Bayesian Inference , Priors

HASCO-2012 Group 1 – Babar Ali, Stefano De Nicola Uppsala, Cambridge

Bayesian vs Frequentist

- We are interested in using a given sample of data to make inferences about a probabilistic model

- In frequentist statistics, probability is interpreted as the frequency of the outcome of a repeatable experiment. **parameter estimation**
- Frequentist statistics provides the usual tools for reporting the outcome of an experiment ***objectively***, without needing to incorporate prior beliefs concerning the parameter being measured or the theory being tested

- In Bayesian statistics, the interpretation of probability is more general and includes ***degree of belief*** called **subjective** probability
- Probability density function (p.d.f.) for a parameter expresses one's state of knowledge about where its true value lies
- they require the ***prior*** p.d.f. as input for the parameters, *i.e.*, the degree of belief about the parameters values before carrying out the measurement

- *" a quantitative measure of strength of our anticipation, founded on the said knowledge, that the event comes true "* (D' Agostini 2003)
- *" ….Probability of an event to be understood as given state of knowledge."*

A and B are two propositions

$$0 \le P(A) \le 1$$
$$P(\Omega) = 1 \quad \textit{(tautology)}$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cap B) = P(A \,|\, B)\, P(B) = P(B \,|\, A)\, P(A)$$
$$P(A) + P(\overline{A}) = 1$$

Set of propositions

$$\begin{aligned} \cup_i H_j &= \Omega \\ H_j \cap H_k &= \emptyset \quad \text{if} \quad j \neq k \\ \sum_j P(H_j) &= 1 \\ P(H_j) &= \sum_i P(E_i, H_j) \\ P(H_j) &= \sum_i P(H_j \,|\, E_i)\, P(E_i) \\ P(E_i) &= \sum_j P(E_i, H_j) \\ P(E_i) &= \sum_j P(E_i \,|\, H_j)\, P(H_j) \end{aligned}$$

In physics experiments, deal with measurement
That are discrete or continuous

For discrete *x*, expression *p(x)* **Probability Function**
For continuous , expression *p(x)* dx **Probability density function**

Bayesian Inference

- Everything we do based on what we know about the physical world
- Conclusions about hypotheses will be based on our general background knowledge
- Dependence of probability on the state of background information, I

$$P(A, B \,|\, I) = P(A \,|\, B, I)\, P(B \,|\, I) = P(B \,|\, A, I)\, P(A \,|\, I)$$

$$\frac{P(H_j \,|\, E_i, I)}{P(H_j \,|\, I)} = \frac{P(E_i \,|\, H_j, I)}{P(E_i \,|\, I)} \qquad P(H_j \,|\, E_i, I) = \frac{P(E_i \,|\, H_j, I)\, P(H_j \,|\, I)}{P(E_i \,|\, I)}$$

Bayes' theorem

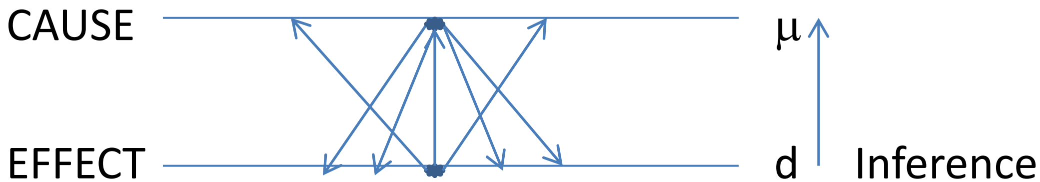
- A logical rule to update our beliefs on the basis of new conditions

$$P(H_j \,|\, E_i, I) = \frac{P(E_i \,|\, H_j, I)\, P(H_j \,|\, I)}{\sum_j P(E_i \,|\, H_j, I)\, P(H_j \,|\, I)}$$

$$P(H_j \,|\, E_i, I) \propto P(E_i \,|\, H_j, I)\, P(H_j \,|\, I)$$

posterior $\propto$ likelihood $\times$ prior

Bayesian Inference, Priors — Gaussian Model



Response signal d from detector

True value μ

Standard deviation σ

- Known as *normal* often assumed that errors are normally distributed according to function

$$p(d\,|\,\mu, I) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(d-\mu)^2}{2\,\sigma^2}\right] \qquad p(\mu\,|\,d, I) = \frac{\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(d-\mu)^2}{2\,\sigma^2}\right] p(\mu\,|\,I)}{\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(d-\mu)^2}{2\,\sigma^2}\right] p(\mu\,|\,I)\,\mathrm{d}\mu}$$

- Considering all values of μ equally likely over a large interval

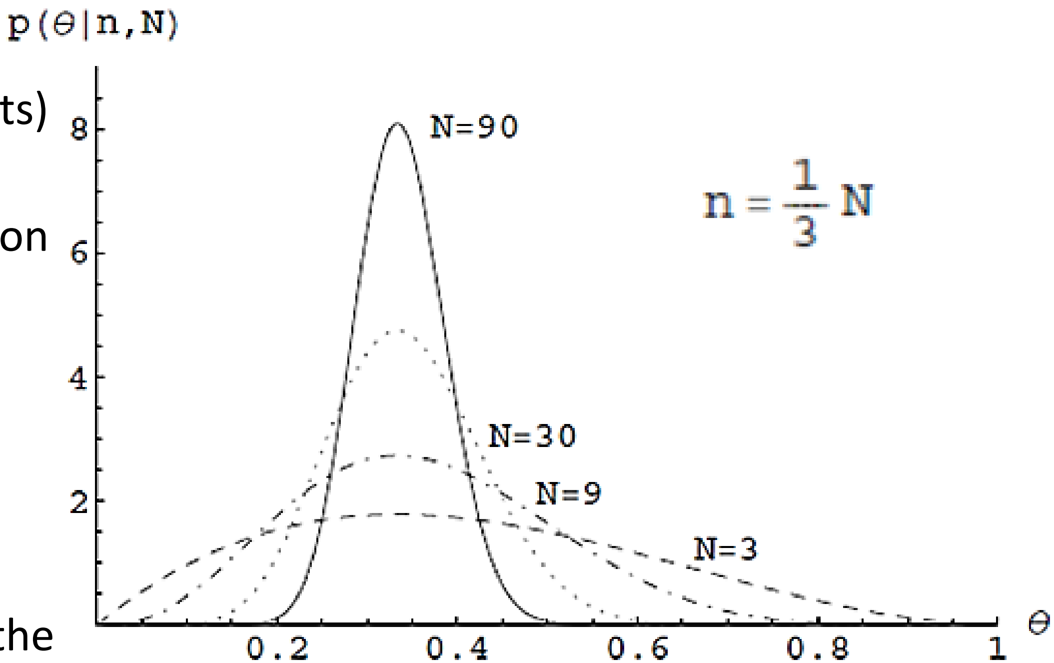
$$p(\mu\,|\,d, I) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(\mu-d)^2}{2\,\sigma^2}\right]$$

- If the quantity is constrained in physical region $\mu \geq 0$, while d falls outside it or at its edge
- Prior (step function)

- In a large class of experiments, the observations consist of counts (events)
- The # of counts described probabilistically by Binomial or Poisson model
- Inference about the efficiency of detector
- Branching ratio in particle decay

- The binomial distribution describes the probability of randomly obtained **n** events (success) in **N** independent trials
- Assume the probability θ that the event will happen

$$p(n\,|\,\theta, N) = \frac{(N+1)!}{n!\,(N-n)!}\,\theta^n (1-\theta)^{N-n}$$



$$p(\theta\,|\,n, N, I) = \frac{\theta^n\,(1-\theta)^{N-n}}{\int_0^1 \theta^n\,(1-\theta)^{N-n}\,d\theta} = \frac{(N+1)!}{n!\,(N-n)!}\,\theta^n\,(1-\theta)^{N-n}$$

$$E(\theta) = \frac{n+1}{N+2}$$

$$\sigma^2(\theta) = \frac{(n+1)(N-n+1)}{(N+3)(N+2)^2} = \frac{E(\theta)\;(1-E(\theta))}{N+3}$$

$$\theta_m = \frac{n}{N}$$

- The Poisson distribution gives the probability of observing **n** counts in fixed time interval

- # of counts to be observed is λ

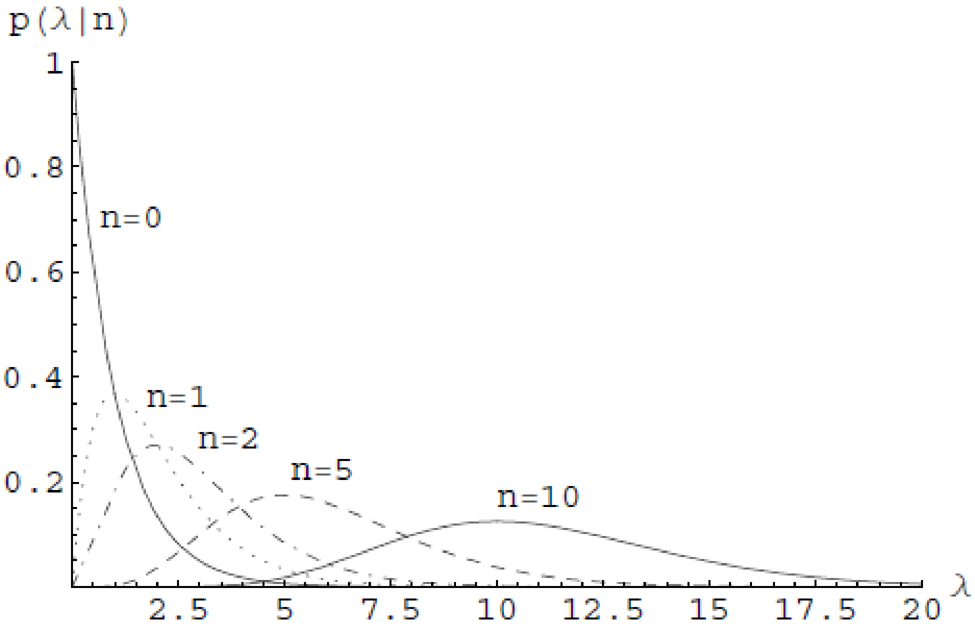
$$p(n\,|\,\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

- To infer λ from **n** counts observed

- By using uniform prior p(λ | I) for λ

$$p(\lambda\,|\,n,I) = \frac{\frac{\lambda^n\,e^{-\lambda}}{n!}}{\int_0^\infty \frac{\lambda^n\,e^{-\lambda}}{n!}\,\mathrm{d}\lambda} = \frac{\lambda^n\,e^{-\lambda}}{n!}$$

- The expectation and variance of λ is n+1, while the most probable value is λm = n



- Choice of priors is crucial in *non-likelihood-dominated situations*, i.e. outcomes not assumed to be equally likely

- Priors are often left implicit or dealt with inappropriately

- Good choice of prior can be a significant advantage of a Bayesian approach over a frequentist one

- A good prior function should:

    - Model the current information on the underlying PDF

    - Be mathematically handy!

Improper priors

- When choosing a function to model a prior, in most cases detailed values (or normalization) don't matter

- In fact, the choice of "*improper*" priors can be extremely advantageous

- Choose a family of functions with suitable adjustable parameters

- Test effect of chosen prior on posterior (*sensitivity tests*)

- Need to model information realistically while keeping the calculation feasible

- The idea arises of choosing a prior such that the posterior is of the same functional family: *Conjugate Priors*

*e.g. Gaussian likelihood * Gaussian prior → Gaussian posterior*

As an expression of the form

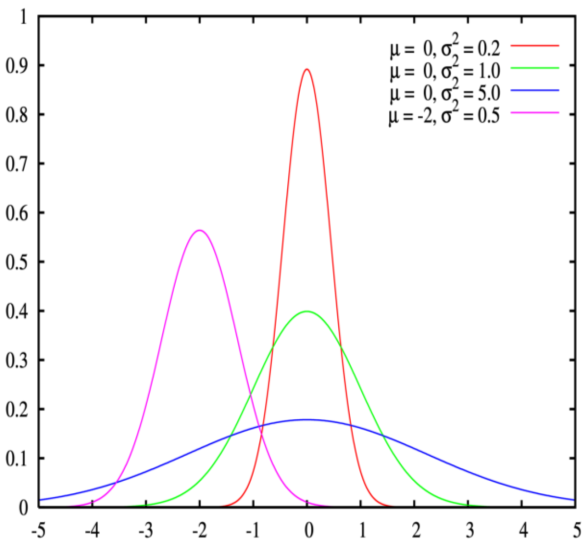
$$\mathrm{K}\exp\left[\frac{-(x_1-\mu)^2}{2\sigma_1^2}-\frac{(x_2-\mu)^2}{2\sigma_2^2}\right]$$

Can always be casted into the form

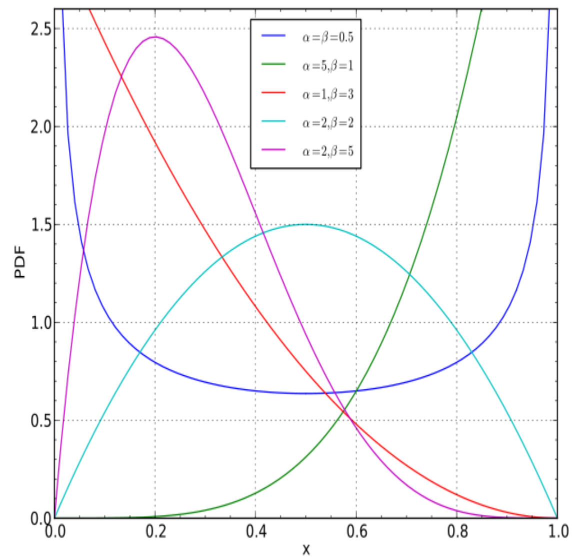
$$K'\exp\left[\frac{-(x'-\mu)^2}{2\sigma'^2}\right]$$

- The conjugated priors for important PDFs can be summarized as follows

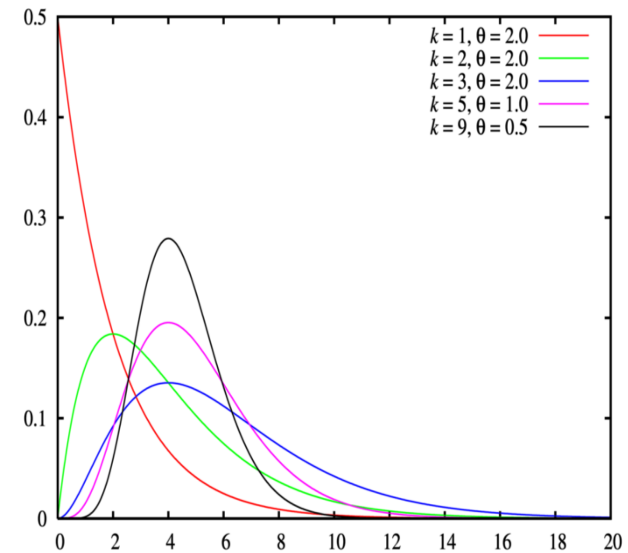
Likelihood	Conjugate prior	Posterior
Binomial (N,p)	Beta (r,s)	Beta (r+N, s+N-n)
Poisson (λ)	Gamma (r,s)	Gamma(r+n, s+1)
Normal (μ, σ)	Normal (μ_0, σ_0)	Normal (μ_1, σ_1)



Normal distribution



Beta distribution



Gamma distribution

- For some applications, useful to determine prior from general principles, keeping "subjective" factor to a minimum

- Some of these rules are obtained by requiring transformation invariance

-> Translational invariance

Requiring that $p(a)da = p(a')da'$ (where a' = a+b)

$$\longrightarrow p(a) = const.$$ *(flat prior)*

-> Scale invariance

Requiring that $p(a)da = \beta p(\beta a)da$ for a scaling factor β

$$\longrightarrow p(a) \propto \frac{1}{a}$$ (*Jeffreys' prior*)

Other family of approaches based on *Maximum Entropy Principle (MEP)*

- Choose prior that maximizes Shannon-Jaynes information entropy, defined as

$$S = -\sum_{i}^{n} (p_i \ln p_i)$$

subjected to what we assume to know about the PDF.

Same principles are recovered. E.g. :

- If there are no constraints, S is maximized by Jeffreys' prior ($p(a) \propto \frac{1}{a}$)

Yet another approach: *Reference priors*

- Maximize *Kullback–Leibler divergence* → amount of information from posterior (i.e. "least informative" prior)
- Most used to tackle multivariate problems, where other priors (Jeffreys') can result in unwanted behaviours

Summary

- Two school of statistics: Bayesian and Frequentist
- The concept of Bayes' theorem
- Bayesian Inference for different PDFs
- Suitable choice of priors
- Conjugated priors
- Obtaining priors from general principles
- Alternative views on the choice of priors
  (e.g. maximum entropy, reference priors)