PARAMETER ESTIMATION IN GRAVITATIONAL-WAVE ASTRONOMY

Matthew Pitkin
PyCBC Inference Workshop

14 May 2019

University of Glasgow

- Introduction
  - Bayes' theorem
  - Marginalisation

- Gravitational-wave inference
  - the likelihood function
  - the priors
  - parameter estimation examples
  - model selection
  - hierarchical inference

There are many textbooks on statistics (and Bayesian statistics in particular), but three I can recommend are:
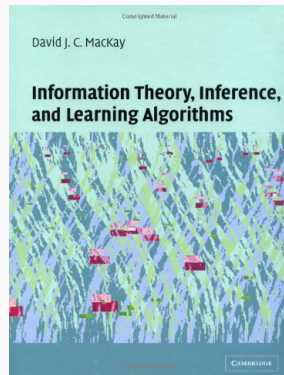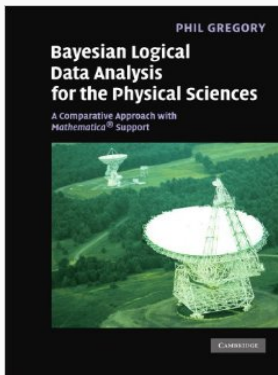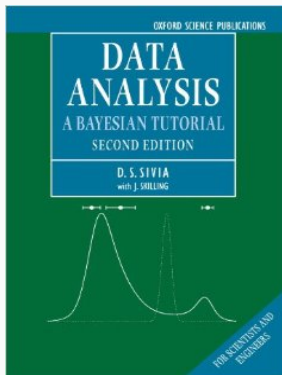
Some useful papers to read are:

- The `LALInference` paper (Veitch et al., 2015).
- The PyCBC Inference paper (Biwer et al., 2019).
- The Bilby paper (Ashton et al., 2019).

From the product rule comes

$$P(B|A, I) = \frac{P(A|B, I)P(B|I)}{P(A|I)}.$$

This is Bayes' theorem.

From now on we will stick to using $P(A, B|I)$ to denoted the probability of *A and B*, where we have also explicitly added the conditioning on background information *I*.

Bayes theorem can be cast in terms of a **model** and some observations, or **data**. It tells us how to update our degree of belief about our model based on new data.

$$P(\text{model}|\text{data}, I) = \frac{\overbrace{P(\text{data}|\text{model}, I)}^{\text{Likelihood}} \overbrace{p(\text{model}|I)}^{\text{Prior}}}{\underbrace{P(\text{data}|I)}_{\text{Evidence}}}$$

$$\underbrace{\phantom{P(\text{model}|\text{data}, I)}}_{\text{Posterior}}$$

We can often calculate the these terms (e.g., analytically or numerically on a computer).

- **Prior**: what we knew, or our degree of belief, about our model before taking data
- **Likelihood**: the influence of the data in updating our degree of belief
- **Evidence**: the "*evidence*" for the data, or the likelihood for the data *marginalised* over the model (we'll explore this later, but at the moment note it as a constant normalisation factor for the posterior)
- **Posterior**: our new degree of belief about our model in light of the data

How can we determine if a coin is fair[1]? We can consider a large number of contiguous propositions over the range in which the bias weighting H of the coin might lie:

· $H = 0$ coin produces a tail every time
· $H = 1$ coin produces a head every time
· $H = 0.5$ is a 'fair' coin with 50:50 chance of head or tail
· continuum of probabilities $0 \leq H \leq 1$

Given some data (an observed number of coin tosses) we can assess how much we believe each of these propositions (e.g. $0 \leq H < 0.01$, $0.01 \leq H < 0.02$, and so on) to be true, e.g.

$$\text{Prob}(0 \leq H < 0.01 | d)$$

In the limiting case where our propositions each lie in the infinitesimal range dH our inference about the bias weighting is summarised by the pdf for the conditional probability $p(H|d, I)$, i.e. the *posterior*. We can use Bayes' theorem to calculate it.

For coin flips, assuming that they are independent events, the probability of obtaining 'r heads in n tosses' is given by the binomial distribution, so our *likelihood* is:

$$p(d|H, I) \propto H^r (1 - H)^{n-r}$$

But, what should we use as our *prior*?

But, what should we use as our *prior*?

Assuming we have no knowledge about the provenance of the coin, or the person tossing it, and want to reflect total ignorance of the possible bias, then a simple probability reflecting this is a **uniform**, or **flat**, pdf:
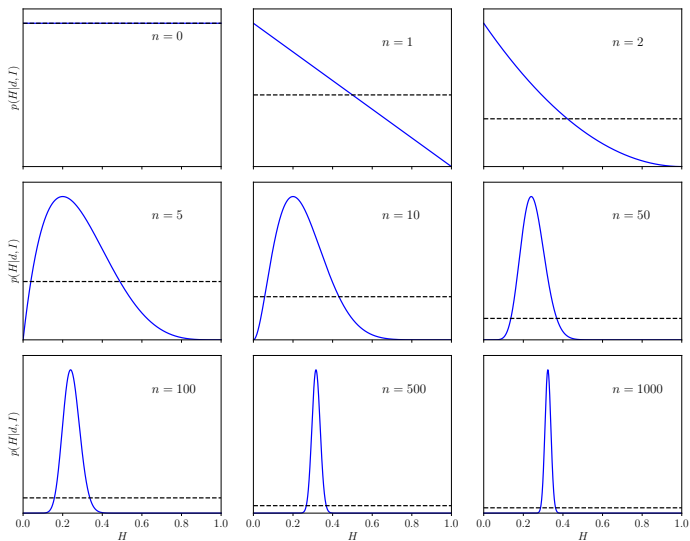
$$p(H|I) = \begin{cases} 1, \text{if } 0 \leq H \leq 1, \\ 0, \text{otherwise.} \end{cases}$$

Using these we can calculate our posterior, $p(H|d, I)$, as we obtain more data (counting r as the number of coin tosses, n, increases).[2]
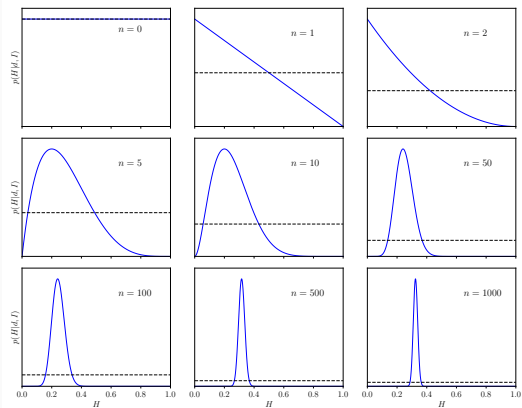
[2]See `coin_toss.py`

9

As the number of coin tosses increases the posterior evolves from the uniform prior to a tight range in H with the most probable value being $H = 0.3$.

What about a *different* prior?

We know that coins are generally fair, so what if we assume this one is too?

We can assign a Gaussian prior distribution that focusses the probability around the expected 'fair coin' value

$$p(H|I) \propto \exp\left(-\frac{1}{2}\frac{(H - \mu_H)^2}{\sigma_H^2}\right),$$

with $\sigma_H = 0.05$ and $\mu_H = 0.5$.[3]

---

[3]See coin_toss_2.py
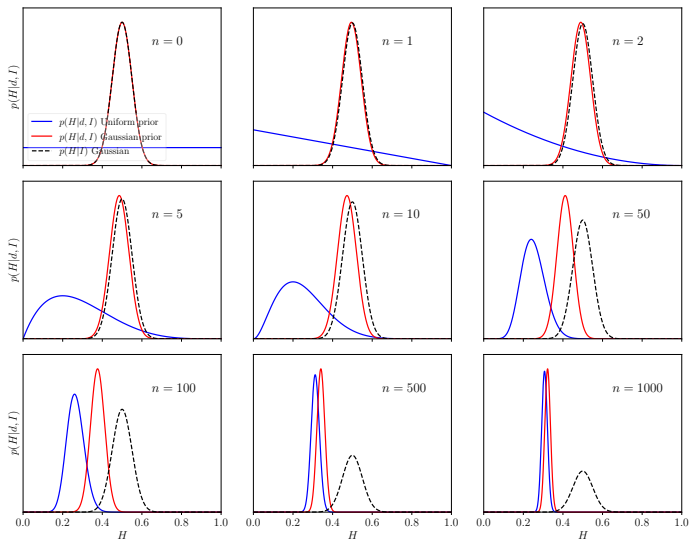
What do we learn from this?

· As our data improve (i.e. we gather more samples), the posterior pdf narrows and becomes less sensitive to our choice of prior (i.e. the likelihood starts to dominate)

· The posterior conveys our (evolving) degree of belief in different values of H given our data

· If we want to express our belief as a **single number** we can adopt e.g., the mean, median or mode

· It is very straightforward to define *Bayesian confidence intervals* (more correctly termed **credible intervals**), to quantify our uncertainty on H.

The probability distribution for a *discrete* parameter is called the probability mass function (PMF). The value of the PMF at a particular value of the parameter (say H) is the *probability* for that value, and we must have:

$$\sum_{i=1}^{N} P(H|I) = 1$$

For a continuous parameter, the probability distribution is called the **probability density function** (PDF). The value of the PDF at a particular parameter value is *not* the probability for the value, it is a probability *density*, and the probability can be calculated only for some range of the allowed parameter values, e.g.

$$P(h_1 \leq H \leq h_2) = \int_{h_1}^{h_2} p(H|I)\mathrm{d}H,$$

provided the probability distribution is properly normalised, such that

$$\int_{-\infty}^{\infty} p(H|I)\mathrm{d}H = 1.$$

16

Posteriors probability distributions have the same dimensionality as the number of parameters you want to infer in your probabilistic model, e.g., if we need to infer the gradient and y-intercept of a straight line (m and c) from some data ($\boldsymbol{d}$), then you have two parameters and your posterior will be two-dimensional: $p(m, c | \boldsymbol{d}, I)$.

You may only be interested in the distribution of one (or some subset) of the parameters, so you **marginalise** (i.e., integrate) over the *nuisance* parameters. E.g., if we're only interested in the gradient of the line then:

$$p(m | \boldsymbol{d}, I) = \int_{-\infty}^{\infty} p(m, c | \boldsymbol{d}, I) \mathrm{d}c.$$

For higher dimensional problems multiple integrals may be required:

$$p(x|\boldsymbol{d}, I) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y, z|\boldsymbol{d}, I) \mathrm{d}y \mathrm{d}z.$$

In some cases marginalisation can be performed analytically. For low dimensional problems, with a well localised posterior (i.e., it doesn't have support out to $\pm\infty$), the posterior can be evaluated on a grid and marginalisation can be performed numerically. For high-dimensional problems this is not viable and we must use stochastic sampling methods, e.g., Markov chain Monte Carlo (see talk be Vivien Raymond) or nested sampling (see talk by John Veitch).

The normalisation constant for the posterior probability is often called the Bayesian evidence, or marginal likelihood,

$$p(\boldsymbol{d}|\mathcal{H}, I) = \int^{\boldsymbol{\theta}} p(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{H}, I)p(\boldsymbol{\theta}|\mathcal{H}, I)\mathrm{d}\boldsymbol{\theta},$$

where the integral is multi-dimensional over all parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$. Here we been explicitly stated that the likelihood and prior are conditional on a given model, or hypothesis, $\mathcal{H}$. Hence, the evidence is the *likelihood* of observing the data for a given model $\mathcal{H}$.

The multi-dimensional integral may difficult/impossible to compute analytically or using standard numerical integration methods, so the nested sampling algorithm may be required (see talk by John Veitch).

If you are purely interested in marginal posterior distributions the normalisation is not important and can must often be ignored. However, the evidence allows you to compare different models (say $\mathcal{H}_1$ and $\mathcal{H}_2$) given the same data. The model odds is:

$$\mathcal{O}_{12} \equiv \underbrace{\frac{p(\mathcal{H}_1|\boldsymbol{d}, I)}{p(\mathcal{H}_2|\boldsymbol{d}, I)}}_{\text{Model Odds}} = \overbrace{\frac{\boldsymbol{d}|p(\mathcal{H}_1, I)}{p(\boldsymbol{d}|\mathcal{H}_2, I)}}^{\text{Bayes Factor}} \overbrace{\frac{p(\mathcal{H}_1|I)}{p(\mathcal{H}_2|I)}}^{\text{Prior Odds}}.$$

Bayes factor is the ratio of the evidences for the two hypotheses. The prior odds defines the *a priori* relative degree-of-belief about each model, which in practice is often set to unity, i.e., neither model is preferred *a priori*.

We define a **credible interval** $[\theta_a, \theta_b]$ as a (*non-unique*) range that a contains a certain amount of posterior probability, X,

$$X = \int_{\theta_a}^{\theta_b} p(\theta|d, I)\mathrm{d}\theta.$$

If $X = 0.95$ then we can find $[\theta_a, \theta_b]$ that e.g. gives the minimum range containing 95% of the probability.

The meaning of this is simple: *we are 95% sure that θ lies between θ_a and θ_b.*

In many situations in physics/astronomy our data consists of the signal with some additive noise, e.g. considering a single data point

$$d_1 = s_1 + n_1.$$

We are interested in the *inverse problem* of inferring the properties of the signal given the data. To do this, and define a likelihood, we need to make some assumptions about the noise properties.

If the noise generating process can be thought of as the sum of independent random processes then by the *central limit theorem* it will tend towards a normal (or Gaussian) distribution. So, we often assume $n \sim N(0, \sigma^2)$ ("*n is drawn from a Normal distribution with mean of zero and variance σ^2*")

Also, for a process where we have the expectation value μ and variance σ^2, the distribution that maximises the entropy, is the least informative (most conservative), is the normal distribution:

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

For our single point of data $d_1 = s_1 + n_1$, if we have a model of our signal parameterised by $\vec{\theta}$, such that $s_1 \equiv s_1(\vec{\theta})$, then due to the additive nature of the noise and signal, the expectation value $\mu = s_1(\vec{\theta})$, and we have:

$$p(d_1|\vec{\theta}, \sigma, I) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(d_1 - s_1(\vec{\theta})\right)^2}{2\sigma^2}\right)$$

Often we have more than one data point! If the noise in the data is independent and identically distributed (*i.i.d.*) you can multiply the likelihoods for each data point to give the *joint* likelihood for all the data $\mathbf{d} = \{d_1, d_2, \ldots, d_N\}$:

$$
p(\mathbf{d}|\vec{\theta}, \sigma, I) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(d_i - s_i(\vec{\theta})\right)^2}{2\sigma^2}\right),
$$

$$
= \left(2\pi\sigma^2\right)^{-N/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{N} \frac{\left(d_i - s_i(\vec{\theta})\right)^2}{\sigma^2}\right)
$$

This is a strictly (or strongly) stationary process.

If the noise process is Gaussian, but the noise is *not* correlated, i.e., each point (or a subset of points) has a different *known* variance $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_N\}$, we have:

$$p(\mathbf{d}|\vec{\theta}, \boldsymbol{\sigma}, I) = \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \right) \exp \left(-\frac{\left(d_i - s_i(\vec{\theta}) \right)^2}{2\sigma_i^2} \right).$$

If the noise process is Gaussian, but the noise is correlated and can be defined by a known (or estimatable) covariance matrix $\boldsymbol{\Sigma}$ (a weakly stationary process), we have a multivariate normal distribution:

$$p(\mathbf{d}|\vec{\theta}, \boldsymbol{\Sigma}, I) = (2\pi)^{n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\left(\mathbf{d} - \mathbf{s}(\vec{\theta})\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{d} - \mathbf{s}(\vec{\theta})\right)\right),$$

(1)

where $\mathbf{s}(\vec{\theta}) = \{s_1(\vec{\theta}), \ldots, s_N(\vec{\theta})\}$. This becomes the previous case if $\boldsymbol{\Sigma}$ is diagonal.

The covariance matrix can be estimated by finding the
autocovariance of the noise (ideally from some data that is
drawn purely from the noise process, i.e., contains not signal).
If we assume N evenly sampled noise data points $\boldsymbol{n}$, then the
autocovariance is:

$$\gamma_j = \frac{1}{N-1} \sum_{i=0}^{N-j} (n_{i+j} - \bar{\boldsymbol{n}}) (n_{i+1} - \bar{\boldsymbol{n}}),$$

with j indices starting at 1, and $\bar{\boldsymbol{n}} = (1/N) \sum_{i=1}^{N} n_i$. This could
be estimated from M multiple stretches of data and averaged,
e.g., $\bar{\gamma}_j = (1/M) \sum_{i=1}^{M} \gamma_{ij}$.

The **autocorrelation function** is then defined as $\rho_j = \frac{\gamma_j}{\gamma_1}$ and, setting $\gamma_1 \equiv \sigma^2$, the covariance matrix is:

$$\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho_2 & \rho_3 & \cdots & \rho_n \\ \rho_2 & 1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_3 & \rho_2 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho_2 \\ \rho_n & \rho_{n-1} & \cdots & \rho_2 & 1 \end{pmatrix}.$$

As an example, we'll examine the problem of fitting a line $\boldsymbol{y} = m\boldsymbol{x} + c$ to data, $\boldsymbol{d}$ (linear regression). We can write the posterior for the parameters

$$p(m, c | \boldsymbol{d}, I) \propto \underbrace{p(\boldsymbol{d} | m, c, I)}_{\text{Likelihood}} \times \underbrace{p(m, c | I)}_{\text{Prior}}.$$

If the prior on the parameters is uniform and independent, so

$$p(m, c | I) = p(m | I) p(c | I) = \text{constant},$$

then the posterior is

$$p(m, c | \boldsymbol{d}, I) \propto p(\boldsymbol{d} | m, c, I).$$

We could, in this case, use the machinery of *maximum likelihood* to estimate the parameters.

However, we will use this to show to general concept of fitting any model (see `bayesian_line_fitting.py`). If the likelihood is Gaussian, with known values of σ_i, then

$$p(m, c | \boldsymbol{d}, I) = p(m, c | I) \left(\frac{1}{2\pi\sigma_i^2} \right)^{n/2} \exp\left(-\sum_{i=1}^{n} \frac{[d_i - (mx_i + c)]^2}{2\sigma_i^2} \right)$$
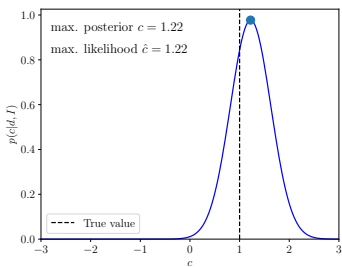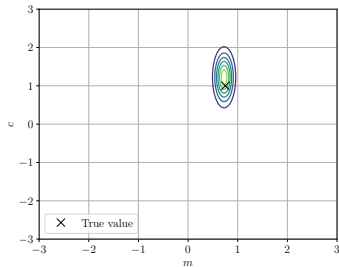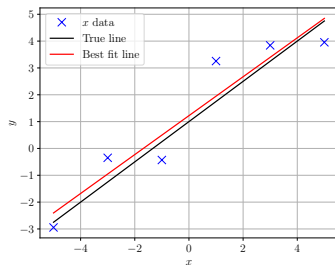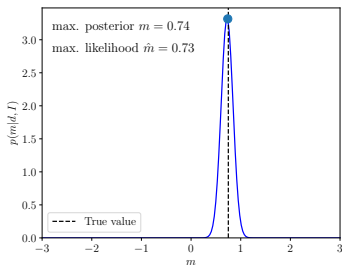
and we can evaluate the posterior over a grid in the parameters m and c.

We can also compute the marginal posteriors on m and c as, e.g.

$$p(m | \boldsymbol{d}, I) = \int_{-\infty}^{\infty} p(m, c | d, I) \mathrm{d}c.$$

In the above example, in practice, when $p(m, c|I) = $ constant and $\sigma_i = \sigma$ are constant, we can just calculate the posterior[4] over a grid in m and c

$$\ell(m_{i_m}, c_{i_c}) = \ln p(m_{i_m}, c_{i_c}|d, I) = -\sum_{i=1}^{n} \frac{[d_i - (m_{i_m} x_i + c_{i_c})]^2}{2\sigma^2}$$

and get the marginal posteriors through numerical integration, e.g.

$$p(m_{i_m}|d, I) \propto \sum_{i_c}^{n_c} \exp\left(\ell(m_{i_m}, c_{i_c}) - \max\ell(m, c)\right)\Delta c$$

where Δc are the grid step sizes in c (or you could use the trapezium rule for more accuracy).

[4] We generally work in natural logarithm space due to numerical precision issues.

33

Gravitational-wave detectors produce a *real* time series of strain measurements $h(t)$ ("*h of t*"). This is the linear combination of noise (assumed to be produced by a weakly stationary process, i.e., coloured Gaussian noise) and the signal as projected onto the detector via its response function:

$$h(t) = n(t) + s(t; \boldsymbol{\theta}),$$

where

$$s(t; \boldsymbol{\theta}) = F_+^D(t; \alpha, \delta, \psi) h_+(t; \boldsymbol{\theta}') + F_\times^D(t; \alpha, \delta, \psi) h_\times(t; \boldsymbol{\theta}').$$

$F_{+/\times}^D$ are the 'plus' and 'cross' polarisation responses of detector D to a source at a sky position given by right ascension α and declination δ and with polarisation angle ψ. $h_{+/\times}$ are the source amplitudes at the Earth defined by the parameters $\boldsymbol{\theta}'$, where $\boldsymbol{\theta} = \{\alpha, \delta, \psi, \boldsymbol{\theta}'\}$.

We are interested in using $h(t)$ (we'll use $\boldsymbol{d}$ for the vector of observed time series *data* points instead of h from now on) to infer the marginal probability distributions of the parameters $\boldsymbol{\theta}$ (or some subset of them). E.g., the posteriors on the sky position of the source

$$p(\alpha, \delta | \boldsymbol{d}, I) \propto \int^{\boldsymbol{\theta}_{\notin \{\alpha, \delta\}}} p(\boldsymbol{d} | \boldsymbol{\theta}, I) p(\boldsymbol{\theta} | I) \mathrm{d}\boldsymbol{\theta}_{\notin \{\alpha, \delta\}},$$

or the source masses for a compact binary coalescence event

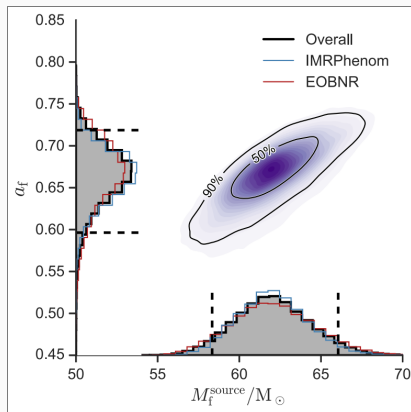$$p(m_1, m_2 | \boldsymbol{d}, I) \propto \int^{\boldsymbol{\theta}_{\notin \{m_1, m_2\}}} p(\boldsymbol{d} | \boldsymbol{\theta}, I) p(\boldsymbol{\theta} | I) \mathrm{d}\boldsymbol{\theta}_{\notin \{m_1, m_2\}}.$$

So, we need to have a likelihood for the data and a prior for the source parameters.
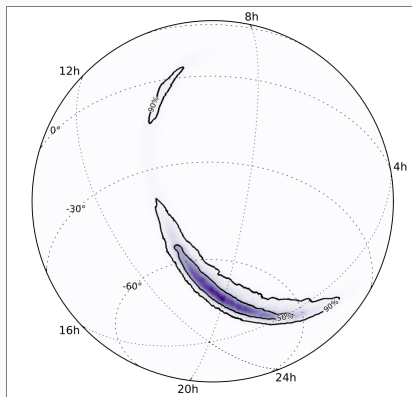
Sky location posterior for GW150914 (Abbott et al., 2016).

Source mass posteriors for GW150914 (Abbott et al., 2016).

A non-exhaustive list examples of where Bayesian inference has been used in (ground-based) gravitational-wave astronomy:

· Searches for continuous (monochromatic) gravitational waves from known pulsars
· Source parameter estimation for CBC signals
· Rapid CBC source sky and distance localisation (BAYESTAR)
· Unmodelled burst event trigger generator (BLOCKNORMAL)
· Unmodelled burst waveform reconstruction, glitch reconstruction, and power spectrum estimation (BAYESWAVE)
· Unmodelled burst parameter estimation (oLIB)
· Supernova signal model comparison (SMEE)
· Hierarchical inference of CBC mass and spin distributions

We've seen the "standard" Gaussian likelihood function, but in GW papers you might see (e.g., Equations 3 and 4 of Biwer et al., 2019, assuming a single detector, and ignoring the normalisation):

$$p(\boldsymbol{d}|\boldsymbol{\theta}, I) \propto \exp\left(-\frac{1}{2} \overbrace{\langle \tilde{d}(f) - \tilde{s}(f; \boldsymbol{\theta}) | \tilde{d}(f) - \tilde{s}(f; \boldsymbol{\theta})\rangle}^{\text{noise weighted innner product}}\right)$$

$$\equiv \exp\left(-\frac{1}{2}\left[4\Re \int_0^\infty \frac{\left(\tilde{d}(f) - \tilde{s}(f; \boldsymbol{\theta})\right)\left(\tilde{d}(f) - \tilde{s}(f; \boldsymbol{\theta})\right)^*}{S_n} df\right]\right),$$

where d is the data, $\boldsymbol{\theta}$ is a set of parameters defining a waveform model s, the tilde represents the Fourier transform, and S_n is the one-sided power spectral density of the noise in the data.

Let's see how this relates to our earlier equation.

First, we actually work with discrete data, and discrete Fourier transforms, so

$$p(\boldsymbol{d}|\boldsymbol{\theta}, I) \propto \exp\left(-\frac{1}{2}\left[4\Re\int_0^\infty \frac{\left(\tilde{d}(f) - \tilde{s}(f;\boldsymbol{\theta})\right)\left(\tilde{d}(f) - \tilde{s}(f;\boldsymbol{\theta})\right)^*}{S_n}(f)\mathrm{d}f\right]\right),$$

becomes

$$p(\boldsymbol{d}|\boldsymbol{\theta}, I) \propto \exp\left(-\frac{1}{2}\left[4\Re\sum_{i=j}^k \frac{\left(\tilde{d}_i - \tilde{s}_i(\boldsymbol{\theta})\right)\left(\tilde{d}_i - \tilde{s}_i(\boldsymbol{\theta})\right)^*}{TS_n^{(i)}}\right]\right),$$

due to $\int \ldots \mathrm{d}f \approx \sum \ldots \Delta f$, and $\Delta f = 4/T$ for data of length T seconds, and i in the index over frequency bins from bin j to k.

Given that for complex $x = a + ib$,
$xx^* = (a + ib)(a - ib) = a^2 + b^2$, we get

$$p(\boldsymbol{d}|\boldsymbol{\theta}, I) \propto \exp\left(-\frac{1}{2}\left[4\sum_{i=j}^{k}\frac{\Re\left(\tilde{d}_i - \tilde{s}_i(\boldsymbol{\theta})\right)^2 + \Im\left(\tilde{d}_i - \tilde{s}_i(\boldsymbol{\theta})\right)^2}{TS_n^{(i)}}\right]\right),$$

$$\equiv \exp\left(-\frac{1}{2}\left[4\sum_{i=j}^{k}\frac{\left|\tilde{d}_i - \tilde{s}_i(\boldsymbol{\theta})\right|^2}{TS_n^{(i)}}\right]\right)$$

This is the same as starting from the assumption that the noise in the real and imaginary parts of the Fourier transform are independent (but drawn from the same noise process), and writing the *joint* the likelihood of these independent data sets.

For numerical reasons we generally work with the natural logarithm of the likelihood (and other probability densities), so

$$\ln p(\boldsymbol{d}|\boldsymbol{\theta}, I) = -\frac{1}{2}\left[4\sum_{i=j}^{k}\frac{\Re\left(\tilde{d}_i - \tilde{s}_i(\boldsymbol{\theta})\right)^2 + \Im\left(\tilde{d}_i - \tilde{s}_i(\boldsymbol{\theta})\right)^2}{TS_n^{(i)}}\right] + C,$$

where C is the normalisation term given by (see Equation 12 of Veitch et al. (2015))

$$C = -\frac{1}{2}\sum_{i=j}^{k}\ln\left(\pi TS_n^{(i)}/2\right).$$

If we expand out the quadratic terms we get:

- the *null* log-likelihood (noise-only log *evidence*):

$$\ln p(\boldsymbol{d}|\boldsymbol{s}(\boldsymbol{\theta}) = 0, I) \equiv \mathcal{L}_n = -(2/T) \sum_i \left(\Re(d_i)^2 + \Im(d_i)^2 \right) / S_n^{(i)} + C,$$

- the *optimal* signal-to-noise:

$$\rho_{\text{opt}}^2 = (4/T) \sum_i \left(\Re(s_i)^2 + \Im(s_i)^2 \right) / S_n^{(i)},$$

- the *matched filter* signal-to-noise:

$$\rho_{\text{mf}}^2 = (4/T) \sum_i \left(\Re(d_i)\Re(s_i) + \Im(d_i)\Im(s_i) \right) / S_n^{(i)}.$$

So, we can write the log-likelihood in terms of:

$$\ln p(d|\boldsymbol{\theta}, I) = -\frac{1}{2}\left(\rho_{\mathrm{opt}}^2(\boldsymbol{\theta}) - 2\rho_{\mathrm{mf}}^2(\boldsymbol{\theta})\right) + \mathcal{L}_n. \tag{2}$$

The log of the likelihood ratio $p(\boldsymbol{d}|\boldsymbol{\theta}, I)/p(\boldsymbol{d}|\boldsymbol{s}(\boldsymbol{\theta}) = 0, I)$ is therefore:

$$\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}_n = -\frac{1}{2}\left(\rho_{\mathrm{opt}}^2(\boldsymbol{\theta}) - 2\rho_{\mathrm{mf}}^2(\boldsymbol{\theta})\right).$$

Evaluating the likelihood over the parameter space effectively required evaluating these ρ^2 terms.

In the above likelihood it assumes the noise in each frequency bin is independent and Gaussian, with a variance defined using the *one-sided* **power spectral density** (PSD), $S_n(f)$, given by[5]

$$S_n(f) = \frac{2}{N^2 \Delta f}|\tilde{n}(f)|^2, \text{ with } \tilde{n}(f_k) = \sum_j n_j e^{-2\pi ijk/N},$$

where N is the number of data points, and $\Delta f = 1/T = 1/(N\Delta t)$ for observation time T (see, e.g. Appendix of Veitch & Vecchio, 2006)

The variance for the real and imaginary components of each frequency bin is given by $\sigma_i^2 = (T/4)S_n^{(i)}$, which when substituted into Equation (2) gives the standard Gaussian log-likelihood.

[5]This is equivalent to the Fourier transform of the noise autocorrelation

In practice we estimate the PSD using, e.g., Welch's method. A stretch of noise-only data is chosen and divided into M overlapping segments (with fractional overlap α) each of the same length, which is the same length N as the data segment to be analysed. Each segments is multiplied by a window, Fourier transformed, and the average power from all segments is used:

$$S_n(f) = \frac{2}{MN^2\Delta f} \sum_{i=0}^{M-1} \left| \text{FFT}(d(t_{1+i\alpha N:N(1+i\alpha)})w(t)) \right|^2.$$

Windowing is *vital* (when Fourier tranforming the analysis segment and for the PSR estimation) to prevent spectral leakage and add in correlations to the data.

For broadband signals, like those from CBC, the coloured nature of the noise generally means it's easier to work in the frequency domain; in the frequency domain we just have vector-vector dot products of the data/signal and PSD rather than a vector-matrix product of the data/signal and the correlation matrix as in Equation 1 (i.e., its just a multiplication in the frequency domain rather than a convolution).

But, some advantage of the time-domain are:

· you do not have to worry about windowing(!);
· time delays are just time delays rather than frequency dependent phase shifts;
· the start and end of signals can be simply defined.

If you have M detectors, assuming the noise in each is independent, you can coherently combine them by taking the product of the likelihoods for each, so

$$p(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta}, I) \propto \prod_{j=1}^{M} p(\boldsymbol{d}_j|\boldsymbol{\theta}, I)$$
$$= \exp\left(-\frac{1}{2}\left[4\sum_{j=1}^{M}\sum_{i=1}^{N_j}\frac{\left|\tilde{d}_{ij} - \tilde{s}_{ij}(\boldsymbol{\theta})\right|^2}{T_j S_{n_j}^{(i)}}\right]\right),$$

where $\boldsymbol{\mathcal{D}} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_M\}$ is the combined data from all detectors.

For the standard GW likelihood function and a certain forms of the signal model, it is possible to analytically marginalise out certain parameters. For example, if the signal consists of a sinusoidal term with an initial phase, e.g., $s \propto e^{i\phi_0}$, then

$$p(\boldsymbol{d}|\boldsymbol{\theta}', I) \propto \int_0^{2\pi} p(\boldsymbol{d}|\boldsymbol{\theta})p(\phi_0|I)\mathrm{d}\phi_0,$$
$$= \exp\left(-\frac{2}{T}\sum_i \frac{|s_i(\boldsymbol{\theta}')|^2 + |d_i|^2)}{S_n^{(i)}}\right) I_0\left(\frac{4}{T}\left|\sum_i \frac{s_i(\boldsymbol{\theta}')d_i^*}{S_n^{(i)}}\right|\right),$$

where $\boldsymbol{\theta}'$ contains all the parameters of $\boldsymbol{\theta}$ except ϕ_0 (see Equation (20) of Veitch et al., 2015).

Add one or two slides on this!

J. Veitch et al. Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library. *Phys. Rev. D*, 91, 042003, 2015.

C. M. Biwer et al. PyCBC Inference: A Python-based parameter estimation toolkit for compact binary coalescence signals. *Publ. Astron. Soc. Pac.*, 131, 024503, 2019.

G. Ashton et al. BILBY: A User-friendly Bayesian Inference Library for Gravitational-wave Astronomy *Astrophys. J. Supplement Series*, 241, 27, 2019.

D. S. Sivia. *Data analysis: A Bayesian Tutorial.* Oxford University Press, 2006.

B. P. Abbott et al. Properties of the Binary Black Hole Merger GW150914 *Phys. Rev. Lett.*, 116, 241102, 2016.

J. Veitch & A. Vecchio Bayesian coherent analysis of in-spiral gravitational wave signals with a detector network. *Phys. Rev. D*, 81, 062003, 2010.

The rules for probabilities of propositions are inherited from classical logic and Boolean algebra:

- *Law of Excluded Middle* $P(A \text{ or } \text{not}(A)) = 1$
- *Law of Non-contradiction* $P(A \text{ and } \text{not}(A)) = 0$
  - i.e. $P(A) + P(\text{not } A) = 1$ (the **sum rule**)
- *Association*
  - $P(A, [B, C]) = P([A, B], C)$
  - $P(A \text{ or } [B \text{ or } C]) = P([A \text{ or } B] \text{ or } C)$
- *Distribution*
  - $P(A, [B \text{ or } C]) = P(A, B \text{ or } A, C)$
  - $P(A \text{ or } [B, C]) = P([A \text{ or } B], [A \text{ or } C])$

- *Commutation*

  - $P(A, B) = P(B, A)$
  - $P(A \text{ or } B) = P(B \text{ or } A)$

- *Duality* (De Morgan's Theorem)

  - $P(\text{not } [A, B]) = P(\text{not}(A) \text{ or } \text{not}(B))$
  - $P(\text{not } [A \text{ or } B]) = P(\text{not}(A), \text{not}(B))$

Note that you may see other notation for probabilities expressed with Boolean logic (this list is not exhaustive)

- Negation (A is false)
  - $P(\text{not } A)$, or $P(\bar{A})$, or $P(\neg A)$
- Logical product (both A *and* B are true)
  - $P(A, B)$, or $P(AB)$, or $P(A \text{ and } B)$, or $P(A \wedge B)$
- Logical sum (at least one of A *or* B is true)
  - $P(A + B)$, or $P(A \text{ or } B)$, or $P(A \vee B)$

From these axioms we can derive:

- The **(Extended) Sum Rule**
  - $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

- The **Product Rule**
  - $P(A \text{ and } B) \equiv p(A, B) = P(A)P(B|A) = P(B)P(A|B)$, where $P(x|y)$ is the probability that x is true given y is true.

These rules apply to probabilities P and also probability density functions (pdfs) p.

Simple demonstration of the extended sum rule.

What is the probability that a card drawn from a standard deck of cards is a spade *or* an ace?

We have $P(\spadesuit) = 13/52 = 1/4$ and $P(\text{ace}) = 4/52 = 1/13$, and $P(\spadesuit \text{ and ace}) = 1/(4 \times 13) = 1/52$. It is reasonably obvious that for $P(\spadesuit \text{ or ace})$ we want to sum the probabilities for both cases, however they both contain the case where $P(\spadesuit \text{ and ace})$, so we have to remove one of those instances

$$P(\spadesuit \text{ or ace}) = \frac{13 + 4 - 1}{52} = \frac{16}{52}$$

What if we don't know σ?

In this case we can treat σ as another unknown variable and marginalise over it, e.g.

$$p(m, c|\boldsymbol{d}, I) = p(m, c|I) \int_0^\infty p(d|m, c, \sigma, I) p(\sigma|I) \mathrm{d}\sigma$$

If the likelihood is Gaussian and we assume a flat prior on all parameters, e.g.

$$p(\sigma|I) = \begin{cases} C, \sigma > 0 \\ 0, \sigma \leq 0 \end{cases}$$

Then we have

$$p(m, c|\boldsymbol{d}, I) \propto \int_0^\infty \sigma^{-n} \exp\left(-\sum_{i=1}^n \frac{[d_i - (mx_i + c)]^2}{2\sigma^2}\right) \mathrm{d}\sigma$$

56

This integral is analytic, and through some substitution (see, e.g., Chap. 3 of Sivia, 2006), becomes

$$p(m, c | \boldsymbol{d}, I) \propto \left(\sum_{i=1}^{n} [d_i - (mx_i + c)]^2 \right)^{-(n-1)/2}$$

This is essentially a **Student's t-distribution** with $\nu = (n-2)$ degrees of freedom.

Note: if we were instead to use a prior on σ of $p(\sigma|I) \propto 1/\sigma$ it would lead to a Student's t-distribution with $\nu = n - 1$ degrees of freedom.