VE414 Lecture 4

Jing Liu

UM-SJTU Joint Institute

May 23, 2019

- Recall using Beta with Binomial results another Beta, it is not the only case.

Definition

A class $\mathcal{P}$ of prior distributions for Y is known to be conjugate to a likelihood

$$\mathcal{L}(y; x) = f_{X|Y}(x \mid y)$$

that is, to a data generating model, if the posterior is also in the same class

$$f_{Y|X} \in \mathcal{P}$$

The prior $f_Y \in \mathcal{P}$ is called a conjugate prior for the likelihood.

$$f_Y(y) \in \mathcal{P} \implies f_{Y|X}(y \mid x) \in \mathcal{P}$$

- It is clear that the beta distribution is a conjugate prior for binomial.

Q: Can you think of any other conjugate distributions?

- Suppose $X \mid \mu \sim \text{Normal}\left(\mu, \sigma^2\right)$, where the variance σ^2 is known, consider

$$\mu \sim \text{Normal}\left(\nu, \omega^2\right)$$

we have the following likelihood function

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

and the prior is taken to be normal as well with the following density function

$$g\left(\mu\right) = \frac{1}{\sqrt{2\pi\omega^2}} \exp\left(-\frac{(\mu-\nu)^2}{2\omega^2}\right)$$

- According to Bayes' theorem, the posterior density function h must satisfies

$$h\left(\mu \mid x\right) \propto \mathcal{L}(\mu; x, \sigma^2) g\left(\mu\right)$$

- Multiplying $\mathcal{L}$ and g together, we have

$$\mathcal{L}(\mu; x, \sigma^2) g(\mu) = \frac{1}{\sqrt{2\pi\sigma^2 \cdot 2\pi\omega^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\nu)^2}{2\omega^2} \right)$$

- Completing the square and dropping the following which does not involve μ,

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

we obtain the posterior PDF of μ up to a multiplicative constant

$$h(\mu \mid x) \propto \exp\left(-\frac{1}{2} \frac{\omega^2 + \sigma^2}{\sigma^2 \omega^2} \left(\mu - \frac{\omega^2 x + \sigma^2 \nu}{\omega^2 + \sigma^2} \right)^2 \right)$$

from which we can conclude the posterior is also a normal distribution

$$\mu \mid x \sim \text{Normal}\left(\frac{\omega^2 x + \sigma^2 \nu}{\omega^2 + \sigma^2}, \frac{\sigma^2 \omega^2}{\omega^2 + \sigma^2} \right)$$

- Now suppose we have a dataset of independent and identically distributed

$$X_i \mid \mu \sim \text{Normal}\left(\mu, \sigma^2\right)$$

where σ^2 is known, and we again consider the prior

$$\mu \sim \text{Normal}\left(\nu, \omega^2\right)$$

- Since X_i are i.i.d., the likelihood of n observations is simply the product

$$\begin{aligned}
\mathcal{L}(\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\left(2\sigma^2\right)^{-1} \sum_{i=1}^{n} (x_i-\mu)^2\right) \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\left(2\sigma^2\right)^{-1} \left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2\right)\right)
\end{aligned}$$

- Dropping multiplicative terms that does not depend on μ, we have

$$\begin{aligned}
\mathcal{L}(\mu) &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\left(2\sigma^2\right)^{-1}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)\right) \\
&\propto \exp\left(-\left(2\sigma^2\right)^{-1}\left(-2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)\right) \\
&\propto \exp\left(-\left(2\sigma^2\right)^{-1} n\left(-2\mu\bar{x} + \mu^2\right)\right) \qquad \text{where} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \\
&\propto \exp\left(-\left(2\sigma^2\right)^{-1} n\left(\bar{x} - \mu\right)^2\right)
\end{aligned}$$

  where the square is completed since the term needed does not depend on μ.

- Firstly notice $\bar{x}$ is sufficient, the individual x_i values are not needed

$$\mathcal{L}\left(\mu; \bar{x}, \sigma^2\right) \propto \exp\left(-\left(2\sigma^2\right)^{-1} n\left(\bar{x} - \mu\right)^2\right)$$

- Second thing to notice from this derivation is the fact that the likelihood

$$\mathcal{L}\left(\mu; \bar{x}, \sigma^2\right) \propto \exp\left(-\frac{\left(\bar{x}-\mu\right)^2}{2\left(\sigma/\sqrt{n}\right)^2}\right)$$

is identical to the likelihood of having $\bar{x}$ as a single realisation from a normal distribution with mean of μ and variance of $\frac{\sigma^2}{n}$, which means having n i.i.d. observations can be treated as having a single observation, for which we have

$$\mu \mid x \sim \text{Normal}\left(\frac{\omega^2 x + \sigma^2 \nu}{\omega^2 + \sigma^2}, \frac{\sigma^2 \omega^2}{\omega^2 + \sigma^2}\right)$$

with the prior $\mu \sim \text{Normal}\left(\nu, \omega^2\right)$, so this prior is again a conjugate prior

$$\mu \mid \bar{x} \sim \text{Normal}\left(\frac{\omega^2 \bar{x} + \nu\sigma^2/n}{\omega^2 + \sigma^2/n}, \frac{\omega^2 \sigma^2/n}{\omega^2 + \sigma^2/n}\right)$$

- For a comprehensive list of other conjugate priors, see the wikipedia table.

- Conjugate priors are often used to avoid explicitly evaluating

$$\text{normalising constant} = \int_{-\infty}^{\infty} f_{X|Y}\left(x \mid y\right) f_Y\left(y\right)\, dy$$

- The integral needs to be evaluated in order to normalise the posterior

$$\begin{aligned} f_{Y|X} &= \text{Posterior} \propto \text{Likelihood} \times \text{Prior} \\ &= f_{X|Y} \times f_Y / \text{normalising constant} \end{aligned}$$

  so that we have a proper density function $f_{Y|X}$.

- However, often having a prior that correctly reflects our prior knowledge out weight the computational saving in modern era since the rise of computers.
- Once we have data, the posterior can be used as the prior for future data.

$$? \longrightarrow f_{Y|X_1} \longrightarrow f_{Y|X_1^*} \longrightarrow f_{Y|X_1^{**}}$$

Q: How about the very first prior?

- Recall to reflect the fact that we have no prior knowledge regarding where the black ball is before rolling any red ball, Bayes advocated using

$$\text{Unif}\,(0,1)$$

as the prior since it does not prefer any particular $p \in [0,1]$ over another.

- Using the uniform prior when no prior knowledge exists seems very nature and convincing, however, there are two issues we need to address.
- The first issue can be identified by considering the following example again

$$X_i \mid \mu \overset{\text{i.i.d.}}{\sim} \text{Normal}\left(\mu, \sigma^2\right)$$

where σ^2 is known and there are n such data points.

- It is clear that the mean could be any real number

$$-\infty < \mu < \infty$$

- Recall the density function of a uniform distribution in general is $f_Y \propto 1$.
- However, there is no constant c such that

$$\int_{-\infty}^{\infty} c\, dy < \infty$$

  so it is not strictly possible to have a uniform distribution over $(-\infty, \infty)$.
- In other words, the uniform prior over $(-\infty, \infty)$ cannot be normalised, such priors, which are strictly not a distribution, are known as improper priors.
- Technically, an improper prior does not lead to any posterior, however,

$$\int_{-\infty}^{\infty} f_{X|Y} f_Y \, dy = \Lambda < \infty$$

  is often computed. If Λ is finite, then $\dfrac{f_{X|Y} f_Y}{\Lambda}$ is interpreted as a posterior in a limiting sense. If Λ is not finite, then the prior should not be used.

- In practice, we simply settle for a flat normal distribution with a big variance

$$\mu \sim \text{Normal}\left(\nu, \omega^2\right), \qquad \text{where } \omega \gg \sigma$$

as the prior. In this case, it means we can exploit the conjugate relationship

$$\mu \mid \bar{x} \sim \text{Normal}\left(\frac{\omega^2 \bar{x} + \nu\sigma^2/n}{\omega^2 + \sigma^2/n}, \frac{\omega^2\sigma^2/n}{\omega^2 + \sigma^2/n}\right)$$

- Since having a large ω reduces the role that the prior has on the result, i.e.

$$\frac{\omega^2 \bar{x} + \nu\sigma^2/n}{\omega^2 + \sigma^2/n} = \frac{\omega^2}{\omega^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\omega^2 + \sigma^2/n}\nu$$

using which reflects a limited prior knowledge at the very beginning.

- However, this approach also suffers the other issue that a uniform prior has.

Q: Does a uniform/flat prior preserve our ignorance under reparametrization?

- To isolate the second issue, suppose we need a prior on a finite interval

$$[a, b]$$

which means the uniform prior is simply

$$f_Y(y) = \frac{1}{b-a} \qquad \text{for} \quad y \in [a, b]$$

- However, consider the following smooth monotone transformation (SMT),

$$u = g(y) = \exp(y) \iff y = g^{-1}(u) = \ln(u)$$

then, according to COVT, the new variable must have the following PDF

$$g_U(u) = \frac{1}{b-a}\frac{1}{u}$$

Q: What does it mean in terms of posteriors $f_{U|X}$ and $f_{Y|X}$?

- The inconsistency in the two posteriors (either of y or u) is one of the main reasons why Bayesian analysis was criticised and dismissed after the death of Laplace, who also advocated using a uniform prior.
- The inconsistency was addressed by **Jeffreys**[1], who proposed using the prior

$$f_Y(y) \propto \sqrt{I(y)}$$

where $I(y)$ is the variance of $\dfrac{\partial \ln \mathcal{L}(y; X)}{\partial y}$ conditional on y

$$I(y) = \operatorname{Var}\left[\frac{\partial \ln \mathcal{L}(y; X)}{\partial y}\middle| y\right] = -\mathbb{E}\left[\frac{\partial^2 \ln \mathcal{L}(y; X)}{\partial y^2}\middle| y\right]$$

- This, which reflects the lack of prior knowledge, is called the Jeffreys Prior

$$\varphi_Y$$

[1] Harold Jeffreys. "An invariant form for the prior probability in estimation problems". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186.1007 (1946), pp. 453–461.

Q: What is the Jeffreys prior for the binomial likelihood?

$$\mathcal{L}(p;x) = \frac{k!}{x!(k-x)!}p^x(1-p)^{k-x}$$

- To see why Jeffreys prior will not create the inconsistency, let the following

$$\mathcal{L}(y;x) = f_{X|Y}(x \mid y)$$

be some likelihood in general, and consider some alternative parametrisation

$$\left(\mathcal{L} \circ g^{-1}\right)(u;x) = \mathcal{L}\left(g^{-1}(u);x\right)$$

where $u = g(y)$ is a SMT, then we have two Jeffrey priors in this setting, i.e.

$$\varphi_Y \propto \sqrt{-\int_{-\infty}^{\infty} \frac{\partial^2 \ln \mathcal{L}(y;x)}{\partial y^2} f_{X|y}(x \mid y)\, dx}$$

$$\varphi_U \propto \sqrt{-\int_{-\infty}^{\infty} \frac{\partial^2 \left(\ln \mathcal{L} \circ g^{-1}\right)(u;x)}{\partial u^2} f_{X|y}(x \mid g^{-1}(u))\, dx}$$

- Note there are **3 ways** to obtain a posterior of U based on Jeffreys priors,

$$f_{U|X} \propto \left(\mathcal{L} \circ g^{-1}\right) \cdot \left(\frac{\varphi_Y}{|g'|} \circ g^{-1}\right)$$

$$f^{*}_{U|X} \propto \left(\mathcal{L} \circ g^{-1}\right) \cdot \varphi_U$$

$$f^{**}_{U|X} \propto \frac{f_{Y|X}}{|g'|} \circ g^{-1} \qquad \text{where} \quad f_{Y|X} \propto \mathcal{L} \cdot \varphi_Y$$

- With some manipulation, we see the first and third ways are always the same

$$f_{U|X} = f^{**}_{U|X} \propto \frac{\mathcal{L} \cdot \varphi_Y}{|g'|} \circ g^{-1}$$

- Understand why the following is true

$$\varphi_U \propto \frac{\varphi_Y}{|g'|} \circ g^{-1}$$

will show why a Jeffreys prior will not create the inconsistency i.e. invariant.

- By considering the integral that defines φ_U, we obtain the desired result

$$\varphi_U \propto \left(-\int_{-\infty}^{\infty} \frac{\partial^2 \left(\ln \mathcal{L} \circ g^{-1} \right)(u; x)}{\partial u^2} f_{X|Y}\left(x \mid g^{-1}(u) \right) \, dx \right)^{1/2}$$

$$= \left(-\int_{-\infty}^{\infty} \left[\frac{\partial^2 \ln \mathcal{L}}{\partial y^2} \left(\frac{dy}{du} \right)^2 + \frac{\partial \ln \mathcal{L}}{\partial y} \frac{d^2 y}{du^2} \right] f_{X|Y}\left(x \mid g^{-1}(u) \right) \, dx \right)^{1/2}$$

$$= \left(-\int_{-\infty}^{\infty} \frac{\partial^2 \ln \mathcal{L}}{\partial y^2} f_{X|Y}\left(x \mid g^{-1}(u) \right) \, dx \right)^{1/2} \left| \frac{dy}{du} \right| = \frac{\varphi_Y}{|g'|} \circ g^{-1}$$

- To see why the 2nd term varnished, notice $\dfrac{d^2 y}{du^2}$ does not depend on x, and

$$\int_{-\infty}^{\infty} \frac{\partial \ln \mathcal{L}}{\partial y} f_{X|Y} \, dx = \int_{-\infty}^{\infty} \left(\frac{1}{\mathcal{L}} \frac{\partial \mathcal{L}}{\partial y} \right) f_{X|Y} \, dx$$

$$= \int_{-\infty}^{\infty} \left(\frac{\partial \mathcal{L}}{\partial y} \frac{1}{f_{X|Y}} \right) f_{X|Y} \, dx = \frac{\partial}{\partial y} \int_{-\infty}^{\infty} f_{X|Y} \, dx = 0$$

- The above shows that φ_Y is invariant, but why does it reflect our ignorance?

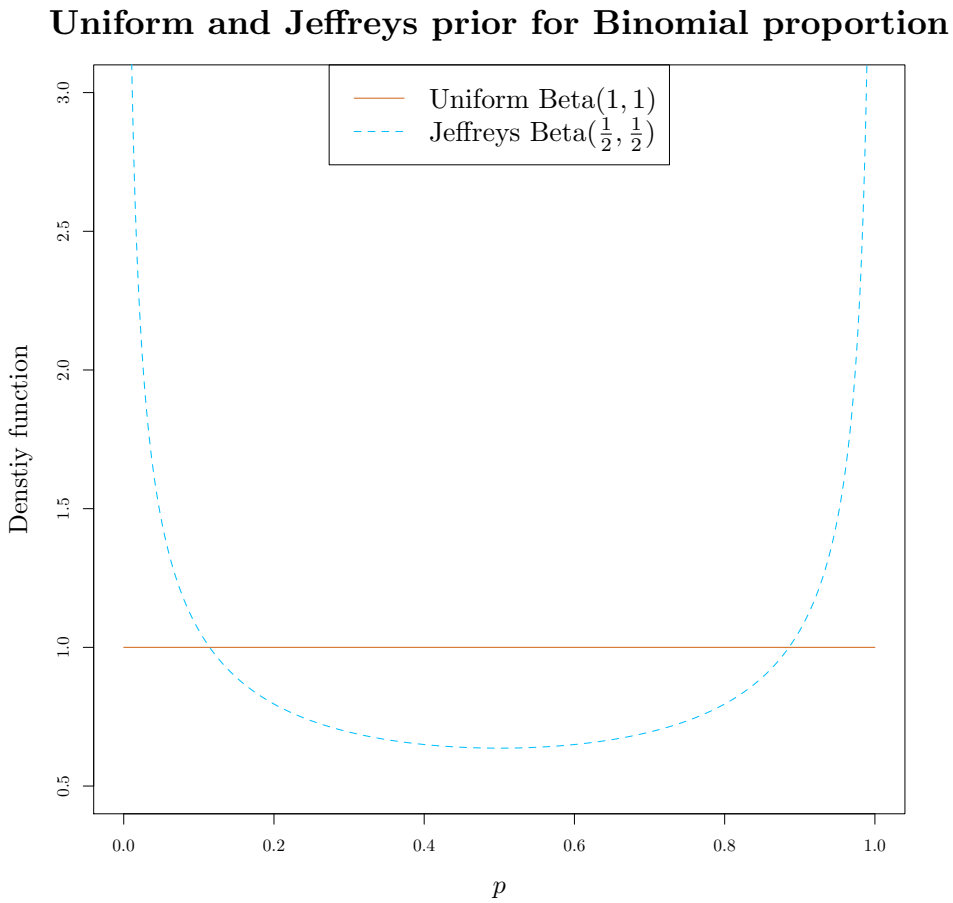

Figure: R Code: `fig_beta_binomial_jeffreys_prior_414.R`

- It was not known to Jeffreys, but the Jeffreys prior reflects our ignorance in the sense that it is the prior that maximises the distance between the prior and posterior, similar priors in high dimension is known as reference priors.

$$\int_{-\infty}^{\infty} f_X \underbrace{\left(\int_{-\infty}^{\infty} f_{Y|X} \ln \left(\frac{f_{Y|X}}{f_Y} \right) dy \right)}_{\text{Kullback-Leibler Divergence}} dx$$

- We will not discuss reference priors further since it is beyond our scope.
- Although we have advanced theorems supporting Jeffreys priors, care must be taken with Jeffreys priors in practice, for example, when σ^2 is known,

$$X_i \mid \mu \overset{\text{i.i.d.}}{\sim} \text{Normal}\left(\mu, \sigma^2\right)$$

I will leave it to you to show the Jeffreys prior in this case is actually uniform

$$\varphi_\mu \propto 1 \qquad \text{for} \quad -\infty < \mu < \infty$$

which means Jeffreys priors sometimes are improper, and care must be taken.