Bayesian model selection and parameter estimation
Bayesian inference

Aim: use available data to

- Construct probability density distributions for parameters associated with these hypotheses
  → Parameter estimation
- Evaluate which out of several hypotheses is the most likely
  → Model selection

Do this while making explicit all extraneous assumptions
Inductive logic

Propositions (i.e. statements, events) denoted by uppercase letters, e.g. A, B, C, ..., X

Boolean algebra:

- Conjunction: *A and B are both true*
  \[ AB \text{ or } A \land B \]

- Disjunction: *At least one of A or B is true*
  \[ A + B \text{ or } A \lor B \]

- Negation: *A is false*
  \[ \overline{A} \text{ or } \neg A \]

- Implication: *From A follows B*
  \[ A \rightarrow B \text{ or } A \Rightarrow B \]
Probabilities for propositions

- Useful to view statements as sets which are subsets of a “Universe”
  - Conjunction: intersection of sets
    \[ AB \quad \text{or} \quad A \land B \]
  - Disjunction: union of sets
    \[ A + B \quad \text{or} \quad A \lor B \]
  - Negation: complement within Universe
    \[ \bar{A} \quad \text{or} \quad \neg A \]

- Each of these sets have a probability associated with them
  - If \( A \subset B \) then \( p(A) \leq p(B) \)
  - If \( A \) and \( B \) disjoint then \( p(A \cup B) = p(A) + p(B) \)
  - The Universe has probability 1, so that e.g.
    \[ p(A) + p(\bar{A}) = 1 \]
Conditional probability

- **Conditional probability:** \( p(A|B) \equiv \frac{p(A \cap B)}{p(B)} \)

- **Product rule:**
  \[
p(A, B) = p(A \cap B) = p(A|B) \cdot p(B)
  \]

- It is customary to explicitly denote probabilities being conditional on “all background information we have”: \( p(A|I), p(B|I), \ldots \)

- All essential formulae unaffected, e.g. product rule:
  \[
p(A, B|I) = p(A|B, I) \cdot P(B|I)
  \]

- From the product rule follows Bayes’ theorem:
  \[
p(A|B, I) = \frac{p(B|A, I) \cdot p(A|I)}{p(B|I)}
  \]
Marginalization

Note that for any $A$ and $B$,

$$A \cap B \quad \text{and} \quad A \cap \bar{B}$$

are disjoint sets whose union is $A$, so

$$p(A|I) = p(A, B|I) + p(A, \bar{B}|I)$$

Consider sets $\{B_k\}$ such that

- They are disjoint: $B_k \cap B_l = \emptyset$, $k \neq l$
- They are exhaustive: $\bigcup_k B_k$ is the Universe, or

$$\sum_k p(B_k|I) = 1$$

Then

$$p(A|I) = \sum_k p(A, B_k|I)$$

**Marginalization rule**
Marginalization over a continuous variable

Consider the proposition

“The continuous variable $x$ has the value $\alpha$”

Then not necessarily a well-defined meaning of probability $p(x = \alpha|I)$

Instead assign probabilities to finite intervals:

$$p(x_1 \leq x \leq x_2|I) = \int_{x_1}^{x_2} \text{pdf}(x)dx$$

where “pdf” is the probability density function

- Exhaustiveness written as

$$\int_{x_{\text{min}}}^{x_{\text{max}}} \text{pdf}(x)dx = 1$$

Marginalization for continuous variables:

$$p(A) = \int_{x_{\text{min}}}^{x_{\text{max}}} \text{pdf}(A, x)dx$$
Parameter estimation

- Experiment performed, data $d$ collected
- Parameter $\theta$ being measured
- Consider a model $H$ that allows to calculate probability of getting data $d$ if parameter $\theta$ is known ("generative model")
  - Can calculate the likelihood $p(d|\theta, H, I)$
- What is wanted is instead posterior probability of $\theta$, $p(\theta|d, H, I)$
- Use Bayes' theorem:
  $$p(\theta|d, H, I) = \frac{p(d|\theta, H, I)p(\theta|H, I)}{p(d|H, I)}$$
  - "Prior" $p(\theta|H, I)$ is our knowledge of $\theta$ before experiment
  - "Evidence" $p(d|H, I)$ doesn't depend on $\theta$, ignore for now
  $$p(\theta|d, H, I) \propto p(d|\theta, H, I)p(\theta|H, I)$$
Parameter estimation

\[ p(\theta|d, H, I) \propto p(d|\theta, H, I)p(\theta|H, I) \]

- Posterior is likelihood weighted by prior
- Conclusions drawn based on:
  - Information available before experiment
  - Experimental data obtained
- Can extend to more parameters: joint posterior \( p(\theta_1, \ldots, \theta_N|d, H, I) \)
- If we want posterior distribution just for variable \( \theta_1 \), \( p(\theta_1|d, H, I) \), then we marginalize:

\[
p(\theta_1|d, H, I) = \int_{\theta_2}^{\theta_2^{\text{max}}} \cdots \int_{\theta_N}^{\theta_N^{\text{max}}} p(\theta_1, \ldots, \theta_N|d, H, I) \, d\theta_2 \cdots d\theta_N
\]
Mean of a 1D posterior:
\[ \mu = E[\theta] \]
\[ = \int_{\theta_{\text{min}}}^{\theta_{\text{max}}} \theta p(\theta|d, H, I) \, d\theta \]

Variance of a 1D posterior:
\[ \sigma^2 = E[(\theta - \mu)^2] \]
\[ = \int_{\theta_{\text{min}}}^{\theta_{\text{max}}} (\theta - \mu)^2 p(\theta|d, H, I) \, d\theta \]

Means for \( N \) variables:
\[ \mu_i = E[\theta_i] \]
\[ = \int_{\theta_{1_{\text{min}}}^{\text{max}}}^{\theta_{1_{\text{max}}}} \ldots \int_{\theta_{N_{\text{min}}}^{\text{max}}}^{\theta_{N_{\text{max}}}} \theta_i p(\theta_1, \ldots, \theta_N|d, H, I) \, d\theta_1 \ldots d\theta_N \]

Covariance matrix:
\[ \Sigma_{ij} \equiv E[(\theta_i - \mu_i)(\theta_j - \mu_j)] \]
\[ = \int_{\theta_{1_{\text{min}}}^{\text{max}}}^{\theta_{1_{\text{max}}}} \ldots \int_{\theta_{N_{\text{min}}}^{\text{max}}}^{\theta_{N_{\text{max}}}} (\theta_i - \mu_i)(\theta_j - \mu_j) p(\theta_1, \ldots, \theta_N|d, H, I) \, d\theta_1 \ldots d\theta_N \]
**Confidence interval** is the smallest interval within whose limits a fraction $\gamma$ of the posterior is contained:

$$
\gamma = \int_{\theta^{\text{lo}}}^{\theta^{\text{hi}}} p(\theta | d, H, I) d\theta
$$

where $\theta^{\text{hi}} - \theta^{\text{lo}}$ is minimal

In most literature $\gamma$ is taken to be 0.68 or 0.95, roughly corresponding to 1-sigma and 2-sigma intervals of Gaussian distribution

**Multi-dimensional confidence intervals:**

$$
\gamma_{\theta_1} = \int_{\theta_{1}^{\text{lo}}}^{\theta_{1}^{\text{hi}}} p(\theta_1 | d, H, I) d\theta_1
$$

$$
= \int_{\theta_{1}^{\text{lo}}}^{\theta_{1}^{\text{hi}}} \int_{\theta_{2}^{\text{min}}}^{\theta_{2}^{\text{max}}} \cdots \int_{\theta_{N}^{\text{min}}}^{\theta_{N}^{\text{max}}} p(\theta_1, \ldots, \theta_N | d, H, I) d\theta_1 \cdots d\theta_N
$$
Hypothesis testing

- Estimating parameters is possible if generative model known
- If we want to compare possible generative models, e.g. X, Y: calculate posterior probabilities \( p(X|d, I) \) and \( p(Y|d, I) \)

Bayes' theorem:

\[
p(X|d, I) = \frac{p(d|X, I)p(X|I)}{p(d|I)}
\]

Compute odds ratio

\[
O_X^Y \equiv \frac{p(X|d, I)}{p(Y|d, I)} = \frac{p(d|X, I)p(X|I)}{p(d|Y, I)p(Y|I)}
\]

where factors of \( p(d|I) \) have canceled out

- \( \frac{p(X|I)}{p(Y|I)} \) ratio of prior odds
- \( \frac{p(d|X, I)}{p(d|Y, I)} \) ratio of evidences, or Bayes factor \( B_Y^X = \frac{p(d|X, I)}{p(d|Y, I)} \)
Hypothesis testing

- Hypotheses usually have parameters associated with them
- Bayes theorem relating posterior to likelihood:

\[ p(\theta|d, H, I) = \frac{p(d|θ, H, I)p(θ|H, I)}{p(d|H, I)} \]

or

\[ p(\theta|d, H, I)p(d|H, I) = p(d|θ, H, I)p(θ|H, I) \]

- Marginalize both sides over parameter(s):

\[ \int p(\theta|d, H, I)p(d|H, I)dθ = \int p(d|θ, H, I)p(θ|H, I)dθ \]

Note that \( p(d|H, I) \) independent of parameter(s), and posterior \( p(\theta|d, H, I) \) normalized by definition, hence left hand side:

\[ \int p(\theta|d, H, I)p(d|H, I)dθ = p(d|H, I) \int p(\theta|d, H, I)dθ = p(d|H, I) \]

Therefore evidence is given by

\[ p(d|H, I) = \int p(d|θ, H, I)p(θ|H, I)dθ \]
Hypothesis testing

Odds ratio

\[ O_Y^X \equiv \frac{p(X|d, I)}{p(Y|d, I)} = \frac{p(d|X, I) p(X|I)}{p(d|Y, I) p(Y|I)} \]

Bayes factor

\[ B_Y^X = \frac{p(d|X, I)}{p(d|Y, I)} \]

Marginalized evidences e.g.

\[ p(d|X, I) = \int p(d|\theta, X, I)p(\theta|X, I)d\theta \]

Hypotheses can have arbitrary number of free parameters

- Does model that fits data the best give the highest evidence?
- If so, model with more parameters would give highest evidence even if incorrect!
Occam's razor

For simplicity, compare two generative hypotheses:

- $X$ has no free parameters
- $Y$ has one free parameter, $\lambda$

Will $Y$ automatically be favored over $X$?

Odds ratio \( O_Y^X = \frac{p(d|X, I) \cdot p(X|I)}{p(d|Y, I) \cdot p(Y|I)} \)

Evidence for $X$ is straightforward, but for $Y$:

\[
p(d|Y, I) = \int p(d|\lambda, Y, I)p(\lambda|Y, I)d\lambda
\]

Assume flat prior for $\lambda \in [\lambda_{\text{min}}, \lambda_{\text{max}}]$:

\[
p(\lambda|Y, I) = \frac{1}{\lambda_{\text{max}} - \lambda_{\text{min}}}, \quad \text{for } \lambda_{\text{min}} \leq \lambda \leq \lambda_{\text{max}}
\]
Occam's razor

Evidence for $Y$:

$$p(d|Y, I) = \int p(d|\lambda, Y, I)p(\lambda|Y, I)d\lambda$$

Flat prior:

$$p(\lambda|Y, I) = \frac{1}{\lambda_{\text{max}} - \lambda_{\text{min}}}, \quad \text{for } \lambda_{\text{min}} \leq \lambda \leq \lambda_{\text{max}}$$

For definiteness, assume likelihood of the form

$$p(d|\lambda, Y, I) = p(d|\lambda_0, Y, I) \exp \left[-\frac{(\lambda - \lambda_0)^2}{2\sigma^2_\lambda}\right]$$

Evidence for $Y$:

$$p(d|Y, I) = \int p(d|\lambda, Y, I)p(\lambda|Y, I)d\lambda$$

$$= \int \frac{1}{\lambda_{\text{max}} - \lambda_{\text{min}}} p(d|\lambda_0, Y, I) \exp \left[-\frac{(\lambda - \lambda_0)^2}{2\sigma^2_\lambda}\right] d\lambda$$

$$= \frac{p(d|\lambda_0, Y, I)}{\lambda_{\text{max}} - \lambda_{\text{min}}} \int \exp \left[-\frac{(\lambda - \lambda_0)^2}{2\sigma^2_\lambda}\right] d\lambda$$

$$= p(d|\lambda_0, Y, I) \frac{\sigma_\lambda \sqrt{2\pi}}{\lambda_{\text{max}} - \lambda_{\text{min}}}.$$
Occam's razor

Evidence for $Y$:

$$p(d|Y, I) = p(d|\lambda_0, Y, I) \frac{\sigma_\lambda \sqrt{2\pi}}{\lambda_{\text{max}} - \lambda_{\text{min}}}$$

Hence odds ratio becomes:

$$O_Y^X = \frac{p(X|I)}{p(Y|I)} \frac{p(d|X, I)}{p(d|\lambda_0, Y, I)} \frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\sigma_\lambda \sqrt{2\pi}}$$

where

- $p(X|I)/p(Y|I)$ ratio of prior odds; can be set to 1 in this example
- $p(d|X, I)/p(d|\lambda_0, Y, I)$ just compares best fits; will usually be $< 1$
- $(\lambda_{\text{max}} - \lambda_{\text{min}})/(\sigma_\lambda \sqrt{2\pi})$ penalizes $Y$ if experimental uncertainty on $\lambda$ much smaller than prior range

- Will tend to be the case if $\lambda$ not needed!

Occam's Razor:

“It is vain to do with more what can be done with fewer”
Nested sampling

Parameter estimation requires computing the posterior density distribution from likelihood and prior using Bayes' theorem:

\[ p(\theta|d, H, I) = \frac{p(d|\theta, H, I)p(\theta|H, I)}{p(d|H, I)} \]

Often the parameter space has high dimensionality (e.g. 15 for quasi-circular binary inspiral), making it computationally challenging to map out the likelihood.

Similarly calculation of evidence integral over high-dimensional space:

\[ p(d|H, I) = \int d^N \theta \ p(d|\theta, H, I)p(\theta|H, I) \]

\[ = \int d^N \theta \ L(\theta)\pi(\theta), \]

Efficient way of obtaining both: nested sampling
Nested sampling: basic idea

\[ p(d|H, I) = \int d^N \theta \ p(d|\theta, H, I)p(\theta|H, I) = \int d^N \theta \ L(\theta)\pi(\theta), \]

Nested sampling computes the evidence by rewriting the above integration in terms of a single scalar called *prior mass* \( X \)

“Fraction of volume with likelihood greater than \( \lambda \)”

Mathematically:

\[ X(\lambda) \equiv \int \int \cdots \int_{L(\theta) > \lambda} \pi(\theta) d^N \theta \]

Element of prior mass: \( dX = \pi(\theta)d^N \theta \)

Since prior is normalized, \( X \in [0, 1] \)

- Lower bound \( X = 0 \):
  
  surface within which no higher likelihood; \( \lambda = L_{\text{max}} \)

- Upper bound \( X = 1 \):
  
  surface within which all points higher likelihood; \( \lambda = L_{\text{min}} \)
Nested sampling: basic idea

\[ p(d|H, I) = \int d^N \theta \, p(d|\theta, H, I)p(\theta|H, I) \]
\[ = \int d^N \theta \, L(\theta)\pi(\theta), \]

Rewrite as

\[ Z = \int \int \cdots \int L(\theta)\pi(\theta)d^N \theta \]
\[ = \int \tilde{L}(X)dX. \]

Posterior obtained trivially from

\[ \tilde{P}(X) = \frac{\tilde{L}(X)}{Z} \]

Idea behind nested sampling: construct the function \( \tilde{L}(X) \) by progressively finding locations in parameter space with higher likelihood and associated progressively smaller prior mass

- Then use above formulae for evidence, posterior
Nested sampling: schematically
Nested sampling: the algorithm

- Drop $M$ samples across parameter space, sampled from the prior. These are called “live points”
  - Each has likelihood associated with it
  - Associated with volume s.t. likelihood lowest at the surface
  - Uniformly sampled in prior mass between 0 and 1
- Discard live point with lowest likelihood $L_0$, i.e. highest prior mass $X_0$
  - Replace by new live point, sampled from the prior, which has higher likelihood
  - New point with lowest likelihood $L_1$ must have $X_1 < X_0$
  - Statistically assign value for $X_1$
- Repeat the step above
Nested sampling: the algorithm

- Having discarded the old lowest-likelihood point with prior mass $X_0$, how do we statistically assign a prior mass $X_1$ to the new lowest-likelihood point?

- Probability that the surface with highest prior mass is at $X = \chi$ is joint probability that none of the samples have prior mass $> \chi$

  $\quad P(X_i < \chi) = \prod_{i=1}^{M} \int_{0}^{\chi} dX_i = \prod_{i=1}^{M} \chi = \chi^M$

- Probability density that highest of $M$ samples has prior mass $\chi$

  $\quad P(\chi, M) = M\chi^{M-1}$

- Define shrinkage ratio between new and old highest prior mass:

  $\quad t = X_1/X_0$

  This has same probability density:

  $\quad P(t, M) = Mt^{M-1}$

- Hence we assign $X_1$ by drawing a shrinkage ratio from the above distribution
**Nested sampling**

- At first step: set $X = 1$
- At $k^{th}$ iteration: live point with largest prior mass has
  \[ X_k = \prod_{j=1}^{k} t_k \]

Recall distribution of shrinkage ratios:
\[ P(t, M) = Mt^{M-1} \]
Mean and standard deviation of $\log(t)$:
\[ \log t = (-1 \pm 1)/M \]

Hence $\log(X_k)$ has mean and stdev
\[ \log X_k = (k \pm \sqrt{k})/M \]
Hence mean values go like
\[ X_k = \exp(-k/M) \]
- Very quickly reaches prior mass where likelihood is largest
- Errors decrease exponentially
- Larger number of live points is better
Nested sampling: termination condition

- No obvious choice for ending the sampling process
  - Use practical guidelines

Estimate *information* as function of evidence and likelihood:

\[
\mathcal{H} = \int P(X) \ln(P(X)) \, dX \\
\approx \sum_k \frac{L_k}{Z} \ln \frac{L_k}{Z} \Delta X_k,
\]

Terminate when \( X = e^{-\mathcal{H}} \)

Or, can estimate amount of evidence yet to be accumulated and compare with evidence already accumulated

Terminate when \( L_{\text{max}} X_{\text{cur}} < \alpha Z_{\text{cur}} \) where \( \alpha \) is user-specified
Nested sampling: accuracy

- Take termination condition
  \[ X = e^{-\mathcal{H}} \]

- Means go like
  \[ X_k = \exp(-k/M) \]

  “Terminate when count \( k \) exceeds \( M\mathcal{H} \)”

- Evidence:
  \[ Z = \int \tilde{L}(X)dX \approx \sum_k L_k \Delta X_k \]

  Recall
  \[ \log X_k = (k \pm \sqrt{k})/M \]

  Hence uncertainty on the evidence:
  \[ \Delta \log Z = \sqrt{\mathcal{H}/M} \]

- In gravitational-wave applications, with a few thousand live points this is typically \( O(10^{-1}) \) whereas for detectable signal \( \log Z = O(10^2) \)
Application to gravitational waves

- Compute evidence for hypothesis that there is a signal in the data, $\mathcal{H}_S$:
  \[ p(d|\mathcal{H}_S, I) = Z = \int \tilde{L}(X) \, dX \approx \sum_k L_k \Delta X_k \]

- Compute posterior density function for signal parameters, $\theta$:
  \[ p(\theta|d, \mathcal{H}_S, I) \approx \frac{L_k}{Z} \Delta X_k \]

- In the case of a coalescing binary (black holes and/or neutron stars):
  \[ \theta = \{t_c, \varphi_c, m_1, m_2, \vec{S}_1, \vec{S}_2, \theta, \phi, \iota, \psi, D\} \]

- Posterior density for a given parameter, e.g. $m_1$:
  - Use some smooth interpolation of the above posterior density
  - Marginalize over all other parameters

  \[ p(\theta_1|d, H, I) = \int_{\theta_2}^{\theta_2^{\text{max}}} \ldots \int_{\theta_N}^{\theta_N^{\text{max}}} p(\theta_1, \ldots, \theta_N|d, H, I) \, d\theta_2 \ldots d\theta_N \]
Application to gravitational waves: GW150914

Application to gravitational waves: GW150914

Application to gravitational waves: GW150914

Application to gravitational waves: GW150914

Masses, spins, distances

https://arxiv.org/abs/1606.04856
Testing general relativity

- Solar system tests of GR:
  - Perihelion precession of Mercury [weak, static field]
  - Deflection of starlight by the Sun [weak, static field]
  - Shapiro time delay [weak, static field]
  - Gravity Probe B
    * Geodetic effect [weak, static field]
    * Frame dragging effect [weak, stationary field]

- Binary neutron star observations (e.g. Hulse-Taylor):
  - Most penetrating tests of GR up to the present time
  - But: dissipative dynamics only to leading order (quadrupole)
  - Dynamical self-interaction of spacetime not being probed

No test of the strong-field dynamics of spacetime
Ideal laboratories: coalescing binary neutron stars and black holes
Any deviations from GR in the shape of the wave?

Orbital phase during inspiral as a function of (ever increasing) orbital speed:

\[ \Phi(\nu) = \left( \frac{\nu}{c} \right)^{-5} \sum_{n=0}^{\infty} \left[ \varphi_n + \varphi_n^{(l)} \ln \left( \frac{\nu}{c} \right) \right] \left( \frac{\nu}{c} \right)^n \]

- Up to factor of 2, this is also the GW signal during inspiral
- In general relativity, the coefficients \( \varphi_n \) and \( \varphi_n^{(l)} \) are known functions
- Can we put bounds on possible deviations from the GR predictions?
Any deviations from GR in the shape of the wave?

\[ \delta \dot{\varphi} \]

-1 0 1 2 3 4

PN order

\((v/c)^0\) \((v/c)^1\) \((v/c)^2\) \((v/c)^3\) \((v/c)^4\) \((v/c)^6\) \((v/c)^7\)

GW150914

J0737-3039A/B
Any deviations from GR in the shape of the wave?
Any deviations from GR in the shape of the wave?
Any deviations from GR in the shape of the wave?

"intermediate"

Any deviations from GR in the shape of the wave?
Does the graviton have mass?

\[ E^2 = p^2 c^2 + m_g^2 c^4 \]

\[ \delta \Phi(f) = -\frac{\pi D c}{\lambda_g^2 (1+z)} f^{-1} \]

\[ \lambda_g = \frac{\hbar}{m_g c} \]

\[ m_g < 10^{-22} \text{ eV}/c^2 \]
