Jeffreys' Prior and Invariance

Theorem 3.1 (Jeffreys' prior invariance) + a worked example

(Lecture slides)

Motivation: "Objective" priors and re-parameterisation

Goal

Choose a prior that does *not* depend on how we parameterise the model.

Motivation: "Objective" priors and re-parameterisation

Goal

Choose a prior that does *not* depend on how we parameterise the model.

- We can describe the same statistical model using different parameters:

$$\theta \in \Theta \quad \longleftrightarrow \quad \psi = g(\theta) \in \Psi \quad \text{(one-to-one transform)}.$$

Motivation: "Objective" priors and re-parameterisation

Goal

Choose a prior that does *not* depend on how we parameterise the model.

- We can describe the same statistical model using different parameters:

$$\theta \in \Theta \quad \longleftrightarrow \quad \psi = g(\theta) \in \Psi \quad \text{(one-to-one transform).}$$

- Desideratum: prior beliefs should not change just because we re-label the parameter.

Motivation: "Objective" priors and re-parameterisation

Goal

Choose a prior that does *not* depend on how we parameterise the model.

- We can describe the same statistical model using different parameters:

$$\theta \in \Theta \quad \longleftrightarrow \quad \psi = g(\theta) \in \Psi \quad \text{(one-to-one transform)}.$$

- Desideratum: prior beliefs should not change just because we re-label the parameter.
- Jeffreys' idea: use the **Fisher information** as a canonical notion of "how sensitive" the likelihood is to changes in the parameter.

Change of variables for priors (recall)

Let $\psi = g(\theta)$ be a one-to-one differentiable transformation with inverse $\theta = g^{-1}(\psi)$.

Change of variables for priors (recall)

Let $\psi = g(\theta)$ be a one-to-one differentiable transformation with inverse $\theta = g^{-1}(\psi)$.

Density transformation rule

If $\pi_\theta(\theta)$ is a density for θ, then the induced density for ψ is

$$\pi_\psi(\psi) = \pi_\theta(\theta(\psi)) \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right| .$$

Change of variables for priors (recall)

Let $\psi = g(\theta)$ be a one-to-one differentiable transformation with inverse $\theta = g^{-1}(\psi)$.

Density transformation rule

If $\pi_\theta(\theta)$ is a density for θ, then the induced density for ψ is

$$\pi_\psi(\psi) = \pi_\theta(\theta(\psi)) \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right| .$$

- This is the standard Jacobian rule for transforming random variables.
- We'll use it to formalise what "invariance" means for priors.

Fisher information (definition)

Suppose we observe data Y with likelihood $\pi(y \mid \theta)$.

Fisher information (definition)

Suppose we observe data Y with likelihood $\pi(y \mid \theta)$.

Definition (Fisher information)

$$I(\theta) \; := \; -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log \pi(Y \mid \theta) \,\middle|\, \theta \right].$$

Fisher information (definition)

Suppose we observe data Y with likelihood $\pi(y \mid \theta)$.

Definition (Fisher information)

$$I(\theta) \;:=\; -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log \pi(Y \mid \theta) \,\middle|\, \theta \right].$$

- The expectation is taken w.r.t. $Y \sim \pi(\cdot \mid \theta)$.
- Intuition: large $I(\theta)$ means the log-likelihood curves sharply near θ.

Theorem 3.1: Jeffreys' prior (invariance theorem)

Theorem 3.1 (Jeffreys' prior invariance)

For a scalar parameter θ, define the prior

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

Then for any one-to-one differentiable re-parameterisation $\psi = g(\theta)$, the induced prior on ψ satisfies

$$\pi_J(\psi) \;\propto\; \sqrt{I(\psi)},$$

i.e. the Jeffreys prior is **invariant** under re-parameterisation.

Theorem 3.1: Jeffreys' prior (invariance theorem)

Theorem 3.1 (Jeffreys' prior invariance)

For a scalar parameter θ, define the prior

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

Then for any one-to-one differentiable re-parameterisation $\psi = g(\theta)$, the induced prior on ψ satisfies

$$\pi_J(\psi) \;\propto\; \sqrt{I(\psi)},$$

i.e. the Jeffreys prior is **invariant** under re-parameterisation.

- "Same prior information" in any coordinate system.
- This is why Jeffreys' prior is often called an **objective** (or parameterisation-invariant) prior.

Key lemma: how Fisher information transforms

Let $\psi = g(\theta)$ be one-to-one differentiable.

Key lemma: how Fisher information transforms

Let $\psi = g(\theta)$ be one-to-one differentiable.

Claim

The Fisher information transforms as

$$I(\psi) = I(\theta)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)^2.$$

Key lemma: how Fisher information transforms

Let $\psi = g(\theta)$ be one-to-one differentiable.

Claim

The Fisher information transforms as

$$I(\psi) = I(\theta)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)^2.$$

If this holds, then taking square roots gives

$$\sqrt{I(\psi)} = \sqrt{I(\theta)}\left|\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right|,$$

which will match the Jacobian rule for densities and prove invariance.

Proof sketch (Step 1): chain rule for the score

Start from the log-likelihood as a function of ψ:

$$\ell(\psi) := \log \pi(Y \mid \psi) = \log \pi(Y \mid \theta(\psi)).$$

Proof sketch (Step 1): chain rule for the score

Start from the log-likelihood as a function of ψ:

$$\ell(\psi) := \log \pi(Y \mid \psi) = \log \pi(Y \mid \theta(\psi)).$$

Chain rule

$$\frac{\partial}{\partial \psi}\ell(\psi) = \frac{\partial}{\partial \theta} \log \pi(Y \mid \theta) \cdot \frac{\mathrm{d}\theta}{\mathrm{d}\psi}.$$

Proof sketch (Step 1): chain rule for the score

Start from the log-likelihood as a function of ψ:

$$\ell(\psi) := \log \pi(Y \mid \psi) = \log \pi(Y \mid \theta(\psi)).$$

Chain rule

$$\frac{\partial}{\partial \psi}\ell(\psi) = \frac{\partial}{\partial \theta}\log \pi(Y \mid \theta) \cdot \frac{\mathrm{d}\theta}{\mathrm{d}\psi}.$$

Here $\frac{\partial}{\partial \theta}\log \pi(Y \mid \theta)$ is the **score function**.

Proof sketch (Step 2): differentiate again (product rule)

Differentiate once more w.r.t. ψ:

Proof sketch (Step 2): differentiate again (product rule)

Differentiate once more w.r.t. ψ:

$$\frac{\partial^2}{\partial \psi^2}\ell(\psi) = \frac{\partial}{\partial \psi}\left(\frac{\partial}{\partial \theta}\log \pi(Y \mid \theta) \cdot \frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right).$$

Proof sketch (Step 2): differentiate again (product rule)

Differentiate once more w.r.t. ψ:

$$\frac{\partial^2}{\partial \psi^2}\ell(\psi) = \frac{\partial}{\partial \psi}\left(\frac{\partial}{\partial \theta}\log \pi(Y \mid \theta) \cdot \frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right).$$

Product rule + chain rule

$$\frac{\partial^2}{\partial \psi^2}\ell(\psi) = \underbrace{\frac{\partial^2}{\partial \theta^2}\log \pi(Y \mid \theta)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)^2}_{\text{(A)}} + \underbrace{\frac{\partial}{\partial \theta}\log \pi(Y \mid \theta) \cdot \frac{\mathrm{d}^2\theta}{\mathrm{d}\psi^2}}_{\text{(B)}}.$$

Proof sketch (Step 3): take expectation and use $\mathbb{E}[\text{score}] = 0$

By definition,

$$I(\psi) = -\mathbb{E}\left[\frac{\partial^2}{\partial\psi^2}\ell(\psi)\right].$$

Proof sketch (Step 3): take expectation and use $\mathbb{E}[\text{score}] = 0$

By definition,

$$I(\psi) = -\mathbb{E}\left[\frac{\partial^2}{\partial \psi^2}\ell(\psi)\right].$$

Insert (A)+(B):

$$I(\psi) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2}\log \pi(Y \mid \theta)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)^2\right] - \mathbb{E}\left[\frac{\partial}{\partial \theta}\log \pi(Y \mid \theta) \cdot \frac{\mathrm{d}^2\theta}{\mathrm{d}\psi^2}\right].$$

Proof sketch (Step 3): take expectation and use $\mathbb{E}[\text{score}] = 0$

By definition,

$$I(\psi) = -\mathbb{E}\left[\frac{\partial^2}{\partial \psi^2}\ell(\psi)\right].$$

Insert (A)+(B):

$$I(\psi) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2}\log \pi(Y \mid \theta)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)^2\right] - \mathbb{E}\left[\frac{\partial}{\partial \theta}\log \pi(Y \mid \theta)\cdot \frac{\mathrm{d}^2\theta}{\mathrm{d}\psi^2}\right].$$

Fact (mean score is zero)

$$\mathbb{E}\left[\frac{\partial}{\partial \theta}\log \pi(Y \mid \theta)\right] = 0.$$

So the entire second term vanishes.

Proof sketch (Step 3): take expectation and use $\mathbb{E}[\text{score}] = 0$

By definition,

$$I(\psi) = -\mathbb{E}\left[\frac{\partial^2}{\partial\psi^2}\ell(\psi)\right].$$

Insert (A)+(B):

$$I(\psi) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log\pi(Y \mid \theta)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)^2\right] - \mathbb{E}\left[\frac{\partial}{\partial\theta}\log\pi(Y \mid \theta)\cdot\frac{\mathrm{d}^2\theta}{\mathrm{d}\psi^2}\right].$$

Fact (mean score is zero)

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log\pi(Y \mid \theta)\right] = 0.$$

So the entire second term vanishes.

Therefore

$$I(\psi) = \left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)^2\left(-\mathbb{E}[\frac{\partial^2}{\partial\theta^2}\log\pi(Y \mid \theta)]\right) = I(\theta)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)^2$$

Finish: invariance of Jeffreys prior

From the transformation rule:

$$\sqrt{I(\psi)} = \sqrt{I(\theta)} \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right|.$$

Finish: invariance of Jeffreys prior

From the transformation rule:

$$\sqrt{I(\psi)} = \sqrt{I(\theta)} \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right| .$$

Now define Jeffreys prior on θ:

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

Finish: invariance of Jeffreys prior

From the transformation rule:

$$\sqrt{I(\psi)} = \sqrt{I(\theta)} \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right|.$$

Now define Jeffreys prior on θ:

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

Transform it to ψ using the Jacobian rule:

$$\pi_J(\psi) = \pi_J(\theta(\psi)) \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right| \propto \sqrt{I(\theta)} \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right| = \sqrt{I(\psi)}.$$

Finish: invariance of Jeffreys prior

From the transformation rule:

$$\sqrt{I(\psi)} = \sqrt{I(\theta)} \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right|.$$

Now define Jeffreys prior on θ:

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

Transform it to ψ using the Jacobian rule:

$$\pi_J(\psi) = \pi_J(\theta(\psi)) \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right| \propto \sqrt{I(\theta)} \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right| = \sqrt{I(\psi)}.$$

Conclusion

Jeffreys' prior is invariant under one-to-one re-parameterisations.

How to compute Jeffreys' prior in practice

Recipe

Given likelihood $\pi(y \mid \theta)$:

1. Write the log-likelihood $\ell(\theta) = \log \pi(y \mid \theta)$.
2. Compute $\frac{\partial^2}{\partial\theta^2}\ell(\theta)$.
3. Take expectation w.r.t. $Y \sim \pi(\cdot \mid \theta)$:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right].$$

4. Set Jeffreys prior: $\pi_J(\theta) \propto \sqrt{I(\theta)}$.

How to compute Jeffreys' prior in practice

Recipe

Given likelihood $\pi(y \mid \theta)$:

1. Write the log-likelihood $\ell(\theta) = \log \pi(y \mid \theta)$.
2. Compute $\frac{\partial^2}{\partial\theta^2}\ell(\theta)$.
3. Take expectation w.r.t. $Y \sim \pi(\cdot \mid \theta)$:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right].$$

4. Set Jeffreys prior: $\pi_J(\theta) \propto \sqrt{I(\theta)}$.

Often, we only need $I(\theta)$ *up to proportionality*.

Example: Binomial model (bots on a platform)

Model (Example 3.1 revisited)

$$Y \mid \theta \sim \mathrm{Bin}(n, \theta), \qquad \theta \in (0, 1).$$

Likelihood:

$$\pi(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

Example: log-likelihood and derivatives

Log-likelihood:

$$\ell(\theta) = \log \binom{n}{y} + y \log \theta + (n - y) \log(1 - \theta).$$

Example: log-likelihood and derivatives

Log-likelihood:

$$\ell(\theta) = \log \binom{n}{y} + y \log \theta + (n - y) \log(1 - \theta).$$

First derivative:

$$\ell'(\theta) = \frac{y}{\theta} - \frac{n - y}{1 - \theta}.$$

Example: log-likelihood and derivatives

Log-likelihood:

$$\ell(\theta) = \log \binom{n}{y} + y \log \theta + (n - y) \log(1 - \theta).$$

First derivative:

$$\ell'(\theta) = \frac{y}{\theta} - \frac{n - y}{1 - \theta}.$$

Second derivative:

$$\ell''(\theta) = -\frac{y}{\theta^2} - \frac{n - y}{(1 - \theta)^2}.$$

Example: Fisher information

By definition,

$$I(\theta) = -\mathbb{E}[\ell''(\theta)].$$

Example: Fisher information

By definition,

$$I(\theta) = -\mathbb{E}[\ell''(\theta)].$$

Compute expectation using $\mathbb{E}[Y] = n\theta$ for $Y \sim \mathrm{Bin}(n, \theta)$:

$$\mathbb{E}[\ell''(\theta)] = -\frac{\mathbb{E}[Y]}{\theta^2} - \frac{n - \mathbb{E}[Y]}{(1-\theta)^2} = -\frac{n\theta}{\theta^2} - \frac{n(1-\theta)}{(1-\theta)^2}.$$

Example: Fisher information

By definition,

$$I(\theta) = -\mathbb{E}[\ell''(\theta)].$$

Compute expectation using $\mathbb{E}[Y] = n\theta$ for $Y \sim \mathrm{Bin}(n, \theta)$:

$$\mathbb{E}[\ell''(\theta)] = -\frac{\mathbb{E}[Y]}{\theta^2} - \frac{n - \mathbb{E}[Y]}{(1-\theta)^2} = -\frac{n\theta}{\theta^2} - \frac{n(1-\theta)}{(1-\theta)^2}.$$

So

$$\mathbb{E}[\ell''(\theta)] = -\frac{n}{\theta} - \frac{n}{1-\theta} = -\frac{n}{\theta(1-\theta)}.$$

Example: Fisher information

By definition,

$$I(\theta) = -\mathbb{E}[\ell''(\theta)].$$

Compute expectation using $\mathbb{E}[Y] = n\theta$ for $Y \sim \mathrm{Bin}(n, \theta)$:

$$\mathbb{E}[\ell''(\theta)] = -\frac{\mathbb{E}[Y]}{\theta^2} - \frac{n - \mathbb{E}[Y]}{(1-\theta)^2} = -\frac{n\theta}{\theta^2} - \frac{n(1-\theta)}{(1-\theta)^2}.$$

So

$$\mathbb{E}[\ell''(\theta)] = -\frac{n}{\theta} - \frac{n}{1-\theta} = -\frac{n}{\theta(1-\theta)}.$$

Therefore

$$I(\theta) = \frac{n}{\theta(1-\theta)}.$$

Example: Jeffreys prior and identification

Jeffreys prior:

$$\pi_J(\theta) \propto \sqrt{I(\theta)} = \sqrt{\frac{n}{\theta(1-\theta)}} \propto \frac{1}{\sqrt{\theta(1-\theta)}}.$$

Example: Jeffreys prior and identification

Jeffreys prior:

$$\pi_J(\theta) \propto \sqrt{I(\theta)} = \sqrt{\frac{n}{\theta(1-\theta)}} \propto \frac{1}{\sqrt{\theta(1-\theta)}}.$$

Rewrite:

$$\pi_J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}.$$

Example: Jeffreys prior and identification

Jeffreys prior:

$$\pi_J(\theta) \propto \sqrt{I(\theta)} = \sqrt{\frac{n}{\theta(1-\theta)}} \propto \frac{1}{\sqrt{\theta(1-\theta)}}.$$

Rewrite:

$$\pi_J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}.$$

Recognise the family

This is the kernel of a Beta distribution:

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

Example: Jeffreys prior and identification

Jeffreys prior:

$$\pi_J(\theta) \propto \sqrt{I(\theta)} = \sqrt{\frac{n}{\theta(1-\theta)}} \propto \frac{1}{\sqrt{\theta(1-\theta)}}.$$

Rewrite:

$$\pi_J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}.$$

Recognise the family

This is the kernel of a Beta distribution:

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

So the **invariant** prior for the binomial probability is $\mathrm{Beta}(1/2, 1/2)$.

Interpretation: what does $\mathrm{Beta}(1/2, 1/2)$ look like?

- $\mathrm{Beta}(1/2, 1/2)$ is U-shaped: it places more mass near 0 and 1 than near 1/2.

Interpretation: what does $\mathrm{Beta}(1/2, 1/2)$ look like?

- $\mathrm{Beta}(1/2, 1/2)$ is U-shaped: it places more mass near 0 and 1 than near 1/2.
- This reflects the geometry of the Bernoulli/binomial likelihood:

$$I(\theta) = \frac{n}{\theta(1-\theta)} \quad \text{blows up near } 0, 1.$$

Interpretation: what does $\mathrm{Beta}(1/2, 1/2)$ look like?

- $\mathrm{Beta}(1/2, 1/2)$ is U-shaped: it places more mass near 0 and 1 than near $1/2$.
- This reflects the geometry of the Bernoulli/binomial likelihood:

$$I(\theta) = \frac{n}{\theta(1-\theta)} \quad \text{blows up near } 0, 1.$$

- "Objective" does not mean "flat": uniform prior $\mathrm{Beta}(1, 1)$ is *not* invariant under re-parameterisation.

Quick recap (what you should remember)

Jeffreys' prior

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

Quick recap (what you should remember)

Jeffreys' prior

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

- Fisher information transforms as

$$I(\psi) = I(\theta) \left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right)^2 .$$

Quick recap (what you should remember)

Jeffreys' prior

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

- Fisher information transforms as

$$I(\psi) = I(\theta)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)^2.$$

- Hence

$$\sqrt{I(\psi)} = \sqrt{I(\theta)}\left|\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right|$$

which matches the Jacobian density transform.

Quick recap (what you should remember)

Jeffreys' prior

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

- Fisher information transforms as

$$I(\psi) = I(\theta) \left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right)^2.$$

- Hence

$$\sqrt{I(\psi)} = \sqrt{I(\theta)} \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right|$$

  which matches the Jacobian density transform.
- Binomial example: $\pi_J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, so $\theta \sim \mathrm{Beta}(1/2, 1/2)$.

History break: Harold Jeffreys (brief)

- Harold Jeffreys (1891–1989): major figure in Bayesian methods and scientific inference.

History break: Harold Jeffreys (brief)

- Harold Jeffreys (1891–1989): major figure in Bayesian methods and scientific inference.
- Worked broadly across geophysics, astronomy, and statistical methodology.

History break: Harold Jeffreys (brief)

- Harold Jeffreys (1891–1989): major figure in Bayesian methods and scientific inference.
- Worked broadly across geophysics, astronomy, and statistical methodology.
- Motivated by making Bayesian inference less dependent on subjective prior choices: Jeffreys' prior is one influential attempt at an *objective* prior.

History break: Harold Jeffreys (brief)

- Harold Jeffreys (1891–1989): major figure in Bayesian methods and scientific inference.
- Worked broadly across geophysics, astronomy, and statistical methodology.
- Motivated by making Bayesian inference less dependent on subjective prior choices: Jeffreys' prior is one influential attempt at an *objective* prior.

Takeaway

Jeffreys' prior is not magic; it is a principled default driven by invariance.

(Optional bridge) Why we care: towards large-sample agreement

- Jeffreys: "choose priors invariantly" (objective defaults).

(Optional bridge) Why we care: towards large-sample agreement

- Jeffreys: "choose priors invariantly" (objective defaults).
- Next big theorem (Bernstein–von Mises): with lots of data, posterior $\approx$ Normal around the MLE (under conditions).

(Optional bridge) Why we care: towards large-sample agreement

- Jeffreys: "choose priors invariantly" (objective defaults).
- Next big theorem (Bernstein–von Mises): with lots of data, posterior $\approx$ Normal around the MLE (under conditions).
- So Bayesian and frequentist answers often agree asymptotically.

End

Questions?