

ST308 - Lent term

Bayesian Inference

Kostas Kalogeropoulos

Hypothesis Testing - Prediction - Monte Carlo

Outline

Topics covered: Bayes factors, Lindley's paradox, unit information prior, predictive distribution, Monte Carlo integration

- 1 Hypothesis testing
- 2 Prediction
- 3 Monte Carlo Integration

Outline

- 1 Hypothesis testing
- 2 Prediction
- 3 Monte Carlo Integration

Hypothesis Testing problem

- Consider the **data** $x = (x_1, \dots, x_n)$.
- Assign **model-likelihood** $f(x|\theta)$ with some unknown parameters θ .
- (In Bayesian Inference) Assign a **prior** on θ .
- Consider $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$. Use the information from x to **choose** between them.
- The above can be extended to the case of **more than two** hypotheses.

Bayes factor

Let $\pi(\theta \in \Theta_0)/\pi(\theta \in \Theta_1)$ and $\pi(\theta \in \Theta_0|x)/\pi(\theta \in \Theta_1|x)$ be the prior and posterior odds of H_0 , respectively.

Bayes factor

The **Bayes factor** in favour of H_0 is the ratio of the corresponding posterior to prior odds

$$B_{01}(x) = \frac{\frac{\pi(\theta \in \Theta_0|x)}{\pi(\theta \in \Theta_1|x)}}{\frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}}$$

Note: It is not hard to see that $B_{10}(x) = 1/B_{01}(x)$.

Bayes factor motivation

For the case of **simple vs simple** hypotheses $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, Bayes factor reduces to the likelihood ratio test, i.e. the **most powerful** test for this case.

$$B_{10}(x) = \frac{\frac{\pi(\theta_1|x)}{\pi(\theta_0|x)}}{\frac{\pi(\theta_1)}{\pi(\theta_0)}} = \frac{\frac{\frac{1}{m(x)}f(x|\theta_1)\pi(\theta_1)}{\frac{1}{m(x)}f(x|\theta_0)\pi(\theta_0)}}{\frac{\pi(\theta_1)}{\pi(\theta_0)}} = \frac{f(x|\theta_1)}{f(x|\theta_0)}$$

In more general cases the above does not hold but it is still considered as the **default** criterion for Bayesian hypothesis testing

Bayes factor - interpretation

In terms of interpretation the following guidelines are available

$1 < B_{10}(x) \leq 3$	evidence against H_0 is poor
$3 < B_{10}(x) \leq 20$	evidence against H_0 is substantial
$20 < B_{10}(x) \leq 150$	evidence against H_0 is strong
$B_{10}(x) > 150$	evidence against H_0 is decisive

Example: IQ scores

Recall the IQ test example from last week. The **prior** was the $N(110, 120)$ and the **posterior** $N(102.8, 48)$.

The student claims it was not his day and his genuine IQ is at least 105. So $H_0 : \theta \geq 105$ vs $H_1 : \theta < 105$.

$$\pi(\theta < 105|x) = \pi\left(Z < \frac{105-102.8}{\sqrt{48}}\right) = \Phi(.3175) = .6246$$

$$\pi(\theta < 105) = \pi\left(Z < \frac{105-110}{\sqrt{80}}\right) = \Phi(-.559) = .4046$$

So the **Bayes factor** against H_0 is **4.11**. Strong evidence against H_0 (student's claim).

General case

Suppose we want to test $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$.

Note that for -say- $\theta \in \mathbb{R}$ or $\theta \in [-1, 1]$, $\pi(\theta = 0) = \pi(\theta = 0|x) = 0$.
Hence the Bayes factor is **indeterminate** in such cases.

A more general expression uses the **model evidence / marginal likelihood**:

$$B_{10}(x) = \frac{\frac{\pi(H_1|x)}{\pi(H_0|x)}}{\frac{\pi(H_1)}{\pi(H_0)}} = \frac{\frac{\pi(H_1|x)f(x)}{\pi(H_1)}}{\frac{\pi(H_0|x)f(x)}{\pi(H_0)}} = \frac{f(x|H_1)}{f(x|H_0)} = \frac{\int_{\Theta_1} f(x|\theta, H_1)\pi(\theta|H_1)d\theta}{\int_{\Theta_0} f(x|\theta, H_0)\pi(\theta|H_0)d\theta}$$

The hypothesis (model) with the **higher** model evidence is chosen.

Notes on Bayes factors

- No control of **type I error** probability.
- **Compare** H_0 with H_1 unlike frequentist inference that focuses on H_0 .
- Labels H_0 or H_1 **do not matter**.
- Except for some specific cases, require **proper** priors.
- It is easy to extend to **more** hypotheses.

Bayesian Occam's razor

Bayesian Occam's razor: Models with more parameters (more complex models) will not necessarily have higher marginal likelihood.

Conservation of probability mass: More complex models will handle more complex datasets adequately. But the probabilities over all these datasets will have to sum to one.

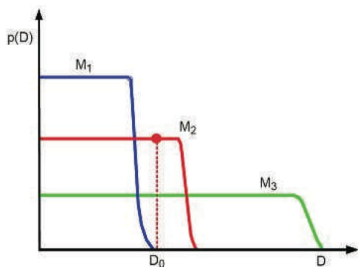


Figure 5.6 A schematic illustration of the Bayesian Occam's razor. The broad (green) curve corresponds to a complex model, the narrow (blue) curve to a simple model, and the middle (red) curve is just right. Based on Figure 3.13 of (Bishop 2006a). See also (Murray and Ghahramani 2005, Figure 2) for a similar plot produced on real data.

Jeffreys-Lindley-Bartlett (Lindley) paradox - example 1

Real data example: A person claimed to possess extrasensory capacities (ESP) and can alter the outcome of a machine that output 0, 1 with probability $\theta = 0.5$ (H_0). H_1 is $\theta \neq 0.5$.

In 104.490.000 trials, there were 52.263.471 ones.



Jeffreys-Lindley-Bartlett (Lindley) paradox - example 1

Under frequentist inference we reject the null and **conclude ESP**; $p\text{-value} \ll 0.01$.

Bayes factor favours H_1 therefore **rejecting** the ESP claim.

Maybe not a **paradox**. Frequentist testing asks the question is $\theta = 0.5$?

Bayesian testing **compares** a model with $\theta = 0.5$ and a model with θ drawn uniformly from $(0, 1)$ as to how they explain the data.

Note on p-value: A more careful study of the problem would also check the power and provide a much lower threshold for the p-value.

Jeffreys-Lindley-Bartlett (Lindley) paradox - example 2

Let $y = (y_1, \dots, y_n)$ iid from the $N(\theta, 1)$ distribution, and consider testing $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$.

The **Bayes factor** in favour of H_0 is

$$B_{01} = \frac{\exp\{-n(\bar{y}_n)^2/2\}}{\int_{-\infty}^{+\infty} \exp\{-n(\bar{y}_n - \theta)^2/2\} \pi(\theta) d\theta}$$

Assume the improper **Jeffreys prior** $p(\theta) = c$. Then

$$B_{01} = \frac{\exp\{-n(\bar{y}_n)^2/2\}}{c \int_{-\infty}^{+\infty} \exp\{-n(\bar{y}_n - \theta)^2/2\} d\theta} = \frac{\exp\{-n(\bar{y}_n)^2/2\}}{c\sqrt{2\pi/n}}$$

The decision depends on the **arbitrary** constant c !

Jeffreys-Lindley-(Bartlett) paradox - example 2 (cont'd)

Consider the **low informative** prior $N(0, \tau^2)$ for some **big** τ^2 . The **Bayes factor** is

$$B_{01} = \frac{\exp\{-n(\bar{y}_n)^2/2\}}{\int_{-\infty}^{+\infty} \exp\{-n(\bar{y}_n - \theta)^2/2\} (2\pi\tau^2)^{-1/2} \exp(-\theta^2/2\tau^2) d\theta}$$

As $\tau \rightarrow \infty$, $B_{01} \rightarrow \infty$ regardless of \bar{y}_n (except if $\bar{y}_n = 0$). So for a near-infinite value of τ^2 we will **always** choose H_0 .

It is therefore clear that **more thought** should be put on the choice of $\pi(\theta)$ when it come to testing.

If we don't have information we still need to put **some** information but not **too much**.

Unit information priors

In the previous example the **unit information** prior is the $N(\mu_0, 1)$, i.e. putting the same prior variance as the variance of each data point.

The posterior is $N(\mu_n, \tau_n^2)$ with

$$\mu_n = \frac{1}{n+1}(\mu_0 + n\bar{y}), \quad \tau_n^2 = \frac{1}{n+1}$$

This prior is like adding **one** more observation equal to μ_0 . In fact σ^2 corresponds to **Fisher** information from **one** data point.

Cheat (add information), but as **little** as possible.

Summary on Bayesian Hypothesis testing

- Use the **Bayes factor**. In most cases it requires **proper** priors.
- Bayes factor can be computed in two ways; either can be used. The **model evidence** is sometimes the only option and can be computationally expensive to compute.
- For testing **simple versus simple** hypothesis the prior plays **no role** so any prior can be used.
- For testing hypotheses with the θ of **equal** dimension, e.g. $H_0 : \theta < 0$ vs $H_1 : \theta \geq 0$, priors with big variance (in some cases even improper priors) are fine.
- But for testing hypotheses of **different** dimension, e.g. $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$, such priors may lead to **Lindley's paradox**. Unit information priors are the recommended option then.

Outline

- 1 Hypothesis testing
- 2 Prediction
- 3 Monte Carlo Integration

Prediction problem

- Consider the **data** $x = (x_1, \dots, x_n)$.
- Assign **model-likelihood** $f(x|\theta)$ with some unknown parameters θ .
- (In Bayesian Inference) Assign a **prior** on θ .
- Consider a **future** observation y from the **same** model $f(y|\theta)$.
Provide
 - ▶ a point estimate (prediction) of y
 - ▶ an interval for y with high probability (prediction interval)
 - ▶ choose between two or more hypotheses regarding y , e.g. $y > 0$ or $y \leq 0$.

Sources of uncertainty in prediction

Even under the assumption that the new observation follows the same adopted model there are still two sources of error:

- 1 Every future value is a random event on its own.
- 2 The parameters are unknown.

Frequentist inference takes into account 1 but it is not clear what to do for 2 (perhaps a bootstrap approach).

Bayesian Inference handles both 1 and 2 via the predictive distribution

$$f(y|x) = \int f(y|\theta)\pi(\theta|x)d\theta$$

In the presence of several models we can treat the model indicator as part of θ . This is known as model averaging.

Example: Exp-Gamma conjugate family

Let $x = (x_1, \dots, x_n)$ be a random sample from an $\text{Exp}(\lambda)$. A $\text{Gamma}(\alpha, \beta)$ **prior** on λ gives the **posterior** $\text{Gamma}(n + \alpha, n\bar{x} + \beta)$.

The **predictive** distribution (for $y > 0$) is

$$\begin{aligned} f(y|x) &= \int \lambda \exp(-\lambda y) \frac{(n\bar{x} + \beta)^{n+\alpha}}{\Gamma(n + \alpha)} \lambda^{n+\alpha-1} \exp(-(n\bar{x} + \beta)\lambda) d\lambda \\ &= \frac{(n\bar{x} + \beta)^{n+\alpha}}{\Gamma(n + \alpha)} \int \lambda^{n+\alpha+1-1} \exp(-(n\bar{x} + \beta + y)\lambda) d\lambda \\ &= \frac{(n\bar{x} + \beta)^{n+\alpha}}{\Gamma(n + \alpha)} \frac{\Gamma(n + \alpha + 1)}{(n\bar{x} + \beta + y)^{n+\alpha+1}} \end{aligned}$$

Outline

- 1 Hypothesis testing
- 2 Prediction
- 3 Monte Carlo Integration

Monte Carlo Integration

Monte Carlo Integral

Let $F(x)$ be a probability distribution and $h(x)$ be a function such that $E_X(h(X)) < \infty$. Also let $x = (x_1, \dots, x_n)$ be a sample from F . Then

$$E_X(h(X)) = \int_{\mathcal{X}} h(x) dF(x) \approx \frac{1}{n} \sum_{i=1}^n h(x_i)$$

Implementation: Draw x_1, \dots, x_n from F and calculate the integral using the above estimator. The error may become arbitrarily small.

Monte Carlo Integration (cont'd)

- Note that $\int_{\mathcal{X}} h(x) dF(x)$ covers both discrete and continuous RV cases. In the former case the integral is a sum and in the latter we can write

$$\int_{\mathcal{X}} h(x) dF(x) = \int_{\mathcal{X}} h(x) f(x) dx$$

- Proof:** Direct application of Strong Law of Large numbers:

$$I_n = \frac{1}{n} \sum_{i=1}^n h(x_i) \xrightarrow{\text{a.s.}} \int_{\mathcal{X}} h(x) dF(x) = I$$

- The speed of convergence depends on the variance of I_n

Importance sampler

Suppose that it is **difficult** to simulate from F (with density f), but it is **easy** to generate from G (with density g).

Importance sampler

Let $F(x)$ be a probability distribution and $h(x)$ be a function such that $E_X(h(X)) < \infty$. Also let $x = (x_1, \dots, x_n)$ be a sample from G . Then

$$E_X(h(X)) = \int_{\mathcal{X}} h(x) \frac{f(x)}{g(x)} g(x) dx \approx \frac{1}{n} \sum_{i=1}^n h(x_i) \frac{f(x_i)}{g(x_i)}$$

Importance sampler will **improve** the stability of Monte Carlo integrals if f and g are similar.

Monte Carlo integration in Bayesian Inference

If we **identify** the posterior distribution and we can **simulate** from it (directly or via importance sampling) Bayesian inference is straightforward.

- **Expectations** of functions of the posterior may be accurately approximated. Note that probabilities can be viewed as expectations of indicator functions.
- **Percentiles** can also be accurately approximated by **sorting** the simulated posterior draws.
- Posterior draws may be inserted in $f(y|\theta)$. This will provide draws from the **predictive distribution**.

Percentiles from Expectations

Consider the **indicator** function $I(\theta \in A)$ that takes the value 1 if $\theta \in A$ and 0 otherwise.

For -say- the median $\theta^{0.5}$ we can use the function $I(\theta < \theta^{0.5})$. Then the value of the **following** is

$$E_{\theta|x}(I(\theta < \theta^{0.5}))P(\theta < \theta^{0.5}|x) = 0.5$$

To find $\theta^{0.5}$ one needs to **solve** the following equation

$$E_{\theta|x}(I(\theta < \theta^{0.5})) = P(\theta < \theta^{0.5}|x) = 0.5 \quad (1)$$

Percentiles from Expectations (cont'd)

Suppose that you have $n = 100,000$ **draws** from $\pi(\theta|x)$, Denote by θ_i for $i = 1, \dots, n$.

Consider the **sample median** as an estimate $\hat{\theta}^{0.5}$ for $\theta^{0.5}$. What is the value of $E_{\theta|x}(I(\theta < \hat{\theta}^{0.5}))$?

$$P(\theta < \hat{\theta}^{0.5}|x) = E_{\theta|x}(I(\theta < \hat{\theta}^{0.5})) \stackrel{\text{Monte Carlo}}{\approx} \frac{1}{n} \sum_{i=1}^n I(\theta_i < \hat{\theta}^{0.5}) = 0.5$$

In other words, $\hat{\theta}^{0.5}$ is a **numerical** solution to (1). For large n the Monte Carlo error goes to 0.

Reading

J.O. Berger:

Sections 2.4.4, 2.4.4, 4.3.3, 4.3.4 and 4.4.3

Gamerman & Lopes:

Sections 3.1 3.2.1 3.2.2 3.4 5.1 5.2