Conjugacy, Likelihood Principle

Posterior mean and posterior variance

What will be discussed…

- More examples on
  - forming the prior distributions
  - finding ‘kernels’
  - computing integral (sum) of the kernel = 1/ normalizing constant

- What are the conjugate priors? How do we find them?

- Posterior mean and posterior variance
  - How are they denoted? How do we find them?

- Informative priors, Reference priors, improper priors

Recap – The Bayesian Framework

i) Suppose we observe an iid sample of data : $\boldsymbol{X} = (X_1,\ldots,X_n)$

Then, our model for the distribution of the data will give us the likelihood

$$L(\theta|\,\boldsymbol{x}) = f\,(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f\,(x_i|\theta)$$

ii) We also must specify $f(\theta)$, the prior distribution for θ, based on any knowledge we have about θ before observing the data.

iii) By Bayes' Rule,

Posterior distribution is $f(\theta|\boldsymbol{x}) = \dfrac{f(\boldsymbol{x}|\theta)f(\theta)}{f(\boldsymbol{x})} = \dfrac{f(\boldsymbol{x}|\theta)f(\theta)}{\int f(\boldsymbol{x}|\theta)f(\theta)d\theta}$

- $f(\boldsymbol{x})$ is a normalizing constant that ensures $f(\theta|\boldsymbol{x})$ integrating to 1. We don't have to calculate it (except posterior probabilities for discrete cases).

$f(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta)f(\theta)$: posterior $\propto$ lkhd x prior

Review: Binomial-Beta example

- Derive posterior by $f(\theta|x) \propto f(x|\theta)f(\theta)$: posterior $\propto$ lkhd x prior

- Suppose that $X|\theta \sim$ Binom(n, θ). Take $\theta \sim$Beta(α, β)

- Lkhd x prior: $f(x|\theta)f(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

Ignoring constants (in θ),

$$f(x|\theta)f(\theta) \propto \theta^x(1-\theta)^{n-x}\ \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$: kernel of Beta(x+α, $n-x$+β)

Then, the posterior distribution is $\theta|x \sim$ Beta(x+α, $n-x$+β)

Remarks

- In a real problem, we need to specify the values of our hyperparameters α and β of our prior.

- Ideally our choices of α and β should reflect our prior beliefs about θ.

  - If we have no prior idea what θ is, we could set α = β = 1, which corresponds to a Uniform(0, 1) prior for θ : completely flat, so that all values of θ are equally likely a priori.

- If we have more informative prior beliefs about the value of θ, we could choose α and β to reflect that.

- Plots of the Beta pdf for various values of α and β can help inform the prior specification (see the plots from the previous lecture).

Forming a Beta prior distribution

- The expected value and variance of a Beta(α, β) r.v. is

$$E[\theta] = \frac{\alpha}{\alpha+\beta} \qquad \text{Var}[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- So if our prior belief is that θ is closer to 0 than to 1, we should choose our hyperparameters α and β such that α < β.

- If our prior belief is strong that θ is near a certain value, we can pick α and β so that this variance is small.

- If our prior belief is less certain, pick α and β so that the variance is large.

- The mode (location where the pdf reaches its maximum) for Beta(α, β) pdf is $\frac{\alpha-1}{\alpha+\beta-2}$

Examples of forming Likelihood functions

- Suppose students in a statistics class conduct a study to estimate the proportion of cars in a Campus Parking lot that are black.

i) Student A observes the first 10 cars and record X, the number that are black : *NB*, *B*, *NB*, *NB*, *NB*, *B*, *NB*, *NB*, *NB*, *B*

$\Longrightarrow x = 3$

ii) Student B decides to to observe cars until the third black one parks and record Y, the total number of cars that drive by until the third black one.

B records $y = 10$.

iii) Student C decides to observe every car and records either 1 or 0 according to whether the *i*'th car is black. $Z : 0\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1$

Examples of forming Likelihood functions

- Suppose students in a statistics class conduct a study to estimate the proportion of cars in a Campus Parking lot that are black.

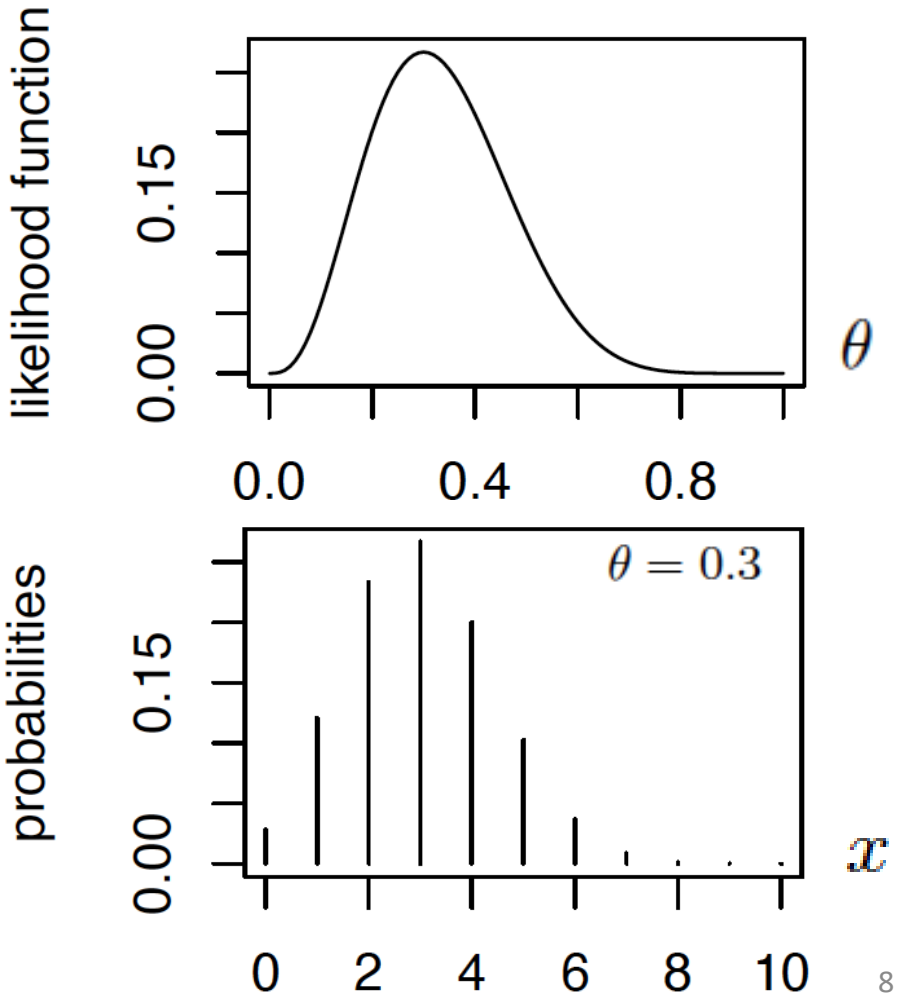
i) Student A observes the first 10 cars and record X, the number that are black :

NB, *B*, *NB*, *NB*, *NB*, *B*, *NB*, *NB*, *NB*, *B*

$\Longrightarrow x = 3$

This is a Binomial experiment
and the statistical model is $X \sim$Bin(10, θ);
the likelihood function is

$$L_A(\theta \mid x) = \binom{10}{3} \theta^3(1-\theta)^7$$



ii) Student B decides to to observe cars until the third black one parks and record Y, the total number of cars that drive by until the third black one.

B records y = 10. For B, the likelihood function is

$L_B(\theta \mid y)$ = P[Y = 10 | θ] = P[2 blacks among first 9 cars] × P[10'th car is black]

$$= \binom{9}{2} \theta^2(1-\theta)^7 \times \theta = \binom{9}{2} \theta^3(1-\theta)^7$$

iii) Student C decides to observe every car and records either 1 or 0 according to whether the *i*'th car is black. Z : 0 1 0 0 0 1 0 0 0 1

The likelihood function is

$$L_C(\theta \mid z) = (1-\theta)\theta\,(1-\theta)(1-\theta)(1-\theta)\,\theta(1-\theta)(1-\theta)(1-\theta)\theta = \theta^3(1-\theta)^7$$

- All experiments yield equal (proportional to) likelihoods and contains the same information. ⟹ They should produce equivalent inference about θ.

Likelihood Principle

1) Suppose that only individual binary values are observed for data:

$X_1,\ldots,X_n \,|\theta \sim$ Bern(θ) : $\quad L(\theta : \boldsymbol{x}) = \prod \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$

2) If we only observe the total number of successes, $S = \sum x_i$

$S|\theta \sim$ Binom(θ) : $\quad L(\theta : \boldsymbol{x}) = \binom{n}{s}\theta^s(1-\theta)^{n-s} \propto \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$

3) Suppose that we sample until we observe 's' number of successes.

(i.e. We do not fix the total number of trials = n). $\quad S = \sum x_i$

$N|\theta \sim$ Negative Binom(θ) : $\quad L(\theta : \boldsymbol{x}) = \binom{n-1}{s-1}\theta^s(1-\theta)^{n-s}$

- In Frequentist approach, different sampling schemes yield different probability models ($\Rightarrow$ different estimates for θ in general)
- But, they all give the same likelihood function $\Rightarrow$ same posterior

Inference for Binary data with Beta prior

- Likelihood : $f(\boldsymbol{x}|\theta) \propto \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$ and Prior: $\theta \sim$ Beta(α, β)

Posterior: $f(\theta|\boldsymbol{x}) \propto f(x|\theta)f(\theta) \propto \theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}$

$\Longrightarrow \quad \theta|\boldsymbol{X} \sim$ Beta

- Posterior mean:

$$\mathrm{E}[\theta|\boldsymbol{X}]=$$

Posterior Variance:

$$\mathrm{Var}[\theta|\boldsymbol{X}]=$$

Inference for Binary data with Beta prior

- Likelihood : $f(\boldsymbol{x}|\theta) \propto \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$ and Prior: $\theta \sim$ Beta(α, β)

Posterior: $f(\theta|\boldsymbol{x}) \propto f(x|\theta)f(\theta) \propto \theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}$

$\Longrightarrow \quad \theta|\boldsymbol{X} \sim$ Beta($\sum x_i$+α, $n-\sum x_i$+β)

- Posterior mean:

$$E[\theta|X] = \frac{\sum x_i + \alpha}{n+\alpha+\beta} = \frac{n\frac{\sum x_i}{n} + \frac{\alpha}{\alpha+\beta}(\alpha+\beta)}{n+\alpha+\beta}$$

: weighted average of $\bar{X}$ & $E[\theta]$

- If $n \longrightarrow \infty$ & α, β fixed, $E[\theta|\boldsymbol{X}] \longrightarrow \bar{X}$ (frequentist estimate).
- If α, β $\longrightarrow \infty$ & $\frac{\beta}{\alpha}$ fixed, $E[\theta|\boldsymbol{X}] \longrightarrow \frac{\alpha}{\alpha+\beta}$ (prior mean)
- If α=β=0, prior dist^n= Beta(0,0)= Unif(0,1), $E[\theta|\boldsymbol{X}] \longrightarrow \bar{X}$

EXAMPLE : Cancer trials

- A new treatment protocol is proposed for a particular form of cancer. The measure of success will be the proportion of patients that survive longer than six months after diagnosis.

- With the present treatment, this success rate is 40%. Letting θ be the success rate of the new treatment, a doctor assesses her prior beliefs about θ as follows. She judges that her expectation of θ is E(θ) = 0.45, and her standard deviation is 0.07. Assume that her beliefs can be represented by a beta distribution.

- A clinical trial of the new treatment protocol is carried out. Out of 70 patients in the trial, 34 survive beyond six months from diagnosis. Is the new treatment protocol better?

Example (analysis)

- $X|\theta \sim$ Binom(n, θ) : n = 70, x = 34
- Prior: $\theta \sim$Beta(α, β) $E[\theta]=\frac{\alpha}{\alpha+\beta}$ $Var[\theta]=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Doctor's experience: $0.45=\frac{\alpha}{\alpha+\beta}$ $0.07^2=\frac{\alpha}{\underbrace{(\alpha+\beta)}_{0.45}}\frac{\beta}{\underbrace{(\alpha+\beta)}_{0.55}}\frac{1}{\underbrace{(\alpha+\beta+1)}}$

$\Rightarrow$ α = 0.45x49.51 ≈ 22.28 β= 27.23

- Posterior: $\theta|X \sim$ Beta(x+α, $n-x$+β) = Beta(56.28, 63.23)
- P[$\theta > 0.4\ |X$]= 0.941 (posterior prob.)

  P[$\theta > 0.4\ ]$ = 0.758 (prior prob.) In R, `>1-pbeta(0.4,22.28,27.23)`

Examples

4) Calculate the integral by identifying a kernel of a pdf

$$\int_0^{\infty} \theta^{\alpha+x-1} e^{-\theta(2\beta)} d\theta$$

- Identify the probability distribution for this kernel of density.

5) $f(x|\theta, \tau) \propto \exp(-\tau x^2 + \theta x) \quad -\infty < x < \infty, \qquad \theta, \tau > 0$

6) $\exp(-2\phi^2 + 12\phi\psi)$

4)

$$\frac{1}{\Gamma(\alpha+x)\left[\frac{1}{2\beta}\right]^{\alpha+x}}\int_0^{\infty}\theta^{\alpha+x-1}e^{-\theta(2\beta)}d\theta=1 \Longrightarrow \int_0^{\infty}\theta^{\alpha+x-1}e^{-\theta(2\beta)}d\theta=\Gamma(\alpha+x)\left[\frac{1}{2\beta}\right]^{\alpha+x}$$

5) $f(x|\theta,\tau) \propto \exp(-\tau x^2+\theta x) = \exp\{-\tau\left(x^2-\frac{\theta x}{\tau}\right)\} \propto \exp\{-\tau\left(x^2-\frac{\theta x}{\tau}+(\frac{\theta}{2\tau})^2\right)\}$

$$= \exp\{-\tau\left(x-\frac{\theta}{2\tau}\right)^2\} = \exp\{-\frac{\left(x-\frac{\theta}{2\tau}\right)^2}{\frac{1}{\tau}\ 2\frac{1}{2}}\}$$: kernel of $N\left(\frac{\theta}{2\tau},\ \frac{1}{2\tau}\right)$

- Note: $\left(x-\frac{\theta}{2\tau}\right)^2 = x^2-\frac{\theta}{\tau}x+\frac{\theta^2}{4\tau^2}$

6) $f(\phi) \propto \exp(-2\phi^2+12\phi\psi) = \exp\{-2(\phi^2-6\phi\psi)\}$

$$\propto \exp\{-2(\phi^2-6\phi\psi+(3\psi)^2\} = \exp\{-2(\phi-3\psi)^2\}$$

$$= \exp\{-\frac{(\phi-3\psi)^2}{\frac{1}{2}\ 2\frac{1}{2}}\}$$: kernel of Normal(3ψ , 1/4)

Conjugacy

- A conjugate prior is the one for which the prior distribution and the posterior distribution have the same family of distributions (same functional form), just with different (updated) parameters.

  - For example, in the Beta-binomial model, the prior is a Beta and the posterior is also a Beta, so this is a conjugate prior.

- Conjugate priors are nice because

  - we can typically derive the posterior without needing any difficult computation (of normalizing constants)

  - it is usually easy to understand the respective contributions of the prior information and the data information to the posterior.

- Examples: Poisson-Gamma, Exponential-Gamma, Normal-Normal, etc.

Example: Poisson-Gamma

- The Poisson distribution is a common model for count data, characterized by a rate λ (expected occurrences per unit time, volume, or other unit), and the number of occurrences x (taking on integers).

- If our data consists of a random sample on n such counts, then the likelihood function is the joint density function $f(\boldsymbol{x}|\lambda) = f(x_1|\lambda)\, f(x_2|\lambda) \cdots f(x_n|\,\lambda)$, since $X_1,\ldots,X_n$ are independent.

- Likelihood: $X_1,\ldots,X_n\,|\lambda \sim^{\text{ind.}} \text{Poisson}(\lambda)$ i.e. $f(x_i|\lambda) = \dfrac{\mathrm{e}^{-\lambda}\,\lambda^{x_i}}{x_i!}$

$$L(\lambda: \boldsymbol{x}) = \prod \frac{\lambda^{x_i}\,\mathrm{e}^{-\lambda}}{x_i!} = \frac{\lambda^{x_1}\cdots\lambda^{x_n}\;\mathrm{e}^{-\lambda}\cdots\mathrm{e}^{-\lambda}}{x_1!\cdots x_n!} = \frac{\lambda^{\sum_{i=1}^{n} x_i}\,\mathrm{e}^{-n\lambda}}{\prod x_i!}$$

- The Gamma distribution is a good choice for the prior, since its support is $(0, \infty)$ as $\lambda > 0$.

- Prior: $\lambda \sim$ Gamma(α, β), $\lambda > 0$

$$f(\lambda) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \lambda^{\alpha-1} \exp\left(-\frac{\lambda}{\beta}\right) \qquad E[\lambda] = \alpha\beta \qquad Var[\lambda] = \alpha\beta^2$$

$$\text{(Then, } \int_0^{\infty} \lambda^{\alpha-1} \exp\left(-\frac{\lambda}{\beta}\right) d\lambda = \Gamma(\alpha)\beta^{\alpha} \text{ : 1/N.C.)}$$

- The parameterization of the Gamma distribution that we will use in this course is a Gamma pdf with a shape parameter α and a scale parameter β.

- Note that the scale parameter is the reciprocal of the rate parameter used in the other parameterization.

Posterior distribution with conjugacy

- Posterior: $f(\lambda|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\lambda)\, f(\lambda)}{\int f(\boldsymbol{x}|\lambda) f(\lambda) d\lambda} \propto f(x|\lambda) f(\lambda)$

$$f(\lambda|\boldsymbol{x}) = \frac{\lambda^{\sum x_i}\, \mathrm{e}^{-n\lambda}}{\prod x_i!} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \lambda^{\alpha-1} \exp\left(\frac{-\lambda}{\beta}\right)$$

$$\propto \lambda^{A-1} \exp\left(-\frac{\lambda}{B}\right)$$

: kernel of ?

Thus, $\lambda|\boldsymbol{x} \sim$?

Posterior distribution with conjugacy

- Posterior: $f(\lambda|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\lambda)\, f(\lambda)}{\int f(\boldsymbol{x}|\lambda) f(\lambda) d\lambda} \propto f(\boldsymbol{x}|\lambda) f(\lambda)$

$$f(\lambda|\boldsymbol{x}) = \frac{\lambda^{\sum x_i}\, \mathrm{e}^{-n\lambda}}{\prod x_i!} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \lambda^{\alpha-1} \exp\left(\frac{-\lambda}{\beta}\right)$$

$$\propto \lambda^{\sum x_i + \alpha - 1} \exp\left(-n\lambda - \lambda\frac{1}{\beta}\right)$$

$$= \lambda^{\sum x_i + \alpha - 1} \exp(-\lambda(n + 1/\beta)) \quad : \text{kernel of Gamma}\left(\sum x_i + \alpha,\ \frac{1}{n + \frac{1}{\beta}}\right)$$

Thus, $\lambda|\boldsymbol{x} \sim \text{Gamma}\left(\sum x_i + \alpha,\ \frac{1}{n + \frac{1}{\beta}}\right)$

Deriving Posterior with all constants – not needed in pracitice (Aside)

$$f(\boldsymbol{x}|\lambda)f(\lambda) = \frac{\lambda^{\sum x_i}\, e^{-n\lambda}}{\prod x_i!}\, \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \lambda^{\alpha-1} \exp\left(-\frac{\lambda}{\beta}\right)$$

$$f(\boldsymbol{x}) = \int_0^{\infty} f(\boldsymbol{x}|\lambda)f(\lambda)d\lambda = \frac{1}{\prod x_i!\, \Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} \lambda^{\sum x_i+\alpha-1} e^{-\lambda(n+1/\beta)}\, d\lambda$$

$$= \frac{1}{\prod x_i! \Gamma(\alpha)\beta^{\alpha}}\, \Gamma(\sum x_i+\alpha) \left[\frac{1}{n+\frac{1}{\beta}}\right]^{\sum x_i+\alpha}$$

$$f(\lambda|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\lambda)\, f(\lambda)}{\int f(\boldsymbol{x}|\lambda)f(\lambda)d\lambda} = \frac{\frac{\lambda^{\sum x_i}\, e^{-n\lambda}}{\prod x_i!}\, \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \lambda^{\alpha-1}\exp(-\lambda/\beta)}{\frac{1}{\prod x_i! \Gamma(\alpha)\beta^{\alpha}}\, \Gamma(\sum x_i+\alpha)\left[\frac{1}{n+\frac{1}{\beta}}\right]^{\sum x_i+\alpha}} = \frac{\lambda^{\sum x_i+\alpha-1}\exp\left(-\lambda\,(n+\frac{1}{\beta})\right)}{\Gamma(\sum x_i+\alpha)\left[\frac{1}{n+\frac{1}{\beta}}\right]^{\sum x_i+\alpha}}$$

Inference with Poisson-Gamma model

- Posterior distribution: $\lambda|\boldsymbol{x} \sim$ Gamma($\sum x_i$ +α, $\frac{1}{n+\frac{1}{\beta}}$)

  - Posterior mean: E[$\lambda|\boldsymbol{x}$]=

    : Posterior mean compared with $\bar{X}$ and E[λ]?

  - Posterior variance: Var[$\lambda|\boldsymbol{x}$] =

Inference with Poisson-Gamma model

- Posterior distribution: $\lambda|\boldsymbol{x} \sim \text{Gamma}(\sum x_i + \alpha, \ \frac{1}{n+\frac{1}{\beta}})$

Posterior mean: $E[\lambda|\boldsymbol{x}] = \frac{\sum x_i + \alpha}{n + \frac{1}{\beta}} = \frac{n\frac{\sum x_i}{n} + \frac{\alpha\beta}{\beta}}{n + \frac{1}{\beta}}$: weighted average of $\bar{X}$ & $E[\lambda]$

- As $n \longrightarrow \infty$, $E[\lambda|\boldsymbol{x}] \longrightarrow \bar{X}$ (frequentist estimate).

  As $\frac{1}{\beta} \longrightarrow \infty$, $E[\lambda|\boldsymbol{x}] \longrightarrow \alpha\beta$ (prior mean)

- $\text{Var}[\bar{X}] = \frac{\lambda}{n} \Rightarrow$ precision of $\bar{X} \propto \frac{n}{\hat{\lambda}}$ Also, precision of $\lambda \propto \frac{1}{\beta}$
- Posterior variance: $\text{Var}[\lambda|\boldsymbol{x}] = (\sum x_i + \alpha)\left[\frac{1}{n+\frac{1}{\beta}}\right]^2 = \frac{\sum x_i + \alpha}{\left[n+\frac{1}{\beta}\right]^2}$