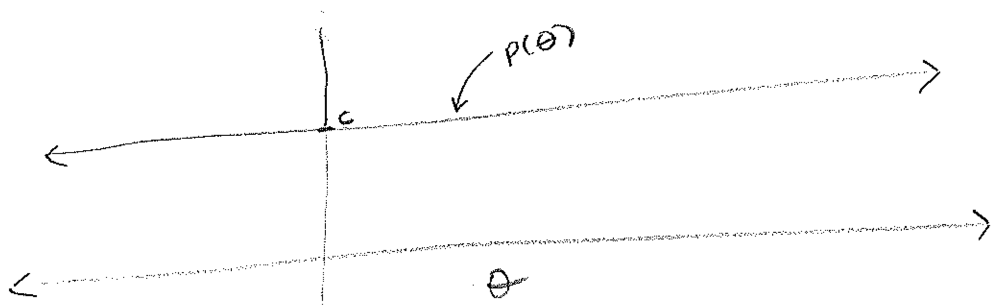
PRIORS

Improper priors

- Most use this phrase to describe prior beliefs, $p(\theta)$, that **cannot** be normalized to integrate to 1.
- **Bayesian Data Analysis** also call data-dependent priors "improper" because they violate the important philosophical principle that a prior, $p(\theta)$, must be specified **before** looking at the data.

Example:

$$p(\theta) \propto \overset{\text{constant}}{c}, \quad \theta \in \mathbb{R}$$



$$\underset{\text{constant}}{\frac{1}{K}} \int_{-\infty}^{\infty} c \, d\theta = \frac{c}{K} \theta \Big|_{-\infty}^{\infty} = \infty$$

and **cannot** be **normalized** for any choice of K.

* Uniform priors **seem** "uninformative" but this is deceiving.

- why does it **seem** uninformative? If $p(\theta) \propto c$, then
$$p(\theta \mid y) \propto p(y \mid \theta).$$

- why is it **not**?

Ex: 

- A cubic box has side length l.
- $l \sim \text{Unif}(0,1)$
  $l^2 \sim \text{Unif}(0,1)$
  $l^3 \sim \text{Unif}(0,1)$

$\Bigg\}$ all different! **Parameterization Matters!**

- Why else is it not "uninformative"?

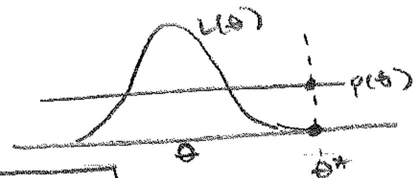
<u>Ex</u>: Let $p(\theta) \propto c$, $\theta \in \mathbb{R}$.

a priori, $\text{prob}(\theta \in (a,b)) = 0 \quad \forall\ a, b < \infty$

<u>Note</u>: The minimal condition that <u>must be satisfied</u> to use an improper prior in Bayesian inference is that the <u>posterior</u> is <u>proper</u>.

i.e. $\int p(y|\theta)\,p(\theta)\,d\theta < \infty$ (the posterior <u>can be</u> normalized).

<u>Picture:</u> $\text{posterior}(\theta) \propto L(\theta) \cdot p(\theta)$



$\text{post}(\theta^*) = L(\theta^*) \cdot p(\theta^*)$

posterior is proper if tails of $L(\theta)$ shrink "fast enough"!

Unit Information Prior

- A <u>data-dependent</u> prior
- Considered "philosophically improper" by the Gelman/BDA definition
- typically is a proper density
- often <u>mathematically convenient</u> (e.g. Zellner's g prior) <u>will talk more on later</u>
- contain same amount of <u>information</u> as a single observation.

<u>Example</u>: $Y_i | \beta, \sigma^2 \sim \text{Normal}(x_i \beta, \sigma^2)$

$\beta | \sigma^2 \sim \text{Normal}(\beta_0, \tau^2)$

<u>Here</u>: $Y_i \in \mathbb{R}$, $\beta \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$ — <u>univariate ex</u>

want β_0, τ^2 such that we have "unit info." about β a priori.

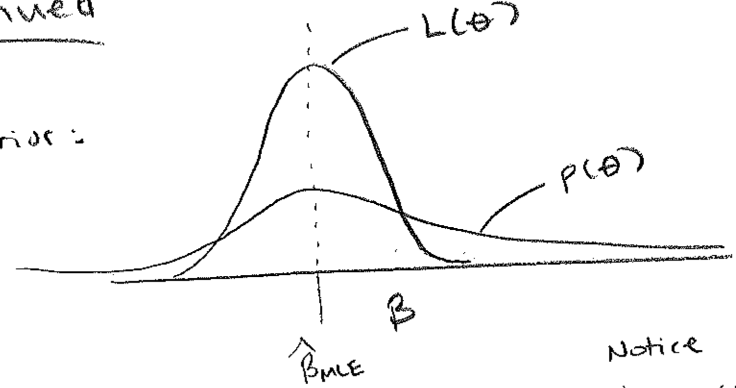
Solution: Let

$$\boxed{\beta_0 = \hat{\beta}_{MLE} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\mathbf{x}^T \mathbf{y}}{\mathbf{x}^T \mathbf{x}} \qquad \tau^2 = \frac{n\sigma^2}{\mathbf{x}^T \mathbf{x}}}$$

Notice: <u>$\mathbf{x}^T\mathbf{x}$</u> is 1×1 (a scalar)

Unit Info. Prior continued

Picture of unit info prior:



Notice wide variance in $p(\theta)$.

Math of unit info idea:

$\underline{y}$:= $n \times 1$ vector of outcomes
$\underline{x}$:= $n \times 1$ vector of predictor

$$p(\underline{y} \mid \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ \frac{-1}{2\sigma^2} (\underline{y} - \underline{X}\beta)^T (\underline{y} - \underline{X}\beta) \right\}$$

$$p(\beta \mid \sigma^2) = \left(2\pi \cdot \frac{n\sigma^2}{\underline{X}^T\underline{X}}\right)^{-1/2} \exp\left\{ \frac{-1}{2n\sigma^2} \cdot \underline{X}^T\underline{X} \left(\beta - \frac{\underline{X}^T\underline{y}}{\underline{X}^T\underline{X}}\right)^T \left(\beta - \frac{\underline{X}^T\underline{y}}{\underline{X}^T\underline{X}}\right) \right\}$$

$$p(\beta \mid \underline{y}, \sigma^2) \underset{\beta}{\propto} \exp\left\{ \frac{-1}{2\sigma^2} \left(\beta^2(\underline{X}^T\underline{X}) - 2\beta \underline{X}^T\underline{y}\right) - \frac{1}{2n\sigma^2} \cdot \underline{X}^T\underline{X} \left(\beta^2 - 2\beta\left(\frac{\underline{X}^T\underline{y}}{\underline{X}^T\underline{X}}\right)\right) \right\}$$

combine terms:

$$\underset{\beta}{\propto} \exp\left\{ \frac{-1}{2\sigma^2} \left[\beta^2\left(\underline{X}^T\underline{X} + \frac{\underline{X}^T\underline{X}}{n}\right) - 2\beta\left(\underline{X}^T\underline{y} + \frac{\underline{X}^T\underline{y}}{n}\right) \right] \right\}$$

From the normal model notes, complete the sq:

$$a = \frac{(n+1)}{n\sigma^2} \underline{X}^T\underline{X} \quad ; \quad b = \frac{(n+1)}{n\sigma^2} \underline{X}^T\underline{y}$$

$$\beta \mid \underline{y}, \sigma^2 \sim N\left(\frac{b}{a}, a^{-1}\right)$$ i.e.

$$\boxed{\beta \mid \underline{y}, \sigma^2 \sim N\left(\frac{\underline{X}^T\underline{y}}{\underline{X}^T\underline{X}}, \frac{n}{n+1} \cdot \sigma^2 \underline{X}^T\underline{X}\right)}$$

still centered on MLE.

slightly smaller variance (equivalent to 1 extra obs).

Exercise: Unit Info. Prior.

Write $\frac{1}{n} \log p(y|\theta)$:

$= \frac{1}{n} \log \left(\theta^{\Sigma y_i} (1-\theta)^{n - \Sigma y_i} \right)$, recall $\Sigma y_i = \bar{y} \cdot n$, expand log(·)

$= \frac{1}{\cancel{n}} \left[\cancel{n} \bar{y} \log(\theta) + (\cancel{n} - \cancel{n} \bar{y}) \log(1-\theta) \right]$

$= \bar{y} \log(\theta) + (1 - \bar{y}) \log(1-\theta) \quad := \text{A}$

Set $\log p(\theta) = \text{A} + c$ ← constant

So that when I exponentiate,

$p(\theta) = \underbrace{e^c}_{\text{constant}} \cdot \theta^{\bar{y}} (1-\theta)^{1-\bar{y}}$, looks like kernel of beta!

$\Rightarrow \boxed{\theta \sim \text{beta}(\bar{y}+1, 2-\bar{y})}$

Jeffreys prior, $J(\theta)$

- is the "reference prior" in 1D parameter models.
- invariant to re-parameterization*
  - * i.e. if I write $J_\theta(\theta)$ or $J_f(f(\theta))$ where $\theta \to f(\theta)$ or f is monotic, and then transform $f(\theta) \to \theta$, I will still have the Jeffreys prior in the new parameterization.
- typically improper (but not always)
- in 1D, often considered "non-informative"

Exercise Jeffreys Prior

$$P(Y|\theta) = \frac{\theta^y e^{-\theta}}{y!}$$

Let $\ell(\theta) \equiv \log p(Y|\theta) = y \log\theta - \theta - \log y!$

then $\frac{d}{d\theta}\ell(\theta) = \frac{y}{\theta} - 1$

and $\frac{d^2}{d\theta^2}\ell(\theta) = \frac{-y}{\theta^2}$

Compute:

$$-\mathbb{E}\left[\frac{-y}{\theta^2}\Big|\theta\right] = \int \frac{y}{\theta^2} p(y|\theta)\,dy$$

$$= \frac{1}{\theta^2}\underbrace{\int y\, p(y|\theta)\,dy}_{\theta,\ \text{since } y|\theta \sim \text{Poisson}(\theta)}$$

$$I(\theta) = \frac{1}{\theta}$$

since $P_{\text{JEFFREYS}}(\theta) \propto \sqrt{I(\theta)}$, $\boxed{p(\theta) \propto \theta^{-1/2}}$ □

Family of posterior: gamma, proof:

$$p(\theta|y) \propto \theta^{\sum y_i} e^{-n\theta} \cdot \theta^{-1/2}$$

$$= \theta^{n\bar{y}-\frac{1}{2}} e^{-n\theta} \quad \leftarrow \text{looks like kernel of gamma!}$$

$$\theta|y \sim \text{gamma}(n\bar{y} + \tfrac{1}{2}, n)$$ □

Bonus Let $\phi = \log\theta \Rightarrow \theta = e^\phi$

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| = e^{-\phi/2} \cdot e^\phi = e^{\phi/2}$$

$$\begin{array}{ccc} Y|\theta & & Y|\phi = f(\theta) \\ \downarrow & & \downarrow \\ J_\theta(\theta) & \longleftrightarrow & J_\phi(\phi) \end{array}$$

variable transformation