

ST308 - Lent term

Bayesian Inference

Kostas Kalogeropoulos

Credible Intervals - Priors - Multiparameter models

Outline

Topics covered: Symmetric Credible Intervals, Highest Region Sets, Prior Elicitation, Jeffreys prior, joint and marginal posteriors

1 Credible Intervals (sets)

2 Priors

3 Multiparameter models

Outline

1 Credible Intervals (sets)

2 Priors

3 Multiparameter models

Interval Estimation problem

- Collect the **data** $x = (x_1, \dots, x_n)$ from an experiment.
- Assign **likelihood** (model).
- (In Bayesian Inference) Assign a **prior** to the unknown parameters θ .
- Based on the above decide on a **'good' interval/set of values** for θ .

Confidence vs Credible Intervals

Confidence Interval with level $1-\alpha$

If the experiment was repeated many times, the interval would contain the true value of θ in $100(1 - \alpha)\%$ of them.

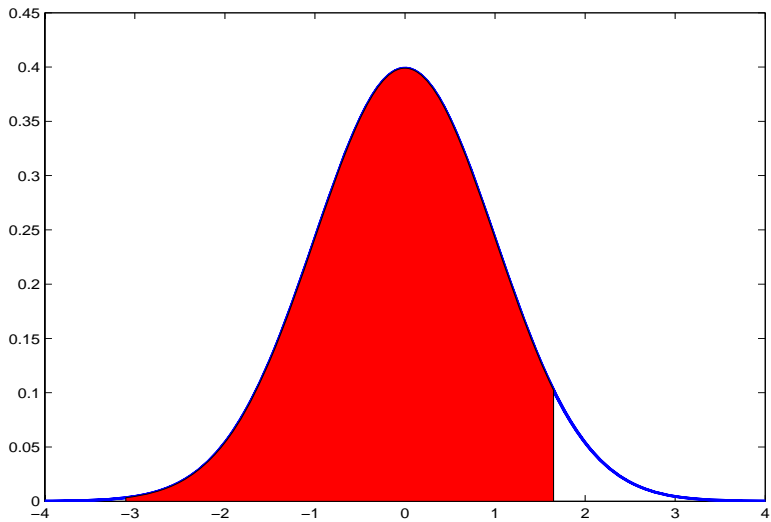
Credible Intervals of $1-\alpha$

The parameter θ is in the credible interval with probability $1 - \alpha$.

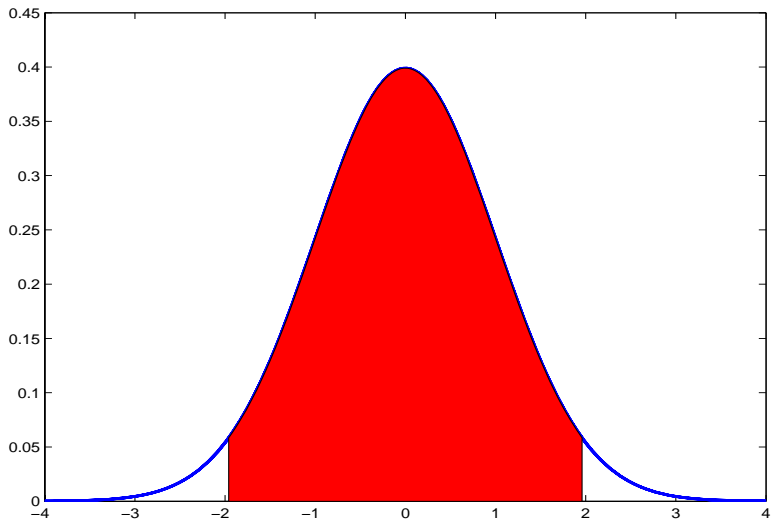
A $100(1 - \alpha)\%$ **credible set** C_α for θ satisfies

$$P(\theta \in C_\alpha | x) = 1 - \alpha$$

A 95% credible set



A shorter 95% credible set



Highest region sets

Highest (posterior density) region sets

Let $H_\gamma = \{\theta : \pi(\theta|x) > \gamma\}$. If we set γ such that

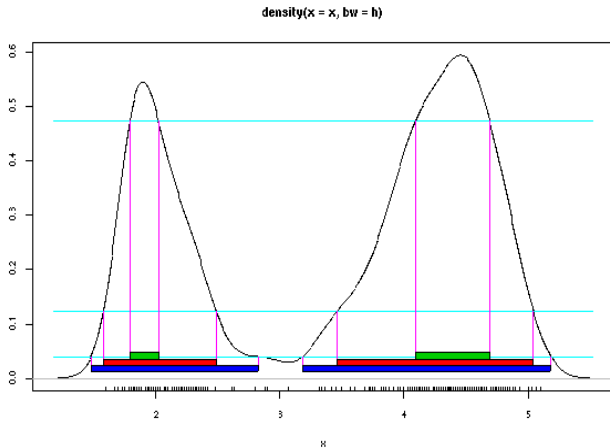
$$P(\theta \in H_\gamma|x) = 1 - \alpha,$$

we get the $100(1 - \alpha)\%$ highest region credible set.

Notes:

- Highest region sets are the **shortest**. Also intuitive, as we want to report the most **plausible** values of θ .
- Not always easy to calculate highest region sets. In most cases **symmetric sets** just report the $\frac{\alpha}{2}$, $1 - \frac{\alpha}{2}$ posterior **percentiles**.

Highest posterior density region sets



Example: $\text{Normal}(\theta, \sigma^2)$ - σ^2 known

Let $x = (x_1, \dots, x_n)$ be a **random sample** from the $N(\theta, \sigma^2)$, with σ^2 known and the **prior** for θ being a $N(\mu, \tau^2)$. The **posterior** for θ is

$$N\left(\frac{\bar{x}\tau^2 + \mu\frac{\sigma^2}{n}}{(\frac{\sigma^2}{n} + \tau^2)}, \frac{\frac{\sigma^2}{n}\tau^2}{(\frac{\sigma^2}{n} + \tau^2)}\right).$$

The symmetric and highest region $100(1 - \alpha)\%$ **credible set** is then

$$\frac{\bar{x}\tau^2 + \mu\frac{\sigma^2}{n}}{(\frac{\sigma^2}{n} + \tau^2)} \pm \mathcal{Z}_{\alpha/2} \sqrt{\frac{\frac{\sigma^2}{n}\tau^2}{(\frac{\sigma^2}{n} + \tau^2)}}.$$

where $\mathcal{Z}_{\alpha/2}$ is the $\frac{\alpha}{2}$ **percentile** of a $N(0, 1)$.

For $\pi(\theta) \propto 1$, we get $\bar{x} \pm \mathcal{Z}_{\alpha/2} \sqrt{\frac{\sigma^2}{n}}$. **Same** as in the frequentist case but with **different** interpretation.

Numerical example: IQ scores

A student uses Bayesian inference in modelling his **IQ score** $x|\theta$ which is $N(\theta, 80)$. His **prior** is $N(110, 120)$. After a score of $x = 98$ the posterior becomes $N(102.8, 48)$.

The frequentist 95% **confidence** interval is

$$\left[98 - 1.96\sqrt{80}, 98 + 1.96\sqrt{80} \right] = [80.5, 115.5]$$

The HPD (and symmetric) 95% **credible** interval is

$$\left[102.8 - 1.96\sqrt{48}, 102.8 + 1.96\sqrt{48} \right] = [89.2, 116.4]$$

which is **shorter** as it includes prior information.

Outline

1 Credible Intervals (sets)

2 Priors

3 Multiparameter models

Prior distributions

- The choice of the prior distribution may be very **influential**.
- In the presence of **information**, the prior should reflect it appropriately.
- If no information is available a vague, **low-informative** prior should be chosen.
- Bayes and Laplace used **uniform** priors - Not always possible/sensible.

Conjugate priors

Posterior and prior have the **same** distribution. Only the parameters are **updated**. Used primarily for **convenience**.

- **Binomial-Beta:** The prior is a $\text{Beta}(\alpha, \beta)$, the data consist of the observation x and the posterior is a $\text{Beta}(\alpha + x, n - x + \beta)$.
- **Poisson-Gamma:** The prior is a $\text{Gamma}(\alpha, \beta)$, the data consist of the random sample $x = (x_1, \dots, x_n)$ and the posterior is a $\text{Gamma}(\alpha + \sum x_i, n + \beta)$.
- **Normal-Normal:** The prior is a $N(\mu, \tau^2)$, the data consist of the random sample $x = (x_1, \dots, x_n)$ and the posterior is a

$$N\left(\frac{\frac{\sigma^2}{n}\mu + \tau^2\bar{x}}{\tau^2 + \frac{\sigma^2}{n}}, \frac{\tau^2\frac{\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}}\right).$$

Prior elicitation

- Information may be **available** prior to the experiment about moments, percentiles, probabilities of certain intervals etc.
- This information can be incorporated in the parameters of the **prior** distribution (prior elicitation).
- If information is **not available** a vague / **low-informative** prior should be chosen.

Example: In a Poisson likelihood experiment we know that the **mean** of λ is around 5 and the **variance** around 4.

A Gamma(α, β) may be chosen. Since $E(\theta) = \frac{a}{b} = 5$ and $\text{Var}(\theta) = \frac{a}{b^2} = 4$, we get that **$\alpha = 6.25$ and $\beta = 1.25$**

Jeffreys' priors

Consider a sample $x = (x_1, \dots, x_n)$ with likelihood $f(x|\theta)$. Fisher information for θ is defined as

$$I(\theta) = E_X \left[\left(\frac{\partial \log f(x|\theta)}{\partial \theta} \right)^2 \right] = -E_X \left(\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} \right)$$

Jeffreys suggested the following prior

$$\pi(\theta) \propto \det(I(\theta))^{1/2}$$

and for the single parameter case

$$\pi(\theta) \propto I(\theta)^{1/2}$$

Invariant to transformations

Jeffreys prior is **invariant to transformations**.

Lemma

Let $\phi = h(\theta)$ or else $\theta = g(\phi)$. Then if $\theta \sim \pi_\theta(\theta)$ the pdf of ϕ is

$$\pi_\phi(\phi) = \pi_\theta(g(\phi)) \left| \frac{\partial g(\phi)}{\partial \phi} \right| = \pi_\theta(\theta) \left| \frac{\partial \theta}{\partial \phi} \right|$$

Now take

$$\begin{aligned} \pi_\phi(\phi) &= \pi_\theta(\theta) \left| \frac{\partial \theta}{\partial \phi} \right| \propto E_X \left[\left(\frac{\partial \log f(x|\theta)}{\partial \theta} \right)^2 \right]^{1/2} \left(\left| \frac{\partial \theta}{\partial \phi} \right|^2 \right)^{1/2} \\ &= E_X \left[\left(\frac{\partial \log f(x|\theta)}{\partial \theta} \frac{\partial \theta}{\partial \phi} \right)^2 \right]^{1/2} = E_X \left[\left(\frac{\partial \log f(x|\phi)}{\partial \phi} \right)^2 \right]^{1/2} = I(\phi)^{1/2} \end{aligned}$$

Example 1: Normal likelihood

Let $x = (x_1, \dots, x_n)'$ be a random sample from $N(\theta, \sigma^2)$, σ^2 known.

$$\log f(x|\theta) = \log \left[\exp \left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2} \right) \right] = -\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2},$$

$$\frac{\partial}{\partial \theta} \log f(x|\theta) = \frac{\sum_{i=1}^n (x_i - \theta)}{\sigma^2} = \frac{\sum_{i=1}^n x_i - n\theta}{\sigma^2} = \frac{n(\bar{x} - \theta)}{\sigma^2}.$$

$$\mathcal{I}(\theta|x) = -E_x \left(\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right) = -E_x \left(-\frac{n}{\sigma^2} \right) = \frac{n}{\sigma^2}$$

Hence the Jeffreys prior is $\pi(\theta) \propto 1$.

The **posterior** will be

$$\pi(\theta|x) \propto \exp \left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2} \right) \propto \exp \left(-\frac{\theta^2 - 2\theta\bar{x}}{2\frac{\sigma^2}{n}} \right) \stackrel{\mathcal{D}}{=} N \left(\bar{x}, \frac{\sigma^2}{n} \right)$$

Example 2: Poisson likelihood

Let $x = (x_1, \dots, x_n)'$ be a random sample from **Poisson**(λ).

$$\log f(x|\lambda) = \log \left[\exp(-n\lambda) \lambda^{\sum x_i} \right] = -n\lambda + \log(\lambda) \sum_{i=1}^n x_i,$$

$$\frac{\partial}{\partial \lambda} \log f(x|\lambda) = \frac{\sum_{i=1}^n x_i}{\lambda} - n.$$

$$\mathcal{I}(\lambda|x) = -E_X \left(\frac{\partial^2}{\partial \lambda^2} \log f(x|\lambda) \right) = -E_X \left(-\frac{\sum_{i=1}^n x_i}{\lambda^2} \right) = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$$

Jeffreys prior: $\pi(\theta) \propto 1/\sqrt{\lambda}$. **Posterior:** $\text{Gamma}(\frac{1}{2} + \sum x_i, n)$.

Pros and Cons

- Jeffreys' prior provides a popular for 'objective' Bayesian inference.
- Usually improper, therefore the propriety of the posterior should also be checked.
- Requires Fisher information that can be very hard to compute, particularly in high dimensions.
- Overall, the problem of prior specification in higher dimensions is quite challenging.

Outline

- 1 Credible Intervals (sets)
- 2 Priors
- 3 Multiparameter models**

Multiparameter models

We may have **more than one** parameters, say $\theta = (\theta_1, \theta_2)$. As before assign a prior $\pi(\theta_1, \theta_2)$ and obtain the posterior

$$\pi(\theta_1, \theta_2 | x) = \frac{f(x | \theta_1, \theta_2) \pi(\theta_1, \theta_2)}{\int \int f(x | \theta_1, \theta_2) \pi(\theta_1, \theta_2) d\theta_1 d\theta_2}$$

If interest mainly lies in θ_1 , the **marginal posterior** of θ_1 may be used by averaging over θ_2 .

$$\pi(\theta_1 | x) = \int \pi(\theta_1, \theta_2 | x) d\theta_2$$

Example 3: Normal likelihood for μ and σ^2

Let $x = (x_1, \dots, x_n)$ be a random sample from a $\mathbf{N}(\theta, \sigma^2)$ with both μ and σ^2 unknown.

Likelihood: The **likelihood** is (s^2 denotes the sample variance)

$$f(x|\theta) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{(n-1)s^2 + n(\bar{x} - \theta)^2}{2\sigma^2}\right)$$

Prior: Consider the **improper** prior is $\pi(\theta, \sigma^2) \propto 1/\sigma^2$

Example 3: Normal likelihood for μ and σ^2 (cont'd)

Posterior: factorised as $\pi(\theta|x, \sigma^2)\pi(\sigma^2|x)$

$$\begin{aligned}\pi(\theta, \sigma^2|x) &\propto (\sigma^2)^{-n/2-1} \exp\left(-\frac{(n-1)s^2 + n(\bar{x} - \theta)^2}{2\sigma^2}\right) \\&= (\sigma^2)^{-n/2-1} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \exp\left(-\frac{(\bar{x} - \theta)^2}{2\frac{\sigma^2}{n}}\right) \\&= \sqrt{2\pi}\left(\frac{\sigma^2}{n}\right)^{1/2}(\sigma^2)^{-n/2-1} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \times \text{N}\left(\bar{x}, \frac{\sigma^2}{n}\right) \\&= \text{IGamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \times \text{N}\left(\bar{x}, \frac{\sigma^2}{n}\right)\end{aligned}$$

(Note: The distributions above represent their densities)

Notes on multiparameter models

- Substantial increase in calculations. Inference is feasible only for a few models.
- This can be resolved using MCMC or approximate methods.
- Specification of priors may still be an issue.

Reading

J.O. Berger Sections 3.3, 4.2 and 4.3.2