Bayesian Inference

It's not about true param., it's about true parameter Distribution!

Bayesian View: Data are fixed in a certain way. We know 100% for sure that the data is there, its all the information we have right now! The uncertainty lies in the parameter which produced the data. ~~True~~ Parameter viewed as a (vector-valued) random variable. Underlying subjective interpretation of probability which can be more intuitive than frequentist (∞ repetitions of experiment impossible in reality)

→ We probabilistically express our knowledge/belief about parameter(s) θ given limited amount of obs. data (sample) data are there, don't change & we don't get more (ex. One case is useless (Clinical Test) where $P(B|A) = P(B|\bar{A}) = 0.5$

Inverse Probability, Bayes Theorem: $P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B)} = \frac{P(B|A)\cdot P(A)}{P(B|A)\cdot P(A) + P(B|\bar{A})\cdot P(\bar{A})} = \frac{P(B\cap A)}{P(B\cap A) + P(B\cap \bar{A})}$

prob. of B given A, prior probability of A. Prior $P(A)=0/1$ => Posterior $P(A|B)$ also 0/1

$= P(A)$ if $B \perp A$ (*) more general. ↳ $P(B|A) = P(B) \Leftrightarrow P(B\cap A) = P(B)P(A)$

posterior prob. of A ≥ P(A)! for disjoint $A_1 \dots A_n$: $P(A_i|B) = \frac{P(B|A_i)\cdot P(A_i)}{\sum_{j=1}^{n} P(B|A_j)\,P(A_j)} = \frac{P(B|A_i)\cdot P(A_i)}{\int_\Omega P(B\cap\tilde{A})\,d\tilde{A}}$

$P(B|A) = 1 - P(\bar{B}|A)$!

Bayes Theorem ≙ Update Rule for subjective probability. This here describes the discrete case in simplified form for two events, B as new evidence

data $x = (x_1, \dots, x_n)$ (π)

Continuous parameter: $p(\pi|x) = \frac{f(x|\pi)\cdot p(\pi)}{f(x)}$ ∝ $f(x|\pi)\cdot p(\pi)$

Likelihood of data x / data density; prior distr. / density (parameter)

$f(x) = \int p(\pi|x)\,d\pi = 1$; $f(x)$: marginal data density

posterior distr. / dens.

Normalization constant w.r.t. π; x fixed not random

$f(x|\pi)$ ≙ likelihood if we "fixate" the observed data x

Likelihood ≠ probability cause it has not been weighted yet by prior and not normalized (in case of discrete π) to interval [0,1]

Posterior up to multiplicative constant

$p(\pi|x) \propto f(x|\pi)\cdot p(\pi) \propto f(x|\pi)$ for const. prior (="flat")

(*) Posterior proportional to Likelihood x Prior

- We could also use notation $f(\pi|x)$, $f(\pi)$ instead. Notation above highlight prior & posterior!
- We should actually write $f(x|\Pi=\pi)$ for discrete RV Π bzw. in general $f_{X|\Pi}(x|\pi)$ (*) can define integral as $\int f(x|\omega)p(\omega)d\omega$ w.r.t. π. Integration over param. space inherently Bayesian

(Conditional) Posterior Distr. $\Pi|X=x$

↳ Then: $f(x) = \sum_\pi f(x|\pi)\,P(\Pi=\pi)$ continuous Π: $f(x) = \int_\Pi f(x|\tilde{\pi})\,p(\tilde{\pi})\,d\tilde{\pi}$ We "integrate out parameter"

f(x) is marginal prob. of observing the data, across whole par. space. data model (Likelihood)

- Form posterior by combining prior knowledge/assumptions/beliefs about parameter with new observed data ⇒ using Bayes Rule reduces our uncertainty about θ, personal confidence↑

[Random distribution of estimates for θ; posterior denotes our new current belief about prob./distr. of param]

Subjectivity is implied. There is never complete certainty/perfect knowledge about anything or else there would be no need for statistics in the first place!!

- Bayesian Learning can be sequential. Posterior becomes new prior ⇝ new data .... new posterior

Bayes-Inference formal

1. Set probabilistic model (distribution) for (i.i.d.) data y: $\mathcal{F} = \{f(\cdot;\theta)\,|\,\theta\in\Theta\}$ We only "know" this conditional data density, $f(y|\theta)$?? From there we derive Likelihood $f(y|\theta)$

2. For θ impose a prior distribution that captures our knowledge/beliefs about param: $\theta \sim p(\cdot,\omega)$ often just $\theta \sim p(\theta)$ for simplicity
which may itself depend on additional hyperparameters ω. $L(\theta) = \exp(\ell(\theta))$ ⇝ then posterior $p(\theta|y)$

3. Calculate θs posterior distribution $P_{\theta|y}(\theta|y,\omega) = \frac{\prod_{i=1}^{n} f(y_i;\theta)\,p(\theta,\omega)}{f(y;\omega)}$ with $f(y;\omega) = \int_{\theta\in\Theta} \prod_{i=1}^{n} f(y_i,\theta)\,p(\theta,\omega)\,d\theta$

Conjugate Prior Distribution Leads to posterior from same distr.-family as prior i.e. $P_{\theta|y} \in \mathcal{P} = \{p(\cdot;\theta)\,|\,\omega\in\Omega\}$ $\mathcal{P}$ ≙ prior structure model class for given probability model $\mathcal{F}$. Not all prob. models in $\mathcal{F}$ have such conjugate prior!

List of Conjugate Priors and respective Posteriors for Common Exponential Family Distributions

	Prob. Model / Likelihood	Parameter	Prior Conjugate	Posterior with updated (Prior) hyperparams	Jeffreys Prior
discrete f	$Bern(\pi)$, $Bin(n,\pi)$	$\pi\in(0,1)$	$Beta(\alpha,\beta)$	$Beta(\alpha+y,\ \beta+n-y)$	$p^*(\pi) \propto \pi^{-1/2}(1-\pi)^{-1/2} \triangleq Beta(\frac{1}{2},\frac{1}{2})$ core
	$Geom(\pi)$	$\pi\in(0,1)$	$Beta(\alpha,\beta)$	$Beta(\alpha+n,\ \beta+\sum y_i)$	
	$Neg.Bin(r,\pi)$	$r>0, \pi\in(0,1)$	$Beta(\alpha,\beta)$	$Beta(\alpha+r,\ \beta+\sum y_i)$	
	$Poisson(\lambda)$	$\lambda>0$	$Gamma(\alpha,\beta)$	$Gamma(\alpha+\sum y_i,\ \beta+n)$	$p^*(\lambda) \propto \lambda^{-1/2}$ ≙ special case of $Ga(\alpha,\beta)$ $\alpha=1/2, \beta\to 0$ (improper conjugate prior)
continuous f	$Exp(\lambda)$	$\lambda>0$	$Gamma(\alpha,\beta)$	$Gamma(\alpha+n,\ \beta+\sum y_i)$ $\sum y_i = n\bar{y}$	$p^*(\lambda) \propto \lambda^{-1}$ ≙ special case of $Ga(\alpha,\beta)$ $\alpha\to 0, \beta\to 0$
	$Gamma(known\ \alpha,\beta)$	$\beta>0$	$Ga(a,b)$	$Gamma(a+n\alpha,\ b+\sum y_i)$	
	$N(\mu, known\ \sigma^2)$	$\mu\in\mathbb{R}$	$N(\mu_0,\tau^2)$	$N\left(\left(\frac{\mu_0}{\tau^2}+\frac{n\bar{y}}{\sigma^2}\right)\Big/\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right),\ 1\Big/\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)\right)$	$p^*(\mu) \propto const.$
	$N(known\ \mu, \sigma^2)$	$\sigma^2>0$	$Inv\text{-}Gamma(\alpha,\beta)$	$Inv\text{-}Gamma\left(\alpha+n/2,\ \beta+\frac{1}{2}\sum(y_i-\mu)^2\right)$	$p^*(\sigma^2) \propto 1/\sigma^2$
	$N(\mu,\sigma^2)$	(μ,σ^2)	Normal Inv-Gamma	Normal Inv-Gamma updated	

(connection betw. Bayes and MLE: For fixated i.i.d. data get posterior $p(\theta|y) = \frac{L(\theta)\,p(\theta)}{f(y)}$, Lets define posterior mode $\hat{\theta}_{post,mode} = \underset{\theta\in\Theta}{\operatorname{argmax}}\ p(\theta|y)$

For constant FLAT prior $p(\theta) \Rightarrow p(\theta|y) \propto L(\theta)$ and posterior mode equals Max. Likelihood estimate $\underset{\theta\in\Theta}{\operatorname{argmax}}\ \ell(\theta)$ [Laplace Prior makes Bayesian basically frequentistic (approx.)]

Proof uses penalized Likelihood $\ell(\theta) + \log p(\theta)$ $\hat{\theta}$ refers to $\hat{\theta}_{MLE}$

Bernstein-von-Mises Theorem: For massive sample size n and appropriate prior it holds $\theta \overset{a}{\sim} N(\hat{\theta}, \mathcal{I}^{-1}(\hat{\theta}))$ ⇒ Bayesian Inference corresponds to ML-Inf.

This however says nothing about finite samples in reality! Big Upside of Bayesian is that it still works when CLT and thus MLE fails, e.g. extreme data or just 1 observ. Posterior always calculable (approx.)

everything proport. to this (e.g. scaled) ≙ **Posterior Core** $p(\theta|x) \propto f(x|\theta)\,p(\theta)$

$\Rightarrow p(\theta|x)$ higher and more dense ∧ vs. ∩ $n\uparrow$ => Prior has less influence, ("weak" uninform. prior requires (less) larger n to carve out precise poster. (strong informative)

Posterior $p(\theta|x) = \frac{f(x|\theta)\,p(\theta)}{\int f(x|\tilde{\theta})\,p(\tilde{\theta})\,d\tilde{\theta}}$ $n\uparrow$ decrease Var(Posterior), we are more certain

contains ALL information about θ after observing data

$f(x) = \int f(x|\theta)\,p(\theta)\,d\theta$ Bayesian Denominator / Normalis. Constant(!) Posterior dens. concentrate around more likely θ values (lower var)

Posterior shape reflects shape of prior & Likelihood. More informative (less flat) prior ⇒ Prior influence↑

Bayes-Principle: All conclusions drawn from posterior distr.

- for multidim. θ we look at marginal distr. $\theta_j|x$
- Point estimates (Posterior-EV/Median, Maximum (Postmode) MAP etc.)
- Interval estimators (Credibility Interval) and Tests
- Model comparison
- Prediction

How to choose Prior

On the one hand we want to consider prior knowledge by using an (informative) prior, learn sequentially etc. then data evidence has only tiny influence

On the other hand the prior choice can heavily skew our results. Too informative prior basically determines posterior already!

Obviously it matters how good our prior knowledge is, how confident are in our assumptions (keep in mind that there is always subjectivity involved)
Its not a good idea to artificially make the prior less informative than an honest assessment calls for!

Subjective Prior sometimes useful. Ideally based on reliable expert knowledge. More on Page Two....

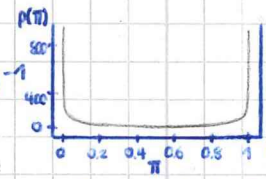
Poisson $\lambda \sim Ga(a,b)$: Choose a,b value s.t. Prior $E(\lambda)=\frac{a}{b}$ matches beliefs vs. for achieve both

Within a specific (!) family of prior distr. we can "measure" informativeness through the variance of param. $Var(\theta)$ given $\theta \overset{prior}{\sim} p(\cdot, w)$ controlled by priors hyperparameters

- Larger Variance $\Longleftrightarrow$ less condensed prior density, meaning less prior information
- Tiny Variance $\Longleftrightarrow$ highly concentrated prior, very informative

Example: Often choose $Ga(a=0.001, b=0.001)$ as prior for Poisson-λ
That leads to high $Var(\lambda) = \frac{a}{b^2} = 1000$ for Prior λ, low informativeness

But what do we do if there is no Prior information at all OR (as often an issue in reality) its not really quantifiable? Then we aim for **Uninformative Prior**
Going back to Billard Ball / Lottery Example: Prior Beta(a=1, b=1) where we interpreted a as "prior Nr. of successes" b as "prior Nr. of failures"
But a=1, b=1 seems to have given some info. Intuitively no information implies a=0, b=0



Start get confused. Its not a valid density so emphasis on rare or frequent occurrence is an illusion! Area ≠ Prob. Mass

$\Longrightarrow$ inputting a=0, b=0 in core of Beta distribution produces the **Haldane Prior** $p(\pi) \propto \pi^{-1}(1-\pi)^{-1}$

Careful! Haldane is an **improper prior** $\int \pi^{-1}(1-\pi)^{-1} d\pi \neq 1$ divergent Normalising $\frac{1}{B(a,b)} = \frac{0!}{0!0!} = 1$ not working here

$\int_0^1 \left(\frac{1}{\pi} - \frac{1}{\pi^2}\right) d\pi = \lim_{a \to 0} \left[\log\pi + 1/\pi\right]_a^1 = \infty$

Its a borderline case derived from Beta(a,b) with both $a,b \to 0$

In general definition of degenerate / improper prior $\int_{\theta \in \Theta} p(\theta; w) \neq 1$ divergent **Many flat priors are improper**

Using Haldane prior $\pi \sim$ "Beta(0,0)" which is not a real (Beta-) Distr. still functions a conjugate Prior $p(\pi|x) \propto \pi^{-1}(1-\pi)^{-1} \cdot \pi^x (1-\pi)^{n-x} = \pi^{x-1}(1-\pi)^{n-x-1}$ (if $x>0, n-x>0$ $\hat{=}$ Beta(x, n-x))

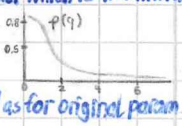
Even with improper Prior, the Posterior is usually well defined ✓ (Convergence to proper Posterior)
For $n \to \infty$ and finite dimensional param. space always the case that Posterior integrates to one.
Norming by data: Sample n large enough, Likelihood (usually sharp locally concentrated) "strong enough" to make $\int f(x|\tilde{\theta}) p(\tilde{\theta}) d\tilde{\theta} < \infty$ finite Bayesian Denominator
Analogy: Flashlight illuminates finite spot in infinite fog, data cut out finite volume from ∞ area

Interesting: Non-informative Prior might be edge case of conjugate Prior

Laplace Prior / Laplace-Principle: Without knowledge should initially assume equal probability for all elementary events $P(w_i) = 1/|\Omega|$ discrete uniform distr.
In general defined as $p(\theta) \propto$ constant $\Rightarrow$ flat prior, but not necessarily totally uninformative as explained above with $Beta(1,1) \Leftrightarrow U(0,1)$ prior
continuous uniform dist. for constrained θ smt. refer to $p(\theta) \propto$ const. as "uniform distr. on $\mathbb{R}$"
If parameter space unconstrained then Laplace Prior is improper

Problem: Laplace-Prior can transformed Parameter can lead to different Prior & Post.-Family (unlike Haldane-Prior which is transf.invariant)

Example] Instead of π use odds $q = \frac{\pi}{1-\pi}$ (density transform with $\pi \sim U(0,1)$) $p(q) = \frac{1}{(1+q)^2}$ $f_q(q) = f_\pi(\pi(q)) \cdot \left|\frac{\partial \pi(q)}{\partial q}\right| = \frac{1}{(1+q)^2} \cdot 1$ $\pi(q) = \frac{q}{q+1}$, $\frac{\partial \pi(q)}{\partial q} = \frac{1}{q+1} - \frac{q}{(q+1)^2} = \frac{1}{(q+1)^2}$



- Transformed Prior not uniform, not constant flat. Thus Posterior looks different, no longer just influenced by Likelihood as for original param
- This also showcases that this Laplace-Prior does contain a tiny bit of information, transformed Prior is partly informative

Jeffreys Prior non-informative Prior, invariant to Parameter-Transf. (bijective ϕ) $p^*(\phi(\theta))$ is again Jeffrey Prior for $\phi(\theta)$ and therefore uninformative

$p^*(\theta) \propto \mathcal{I}^{1/2}(\theta)$ Jeffreys prior defined as (multipl. of) sq. root of Fisher Information Matrix Jeffrey Prior can, but doesn't have to be improper

$\mathcal{I}(\theta) = -E_x\left(\frac{\partial^2 \log f(X|\theta)}{\partial\theta\,\partial\theta^T}\right)$ in 1D $= E_x\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]$ Interpretable as expected curvature of log-Likelihood

Frequentist: integration over data x !!

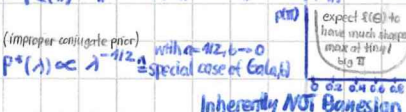
under certain conditions: $\mathcal{I}(\theta) = Var_x(s(x;\theta)) = E_x[s(x)^2]$ as $E(s(x))=0$
score-fct. $\frac{\partial}{\partial\theta}\log f(x|\theta)$

Fisher Info indicates degree of param. information in a random sample of data, param. viewed as fixed

Jeffreys Prior for Binomial Distribution: $X \sim B(\pi)$ | $f(x|\pi) = \pi^x (1-\pi)^{1-x}$ | $\log(f(x|\pi)) = x\log\pi + (1-x)\log(1-\pi)$, $s_\pi(x) = \frac{x}{\pi} - \frac{1-x}{1-\pi}$

$\mathcal{I}(\pi) = E_x(s^2(x)) = \left(\frac{1}{\pi} - \frac{1-1}{1-\pi}\right)^2 \pi + \left(\frac{0}{\pi} - \frac{1-0}{1-\pi}\right)^2 (1-\pi) = \frac{1}{\pi} + \frac{1}{1-\pi} = \frac{1}{\pi(1-\pi)}$ (discrete distr. $x \in \{0,1\}$)

$p^*(\pi) \propto \sqrt{\mathcal{I}(\pi)} = \pi^{-1/2}(1-\pi)^{-1/2} = \pi^{(1/2)-1}(1-\pi)^{(1/2)-1} \hat{=} Beta(\frac{1}{2},\frac{1}{2})$ density core



(would mean $p(\theta) = \text{const.}$ assuming just one event!) $X \sim Exp(\lambda)$ | $f(x|\lambda) = \lambda\exp(-\lambda x)$

Jeffreys Prior for Exp. Distribution: $\log(f(x|\lambda)) = \log\lambda - \lambda x$, $s_\lambda(x) = \frac{1}{\lambda} - x$

$\mathcal{I}(\lambda) = Var(s_\lambda) = Var(\frac{1}{\lambda} - x) = 0 + Var(x) = 1/\lambda^2 \rightsquigarrow p^*(\lambda) \propto \sqrt{1/\lambda^2} = \lambda^{-1} \hat{=}$ special case of $Ga(a,b)$ with $a \to 0, b \to 0$ (improper, edge case conjugate)

Poisson: $p^*(\lambda) \propto \lambda^{-1/2} \hat{=}$ special case of $Ga(a,b)$ with $a \to 1/2, b \to 0$ (improper conjugate prior)

Inherently NOT Bayesian **Empirical Bayes**

Guidelines / Approaches 1. Prior-Hyperparam. choice should closely reflect Prior Knowledge about θ (subjective! maybe hard...)
2. Prior should be as uninformative / flat as possible (esp. if no knowledge) and/or be independant of parametrification
Sometimes we prefer rel. "flat" but proper Priors over improper Laplace or improper Jeffrey (choose extremely high Var Prior)
3. Estimate Prior-Hyperparams from Data: $\bar{f}(y|w) = \int_{\theta\in\Theta} f(y;\theta)\, p(\theta|w)\, d\theta$ (pretend w still fixed) marginal data distr. given w
↳ new Likelihood $\bar{\mathcal{L}}(w) = \prod_i \int_{\theta\in\Theta} f(y_i;\theta) p(\theta|w) d\theta$

Subjective Priors

One issue: Expert knowledge/ strong beliefs typically not given with a specific param. density

⇓

Solution: • For discrete Params should be pretty easy to formulate Probabilities that add up to one

• For continuos Params we could discretize it into intervals, find parametric density that "matches" knowledge or use sth. like Maximum Entropy Prior

Parametric subjective Prior i) Choose matching param distribution (symmetric gaussian, skewed gamma, beta etc.)
In the easiest case its the conjugate Prior

ii) Set Prior Hyperparameters so that Prior Knowledge is captured with unique theoretical Hyperparam. combination
(mean, variance or more robust statistics like median, quantiles & perhaps credibility interval)

If there are several Hyperparams we need more than just a single information to get unique Hyperparams
Lack of additional information? Make reasonable assumption or use information criterion under constraint of given knowledge

Example • Prior knowledge on Bernoulli π is mean = 0.3 and 95% credibility interval $[0.1, 0.5]$ i.e. $\mathbb{P}(0.1 < \pi < 0.5) = 0.95$
Conjugate Prior $\pi \sim \text{Beta}(a,b)$ with EV $\frac{a}{a+b}$ ⇒ map it to mean: $\frac{a}{a+b} = 0.3 \Leftrightarrow b = \frac{7}{3}a$
Additional info about Cred. Interv: Numerically solve Min. deviation of $(\text{pbeta}(0.5, a, \frac{7}{3}a) - \text{pbeta}(0.1, a, \frac{7}{3}a))$ from 0.95 regarding e.g. L2 Loss
⇒ In R we can use optimize(f, interval) to search interval $c(0, 1000)$ which yields $a \approx 5.45 \rightsquigarrow b \approx 12.72$

• Same as above but only info about median = 0.36. This gets much more complex, also because Median(Beta) has no general form and needs to be approx.
We could numerically solve $\text{Med}(\text{Beta}(a,b)) = 0.36$ under $a + b = \delta$ with δ measuring degree of confidence in knowledge (low $\delta = 2$, medium $\delta = 10$, high $\delta = 50$)
(Another option is a very conservative Minimal Information prior that maximizes Entropy under Median-constraint? Applying MEP requires using $\hat{E}(\theta) = \text{median}_{\text{robust}}$)

Maximum Entropy Prior Basis idea is to choose minimally informative prior distr. that still uses Prior Information

Definition Shannon-Entropy: $H(\theta) = \mathbb{E}[-\log f(\theta)] = -\int_\Theta \log(p(\theta))\, p(\theta)\, d\theta = -\sum_\Theta \log(p(\theta))\, p(\theta)$ for discrete θ
(regarding prior)

Prior Knowledge appears in form of certain of certain moments $\mathbb{E}(g_k(\theta))$ of random variable, here parameter θ

⇓

There exists a Max. Entropy Prior (MEP) $p_{MEP}(\theta) = \underbrace{V}_{\text{normalisation const.}} \cdot \exp\left(\sum_{k=1}^{K} \underbrace{c_k}_{\text{coeff.}} \underbrace{g_k(\theta)}_{\theta^k}\right)$ which preserves the moments
$\mathbb{E}(g_1(\theta)) = m_1(\theta) = \mathbb{E}(\theta)$ etc.