Prior Choice

AHMAD PARSIAN

SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE
UNIVERSITY OF TEHRAN

Different types of Bayesians

- Classical Bayesians,
- Modern Parametric Bayesians,
- Subjective Bayesians.

Different types of Bayesians

- Classical Bayesians,
- Modern Parametric Bayesians,
- Subjective Bayesians.

Prior Choice

- Informative prior based on,
    - Expert knowledge (subjective),
    - Historical data (objective).

  Subjective information is based on personal opinions and feelings rather than facts.
  Objective information is based on facts.

Different types of Bayesians

- Classical Bayesians,
- Modern Parametric Bayesians,
- Subjective Bayesians.

Prior Choice

- Informative prior based on,
  - Expert knowledge (subjective),
  - Historical data (objective).

  Subjective information is based on personal opinions and feelings rather than facts.
  Objective information is based on facts.

- Uninformative prior, representing ignorance,
  - Jeffreys prior,
  - Based on data in some way (reference prior).

Classical Bayesians

- The prior is a necessary evil,
- Choose priors that interject the least information possible.

  The least = the minimum that should done in a situation.

Classical Bayesians

- The prior is a necessary evil,
- Choose priors that interject the least information possible.

  The least = the minimum that should done in a situation.

Modern Parametric Bayesians

- The prior is a useful convenience.
- Choose prior distributions with desirable properties (e.g.: conjugacy).
- Given a distributional choice, prior parameters are chosen to interject the least information.

Classical Bayesians

- The prior is a necessary evil,
- Choose priors that interject the least information possible.

  The least = the minimum that should done in a situation.

Modern Parametric Bayesians

- The prior is a useful convenience.
- Choose prior distributions with desirable properties (e.g.: conjugacy).
- Given a distributional choice, prior parameters are chosen to interject the least information.

Subjective Bayesians

- The prior is a summary of old beliefs.
- Choose prior distributions based on previous knowledge (either the results of earlier studies or non-scientific opinion.)

Example

Modern Parametric Bayesians

Suppose $X \sim N(\theta, \sigma^2)$. Let $\tau = 1/\sigma^2$.

Example

Modern Parametric Bayesians

Suppose $X \sim N(\theta, \sigma^2)$. Let $\tau = 1/\sigma^2$.

Q: What prior distribution would a Modern Parametric Bayesians choose to satisfy the demand of convenience?

Example

Modern Parametric Bayesians

Suppose $X \sim N(\theta, \sigma^2)$. Let $\tau = 1/\sigma^2$.

Q: What prior distribution would a Modern Parametric Bayesians choose to satisfy the demand of convenience?

A: Using the definition $\pi(\theta, \tau) = \pi(\theta|\tau)\pi(\tau)$,

Example

Modern Parametric Bayesians

Suppose $X \sim N(\theta, \sigma^2)$. Let $\tau = 1/\sigma^2$.

Q: What prior distribution would a Modern Parametric Bayesians choose to satisfy the demand of convenience?

A: Using the definition $\pi(\theta, \tau) = \pi(\theta|\tau)\pi(\tau)$,

Prior choice is

$$\begin{aligned} \theta|\tau &\sim N(\mu, \sigma_0^2) \\ \tau &\sim Gamma(\alpha, \beta) \end{aligned}$$

And you know that

$$\begin{aligned} \theta|\tau, x &\sim Normal \\ \tau|x &\sim Gamma \end{aligned}$$

Example

(Continued)

Q: What prior distribution would a Lazy Modern Parametric Bayesians choose to satisfy the demand of convenience?

Example

(Continued)

Q: What prior distribution would a Lazy Modern Parametric Bayesians choose to satisfy the demand of convenience?

A: Using the fact (suppose you do not want to think too hard about the prior)

$$\pi(\theta, \tau) = \pi(\theta)\pi(\tau),$$

Example

(Continued)

Q: What prior distribution would a Lazy Modern Parametric Bayesians choose to satisfy the demand of convenience?

A: Using the fact (suppose you do not want to think too hard about the prior)

$$\pi(\theta, \tau) = \pi(\theta)\pi(\tau),$$

Prior choice is

$$\begin{aligned} \theta|\tau &\sim N(0, t) \\ \tau &\sim Gamma(\alpha, \beta) \end{aligned}$$

Obviously, the marginal posterior from this model would be a bit difficult

analytically (in general), but it is easy to implement the Gibbs Sampler.

The Main Talk

$$X = (X_1, , X_n) \quad \sim \quad f_\theta(x)$$

The Main Talk

$$X = (X_1, , X_n) \quad \sim \quad f_\theta(x)$$

$$\theta \quad \sim \quad \pi(\theta)$$

The Main Talk

$$X = (X_1, , X_n) \quad \sim \quad f_\theta(x)$$

$$\theta \quad \sim \quad \pi(\theta)$$

$$\theta|x \quad \sim \quad \pi(\theta|x)$$

$$\pi(\theta|x) \quad = \quad \frac{f_\theta(x)\pi(\theta)}{m(x)},$$

Where $m(x) = \int f_\theta(x)\pi(\theta)d\theta$ is marginal dist. of X.

Let us concentrate on the following problem.

Suppose $X_1, , X_n$ be i.i.d. $B(1, \theta)$, then $Y = \sum X_i \sim B(n, \theta)$

Need a prior on θ:

Let us concentrate on the following problem.

Suppose $X_1, , X_n$ be i.i.d. $B(1, \theta)$, then $Y = \sum X_i \sim B(n, \theta)$

Need a prior on θ:

Take $\theta \sim Beta(\alpha, \beta)$ (Remember that this is a perfectly Subjective choice and anybody can use their own.) So, $\theta | y \sim Beta(y + \alpha, n - y + \beta)$.

Let us concentrate on the following problem.

Suppose $X_1, , X_n$ be i.i.d. $B(1, \theta)$, then $Y = \sum X_i \sim B(n, \theta)$

Need a prior on θ:

Take $\theta \sim Beta(\alpha, \beta)$ (Remember that this is a perfectly Subjective choice and anybody can use their own.) So, $\theta | y \sim Beta(y + \alpha, n - y + \beta)$.

Under Squared Error Loss (SEL), the Bayes estimate is

$$\begin{aligned}
\delta_\pi(y) &= \frac{y + \alpha}{n + \alpha + \beta} \\
&= \frac{n}{n + \alpha + \beta} \frac{y}{n} + \frac{\alpha + \beta}{n + \alpha + \beta} \frac{\alpha}{\alpha + \beta}
\end{aligned}$$

Which is a linear combination of sample mean and prior mean.

We have a coin. Is this a fair coin? i.e., is $\theta = \frac{1}{2}$?

We have a coin. Is this a fair coin? i.e., is $\theta = \frac{1}{2}$?

Suppose you flip it 10 times, and it comes up heads 3 times.

We have a coin. Is this a fair coin? i.e., is $\theta = \frac{1}{2}$?

Suppose you flip it 10 times, and it comes up heads 3 times.

As a frequentist: We use the sample mean, i.e., $\hat{\theta} = \frac{3}{10} = 0.3$.

We have a coin. Is this a fair coin? i.e., is $\theta = \frac{1}{2}$?

Suppose you flip it 10 times, and it comes up heads 3 times.

As a frequentist: We use the sample mean, i.e., $\hat{\theta} = \frac{3}{10} = 0.3$.

As a Bayesian: We have to completely specify the prior distribution, i.e., we have to choose α and β. The Choice again depends on our belief.

Notice that:

- To estimate θ, a Bayesian analyst would put a prior dist. on θ and use the posterior dist. of θ to draw various conclusions: estimating θ with posterior mean.
- When there is no strong prior opinion on what θ is, it is desirable to pick a prior that is NON-INFORMATIVE.

If we feel strongly that this coin is like any other coin and therefore really should be a fair coin, we should choose α and β so that the prior puts all its weight at around $\frac{1}{2}$.

If we feel strongly that this coin is like any other coin and therefore really should be a fair coin, we should choose α and β so that the prior puts all its weight at around $\frac{1}{2}$.

e.g., $\alpha = \beta = 100$, then $E(\theta) = \frac{\alpha}{\alpha+\beta} = \frac{1}{2}$

and $Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} = 0.0016$

Therefore,

$$\delta_\pi(3) = \frac{(3 + 100)}{(10 + 100 + 100)} = 0.4905$$

If we feel strongly that this coin is like any other coin and therefore really should be a fair coin, we should choose α and β so that the prior puts all its weight at around $\frac{1}{2}$.

e.g., $\alpha = \beta = 100$, then $E(\theta) = \frac{\alpha}{\alpha+\beta} = \frac{1}{2}$

and $$Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} = 0.0016$$

Therefore,

$$\delta_\pi(3) = \frac{(3+100)}{(10+100+100)} = 0.4905$$

Clearly for such a strong prior the actual sample almost does not matter:

$y = 0 \rightarrow \delta_\pi(0) = \frac{(0+100)}{(10+100+100)} = 0.476$

$\vdots$

$y = 10 \rightarrow \delta_\pi(10) = \frac{(10+100)}{(10+100+100)} = 0.524$

Wrong Conclusion:

Suppose we have never even heard the word coin and have no idea what one looks like.
Let alone what probability of heads might be?

Wrong Conclusion:

Suppose we have never even heard the word coin and have no idea what one looks like.
Let alone what probability of heads might be?

We could choose $\alpha = \beta = 1$, i.e., a uniform prior distribution
(Really this would indicate our complete lack of knowledge regarding θ, this is called an uninformative prior.)

As it is seen, in this simple case, it is most intuitive to use the uniform distribution on $[0, 1]$ as a non-informative prior.
it is non-informative because it says that all possible values of θ are equally likely *a priori*.

However, a non-informative prior constructed using Jeffreys' rule is of the form

$$\begin{aligned} \pi(\theta) &\propto \frac{1}{\sqrt{(\theta(1-(\theta)}} \\ &= \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \\ &= \theta^{\frac{1}{2}-1}(1-\theta)^{\frac{1}{2}-1} \end{aligned} \quad (1)$$

However, a non-informative prior constructed using Jeffreys' rule is of the form

$$\begin{aligned}\pi(\theta) &\propto \frac{1}{\sqrt{(\theta(1-(\theta)}} \\ &= \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \\ &= \theta^{\frac{1}{2}-1}(1-\theta)^{\frac{1}{2}-1} \qquad (1)\end{aligned}$$

Jefferys' rule is motivated by an invariance argument:

In order for $\pi_\theta(\theta)$ to be non-informative, it is argued that the parameterization must not influence the choice of $\pi_\theta(\theta)$, i.e., if one re-parameterizes the problem in terms of $\tau = h(\theta)$ then the rule must pick $\pi_\tau(\tau) = |\frac{\partial\theta}{\partial\tau}|\pi_\theta(h^{-1}(\tau))$ as the prior for τ.

Notice that Jefferys' rule is to pick $\pi_\theta(\theta) \propto [I(\theta)]^{\frac{1}{2}}$, as a prior for θ.

As you may realize, Jefferys' prior for this simple problem can be quite couter-intuitive.

Notice that Jefferys' rule is to pick $\pi_\theta(\theta) \propto [I(\theta)]^{\frac{1}{2}}$, as a prior for θ.

As you may realize, Jefferys' prior for this simple problem can be quite couter-intuitive.

Under the prior in (1) it appears that some values of θ are more likely than others (see the figure)

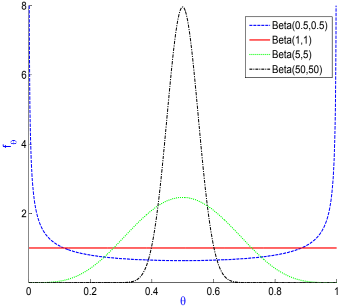

Figure: GRAPHs of Beta(0.5, 0.5), Beta(1,1), Beta(5,5) and Beta(50,50).

Therefore, intuitively, it appears that this prior is actually quite informative.

Q1: What is the goal?

Q1: What is the goal?

A1: We are going to construct a simple argument and illustrate why the uniform prior is not necessarily the most non-informative.

Q1: What is the goal?

A1: We are going to construct a simple argument and illustrate why the uniform prior is not necessarily the most non-informative.

Q2: How do the parameters α and β affect the outcome?

Q1: What is the goal?

A1: We are going to construct a simple argument and illustrate why the uniform prior is not necessarily the most non-informative.

Q2: How do the parameters α and β affect the outcome?

A2: For a partial answer, we focus on a particular subfamily of Beta-distributions with $\alpha = \beta = c$, i.e., $\theta \sim Beta(c, c)$.
Then $E(\theta) = \frac{1}{2}$ and $Var(\theta) = \frac{c^2}{4c^2(2c+1)} = \frac{1}{4(2c+1)}$.

Notice that, then the Bayes estimator is

$$\delta_\pi(Y) = \frac{Y + c}{n + 2c}$$

Notice that, then the Bayes estimator is

$$\delta_{\pi}(Y) = \frac{Y + c}{n + 2c}$$

It is clear from $\delta_{\pi}(Y)$ that the prior parameter c influences the posterior mean as if an extra $2c$ observations, equally split between zero's (tails) and one's (heads), were added to the sample.

Therefore, the larger c is the more influence the prior will have on the posterior mean.

Notice that, then the Bayes estimator is

$$\delta_\pi(Y) = \frac{Y + c}{n + 2c}$$

It is clear from $\delta_\pi(Y)$ that the prior parameter c influences the posterior mean as if an extra $2c$ observations, equally split between zero's (tails) and one's (heads), were added to the sample.

Therefore, the larger c is the more influence the prior will have on the posterior mean.

The Uniform Prior=$Beta(1,1)$, $(c = 1)$, adds two extra observations.

Jeffreys' prior= $Beta(\frac{1}{2}, \frac{1}{2})$, $(c = \frac{1}{2})$, adds one extra observation.

It is in this sense that Jeffreys' prior is actually less influential than the Uniform prior.

Q3: What Next?

Q3: What Next?

A3: Look at $Var(\theta) = \frac{1}{4(2c+1)}$ which is $\downarrow$ in c.

This also says that the larger the prior variance, the less influential the prior is, which makes intuitive sense:

Q3: What Next?

A3: Look at $Var(\theta) = \frac{1}{4(2c+1)}$ which is $\downarrow$ in c.

This also says that the larger the prior variance, the less influential the prior is, which makes intuitive sense:

A larger Prior Variance would normally indicate a relatively weak prior opinion. In view of this, two extreme cases become quite interesting:

i) $c \to +\infty$

ii) $c \to 0$??

i) If $c \to +\infty$, then $\delta_\pi(Y) = \frac{Y+c}{n+2c} \to \frac{1}{2}$, which is the same as prior mean regardless of what the observed outcome are.

In other words, our prior opinion of θ is so strong that it can not be changed by the observed outcomes.

i) If $c \to +\infty$, then $\delta_\pi(Y) = \frac{Y+c}{n+2c} \to \frac{1}{2}$, which is the same as prior mean regardless of what the observed outcome are.

In other words, our prior opinion of θ is so strong that it can not be changed by the observed outcomes.

Also, $Var(\theta) = \frac{1}{4(2c+1)} \to 0$ as $c \to +\infty$. This is, again, consistent with our intuition:

The small prior variance means that one's prior belief is heavily concentrated on the point $\theta = \frac{1}{2}$, so heavy that the observed outcomes could not alter this belief in any way!

ii) If $c \to 0$, then $\delta_\pi(Y) = \frac{Y+c}{n+2c} \to \frac{Y}{n}$, which is the same as the least influential prior in our sub-family would have been the one with $c = 0$.

ii) If $c \to 0$, then $\delta_\pi(Y) = \frac{Y+c}{n+2c} \to \frac{Y}{n}$, which is the same as the least influential prior in our sub-family would have been the one with $c = 0$.

Using such a prior, the posterior mean would have been the same as the MLE, i.e., it would have been entirely determined by the observed outcomes. But notice that *Beta*$(0, 0)$-distribution is not defined.

ii) If $c \to 0$, then $\delta_\pi(Y) = \frac{Y+c}{n+2c} \to \frac{Y}{n}$, which is the same as the least influential prior in our sub-family would have been the one with $c = 0$.

Using such a prior, the posterior mean would have been the same as the MLE, i.e., it would have been entirely determined by the observed outcomes. But notice that $Beta(0, 0)$-distribution is not defined.

To understand the behavior of this distribution, we can examine the limiting distribution as $c \to 0$, i.e.,

$$B_{0,0} = \lim_{c \to 0} Beta(c, c).$$

Theorem

The limiting distribution $B_{0,0}$ consists of two equal point masses at 0 and 1.

Numerical Example

Numerical Example

- Notice that the variance of $B_{0,0}$ is $\frac{1}{4}$.

- Notice that the variance of $B_{0,0}$ is $\frac{1}{4}$.
- Theorem says that the prior distribution $Beta(\epsilon, \epsilon)$ with arbitrary small $\epsilon > 0$ approaches two point masses at 0 and 1.

- Notice that the variance of $B_{0,0}$ is $\frac{1}{4}$.
- Theorem says that the prior distribution $Beta(\epsilon, \epsilon)$ with arbitrary small $\epsilon > 0$ approaches two point masses at 0 and 1.
- Such a prior belief, of course, seems extremely strong, since it says θ is essentially either 0 or 1.

Numerical Example

- Notice that the variance of $B_{0,0}$ is $\frac{1}{4}$.
- Theorem says that the prior distribution $Beta(\epsilon, \epsilon)$ with arbitrary small $\epsilon > 0$ approaches two point masses at 0 and 1.
- Such a prior belief, of course, seems extremely strong, since it says θ is essentially either 0 or 1.
- Intuitively, one would consider such a strong prior belief to be extremely unreasonable, but this is the prior that would yield a posterior mean as close as possible to the MLE.

- Notice that the variance of $B_{0,0}$ is $\frac{1}{4}$.
- Theorem says that the prior distribution $Beta(\epsilon, \epsilon)$ with arbitrary small $\epsilon > 0$ approaches two point masses at 0 and 1.
- Such a prior belief, of course, seems extremely strong, since it says θ is essentially either 0 or 1.
- Intuitively, one would consider such a strong prior belief to be extremely unreasonable, but this is the prior that would yield a posterior mean as close as possible to the MLE.
- In this sense, the prior $Beta(\epsilon, \epsilon)$, $\epsilon > 0$, which would otherwise appear strong, could actually be regarded as the least influential prior in this family.

NUMERICAL EXAMPLE

- Notice that the variance of $B_{0,0}$ is $\frac{1}{4}$.
- Theorem says that the prior distribution $Beta(\epsilon, \epsilon)$ with arbitrary small $\epsilon > 0$ approaches two point masses at 0 and 1.
- Such a prior belief, of course, seems extremely strong, since it says θ is essentially either 0 or 1.
- Intuitively, one would consider such a strong prior belief to be extremely unreasonable, but this is the prior that would yield a posterior mean as close as possible to the MLE.
- In this sense, the prior $Beta(\epsilon, \epsilon)$, $\epsilon > 0$, which would otherwise appear strong, could actually be regarded as the least influential prior in this family.

- Theorem states that the limiting distribution $B_{0,0}$ is $B(1, \frac{1}{2})$-distribution, which strictly speaking, is not a member of the Beta Family.

Numerical Example

- Notice that the variance of $B_{0,0}$ is $\frac{1}{4}$.
- Theorem says that the prior distribution $Beta(\epsilon, \epsilon)$ with arbitrary small $\epsilon > 0$ approaches two point masses at 0 and 1.
- Such a prior belief, of course, seems extremely strong, since it says θ is essentially either 0 or 1.
- Intuitively, one would consider such a strong prior belief to be extremely unreasonable, but this is the prior that would yield a posterior mean as close as possible to the MLE.
- In this sense, the prior $Beta(\epsilon, \epsilon)$, $\epsilon > 0$, which would otherwise appear strong, could actually be regarded as the least influential prior in this family.

- Theorem states that the limiting distribution $B_{0,0}$ is $B(1, \frac{1}{2})$-distribution, which strictly speaking, is not a member of the Beta Family.
- Moreover, if $B_{0,0}$ is actually used as a prior, then the posterior distribution is not defined unless all the observations $X_1, \ldots, X_n$ are identical.

- Notice that the variance of $B_{0,0}$ is $\frac{1}{4}$.
- Theorem says that the prior distribution $Beta(\epsilon, \epsilon)$ with arbitrary small $\epsilon > 0$ approaches two point masses at 0 and 1.
- Such a prior belief, of course, seems extremely strong, since it says θ is essentially either 0 or 1.
- Intuitively, one would consider such a strong prior belief to be extremely unreasonable, but this is the prior that would yield a posterior mean as close as possible to the MLE.
- In this sense, the prior $Beta(\epsilon, \epsilon)$, $\epsilon > 0$, which would otherwise appear strong, could actually be regarded as the least influential prior in this family.

- Theorem states that the limiting distribution $B_{0,0}$ is $B(1, \frac{1}{2})$-distribution, which strictly speaking, is not a member of the Beta Family.
- Moreover, if $B_{0,0}$ is actually used as a prior, then the posterior distribution is not defined unless all the observations $X_1, \ldots, X_n$ are identical.
- Hence $B_{0,0}$ is in itself quite an influential prior, but $Beta(\epsilon, \epsilon)$, $\epsilon > 0$, is not, although for arbitrary small $\epsilon > 0$, it encodes essentially the same prior opinion as $B_{0,0}$, whose predictive distribution puts half probability on all ones and half on all zeros.

THE LESSONS OF THIS DISCUSSION:

THE LESSONS OF THIS DISCUSSION:

- It tells us that flat priors, such as Uniform prior, are not always the same thing as non-informative.

THE LESSONS OF THIS DISCUSSION:

- It tells us that flat priors, such as Uniform prior, are not always the same thing as non-informative.
- A seemingly informative prior can actually be quite weak in that sense that it does not influence the posterior opinion very much.

THE LESSONS OF THIS DISCUSSION:

- It tells us that flat priors, such as Uniform prior, are not always the same thing as non-informative.
- A seemingly informative prior can actually be quite weak in that sense that it does not influence the posterior opinion very much.
- It is clear, in our example, that the MLE is the result of using a weak prior, whereas the most intuitive non-informative prior, the Uniform prior, is not as weak or non-informative as one would have thought.

THE LESSONS OF THIS DISCUSSION:

- It tells us that flat priors, such as Uniform prior, are not always the same thing as non-informative.
- A seemingly informative prior can actually be quite weak in that sense that it does not influence the posterior opinion very much.
- It is clear, in our example, that the MLE is the result of using a weak prior, whereas the most intuitive non-informative prior, the Uniform prior, is not as weak or non-informative as one would have thought.

THANKS