Non-informative Priors, Jeffreys Prior, and Frequentist Properties

Sections 3.5–3.7

Roadmap

- **3.5 Non-informative priors:** why "uniform" can be misleading
- **Jeffreys prior:** definition via Fisher information
- **Invariance theorem:** why Jeffreys prior is reparametrisation-invariant
- Examples: Binomial model; Normal mean and *improper* priors
- **3.6 Frequentist view:** bias/variance of Bayesian estimators; coverage of credible intervals
- **3.7 Hierarchical models:** hyperpriors and conditional posteriors (Gibbs sampling motivation)

Motivation: prior choice can matter

- We have seen in examples that **prior choice** (and prior parameters) affects:
  - posterior distributions,
  - posterior summaries (mean/MAP),
  - and practical conclusions.
- A classic criticism of Bayesian inference: the prior is **subjective**.
- One response: use a prior that reflects **lack of information** about θ.

A very simple model: Bernoulli

Assume

$$X \mid \theta \sim \mathrm{Bern}(\theta), \qquad \pi(x \mid \theta) = \theta^x (1-\theta)^{1-x}, \quad \theta \in [0,1],\ x \in \{0,1\}.$$

- First instinct for "no information": choose $\theta \sim \mathrm{Unif}[0,1]$.
- But "uniform" depends on how we parametrise the model.

Uniform is not invariant: a reparametrisation

Reparametrise the model using

$$\psi = \theta^2 \in [0, 1], \qquad \theta = \sqrt{\psi}.$$

Then the likelihood becomes

$$\pi(x \mid \psi) = (\sqrt{\psi})^x (1 - \sqrt{\psi})^{1-x}.$$

Uniform is not invariant: a reparametrisation

Reparametrise the model using

$$\psi = \theta^2 \in [0, 1], \qquad \theta = \sqrt{\psi}.$$

Then the likelihood becomes

$$\pi(x \mid \psi) = (\sqrt{\psi})^x (1 - \sqrt{\psi})^{1-x}.$$

If we place a uniform prior on θ:

$$\pi(\theta) = 1, \quad \theta \in [0, 1],$$

what prior does this imply for ψ?

Change of variables: induced prior on ψ

Using the change-of-variable rule:

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

Here $\theta(\psi) = \sqrt{\psi}$, so

$$\frac{\mathrm{d}\theta}{\mathrm{d}\psi} = \frac{1}{2\sqrt{\psi}} \quad \Rightarrow \quad \pi(\psi) = 1 \cdot \frac{1}{2\sqrt{\psi}}.$$

- This density is **not uniform** on $[0, 1]$.
- It places **more mass near** $\psi = 0$ (equivalently near $\theta = 0$).

Conclusion: uniform priors are not invariant to reparametrisation.

Jeffreys' principle

Sir Harold Jeffreys argued:

> *If there are two sensible ways to parametrise a model, priors under these parametrisations should be* **equivalent***.*

- Goal: define a "non-informative" prior that is **invariant** under smooth 1–1 transformations.
- Tool: Fisher information.

Definition: Fisher information

Given a model $Y \mid \theta \sim \pi(y \mid \theta)$, define the Fisher information

$$I_Y(\theta) = \mathrm{Var}\left[\frac{\partial}{\partial\theta} \log \pi(Y \mid \theta)\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log \pi(Y \mid \theta)\right], \qquad Y \sim \pi(y \mid \theta).$$

- Both expressions are equal under standard regularity conditions.
- Intuition: $I_Y(\theta)$ measures how **informative** the data distribution is about θ.

Definition 3.3: Jeffreys invariant prior

Jeffreys prior is defined as

$$\pi(\theta) \;\propto\; \sqrt{I_Y(\theta)}.$$

- Depends on the likelihood through $I_Y(\theta)$.
- Designed to be invariant under smooth one-to-one reparametrisations.
- May be **improper** (does not integrate to 1), but can still yield a proper posterior.

Theorem 3.1: invariance statement

Let $Y \mid \theta \sim \pi(y \mid \theta)$ and reparametrise by

$$\psi = h(\theta),$$

where h is smooth and strictly monotone (so $\theta = h^{-1}(\psi)$ exists). Then Jeffreys prior is invariant in the sense that

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

Proof idea (high level)

Start from

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

So it suffices to show

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

We do this by computing $I_Y(\psi)$ from the definition using the chain rule.

Proof sketch (chain rule + score mean zero)

Compute the second derivative:

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

Product rule gives two terms:

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

Proof sketch (chain rule + score mean zero)

Compute the second derivative:

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

Product rule gives two terms:

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

Key identity (score has mean zero):

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

So the second term vanishes and

$$I_Y(\psi) = I_Y(\theta)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)^2 \quad\Rightarrow\quad \sqrt{I_Y(\psi)} = \sqrt{I_Y(\theta)}\left|\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right|.$$

Example 3.8: Binomial model $\Rightarrow$ Beta$(\frac{1}{2}, \frac{1}{2})$

Let $Y \mid \theta \sim \mathrm{Bin}(n, \theta)$, with pmf

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

Log-likelihood:

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

Derivatives:

$$\frac{\partial}{\partial \theta} \log \pi(y \mid \theta) = \frac{y}{\theta} - \frac{n-y}{1-\theta}, \quad \frac{\partial^2}{\partial \theta^2} \log \pi(y \mid \theta) = -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}.$$

Example 3.8 continued: compute Fisher information

Using $I_Y(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log \pi(Y \mid \theta)\right]$:

$$I_Y(\theta) = \mathbb{E}\left[\frac{Y}{\theta^2} + \frac{n-Y}{(1-\theta)^2}\right] = \frac{\mathbb{E}[Y]}{\theta^2} + \frac{n - \mathbb{E}[Y]}{(1-\theta)^2}.$$

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

$$I_Y(\theta) = \frac{n\theta}{\theta^2} + \frac{n - n\theta}{(1-\theta)^2} = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)}.$$

Thus

$$\pi(\theta) \propto \sqrt{I_Y(\theta)} \propto \theta^{-1/2}(1-\theta)^{-1/2},$$

i.e.

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

Example 3.9: Normal mean $\Rightarrow$ improper prior

Let

$$Y \mid \mu \sim \mathcal{N}(\mu, \sigma^2), \quad \sigma > 0 \text{ known}, \ \mu \in \mathbb{R}.$$

One can show

$$I_Y(\mu) = \frac{1}{\sigma^2}.$$

Therefore Jeffreys prior:

$$\pi(\mu) \propto \sqrt{I_Y(\mu)} = \frac{1}{\sigma} \ \propto \ 1.$$

Example 3.9: Normal mean $\Rightarrow$ improper prior

Let

$$Y \mid \mu \sim \mathcal{N}(\mu, \sigma^2), \quad \sigma > 0 \text{ known}, \ \mu \in \mathbb{R}.$$

One can show

$$I_Y(\mu) = \frac{1}{\sigma^2}.$$

Therefore Jeffreys prior:

$$\pi(\mu) \propto \sqrt{I_Y(\mu)} = \frac{1}{\sigma} \ \propto \ 1.$$

But on $\mathbb{R}$,

$$\int_{-\infty}^{\infty} \pi(\mu) \, \mathrm{d}\mu = \infty,$$

so $\pi(\mu) \propto 1$ is improper.

Improper priors: what and why?

Definition (improper prior). A prior $\pi(\theta)$ on Θ is improper if

$$\int_{\Theta} \pi(\theta)\,\mathrm{d}\theta = \infty.$$

- Improper priors are commonly used to express "very weak information".
- They are acceptable **if the posterior is proper** (normalisable):

$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta) \quad \text{and} \quad \int_{\Theta} \pi(\theta \mid y)\,\mathrm{d}\theta < \infty.$$

- Always check posterior propriety when using improper priors.

Frequentist vs Bayesian viewpoint (conceptual)

- **Frequentist:** there is a true (fixed) parameter $\theta^\star$ generating the data.
- **Bayesian:** the parameter θ is modelled as random with prior $\pi(\theta)$.

Frequentist vs Bayesian viewpoint (conceptual)

- **Frequentist:** there is a true (fixed) parameter $\theta^\star$ generating the data.
- **Bayesian:** the parameter θ is modelled as random with prior $\pi(\theta)$.

A useful bridge:

- Treat Bayesian inference as a method to produce estimators/intervals.
- Then analyse their **frequentist properties** under data generated at $\theta^\star$:
  - bias, variance, MSE,
  - coverage of credible intervals.

Example 3.10: conjugate normal model

Model + prior:

$$X_1, \ldots, X_n \mid \theta \overset{i.i.d}{\sim} \mathcal{N}(\theta, 1), \qquad \theta \sim \mathcal{N}(0, 1).$$

Posterior (from conjugacy):

$$\theta \mid X \sim \mathcal{N}\left(\frac{n}{n+1}\overline{X}_n,\ \frac{1}{n+1}\right), \qquad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Posterior mean estimator:

$$\widehat{\theta}_n := \mathbb{E}[\theta \mid X] = \frac{n}{n+1}\overline{X}_n.$$

Frequentist analysis: assume a true $\theta^\star$

Now assume (hypothetically)

$$X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\theta^\star, 1),$$

with fixed $\theta^\star$.

Compute bias and variance of $\widehat{\theta}_n = \frac{n}{n+1}\overline{X}_n$:

$$\mathbb{E}[\widehat{\theta}_n] = \frac{n}{n+1}\mathbb{E}[\overline{X}_n] = \frac{n}{n+1}\theta^\star \quad \Rightarrow \quad \mathrm{Bias}(\widehat{\theta}_n) = \mathbb{E}[\widehat{\theta}_n] - \theta^\star = -\frac{1}{n+1}\theta^\star.$$

$$\mathrm{Var}(\widehat{\theta}_n) = \left(\frac{n}{n+1}\right)^2 \mathrm{Var}(\overline{X}_n) = \left(\frac{n}{n+1}\right)^2 \frac{1}{n} = \frac{n}{(n+1)^2}.$$

Compare to MLE and asymptotic agreement

For this model, the MLE is $\overline{X}_n$ with

$$\mathrm{Bias}(\overline{X}_n) = 0, \qquad \mathrm{Var}(\overline{X}_n) = \frac{1}{n}.$$

Difference:

$$\overline{X}_n - \widehat{\theta}_n = \overline{X}_n - \frac{n}{n+1}\overline{X}_n = \frac{1}{n+1}\overline{X}_n \xrightarrow{P} 0,$$

since $\overline{X}_n \xrightarrow{P} \theta^\star$ and $(n+1)^{-1} \to 0$.

Compare to MLE and asymptotic agreement

For this model, the MLE is $\overline{X}_n$ with

$$\mathrm{Bias}(\overline{X}_n) = 0, \qquad \mathrm{Var}(\overline{X}_n) = \frac{1}{n}.$$

Difference:

$$\overline{X}_n - \widehat{\theta}_n = \overline{X}_n - \frac{n}{n+1}\overline{X}_n = \frac{1}{n+1}\overline{X}_n \xrightarrow{P} 0,$$

since $\overline{X}_n \xrightarrow{P} \theta^\star$ and $(n+1)^{-1} \to 0$.

So posterior mean and MLE **agree asymptotically**.

Credible interval and frequentist coverage (idea)

A $(1-\alpha)$ credible interval from the posterior is

$$C_n(X) = \widehat{\theta}_n \pm \frac{1}{\sqrt{n+1}} \Phi^{-1}(1-\alpha/2),$$

where Φ^{-1} is the standard Normal quantile function.

Frequentist coverage asks:

$$\mathbb{P}_{\theta^\star}\big(\theta^\star \in C_n(X)\big), \quad \text{under } X_i \overset{i.i.d}{\sim} \mathcal{N}(\theta^\star, 1).$$

Using asymptotic normality and Slutsky-type arguments, one obtains

$$\mathbb{P}_{\theta^\star}\big(\theta^\star \in C_n(X)\big) \;\longrightarrow\; 1-\alpha \quad \text{as } n \to \infty.$$

So (in this model) Bayesian credible intervals are asymptotically valid confidence intervals.

Theorem 3.2: Bernstein–von Mises (statement)

Let $X_1, \dots, X_n \overset{i.i.d}{\sim} P_{\theta^\star}$ with $\theta^\star \in \Theta \subseteq \mathbb{R}$. Let $\widehat{\theta}_{\mathrm{MLE}}$ be the MLE and $I(\theta^\star)$ the Fisher information.

Under mild regularity conditions and a prior $\pi(\theta)$ that is **positive near** $\widehat{\theta}_{\mathrm{MLE}}$, the posterior is asymptotically normal:

$$\theta \mid X \;\approx\; \mathcal{N}\left(\widehat{\theta}_{\mathrm{MLE}},\; (nI(\theta^\star))^{-1}\right), \qquad n \to \infty,$$

more precisely (in total variation distance):

$$\frac{1}{2}\int |\pi(\theta \mid X) - \widehat{\varphi}_n(\theta)| \, \mathrm{d}\theta \xrightarrow{a.s.} 0,$$

where $\widehat{\varphi}_n$ is the density of $\mathcal{N}(\widehat{\theta}_{\mathrm{MLE}}, (nI(\theta^\star))^{-1})$.

Consequence: asymptotic credible interval matches CI

When $\theta \in \mathbb{R}$, BvM implies an approximate $(1-\alpha)$ credible interval:

$$C_n(X) = \widehat{\theta}_{\mathrm{MLE}} \pm \frac{1}{\sqrt{n\, I(\widehat{\theta}_{\mathrm{MLE}})}} \Phi^{-1}(1-\alpha/2).$$

- This matches the usual **asymptotic confidence interval** from MLE theory.
- Hence coverage satisfies

$$\mathbb{P}_{\theta^\star}\left(\theta^\star \in C_n(X)\right) \to 1-\alpha.$$

Why hierarchical models?

In many problems:

- We have **multiple parameters** that relate to each other.
- Prior parameters may themselves be uncertain.

Why hierarchical models?

In many problems:

- We have **multiple parameters** that relate to each other.
- Prior parameters may themselves be uncertain.

A **hierarchical model** builds layers:

$$\text{hyperparameters} \rightarrow \text{parameters} \rightarrow \text{data.}$$

Benefits:

- more flexible modelling of uncertainty,
- reduced sensitivity to fixed prior hyperparameters,
- enables structured inference (e.g. Gibbs sampling).

Example 3.11: Exponential likelihood with hyperprior

Recall (Example 3.4): data $y = (y_1, \dots, y_n)$ with

$$Y_i \mid \lambda \overset{i.i.d}{\sim} \mathrm{Exp}(\lambda), \qquad \lambda > 0.$$

Previously we set a fixed prior $\lambda \sim \mathrm{Exp}(\gamma)$ and obtained

$$\lambda \mid y \sim \mathrm{Gamma}\left(n + 1, \sum_{i=1}^{n} y_i + \gamma\right).$$

Example 3.11: Exponential likelihood with hyperprior

Recall (Example 3.4): data $y = (y_1, \ldots, y_n)$ with

$$Y_i \mid \lambda \overset{i.i.d}{\sim} \mathrm{Exp}(\lambda), \qquad \lambda > 0.$$

Previously we set a fixed prior $\lambda \sim \mathrm{Exp}(\gamma)$ and obtained

$$\lambda \mid y \sim \mathrm{Gamma}\left(n+1,\ \sum_{i=1}^{n} y_i + \gamma\right).$$

Now treat γ as unknown by putting a **hyperprior**:

$$\gamma \sim \mathrm{Exp}(\nu).$$

Hierarchy diagram and joint posterior

Hierarchy:

$$Y_1, \ldots, Y_n \mid \lambda \sim \mathrm{Exp}(\lambda) \quad \text{(likelihood)}$$

$$\lambda \mid \gamma \sim \mathrm{Exp}(\gamma) \quad \text{(prior)}$$

$$\gamma \sim \mathrm{Exp}(\nu) \quad \text{(hyperprior)}$$

Diagram: $\gamma \to \lambda \to \{Y_i\}_{i=1}^n$.

Joint posterior:

$$\pi(\lambda, \gamma \mid y) \propto \pi(y \mid \lambda)\pi(\lambda \mid \gamma)\pi(\gamma).$$

Up to proportionality (collecting kernel terms):

$$\pi(\lambda, \gamma \mid y) \propto \lambda^n \gamma \exp\Big(-\lambda\big(\sum_{i=1}^n y_i + \gamma\big)\Big) \exp(-\nu\gamma), \quad \lambda, \gamma > 0.$$

Conditional posteriors (key for Gibbs sampling)

Use the identity

$$\pi(\lambda, \gamma \mid y) = \pi(\lambda \mid y, \gamma)\pi(\gamma \mid y) = \pi(\gamma \mid y, \lambda)\pi(\lambda \mid y).$$

To get $\pi(\lambda \mid y, \gamma)$, keep only terms depending on λ:

$$\pi(\lambda \mid y, \gamma) \propto \lambda^n \exp\Big(-\lambda\big(\sum_{i=1}^{n} y_i + \gamma\big)\Big),$$

so

$$\lambda \mid y, \gamma \sim \mathrm{Gamma}\left(n+1,\ \sum_{i=1}^{n} y_i + \gamma\right).$$

Similarly, keep only terms depending on γ:

$$\pi(\gamma \mid y, \lambda) \propto \gamma \exp\left(-(\lambda+\nu)\gamma\right),$$

so

$$\gamma \mid y, \lambda \sim \mathrm{Gamma}(2,\ \lambda+\nu).$$

Why these conditionals matter

- The conditional posteriors have **standard forms** (Gamma distributions).
- This makes simulation-based inference straightforward:
  - sample $\lambda^{(t+1)} \sim \pi(\lambda \mid y, \gamma^{(t)})$,
  - sample $\gamma^{(t+1)} \sim \pi(\gamma \mid y, \lambda^{(t+1)})$,

  which is exactly the structure used by **Gibbs sampling** (an MCMC method).

Summary

- Uniform priors are **not** invariant: "non-informative" depends on parametrisation.
- Jeffreys prior uses Fisher information:

$$\pi(\theta) \propto \sqrt{I_Y(\theta)},$$

  and is invariant to smooth 1–1 reparametrisations.
- Examples:
  - Binomial $\Rightarrow \mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$,
  - Normal mean $\Rightarrow \pi(\mu) \propto 1$ (improper).
- Frequentist analysis can study bias/variance/coverage of Bayesian outputs.
- Hierarchical models add hyperpriors; conditional posteriors enable Gibbs sampling.