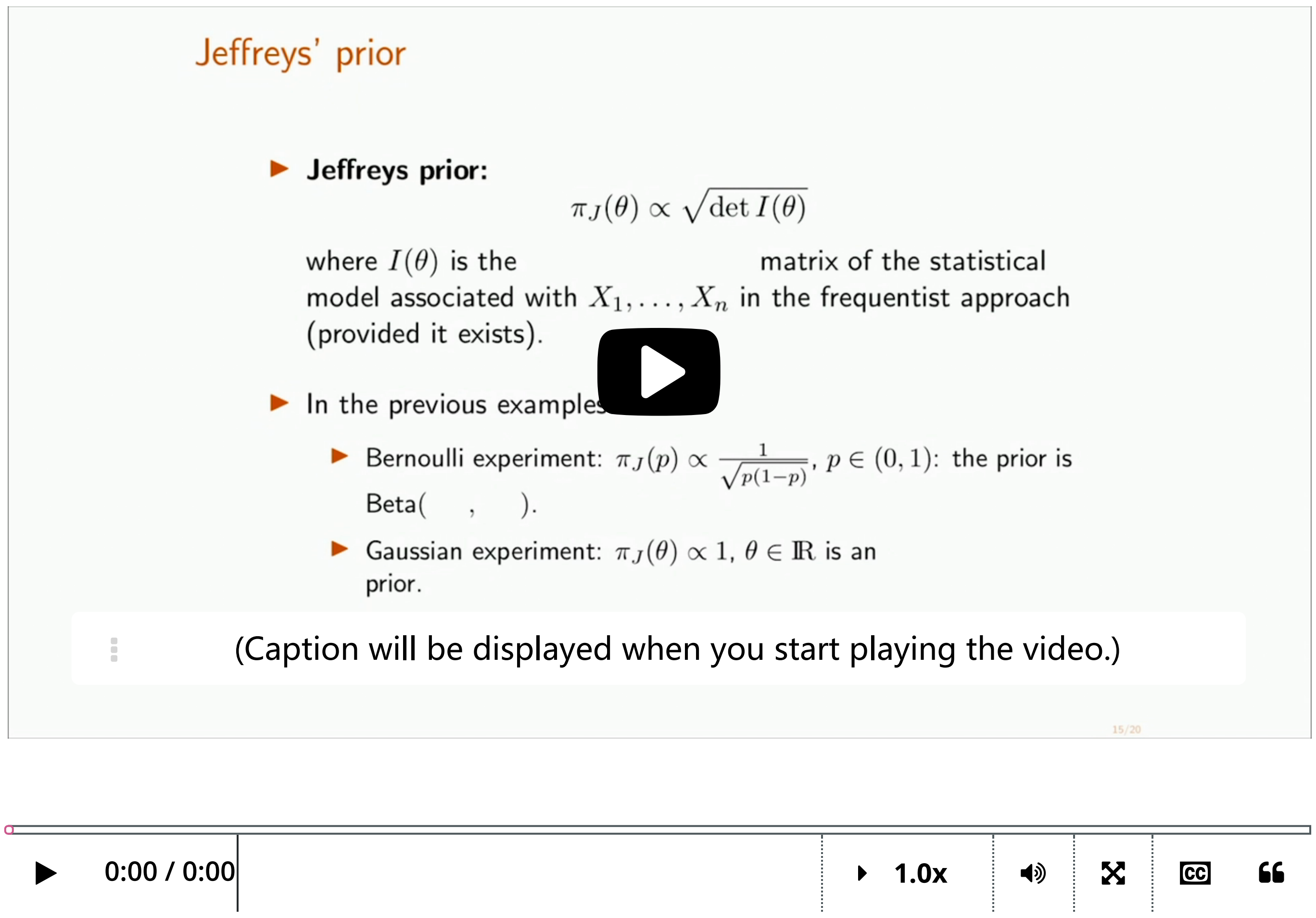
6. Jeffreys Prior I: Definition

Jeffreys Prior I: Definition



Start of transcript. Skip to the end.

To conclude, I'd like to talk about a different prior, which

is the Jeffreys prior.

So this is a prior that's also part of the non-informative,

or sometimes called "objective" priors.

And the idea is to say, can I build

a prior which is invariant under rescaling of the parameter.

So what do I mean by this?

Well, if I think about the uniform prior on p,

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Jeffreys Prior

Jeffreys Prior is an attempt to incorporate frequentist ideas of likelihood in the Bayesian framework, as well as an example of a *non-informative prior*. This prior depends on the statistical model used for the observation data and the likelihood function. Mathematically, it is the prior $\pi_J(\theta)$ that satisfies

$$\pi_J(\theta) \propto \sqrt{\det I(\theta)},$$

where $I(\theta)$ is the **Fisher Information matrix** of the statistical model associated with $X_1, \ldots, X_n$ in the frequentist approach, provided that it exists.

In the one-variable case, Jeffreys prior reduces to

$$\pi_J(\theta) \propto \sqrt{I(\theta)}.$$

The Fisher information matrix $I(\theta)$ here is treated as a *linear transformation* matrix which maps one coordinate space to another (the logic behind such a mapping would be explained soon). In linear transformation terms, taking the determinant represents the ratio of volumes of corresponding spaces between coordinate system, which explains the intuition behind the use of $\det I(\theta)$.

Recitation Note : Two newly recorded recitations on Jeffreys prior are now available in the tabs after homework 9 and after midterm 2. The concepts discussed may be helpful for the upcoming lecture exercises.

Fisher Information and MLE Interpretation

0/1 point (graded)

Let our parameter of interest be θ. As computing Jeffreys prior makes use of the Fisher information $I(\theta)$, it is somehow related to the frequentist MLE approach (which has variance $I(\theta)^{-1}$). This yields interpretations of Jeffreys prior in terms of frequentist notions of estimation, uncertainty, and information.

For each statement, fill in the blank with the appropriate choice (more / less), then choose the option that represents your answers in order.

1. The Jeffreys prior gives more weight to values of θ whose MLE estimate has ____ uncertainty.
2. As a result, the Jeffreys prior yields more weight to values of θ where the data has ____ information towards deciding the parameter.
3. The Fisher information can be taken as a proxy for how much, at a particular parameter value θ, would equivalent shifts to the parameter influence the data. Thus, Jeffreys prior gives more weight to regions where the potential outcomes are ____ sensitive to slight changes in θ.

- more, more, more
- more, more, less
- more, less, more
- more, less, less
- less, more, more ✔
- less, more, less
- less, less, more
- less, less, less

Solution:

- The weight given to a parameter value θ is the square root of its Fisher information $I(\theta)$, so more weight is given when $I(\theta)$ is high. The Fisher information is also the reciprocal of the MLE variance, so when the Fisher information is high, the MLE variance is low and thus the MLE has less uncertainty. Combining, we get that the Jeffreys prior gives more weight to values of θ whose MLE estimate has **less** uncertainty.
- Continuing from the above reasoning, when the MLE estimate has less uncertainty and we are able to estimate it more precisely. This corresponds to the data giving **more** information about the parameter when the Jeffreys prior yields larger values.

方差越小，信噪比越高，所有信息就越多

- Again, Jeffreys prior gives more weight to regions with high Fisher informations. By the given interpretation for the Fisher information, this means that at these areas, a small change to θ will influence the data relatively more, or in other words, potential outcomes are **more** sensitive to slight changes in θ.

前者是给了更大权重的部分，更影响theta。
给了更大权重的部分，本来方方差就小。
因为方差小，所以有一些改变，整体就有变化了。

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Area Interpretation of Jeffreys Prior

0/1 point (graded)

We start with a fixed one-parameter statistical model where we use the MLE as our estimate, and consider the case where the number of samples n gets large. For each potential estimate θ, we construct using the asymptotic MLE variance the 95% confidence interval $X(\theta)$ centered at θ. Then, we consider the area over the interval $X(\theta)$ under the curve based on the Jeffreys prior. This area is is ____.

- the same regardless of θ. ✔
- significantly larger for values of θ where $I(\theta)$ is large. ✖
- significantly larger for values of θ where $I(\theta)$ is small.

Solution:

We work in the asymptotic view, where $n \to \infty$. Regardless of n, the Jeffreys prior is fixed, so as the interval $X(\theta)$ gets small (which always happens as $n \to \infty$ by the , we can assume that the weight of Jeffreys prior over the interval is constant at $\sqrt{I(\theta)}$

Jeffreys prior is constant regardless of the value of n, the height can be treated as similar or essentially the same over the small interval.

The width of the interval is approximately $2 \cdot 1.96\sqrt{\frac{I(\theta)^{-1}}{n}}$ as the MLE has asymptotic variance $I(\theta)^{-1}$. Hence, the area under the interval based on Jeffreys prior pdf is $(\sqrt{I(\theta)})\,(3.92\sqrt{\frac{I(\theta)^{-1}}{n}}) = \frac{3.92}{\sqrt{n}}$, which is **the same regardless of θ**.

prior * likelihood

Submit You have used 2 of 2 attempts

Jeffrey's prior is a bad idea.

discussion posted 8 days ago by SergK (Community TA)

After seeing the Jeffrey's prior for Bernoulli trials

$$\pi_J(p) \propto \frac{1}{\sqrt{p(1-p)}}$$

I can't stop thinking that the very idea of Jeffrey's prior is bad because in this case it assigns most weight on least likely values of an unknown parameter. Though mathematically possible, Jeffrey's priors should be avoided in practice. I'd better use

$$\pi(p) \propto p(1-p)$$

as a prior for Bernoulli trials.

Show Discussion

ptressel

about 23 hours ago

+

...

I was about to write a post about the machine learning procedure called "boosting", in which one repeatedly re-runs the training process, and on each pass, assigns more weight to the samples that were incorrectly classified on the previous round. Here, "more weight" means, artificially pretend we got more samples like those. That's done to make the classifier put more effort into classifying those "hard to classify" samples. I was thinking, from the verbal description in lecture, that the Jeffreys prior was doing the opposite, preferring the "easy" cases. Then I saw the Beta(1/2, 1/2) prior for the Bernoulli trials, that blows up at 0 and at 1. But maybe near 0 and near 1 *are* the easy cases for Bernoulli trials -- those are the low entropy cases. In any case, I wasn't going to make a point about the Jeffreys prior being bad, or weird -- was just going to mention this as a curiosity -- ha ha, it's doing the *opposite* of what we do in ML. But now I'm not sure, and I'd have to go off in a corner and glare at it. And there's no time for that now...