

Bayesian Statistics

STATISTICAL COMPUTING WORKSHOP, ISI DELHI

DAY 2



The Bayesian Philosophy

- We often have some idea/belief about parameters before seeing any data
- For example, rate of twin births in India
- After we see the data, that belief will potentially change based on the data
- The Bayesian philosophy seeks to quantify these beliefs (or lack thereof) in terms of random variables

Prior and Posterior

- The certainty (or uncertainty) about the parameter I have BEFORE seeing the data can be quantified using a distribution
- This is my prior distribution
- The updated uncertainty about the parameter I have AFTER seeing the data can also be quantified using another distribution
- This is my posterior distribution

CAUTION

- The Bayesian philosophy does NOT say that there is no true value for the parameter θ !
 - There is a true value of the parameter θ_0 that generated the data
 - Our uncertainty about the parameter θ is quantified by the prior and posterior distributions.
-
- But, why bother with this?

The Theoretical Answer

- There are scenarios when the answer from the Frequentist approach is nonsensical.
- If you see in your data that no twin births happened in the last month, the frequentist answer would be that the rate of twin births is 0
- This does not make sense, since twin births happen.
- The Bayesian estimate on the other hand will never be zero, because it takes into account your prior belief (or knowledge) that twin births happen rarely.

The Practical Answer

- There is always uncertainty associated with an estimate
- We give standard deviations or confidence intervals to quantify this
- Often using asymptotic results
- The Bayesian methods does this uncertainty quantification by design
- We are not interested in a single estimator
- We are interested in the entire distribution of my belief about the parameter
- So uncertainty quantification is natural and does NOT require asymptotic results

History of Bayesian Statistics

- Named after Thomas Bayes (1701-1761)
- Developed the Bayes theorem: $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$
- Developed independently by Pierre-Simon Laplace (1749-1827)
- These methods were coined “Bayesian” only around 1950s
- They were generally unfavoured out of philosophical and practical consideration
- Advent of computers and algorithms popularized them around 1970s



Bayesian Modelling

- Observe data $X_1, X_2, \dots, X_n \sim F_\theta$
- Quantify uncertainty about the parameter using a random variable θ with distribution $\pi(\theta)$; i.e., $\theta \sim \pi(\theta)$
- Now you want the uncertainty about the parameter, given you have observed the data, i.e., $\pi(\theta | X_1, \dots, X_n)$
- Using the Bayes Theorem: $\pi(\theta | \mathbf{X}) = \frac{f(\mathbf{X} | \theta) \pi(\theta)}{f(\mathbf{X})}$
- How do we get $f(\mathbf{X})$?

Example

- Suppose $X_1, \dots, X_n \sim \text{Ber}(p)$
- Also suppose p is equally likely to take values 0.1, 0.5, and 0.9.

- So, $\pi(p) = \begin{cases} \frac{1}{3}, & p \in \{0.1, 0.5, 0.9\} \\ 0, & p \in \{0.1, 0.5, 0.9\}^c \end{cases}$

- Then we can compute

$$\pi(p|\mathbf{X}) = \frac{f(\mathbf{X}|p)\pi(p)}{f(\mathbf{X}|0.1)\pi(0.1) + f(\mathbf{X}|0.5)\pi(0.5) + f(\mathbf{X}|0.9)\pi(0.9)}$$

- Quick observation: $\pi(p|X) = 0$ if $\pi(p) = 0$

Another Scenario

- Suppose $X_1, X_2, \dots, X_n \sim F_\theta$
- And suppose $\theta \sim \Pi$
- Then, we can compute

$$\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{\int f(X|\theta)\pi(\theta) d\theta}$$

- The denominator may not be tractable!

Choice of Priors

- A largely debated area of Bayesian philosophy
- Priors should match the belief/information you have
- When you do not have any information, the prior should reflect that
- How should it reflect that?
- Priors which include specific, definite information about the parameter, are called informative priors
- Priors which convey partial information about the parameter, are called weakly informative priors
- Priors which carry no information, are called uninformative priors

Informative Priors

- Assume that we are interested in tomorrow's temperature at 8 AM in Delhi.
- We know today's temperature at 8 AM, and we know that day-to-day variation in temperature in Delhi is about 2°C
- So, I can take the my prior to be normally distributed with today's temperature as mean and standard deviation being 2°C
- Then I collect data hourly from now onwards till the afternoon and infer about it based the data and my prior.
- This would be an informative prior, since I am using very specific and definite information about the temperature.

Weakly Informative Priors

- Informative priors tend to put a lot of weight on the prior and influence the inference heavily.
- They also do not allow for much flexibility.
- Weakly informative priors use information that is relevant but not too specific to allow the data to guide the inference with a little nudge from the prior.
- An example would be to use a normally distributed prior for the previous problem with historically average temperature as the mean and monthly variation in temperature as the standard deviation.

Uninformative Priors

- Uninformative priors are priors that carry no information about the parameter we want to estimate.
- You may think that putting a $U(-273, 200)$ prior on the temperature would be uninformative
- But, strictly speaking, it is not!
- Uninformative priors are subject to regular philosophical debate for Bayesians
- Most uninformative priors tend to be improper priors
- One very common such choice is the flat prior

Improper Priors

- Improper priors are priors that have a “density” function which is not integrable
- The flat prior is $\pi(\theta) = 1, \forall \theta \in \mathbb{R}$
- The resulting posterior from these priors may be improper themselves
- In such a case, most traditional Bayesian results may not hold
- Hence, anytime one uses an improper prior, one *MUST* verify that the resulting posterior is proper before proceeding to inference.

Conjugate Priors

- Conjugate priors are priors often used for convenience
- Conjugate priors have the specialty that if a conjugate prior is used for a parameter with a certain data distribution, then the posterior distribution of the parameter will also be the same as the prior distribution (albeit different hyperparameters)
- Examples: When $X_1, \dots, X_n | \mu \sim N(\mu, 1)$, assuming $\mu \sim N(\mu_0, \sigma_0^2)$ leads to

$$\mu | X_1, \dots, X_n \sim N \left(\frac{\frac{\mu_0}{\sigma_0^2} + n\bar{x}}{n + \frac{1}{\sigma_0^2}}, \frac{1}{n + \frac{1}{\sigma_0^2}} \right)$$

Example of Conjugate Priors

Data Distribution	Parameter	Conjugate Prior	Posterior
Binomial(m, p)	p	Beta(a, b)	Beta($a + n\bar{x}, b + mn - n\bar{x}$)
Poisson(λ)	λ	Gamma(α, β)	Gamma($\alpha + n\bar{x}, \beta + n$)
Normal(μ, σ^2) with known σ^2	μ	Normal(μ_0, σ_0^2)	Normal $\left(\frac{\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right)}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\right)$
Normal(μ, σ^2) with known μ	σ^2	Inverse-Gamma(α, β)	Inverse-Gamma $\left(\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}\right)$
Gamma(α, β) with known α	β	Gamma(α_0, β_0)	Gamma($\alpha_0 + n\alpha, \beta_0 + n\bar{x}$)

Inference: MAP Estimators

- Given a data likelihood and a prior distribution, the posterior can be computed as
$$\pi(\theta|X) \propto f(X|\theta)\pi(\theta)$$
- For inference, we could simply maximize this as a function of θ
- This is called Maximize *aposteriori* or MAP estimator
- Note that taking $\pi(\theta) = 1$ makes MAP and MLE same!
- However, MAP estimators are discouraged as the objective is not the posterior mode, but the entire distribution
- Ideally, we should identify the posterior distribution OR draw samples from it

Bayes Risk

- For any parameter θ , its estimator $\hat{\theta}$, we want to minimize some loss function $L(\hat{\theta}, \theta)$ over all θ and $\hat{\theta}$
- Squared error loss: $(\hat{\theta} - \theta)^2$, Absolute error loss: $|\hat{\theta} - \theta|$, 0-1 Loss: $\mathbb{I}(\hat{\theta} \neq \theta)$
- Since it is impossible, we instead minimize a Risk function: $E(L(\hat{\theta}, \theta))$
- Bayes risk is when the expectation is computed with respect to the posterior distribution
- Bayes risk: $E_{\pi|X}(L(\hat{\theta}, \theta))$

Bayes Estimators

- If you use squared error loss, the Bayes estimator is the posterior mean
- For absolute error loss, the Bayes estimator is the posterior median
- For 0-1 loss, the Bayes estimator is the posterior mode

Examples

- For Binomial(m, p) data with Beta(a, b) prior for p , the Bayes estimator would be the posterior mean for the squared error loss.

- Then $\hat{p}_B = \frac{a+n\bar{x}}{a+b+mn}$

- For $N(\mu, 1)$ data with $N(\mu_0, \sigma_0^2)$ prior for μ , the Bayes estimator would be the posterior median for the absolute error loss.

- Then $\hat{\mu}_B = \frac{\frac{\mu_0}{\sigma_0^2} + n\bar{x}}{\frac{1}{\sigma_0^2} + n}$

Unbiasedness, Consistency, Efficiency

- Bayes Estimators are usually not unbiased
- Bayes estimators are usually consistent
- Bayes estimators are usually asymptotically efficient
- [Bernstein-von Mises]: If Π_n is the sequence of posterior distributions for θ with the MLE $\hat{\theta}_n$, then under some mild conditions,

$$\left\| \Pi_n - N \left(\hat{\theta}_n, I_n^{-1}(\hat{\theta}_n) \right) \right\|_{TV} \xrightarrow{P_{\theta_0}} 0$$

as $n \rightarrow \infty$

Uncertainty Quantification

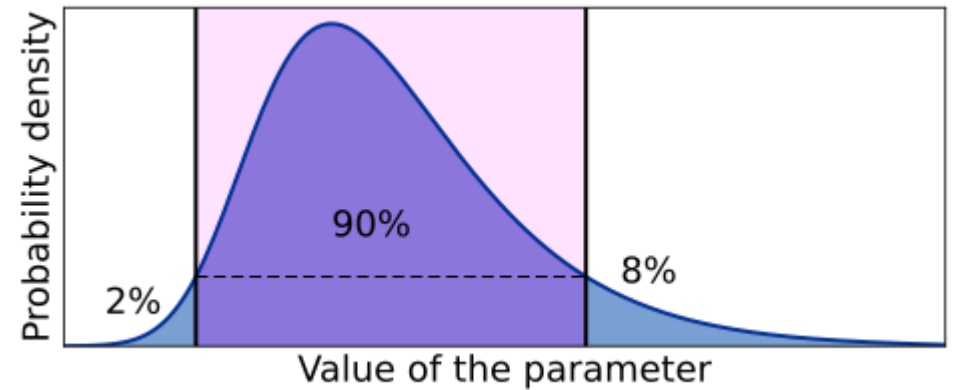
- Computing the posterior standard deviation or posterior quantiles are good ideas for uncertainty quantification about the parameter.
- If the posterior distribution is known, we can simply use the standard deviation of the said distribution as a measure of uncertainty.
- If we have samples from this distribution, we can also compute the sample variance based on this sample and use that as a measure.
- For interval estimators, the Bayesian counterpart is called Credible Region

Credible Regions

- For any parameter θ , a region R is said to be a γ credible region if
$$P(\theta \in R | X) = \gamma$$
- There can be infinitely many γ -credible regions
- For univariate parameter θ we usually use either
 - Highest Density Region (HDR)
OR
 - Equi-tailed Credible Interval

Highest Density Region

- The highest density region is the smallest region that contains $100\gamma\%$ of the posterior probability mass
- Can be a union of disjoint intervals for when the data is multimodal



Equi-tailed Credible Interval

- This is computed by computing the posterior quantiles for $\frac{1-\gamma}{2}$ and $\frac{1+\gamma}{2}$
- If the corresponding quantiles are $q_{\frac{1-\gamma}{2}}$ and $q_{\frac{1+\gamma}{2}}$, then the corresponding credible interval is $[q_{\frac{1-\gamma}{2}}, q_{\frac{1+\gamma}{2}}]$