MT4531/MT5731: (Advanced) Bayesian Inference

Non-informative priors

Nicolò Margaritella

School of Mathematics and Statistics, University of St Andrews

–

University of St Andrews

Outline

Outline

1. Prior distributions
2. Non-informative priors
3. Jeffreys' priors
4. Example

On prior distributions

- How should we assign the prior $p(\theta)$?
- There is no such thing as the *correct choice* of $p(\theta)$ for a given problem.
- The choice of prior lies entirely with the statistician and the information and experience they have.
- The investigator should be able to persuade their audience that their choice of prior is sensible.
- In the next lecture we will see how to assign a prior when expert or strong prior information is available.
- In this set of slides, we will discuss assigning a prior when there is no prior information on the parameter or quantity of interest.

On prior distributions

- How should we assign the prior $p(\theta)$?
- There is no such thing as the *correct choice* of $p(\theta)$ for a given problem.
- The choice of prior lies entirely with the statistician and the information and experience they have.
- The investigator should be able to persuade their audience that their choice of prior is sensible.
- In the next lecture we will see how to assign a prior when expert or strong prior information is available.
- In this set of slides, we will discuss assigning a prior when there is no prior information on the parameter or quantity of interest.

On prior distributions

- How should we assign the prior $p(\theta)$?
- There is no such thing as the *correct choice* of $p(\theta)$ for a given problem.
- The choice of prior lies entirely with the statistician and the information and experience they have.
- The investigator should be able to persuade their audience that their choice of prior is sensible.
- In the next lecture we will see how to assign a prior when expert or strong prior information is available.
- In this set of slides, we will discuss assigning a prior when there is no prior information on the parameter or quantity of interest.

On prior distributions

- How should we assign the prior $p(\theta)$?
- There is no such thing as the *correct choice* of $p(\theta)$ for a given problem.
- The choice of prior lies entirely with the statistician and the information and experience they have.
- The investigator should be able to persuade their audience that their choice of prior is sensible.
- In the next lecture we will see how to assign a prior when expert or strong prior information is available.
- In this set of slides, we will discuss assigning a prior when there is no prior information on the parameter or quantity of interest.

On prior distributions

- How should we assign the prior $p(\theta)$?
- There is no such thing as the *correct choice* of $p(\theta)$ for a given problem.
- The choice of prior lies entirely with the statistician and the information and experience they have.
- The investigator should be able to persuade their audience that their choice of prior is sensible.
- In the next lecture we will see how to assign a prior when expert or strong prior information is available.
- In this set of slides, we will discuss assigning a prior when there is no prior information on the parameter or quantity of interest.

On prior distributions

- How should we assign the prior $p(\theta)$?
- There is no such thing as the *correct choice* of $p(\theta)$ for a given problem.
- The choice of prior lies entirely with the statistician and the information and experience they have.
- The investigator should be able to persuade their audience that their choice of prior is sensible.
- In the next lecture we will see how to assign a prior when expert or strong prior information is available.
- In this set of slides, we will discuss assigning a prior when there is no prior information on the parameter or quantity of interest.

Outline

Non-informative priors (1)

- In practice, without prior information, it is often adequate to assign a very flat prior, or one with a very large variance.
- For example, for the prior mean of a Normal distribution, we may assign a Normal prior with variance of order 10^5 or 10^6.
- We will see later how such choices can affect the comparison of different statistical models.
- If the parameter has a finite range, then we could assign a uniform prior, such as the $U(0, 1)$ for the probability of success in a Binomial experiment.
- However, there can be problems with this 'natural' choice. (Bayes himself suggested such a Uniform prior.)
- The problems arise if one is interested in making inferences for transformations of the parameter of interest. Specifically...

Non-informative priors (1)

- In practice, without prior information, it is often adequate to assign a very flat prior, or one with a very large variance.
- For example, for the prior mean of a Normal distribution, we may assign a Normal prior with variance of order 10^5 or 10^6.
- We will see later how such choices can affect the comparison of different statistical models.
- If the parameter has a finite range, then we could assign a uniform prior, such as the $U(0, 1)$ for the probability of success in a Binomial experiment.
- However, there can be problems with this 'natural' choice. (Bayes himself suggested such a Uniform prior.)
- The problems arise if one is interested in making inferences for transformations of the parameter of interest. Specifically...

Non-informative priors (1)

- In practice, without prior information, it is often adequate to assign a very flat prior, or one with a very large variance.
- For example, for the prior mean of a Normal distribution, we may assign a Normal prior with variance of order 10^5 or 10^6.
- We will see later how such choices can affect the comparison of different statistical models.
- If the parameter has a finite range, then we could assign a uniform prior, such as the $U(0,1)$ for the probability of success in a Binomial experiment.
- However, there can be problems with this 'natural' choice. (Bayes himself suggested such a Uniform prior.)
- The problems arise if one is interested in making inferences for transformations of the parameter of interest. Specifically...

Non-informative priors (1)

- In practice, without prior information, it is often adequate to assign a very flat prior, or one with a very large variance.
- For example, for the prior mean of a Normal distribution, we may assign a Normal prior with variance of order 10^5 or 10^6.
- We will see later how such choices can affect the comparison of different statistical models.
- If the parameter has a finite range, then we could assign a uniform prior, such as the $U(0, 1)$ for the probability of success in a Binomial experiment.
- However, there can be problems with this 'natural' choice. (Bayes himself suggested such a Uniform prior.)
- The problems arise if one is interested in making inferences for transformations of the parameter of interest. Specifically...

Non-informative priors (1)

- In practice, without prior information, it is often adequate to assign a very flat prior, or one with a very large variance.
- For example, for the prior mean of a Normal distribution, we may assign a Normal prior with variance of order 10^5 or 10^6.
- We will see later how such choices can affect the comparison of different statistical models.
- If the parameter has a finite range, then we could assign a uniform prior, such as the $U(0, 1)$ for the probability of success in a Binomial experiment.
- However, there can be problems with this 'natural' choice. (Bayes himself suggested such a Uniform prior.)
- The problems arise if one is interested in making inferences for transformations of the parameter of interest. Specifically...

Non-informative priors (1)

- In practice, without prior information, it is often adequate to assign a very flat prior, or one with a very large variance.
- For example, for the prior mean of a Normal distribution, we may assign a Normal prior with variance of order 10^5 or 10^6.
- We will see later how such choices can affect the comparison of different statistical models.
- If the parameter has a finite range, then we could assign a uniform prior, such as the $U(0, 1)$ for the probability of success in a Binomial experiment.
- However, there can be problems with this 'natural' choice. (Bayes himself suggested such a Uniform prior.)
- The problems arise if one is interested in making inferences for transformations of the parameter of interest. Specifically...

Non-informative priors (2)

- Suppose that we have no information about a parameter θ in $[0, 1]$.
- We place a Uniform prior, $p(\theta) = 1$, $\theta \in [0, 1]$.
- Then the corresponding prior on $\phi = h(\theta) = \theta^2$ is non-Uniform. Note that according to the change of variable rule,
-
$$p_\phi(\phi) = p_\theta\left(h^{-1}(\phi)\right)\left|\frac{dh^{-1}(\phi)}{d\phi}\right| = p_\theta\left(\sqrt{\phi}\right)\left|\frac{d\sqrt{\phi}}{d\phi}\right|$$
$$= \left|\frac{d\sqrt{\phi}}{d\phi}\right| = \frac{1}{2\sqrt{\phi}}, \qquad \phi \in [0, 1].$$

Non-informative priors (2)

- Suppose that we have no information about a parameter θ in $[0,1]$.
- We place a Uniform prior, $p(\theta) = 1$, $\theta \in [0,1]$.
- Then the corresponding prior on $\phi = h(\theta) = \theta^2$ is non-Uniform. Note that according to the change of variable rule,
-
$$p_\phi(\phi) = p_\theta\left(h^{-1}(\phi)\right)\left|\frac{dh^{-1}(\phi)}{d\phi}\right| = p_\theta\left(\sqrt{\phi}\right)\left|\frac{d\sqrt{\phi}}{d\phi}\right|$$
$$= \left|\frac{d\sqrt{\phi}}{d\phi}\right| = \frac{1}{2\sqrt{\phi}}, \qquad \phi \in [0,1].$$

Non-informative priors (2)

- Suppose that we have no information about a parameter θ in $[0,1]$.
- We place a Uniform prior, $p(\theta) = 1$, $\theta \in [0,1]$.
- Then the corresponding prior on $\phi = h(\theta) = \theta^2$ is non-Uniform. Note that according to the change of variable rule,
-
$$p_\phi(\phi) = p_\theta\left(h^{-1}(\phi)\right)\left|\frac{dh^{-1}(\phi)}{d\phi}\right| = p_\theta\left(\sqrt{\phi}\right)\left|\frac{d\sqrt{\phi}}{d\phi}\right|$$
$$= \left|\frac{d\sqrt{\phi}}{d\phi}\right| = \frac{1}{2\sqrt{\phi}}, \qquad \phi \in [0,1].$$

Non-informative priors (2)

- Suppose that we have no information about a parameter θ in $[0, 1]$.
- We place a Uniform prior, $p(\theta) = 1$, $\theta \in [0, 1]$.
- Then the corresponding prior on $\phi = h(\theta) = \theta^2$ is non-Uniform. Note that according to the change of variable rule,
- $$p_\phi(\phi) = p_\theta\left(h^{-1}(\phi)\right)\left|\frac{dh^{-1}(\phi)}{d\phi}\right| = p_\theta\left(\sqrt{\phi}\right)\left|\frac{d\sqrt{\phi}}{d\phi}\right|$$
  $$= \left|\frac{d\sqrt{\phi}}{d\phi}\right| = \frac{1}{2\sqrt{\phi}}, \qquad \phi \in [0, 1].$$

Non-informative priors (3)

- However, one may expect that ignorance about the value of θ would imply ignorance about $\phi = \theta^2$.
- More generally, it might be beneficial to be able to define a non-informative prior, $p(\theta)$, so that the prior for $\phi = h(\theta)$ is non-informative for ϕ **in the same manner** in which the prior for θ is not informative for θ.

Non-informative priors (3)

- However, one may expect that ignorance about the value of θ would imply ignorance about $\phi = \theta^2$.
- More generally, it might be beneficial to be able to define a non-informative prior, $p(\theta)$, so that the prior for $\phi = h(\theta)$ is non-informative for ϕ **in the same manner** in which the prior for θ is not informative for θ.

Outline

Jeffreys' priors (1)

- Jeffreys suggested a prior based on an invariance rule for one-to-one (bijective) transformations.
- The idea is to derive a prior for θ so that for any $\phi = h(\theta)$ (h: bijective function) computing the prior for ϕ produces a prior that is uninformative for ϕ in exactly the same manner as the prior for θ is uninformative for θ.

Jeffreys' priors (1)

- Jeffreys suggested a prior based on an invariance rule for one-to-one (bijective) transformations.
- The idea is to derive a prior for θ so that for any $\phi = h(\theta)$ (h: bijective function) computing the prior for ϕ produces a prior that is uninformative for ϕ in exactly the same manner as the prior for θ is uninformative for θ.

Jeffreys' priors (2)

- Jeffreys' prior is given by,

$$p(\theta) \propto \sqrt{I(\theta|\boldsymbol{x})},$$

- where $I(\theta|\boldsymbol{x})$ is the Fisher Information

$$I(\theta|\boldsymbol{x}) = \mathbb{E}_x \left(\frac{d \log f(\boldsymbol{x}|\theta)}{d\theta} \right)^2 .$$

- Essentially, Fisher information is an indicator of the amount of information supplied by the model and observations about an unknown parameter θ.

Jeffreys' priors (2)

- Jeffreys' prior is given by,

$$p(\theta) \propto \sqrt{I(\theta|\boldsymbol{x})},$$

- where $I(\theta|\boldsymbol{x})$ is the Fisher Information

$$I(\theta|\boldsymbol{x}) = \mathbb{E}_x \left(\frac{d \log f(\boldsymbol{x}|\theta)}{d\theta} \right)^2 .$$

- Essentially, Fisher information is an indicator of the amount of information supplied by the model and observations about an unknown parameter θ.

Jeffreys' priors (2)

- Jeffreys' prior is given by,

$$p(\theta) \propto \sqrt{I(\theta|\boldsymbol{x})},$$

- where $I(\theta|\boldsymbol{x})$ is the Fisher Information

$$I(\theta|\boldsymbol{x}) = \mathbb{E}_x \left(\frac{d \log f(\boldsymbol{x}|\theta)}{d\theta} \right)^2 .$$

- Essentially, Fisher information is an indicator of the amount of information supplied by the model and observations about an unknown parameter θ.

Jeffreys' priors (3)

- Looking Fisher information as a function of θ, at the regions of the parameter space where it obtains high values, the amount of information brought by the data is high.
- If we use this function/curve as a prior for θ, we favour the values of θ for which $I(\theta|\boldsymbol{x})$ is large, i.e., we minimize the influence of the prior.
- Under certain regularity conditions, Fisher's information can also be expressed in the following form (see proof in pdf document on Moodle),

$$I(\theta|\boldsymbol{x}) = -\mathbb{E}_x\left[\frac{d^2 \log f(\boldsymbol{x}|\theta)}{d\theta^2}\right].$$

- Fisher's information is generally more easily calculated using this latter expression.

Jeffreys' priors (3)

- Looking Fisher information as a function of θ, at the regions of the parameter space where it obtains high values, the amount of information brought by the data is high.
- If we use this function/curve as a prior for θ, we favour the values of θ for which $I(\theta|\boldsymbol{x})$ is large, i.e., we minimize the influence of the prior.
- Under certain regularity conditions, Fisher's information can also be expressed in the following form (see proof in pdf document on Moodle),

$$I(\theta|\boldsymbol{x}) = -\mathbb{E}_x\left[\frac{d^2 \log f(\boldsymbol{x}|\theta)}{d\theta^2}\right].$$

- Fisher's information is generally more easily calculated using this latter expression.

Jeffreys' priors (3)

- Looking Fisher information as a function of θ, at the regions of the parameter space where it obtains high values, the amount of information brought by the data is high.
- If we use this function/curve as a prior for θ, we favour the values of θ for which $I(\theta|\boldsymbol{x})$ is large, i.e., we minimize the influence of the prior.
- Under certain regularity conditions, Fisher's information can also be expressed in the following form (see proof in pdf document on Moodle),

$$I(\theta|\boldsymbol{x}) = -\mathbb{E}_x\left[\frac{d^2 \log f(\boldsymbol{x}|\theta)}{d\theta^2}\right].$$

- Fisher's information is generally more easily calculated using this latter expression.

Jeffreys' priors (3)

- Looking Fisher information as a function of θ, at the regions of the parameter space where it obtains high values, the amount of information brought by the data is high.
- If we use this function/curve as a prior for θ, we favour the values of θ for which $I(\theta|\boldsymbol{x})$ is large, i.e., we minimize the influence of the prior.
- Under certain regularity conditions, Fisher's information can also be expressed in the following form (see proof in pdf document on Moodle),

$$I(\theta|\boldsymbol{x}) = -\mathbb{E}_x\left[\frac{d^2 \log f(\boldsymbol{x}|\theta)}{d\theta^2}\right].$$

- Fisher's information is generally more easily calculated using this latter expression.

Jeffreys' priors (4)

- Suppose ϕ, is a bijective transformation of θ, so that, $\phi = h(\theta)$. Then (**this is the bit that shows that the two priors are non-informative in the same manner!**)

$$p(\theta) \propto \sqrt{I(\theta|\boldsymbol{x})}$$

then,

$$p(\phi) \propto \sqrt{I(\phi|\boldsymbol{x})}.$$

Proof:

- **Proof:** Remember that for a transformed X, so that $Y = h(X)$, where h is bijective,

$$f_Y(y) = f_X(x)|\frac{dx}{dy}| = f_X(h^{-1}(y))\left|\frac{dh^{-1}(y)}{dy}\right|.$$

Remember also that,

$$\frac{d}{dx}f(h(x)) = \frac{df(y)}{dy}\frac{dy}{dx}.$$

Notice now that Fisher's information is

$$I(\theta|\mathbf{x}) = \mathbb{E}\left(\frac{d\log f(\mathbf{x}|\theta)}{d\theta}\right)^2 = \mathbb{E}\left(\frac{d\log f(\mathbf{x}|\theta = h^{-1}(\phi))}{d\theta}\right)^2$$

$$= \mathbb{E}\left(\frac{d\log f(\mathbf{x}|\phi)}{d\phi} \times \frac{d\phi}{d\theta}\right)^2, \quad \text{as } h \text{ is bijective}$$

$$= \left|\frac{d\phi}{d\theta}\right|^2 \mathbb{E}\left(\frac{d\log f(\mathbf{x}|\phi)}{d\phi}\right)^2 = I(\phi|\mathbf{x})\left|\frac{d\phi}{d\theta}\right|^2.$$

Assume that the prior on θ is specified as:

$$p(\theta) \propto \sqrt{I(\theta|\mathbf{x})}$$

Using the transformation of variable rule, we have that,

$$p(\phi) \propto \sqrt{I(\phi|\mathbf{x})\left|\frac{d\phi}{d\theta}\right|^2} \times \left|\frac{d\theta}{d\phi}\right| = \sqrt{I(\phi|\mathbf{x})},$$

as required.

Useful notes (1)

- For n independent observations $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from the same distribution f, Fisher's information is given by,

$$I(\theta|\boldsymbol{x}) = nI(\theta|x),$$

where $X \sim f$.

Useful notes (2)

- Jeffreys' prior can be extended to the case where there are several unknown parameters.
- Then, Fisher's information is defined as the matrix, with the element in row i and column j given by,

$$(I(\theta|\boldsymbol{x}))_{i,j} = \mathbb{E}_{\boldsymbol{x}} \left(\frac{d^2 \log f(\boldsymbol{x}|\theta)}{d\theta_i d\theta_j} \right).$$

- Then, the prior is specified as,

$$p(\theta) \propto \sqrt{\det I(\theta|\boldsymbol{x})}.$$

Useful notes (2)

- Jeffreys' prior can be extended to the case where there are several unknown parameters.
- Then, Fisher's information is defined as the matrix, with the element in row i and column j given by,

$$(I(\boldsymbol{\theta}|\boldsymbol{x}))_{i,j} = \mathbb{E}_{\boldsymbol{x}}\left(\frac{d^2 \log f(\boldsymbol{x}|\boldsymbol{\theta})}{d\theta_i d\theta_j}\right).$$

- Then, the prior is specified as,

$$p(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta}|\boldsymbol{x})}.$$

Useful notes (2)

- Jeffreys' prior can be extended to the case where there are several unknown parameters.
- Then, Fisher's information is defined as the matrix, with the element in row i and column j given by,

$$(I(\boldsymbol{\theta}|\boldsymbol{x}))_{i,j} = \mathbb{E}_{\boldsymbol{x}}\left(\frac{d^2 \log f(\boldsymbol{x}|\boldsymbol{\theta})}{d\theta_i d\theta_j}\right).$$

- Then, the prior is specified as,

$$p(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta}|\boldsymbol{x})}.$$

Useful notes (3)

- The most important objection to Jeffrey's prior is that it does not satisfy the Likelihood Principle. (That is, probability models that lead to the same likelihood for the data should give the same inferences for θ.)
- Depending on the design of the experiment or the stopping rule, Jeffreys' prior may be different even if the likelihood for the observed data is the same, leading to different posterior distributions.
- For example, there are different Jeffrey's priors for Binomial and Negative Binomial experiments (see Example in this lecture and Q4 in Tutorial 2), and posterior inference using Jeffrey's prior in each case will violate the Likelihood Principle.

Useful notes (3)

- The most important objection to Jeffrey's prior is that it does not satisfy the Likelihood Principle. (That is, probability models that lead to the same likelihood for the data should give the same inferences for θ.)
- Depending on the design of the experiment or the stopping rule, Jeffreys' prior may be different even if the likelihood for the observed data is the same, leading to different posterior distributions.
- For example, there are different Jeffrey's priors for Binomial and Negative Binomial experiments (see Example in this lecture and Q4 in Tutorial 2), and posterior inference using Jeffrey's prior in each case will violate the Likelihood Principle.

Useful notes (3)

- The most important objection to Jeffrey's prior is that it does not satisfy the Likelihood Principle. (That is, probability models that lead to the same likelihood for the data should give the same inferences for θ.)
- Depending on the design of the experiment or the stopping rule, Jeffreys' prior may be different even if the likelihood for the observed data is the same, leading to different posterior distributions.
- For example, there are different Jeffrey's priors for Binomial and Negative Binomial experiments (see Example in this lecture and Q4 in Tutorial 2), and posterior inference using Jeffrey's prior in each case will violate the Likelihood Principle.

Outline

Example:

- Let X denote the number of defective items in a batch of n fudge doughnuts
- Each doughnut is defective with probability θ, independently of each other, given θ.
- Jeffreys' prior for θ is derived as follows...

Example:

- Let X denote the number of defective items in a batch of n fudge doughnuts
- Each doughnut is defective with probability θ, independently of each other, given θ.
- Jeffreys' prior for θ is derived as follows...

Example:

- Let X denote the number of defective items in a batch of n fudge doughnuts
- Each doughnut is defective with probability θ, independently of each other, given θ.
- Jeffreys' prior for θ is derived as follows...

Example: Derivation of Jeffreys' prior

- $X \sim Bin(n, \theta)$ and, $f(x|\theta) \propto \theta^x (1-\theta)^{n-x}$, so that,

$$\log f(x|\theta) = x \log \theta + (n-x)\log(1-\theta) + C,$$

so that,

$$\frac{d^2 \log f(x|\theta)}{d\theta^2} = -\frac{x}{\theta^2} - \frac{(n-x)}{(1-\theta)^2}.$$

Then,

$$\begin{aligned} I(\theta|x) &= -\mathbb{E}\left(\frac{d^2 \log f(x|\theta)}{d\theta^2}\right) \\ &= \mathbb{E}\left(\frac{x}{\theta^2}\right) + \mathbb{E}\left(\frac{(n-x)}{(1-\theta)^2}\right) \\ &= \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} \\ &= \frac{n}{\theta(1-\theta)}. \end{aligned}$$

So that, Jeffreys' prior for the probability parameter θ is,

$$p(\theta) \propto \sqrt{I(\theta|x)} \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}.$$

In other words, $\theta \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right)$.

Finally...

- Note that vague prior distributions do not have to be proper distributions, in the sense that they integrate to one.
- As long as the posterior distribution is proper, then it is acceptable to use an improper vague prior distribution.
- **Task**: Read Section 1.4.1 of the lecture notes and complete the relative exercise.

Finally...

- Note that vague prior distributions do not have to be proper distributions, in the sense that they integrate to one.
- As long as the posterior distribution is proper, then it is acceptable to use an improper vague prior distribution.
- **Task**: Read Section 1.4.1 of the lecture notes and complete the relative exercise.

Finally...

- Note that vague prior distributions do not have to be proper distributions, in the sense that they integrate to one.
- As long as the posterior distribution is proper, then it is acceptable to use an improper vague prior distribution.
- **Task**: Read Section 1.4.1 of the lecture notes and complete the relative exercise.