# Outline

1) Where does the prior come from?

2) Conjugate Priors

3) Bayesian pros and cons

# Where does prior come from?

Biggest issue with Bayes in practice is
how to choose prior

In general, can't check goodness of fit:
1 draw of $\theta \sim \Lambda$, not even directly observed.

Various ideas of how to do it:

## 1) Prior experience:

Ex. A/B testing in tech. co.s,
estimating eff. of 400,000 SNPs on trait,
estimating "true" 3PT% for basketball players

Prior is (relatively) non-controversial

- Can fit from data
  (leads to hierarchical / Empirical Bayes)
- Can test validity of prior b/c we have many draws from it

Works when encountering similar problems repeatedly (but are they really similar?)

## 2) Subjective beliefs:

Prior reflects epistemic uncertainty

Posterior = rational updating of beliefs.

**Pros:**

- Can't be wrong about your own opinion!
- Can bring to bear hard-to-formalize knowledge from outside the data

**Issues:**

- Philosophical conundrum: is the mass of a particle really "random"?
- Scientists find subjectivity offputting (reporting posterior is just reporting an opinion)
- Generally impossible to write down your beliefs about joint dist. of $\theta \in \mathbb{R}^{10}$
- What if people are systematically overconfident?

But This is the most philosophically coherent account of statistics. (Coin flip demo)

## 3) Convenience Prior

Bayes computations can be very hard:
generically, it's very hard to compute the normalizing constant $\int_\Omega \lambda(\theta) p_\theta(x) d\theta$

If $\dim(\Omega)$ large, posterior $\approx 0$ for most of $\Omega$

Lots of Bayesian research is computational (e.g., MCMC) and great progress has been made.

Helps to use conjugate priors where possible

<u>But</u> then the subjective account basically falls apart

# 4) "Objective" Priors

Suppose $X_i | \theta \overset{iid}{\sim} N(\theta, 1) \quad i = 1, \ldots, n$

"Natural" choice is $\theta \sim$ "flat prior"

$$\lambda(\theta) \propto_\theta 1$$

This prior is improper but it's ok:

$$\lambda(\theta | x) \propto_\theta e^{\theta \Sigma x_i - n\theta^2/2}$$

$$\propto_\theta N(\bar{x}, n^{-1})$$

More generally, could always use flat prior

Arises naturally as limit of $\theta \sim N(0, \tau^2)$, $\tau^2 \to \infty$

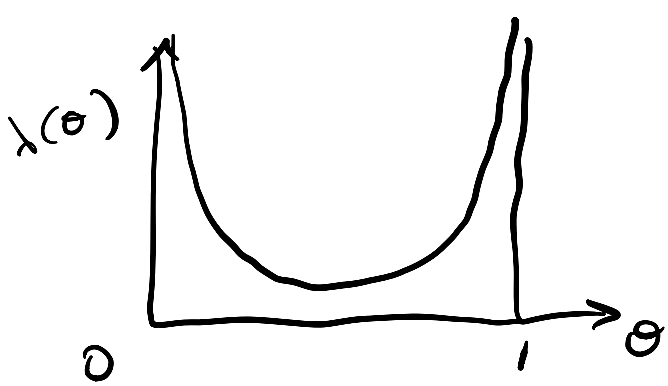
<u>Problem</u>: Flat prior is not flat anymore if we reparameterize $\theta$!

Jeffereys proposed using $\lambda(\theta) \propto_\theta |J(\theta)|^{1/2}$

This is also the Jeffereys prior after any reparameterization

Binomial: Jeffereys' prior is

$$\text{Beta}\left(\tfrac{1}{2}, \tfrac{1}{2}\right) \propto_\theta \theta^{-1/2} (1-\theta)^{-1/2}$$




"Objective"??

# Conjugate Priors

If the posterior is from the same family as the prior, we say the prior is **conjugate** to the likelihood.

Most common in exp. fam.s: Suppose

$$X_i \mid \eta \overset{iid}{\sim} p_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x) \qquad \begin{matrix} \eta \in \Xi \subseteq \mathbb{R}^s \\ i = 1, \dots, n \end{matrix}$$

For carrier $\lambda_0(\eta)$, define $s+1$-dim family:

$$\lambda_{k\mu, k}(\eta) = e^{k\mu'\eta - kA(\eta) - B(k\mu, k)} \lambda_0(\eta)$$

Suff. stat $\begin{pmatrix} \eta \\ -A(\eta) \end{pmatrix} \in \mathbb{R}^{s+1}$

Nat. param. $\begin{pmatrix} k\mu \\ k \end{pmatrix}$

Then

$$\lambda(\eta \mid x_1, \dots, x_n) \propto_\eta \left( \prod_{i=1}^n e^{\eta' T(x_i) - A(\eta)} h(x_i) \right)$$

$$\cdot\ e^{k\mu'\eta - kA(\eta) - B(k\mu, k)} \lambda_0(\eta)$$

$$\propto_\eta e^{(k\mu + \Sigma T(x_i))'\eta - (k+n)A(\eta)} \lambda_0(\eta)$$

$$= \lambda_{k\mu + n\bar{T},\, k+n}(\eta)$$

where $\bar{T}(X) = \frac{1}{n} \sum_{i=1}^n T(x_i)$

Interp.: If we:

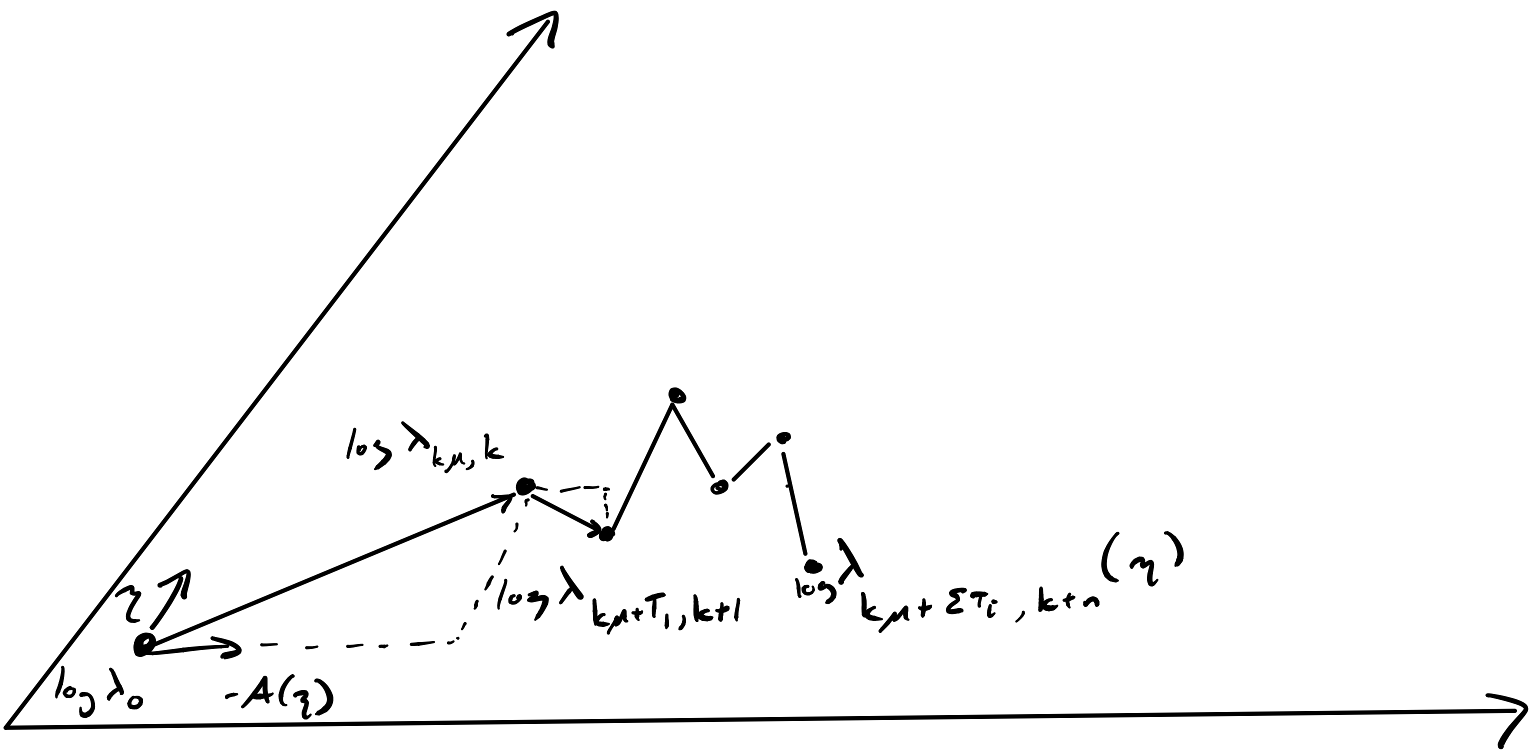
1) Take prior $\lambda_{k\mu, k}$, observe avg. suff. stat. $\bar{T}$ on sample size $n$

OR 2) Take prior $\lambda_0$, observe avg. suff. stat.

a) $\mu$ on sample size $k$ (pseudo-data)

b) $\bar{T}$ on sample size $n$

We get posterior $\lambda_{k\mu + n\bar{T}, k+n}$




Often $\dfrac{k\mu + n\bar{T}}{k+n}$ is Bayes est. for $\mathbb{E}_\eta T$,

then

$$\hat{\mu}_{post} = \bar{T} \cdot \frac{n}{k+n} + \mu \cdot \frac{k}{k+n}$$

UMVUE from data

UMVUE from pseudo data

(If so then $\lambda_0$ not a proper prior. Why?)

# Conjugate Prior Examples

| Likelihood | Prior |
|---|---|
| $X_i \mid \theta \sim \text{Binom}(1, \theta)$ <br> $= \theta^x (1-\theta)^{n-x} \binom{n}{x}$ | $\theta \sim \text{Beta}(\alpha, \beta)$ <br> $= \theta^{\alpha-1} (1-\theta)^{\beta-1} \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ |
| $X_i \mid \theta \sim N(\theta, \sigma^2)$ ($\sigma^2$ known) <br> $= \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-(\theta-x)^2/2\sigma^2}$ | $\theta \sim N(\mu, \tau^2)$ <br> $= \dfrac{1}{\sqrt{2\pi\tau^2}} e^{(\theta-\mu)^2/2\tau^2}$ |
| $X_i \mid \theta \sim \text{Pois}(\theta)$ $x = 0, 1, \ldots$ <br> $= \dfrac{\theta^x e^{-\theta}}{x!}$ | $\theta \sim \text{Gamma}(\nu, s)$ $\theta > 0$ <br> $= \dfrac{1}{\Gamma(\nu) s^2} \theta^{\nu-1} e^{-\theta/s}$ |

## Gamma / Poisson:

$$\lambda(\theta \mid x) \propto_\theta \theta^{\nu - 1 + \Sigma x_i} e^{-(s^{-1} + n)\theta}$$

$$= \text{Gamma}(\nu + \Sigma x_i, (s^{-1} + n)^{-1})$$

$$\Rightarrow k = s^{-1}, \quad \mu = \nu s$$

$$\lambda_0(\theta) = \theta^{-1} \quad \text{(not normalizable)}$$

# Gaussian Sequence Model

$X \sim N_d(\mu, I_d) \qquad \mu \in \mathbb{R}^d$

Jeffereys prior is flat: $\lambda(\mu) \equiv 1$

$$\lambda(\mu \mid x) = N_d(X, I_d)$$

$$\Rightarrow \mathbb{E}[\mu \mid X] = X$$

Reasonable (but inadmissible) estimator: coincides with UMVU, MLE, minimax, ...

What about $\rho^2 = \|\mu\|^2$ ?

$$\mu \sim N_d(X, I_d) \Rightarrow \mathbb{E}\left[\|\mu\|^2 \mid X\right] = \|X\|^2 + d$$

Recall UMVUE was $\|X\|^2 - d$

So $\quad$ Bayes est. $=$ UMVUE $+ \, 2d$

$$\mathrm{MSE}(\theta; \delta_\Lambda) = \mathrm{Var}_\theta(\delta_{\text{UMVU}}) + 4d^2$$

What went wrong?

$$\mu \sim N_d(0, \tau^2 I_d) \Rightarrow \rho^2 \sim \tau^2 \chi^2_d$$

Jeffereys prior takes $\tau^2 \to \infty$

$$\lambda(\rho^2) \propto_{\rho^2} (\rho^2)^{(d-1)/2}$$


"Agnostic" ?

# Advantages of Bayes

Despite difficulties above, Bayes has some major advantages over other approaches:

1) Estimator is defined straightforwardly:

$$\delta_\Lambda(x) = \arg\min_d \int L(\theta, d)\, \lambda(\theta | x)\, d\theta$$

Problem is reduced entirely to computation

May be difficult to compute but in principle we can find it for generic $L, \lambda, \mathcal{P}, g(\theta)$

Don't need to rely on special structure like complete suff. stat, U-estimable $g$, exp. fam., simple $L$

Gives us freedom to:

- Use highly expressive & complex models
- Use the $L$ we actually care about
- Incorporate background subj. matter knowledge

Unparalleled expressive power for systems we know a lot abt. already

2) Appealing optimality property:

Even if we don't "believe" prior, Bayes est. has best avg. case risk

Bayes estimators are usually admissible

3) Detailed output: entire posterior, joint distr. over all parameters

One computation (e.g. large sample from posterior) leads to estimates of any $g(\theta)$ we can think of

.... many more

# Cons

1) Difficulty of choosing $\Lambda$, esp. in high dim.

Avg.-case performance doesn't ensure good performance for the real (distribution of) $\theta$

If we choose $\Lambda$ poorly, we might get no mass near true $\theta$

2) Flipside of ability to specify model in full detail is <u>requirement</u> that we must do so.

e.g. nonparametric estimation of $g(P) = \mathbb{E}_P X_i$

$X_i \overset{iid}{\sim} P$. $\bar{X}$ is UMVUE, natural choice.

To get started on Bayes, must define prior over all distr. on $\mathbb{R}$.

Frequentist approaches let us stay more parsimonious with our assumptions.

.... many more