Bayesian Statistics
Bayesian Test

Shaobo Jin

Department of Mathematics

Hypothesis

Consider a statistical model $f(x \mid \theta)$ with $\theta \in \Theta$. We often want to investigate whether θ belongs to a subset of interest of Θ: $\theta \in \Theta_0$.

- For example, whether $\theta = 0$, or $\theta \leq \theta_0$.

The parameter space Θ is partitioned into two subsets,

- null hypothesis $H_0 : \theta \in \Theta_0$
- alternative hypothesis $H_1 : \theta \in \Theta_1$

such that $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

General Hypothesis

- The hypotheses are often formulated based on the restrictions on the parameter values, e.g., $\theta \in \Theta_0$.
- In general, a hypothesis formulates a class of statistical models.

Example

Suppose that we have randomly chosen n patients and want to analyze their blood samples in order to test drug resistance. Let X be the number of patients with positive test result. Two models are under consideration

$$\begin{aligned} H_0: &\quad X \sim \text{Binomial}(n, p), \ p \sim \text{Beta}(a_0, b_0) \\ H_1: &\quad X \sim \text{Poisson}(\lambda), \ \lambda \sim \text{Gamma}(a_1, b_1). \end{aligned}$$

Statistical Hypothesis Test

Definition

A nonrandomized test ϕ is a statistic from the sample space $\mathcal{X}$ to $\{0,1\}$:

$$\phi(x) = \begin{cases} 1, & \text{if } x \in C_1, \text{ (reject } H_0) \\ 0, & \text{if } x \in C_0, \text{ (do not reject } H_0) \end{cases}$$

where $\mathcal{X} = C_1 \cup C_0$ with $C_1 \cap C_0 = \emptyset$. A randomized test ϕ is a statistic from the sample space $\mathcal{X}$ to $[0,1]$:

$$\phi(x) = \begin{cases} 1, & \text{if } x \in C_1, \text{ (reject } H_0) \\ r, & \text{if } x \in C_=, \text{ (reject } H_0 \text{ with probability } r) \\ 0, & \text{if } x \in C_0, \text{ (do not reject } H_0) \end{cases}$$

where $\mathcal{X} = C_1 \cup C_= \cup C_0$ and $\{C_1, C_=, C_0\}$ are disjoint.

Type I Error and Type II Error

Statistical hypothesis testing is subject to errors.

	Truth	
Decision	H_0	H_1
H_0	Correct decision	Type II error
H_1	Type I error	Correct decision

- In general, a small Type I error probability yields a large Type II error probability, and vice versa.
- The theory of frequentist hypothesis testing is often built on the idea of most powerful test.
  - Among the tests that have low Type I error probability, we want to find the test that has the smallest Type II error.

Neyman-Pearson Test

Definition

Consider testing $H_0 : \mathrm{P}_0$ versus $H_1 : \mathrm{P}_1$. For $k \geq 0$, the randomized Neyman-Pearson test is

$$\phi(x) = \begin{cases} 1, & \text{if } f_0(x) < k f_1(x), \\ r, & \text{if } f_0(x) = k f_1(x), \\ 0, & \text{if } f_0(x) > k f_1(x), \end{cases}$$

where f_0 and f_1 are the density functions related to P_0 and P_1, respectively.

The Neyman-Pearson test is more powerful than any other test ϕ^* of level α, i.e., $\mathrm{E}[\phi^*(X) \mid H_0] \leq \alpha$.

Optimal Bayes Test

We can formulate the nonrandomized test as a decision problem. Suppose that the loss of the wrong decision is

	Truth	
Decision	H_0	H_1
H_0	0	a_1
H_1	a_0	0

such that $a_0 + a_1 > 0$.

Result

The optimal Bayes test that minimizes the posterior expected loss $\mathrm{E}[L \mid x]$ and the expected loss $\mathrm{E}[L]$ is

$$\phi(x) = \begin{cases} 1, & \text{if } \mathrm{P}(H_0 \mid x) < \frac{a_1}{a_0+a_1}, \\ 0, & \text{if } \mathrm{P}(H_0 \mid x) \geq \frac{a_1}{a_0+a_1}. \end{cases}$$

Bayes Test: Example

Example

1. Suppose that $X \mid \theta \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(a, b)$. We are interested in testing
$$H_0 : \theta \geq \frac{1}{2}, \quad \text{versus} \quad H_1 : \theta < \frac{1}{2}.$$
2. Suppose that independent $X_i \mid \theta \sim N\left(\theta, \sigma^2\right)$ for $i = 1, ..., n$, where σ^2 is known. The prior is $\theta \sim N\left(\mu_0, \sigma_0^2\right)$. We are interested in testing
$$H_0 : \theta \leq 0, \quad \text{versus} \quad H_1 : \theta > 0.$$

$0-1$ Loss

In the special case where $a_0 = a_1$, it is the same as the $0-1$ loss.

- Suppose that $\lambda = 0$ means that H_0 is the truth and $\lambda = 1$ that H_1 is the truth.
- The loss function is

$$L = \begin{cases} 1, & \text{if } \phi \neq \lambda, \\ 0, & \text{if } \phi = \lambda. \end{cases}$$

- The optimal Bayes test is equivalent to

$$\phi(x) = \begin{cases} 1, & \text{if } \mathrm{P}(H_0 \mid x) < \mathrm{P}(H_1 \mid x), \\ 0, & \text{if } \mathrm{P}(H_0 \mid x) \geq \mathrm{P}(H_1 \mid x). \end{cases}$$

Bayes Theorem for Simple Hypotheses

Consider two simple hypotheses:

$$H_0: \theta = \theta_0 \quad \text{vs} \quad H_1: \theta = \theta_1.$$

The Bayes theorem yields

$$\mathrm{P}\left(H_k \mid x\right) = \frac{\mathrm{P}\left(H_k\right) f\left(x \mid \theta_k\right)}{\mathrm{P}\left(H_0\right) f\left(x \mid \theta_0\right) + \mathrm{P}\left(H_1\right) f\left(x \mid \theta_1\right)},$$

where $\mathrm{P}\left(H_k\right)$ is the prior probability that hypothesis H_k is true. Hence,

$$\frac{\mathrm{P}\left(H_0 \mid x\right)/\mathrm{P}\left(H_1 \mid x\right)}{\mathrm{P}\left(H_0\right)/\mathrm{P}\left(H_1\right)} = \frac{f\left(x \mid \theta_0\right)}{f\left(x \mid \theta_1\right)}.$$

Bayes Factor

In the case of simple hypotheses, comparing the odds

$$\frac{\mathrm{P}\left(H_0 \mid x\right)/\mathrm{P}\left(H_1 \mid x\right)}{\mathrm{P}\left(H_0\right)/\mathrm{P}\left(H_1\right)}$$

is equivalent to comparing the likelihood values. But we can still apply this ratio to more general cases.

Definition

Consider testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. The Bayes factor is defined to be

$$B_{01}\left(x\right) = \frac{\mathrm{P}\left(H_0 \mid x\right)/\mathrm{P}\left(H_1 \mid x\right)}{\mathrm{P}\left(H_0\right)/\mathrm{P}\left(H_1\right)}.$$

Bayes Factor and Marginal Likelihood

Result

For $k \in \{0,1\}$, let $\pi_k(\theta)$ and $f_k(x \mid \theta)$ be the prior for θ and the likelihood under the hypothesis H_k, respectively. Let $\mathrm{P}(H_k)$ is the prior probability that hypothesis H_k is true. Then,

$$\underbrace{\frac{\mathrm{P}(H_0 \mid x)}{\mathrm{P}(H_1 \mid x)}}_{\text{posterior odds}} = \underbrace{\frac{\int_{\Theta_0} f_0(x \mid \theta)\,\pi_0(\theta)\,d\theta}{\int_{\Theta_1} f_1(x \mid \theta)\,\pi_1(\theta)\,d\theta}}_{\text{Bayes factor}} \underbrace{\frac{\mathrm{P}(H_0)}{\mathrm{P}(H_1)}}_{\text{prior odds}}.$$

Here,

$$m_k(x) = \int_{\Theta_k} f_k(x \mid \theta)\,\pi_k(\theta)\,d\theta$$

is the marginal likelihood under hypothesis H_k. Hence, the Bayes factor is also the ratio of marginal likelihoods.

Bayes Factor vs Likelihood Ratio Test

The Bayes factor is a ratio of marginal likelihoods

$$B_{01}(x) = \frac{\int_{\Theta_0} f_0(x \mid \theta)\, \pi_0(\theta)\, d\theta}{\int_{\Theta_1} f_1(x \mid \theta)\, \pi_1(\theta)\, d\theta}.$$

The likelihood ratio test computes the ratio of maximum likelihoods

$$\lambda(x) = \frac{\sup_{\Theta_0} f(x \mid \theta)}{\sup_{\Theta} f(x \mid \theta)} = \frac{f\left(x \mid \hat{\theta}_0\right)}{f\left(x \mid \hat{\theta}\right)},$$

where $\hat{\theta}_0$ is the MLE with the restriction $\theta \in \Theta_0$ and $\hat{\theta}$ is the MLE without such restriction.

Rule-of-Thumb

A large $B_{01}(x)$ indicates that the marginal likelihood under H_0 is higher than that under H_1. A rule-of-thumb to interpret the value of Bayes factor B_{10} (instead of B_{01}) is as follows.

B_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
>150	Very strong

A side note is that the marginal likelihood $m(x)$ is the omitted normalizing constant when we use $\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$. We have to keep track all constants now!

Compute Bayes Factor: Simple H_0

Example

Compute Bayes factor.

1. Suppose that $X \mid \theta \sim \text{Binomial}\left(n, \theta\right)$ and $\theta \sim \text{Beta}\left(a, b\right)$. We are interested in testing

$$H_0 : \theta = \frac{1}{2}, \quad \text{versus} \quad H_1 : \theta \neq \frac{1}{2}.$$

   The prior under the alternative hypothesis is Uniform $[0, 1]$.

2. Suppose that $X_i, ..., X_n \mid \theta, \sigma^2$ be iid $N\left(\theta, \sigma^2\right)$, where both θ and σ^2 are unknown. We are interested in testing

$$H_0 : \theta = 0, \quad \text{versus} \quad H_1 : \theta \neq 0.$$

   The prior for $\theta \mid \sigma^2$ under the alternative hypothesis is $N\left(0, \sigma^2\right)$. The prior for σ^2 is σ^{-2}.

Compute Bayes Factor: Complicated Example

Example (Two-sample t test)

Suppose that we have two independent samples, $X_i \sim N\left(\mu_1, \sigma^2\right)$, $i = 1, ..., n_1$ and $Y_j \sim N\left(\mu_2, \sigma^2\right)$, $j = 1, ..., n_2$. We want to test whether their expectations are the same. We can reparametrize the distributions as

$$X_i \sim N\left(\mu + 2^{-1}\delta, \sigma^2\right) \qquad Y_j \sim N\left(\mu - 2^{-1}\delta, \sigma^2\right),$$

with parameters $\left(\mu, \delta, \sigma^2\right)$, where $\delta = \mu_1 - \mu_2$, $\mu = \left(\mu_1 + \mu_2\right)/2$. The hypotheses are

$$H_0 : \delta = 0 \quad \text{versus} \quad H_1 : \delta \neq 0.$$

Let $\pi_0\left(\mu, \sigma^2\right) = \sigma^{-2}$ for both hypotheses and $\delta \mid \mu, \sigma^2 \sim N\left(0, \sigma_0^2\sigma^2\right)$ for H_1. Find the Bayes factor.

Bayes Factor and Optimal Bayes Test

Recall that the optimal Bayes test is

$$\phi(x) = \begin{cases} 1, & \text{if } \mathrm{P}(\theta \in \Theta_0 \mid x) < \frac{a_1}{a_0+a_1}, \\ 0, & \text{if } \mathrm{P}(\theta \in \Theta_0 \mid x) \geq \frac{a_1}{a_0+a_1}. \end{cases}$$

From the expression of the Bayes factor, we obtain

$$\mathrm{P}(\theta \in \Theta_0 \mid x) = \frac{B_{01}\mathrm{P}(\theta \in \Theta_0)}{\mathrm{P}(\theta \in \Theta_1) + B_{01}\mathrm{P}(\theta \in \Theta_0)}.$$

Hence, rejecting H_0 by the optimal Bayes test is equivalent to rejecting H_0 if

$$B_{01} < \frac{a_1\mathrm{P}(\theta \in \Theta_1)}{a_0\mathrm{P}(\theta \in \Theta_0)}.$$

Bayes Factor with Improper Prior

As long as the posterior is proper, we can use an improper prior in estimation. However, we need to be careful when using improper prior with Bayes factors.

- Suppose that we have one observation $X \sim N(\theta, 1)$ and consider the Jeffreys prior $\pi(\theta) \propto 1$.
- We want to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$.
- The Bayes factor is

$$
\begin{aligned}
B_{10} &= \frac{\int_{\theta \in \Theta_1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right) \cdot 1 d\theta}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)} \\
&= \sqrt{2\pi} \exp\left(\frac{x^2}{2}\right) \geq \sqrt{2\pi} = 2.5066, \quad \forall x.
\end{aligned}
$$

- The Bayes factor is biased towards favoring H_1.

Jeffreys-Lindley Paradox

For the same example as above, consider the prior $\theta \sim N\left(0, \sigma_0^2\right)$.

- Intuitively speaking, as σ_0^2 increases, the prior becomes less informative.
- As $\sigma_0^2 \to \infty$, we should approximate an uninformative prior.
- The Bayes factor satisfies

$$B_{01} = \sqrt{\sigma_0^2 + 1} \exp\left(\frac{x^2}{2\left(\sigma_0^2+1\right)} - \frac{x^2}{2}\right) \to \infty,$$

  if $\sigma_0^2 \to \infty$, for any fixed x. Hence, it favors H_0 instead.

The Jeffreys-Lindley paradox says that the test based on an improper prior cannot be approximated by tests based on priors with increasing variances.

Training Sample

The intrinsic Bayes factor is a possible way out from the problems associated with improper priors.

Definition

Given an improper prior π, a sample $(x_1, ..., x_n)$ is called a training sample if the corresponding posterior $\pi(\theta \mid x_1, ..., x_n)$ is proper. The sample is a minimal training sample if no subsample is a training sample.

Example

Suppose that we have an iid sample $(x_1, ..., x_n)$ from $N(\mu, \sigma^2)$. Consider the Jeffreys prior.

1. If μ is unknown, but σ^2 is known, then the minimal training sample size is 1.
2. If both μ and σ^2 are unknown, then the minimal training sample size is 2.

Intrinsic Bayes Factor

The idea is to

1. use a training sample to produce a proper posterior from an improper prior,
2. use the resulting posterior as if it were a proper prior for the rest of the sample.

Definition

Let $x_{(l)}$ be a training sample and $x_{-(l)}$ be the rest of the sample. The intrinsic Bayes factor is

$$B_{01}^{I} = \frac{\int_{\Theta_0} f_0\left(x_{-(l)} \mid \theta\right) \pi_0\left(\theta \mid x_{(l)}\right) d\theta}{\int_{\Theta_1} f_1\left(x_{-(l)} \mid \theta\right) \pi_1\left(\theta \mid x_{(l)}\right) d\theta}.$$

Equivalent Representation of Intrinsic Bayes Factor

Result

Suppose that we have an independent sample $(x_1, ..., x_n)$. The intrinsic Bayes factor can be written as

$$B_{01}^{I}(x) = \underbrace{\frac{\int_{\Theta_0} f_0(x \mid \theta)\, \pi_0(\theta)\, d\theta}{\int_{\Theta_1} f_1(x \mid \theta)\, \pi_1(\theta)\, d\theta}}_{B_{01}(x)} \underbrace{\frac{\int_{\Theta_1} f_0(x_l \mid \theta)\, \pi_1(\theta)\, d\theta}{\int_{\Theta_0} f_1(x_l \mid \theta)\, \pi_0(\theta)\, d\theta}}_{B_{10}(x_{(l)})}.$$

The choice of training sample can influence the Bayes factor.

- The training sample should be chosen as small as possible, e.g., minimal training sample.
- Since we split the data into two parts, we avoid using the same data twice.

Credible Set

Definition

A set $C(x)$ is a α-credible set if the posterior distribution satisfies

$$\mathrm{P}(\theta \in C(x) \mid x) \quad \geq \quad 1-\alpha, \quad \alpha \in [0,1].$$

1. It is highest posterior density (HPD) if it can be written as

   $$\{\theta : \pi(\theta \mid x) > k_\alpha\} \subseteq C(x) \subseteq \{\theta : \pi(\theta \mid x) \geq k_\alpha\},$$

   where k_α is the largest bound such that

   $$\mathrm{P}(\theta \in C(x) \mid x) \quad \geq \quad 1-\alpha.$$

2. It is an equal tailed credible interval if the lower and upper bounds satisfy

   $$\mathrm{P}(\theta \leq L(x) \mid x) \quad = \quad \mathrm{P}(\theta \geq U(x) \mid x) = \alpha/2.$$

Find Credible Set

If the posterior is a continuous distribution with density $\pi(\theta \mid x)$, then the HPD credible set is

$$C(x) = \{\theta : \pi(\theta \mid x) \geq k_\alpha\}$$

such that $\mathrm{P}(\theta \in C(x) \mid x) \geq 1 - \alpha$.

Example

Find the credible set.

1. Let X_1, ..., X_n be iid $N(0, \sigma^2)$. We assume that the prior of σ^2 is $\mathrm{InvGamma}(a_0, b_0)$.
2. Let X_1, ..., X_n be iid $N(\theta, 1)$. We assume that the prior of θ is a normal mixture of $N(m_1, \tau_1^2)$ and $N(m_2, \tau_2^2)$.

Some Remarks

To consider the HPD credible sets is motivated by the fact that they minimize the volume among α-credible sets.

- The equal tailed credible interval is easy to work with but may contain regions with low posterior.
- The HPD credible set is not guaranteed to be an interval.

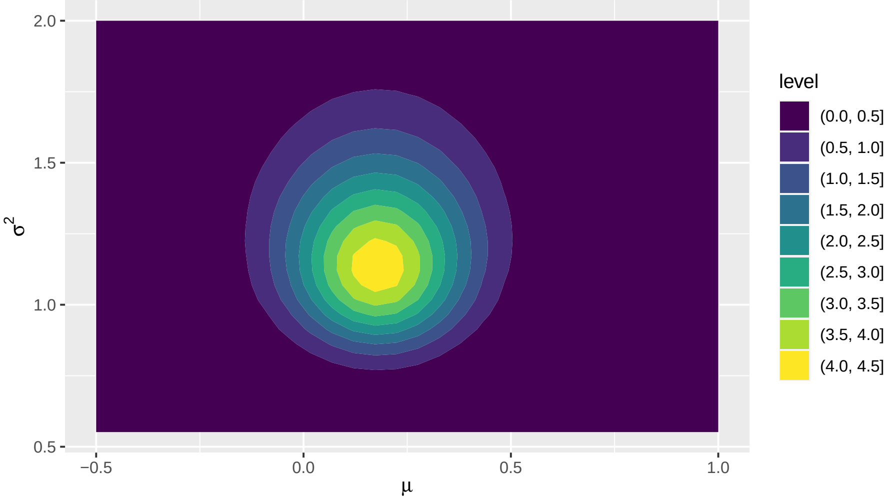
Improper priors can be used to find credible sets, as long as the posterior is proper.

Example

Let X_1, ..., X_n be iid $N\left(\mu, \sigma^2\right)$. We consider the prior $\pi\left(\mu, \sigma^2\right) = \sigma^{-2}$. Find the simultaneous credible set and the marginal credible interval.

Contour Plot of Example

Suppose that we observe $n = 50, \bar{x} = 0.18$, and $\sum_{i=1}^{n} x_i^2 = 60.53$.



Recall: Posterior in Linear model

Consider the linear model

$$Y = X\beta + \epsilon, \qquad \epsilon \mid \sigma^2 \sim N_n\left(0,\ \sigma^2 I_n\right).$$

Under the conjugate prior,

$$\beta \mid \sigma^2 \sim N_p\left(\mu_0, \sigma^2 \Lambda_0^{-1}\right), \qquad \sigma^2 \sim \text{InvGamma}\left(a_0, b_0\right)$$

the posterior is

$$\beta \mid y, \sigma^2, N\left(\mu_n,\ \sigma^2 \Lambda_n^{-1}\right) \qquad \sigma^2 \mid y \sim \text{InvGamma}\left(a_n,\ b_n\right).$$

The marginal posterior of β is

$$\beta \mid y \quad \sim \quad t_{2a_n}\left(\mu_n,\ \frac{b_n}{a_n}\Lambda_n^{-1}\right).$$

Example: Credible set in Linear model

The credible interval for σ^2 can be easily obtained from the marginal posterior

$$\sigma^2 \mid y \quad \sim \quad \text{InvGamma}\left(a_n,\, b_n\right).$$

Lemma

If a $p \times 1$ random vector $X \sim t_v\left(\mu, \Sigma\right)$, then

$$\frac{1}{p}\left(X-\mu\right)^T \Sigma^{-1}\left(X-\mu\right) \quad \sim \quad F\left(p, v\right).$$

The lemma suggests that a credible set for β is

$$\left\{\beta :\ \frac{1}{p}\left(\beta-\mu_n\right)^T \left(\frac{b_n}{a_n}\Lambda_n^{-1}\right)^{-1}\left(\beta-\mu_n\right) \leq F\left(1-\alpha;\, p, v\right)\right\}.$$