Bayesian parameter inference

Jean-Michel Marin

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck (IMAG)

HMMA401 / 2020-2021

The Bayesian paradigm

Bayes theorem = Inversion of probabilities

If A and B are events such that $\mathbb{P}(B) \neq 0$,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} =$$

$$\frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(\bar{A})\mathbb{P}(B|\bar{A})}$$

The Bayesian paradigm

Subjectivism

Frank Plumpton Ramsey (1903-1930)

Bruno de Finetti (1906-1985)

Leonard Jimmie Savage (1921-1971)

The Bayesian paradigm

Given an iid sample $\mathscr{D}_n = (x_1, \ldots, x_n)$ from a density $f(x|\theta)$, depending upon an unknown parameter $\theta \in \Theta$, the associated likelihood function is

$$\ell(\theta|\mathscr{D}_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

The Bayesian paradigm

When $\mathscr{D}_n$ is a normal $\mathscr{N}(\mu, \sigma^2)$ sample of size n and $\theta = (\mu, \sigma^2)$, we get

$$\begin{aligned}
\ell(\theta|\mathscr{D}_n) &= \prod_{i=1}^{n} \exp\{-(x_i - \mu)^2/2\sigma^2\}/\sqrt{2\pi}\sigma \\
&\propto \exp\left\{-\sum_{i=1}(x_i - \mu)^2/2\sigma^2\right\}/\sigma^n \\
&\propto \exp\left\{-\left(n\mu^2 - 2n\bar{x}\mu + \sum_{i=1} x_i^2\right)/2\sigma^2\right\}/\sigma^n \\
&\propto \exp\left\{-\left[n(\mu - \bar{x})^2 + s^2\right]/2\sigma^2\right\}/\sigma^n,
\end{aligned}$$

$\bar{x}$ denotes the empirical mean and $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$

The Bayesian paradigm

In the Bayesian approach θ is considered as a random variable

The Bayesian paradigm

In the Bayesian approach θ is considered as a random variable

In some sense, the likelihood function is transformed into a *posterior* distribution, which is a valid probability distribution on Θ

$$\pi(\theta|\mathscr{D}_n) = \frac{\ell(\theta|\mathscr{D}_n)\pi(\theta)}{\int \ell(\theta|\mathscr{D}_n)\pi(\theta)\, d\theta}$$

The Bayesian paradigm

In the Bayesian approach θ is considered as a random variable

In some sense, the likelihood function is transformed into a *posterior* distribution, which is a valid probability distribution on Θ

$$\pi(\theta|\mathscr{D}_n) = \frac{\ell(\theta|\mathscr{D}_n)\pi(\theta)}{\int \ell(\theta|\mathscr{D}_n)\pi(\theta)\,\mathrm{d}\theta}$$

$\pi(\theta)$ is called the *prior* distribution and it has to be chosen to start the analysis

The Bayesian paradigm

The posterior density is a probability density on the parameter, which does not mean the parameter θ need be a genuine random variable

The Bayesian paradigm

The posterior density is a probability density on the parameter, which does not mean the parameter θ need be a genuine random variable

This density is used as an inferential tool, not as a truthful representation

The Bayesian paradigm

Two motivations:

- the prior distribution summarizes the *prior information* on θ. However, the choice of $\pi(\theta)$ is often decided on practical grounds rather than strong subjective beliefs
- the Bayesian approach provides a fully probabilistic framework for the inferential analysis, with respect to a reference measure $\pi(\theta)$

The Bayesian paradigm

Suppose $\mathscr{D}_n$ is a normal $\mathscr{N}(\mu, \sigma^2)$ sample of size n

When σ^2 is known, if $\mu \sim \mathscr{N}\left(0, \sigma^2\right)$, then

$$
\begin{aligned}
\pi(\mu|\mathscr{D}_n) &\propto \pi(\mu)\,\ell(\theta|\mathscr{D}_n) \\
&\propto \exp\{-\mu^2/2\sigma^2\}\,\exp\left\{-n(\bar{x}-\mu)^2/2\sigma^2\right\} \\
&\propto \exp\left\{-(n+1)\mu^2/2\sigma^2 + 2n\mu\bar{x}/2\sigma^2\right\} \\
&\propto \exp\left\{-(n+1)[\mu - n\bar{x}/(n+1)]^2/2\sigma^2\right\}
\end{aligned}
$$

The Bayesian paradigm

Suppose $\mathscr{D}_n$ is a normal $\mathscr{N}(\mu, \sigma^2)$ sample of size n

When σ^2 is known, if $\mu \sim \mathscr{N}\left(0, \sigma^2\right)$, then

$$
\begin{aligned}
\pi(\mu|\mathscr{D}_n) &\propto \pi(\mu)\,\ell(\theta|\mathscr{D}_n) \\
&\propto \exp\{-\mu^2/2\sigma^2\}\,\exp\left\{-n(\bar{x}-\mu)^2/2\sigma^2\right\} \\
&\propto \exp\left\{-(n+1)\mu^2/2\sigma^2 + 2n\mu\bar{x}/2\sigma^2\right\} \\
&\propto \exp\left\{-(n+1)[\mu - n\bar{x}/(n+1)]^2/2\sigma^2\right\}
\end{aligned}
$$

$$
\boxed{\mu|\mathscr{D}_n \sim \mathscr{N}\left(n\bar{x}/(n+1), \sigma^2/(n+1)\right)}
$$

The Bayesian paradigm

When σ^2 is unknown, $\theta = (\mu, \sigma^2)$, if $\mu|\sigma^2 \sim \mathscr{N}\left(0, \sigma^2\right)$ and $\sigma^2 \sim \mathscr{I}\mathscr{G}(1, 1)$, then $\pi((\mu, \sigma^2)|\mathscr{D}_n) \propto \pi(\sigma^2) \times \pi(\mu|\sigma^2) \times f(\mathscr{D}_n|\mu, \sigma^2)$

$$\propto (\sigma^{-2})^{1/2+2} \exp\left\{-(\mu^2 + 2)/2\sigma^2\right\} \mathbf{1}_{\sigma^2>0}$$

$$(\sigma^{-2})^{n/2} \exp\left\{-\left(n(\mu - \overline{x})^2 + s^2\right)/2\sigma^2\right\}$$

The Bayesian paradigm

When σ^2 is unknown, $\theta = (\mu, \sigma^2)$, if $\mu|\sigma^2 \sim \mathscr{N}\left(0, \sigma^2\right)$ and $\sigma^2 \sim \mathscr{IG}(1,1)$, then $\pi((\mu, \sigma^2)|\mathscr{D}_n) \propto \pi(\sigma^2) \times \pi(\mu|\sigma^2) \times f(\mathscr{D}_n|\mu, \sigma^2)$

$$\propto (\sigma^{-2})^{1/2+2} \exp\left\{-(\mu^2+2)/2\sigma^2\right\} \mathbf{1}_{\sigma^2>0}$$

$$(\sigma^{-2})^{n/2} \exp\left\{-\left(n(\mu-\bar{x})^2+s^2\right)/2\sigma^2\right\}$$

$$\mu|\mathscr{D}_n, \sigma^2 \sim \mathscr{N}\left(\frac{n\bar{x}}{n+1}, \frac{\sigma^2}{n+1}\right)$$

$$\sigma^2|\mathscr{D}_n \sim \mathscr{IG}\left(\left\{1+\frac{n}{2}\right\}, \left\{1+\frac{s^2}{2}+\frac{n\bar{x}}{2(n+1)}\right\}\right)$$

The Bayesian paradigm

Variability in σ^2 induces more variability in μ, the marginal posterior in μ being then a Student's t distribution

The Bayesian paradigm

Variability in σ^2 induces more variability in μ, the marginal posterior in μ being then a Student's t distribution

$$\mu|\mathscr{D}_n \sim \mathscr{T}\left(n+2, \frac{n\bar{x}}{n+1}, \frac{2+s^2+(n\bar{x})/(n+1)}{(n+1)(n+2)}\right)$$

Bayesian estimates

For a given loss function $L\left(\theta, \hat{\theta}(\mathscr{D}_n)\right)$, we deduce a Bayesian estimate by minimizing the posterior expected loss:

$$\mathbb{E}^{\pi}_{\theta|\mathscr{D}_n}\left(L\left(\theta, \hat{\theta}(\mathscr{D}_n)\right)\right)$$

Bayesian estimates

For a given loss function $\mathrm{L}\left(\theta, \hat{\theta}(\mathscr{D}_n)\right)$, we deduce a Bayesian estimate by minimizing the posterior expected loss:

$$\mathbb{E}^{\pi}_{\theta|\mathscr{D}_n}\left(\mathrm{L}\left(\theta, \hat{\theta}(\mathscr{D}_n)\right)\right)$$

To minimize the posterior expected loss is equivalent to minimize the Bayes risk, the frequentist risk integrated over the prior distribution

Bayesian estimates

For instance, for the L_2 loss function, the corresponding Bayes optimum is the expected value of θ under the posterior distribution,

$$\hat{\theta}(\mathscr{D}_n) = \int \theta \, \pi(\theta|\mathscr{D}_n) \, d\theta = \frac{\int \theta \, \ell(\theta|\mathscr{D}_n) \, \pi(\theta) \, d\theta}{\int \ell(\theta|\mathscr{D}_n) \, \pi(\theta) \, d\theta}$$

Bayesian estimates

When no specific penalty criterion is available, the posterior expectation is often used as a default estimator, although alternatives are also available. For instance, the *maximum a posteriori estimator* (MAP) is defined as

$$\hat{\theta}(\mathscr{D}_n) \in \text{argmax}_\theta \quad \pi(\theta|\mathscr{D}_n)$$

Bayesian estimates

When no specific penalty criterion is available, the posterior expectation is often used as a default estimator, although alternatives are also available. For instance, the *maximum a posteriori estimator* (MAP) is defined as

$$\hat{\theta}(\mathscr{D}_n) \in \text{argmax}_\theta \quad \pi(\theta|\mathscr{D}_n)$$

Similarity of with the maximum likelihood estimator: the influence of the prior distribution $\pi(\theta)$ on the estimate progressively disappears as the number of observations n increases

Conjugate prior

The selection of the prior distribution is an important issue in Bayesian statistics

Conjugate prior

The selection of the prior distribution is an important issue in Bayesian statistics

When prior information is available about the data or the model, it can be used in building the prior

Conjugate prior

The selection of the prior distribution is an important issue in Bayesian statistics

When prior information is available about the data or the model, it can be used in building the prior

In many situations, however, the selection of the prior distribution is quite delicate

Conjugate prior

The selection of the prior distribution is an important issue in Bayesian statistics

When prior information is available about the data or the model, it can be used in building the prior

In many situations, however, the selection of the prior distribution is quite delicate

Since the choice of the prior distribution has a considerable influence on the resulting inference, this inferential step must be conducted with the utmost care

Conjugate prior

Conjugate priors are such that the prior and posterior densities belong to the same parametric family

Conjugate prior

Conjugate priors are such that the prior and posterior densities belong to the same parametric family

An advantage when using a conjugate prior, is that one has to select only a few parameters to determine the prior distribution

Conjugate prior

Conjugate priors are such that the prior and posterior densities belong to the same parametric family

An advantage when using a conjugate prior, is that one has to select only a few parameters to determine the prior distribution

But the information known a priori may be either insufficient or incompatible with the structure imposed by conjugacy

Conjugate prior

Justifications

- Device of virtual past observations
- First approximations to adequate priors, backed up by robustness analysis
- But mostly... tractability and simplicity

Conjugate prior

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta,\sigma^2)$	Normal $\mathcal{N}(\mu,\tau^2)$	$\mathcal{N}(\rho(\sigma^2\mu+\tau^2 x),\rho\sigma^2\tau^2)$ $\rho^{-1}=\sigma^2+\tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+x,\beta+1)$
Gamma $\mathcal{G}(\nu,\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+\nu,\beta+x)$
Binomial $\mathcal{B}(n,\theta)$	Beta $\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+x,\beta+n-x)$

Conjugate prior

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Negative Binomial $\mathcal{N}eg(m,\theta)$	Beta $\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+m,\beta+x)$
Multinomial $\mathcal{M}_k(\theta_1,\ldots,\theta_k)$	Dirichlet $\mathcal{D}(\alpha_1,\ldots,\alpha_k)$	$\mathcal{D}(\alpha_1+x_1,\ldots,\alpha_k+x_k)$
Normal $\mathcal{N}(\mu,1/\theta)$	Gamma $\mathcal{G}a(\alpha,\beta)$	$\mathcal{G}(\alpha+0.5,\beta+(\mu-x)^2/2)$

Noninformative prior

Conjugate priors are nice to work with, but require hyperparameters's determination

Noninformative prior

Conjugate priors are nice to work with, but require hyperparameters's determination

One can opt for a completely different perspective and rely on so-called *noninformative* priors that aim at attenuating the impact of the prior on the resulting inference

Noninformative prior

Conjugate priors are nice to work with, but require hyperparameters's determination

One can opt for a completely different perspective and rely on so-called *noninformative* priors that aim at attenuating the impact of the prior on the resulting inference

These priors are fundamentally defined as coherent extensions of the uniform distribution

Noninformative prior

For unbounded parameter spaces, the densities of noninformative priors actually may fail to integrate to a finite number and they are defined instead as positive measures

Noninformative prior

For unbounded parameter spaces, the densities of noninformative priors actually may fail to integrate to a finite number and they are defined instead as positive measures

Generalized Bayesian estimators with improper prior distributions

Noninformative prior

Location models $x|\theta \sim f(x-\theta)$ are usually associated with flat priors $\pi(\theta) \propto 1$

Noninformative prior

Location models $x|\theta \sim f(x-\theta)$ are usually associated with flat priors $\pi(\theta) \propto 1$

Scale models $x|\theta \sim \frac{1}{\theta} f\left(\frac{x}{\theta}\right)$ are usually associated with the log-transform of a flat prior, that is, $\pi(\theta) \propto 1/\theta \times \mathbf{1}_{\theta>0}$

Jeffreys prior

In a more general setting, the noninformative prior favored by most Bayesians is the so-called **Jeffreys prior** which is related to the Fisher information matrix

$$I_x^F(\theta) = -\mathbb{E}\left(\frac{\partial^2 \log f(x|\theta)}{(\partial\theta)^2}\right)$$

by

$$\pi^J(\theta) \propto \sqrt{|I_x^F(\theta)|} \times \mathbf{1}_{\theta\in\Theta}\,,$$

where $|I|$ denotes the determinant of the matrix I

Jeffreys prior

Suppose $\mathscr{D}_n$ is a normal $\mathscr{N}(\mu, \sigma^2)$ sample of size n and $\theta = (\mu, \sigma^2)$

Jeffreys prior

Suppose $\mathscr{D}_n$ is a normal $\mathscr{N}(\mu, \sigma^2)$ sample of size n and $\theta = (\mu, \sigma^2)$

The Fisher information matrix leads to the Jeffreys prior

$$\pi^J(\mu, \sigma^2) \propto 1/\{(\sigma^2)\}^{3/2}\mathbf{1}_{\sigma^2>0}$$

Jeffreys prior

Suppose $\mathscr{D}_n$ is a normal $\mathscr{N}(\mu, \sigma^2)$ sample of size n and $\theta = (\mu, \sigma^2)$

The Fisher information matrix leads to the Jeffreys prior

$$\pi^J(\mu, \sigma^2) \propto 1/\{(\sigma^2)\}^{3/2}\mathbf{1}_{\sigma^2>0}$$

$$\boxed{\mu|\sigma^2, \mathscr{D}_n \sim \mathscr{N}\left(\bar{x}, \sigma^2/n\right)}$$

$$\boxed{\sigma^2|\mathscr{D}_n \sim \mathscr{IG}\left(n/2, s^2/2\right)}$$

$$\boxed{\mu|\mathscr{D}_n \sim \mathscr{T}\left(n, \bar{x}, s^2/n\right)}$$

Bayesian Credible Intervals

Since the Bayesian approach processes θ as a random variable, a natural definition of a confidence region on θ is to determine $C(\mathscr{D}_n)$ such that

$$\pi(\theta \in C(\mathscr{D}_n)|\mathscr{D}_n) = 1 - \alpha$$

where α is a predetermined level

Bayesian Credible Intervals

Since the Bayesian approach processes θ as a random variable, a natural definition of a confidence region on θ is to determine $C(\mathscr{D}_n)$ such that

$$\pi(\theta \in C(\mathscr{D}_n)|\mathscr{D}_n) = 1 - \alpha$$

where α is a predetermined level

The integration is done over the parameter space, rather than over the observation space

Bayesian Credible Intervals

Since the Bayesian approach processes θ as a random variable, a natural definition of a confidence region on θ is to determine $C(\mathscr{D}_n)$ such that

$$\pi(\theta \in C(\mathscr{D}_n)|\mathscr{D}_n) = 1 - \alpha$$

where α is a predetermined level

The integration is done over the parameter space, rather than over the observation space

The quantity $1 - \alpha$ thus corresponds to the probability that a random θ belongs to this set $C(\mathscr{D}_n)$, rather than to the probability that the random set contains the true value of θ

Bayesian Credible Intervals

Given this drift in the interpretation of a confidence set is called a *credible set* by Bayesians.

Bayesian Credible Intervals

Given this drift in the interpretation of a confidence set is called a *credible set* by Bayesians.

A standard credible set corresponds to the values of θ with the highest posterior values,

$$C(\mathscr{D}_n) = \{\theta;\, \pi(\theta|\mathscr{D}_n) \geqslant k_\alpha\}$$

where k_α is determined by the coverage constraint

Bayesian Credible Intervals

Given this drift in the interpretation of a confidence set is called a *credible set* by Bayesians.

A standard credible set corresponds to the values of θ with the highest posterior values,

$$C(\mathscr{D}_n) = \{\theta;\, \pi(\theta|\mathscr{D}_n) \geqslant k_\alpha\}$$

where k_α is determined by the coverage constraint

This region is called the **Highest Posterior Density** (HPD) region

Bayesian Credible Intervals

Once again, suppose $\mathscr{D}_n$ is a normal $\mathscr{N}(\mu, \sigma^2)$ sample of size n and $\theta = (\mu, \sigma^2)$

$$\boxed{\mu | \sigma^2, \mathscr{D}_n \sim \mathscr{N}\left(\bar{x}, \sigma^2/n\right)}$$

$$\boxed{\sigma^2 | \mathscr{D}_n \sim \mathscr{IG}\left(n/2, s^2/2\right)}$$

$$\boxed{\mu | \mathscr{D}_n \sim \mathscr{T}\left(n, \bar{x}, s^2/n\right)}$$

Bayesian Credible Intervals

Once again, suppose $\mathscr{D}_n$ is a normal $\mathscr{N}(\mu, \sigma^2)$ sample of size n and $\theta = (\mu, \sigma^2)$

$$\boxed{\mu|\sigma^2, \mathscr{D}_n \sim \mathscr{N}\left(\bar{x}, \sigma^2/n\right)}$$

$$\boxed{\sigma^2|\mathscr{D}_n \sim \mathscr{IG}\left(n/2, s^2/2\right)}$$

$$\boxed{\mu|\mathscr{D}_n \sim \mathscr{T}\left(n, \bar{x}, s^2/n\right)}$$

Therefore, the credible interval of probability $1-\alpha$ on μ is

$$[\bar{x} - t_{1-\alpha/2,n}\sqrt{s^2/n}, \bar{x} + t_{1-\alpha/2,n}\sqrt{s^2/n}]$$