BAYESIAN PARAMETER ESTIMATION IN GRAVITATIONAL-WAVE ASTRONOMY

Matthew Pitkin

PyCBC Inference Workshop

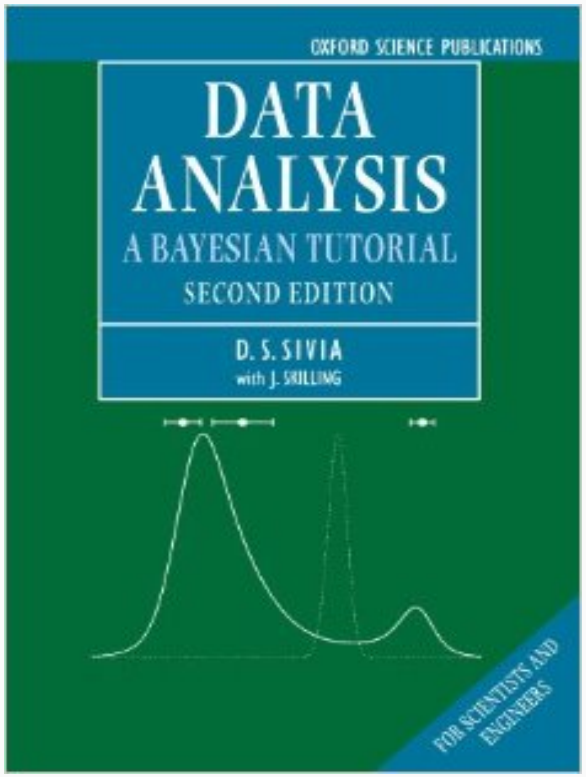
14 May 2019

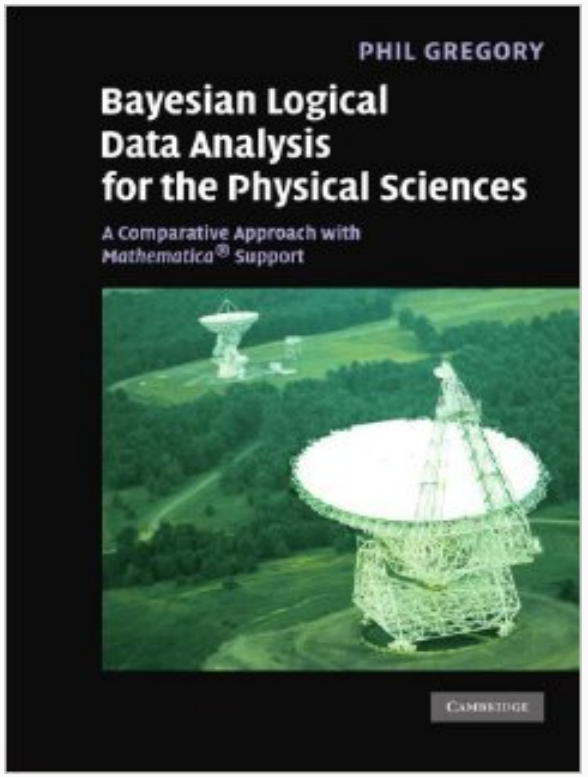
University of Glasgow

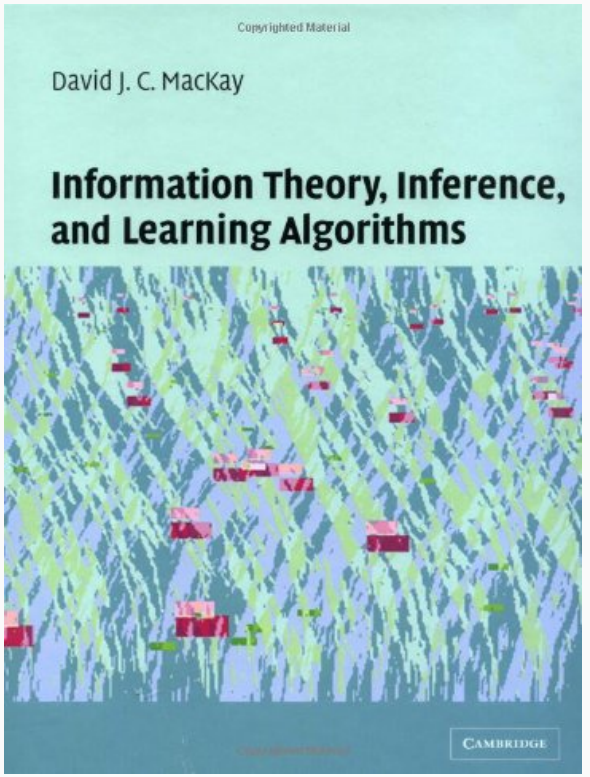
OVERVIEW

- Introduction
  - Bayes' theorem:
    - probability density, marginalisation, evidence, model selection, credible intervals
  - Gaussian likelihood
  - linear regression example
- Gravitational-wave inference
  - Examples
  - the likelihood function
    - the power spectral density
  - hierarchical inference

There are many textbooks on statistics (and Bayesian statistics in particular), but three I can recommend are:







PAPERS TO READ

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)
- *Data analysis recipes: Fitting a model to data*, Hogg, Bovy & Lang (2010)
- *An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models*, Thrane & Talbot (2019)

There are many more...

From the product rule of probability it can be shown that

$$P(A \text{ and } B) \equiv P(A, B) = P(A)P(B|A) = P(B)P(A|B),$$

where $P(x|y)$ means "the probability that x is true given y is true".[1] Rearranging this gives

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

This is Bayes' theorem.

Note: we we have explicitly added the conditioning on background information I.

[1] sometimes "P" is used for "probability" and "p" is use for a "probability density" - the same rules apply in both cases.

BAYES' THEOREM

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

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

For practical problems we need to be able to calculate (*most of*) the these terms, e.g., analytically or numerically on a computer.

BAYES' THEOREM

- 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 (or marginal likelihood): the "*evidence*" for the data, or the likelihood for the data *marginalised* over the model (we'll explore this briefly later, but at the moment note it as the normalisation factor for the posterior)
- Posterior: our new degree of belief about our model in light of the data

A BAYESIAN EXAMPLE: IS A COIN FAIR?

How can we determine if a coin is fair?[2] 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 heads or tails
- 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.

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

[2] See e.g. Chap. 2 of Sivia (2006).

In the limiting case where our propositions each lie in the infinitesimal range $\mathrm{d}H$ our inference about the bias weighting is summarised by the probability density function (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*?

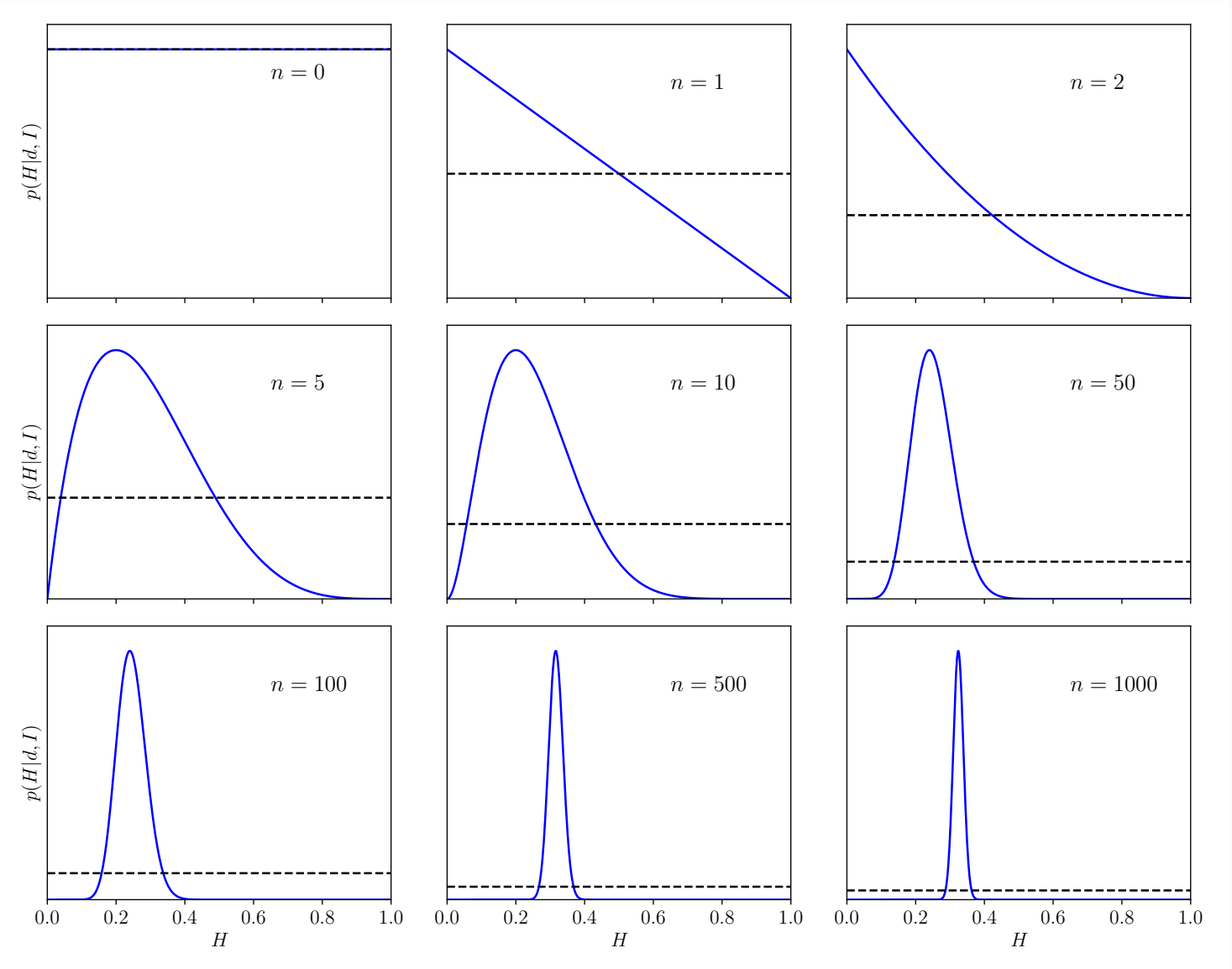
Assuming we have no knowledge about the provenance of the coin, or the person tossing it, and want to reflect total ignorance of the possible bias, then a simple probability reflecting this is a uniform, or *flat*, pdf:

$$p(H|I) = \begin{cases} 1, \text{if } 0 \leq H \leq 1, \\ 0, \text{otherwise.} \end{cases}$$

Using the likelihood and the prior we can calculate our posterior, $p(H|d, I)$, as we obtain more data (counting r as the number of coin tosses, n, increases).[3]

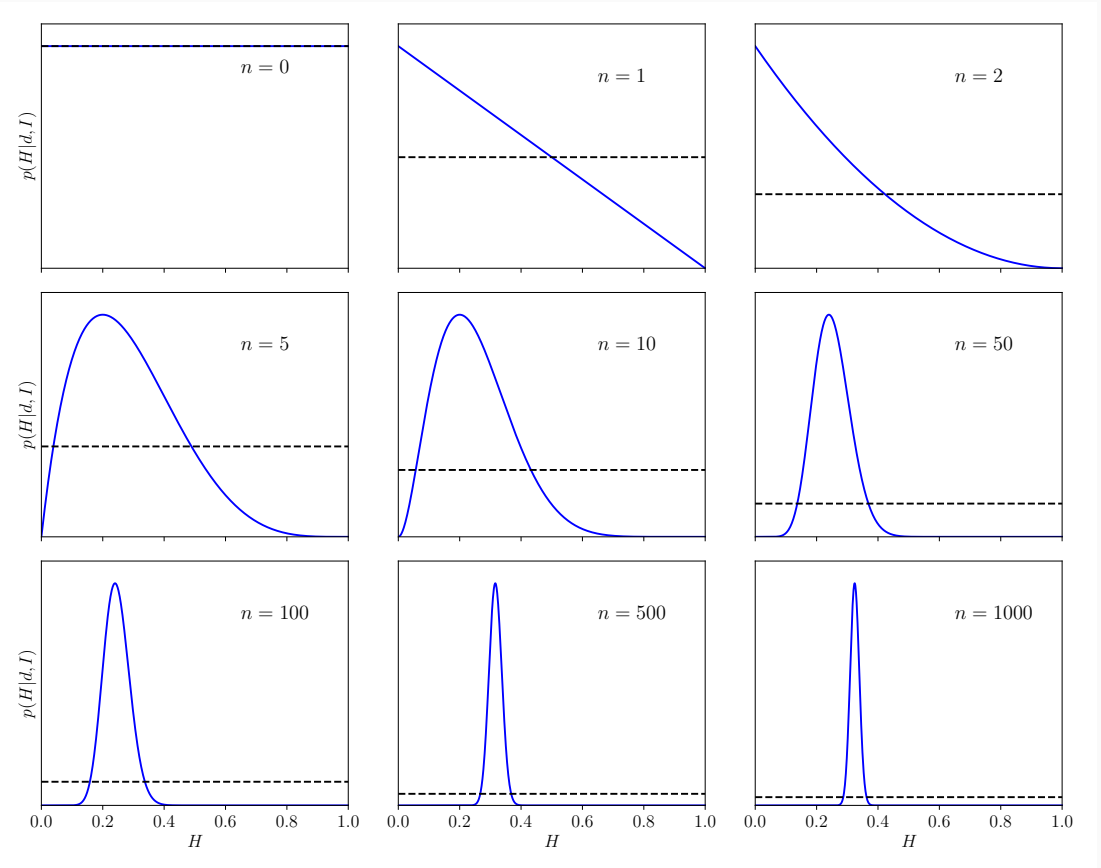
[3] See `coin_toss.py`

A BAYESIAN EXAMPLE: IS A COIN FAIR?



A BAYESIAN EXAMPLE: IS A COIN FAIR?

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?

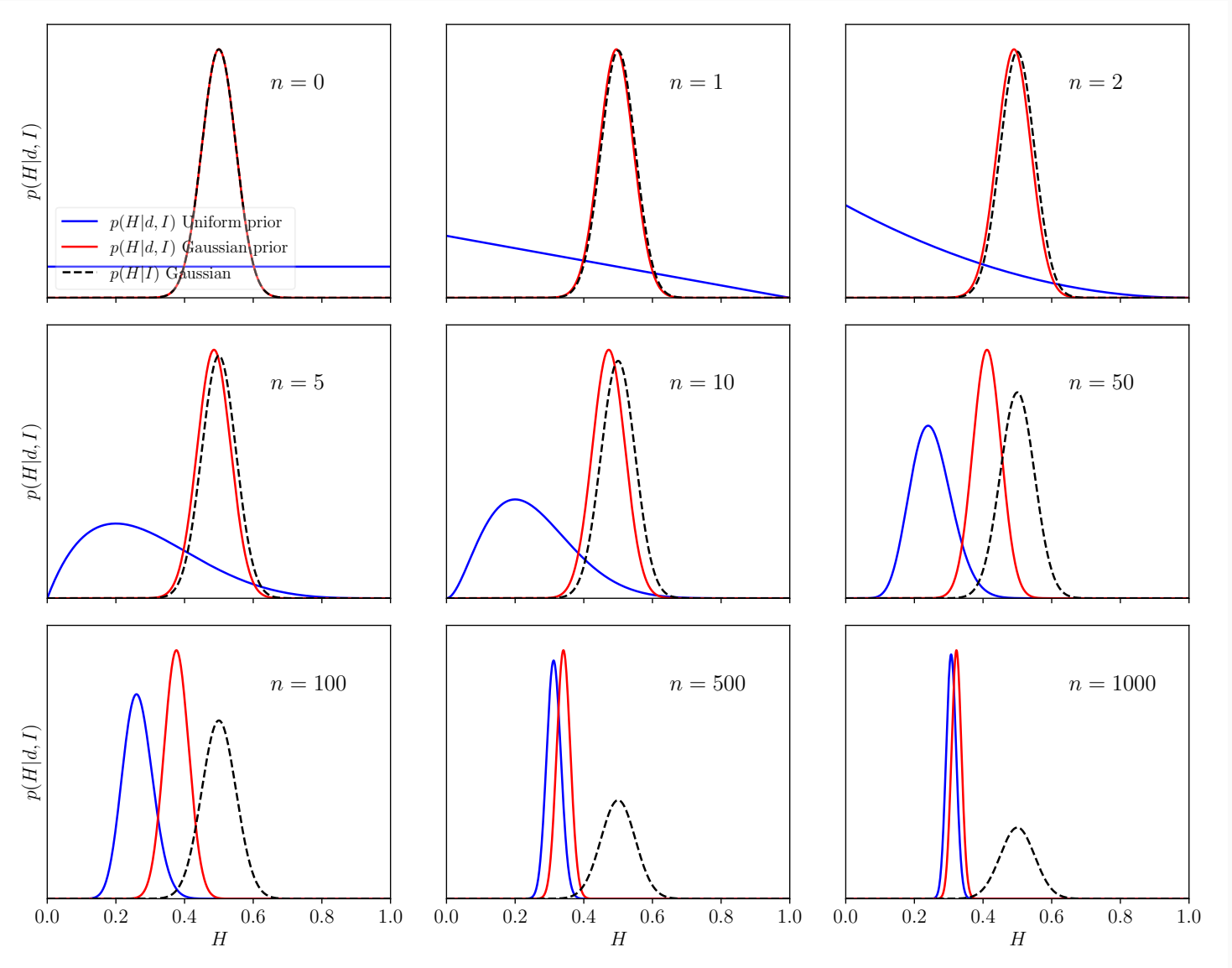
We can assign a Gaussian prior distribution that focuses the probability around the expected 'fair coin' value

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

with $\sigma_H = 0.05$ and $\mu_H = 0.5$.[4]

[4] See `coin_toss_2.py`

A BAYESIAN EXAMPLE: IS A COIN FAIR?



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 the probability *density*, and the probability can be calculated only for some range of the allowed parameter values, e.g.,

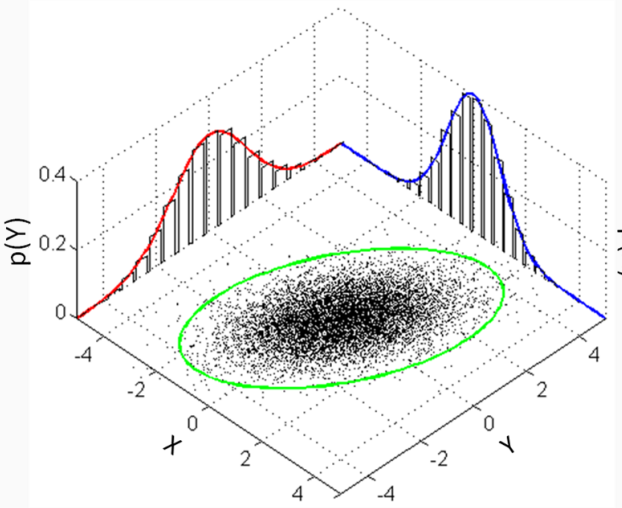
$$P(h_1 \leq H \leq h_2) = \int_{h_1}^{h_2} p(H|I)\mathrm{d}H,$$

provided the probability distribution is properly normalised, such that

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

MARGINALISATION

Posteriors probability distributions have the same dimensionality as the number of parameters we want to infer in our 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 we have two parameters and our posterior will be two-dimensional: $p(m, c|\boldsymbol{d}, I)$.



Example of marginalisation. Credit: Bscan, CC0.

We may only be interested in the distribution of one (or some subset) of the parameters, so we 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 by 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 model parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$. Here we have 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 $\mathcal{H}, marginalised over its parameters.$

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 we are purely interested in marginal posterior distributions the normalisation is not important and can most 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:

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

The Bayes factor is the ratio of the evidences for the two hypotheses (see also the likelihood ratio). 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.

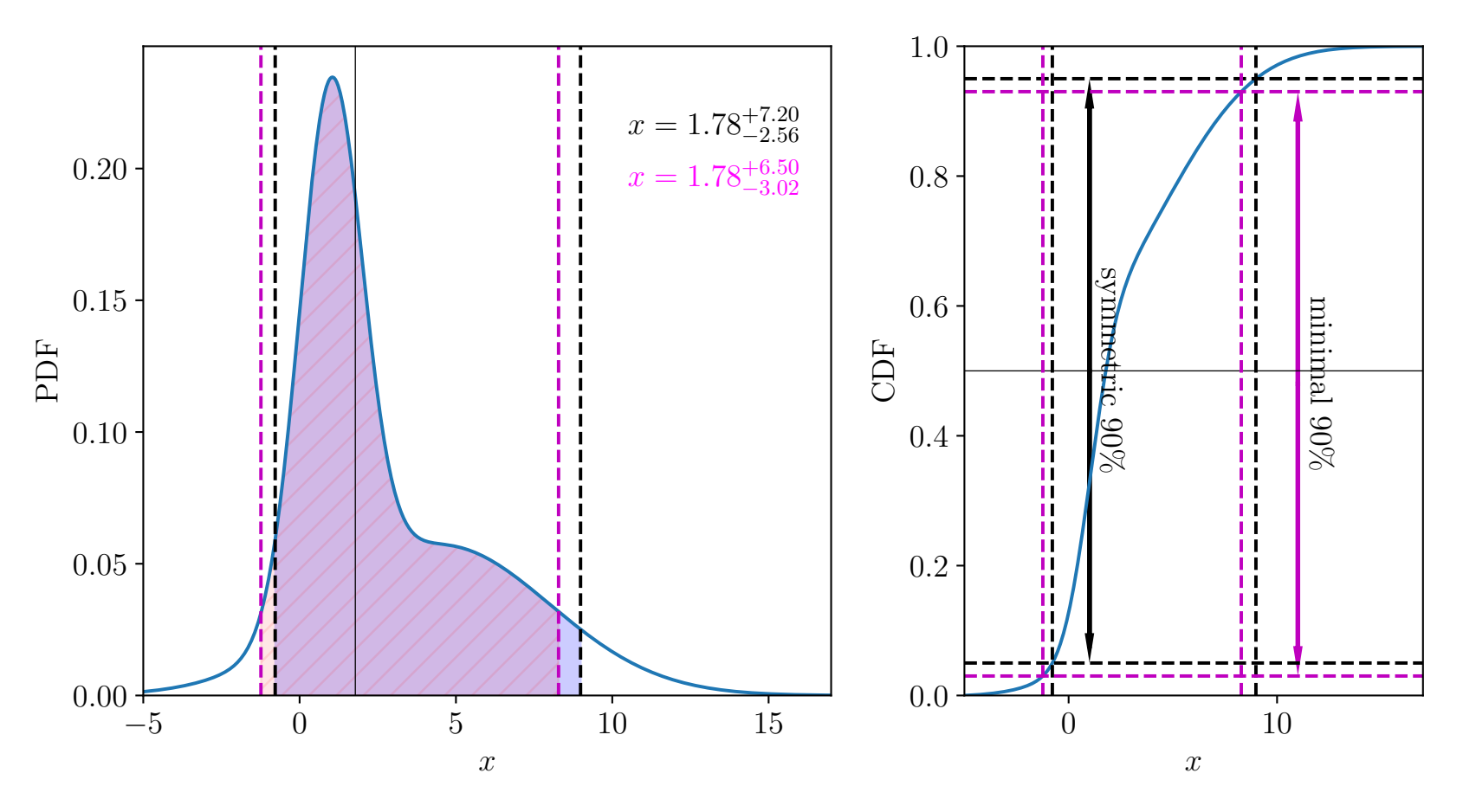
We define a **credible interval** $[\theta_a, \theta_b]$ as a (*non-unique*) range that a contains a certain amount of posterior probability, X,

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

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.*

BAYESIAN CREDIBLE INTERVAL



Two 90% credible intervals: symmetric in probability about the median (black), and the interval that spans the minimum range in x (magenta) (see `cred_int.py`).

In many situations in physics and astronomy our data d consists of the signal s with some additive noise n, 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 know only the expectation value μ and variance σ^2, the distribution that maximises the entropy, i.e., is the least informative, 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 $\boldsymbol{\theta}$, such that $s_1 \equiv s_1(\boldsymbol{\theta})$, then due to the additive nature of the noise and signal, the expectation value $\mu = s_1(\boldsymbol{\theta})$, and we have our Gaussian likelihood

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

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 $\mathbf{d} = \{d_1, d_2, \ldots, d_N\}$

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

In this case, with a fixed value of σ, the noise is drawn from a strictly (or strongly) stationary process.

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 as the likelihood:

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

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

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 no signal). If we assume N evenly sampled noise data points $\boldsymbol{n}$, then the autocovariance is:

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

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., $\bar{\gamma}_j = (1/M) \sum_{i=1}^{M} \gamma_{ji}$.

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:

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

EXAMPLE: FITTING A LINE

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 estimation, e.g., least squares fitting, 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

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

and we can evaluate the posterior for 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 practice, when $p(m, c|I) =$ constant and $\sigma_i = \sigma$ are constant, we can just calculate the posterior[5] over a grid in $\boldsymbol{m} = \{m_1, \ldots, m_j, \ldots, m_{N_m}\}$ and $\boldsymbol{c} = \{c_1, \ldots, c_k, \ldots, c_{N_C}\}$

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

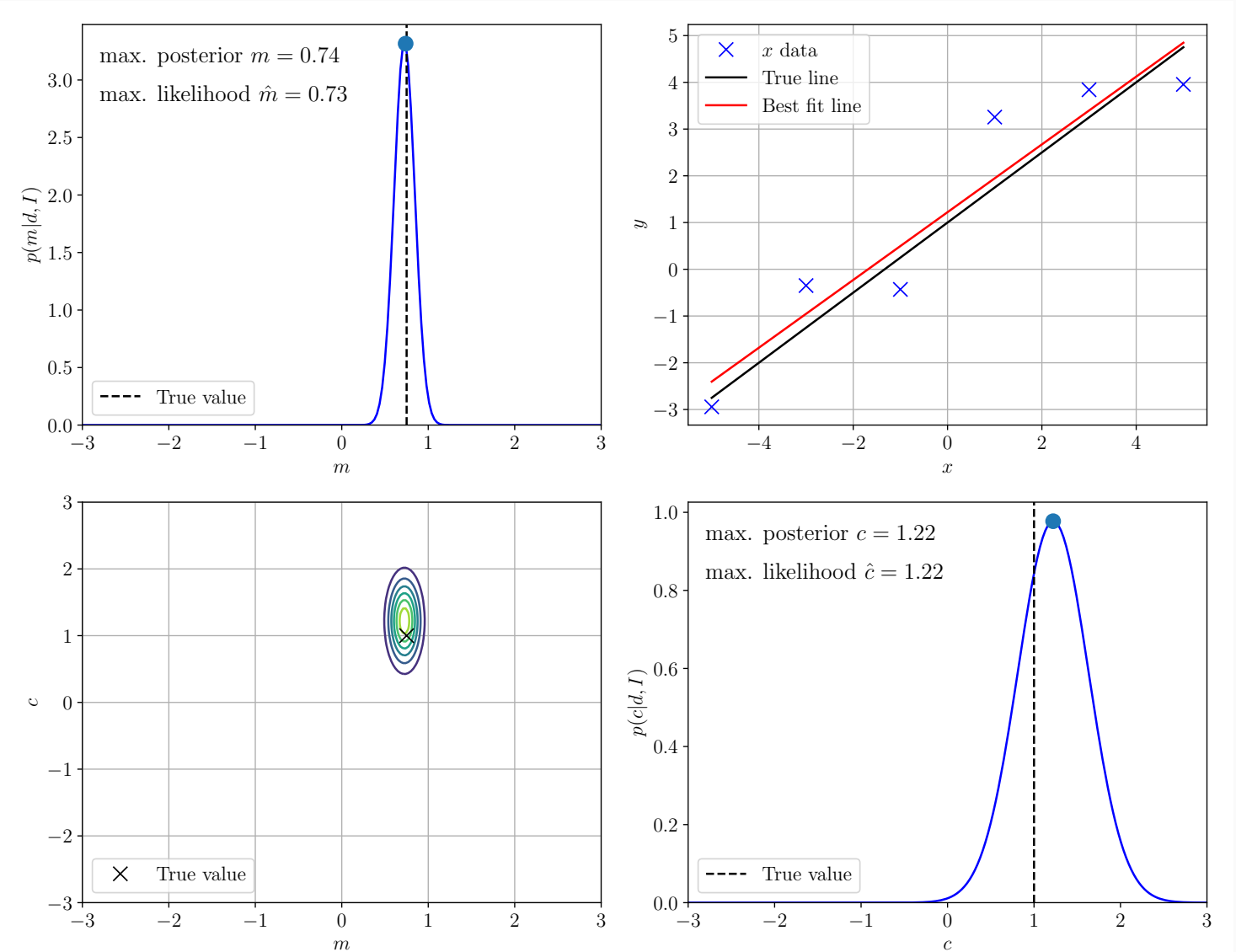
and get the marginal posteriors through numerical integration, e.g.,

$$p(m_j|d, I) \propto \sum_{k}^{N_c} \exp\left(\ell(m_j, c_k) - \max\ell(m, c)\right)\Delta c,$$

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

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

EXAMPLE: FITTING A LINE



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., it is 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 $\boldsymbol{\theta} = \{\alpha, \delta, \psi, \boldsymbol{\theta'}\}$ (see talk by Sebastian Khan).

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

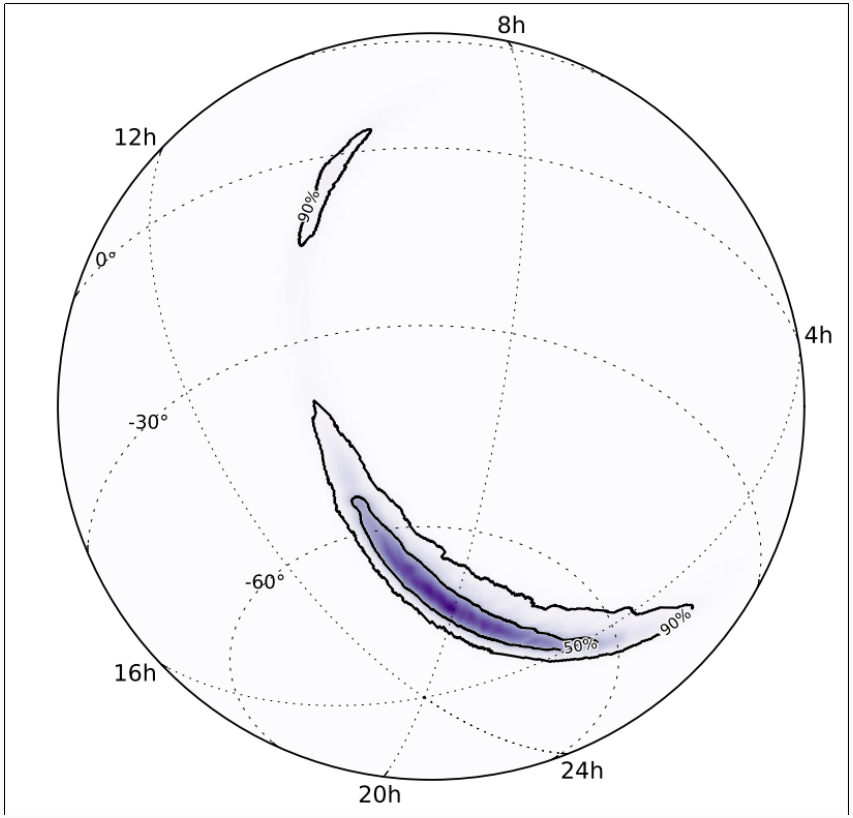
$$p(\alpha, \delta|\boldsymbol{d}, I) \propto \int^{\boldsymbol{\theta}_{\notin\{\alpha,\delta\}}} p(\boldsymbol{d}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)\mathrm{d}\boldsymbol{\theta}_{\notin\{\alpha,\delta\}},$$

or the source masses for a compact binary coalescence event

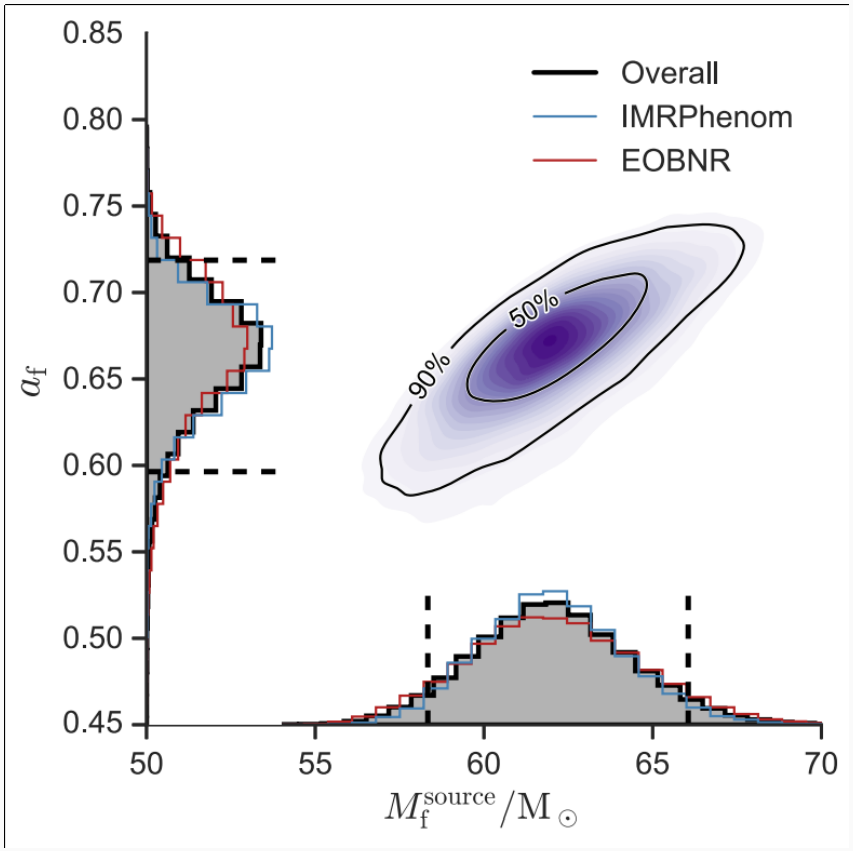
$$p(m_1, m_2|\boldsymbol{d}, I) \propto \int^{\boldsymbol{\theta}_{\notin\{m_1,m_2\}}} p(\boldsymbol{d}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)\mathrm{d}\boldsymbol{\theta}_{\notin\{m_1,m_2\}}.$$

So, we need to have a likelihood for the data and a prior for the source parameters.

INFERENCE IN GRAVITATIONAL-WAVE ASTRONOMY



Sky location posterior for GW150914 (Abbott et al., 2016).



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

Examples of inference in gravitational-wave astronomy

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

- 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 (particularly for transient sources) you might see (e.g., Equations 3 and 4 of Biwer et al., 2019, assuming a single detector, and ignoring the normalisation)

$$p(\boldsymbol{d}|\boldsymbol{\theta}, I) \propto \exp\left(-\frac{1}{2}\overbrace{\langle \tilde{d}(f) - \tilde{s}(f;\boldsymbol{\theta})|\tilde{d}(f) - \tilde{s}(f;\boldsymbol{\theta})\rangle}^{\text{noise weighted innner product}}\right)$$
$$\equiv \exp\left(-\frac{1}{2}\left[4\Re\int_0^\infty \frac{\left(\tilde{d}(f) - \tilde{s}(f;\boldsymbol{\theta})\right)\left(\tilde{d}(f) - \tilde{s}(f;\boldsymbol{\theta})\right)^*}{S_n(f)}\mathrm{d}f\right]\right),$$

where $\boldsymbol{d}$ is the data, $\boldsymbol{\theta}$ is a set of parameters defining a waveform model s, the tilde represents the Fourier transform (i.e., it is working in the frequency domain rather than with the original time domain data), and $S_n(f)$ 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

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

becomes

$$p(\boldsymbol{d}|\boldsymbol{\theta}, I) \propto \exp\left(-\frac{1}{2}\left[4\Re\sum_{i=0}^{k} \frac{\left(\tilde{d}_i - \tilde{s}_i(\boldsymbol{\theta})\right)\left(\tilde{d}_i - \tilde{s}_i(\boldsymbol{\theta})\right)^*}{TS_n(f_i)}\right]\right),$$

due to $\int \ldots \mathrm{d}f \approx \sum \ldots \Delta f$, and $\Delta f = 1/T$ for data of length T seconds, and i is the index over frequency bins.

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

$$p(\boldsymbol{d}|\boldsymbol{\theta}, I) \propto \exp\left(-\frac{1}{2}\left[4\sum_{i=0}^{k}\frac{\left(\Re(\tilde{d})_i - \Re(\tilde{s}_i(\boldsymbol{\theta}))\right)^2 + \left(\Im(\tilde{d}_i) - \Im(\tilde{s}_i(\boldsymbol{\theta}))\right)^2}{TS_n(f_i)}\right]\right),$$

$$\equiv \exp\left(-\frac{1}{2}\left[4\sum_{i=0}^{k}\frac{\left|\tilde{d}_i - \tilde{s}_i(\boldsymbol{\theta})\right|^2}{TS_n(f_i)}\right]\right)$$

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* likelihood of these independent data sets.

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

$$\ln p(\boldsymbol{d}|\boldsymbol{\theta}, I) = -\frac{1}{2}\left[4\sum_{i=0}^{k}\frac{\left(\Re(\tilde{d}_i) - \Re(\tilde{s}_i(\boldsymbol{\theta}))\right)^2 + \left(\Im(\tilde{d}_i) - \Im(\tilde{s}_i(\boldsymbol{\theta}))\right)^2}{TS_n(f_i)}\right] + C, \tag{2}$$

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

$$C = -\frac{1}{2}\sum_{i=0}^{k}\ln\left(\pi\, TS_n(f_i)/2\right).$$

If we expand out the quadratic terms we get:

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

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

- the *optimal* signal-to-noise:

$$\rho^2_{\rm opt} = (4/T)\sum_i \left(\Re(s_i)^2 + \Im(s_i)^2\right)/S_n(f_i),$$

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

$$\rho^2_{\rm mf} = (4/T)\sum_i \left(\Re(d_i)\Re(s_i) + \Im(d_i)\Im(s_i)\right)/S_n(f_i).$$

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

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

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

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

Evaluating the likelihood over the parameter space essentially requires evaluating these ρ^2 terms. For $s \gg n$, we see that the log likelihood ratio tends to $\rho^2/2$.

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[6]

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

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 $\sigma_i^2 = (T/4) S_n(f_i)$, which when substituted into Equation (2) gives the standard Gaussian log-likelihood.

[6] This is equivalent to the Fourier transform of the noise autocovariance function.

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 segment is multiplied by a window, Fourier transformed, and the average power from all segments is used:

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

Windowing is *vital* (when Fourier tranforming the analysis segment *and* for the PSD estimation) to prevent spectral leakage and the addition of correlations to the data.

TIME DOMAIN LIKELIHOOD

For broadband signals, like those from compact binary coalescences, 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 in the likelihood, 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 advantages of the time-domain are:

- you do not have to worry about windowing(!);
- time delays between detectors are frequency indepdent, rather than applied as 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

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

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

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

where $\boldsymbol{\theta}'$ contains all the parameters of $\boldsymbol{\theta}$ except ϕ_0, and I_0 is the modified Bessel function of the first kind (see, e.g., Equation (20) of Veitch et al., 2015).

What if one, or several, components of our prior are themselves parameterised and we also want to infer these parameters (known as hyperparameters)? For example, suppose we have a parameter m for which our prior is a Gaussian distribution, with an *a priori* unknown the mean and standard deviation, then our posterior would be (using the product rule):

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

We now, of course need to define priors on the hyperparameters!

This can be used for population-level inference. For example, if you want to known the underlying distribution of primary black hole masses, m_1, in binary black hole systems then you might define a parameterised prior on m_1, such as a power law, then define a prior on the power law spectral index α, and use a joint likelihood that is the product of likelihoods for many (say N) individual observed signals:

$$p(\alpha|\boldsymbol{\mathcal{D}}, I) = \left[\prod_{i=1}^{N} \int^{\boldsymbol{\theta}_i} p(\boldsymbol{d}_i|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}'_i|I) p(m_{1\,i}|\alpha, I) \mathrm{d}\boldsymbol{\theta}_i\right] p(\alpha|I),$$

where $\boldsymbol{\theta}'_i$ are the parameters for each individual source, and contain all the parameters except $m_{1\,i}$.

BIBLIOGRAPHY

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. W. Hogg, J. Bovy, D. Lang Data analysis recipes: Fitting a model to data, arXiv:1008.4686, 2010

E. Thrane & C. Talbot An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models *Publ. Astron. Soc. Aus.*, 36, e010, 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* $P(A \text{ or not}(A)) = 1$
- *Law of Non-contradiction* $P(A \text{ and not}(A)) = 0$
  - 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)
  - $P(\text{not } [A, B]) = P(\text{not}(A) \text{ or not}(B))$
  - $P(\text{not } [A \text{ or } B]) = P(\text{not}(A), \text{not}(B))$

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

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

APPENDIX: FITTING A LINE (UNKNOWN σ)

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

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

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.

APPENDIX: FITTING A LINE (UNKNOWN σ)