Lecture 8A
MTH6102: Bayesian Statistical Methods

Eftychia Solea

Queen Mary University of London

2023

Today's agenda

Today's lecture will

- Review of noninformative priors
- Learn informative priors
- Be able to make a reasonable choice of informative prior, based on external data.
- Learn that the choice of prior affects the posterior.
- See that more data lessens the dependence of the posterior on the prior.

Prior distributions

- The prior distribution plays a defining role in Bayesian analysis.
- There are two types of priors: noninformative and informative.

Noninformative prior

- A noninformative prior represents our ignorance or lack of information about θ before the experiment.
- Non-informative prior: “let the data speak for themselves".
- An obvious candidate for a noninformative prior is to use a flat or uniform prior over some range, $p(\theta) \propto c$.
- It is flat relative to the range of the likelihood. Assumes that every hypothesis is equally probable.
- Flat priors are not invariant under nonlinear one-to-one transformations g

Noninformative Jeffreys prior

- Another example of noninformative prior is Jeffreys prior defined as

$$p_J(\theta) \propto \sqrt{I(\theta)},$$

where $I(\theta)$ is the Fisher information function given by (under some regularity conditions)

$$I(\theta) = -E\Big[\frac{d^2}{d\theta^2}\log p(X|\theta)\Big],$$

and $p(X|\theta)$ is the likelihood function.

. $X \sim \text{binomial}(n, \theta)$, $p_J(\theta) \sim \text{beta}(\frac{1}{2}, \frac{1}{2})$

Noninformative Jeffreys prior

- Jeffreys prior is invariant under nonlinear, smooth, one-to-one transformations g because

$$\underbrace{I(\psi) = I(\theta)\left(\frac{d\theta}{d\psi}\right)^2}, \quad \text{invariance property}$$

where $\psi = g(\theta)$.

Noninformative prior

- **Advantages**
  1. sometimes used as benchmark that will not reflect the bias of the analyst.
  2. appropriate when little is known on the underlying processes.
  3. can be used in situations where scientific objectivity is at a premium, for example, when presenting results to a regulator or in a scientific journal,
- **Disadvantages**
  1. may lead to improper priors
  2. Flat priors are not invariant under nonlinear one-to-one transformations g
  3. the definition of knowing little may depend on different parameterizations (should θ be assumed to be uniform or should perhaps the logarithm of θ be assumed to be uniform?)

Informative prior

- Informative priors include some judgement concerning plausible values of the parameters based on external information.
- Informative priors can be based on pure judgement, a mixture of data and judgement, or external data alone.
- An informative prior distribution is one in which the probability mass is concentrated in some subset of the possible range for the parameters.

Informative prior

- There are many ways to build an informative prior. For example, using summary statistics, published estimates, intervals or standard errors.

- We can match these quantities to the mean, median standard deviation or percentiles of the prior distribution.

Example: Building an informative prior

Exponential/Gamma example

- Let $t_1, \ldots, t_n \sim \text{Exp}(\lambda)$ denote the lifetimes of lightbulbs.
- The gamma distribution is conjugate to the exponential likelihood for λ (failure rate).

- Suppose we have external information from other similar bulbs with observed failure rates $r_1, \ldots, r_K$.
- Let m and u be the mean and variance of $r_1, \ldots, r_K$, respectively.

- **Goal:** Build a prior $\text{gamma}(\alpha, \beta)$ distribution for λ using external information information.

Exponential/Gamma example

- We can use the method of moments to match the mean and the variance of the prior gamma distribution with the corresponding m and u

sample mean

- That is

Gamma mean

Gamma variance

$$m = \frac{\alpha}{\beta}, \quad u = \frac{\alpha}{\beta^2}$$

m, u observed

$m = \frac{a}{\beta} \Rightarrow a = m\beta$

Then,

$u = \frac{m\cancel{\beta}}{\beta\cancel{\beta}}$

$\Rightarrow \beta = \frac{m}{u}$

$\alpha = \frac{m^2}{u}$

- Solve for α and β

$$\beta = \frac{m}{u}, \quad \alpha = \frac{m^2}{u}.$$

- Thus, our prior for λ is gamma$(\frac{m^2}{u}, \frac{m}{u})$.

Informative prior

- **Advantages**
  1. often analytically convenient (esp for conjugate priors).
  2. can take advantage of your informed understanding, beliefs, experience and external data
- **Disadvantages**
  1. not always easy to quantify the state of knowledge

Board example: Building an informative prior

Binomial/beta example

- Suppose we flip the coin n times and observe k heads with q the probability of heads.
- A beta(α, β) distribution is chosen as the prior distribution for q. Why?
- Based on external information and published statistics, the prior mean is 0.4 and the prior standard deviation is 0.2.
- Find the prior distribution corresponding to this belief.

- See also **Question 2, final exam 2020** prior mean $\frac{a}{a+b}$

prior variance $\frac{ab}{(a+b)^2(a+b+1)}$

We want to find $a>0$ and $b>0$ such that

$$\frac{a}{a+b} = \frac{4}{10} = \frac{2}{5} \quad (1)$$

$$\frac{ab}{(a+b)^2(a+b+1)} = \left(\frac{2}{10}\right)^2 = \frac{4}{100} = \frac{1}{25} \quad (2)$$

Solve for a and b.

From (1) $\quad 5a = 2a+2b \Rightarrow 3a = 2b \Rightarrow a = \frac{2b}{3} \quad (3)$

Replace (3) into (2) to find,

$$\frac{\left(\frac{2b}{3}\right)b}{\left(\frac{2b}{3}+b\right)^2\left(\frac{2b}{3}+b+1\right)} = \frac{1}{25}$$

$$\Rightarrow \frac{\frac{2\cancel{b^2}}{3}}{\frac{25\cancel{b^2}}{9}\left(\frac{5b}{3}+1\right)} = \frac{1}{25}$$

$$\Rightarrow \frac{2\cdot\cancel{9}3}{\cancel{3}\cdot\cancel{25}\left(\frac{5b}{3}+1\right)} = \frac{1}{\cancel{25}} \Rightarrow 6 = \frac{5b}{3}+1$$

$$\Rightarrow \boxed{b=3} \quad \boxed{a=2}$$

Weakly informative prior distributions

- Instead of trying to make the prior completely uniformative, an alternative is to convey some information about the plausible range of the parameters, e.g., exclude implausible values.
- Otherwise let the data speak for themselves.
- For models with large numbers of parameters, adding a little prior information may help with numerical stability.

The choice of prior affects the posterior

- In the Bayesian framework, all our inferences about θ are based on the posterior distribution $p(\theta \mid y)$.

$$p(\theta \mid y) \propto p(\theta)\, p(y \mid \theta)$$

Posterior distribution $\propto$ prior distribution $\times$ likelihood

- Including summaries such as point estimates and credible intervals.
- So our inference depends on the prior distribution as well as the data via the likelihood.
- The choice of prior affects the posterior.
- More data, lessens the dependence of the posterior on the prior.

Normal example, known variance

- Observed data $y_1, \dots, y_n \sim N(\mu, \sigma^2)$.
- Prior distribution $\mu \sim N(\mu_0, \sigma_0^2)$.
- The posterior distribution is

$$\mu \mid y \sim N(\mu_1, \sigma_1^2)$$

$$\mu_1 = \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}\right) \bigg/ \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)$$

$$\sigma_1^2 = 1 \bigg/ \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)$$

Normal example: Posterior mean

- The posterior mean μ_1 can be written as a weighted average of the prior mean μ_0 and the sample mean $\bar{y}$

$$\mu_1 = (1 - w)\mu_0 + w\bar{y},$$

where

$$w = \frac{n}{\sigma^2} \Big/ \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) = \frac{\sigma_0^2}{\frac{\sigma^2}{n} + \sigma_0^2}.$$

$w \to 1$ as $n \to \infty$ or $\sigma_0 \to \infty$, so the posterior mean approaches the sample mean.

$n \to \infty \quad \mu_1 \to \bar{y}$

$\sigma_0 \to \infty \quad \mu_1 \to \bar{y}$

Normal example: Likelihood and prior

- When deriving the posterior distribution, we saw that the likelihood

$$p(y \mid \mu)$$

is proportional to a

$$N\left(\bar{y}, \frac{\sigma^2}{n}\right) \text{ pdf for } \mu$$

if considered as a function of μ.

.

- If we compare $\frac{\sigma^2}{n}$ to the prior variance σ_0^2, this helps to understand how the posterior behaves.

$\underline{p(\mu|y)} \propto$ likelihood $\times$ prior

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\underbrace{(y_i-\mu)^2}\right\} \exp\left\{\frac{-1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left(n\mu^2 - 2n\bar{y}\mu + \underbrace{\sum y_i^2}_{\text{constant}}\right)\right\} \mathcal{N}(\mu_0, \sigma_0^2)$$

function of μ

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\underbrace{(n\mu^2 - 2n\bar{y}\mu)}_{\text{function of } \mu}\right\} \cdot \mathcal{N}(\mu_0, \sigma_0^2)$$

(complete the square)

$$= \exp\left\{-\frac{n}{2\sigma^2}(\mu^2 - 2\bar{y}\mu)\right\} \cdot \mathcal{N}(\mu_0, \sigma_0^2)$$

$$= \exp\left\{-\frac{n}{2\sigma^2}(\mu^2 - 2\bar{y}\mu + \bar{y}^2 \underbrace{- \bar{y}^2}_{\text{constant}})\right\} \mathcal{N}(\mu_0, \sigma_0^2)$$

$$\propto \exp\left\{-\frac{n}{2\sigma^2}\left(\underbrace{\mu^2 - 2\bar{y}\mu + \bar{y}^2}_{(\mu-\bar{y})^2}\right)\right\} \mathcal{N}(\mu_0, \sigma_0^2)$$

$$= \exp\left\{-\frac{n}{2\sigma^2}(\mu-\bar{y})^2\right\} \mathcal{N}(\mu_0, \sigma_0^2)$$

$$= \exp\left\{-\frac{1}{2\sigma^2/n}(\mu-\bar{y})^2\right\} \mathcal{N}(\mu_0, \sigma_0^2)$$

Normal likelihood as a function of μ

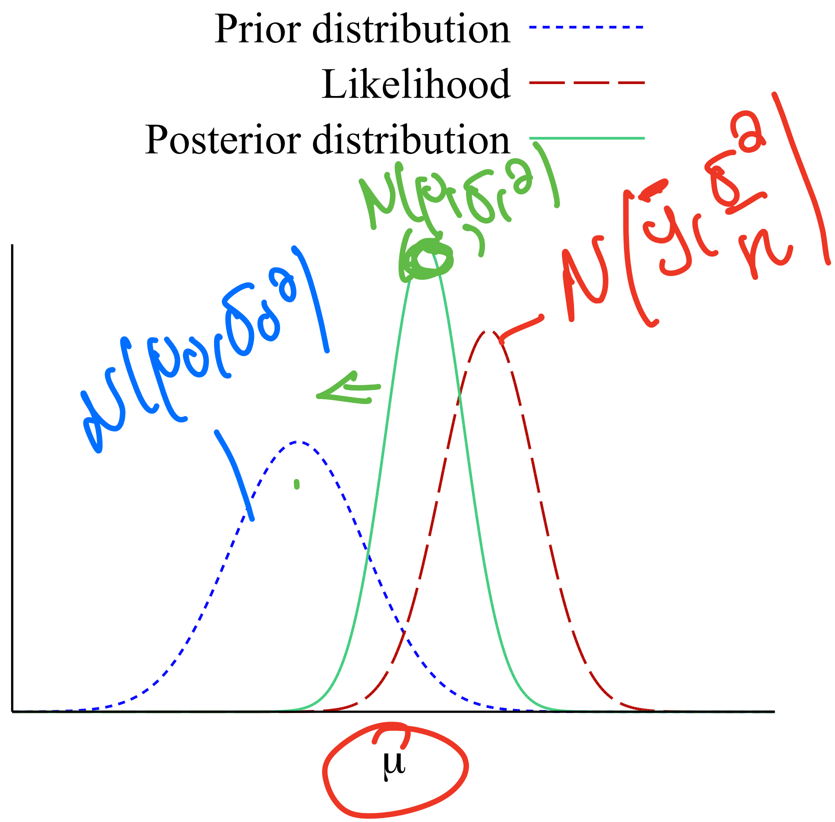
$$\underbrace{\mathcal{N}\left(\bar{y}, \frac{\sigma^2}{n}\right)} \times \underbrace{\mathcal{N}(\mu_0, \sigma_0^2)}$$

Informative and uninformative prior distributions

Normal example with known variance

- An informative prior distribution is strongly peaked around some value.
- Prior changes its value over the range of the likelihood.
- Posterior is shifted relative to likelihood.

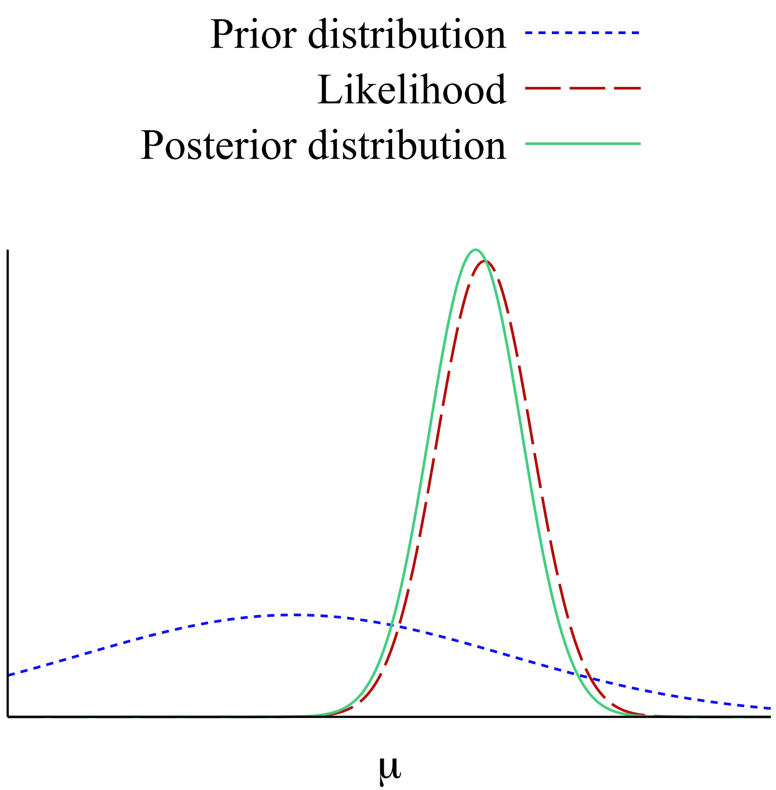
$$\sigma_0^2 = 0.5, \frac{\sigma^2}{n} = 0.25$$



Normal example with known variance

- A weakly or slightly informative prior.
- Only changing gradually over the range of the likelihood.
- When the data provide a lot more information than the prior.
- Posterior is only slightly shifted relative to likelihood.

$$\sigma_0^2 = 5, \frac{\sigma^2}{n} = 0.25$$



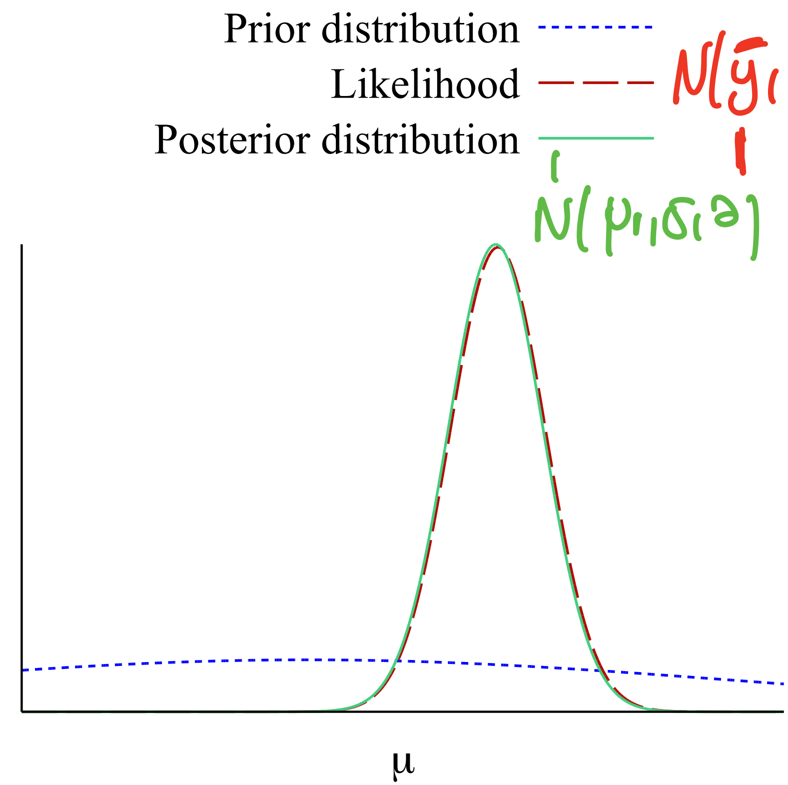
This prior is dominated by the likelihood and they give similar posterior.

Normal example with known variance

- A very weakly informative prior, almost flat prior
- Almost flat over the range of the likelihood
- Posterior practically proportional to likelihood.

$$\sigma_0^2 = 20, \frac{\sigma^2}{n} = 0.25$$

$p(\mu|y) \propto$ likelihood
$\propto N(\bar{y}, \frac{\sigma^2}{n})$



$N(\bar{y}, \frac{\sigma^2}{n})$

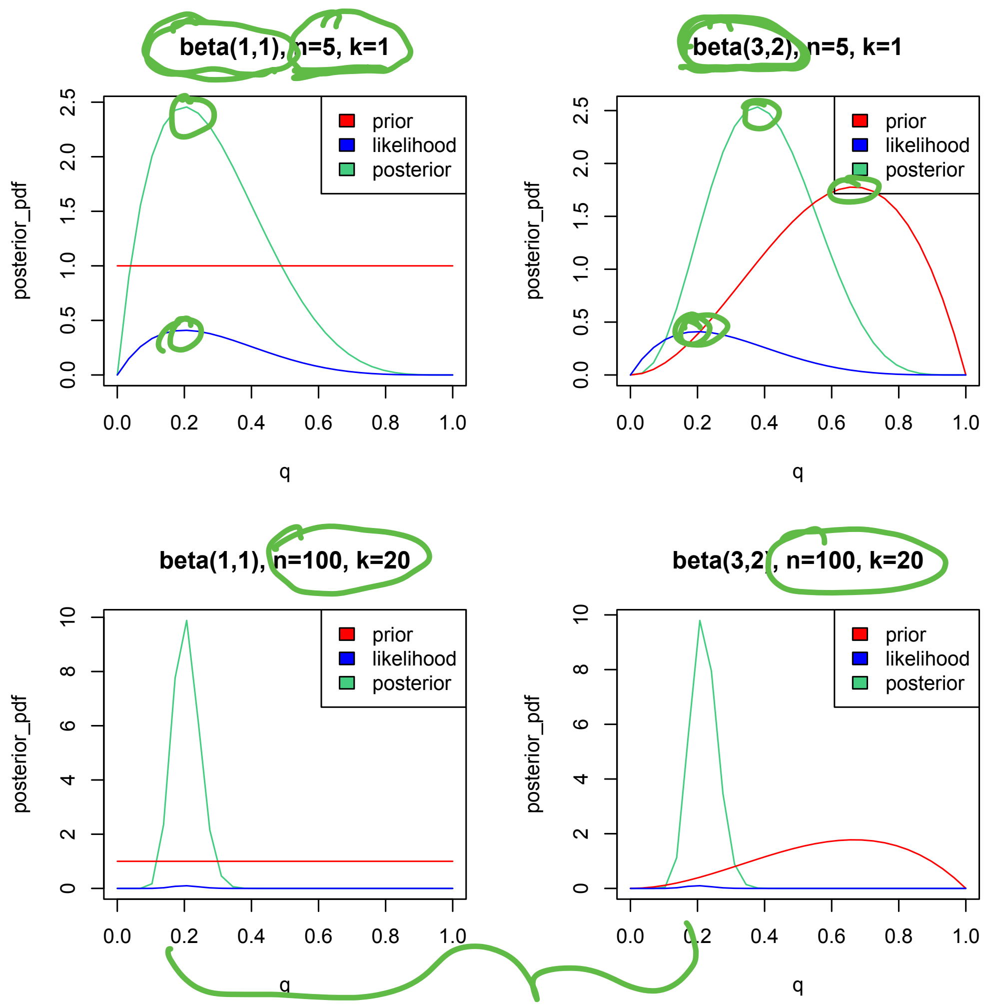
$N(\mu_1, \sigma_1^2)$

Informative and uninformative prior distributions

Beta prior/binomial data example

- Likelihood: $k \sim \text{binom}(n, q)$
- Prior on q: $p(q) \sim \text{beta}(\alpha, \beta)$, $q \in (0, 1)$.
- Posterior, $p(q|k) = \text{beta}(\alpha + k, \beta + n - k)$.
- When $\alpha = \beta = 1$, $q \sim$U$[0, 1]$ or beta$(1, 1)$.

Beta prior/binomial data example



more data lessens the dependence of the posterior on the prior.

Concept question: Normal/normal example

Question 2(a) from final exam Jan 2021

1. We have data $y = (y_1, \dots, y_n)$ from $N(\theta, \sigma^2)$, where $\sigma = 2$.
2. Prior distribution, $p(\theta) \sim N(0, \sigma_0^2)$.
3. **Question:** For an uninformative prior, do we need a large or small value for the prior standard deviation σ_0?

A larger standard deviation corresponds to a less informative prior or uninformative prior.

Concept question: Normal/normal example

Question 2(a) from final exam Jan 2023

1. Same normal/normal example with previous example~~s~~.
2. **Question:** As the prior distribution becomes less informative, what value does the posterior mean for θ approach? As the prior distribution becomes more informative, what value does the posterior mean for θ approach?

A larger value for σ_0 corresponds to a less informative prior. As σ_0 becomes larger, the posterior mean, μ_1, approaches the MLE, $\bar{y}$.
As σ_0 becomes smaller, μ_1 approaches the prior mean