)	Pareto for θ
Exponential	Gamma
Gamma(β is unknown, α is known)	Gamma

Choosing Prior

- Cannot mix + match priors + data dists just b/c they are in exponential family \Rightarrow choices have to follow the table
- So best practice is to choose a prior that meets your conditions so that the posterior will have a nice form
 - Conjugate priors guarantee the posterior has the same parametric form, which is obviously good
 - But, this also locks in the shape of the posterior (even if we chose a prior w/ a large variance) \Rightarrow not flexible again
- Ultimately, choice of prior is subjective \Rightarrow this is one of the main criticisms of Bayesian methods

2.5 Noninformative priors

1. Uniform prior \rightarrow This is one example of an uninformative prior b/c we think all possible values of the parameter are equally likely

- Example: If we toss a coin, $Y \sim \text{Bernoulli}(p)$; $p \sim \text{Uniform}(0, 1)$

(reparameterizing
(to illustrate a problem))

$$\begin{aligned} & \rightarrow \text{Let } \pi(\alpha) = 1 \quad 0 < \alpha < 1 \\ & \rightarrow \text{If considering transformation } \alpha = \frac{p}{1-p} \quad (+>0) \\ & \Rightarrow \pi(\alpha) = ? \quad (\text{goal is to find this}) \quad \Rightarrow p = \frac{\alpha}{\alpha+1} \quad \Rightarrow \frac{dp}{d\alpha} = \frac{1}{(\alpha+1)^2} \\ & \rightarrow \text{need to calculate } |J| \quad (\text{Jacobian}) = \left| \frac{dp}{d\alpha} \right| = \frac{1}{(\alpha+1)^2} \\ & \pi(\alpha) = \pi(p) |J| = 1 \cdot \frac{1}{(\alpha+1)^2} = \frac{1}{(\alpha+1)^2} \end{aligned}$$

Problem \rightarrow After doing the transformation, not all α values are equally likely (ie properties of the prior changed)
unless \rightarrow generally true after a transformation

2. Jeffrey's prior: Fisher information

\rightarrow solves problem

$$I(\theta) = J(\theta) = E \left[\left(\frac{\partial \ln L(y|\theta)}{\partial \theta} \right)^2 \right] = -E \left[\frac{\partial^2 \ln L(y|\theta)}{\partial \theta^2} \right] = -E \left[\frac{\partial^2 \ln \pi(\theta)}{\partial \theta^2} \right]$$

Then the corresponding Jeffrey's prior is $p(\theta) \propto \sqrt{I(\theta)}$

\rightarrow Consider $\varphi = h(\theta)$
new parameterization (some function of original parameter)

\rightarrow It turns out $\pi_1(\varphi) \propto \sqrt{I(\varphi)}$ \rightarrow new prior

\Rightarrow These priors are related by the usual change of variables theorem
 $p(y) = p(\varphi) \left| \frac{d\varphi}{dy} \right|$

- Example: Binomial data $Y|\theta \sim \text{Binomial}(n, \theta)$

$$\rightarrow p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

$$\rightarrow \ln(p(y|\theta)) = \ln(\binom{n}{y}) + y \ln(\theta) + (n-y) \ln(1-\theta)$$

$$\rightarrow \frac{\partial \ln(p(y|\theta))}{\partial \theta} = \frac{y}{\theta} - \frac{n-y}{1-\theta}$$

$$\rightarrow \frac{\partial^2 \ln(p(y|\theta))}{\partial \theta^2} = \frac{-y}{\theta^2} - \frac{n-y}{(1-\theta)^2}$$

$$\rightarrow I(\theta) = -E \left[\frac{\partial^2 \ln(p(y|\theta))}{\partial \theta^2} \right] = -E \left[\frac{-y}{\theta^2} - \frac{n-y}{(1-\theta)^2} \right]$$

$$= \frac{1}{\theta^2} E(y) + \frac{1}{(1-\theta)^2} E(n-y) \quad \rightarrow E(y) = n\theta$$

$$= \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = n\theta$$

$$= \frac{n(1-\theta) + n\theta}{\theta(1-\theta)} = \frac{n}{\theta(1-\theta)}$$

$$= \frac{n - n\theta + n\theta}{\theta(1-\theta)} = \frac{n}{\theta(1-\theta)}$$

$$= \frac{n}{\theta(1-\theta)}$$

lots of algebra

\Rightarrow The Jeffrey's prior is $\pi(\theta) \propto \sqrt{I(\theta)}$

$$= \sqrt{\frac{n}{\theta(1-\theta)}}$$

$$\propto \theta^{-1/2} (1-\theta)^{-1/2}$$

$$\sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$$