Section 4

Selecting priors

Selecting priors

- Selecting the prior is one of the most important steps in a Bayesian analysis

- There is no “right” way to select a prior

- The choices often depend on the objective of the study and the nature of the data
  1. Conjugate versus non-conjugate
  2. Informative versus uninformative
  3. Proper versus improper
  4. Subjective versus objective

Outline

These notes cover Chapter 2

- Conjugate priors
  - Beta/binomial model for a proportion
  - Poisson/gamma model for a rate
  - Normal/normal model for a mean
  - Normal/inverse-gamma model for a variance
- Prior elicitation
- Improper priors
- Objective Bayesian priors
  - Empirical Bayes
  - Jeffrey's prior
  - Others

Conjugate priors

- A prior is **conjugate** if the posterior is a member of the same parametric family
- We have seen that if the response is binomial and we use a beta prior, the posterior is also a beta
- This requires a pairing of the likelihood and prior
- There is a long list of conjugate priors `https://en.wikipedia.org/wiki/Conjugate_prior`
- The advantage of a conjugate prior is that the posterior is available in closed form
- This is a window into Bayes learning and the prior effect

Conjugate priors

- Here is an example of a non-conjugate prior
- Say $Y \sim \textit{Poisson}(\lambda)$ and $\lambda \sim \text{Beta}(a, b)$
- The posterior is

$$f(\lambda|Y) \propto \exp(-\lambda)\lambda^{Y}\lambda^{a-1}(1-\lambda)^{b-1}$$

- This is not a beta PDF, so the prior is not conjugate
- In fact, this is not a member of any known (to me at least) family of distributions
- For some likelihoods/parameters there is no known conjugate prior

Estimating a proportion using the beta/binomial model

- A fundamental task in statistics is to estimate a proportion using a series of trials:
  - What is the success probability of a new cancer treatment?
  - What proportion of voters support my candidate?
  - What proportion of the population has a rare gene?
- Let $\theta \in [0, 1]$ be the proportion we are trying to estimate (e.g., the success probability).
- We conduct *n* independent trial, each with success probability θ, and observe $Y \in \{0, ..., n\}$ successes.
- We would like obtain the posterior of θ, a 95% interval, and a test that θ equals some predetermined value θ_0.

Frequentist analysis

- The maximum likelihood estimate is the sample proportion

$$\hat{\theta} = Y/n$$

- For large Y and $n - Y$, the sampling distribution of $\hat{\theta}$ is approximately

$$\hat{\theta} \sim \text{Normal}\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$

- The standard error (standard deviation of the sampling distribution) is approximated as

$$\text{SE}(\hat{\theta}) \approx \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

- A 95% CI is then

$$\hat{\theta} \pm 2\text{SE}(\hat{\theta})$$

Bayesian analysis - Likelihood

- Since Y is the number of successes in n independent trials, each with success probability θ, its distribution is

$$Y|\theta \sim Binomial(n, \theta)$$

- PMF: $P(Y = y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$

- Mean: $\mathrm{E}(Y|\theta) = n\theta$

- Variance: $\mathrm{V}(Y|\theta) = n\theta(1-\theta)$

Bayesian analysis - Prior

- The parameter θ is continuous and between 0 and 1, therefore a natural prior is

$$\theta \sim \text{Beta}(a, b)$$

- PDF: $f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$

- Mean: $\text{E}(\theta) = \frac{a}{a+b}$

- Variance: $\text{V}(\theta) = \frac{ab}{(a+b)^2(a+b+1)}$

Derivation of the posterior

- The posterior is $\theta|Y \sim \text{Beta}(a + Y, b + n - Y)$

- See “Beta-binomial” in the online derivations

Shrinkage

- The posterior mean is

$$\hat{\theta}_B = \mathrm{E}(\theta|Y) = \frac{y+a}{n+a+b}$$

- The posterior mean is between the sample proportion Y/n and the prior mean $a/(a+b)$:

$$\hat{\theta}_B = w\frac{Y}{n} + (1-w)\frac{a}{a+b}$$

  where the weight on the sample proportion is $w = \frac{n}{n+a+b}$
- When (in terms of n, a and b) is the $\hat{\theta}_B$ close to Y/n?

- When is the $\hat{\theta}_B$ shrunk towards the prior mean $a/(a+b)$?

Selecting the prior

- The posterior is $\theta|Y \sim \text{Beta}(a + Y, b + n - Y)$

- Therefore, a and b can be interpreted as the “prior number of success and failures”

- This is useful for specifying the prior

- What prior to select if we have no information about θ before collecting data?

- What prior to select if historical data/expert opinion indicates that θ is likely between 0.6 and 0.8?

Related problem

- The success probability of independent trials is θ
- Y is the number of successes before we observe n failures
- Then $Y|\theta \sim \text{NegativeBinomial}(n, \theta)$ and

$$\text{Prob}(Y = y|\theta) = \binom{y+n+1}{y}\theta^y(1-\theta)^n$$

- Assume the prior $\theta \sim \text{Beta}(a, b)$ and find the posterior

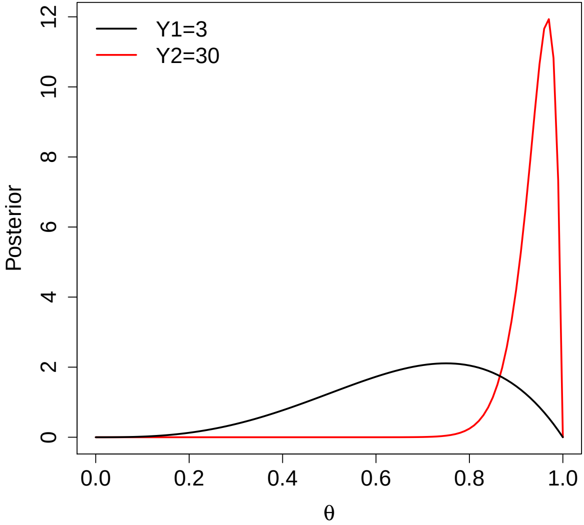
Smoking example

- Two smokers have just quit
- Say subject *i* has probability θ_i of abstaining each day
- The number of days until relapse for two patients is 3 and 30 days
- Can we conclude the patients have different probabilities of relapse?
- What is probability that their next attempts will exceed 30 days?

Smoking example

- The likelihood is $Y_i \sim \text{NegativeBinomial}(1, \theta_i)$
- Assume uniform priors $\theta_i \sim \text{Beta}(1, 1)$
- The posteriors are $\theta_i | Y_i \sim \text{Beta}(Y_i + 1, 2)$
- The posterior are plotted on the next slide
- The following slide uses Monte Carlo sampling to address the two motivating questions

Smoking example



Smoking example

```
> S      <- 1000000
> theta1 <- rbeta(S,3+1,2)
> theta2 <- rbeta(S,30+1,2)
> mean(theta2>theta1)
[1] 0.957222
>
> samp1 <- rnbinom(S,1,prob=1-theta1)
> samp2 <- rnbinom(S,1,prob=1-theta2)
> quantile(samp1,c(0.05,0.5,0.95))
5% 50% 95%
0   1  15
> quantile(samp2,c(0.05,0.5,0.95))
5% 50% 95%
0  13 109
> mean(samp1>30); mean(samp2>30)
[1] 0.015781
[1] 0.254129
```

Estimating a rate using the Poisson/gamma model

- Estimating a rate has many applications:
  - Number of virus attacks per day on a computer network
  - Number of Ebola cases per day
  - Number of diseased trees per square mile in a forest
- Let $\lambda > 0$ be the rate we are trying to estimate
- We make observations over a period (or region) of length (or area) N and observe $Y \in \{0, 1, 2, ...\}$ events
- The expected number of events is $N\lambda$ so that λ is the expected number of events per time unit
- MLE: $\hat{\lambda} = Y/N$ is the sample rate
- We would like obtain the posterior of λ

Bayesian analysis - Likelihood

- Since Y is a count with mean $N\lambda$, a natural model is

$$Y|\lambda \sim \mathit{Poisson}(N\lambda)$$

- PMF: $P(Y = y|\lambda) = \frac{\exp(-N\lambda)(N\lambda)^y}{y!}$

- Mean: $\mathrm{E}(Y|\lambda) = N\lambda$

- Variance: $\mathrm{V}(Y|\lambda) = N\lambda$

Bayesian analysis - Prior

- The parameter λ is continuous and positive, therefore a natural prior is
$$\lambda \sim \text{Gamma}(a, b)$$
- PDF: $f(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda)$
- Mean: $\text{E}(\lambda) = \frac{a}{b}$
- Variance: $\text{V}(\lambda) = \frac{a}{b^2}$

Derivation of the posterior

- The posterior is $\lambda | Y \sim \text{Gamma}(a + Y, b + N)$

- See “Poisson-gamma” in the online derivations

Shrinkage

- The posterior mean is

$$\hat{\lambda}_B = \mathrm{E}(\lambda | Y) = \frac{Y + a}{N + b}$$

- The posterior mean is between the sample rate Y/n and the prior mean a/b:

$$\hat{\theta}_B = w\frac{Y}{n} + (1 - w)\frac{a}{b}$$

  where the weight on the sample rate is $w = \frac{n}{n+b}$
- When (in terms of N, a and b) is the $\hat{\lambda}_B$ close to Y/n?
- When is the $\hat{\lambda}_B$ shrunk towards the prior mean a/b?

Selecting the prior

- The posterior is $\lambda|Y \sim \text{Gamma}(a + Y, b + N)$

- Therefore, *a* and *b* can be interpreted as the “prior number of events and observation time”

- This is useful for specifying the prior

- What prior to select if we have no information about θ before collecting data?

- What prior to select if historical data/expert opinion indicates that λ is likely between 0.6 and 0.8?

Posterior with two observations

- Derive the posterior if $Y_1 \sim \text{Poisson}(N_1\lambda)$; $Y_2 \sim \text{Poisson}(N_2\lambda)$; and $\lambda \sim \text{Gamma}(a, b)$

- See “Poisson-gamma” in the online derivations

Posterior with *m* observations

- Derive the posterior if $Y_i, ..., Y_m \sim$ Poisson($N\lambda$) and $\lambda \sim$ Gamma(a, b)

- See “Poisson-gamma” in the online derivations

Concussions example

- Consider the NFL concussion data at `http://www.pbs.org/wgbh/pages/frontline/concussion-watch/`

- Project 1: Estimate and compare the number of concussions per game (16 games per year) for each NFL team.

- Project 2: Combining teams, estimate the number of concussions per game for 2012 and 2013, separately, to determine if there is a change over time.

- Analysis is on the course webpage

Gaussian models

- The final distribution we'll discuss is the Gaussian (normal) distribution, $Y \sim \text{Normal}(\mu, \sigma^2)$
  - Domain: $Y \in (-\infty, \infty)$
  - PDF: $f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$
  - Mean: $\text{E}(Y) = \mu$
  - Variance: $\text{V}(Y) = \sigma^2$
- In this section, we will discuss:
  - Estimating the mean assuming the variance is known.
  - Estimating the variance assuming the mean is known.

Estimating a normal mean - Likelihood

- We assume the data consist of n independent and identically distributed observations $Y_1, ..., Y_n$.

- Each is Gaussian,

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

  where σ is known

- The likelihood is then

$$\prod_{i=1}^{n} f(y_i|\mu) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right]$$

Bayesian analysis - Prior

- The parameter μ is continuous over the entire real line, therefore a natural prior is

$$\mu \sim \text{Normal}(\theta, \tau^2)$$

- The prior mean θ is the best guess before we observe data

- The math is slightly more interpretable if we set $\tau^2 = \frac{\sigma^2}{m}$

- As we'll see, the prior variance via $m > 0$ controls the strength of the prior

Derivation of the posterior

- Then the posterior is ($w = n/(n+m)$)

$$\mu | Y_1, ..., Y_n \sim \text{Normal}\left(w\bar{Y} + (1-w)\theta, \frac{\sigma^2}{n+m} \right)$$

- See “normal-normal” in the online derivations

Shrinkage

- The posterior mean is

  $$\hat{\mu}_B = \mathrm{E}(\mu | Y_1, ..., Y_n) = w\bar{Y} + (1 - w)\theta$$

  where $w = n/(n + m)$

- Therefore, if m is small then $\hat{\mu}_B \approx \bar{Y}$, and if m is large $\hat{\mu}_B \approx \theta$

- If no prior information is available, take m to be small and thus the prior is uninformative

- Small m gives large prior variance (relative to σ)

Shrinkage

- The posterior variance is

$$\mathrm{V}(\mu | Y_1, ..., Y_n) = \frac{\sigma^2}{n + m}$$

- The sampling variance of $\bar{Y}$ is $\frac{\sigma^2}{n}$

- Therefore, we can loosely interpret *m* as the “prior number of observations”

Estimating a normal variance - Likelihood

- We assume the data consist of n independent and identically distributed observations $Y_1, ..., Y_n$.

- Each is Gaussian,

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

  where μ is known

- The likelihood is then

$$\prod_{i=1}^{n} f(y_i|\mu) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right]$$

Bayesian analysis - Prior

- The parameter σ^2 is continuous over $(0, \infty)$, therefore a natural prior is $\sigma^2 \sim \text{Gamma}(a, b)$

- However, the math is easier if we pick a gamma prior for the inverse variance (precision) $1/\sigma^2$

- If $1/\sigma^2 \sim \text{Gamma}(a, b)$ then $\sigma^2 \sim \text{InverseGamma}(a, b)$

- This is the definition of the inverse gamma distribution

- The inverse gamma prior for σ^2 is PDF

$$f(\sigma^2) = \frac{(\sigma^2)^{-a-1} \exp(-b/\sigma^2)}{b^a \Gamma(a)}$$

Derivation of the posterior

- The posterior is

$$\sigma^2 | Y_1, ..., Y_n \sim \text{InverseGamma}\left(n/2 + a, SSE/2 + b\right)$$

where $SSE = \sum_{i=1}^{n}(Y_i - \mu)^2$

- See “normal-inverse-gamma” in the online derivations

Shrinkage

- The mean of an InverseGamma(a, b) distribution only exists if $a > 1$

- The prior mean (if it exists) is $b/(a-1)$

- The posterior mean is

$$\frac{SSE + b}{n + 2a - 2}$$

- It is common to take a and b to be small to give an uninformative prior

- Then the posterior mean approximates the sample variance $SSE/(n-1)$

Conjugate prior for a normal precision

- The precision is the inverse variance, $\tau = 1/\sigma^2$

- If Y_i have mean μ and precision τ, the likelihood is proportional to

$$\prod_{i=1}^{n} f(y_i|\mu) \propto \tau^{n/2} \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right]$$

- If $\tau \sim \text{Gamma}(a, b)$, then

$$\tau|Y \sim \text{Gamma}(n/2 + a, SSE/2 + b)$$

- This is the exact same analysis as the inverse gamma prior for the variance

Informative versus uninformative priors

- In some cases informative priors are available
- Potential sources include: literature reviews; pilot studies; expert opinions; etc
- **Prior eliciation** is the process of converting expert information to prior distribution
- For example, the expert might not comprehend an inverse gamma PDF, but if they give you an estimate and a spread you can back out *a* and *b*.
- There are tools for this, such as `https://jeremy-oakley.shinyapps.io/SHELF-single/`

Informative versus uninformative priors

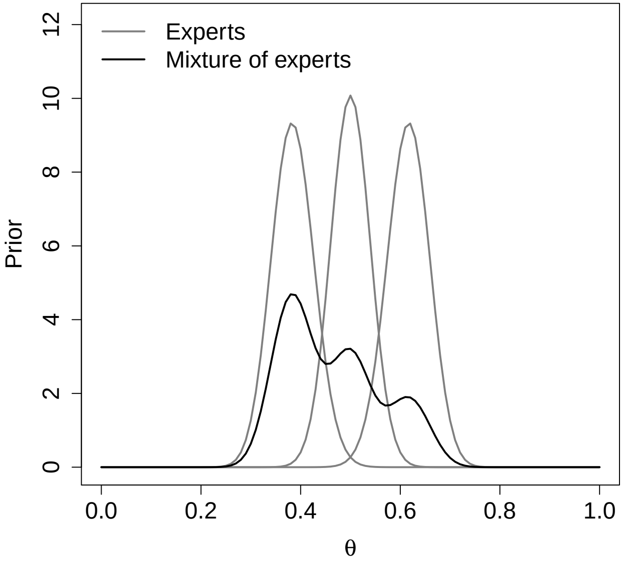
- Strong priors for the parameters of interest can be hard to defend
- Strong priors for nuisance parameters are more common
- For example, say you are doing a Bayesian t-test to study the mean μ, you might use an informative prior for the nuisance parameter σ^2
- Any time informative priors are used you should conduct a **sensitivity analysis**
- That is, compare the posterior for several priors
- Example: `https://www.ncbi.nlm.nih.gov/pubmed/19010642`

Prior elicitation

- Prior information can come from a variety of sources
- Say that source j (e.g., journal article, expert, pilot study) suggests prior $\pi_j(\theta)$
- A **mixture of experts** prior combines the J sources into a single prior
- Say source j is given weight $w_j > 0$ with $\sum_{j=1}^{J} w_j = 1$
- The mixture of experts prior is

$$\pi(\theta) = \sum_{j=1}^{J} w_j \pi_j(\theta)$$

Mixture of experts prior ($w_j \in \{0.2, 0.3, 0.5\}$)



Prior elicitation

- Another powerful idea is to elicit prior information using real-life scenarios
- “Do you expect more events on hot days or cold days?”
- “Would you expect more events on a rainy day August or a sunny day in December?”
- Later you can frame these questions using the parameters in your model and back out the prior that best matches the expert responses

Informative versus uninformative priors

- In most cases prior information is not available and so uninformative priors are used
- Other names: “vague”, “weak”, “flat”, “diffuse”, etc.
- These all refer to priors with large variance
- Examples: $\theta \sim \text{Uniform}(0, 1)$ or $\mu \sim \text{Normal}(0, 1000^2)$
- Uninformative priors can be conjugate or not conjugate
- The idea is that the likelihood overwhelms the prior
- You should verify this with a sensitivity analysis

Improper priors

- Extreme case: $\mu \sim$ Normal$(0, \tau^2)$ and we set $\tau = \infty$
- A “prior” that doesn’t integrate to one is called an **improper**
- Example: $\pi(\mu) = 1$ for all $\mu \in \mathcal{R}$
- It’s OK to use an improper prior so long as you verify that the posterior integrates to one
- For example, in linear regression an improper prior can be used for the slopes as long as the number of observations exceeds the number of covariates and there are no redundant predictors
- Subjective Bayesian interpretation is tricky

Subjective versus objective priors

- A subjective Bayesian picks a prior that corresponds to their current state of knowledge before collecting data
- Of course, if the reader does not share this prior then they might not accept the analysis, and so uninformative priors and a sensitivity analysis are common
- An **objective analysis** is one that requires no subjective decisions by the analyst
- Subjective decisions include picking the likelihood, treatment of outliers, transformations, ... and prior specification
- A completely objective analysis may be feasible in tightly controlled experiments, but is impossible in many analyses

Objective Bayes

- An objective Bayesian attempts to replace the subjective choice of prior with an algorithm that determines the prior
- There are many approaches: Jeffreys, reference, probability matching, maximum entropy, empirical Bayes, penalized complexity, etc.
- Jeffreys priors are the most common and we'll study these in some detail
- The others we will mention superficially
- Many of these priors are **improper** and so you have to check that the posterior is proper

Objective Bayes

- One subjective decision is the parameterization to use
- For example, should we select $\sigma \sim \text{Uniform}(0, 10)$ or $\sigma^2 \sim \text{Uniform}(0, 100)$?
- These are quite different; $\sigma \sim \text{Uniform} \rightarrow p(\sigma^2) \propto 1/\sigma$
- Any objective prior must give the same results, e.g., posterior mean for σ, for any parameterization
- The Jeffreys prior is invariant to transformations

Jeffreys prior

- The Jeffreys prior for parameter θ is $p(\theta) = \sqrt{I(\theta)}$

- The Fisher information is $I(\theta) = -\mathrm{E}_{Y|\theta}\left[\frac{d^2}{d\theta^2}\log p(Y|\theta)\right]$

- Once you have specified the likelihood the Jeffreys prior is determined with no additional input

- Therefore you do not have to make a subjective decision about the prior (other than “subjective” decision to use a Jeffreys prior)

Examples of Jeffreys priors

Likelihood	Jeffreys prior
$Y \sim \text{Binomial}(n, \theta)$	$\theta \sim \text{Beta}(1/2, 1/2)$
$Y \sim \text{N}(\mu, 1)$	$p(\mu) \propto 1$
$Y \sim \text{N}(0, \sigma^2)$	$p(\sigma) \propto 1/\sigma$

- See “Jeffreys” in the online derivations

Reference priors

- These priors try to formally be “uninformative”
- As with Jeffreys they are objective
- They are defined as the prior that gives the maximum expected “distance” between the prior and posterior
- For univariate models they give Jeffreys priors
- For multivariate models they are different
- Reference priors are harder to compute than Jeffreys

Probability matching priors (PMP)

- The PMP is designed so that posterior credible intervals have correct frequentist coverage
- There are only a few cases where this can be done exactly
- For example, if $Y_i|\mu \sim \text{Normal}(\mu, 1)$, the PMP is $p(\mu) = 1$
- The posterior is $\mu|\mathbf{Y} \sim \text{Normal}(\bar{Y}, 1/n)$
- In this case, credible sets have correct frequentist coverage
- Approximations are available for medium-complexity cases

Empirical Bayes

- Empirical Bayes is also objective
- Here you pick the priors based on the data
- Example: maybe σ^2 has prior mean s^2
- Example: σ^2 is fixed at s^2 in the Bayesian analysis of μ
- More formally, you can use marginal maximum likelihood to fix nuisance parameters
- The analysis then proceeds as a usual Bayesian analysis
- Often criticized for “using the data twice”

Penalized complexity priors

- A PCP prior begins with a simple base model, e.g., linear regression with all the slopes equal zero
- The full model, e.g., regression with non-zero slopes, is shrunk towards to the base model
- The “distance” of the full model to the base model has exponential prior to penalized the more complex model from deviating from the base model
- Requires picking the parameter in the exponential prior and setting priors for the parameters in the base model
- So technically this is not purely objective

Maximum entropy priors

- In a maximum entropy prior you fix a few quantities of the prior distribution, e.g., $\mathrm{E}(\theta) = 0.5$

- The prior is taken to be the distribution with the maximum entropy of these that satisfy the known constraints

- “Entropy” is a measure of uncertainty, e.g., the entropy of the PMF $f(x)$ is
$$-\sum_{x \in \mathcal{S}} f(x) \log[f(x)]$$

- If θ has support $\mathcal{R}$ and the mean and variance are known, the maximum entropy prior is Gaussian

- Not purely objective because you have to set the constraints