Lecture 8

Bayesian Statistical Inference

- The basic premise of Bayesian analysis is that probability statements are not limited to data, but can be made about model parameters & models themselves.

• Note that classical & Bayesian stats share a crucial ingredient: the likelihood.

• Yet, the $\mathcal{L}$ cannot be interpreted as the probability of model parameters, since that is not even a valid concept in classical stats.

• The Bayesian paradigm extends the concept of the $\mathcal{L}$ by adding extra "prior" information to the analysis and assigning probabilities to all parameters & models.

- Bayesian analysis is based on Bayes' rule:

$$\boxed{p(M/D) = \frac{P(D|M)\,p(M)}{p(D)}}$$

where $D \to$ data, $M \to$ model.

• In words, we quantify the rule for

"combining an initial belief with new data to arrive at an improved belief" and says that "improved belief" is proportional to the product of "initial belief" and probability that "initial belief" generated the observed data.

- We can explicitly write in the prior information into Bayes' rule:

$$p(M,\vec{\theta}|D,I) = \frac{p(D|M,\vec{\theta},I)\,p(M,\vec{\theta}|I)}{p(D|I)}$$

where

$p(M,\vec{\theta}|D,I)$ → posterior pdf for model M with θ_p parameters, given data D, & information I.

$p(D|M,\vec{\theta},I)$ → likelihood of data given M, $\vec{\theta}$ & I.

$p(M,\vec{\theta}|I)$ → joint prior for model M & $\vec{\theta}$ given I.

Note $$p(M,\vec{\theta}|I) = p(\vec{\theta}|M,I)\,p(M|I)$$

only specify this in parameter estimation problems.

$p(D|I)$ → the probability of data, or prior predictive probability for D.

↳ normalizing constant for parameter estimation; SUPER IMPORTANT for model selection.

Summary of Bayesian workflow

① Formulate the data likelihood, $p(D|M,I)$.

② Make a choice of prior, $p(\vec{\theta}|M,I)$, which incorporates all other knowledge, *that is <u>not used</u> when computing the likelihood.*

③ Determine the posterior, $p(M|D,I)$, using Bayes' theorem.

- Usually computationally expensive in multi-dimensional problems.
- $p(D|I)$ can be ignored through a proper re-normalization of $p(M|D,I)$.

④ A "true Bayesian result" would be to report the full posterior, $p(M|D,I)$, or a suitably marginalized posterior focusing on the parameters of interest.

- In practice, it may be more prudent to report/find the parameters $\vec{\theta}_{MAP}$ that maximize $p(M|D,I)$, i.e. the maximum a posteriori (MAP) estimate.
- Another estimate might be the posterior mean,

$$\bar{\theta} = \int \theta \, p(\theta|D) \, d\theta$$

⑤ Quantify uncertainty on the parameters via credible regions.

- Can be obtained analytically or via numerical techniques like bootstrapping from the posterior.

⑥ Hypothesis testing as needed to make other conclusions about the model or hypothesis.

Bayesian priors

– Broadly, two types:

1] Informative priors: leverage prior information/results (not based on current data).

2] Non(weakly)-informative priors: used when no previous information available.

– Usually, we want our inference to be data-driven or data-dominated, so weakly informative priors are more common in literature.

– Rules/principles for assigning uninformative priors:

① Principle of indifference:

- A set of basic, mutually exclusive

possibilities need to be assigned equal probability.

- This results in a flat, or uniform, prior

$$p(\theta|I) \propto C$$

→ note that since $\int p(\theta|I)\, d\theta = \infty$ here, this is called an "improper" prior.

→ Such a prior can be fine as long as the resulting posterior pdf is well-defined.

② Principle of consistency:

- Based on transformation groups, it demands that the prior for a location parameter should not change with translation of the co-ordinate system.
- This also yields a flat prior.
- Note that for a scale parameter, like σ, this gives us $p(\theta|I) \propto \frac{1}{\theta}$, which implies a uniform prior on $\ln(\theta)$.

③ Principle of maximum entropy:-

- Entropy measures the information content of a pdf.
- The idea here is that by maximizing entropy over a suitable set of pdfs, we find the distribution that is least informative (given constraints).