Parameter Estimation and Information Theory

1 Oct 2021

Reed Essick

Learning Objectives:

- Review how the posterior is constructed from Bayes Theorem
- Interpreting the posterior: median vs. mean
  symmetric vs. "HPD" credible regions
  correct coverage
- Approximations for the Posterior
  - Fisher Information Matrix (Geometric Info. Theory)
  - Laplace Approximation
- Sampling Techniques: when to use different algorithms
  - Rejection Sampling
  - Inverse Transform Sampling
  - Gibbs Samplers
  - Metropolis-Hastings Algorithm
  - Monte Carlo Summation, Importance Sampling

Let's begin by reviewing the components that make up the posterior for the signal (s) given the data (h) and an underlying model (H)

$$p(s|h,H) = \frac{p(h|s,H)\,p(s|H)}{p(h|H)}$$

$p(s|h,H)$: posterior

$p(h|s,H)$: likelihood

$p(s|H)$: prior

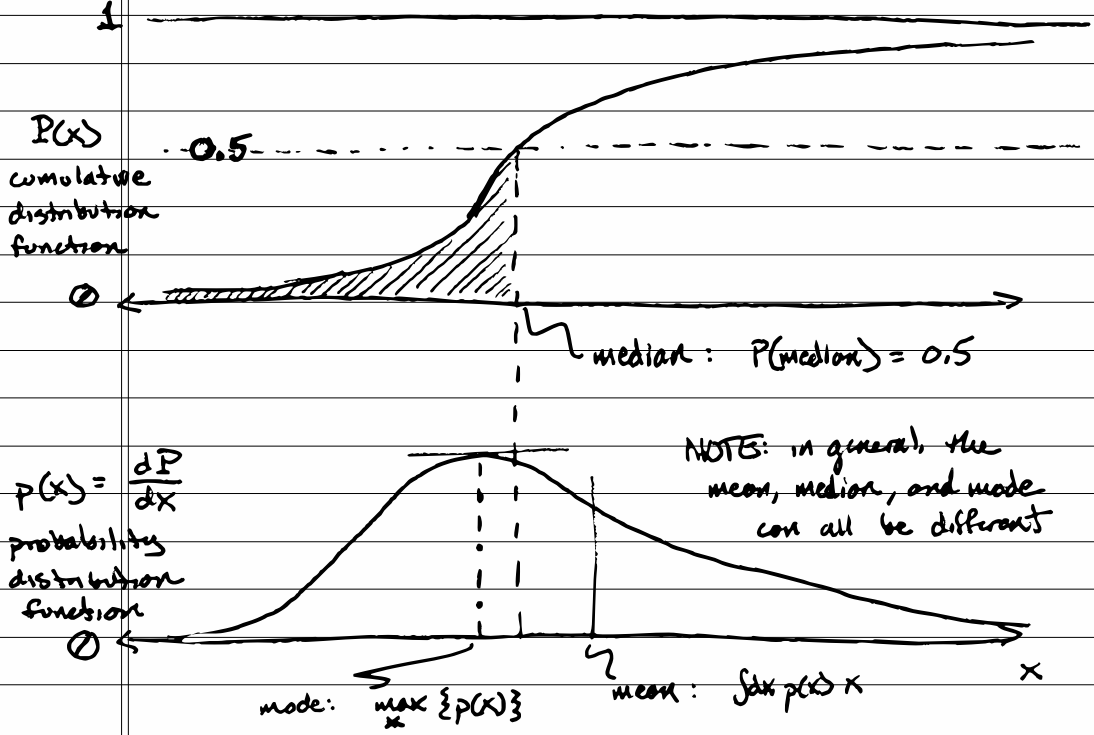
$p(h|H)$: evidence

While the evidence (marginal likelihood) is important in general, it does not affect the shape of the posterior and is often neglected when drawing samples from the posterior. We will therefore primarily use

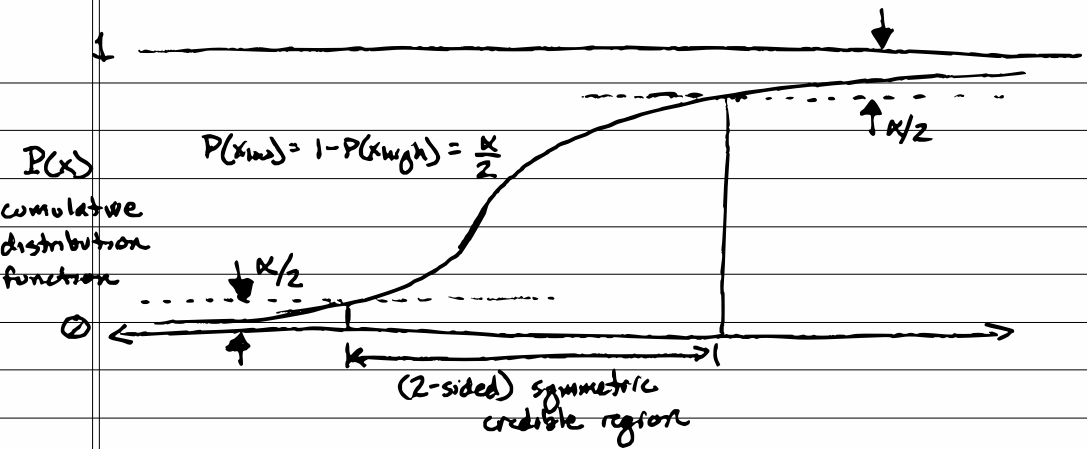
$$p(s|h, H) \propto p(h|s, H)\, p(s|H)$$

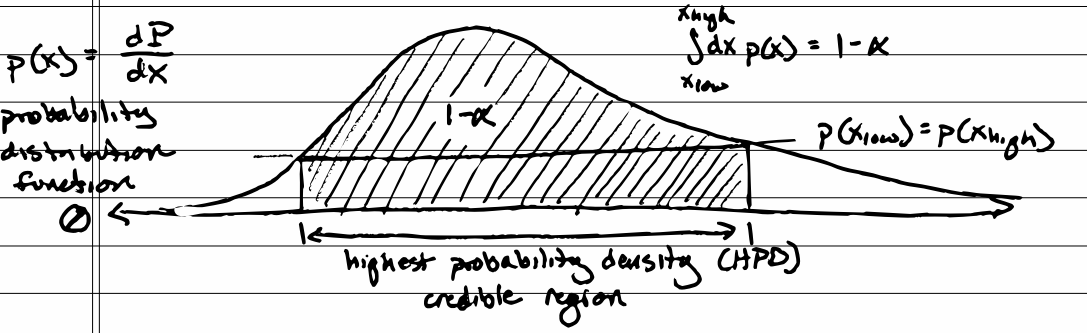
This shows how the posterior is a combination of the likelihood (what the observed data suggests) and the prior (what we believed before the experiment).

Let's assume we can compute the posterior efficiently for the moment & focus on how we should interpret it. We begin w/ a simple (cartoon) univariate distribution

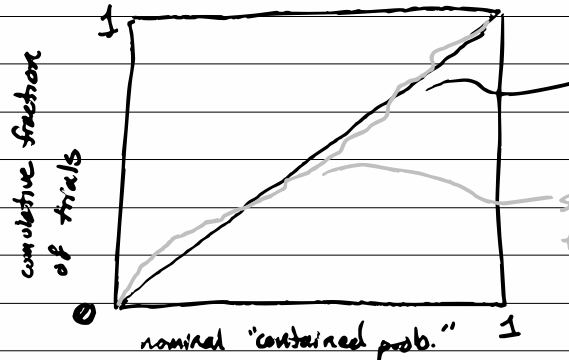


See Feldman+Cousins, PRD 57, 3873 (1998)





credible regions are meant capture the uncertainty in the distribution. For a posterior's CRs to be meaningful, we want the "true parameters" to fall within the X% CR in X% of trials (different possible noise realizations). This is gauged w/ a coverage plot (or p-p or q-q plot)

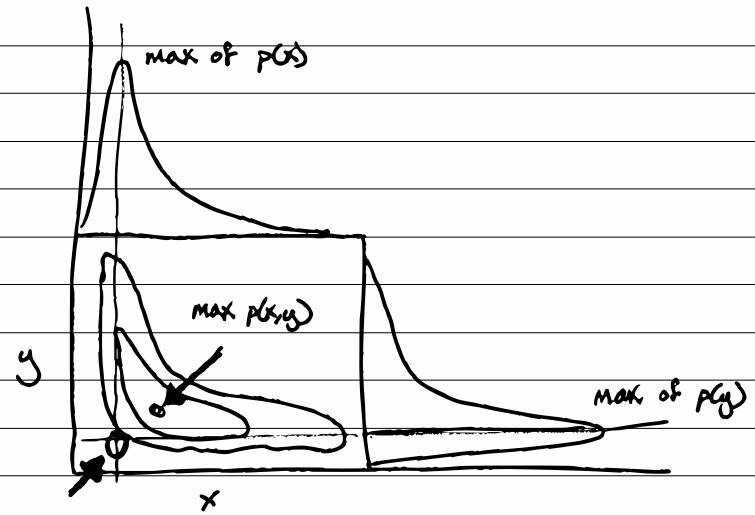


$\int dx\, p(x)\, \Theta(x : p(x) \geq p(x_\alpha))$

Coverage measures the consistency between the distribution's accuracy and precision. However, it does not guarantee that the posterior is particularly accurate (or precise).

Note that the mode of the posterior (aka max. a posteriori or "MAP") may be different than the max. likelihood point. The maximum likelihood point is an unbiased estimator in the frequentist sense. The MAP will be biased in this sense (in general) because of the prior. However, in almost all situations we are not only interested in "point-estimates" but in the full posterior distribution, which requires the specification of a prior.

Note also that the maxima of marginal distributions do not (in general) correspond to the maxima of the joint distribution. The "best fit" point is then not the set of parameters that maximize marginal distribs. separately.



In general, the posterior may be difficult to compute or deal with analytically. A common approximation is the LAPLACE APPROXIMATION, which is a good approx for loud signals (big signal-to-noise ratios) when the likelihood is sharply peaked.

Consider the likelihood for a parametrized model $s(\theta)$. Expand this around the max. likelihood parameters (θ_{MLE})

$$\log p(h|\theta) = \log p(h|\theta_{MLE}) + \partial_i \log p\Big|_{\theta_{MLE}} (\theta-\theta_{MLE})_i + \frac{1}{2}\partial_i\partial_j \log p\Big|_{\theta_{MLE}} (\theta-\theta_{MLE})_i (\theta-\theta_{MLE})_j + \mathcal{O}((\theta-\theta_{MLE})^3)$$

by assumption, $\partial_i \log p\Big|_{\theta_{MLE}} = 0$ so that the leading order nontrivial term is

$$\log p(h|\theta) \sim \frac{1}{2}\left(\partial_i\partial_j \log p\Big|_{\theta_{MLE}}\right)(\theta-\theta_{MLE})_i (\theta-\theta_{MLE})_j$$

This is just a Gaussian distribution with inverse covariance matrix given by

$$\Gamma_{ij} = (Cov^{-1})_{ij} = -\partial_i\partial_j \log p(h|\theta)\Big|_{\theta_{MLE}}$$

this is often called the: Fisher Information Matrix

Higher order terms in the approximation typically scale as $(1/SNR)$, which is why we can neglect them for loud signals.

The Fisher Info Matrix tells us the correlations between parameters. It also provides a lower-bound on the uncertainty; the real covariance is greater or equal to the inverse Fisher matrix (Cramer-Rao Bound).

Note also that the Fisher Info. Matrix is often defined as an expected value w/r/t the data

$$\Gamma_{ij} = -\int dx\, p(x|\theta)\left(\partial_i\partial_j \log p(x|\theta)\big|_\theta\right)$$

$$= \int dx\, p(x|\theta)\left(\partial_i \log p(x|\theta)\big|_\theta\right)\left(\partial_j \log p(x|\theta)\big|_\theta\right)$$

via integration by parts (assuming vanishing boundary terms)

This definition comes from information theory, but many authors neglect the integral, approximating $p(x|\theta) \approx \delta(\theta - \theta_0)$ for high SNR signals.

FURTHER READING

Geometric Information theory

definition of:

- probabilistic manifold (for parametric models)
- divergences (Kullback-Leibler) between prob. distribs.
- information (Shannon)
- Fisher Info. Matrix as a metric on the probabilistic manifold

[INFORMATION THEORY]

consider 2 distribs: $p(x)$ & $q(x)$.

A directional measure of how different these are is the Kullback-Leibler Divergence

$$D_{KL}(p||q) \equiv \int dx\, p(x) \log\left(\frac{p(x)}{q(x)}\right) \geq 0$$

where $\lim_{p\to 0} p \log p = 0$ by convention.

FISHER INFO MATRIX defines a Riemannian metric on the manifold of prob. distrib. induced by the KL Divergence

If we consider q that are similar to p & expand

$$D_{KL}(p||p+\delta p) = \int dx\, p(x)(\log p - \log(q))$$

$$= \int dx\, p(x)\left(\log p - \left[\log p + \frac{\partial \log p}{\partial \theta}\cdot \delta\theta + \frac{1}{2}\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\log p\, \delta\theta_i \delta\theta_j + \cdots\right]\right)$$

$$= -\frac{1}{2}\left[\int dx\, p(x)\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\log p(x)\right]\delta\theta_i\delta\theta_j$$

since $\int dx\, p(x)\frac{\partial \log p}{\partial\theta} = \int dx \frac{\partial p}{\partial\theta} = \frac{\partial}{\partial\theta}\left(\int dx\, p\right) = \frac{\partial}{\partial\theta}(1) = 0$

WHY DOES FISHER SHOW UP IN LAPLACE APPROX?

The likelihood measures the relative probability of getting the observed data given a signal model. As we change the signal model, we are de facto generating different distributions over the observed data. We are interested in quantifying how different those distributions are as a function of the change in the signal model, & the KL Divergence → Fisher Info. Matrix is the natural way to compare distributions.

[SAMPLING FROM TARGET DISTRIBUTIONS]

We can also represent the posterior by a set of samples drawn from the distribution. This works because we can either

- estimate the posterior via a histogram or other fancier techniques, such as kernel density estimates (KDE) or renormalizing flows
- estimate moments via Monte-Carlo summation

$$\int dx\, p(x)\, f(x) \approx \frac{1}{N} \sum_i^N f(x_i) \quad | \quad x_i \sim p(x)$$

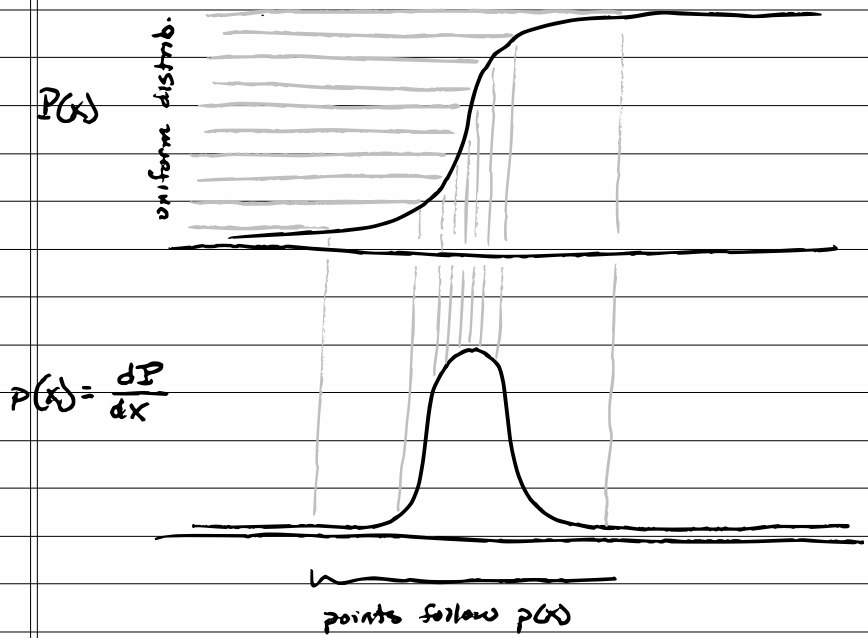
Many techniques exist to sample from distributions. You often will want the simplest that still fits your problem

INVERSE TRANSFORM SAMPLING
(only works for 1D distributions)

One can draw samples from an arbitrary 1D distribution with knowledge of the inverse of the cumulative distribution function: $P^{-1}: [0,1] \to \mathbb{R}$

This works by: ① drawing a random number uniformly distributed w/in $[0,1]$: $r \sim U([0,1])$
② computing $x = P^{-1}(r)$
③ repeating

This is easy to conceptualize with the following cartoon



REJECTION SAMPLING

this works when you can compute your posterior in terms of a distribution that is simpler to sample from.

Assume that our target distrib is $p(x)$ but we can sample from $q(x)$ easily. We can then repeatedly sample from $q(x)$ and keep only those samples that satisfy

$$\frac{p(x_i)}{q(x_i)} \geq r_i$$

where r_i is a (uniformly distrib) random number generated separately for each trial. Importantly, the range of r_i needs to be @ least as large as $\max_x \{p(x)/q(x)\}$

We note that instead of "rejecting" samples, we can also record weights for each ($w_i = p(x_i)/q(x_i)$) and use these w/in our representation of the target distrib (see Importance Sampling).

Note, rejection (and Monte-Carlo) sampling are extremely general but they can be (very!) inefficient.

METROPOLIS-HASTINGS ALGORITHM
(fancy rejection sampling)

This algorithm works like rejection sampling, except the proposal distribution "q" is allowed to depend on the position of the previous sample (generating a MARKOV CHAIN). That is, given the current location "x_i" we generate a new proposed location

$$x_{i+1} \sim q(x \mid x_i)$$

and accept the new position if

$$\min\left\{ \frac{p(x_{i+1})}{p(x_i)} \left(\frac{q(x_i \mid x_{i+1})}{q(x_{i+1} \mid x_i)} \right), 1 \right\} \geq u_i$$

u_i: uniformly distrib. random variable between $[0, 1]$

if we do not accept the proposal, we set $x_{i+1} = x_i$.
Iterating builds up a "chain" of Monte Carlo samples.

We note that MCMC chains contain correlations btwn consecutive samples (induced by the fact that q depends on the previous sample). One therefore must take care to measure the autocorrelation w/in the chains & downsample accordingly (retain only 1 sample out of several autocorrelation lengths). This guarantees the resulting samples are independent fair draws from the posterior.

[FURTHER READING]
estimating autocorrelation length of MCMC chains
"burn-in" and MCMC initialization
parallel tempering

— IMPORTANCE SAMPLING —

Finally, we note that it can be costly to draw a large number of samples. One may want to re-use existing samples to approximate what would be obtained by drawing new samples from a different distrib.

Consider the basic Monte-Carlo approx. to an integral

$$\int dx\, p(x) F(x) \approx \frac{1}{N} \sum_i^N F(x_i) \quad | \quad x_i \sim p(x)$$

Now, if we wish to approx. $\int dx\, q(x) F(x)$, we note

$$\int dx\, q(x) F(x) = \int dx\, p(x) \left(\frac{q(x)}{p(x)} F(x) \right) \approx \frac{1}{N} \sum \frac{q(x_i)}{p(x_i)} F(x_i)$$

as long as the support of p entirely contains the support of q. We effectively just reweigh samples by their importance under q.