Advanced Bayesian Methods: Theory and Applications in R

Principles of Bayesian Inference

Nikolaus Umlauf
https://nikum.org/abm.html

Principles of Bayesian Inference

Aims of sections 1-4:

- Introduce the foundations of Bayesian inference and compare it to frequentist maximum likelihood.
- Motivate how Markov chain Monte Carlo (MCMC) simulations provide numerical access to the posterior distributions.
- Discuss practical aspects of working with MCMC simulations.

Bayes' Theorem

- Two central components of a Bayesian model formulation.
  - Observation model $p(\mathbf{y}|\theta)$, which describes how the data $\mathbf{y}$ are generated for given model parameters θ.
  - Prior distribution $p(\theta)$ representing prior beliefs about the parameter vector θ
- Bayesian learning updates prior beliefs on θ based on information in the data $\mathbf{y}$ using Bayes' theorem

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int p(\mathbf{y}|\theta)p(\theta)d\theta},$$

where $p(\mathbf{y})$ is the marginal density of the data.

Example

- Data $y \overset{i.i.d.}{\sim} Be(\pi)$ with unknown success probability $\pi \in (0, 1)$.
- We consider $n = 10$ trials with one success (1) and nine failures (0).
- The likelihood is the product of each the probabilities of each individual Bernoulli trial

$$\mathcal{L}(\pi|\mathbf{y}) = \prod_{i=1}^{n} p(\pi|y_i) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i} = \pi^{\sum_{i=1}^{n} y_i}(1-\pi)^{n-\sum_{i=1}^{n} y_i}.$$

- For the prior distribution of π, we use a Beta distribution with parameters $a > 0$ and $b > 0$:

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

Example

- The density function of the Beta distribution is:

$$p(\pi|a,b) = \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)}, \text{ where } B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

  Here, $\Gamma(\cdot)$ is the Gamma function.

- The posterior distribution combines the likelihood function and the prior

$$\begin{aligned} p(\pi|\mathbf{y}) &\propto \mathcal{L}(\pi|\mathbf{y}) \cdot p(\pi|a,b) \\ &= \left(\prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}\right) \cdot \pi^{a-1}(1-\pi)^{b-1} \\ &= \pi^{\sum_{i=1}^{n} y_i + a - 1}(1-\pi)^{n - \sum_{i=1}^{n} y_i + b - 1} \end{aligned}$$

Example

- Hence, this is the kernel of a Beta distribution. Therefore, the posterior distribution for π is:

$$\pi|\mathbf{y} \sim \text{Beta}\left(a + \sum_{i=1}^{n} y_i, b + n - \sum_{i=1}^{n} y_i\right).$$

- We can express the parameters of the posterior Beta distribution as:

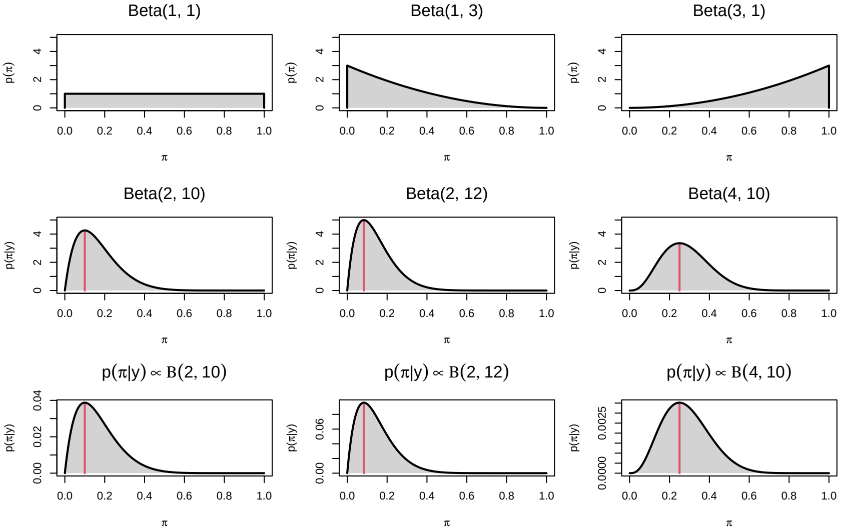
$$\tilde{a} = a + \sum_{i=1}^{n} y_i, \quad \tilde{b} = b + n - \sum_{i=1}^{n} y_i.$$

Example

In R:

```
R> y <- c(1, rep(0, 9))
R> prior <- function(p, a, b, ...) {
+    p^(a - 1) * (1 - p)^(b - 1) / beta(a, b)
+  }
R> likelihood <- function(p, ...) {
+    p^(sum(y)) * (1 - p)^(length(y) - sum(y))
+  }
R> posterior <- function(p, a, b, nc = TRUE, ...) {
+    if(nc) {
+      a <- a + sum(y)
+      b <- b + length(y) - sum(y)
+      pv <- p^(a - 1) * (1 - p)^(b - 1) / beta(a, b)
+    } else {
+      pv <- likelihood(p) * prior(p, a, b)
+    }
+    return(pv)
+  }
```

Example



Beta(1, 1)
Beta(1, 3)
Beta(3, 1)
Beta(2, 10)
Beta(2, 12)
Beta(4, 10)
p(π|y) ∝ B(2, 10)
p(π|y) ∝ B(2, 12)
p(π|y) ∝ B(4, 10)
p(π)
p(π|y)
π

Relation to Maximum Likelihood Estimation

- If the prior distribution is flat, i.e.

$$p(\boldsymbol{\theta}) \propto \text{const},$$

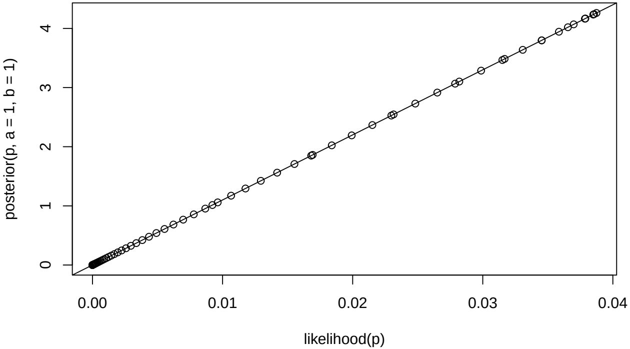
  the posterior is proportional to the likelihood:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \propto p(\mathbf{y}|\boldsymbol{\theta}).$$

- Hence, the mode of the posterior coincides with the maximum likelihood estimate.

Relation to Maximum Likelihood Estimation

```
R> p <- seq(0, 1, length = 100)
R> par(mar = c(4, 4, 1, 1))
R> plot(likelihood(p), posterior(p, a = 1, b = 1))
R> abline(lm(posterior(p, a = 1, b = 1) ~ likelihood(p)))
```



Relation to Maximum Likelihood Estimation

In general,

- the likelihood is a central part of Bayes' theorem that quantifies the information coming from the data and
- the posterior forms a compromise between data (likelihood) and prior beliefs (prior).

Prior Beliefs and Prior Elicitation

- Main conceptual difference between likelihood-based and Bayesian inference: Coming up with a sensible prior distribution.
- The prior should reflect your prior beliefs about the parameter of interest.
- Very common practice:
  - Pick a mathematically convenient class of distributions for the prior and
  - only decide on the parameter of this prior distribution.
- For example, one can formulate belief statements such as

$$P(c_1 \leq \theta \leq c_2) = 1 - \alpha,$$

  where c_1 and c_2 are pre-specified constants from which the prior parameters are determined.
- It is also very common to run analyses for a variety of different priors to study prior sensitivity.

Prior Beliefs and Prior Elicitation

Example in R:

```
R> foo <- function(par, level = 0.9) {
+    p <- pbeta(0.7, par[1], par[2]) - pbeta(0.5, par[1], par[2])
+    (p - level)^2
+  }
R> opt <- optim(c(1, 1), fn = foo, method = "L-BFGS-B", lower = 1, upper = 100)
R> a <- opt$par[1]; b <- opt$par[2]
R> print(a)
```

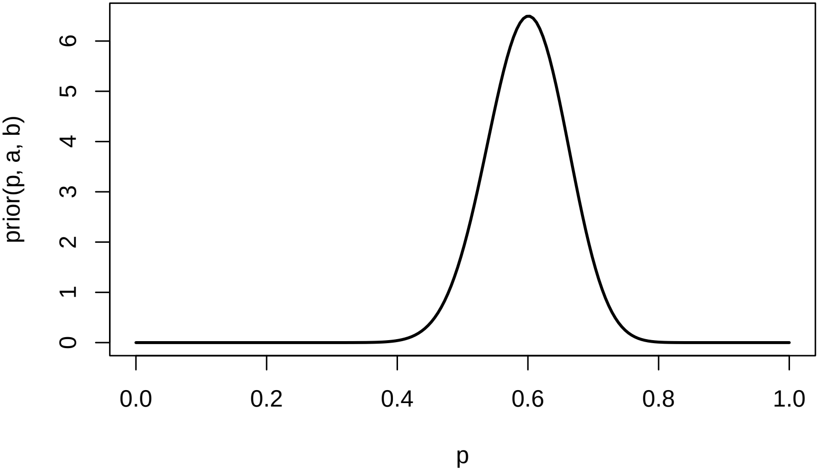
```
[1] 38.34065
```

```
R> print(b)
```

```
[1] 25.81303
```

```
R> p <- seq(0, 1, length = 200)
R> par(mar = c(4, 4, 1, 1))
R> plot(p, prior(p, a, b), type = "l", lwd = 2)
```

Prior Beliefs and Prior Elicitation



Noninformative Prior Specifications

- Flat priors

$$f(\theta) \propto \text{const}$$

  are a popular choice to implement noninformative priors (no value of the parameter is favored a priori).

- Conceptual difficulties:
  - For non-bounded parameter spaces, flat priors are not actual probability distributions.
  - Flat priors are not invariant under transformations of the parameter of interest.

Noninformative Prior Specifications

- An alternative are reference priors for which the prior has the smallest possible influence on the posterior (i.e., it maximizes the Kullback-Leibler discrepancy between the prior and the posterior for given data).

- Another option is Jeffreys' invariant prior:

$$p(\theta) \propto \sqrt{|F(\theta)|}$$

  with expected Fisher information $F(\theta)$.

- For scalar parameters, Jeffreys' prior is equivalent to the reference prior approach.

Bernoulli Experiment: Likelihood and Log-Likelihood

- Consider a Bernoulli experiment where the outcome y is 1 (success) with probability θ and 0 (failure) with probability $1 - \theta$.
- The likelihood function for the parameter θ given the outcome y is:

$$p(y \mid \theta) = \theta^y (1 - \theta)^{1-y}.$$

- The log-likelihood function $\ell(\theta)$ is:

$$\ell(\theta) = \log p(y \mid \theta) = y \log \theta + (1 - y) \log(1 - \theta).$$

Jeffreys' Prior for Bernoulli Experiment

- Compute the first derivative of the log-likelihood function:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{y}{\theta} - \frac{1-y}{1-\theta}.$$

- Compute the second derivative of the log-likelihood function:

$$\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = -\frac{y}{\theta^2} - \frac{1-y}{(1-\theta)^2}.$$

- The Fisher information $F(\theta)$ is the negative expected value of the second derivative:

$$F(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right].$$

Jeffreys' Prior for Bernoulli Experiment

- Substituting the second derivative:

$$F(\theta) = \mathbb{E}\left[\frac{y}{\theta^2} + \frac{1-y}{(1-\theta)^2}\right].$$

- Since y follows a Bernoulli distribution:

$$E[y] = \theta \text{ and } E[1-y] = 1-\theta,$$

so

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

Jeffreys' Prior for Bernoulli Experiment

- Jeffreys' prior is proportional to the square root of the Fisher information:

$$p(\theta) \propto \sqrt{F(\theta)}.$$

- Therefore:

$$p(\theta) \propto \sqrt{\frac{1}{\theta(1-\theta)}}.$$

- Simplifying:

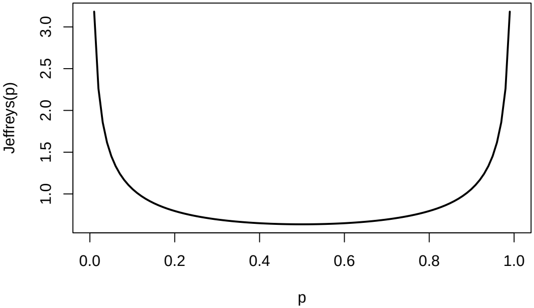
$$p(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}} = \theta^{0.5-1}(1-\theta)^{0.5-1}.$$

- This is equivalent to the Beta$(0.5, 0.5)$ distribution, which is a noninformative prior reflecting minimal prior knowledge about θ.

Jeffreys' Prior for Bernoulli Experiment

Example in R:

```
R> Jeffreys <- function(p) { dbeta(p, 0.5, 0.5) }
R> p <- seq(0, 1, length = 100)
R> par(mar = c(4, 4, 1, 1))
R> plot(p, Jeffreys(p), type = "l", lwd = 2)
```



Priors for the Success Probability

- The beta distribution is conjugate to the Bernoulli observation model, i.e., the posterior is then also a beta distribution with updated parameters.
- Elicit the hyperparameters $a > 0$ and $b > 0$ based on prior statements, e.g., the prior expectation, variance, quantiles, probabilities, etc.
- A flat prior is $\pi \sim \mathrm{U}(0, 1)$, which is also a beta distribution with $a = b = 1$.
- Jeffreys' prior is a beta distribution with $a = b = 0.5$.

Priors for the Success Probability

- A typical discussion on Bayesian inference is that:
  - *Frequentist inference* assumes a true, fixed parameter value, whereas
  - *Bayesian inference* assumes the parameter to be a random variable.
- This is, in general, misleading since the prior is merely used to reflect prior (un)certainty about the parameter of interest.
- The underlying philosophical question is whether this can be done in a sensible way . . .

Posterior Mean and 95% Credible Interval

Model Setup:

- Likelihood: $y \mid \theta \sim \text{Bernoulli}(\theta)$
- Prior: $\theta \sim \text{Beta}(\alpha, \beta)$
- Posterior: $\theta \mid y \sim \text{Beta}(y + \alpha, n - y + \beta)$

Example Parameters:

- Number of trials $n = 10$
- Number of successes $y = 7$
- Prior parameters: $\alpha = 2$, $\beta = 2$

Posterior Mean and 95% Credible Interval

Posterior Distribution:

$$\theta \mid y \sim \text{Beta}(9, 5)$$

Posterior Mean:

$$\text{Mean} = \frac{\alpha'}{\alpha' + \beta'} = \frac{9}{9+5} = \frac{9}{14} \approx 0.643$$

95% Credible Interval:

- Compute quantiles using the Beta distribution:

$$\text{Lower Bound} = \text{Beta}^{-1}(0.025; 9, 5)$$

$$\text{Upper Bound} = \text{Beta}^{-1}(0.975; 9, 5)$$

Posterior Mean and 95% Credible Interval

- Numerical values:

```
R> qbeta(0.025, 9, 5)
[1] 0.3857383

R> qbeta(0.975, 9, 5)
[1] 0.8614207
```

Result: The 95% credible interval for θ is approximately $(0.386, 0.861)$.

Challenges with Non-Conjugate Prior

Challenges:

- **No Closed-Form Solution**: What if the posterior distribution $p(\theta \mid \mathbf{y})$ does not simplify to a standard form?
- **Numerical Approximation Required**: Direct calculation of posterior mean and credible intervals is not feasible.
- **MCMC Methods**: To approximate the posterior mean and credible interval, MCMC methods (e.g., Metropolis-Hastings, Gibbs sampling) must be used.

Summary: Non-conjugate priors may lead to complex posterior distributions that require advanced numerical techniques for estimation.

Frequentist vs. Bayesian Inference

- **Frequentist Inference**:
  - Assumes a **fixed, true parameter** in the population.
  - Estimation through **repeated sampling**.
- **Bayesian Inference**:
  - Treats the parameter as a **random variable**, reflecting **uncertainty**.
  - Combines **prior beliefs** with observed data to update beliefs (posterior distribution).
- **Misconception**: Bayesian inference does not imply the parameter is truly random, but reflects uncertainty.
- **Role of Prior**: Encapsulates prior knowledge or uncertainty, updated with data.
- **Philosophical Debate**: How to sensibly define objective priors?