Part II: Bayesian Statistics

5 Bayesian Theory

As we have seen, in frequentist statistics statements are made with reference to repetitions of the same experiment with parameters fixed. In Bayesian statistics, parameters are no longer regarded as fixed, but are themselves random variables. The probability distribution of the parameter values before taking data, the **prior distribution**, is updated to a probability distribution after taking data, the **posterior distribution**, through the likelihood of the observed data. This update is achieved through **Bayes' Theorem**. Bayesian inference attempts to say as much as possible about the unknown parameter distribution based on the observed data only, without reference to future repetitions of the same experiment. Bayesian posteriors are probability distributions on the unknown parameter and can be interpreted and manipulated in that way, as statements about the relative probability that the parameter takes different values.

The derivation of Bayes' theorem is a mathematical result that follows from the definition of conditional probability, as we will see below, but it is how this result is applied to interpret data, and the philosophical distinction in the interpretation of the parameter values that distinguishes the frequentist and Bayesian approach. Typically, in any given observation, the actual parameter values that led to the generation of the observed data are fixed, not random, but the Bayesian interpretation is that you can never by sure of what the unknown parameter is, and so it is appropriate to consider it to be a random variable. In many cases you will not be able to repeat a particular experiment. Gravitational wave observations are a good example of this — we cannot choose what events occur in the Universe, so every observed event is a unique, non-repeatable, experiment. In such contexts, the frequentist approach of referencing theoretical repetitions cannot really be seen as representative of reality. In cases where it is possible to repeat an experiment with the unknown parameters fixed, the Bayesian posterior converges to the true parameter value asymptotically and so can still be used to represent the current level of uncertainty in the parameter.

Frequentist concepts such as significance and hypothesis testing have been incorporated into the Bayesian framework, but the interpretation in the latter context is not always clean. It is therefore useful to have familiarity with both sets of tools to be fully quipped to handle any kind of data analysis problem.

5.1 Conditional probability

It is often the case that a process generates more than one potentially measurable random output, but only a subset of these are measurable. If the variables are independent then measuring one would not provide any information about the others, but when there are inter-dependencies the observation of a random variable can provide information about other variables with which it is correlated. For example, suppose we have a bag containing 100 balsa, of which 10 are red and stripy, 20 are blue and stripy, 30 are red and spotted and 40 are blue and spotted. In total there are 30 stripy balls out of the 100 and therefore the probability that a randomly chosen ball is stripy is 3/10. However, out of the 40 red balls there are only 10 that are stripy, and so if we have observed that the ball is red the probability that it is also stripy is now 1/4.

The **conditional probability** of an event A, given some other event B is defined as

$$p(A|B) = \frac{p(A \cap B)}{B}.$$

In other words, this is the fraction that both A and B occur, our of all the times that B occurs. This can be rewritten in two different ways by interchanging A and B

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A).$$

Rearranging this identity we obtain **Bayes' Theorem**

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

5.2 Bayesian inference

Bayes' Theorem is a mathematical identity, but it becomes philosophically distinct from frequentist approaches when it is applied to inference. In Bayesian inference, the event A is taken to be an observation of data, $\mathbf{x}$, and the event B is taken to be the value of some unknown parameters, $\vec{\theta}$, characterising the system being observed. Bayes' Theorem becomes

$$p(\vec{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\vec{\theta})p(\vec{\theta})}{p(\mathbf{x})}.$$

In this context $p(\mathbf{x}|\vec{\theta})$ is the likelihood (the same function of data and parameters as in the frequentist case), $p(\vec{\theta})$ is the **prior** distribution of source parameter values, $p(\vec{\theta}|\mathbf{x})$ is the **posterior** distribution on the source parameter values and $p(\mathbf{x})$ is the **evidence** for the model under consideration. In a parameter estimation context, the evidence, which does not depend on parameter values, is a normalisation constant that can be ignored. However, it plays an important role in Bayesian hypothesis testing, which will be discussed in section 5.6.

Example: Medical testing We suppose that a medical test for a disease is 95% effective but has a 1% false alarm rate and the prevalence of the disease in the population is 0.5%. You test positive for the disease. What is the probability you do in fact have it?

The term "95% effective" means that if you have the disease the test gives a positive result 95% of the time. The term 1% false alarm rate means that if you do not have the disease you test positive 1% of the time. We can now apply Bayes theorem with data $\mathbf{x}$ = 'positive test' and parameter θ ='disease status' taking values 'infected' or 'not infected'. The likelihood is

$$p(\text{positive}|\text{infected}) = 0.95, \qquad p(\text{positive}|\text{not infected}) = 0.01.$$

The prior is based on the known prevalence in the population

$$p(\text{infected}) = 1 - p(\text{not infected}) = 0.005.$$

The posterior is then

$$\begin{aligned} p(\text{infected}|\text{positive}) &= \frac{p(\text{positive}|\text{infected})p(\text{infected})}{p(\text{positive}|\text{infected})p(\text{infected}) + p(\text{positive}|\text{not infected})p(\text{not infected})} \\ &= \frac{0.95 * 0.005}{0.95 * 0.005 + 0.01 * 0.995} = 0.323. \end{aligned} \tag{69}$$

So you are more likely not to be infected than to be infected if you get a positive test result. The solution is to get a second opinion. If you take a second (independent) test and it is also positive your posterior probably of being infected is now

$$p(\text{infected}|\text{2nd positive}) = \frac{0.95 * 0.323}{0.95 * 0.323 + 0.01 * 0.677} = 0.978 = \frac{0.95^2 * 0.005}{0.95^2 * 0.005 + 0.01^2 * 0.995}.$$

The first of these two results follows from using the posterior from the first test as a prior for the second. The second result follows from regarding the observed data as "two independent positive tests".

Example: Blood evidence Based on other evidence, a detective is 50% sure that a particular suspect has committed a murder. Then new evidence comes to light. A small amount of blood, of type B, is found at the scene. This is not the victim's blood type, but it is the blood type of the suspect. Such a blood type has a prevalence of 2% in the population. What is the detective's confidence in the guilt of the suspect in light of this new evidence?

The likelihood is

$$p(\text{type B blood}|\text{guilty}) = 1, \qquad p(\text{type B blood}|\text{not guilty}) = 0.02.$$

The prior is $p(\text{guilty}) = 0.5$ and so the posterior is

$$\begin{aligned} p(\text{guilty}|\text{type B blood}) &= \frac{p(\text{type B blood}|\text{guilty})p(\text{guilty})}{p(\text{type B blood}|\text{guilty})p(\text{guilty}) + p(\text{type B blood}|\text{not guilty})p(\text{not guilty})} \\ &= \frac{0.5}{0.5 + 0.01} = 0.98. \end{aligned} \tag{70}$$

5.3 Choice of prior

The prior plays a key role in Bayesian parameter inference. It expresses the current state of our understanding about parameter values, and it is updated to the posterior using data via the likelihood. Mathematically, the prior represents the distribution of the unknown parameter value in nature, but usually this is not known. In that case, the prior reflects the current state of knowledge about the parameter values, which may come from previous experiments or expert opinion or not be known.

5.3.1 Informative/expert priors

If information is available, it is appropriate to use informative priors. For example, if previous measurements have been made of a quantity it is reasonable to use the posterior from those measurements as a prior for the next measurement, as we saw in the medical test example above. Alternatively, even if a measurement has not been made directly, "experts" may be able to give a reasonable range or distribution for the parameter based on experience in other situations. One criticism that is often levelled at Bayesian inference is that the result can depend on the assumed prior. However, the Bayesian response is that this is desired behaviour — if we have additional information from prior knowledge, then it is the correct thing to do to include that in our conclusions based on subsequent observed data.

The process of constructing a prior based on the opinion of experts is known as **elicitation**. Sometimes, elicitation may result in different priors from different experts. In that

case a **mixture prior** can be constructed

$$p(\vec{\theta}) = \sum_{j=1}^{J} \omega_j p_j(\vec{\theta})$$

where j labels which of the J experts we are referring to, $p_j(\vec{\theta})$ is the prior elicited from that expert, and ω_j is the weight given to that expert (or set of experts).

If the prior is based on the posterior from previous observations it is normally clear how to fold this in. If the prior comes from expert opinion, it may be possible to use this in several different ways. In that case, care must be taken to be as conservative as is reasonably possible in the use of that prior information, to avoid making conclusions form the data that are too strong.

5.3.2 Conjugate priors

It is convenient to choose a form for the prior that ensures the posterior takes the same form. In such situations, the posterior from an experiment can be directly be used as a prior for the next experiment and so on. Such a prior is called **conjugate**.

Definition: A family of distributions, $\mathcal{F}$, is **conjugate** to a family of sampling distributions, $\mathcal{P}$, if, whenever the prior belongs to the family $\mathcal{F}$, the posterior belongs to the same family, for any number and value of observations from $\mathcal{P}$.

The form of the conjugate prior depends on the nature of the probability distribution, $\mathcal{P}$, from which the observed data is drawn. This gives rise to a number of conjugate families. In particular, any distribution in the exponential family

$$p(x|\,\theta) = \exp\left\{\sum_{j=1}^{K} A_j(x)B_j(\vec{\theta}) + C(\vec{\theta}) + D(x)\right\} \;\forall x, \vec{\theta}$$

has a conjugate prior in the exponential family of the form

$$p(\vec{\theta}|\vec{\chi},\nu) = p(\vec{\chi},\nu)\exp\left[\vec{\theta}^T\vec{\chi} - \nu A(\vec{\theta})\right] \tag{71}$$

where ν and $\vec{\chi}$ are the hyperparameters of the prior distribution.

A full list of conjugate priors can be found in the conjugate prior entry on wikipedia, but the three most widely used are the Beta-Binomial, Poisson-Gamma and Normal-Normal families, and we will discuss these further here.

Beta-Binomial model Suppose our observed data $\mathbf{X} \sim$Bin(n,p) with likelihood

$$p(\mathbf{x}|p) = \left(\begin{array}{c} n \\ x \end{array}\right) p^x(1-p)^{n-x}.$$

The conjugate prior is the Beta(a,b) distribution with density

$$p(p) = \frac{1}{B(a,b)}p^{a-1}(1-p)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}.$$

Observing binomial distributed data and using the Beta prior gives a posterior

$$\begin{aligned} p(p \mid x) &\propto p(x \mid p)p(p) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \\ &\propto p^{a+x-1}(1-p)^{b+n-x-1}. \end{aligned}$$

So the posterior is also a Beta distribution

$$p(p \mid x) = \text{Beta}(a+x, b+n-x).$$

The mean and variance of a Beta(a, b) distribution are

$$\mathbb{E}(\mathbf{X}) = \frac{a}{a+b}, \qquad \text{var}(\mathbf{X}) = \frac{ab}{(a+b)^2(a+b+1)}.$$

The posterior mean is therefore

$$\mathbb{E}(p|x) = \frac{a+x}{a+b+n}$$

which we compare to the mean in the observed data of x/n. One interpretation of the prior data is that it represents having observed $a-1$ events in $a+b-2$ previous trials. If a and b are kept fixed and $n, x \to \infty$ the posterior mean tends to the maximum likelihood estimator x/n and the posterior variance tends to zero.

Poisson-Gamma model Suppose now that we are observing data, $X_1, \ldots, X_n$, from a Poisson distribution, $\mathbf{X} \sim$Pois(λ), with likelihood

$$p(\mathbf{x} \mid \lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i} \mathrm{e}^{-\lambda}}{x_i!} \right\}.$$

The conjugate prior is the Gamma(m, μ) distribution

$$p(\lambda|m, \mu) = \frac{1}{\Gamma(m)} \mu^m \lambda^{m-1} \mathrm{e}^{-\mu\lambda},$$

which has mean m/μ and variance m/μ^2. With this prior the posterior is

$$\begin{aligned} p(\lambda \mid \mathbf{x}) &\propto p(\mathbf{x}|\lambda)p(\lambda) \\ &= \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i} \mathrm{e}^{-\lambda}}{x_i!} \right\} \frac{1}{\Gamma(m)} \mu^m \lambda^{m-1} \mathrm{e}^{-\mu\lambda} \\ &\propto \mathrm{e}^{-n\lambda-\mu\lambda} \lambda^{\sum_{i=1}^n x_i + m - 1} \\ &\propto \text{Gamma}(m + n\bar{x}, \mu + n). \end{aligned} \tag{72}$$

The posterior mean can be seen to equal

$$\mathbb{E}(p(\lambda \mid \mathbf{x})) = \frac{m + n\bar{x}}{m+n} = \bar{x} \left(\frac{n}{n+m} \right) + \frac{m}{\mu} \left(1 - \frac{n}{n+m} \right),$$

i.e., it is a compromise between the prior mean, m/μ, and the maximum likelihood estimator $\bar{x}$. As the number of samples increases, more weight is placed on the data and less on the prior, as expected.

Normal-Normal/Normal-Gamma model Now we consider $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, and likelihood

$$p(\mathbf{x}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right].$$

We assume first that σ^2 is known. The conjugate prior in this case is the Normal distribution, $N(\mu_0, \sigma_0^2)$,

$$p(\mu \mid \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right].$$

The posterior is

$$\begin{aligned} p(\mu \mid \mathbf{x}, \sigma^2) &\propto p(\mathbf{x} \mid \mu, \sigma^2) p(\mu|\mu_0, \sigma_0^2) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right\} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2\sigma_0^2}\left[\mu^2(n\sigma_0^2 + \sigma^2) - 2\mu(n\bar{y}\sigma_0^2 + \mu_0\sigma^2)\right]\right\}, \end{aligned}$$

which can be recognized as a $N(\mu_n, \sigma_n^2)$ distribution, where

$$\mu_n = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \qquad \sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}. \tag{73}$$

Writing these results in terms of $\tau = 1/\sigma^2$, which is called the **precision** of the Normal distribution we can see

$$\mu_n = \frac{\tau_0}{\tau_0 + n\tau}\mu_0 + \frac{n\tau}{\tau_0 + n\tau}\bar{y}$$

so once again the posterior mean is a balance between the prior mean and the sample mean, with the relative weighting determined by both the number of observations and the relative precision of the observations and the prior.

If we suppose that μ is known (which is an unrealistic assumption in practice), but the variance is uncertain, then we can obtain a conjugate prior by using a Gamma(a, b) prior on the precision

$$p(\tau|a, b) \propto \tau^{a-1}e^{-b\tau}$$

and obtain the posterior

$$\begin{aligned} p(\tau \mid \mathbf{x}, \mu) &\propto p(\mathbf{x} \mid \mu, \tau) p(\tau|a, b) \\ &\propto \tau^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \tau^{a-1}e^{-b\tau} \\ &= \tau^{a+n/2-1} \exp\left\{-\tau\left(b + \frac{1}{2}\sum_i (x_i - \mu)^2\right)\right\} \\ &\sim \text{Gamma}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right). \end{aligned}$$

It is common practice to take the limit in which a and b are both very small and then the posterior becomes

$$p(\tau \mid \mathbf{x}, \mu) = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \quad \Rightarrow \quad \mathbb{E}\left[\tau \mid \mathbf{x}, \mu\right] = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right)^{-1},$$

so the posterior expectation of the precision is approximately the same as the (frequentist) sample precision (up to a factor of $n/(n-1)$).

Finally we assume that both μ and σ^2 are unknown. It would be reasonable to just multiply together the two previous priors, but this does not result in a conjugate prior, essentially because the posterior on μ in the first case depends on the known variance σ^2. However, we can find a correlated conjugate prior (writing $\tau = 1/\sigma^2$ as before) by writing

$$\mu \sim N(\mu_0, 1/(n_0\tau)), \quad \tau \sim \text{Gamma}(a, b),$$

or, explicitly,

$$p(\mu, \tau|\mu_0, n_0, a, b) \propto \left(\frac{n_0\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{n_0\tau}{2}(\mu - \mu_0)^2\right] \tau^{a-1}e^{-b\tau}.$$

The posterior on μ, conditioned on τ, $p(\mu|\tau, \mathbf{x})$, is given by the same expression as before

$$p(\mu|\tau, \mathbf{x}) \sim N\left(\frac{n_0\mu_0 + n\bar{x}}{n_0 + n}, \frac{1}{(n_0 + n)\tau}\right).$$

The posterior on τ can be found by considering the combined posterior, being careful not to drop any terms that depend on μ or τ

$$\begin{aligned} p(\mu, \tau|\mathbf{x}) &\propto \sqrt{\tau}\exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \tau^{\frac{n}{2}} \exp\left[-\frac{n_0\tau}{2}(\mu - \mu_0)^2\right] \tau^{a-1}e^{-b\tau} \\ &= \tau^{a+\frac{n}{2}-1}\exp\left[-\left(b - \frac{(n\bar{x} + n_0\mu_0)^2}{2(n + n_0)} + \frac{1}{2}n_0\mu_0^2 + \frac{1}{2}\sum x_i^2\right)\tau\right] \times \\ &\quad \times \left(\sqrt{\frac{(n + n_0)\tau}{2\pi}}\exp\left[-\frac{(n + n_0)\tau}{2}\left(\mu - \frac{(n\bar{x} + n_0\mu_0)}{n + n_0}\right)^2\right]\right). \end{aligned} \tag{74}$$

If we now marginalise over μ, the round bracketed term on the final line integrates to a constant, independent of τ, and the term inside the exponent on the penultimate line can be simplified to obtain

$$\begin{aligned} p(\tau|\mathbf{x}) &\propto \tau^{a+\frac{n}{2}-1}\exp\left[-\left(b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{nn_0}{2(n + n_0)}(\mu_0 - \bar{x})^2\right)\tau\right] \\ &\Rightarrow p(\tau|\mathbf{x}) \sim \text{Gamma}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{nn_0}{2(n + n_0)}(\mu_0 - \bar{x})^2\right). \end{aligned} \tag{75}$$

And so this is also a conjugate prior model, called the Normal-Gamma model.

5.3.3 Using expert information with conjugate priors

If expert prior information is in the form of a posterior from a previous experiment the form of the distribution is fixed. However, in other circumstances it can be possible to express the prior information in the form of a particular choice of parameters for a conjugate prior. This is most clearly seen with an example.

Example: Consider a drug to be given for relief of chronic pain. Experience with similar compounds has suggested that response rates, p, between 0.2 and 0.6 could be feasible. We plan to observe the response rate in n patients and want to infer a posterior on p. Propose a suitable conjugate prior for p based on the available information.

A response rate between 0.2 and 0.6 could be used to set a uniform prior in that range. However, this is not conjugate to the binomial distribution that determines the observed data. Therefore, it would be better to use a conjugate prior. A $U[0.2, 0.6]$ distribution has mean 0.4 and standard deviation of 0.1. We can find a Beta distribution that has the same mean and standard deviation. Rearranging the equations given earlier we deduce $\text{Beta}(a = 9.2, b = 13.8)$ has the desired mean and variance. This prior is conjugate and reflects the expert opinion as regards the expected response rate for the drug. Suppose now we observe $n = 20$ patients and $x = 15$ respond positively. The posterior is then $\text{Beta}(9.2 + 15, 13.8 + 5) = \text{Beta}(24.2, 18.8)$. The prior, (scaled) likelihood and posterior are illustrated in Figure 8.

5.3.4 Mixture priors

The use of a conjugate prior can be somewhat restrictive as there is limited flexibility within the prior family. However, one way to get around this is by using **mixture priors**. A mixture prior is of the form

$$p(\vec{\theta}) = \sum_{j=i}^{J} \pi_j p(\vec{\theta} \mid \vec{\psi}_j), \quad \sum_{j=1}^{J} \pi_j = 1. \tag{76}$$

Here $\{\pi_j\}$ are called the mixture weights and it is assumed that the hyperparameters, ψ_j, are different in each component. If the mixture components are all drawn from the conjugate prior family, then the mixture prior is also conjugate.

Example: Beta-Binomial mixture prior Suppose $X \sim \text{Bin}(n, p)$ and we use a prior on p that is a mixture distribution

$$p(p|a_1, b_1, a_2, b_2) = \pi \text{Beta}(a_1, b_1) + (1 - \pi)\text{Beta}(a_2, b_2).$$

What is the posterior distribution for p?

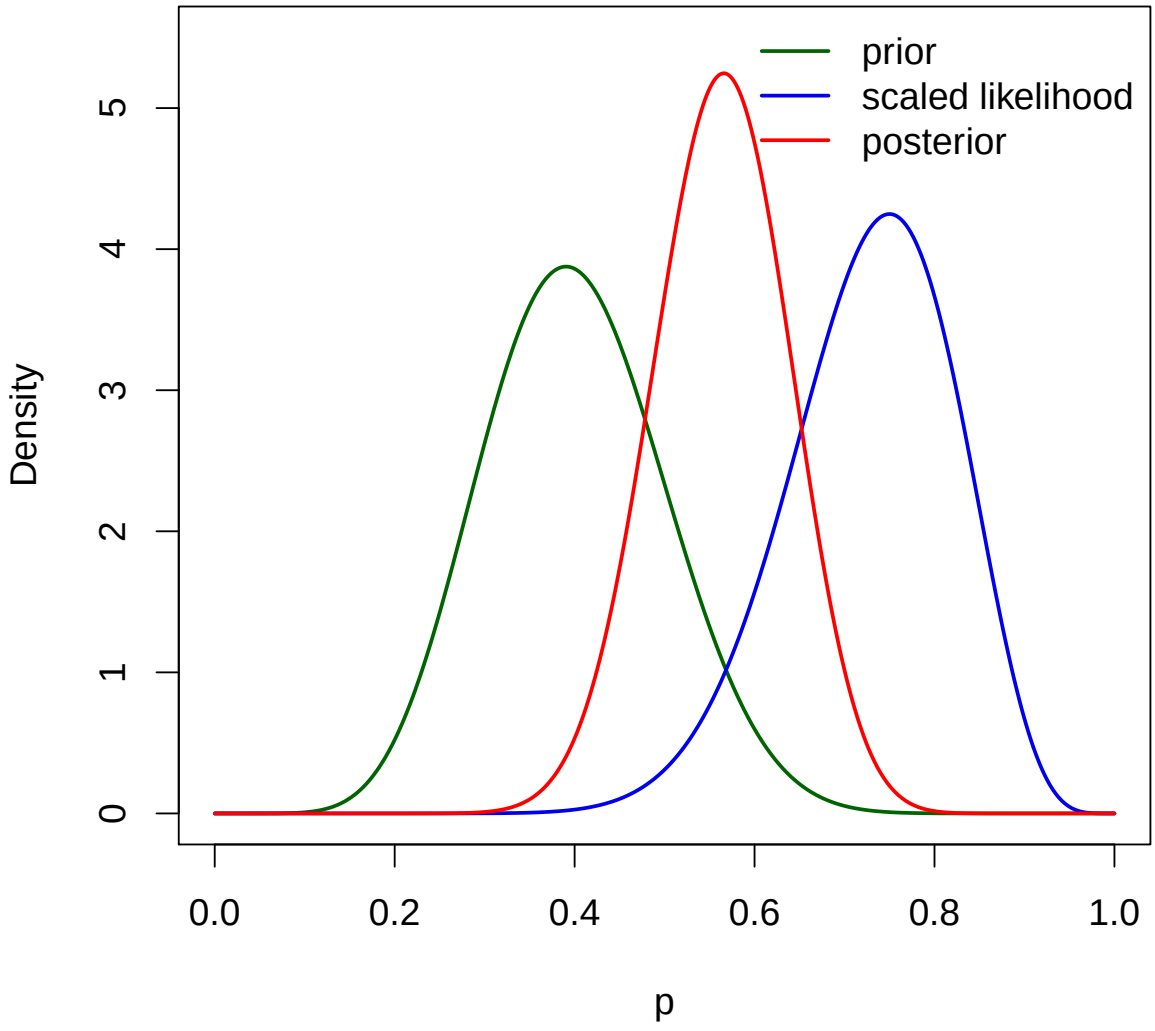


Figure 8: Conjugate prior, Beta$(9.2, 13.8)$, likelihood, Bin$(20, p)$, and posterior, Beta$(24.2, 18.8)$ for the drug response problem described in the text. The likelihood has been rescaled to ensure it has a similar height to the prior and posterior distributions.

Solution: We find the posterior as follows

$$\begin{aligned}
p(p \mid x) &\propto \binom{n}{x} p^x (1-p)^{n-x} \left\{ \pi \frac{1}{B(a_1, b_1)} p^{a_1 - 1} (1-p)^{b_1 - 1} + (1-\pi) \frac{1}{B(a_2, b_2)} p^{a_2 - 1} (1-p)^{b_2 - 1} \right\} \\
&\propto \pi \frac{1}{B(a_1, b_1)} p^{a_1 + x - 1} (1-p)^{b_1 + n - x - 1} + (1-\pi) \frac{1}{B(a_2, b_2)} p^{a_2 + x - 1} (1-p)^{b_2 + n - x - 1} \\
&= \pi \frac{B(a_1 + x, b_1 + n - x)}{B(a_1, b_1)} \frac{1}{B(a_1 + x, b_1 + n - x)} p^{a_1 + x - 1} (1-p)^{b_1 + n - x - 1} \\
&\quad + (1-\pi) \frac{B(a_2 + x, b_2 + n - x)}{B(a_2, b_2)} \frac{1}{B(a_2 + x, b_2 + n - x)} p^{a_2 + x - 1} (1-p)^{b_2 + n - x - 1} \\
&= \pi \frac{B(a_1 + x, b_1 + n - x)}{B(a_1, b_1)} \text{Beta}(p \mid a_1 + x, b_1 + n - x) \\
&\quad + (1-\pi) \frac{B(a_2 + x, b_2 + n - x)}{B(a_2, b_2)} \text{Beta}(p \mid a_2 + x, b_2 + n - x).
\end{aligned}$$

We finish by normalising the weights to obtain

$$p \mid x \sim \omega_1 \text{Beta}(p \mid a_1 + x, b_1 + n - x) + (1 - \omega_1) \text{Beta}(p \mid a_2 + x, b_2 + n - x)$$

with

$$\omega_1 = \pi \frac{B(a_1 + x, b_1 + n - x)}{B(a_1, b_1)} \left(\pi \frac{B(a_1 + x, b_1 + n - x)}{B(a_1, b_1)} + (1-\pi) \frac{B(a_2 + x, b_2 + n - x)}{B(a_2, b_2)} \right)^{-1}$$

So the posterior is also a mixture of Beta distributions.

5.3.5 Jeffreys prior

If we do not have any prior information, it is normal to use an "uninformative" prior, i.e., a prior that assumes as little as possible about the parameter values. It is common to use uniform priors as uninformative priors, so that the posterior basically corresponds to the likelihood of the data. This is approach taken for many parameters in parameter estimation of gravitational wave data and was in fact the approach that Bayes himself advocated. However, uniform priors are not invariant under re-parameterisation. If one is ignorant about the value of θ, one is also ignorant about the value of θ^2 or any other function of θ. Therefore, any uninformative prior should induce the same form of uninformative prior on any other variables defined by transformation. Jeffreys (1961) proposed a class of priors that are invariant under re-parameterisations. By identifying the probability density with a metric on parameter space he argued that the prior should take the form $[\det(g_{ij})]^{1/2}$ where the metric

$$g_{ij}(\vec{\theta}) = \frac{1}{f(\vec{\theta})} \frac{\partial f}{\partial \theta_i} \frac{\partial f}{\partial \theta_j}.$$

This would lead to an invariant prior for any scalar function $f(\vec{\theta})$. Jeffreys advocated the use of the likelihood, which introduces a data dependence into the expression, that can be eliminated by taking the expectation over realisations of the data. This procedure leads to **Jeffreys prior** which is

$$p(\vec{\theta}) \propto \sqrt{\det[I(\vec{\theta})]}, \qquad \text{where } I(\vec{\theta})_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} \right]$$

for $l = \log p(\mathbf{x}|\vec{\theta})$ the log-likelihood is the Fisher information matrix.

Jeffreys prior is "uninformative" because it can be interpreted as being as close as possible to the likelihood function and it is invariant under re-parameterisation. However, it is rarely a member of the conjugate family of distributions or of some other convenient form which is why it is not always convenient to use it in practice. Note also that the Jeffreys prior is not always **proper**, i.e., it does not always have a finite integral and therefore may not be normalisable.

Example: Poisson distribution For a single observation, x, from the Poisson(λ) distribution with pmf

$$p(x|\lambda) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}$$

we have

$$\frac{\partial \log p}{\partial \lambda} = \frac{x}{\lambda} - 1, \quad \frac{\partial^2 \log p}{\partial \lambda^2} = -\frac{x}{\lambda^2} \quad \Rightarrow \quad I(\lambda) \equiv \mathbb{E}\left[-\frac{\partial^2 \log p}{\partial \lambda^2}\right] = \frac{1}{\lambda}.$$

The Jeffreys prior for the Poisson distribution is therefore $p(\lambda) \propto 1/\sqrt{\lambda}$. This is an example of an **improper** prior, since it cannot be normalised to integrate to 1 unless the range of rates is restricted.

5.4 Posterior summary statistics

The result of a Bayesian inference calculation is a probability distribution, the full posterior probability distribution of the parameters, $p(\vec{\theta}|\mathbf{x})$. This is not only difficult to calculate in many cases, it is also unwieldy to manipulate and so it is common to use quantities that summarise the properties of the distribution. These are all of the summary statistics that we encountered in the first chapter of the course.

5.4.1 Point estimates

To obtain point estimates of a parameter value, θ_1 say, one typically works with the **marginalised** distribution for that parameter, defined by

$$p_{\mathrm{marg}}(\theta_1|\mathbf{x}) = \int p(\vec{\theta}|\mathbf{x}) \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_m.$$

From this marginal distribution, we can evaluate the **posterior mean**

$$\mu = \int_{-\infty}^{\infty} \theta_1 p_{\mathrm{marg}}(\theta_1|\mathbf{x}) \mathrm{d}\theta_1$$

or the **posterior median**, m, defined such that

$$\int_{-\infty}^{m} p_{\mathrm{marg}}(\theta_1|\mathbf{x}) \mathrm{d}\theta_1 = 0.5 = \int_{m}^{\infty} p_{\mathrm{marg}}(\theta_1|\mathbf{x}) \mathrm{d}\theta_1$$

or the **posterior mode**

$$M = \mathrm{argmax}\ p_{\mathrm{marg}}(\theta_1|\mathbf{x}).$$

The posterior mean and mode can be defined unambiguously over the full distribution as well. The posterior mean is the same whether computed over the marginal distribution or the full distribution, but the mode typically changes. The median is not unambiguously defined on the whole distribution, as there are infinitely many ways to partition the full parameter space into equal probability subsets.

5.4.2 Credible intervals

To move beyond point estimates, it is natural to want to describe ranges in which parameter values are estimated to lie. The Bayesian equivalent of a frequentist confidence interval is a **credible interval**. This is defined as

Definition: An interval (a, b) is a $100(1-\alpha)\%$ posterior credible interval for θ_1 if

$$\int_a^b p_{\rm marg}(\theta_1|\mathbf{x})\mathrm{d}\theta_1 = (1-\alpha), \quad 0 \leq \alpha \leq 1.$$

A **credible region** can be defined in a similar way. This is any partition of parameter space that contains $100(1-\alpha)\%$ of the total posterior probability. Clearly credible intervals and regions are not unique, but there are two types of credible interval that are commonly used.

Definition: An interval (a, b) is a **symmetric** $100(1-\alpha)\%$ posterior credible interval for θ_1 if

$$\int_{-\infty}^a p_{\rm marg}(\theta_1|\mathbf{x})\mathrm{d}\theta_1 = \frac{\alpha}{2} = \int_b^{\infty} p_{\rm marg}(\theta_1|\mathbf{x})\mathrm{d}\theta_1.$$

Definition: An interval (a, b) is a $100(1-\alpha)\%$ **highest posterior density (HPD) interval** for θ_1 if

1. $[a, b]$ is a $100(1-\alpha)\%$ credible interval for θ_1;

2. for all $\theta \in [a, b]$ and $\theta' \notin [a, b]$ we have $p_{\rm marg}(\theta|\mathbf{x}) \geq p_{\rm marg}(\theta'|\mathbf{x})$.

Credible intervals are more intuitive than confidence intervals as they make an explicit statement about the probability that the parameter takes values in the range, rather than referencing an ensemble of similar experiments.

5.4.3 Posterior samples

Summary statistics provide a useful way to summarise and compare distributions, but they inevitably discard information. To retain full information about the parameters we need the full posterior. Often this cannot be written down in a simple analytic form, but it can be summarised by drawing a set of samples $\{\vec{\theta}_1, \ldots, \vec{\theta}_M\}$ randomly from the posterior. Such samples can then be used to compute integrals over the posterior

$$\int f(\vec{\theta})p(\vec{\theta}|\mathbf{x})\mathrm{d}\vec{\theta} \approx \frac{1}{M}\sum_{i=1}^{M} f(\vec{\theta}_i).$$

Most quantities that one might wish to compute from a posterior distribution can be expressed as integrals of this form, and so generation of such samples is the most complete way to represent posterior distributions. Efficient production of samples is non-trivial and will be the topic of the next chapter of these notes.

5.5 Interpreting summary statistics*

5.5.1 Decision theory

The posterior mean, mode and median are all valid ways to summarise a posterior distribution. One way to motivate these (and other possible) choices is through **decision theory**. In decision theory, understanding which decision is the best is motivated by introducing a **loss function** which characterises the cost or penalty of making a particular decision. Formally we define various quantities

- The **sample space** $\mathcal{X}$ denotes the possible values for the observed data, $\mathbf{x}$.
- The **parameter space**, Ω_θ, denotes possible (unknown) states of nature (or parameter values characterising the true pdf of observed data sets).
- We define a **family of probability distributions**, $\{\mathbb{P}_\theta(x) : x \in \mathcal{X}, \theta \in \Omega_\theta\}$, which describe how the observed data is generated in the possible states of nature.
- The **action space**, $\mathcal{A}$, is the set of actions that an experimenter can take after observing data, e.g., reject or accept a null hypothesis, assign an estimate to the value of θ etc.
- The **loss function**, $L : \Omega_\theta \times \mathcal{A} \to \mathbb{R}$, is a mapping from the space of actions and parameters to the real numbers, such that $L(a, \theta)$ is the loss associated with taking the action a when the true state of nature is θ.
- The set of **decision rules**, $\mathcal{D}$, is a set of mappings from data to actions. Each element $d \in \mathcal{D}$ is a function $d : \mathcal{X} \to \mathcal{A}$ that associates a particular action with each possible observed data set.

For a parameter value $\theta \in \Omega_\theta$, the risk of a decision rule, d, is defined as

$$R(\theta, d) = \mathbb{E}_\theta L(\theta, d(X)) = \begin{cases} \sum_{x \in \mathcal{X}} L(\theta, d(x)) p(x; \theta) & \text{for discrete } X \\ \int_{\mathcal{X}} L(\theta, d(x)) p(x; \theta) \mathrm{d}x & \text{for continuous } \mathcal{X}. \end{cases}$$

In other words, the risk is the expected loss of a particular decision rule when the true value of the unknown parameter is θ. Note that this is fundamentally a frequentist concept, since the definition implicitly invokes the idea of repeated samples from the parameter space $\mathcal{X}$ and computes the average loss over these hypothetical repetitions. However, it is possible to extend these ideas to a Bayesian framework by defining a prior, $\pi(\theta)$, over the parameters of the distribution. The **Bayes risk** of a decision rule, d, is then defined as

$$r(\pi, d) = \int_{\theta \in \Omega_\theta} R(\theta, d) \pi(\theta) \mathrm{d}\theta,$$

or by a sum in the case of a discrete-valued probability distribution. A decision rule is **a Bayes rule** with respect to the prior $\pi(\cdot)$ if it minimizes the Bayes risk, i.e.,

$$r(\pi, d) = \inf_{d' \in \mathcal{D}} r(\pi, d') = m_\pi, \text{ say.}$$

Note that, as usual in a Bayesian context, the Bayes rule depends on the specification of the prior and therefore there will be infinitely many Bayes rules for any particular problem. A

useful choice of prior is the one that is most conservative in its estimate of risk. This gives rise to the concept of **a least favourable prior**. The prior $\pi(\theta)$ is least favourable if, for any other prior $\pi'(\theta)$ we have

$$r(\pi, d_\pi) \geq r(\pi', d_{\pi'})$$

where d_π, $d_{\pi'}$ are the Bayes rules corresponding to $\pi(\cdot)$ and $\pi'(\cdot)$ respectively.

5.5.2 Bayes rules as minimizers of posterior expected loss

The Bayes risk can be written as

$$\begin{aligned} r(\pi, d) &= \int_{\Omega_\theta} R(\theta, d)\pi(\theta)\mathrm{d}\theta \\ &= \int_{\Omega_\theta}\int_{\mathcal{X}} L(\theta, d(x))p(x|\theta)\pi(\theta)\mathrm{d}x\mathrm{d}\theta \\ &= \int_{\Omega_\theta}\int_{\mathcal{X}} L(\theta, d(x))p(\theta|x)p(x)\mathrm{d}x\mathrm{d}\theta \\ &= \int_{\mathcal{X}} p(x)\left\{\int_{\Omega_\theta} L(\theta, d(x))p(\theta|x)\mathrm{d}\theta\right\}\mathrm{d}x \end{aligned}$$

where the second line follows from the definition of the risk function and the third line follows by using Bayes' theorem to write $p(x|\theta)\pi(\theta) = p(\theta|x)p(x)$ in terms of the posterior $p(\theta|x)$ and the evidence $p(x)$. The Bayes rule minimizes the Bayes risk. We see that this minimum is achieved for a particular value of x by making the decision that minimizes the expression in curly brackets. This is the **expected posterior loss** associated with the observed x. This observation simplifies the calculation in many cases and also illustrates the general property of Bayesian procedures, namely that the decision depends only on the observed data and not on potential unobserved data sets.

We will illustrate this with four examples. In the first three examples, we are attempting to make a point estimate and so the decision is an assignment of the value of the parameter $d = \hat{\theta}$.

Example: Point estimation with squared error loss Suppose we want to make a point estimate of a parameter and we use a squared error loss function, $L(\theta, d) = (\theta - d)^2$. Find the Bayes rule.

Solution

The Bayes rule chooses $d(Y)$ to minimize

$$\int_{\Omega_\theta} (\theta - d)^2 p(\theta|y)\mathrm{d}\theta.$$

Differentiating with respect to d and setting this to zero gives

$$\int_{\Omega_\theta} (\theta - d)p(\theta|x)\mathrm{d}\theta = 0 \quad \Rightarrow \quad d = \int_{\Omega_\theta} \theta p(\theta|x)\mathrm{d}\theta.$$

In other words, the Bayes estimator of θ, with squared error loss, is the **posterior mean**.

Example: Point estimation with absolute magnitude error loss
Suppose we instead use the loss function $L(\theta, d) = |\theta - d|$. Find the new Bayes rule.

Solution
In this case, the Bayes rule minimizes

$$\int_{-\infty}^{d} (d - \theta) p(\theta|x) \mathrm{d}\theta + \int_{d}^{\infty} (\theta - d) p(\theta|x) \mathrm{d}\theta.$$

Setting the derivative with respect to d to zero now gives

$$\int_{-\infty}^{d} p(\theta|x) \mathrm{d}\theta - \int_{d}^{\infty} p(\theta|x) \mathrm{d}\theta = 0 \quad \Rightarrow \quad \int_{-\infty}^{d} p(\theta|x) \mathrm{d}\theta = \int_{d}^{\infty} p(\theta|x) \mathrm{d}\theta = \frac{1}{2}.$$

In other words, the Bayes estimator of θ, with absolute magnitude error loss, is the **posterior median**.

Example: Point estimation with delta-function gain
Suppose we instead use the loss function

$$L(\theta, d) = \begin{cases} -\delta(\theta - d) & \text{if } d = \theta \\ 0 & \text{if } d \neq \theta \end{cases}.$$

In other words, the loss is infinitely higher for any value except the correct one. Find the new Bayes rule.

Solution
In this case, the Bayes rule minimizes

$$-\int_{-\infty}^{\infty} \delta(\theta - d) p(\theta|x) \mathrm{d}\theta = -p(d|x).$$

The minimum loss is obtained by setting

$$d = \mathrm{argmaxp(d|x)},$$

i.e., the posterior mode.

Example: Interval estimation
Suppose we have a loss function of the form

$$L(\theta, d) = \begin{cases} 0 & \text{if } |\theta - d| \leq \delta \\ 1 & \text{if } |\theta - d| > \delta \end{cases}$$

for specified $\delta > 0$. What is the Bayes rule?

Solution

The expected posterior loss in this case is the posterior probability that $|\theta - d| > \delta$. The interval that minimises this loss, among intervals of fixed length 2δ, is the interval that contains the highest posterior probability. This is called the *highest posterior density* interval.

We see that all of the "natural" ways to obtain a point estimate from a Bayesian posterior can be interpreted in terms of Bayes rule's with different loss functions.

5.6 Bayesian hypothesis testing

The denominator that appears in Bayes' theorem is the Bayesian evidence and can be computed via

$$\mathcal{Z} = p(\mathbf{x}) = \int p(\mathbf{x} \mid \vec{\theta}) p(\vec{\theta}) \mathrm{d}\vec{\theta}. \tag{77}$$

When writing down Bayes' theorem we suppressed the fact that all of the quantities were conditioned on the particular model we were assuming for the data generating process. Explicitly reintroducing the dependence on the model, M, we have

$$p(\vec{\theta}|\mathbf{x}, M) = \frac{p(\mathbf{x}|\vec{\theta}, M)p(\vec{\theta}|M)}{p(\mathbf{x}|\mathbf{M})}.$$

This makes it clear that the evidence, $p(\mathbf{x}|\mathbf{M})$, represents the *probability of seeing the model data under model M* and can be thought of as the likelihood for the model given the observed data. If we now have more than one model, M_1 and M_2 say, that we believe could describe the data, we can compute the **posterior odds ratio** for M_1 over M_2

$$O_{12} = \frac{p(\mathbf{x}|\mathbf{M_1})}{p(\mathbf{x}|\mathbf{M_2})} \frac{p(M_1)}{p(M_2)}.$$

The first term is called the **Bayes factor** and is the ratio of the model likelihoods. The second term is the **prior odds ratio**, which represents our prior belief about the relative probability of the two models. The posterior odds is the ratio of model probabilities based on the observed data and is the basis for Bayesian hypothesis testing. For $O_{12} \gg 1$ we favour model M_1, while for $O_{12} \ll 1$ we favour M_2.

In the case of a flat prior on models the prior odds ratio is just 1 and decisions are based on the Bayes factor. Kass and Rafferty (1995) described a 'rule of thumb' for interpreting Bayes' factors. This is summarised in Table 1. This Table can be used to interpret the results of Bayesian hypothesis tests. Alternatively, the distribution of the Bayes factor can be computed under the null hypothesis and used, in a frequentist way, to produce a mapping between p-values and Bayesian posterior odds ratios.

The models M_1 and M_2 need not be very different, but could, for example, represent different regions of the parameter space of a distribution, e.g., $M_1 : \theta \in \Theta_1$ versus $M_2 : \theta \in \Theta_2$. If the two hypotheses are both simple then the Bayes factor reduces to the likelihood ratio, which we saw was the optimal test statistic in the frequentist hypothesis testing context.

Computation of the Bayesian evidence is challenging. Most sampling algorithms that return independent samples from the posterior ignore the evidence as it is just a normalisation constant. The evidence can be written as an integral over the posterior which can be

Bayes Factor	Interpretation
< 3	No evidence of M_1 over M_2
> 3	Positive evidence for M_1
> 20	Strong evidence for M_1
> 150	Very strong evidence for M_1

Table 1: Table for intepretation of Bayes' factors, as presented in Kass and Rafferty (1995).

approximated by a sum over samples

$$\frac{1}{\mathcal{Z}} = \int \frac{1}{p(\mathbf{x} \mid \vec{\theta})} \frac{p(\mathbf{x} \mid \vec{\theta})p(\vec{\theta})}{\mathcal{Z}} \mathrm{d}\vec{\theta} \approx \frac{1}{M} \sum_{i=1}^{M} \frac{1}{p(\mathbf{x} \mid \vec{\theta}_i)}.$$

In other words it is the harmonic mean of the likelihoods of the samples. This is an extremely unstable approximation, however, as this sum is dominated by points with small likelihoods, but these are precisely the regions where there will be fewer samples and hence larger Monte Carlo error. Other techniques, such as nested sampling, can be used to compute evidences more accurately and these will be discussed in the next chapter.

Example: Suppose we have a two dimensional Normal likelihood of the form

$$p(\mathbf{x}|\vec{\theta}) = \frac{\sqrt{1-\rho^2}}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{1}{-}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + 2\frac{\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right] \tag{78}$$

and use priors for the parameters μ_1 and μ_2 of the form

$$p(\mu_1) = \frac{1}{\Sigma_1\sqrt{2\pi}} \exp\left[-\frac{1}{2\Sigma_1^2}\mu_1^2\right], \qquad p(\mu_2) = \frac{1}{\Sigma_2\sqrt{2\pi}} \exp\left[-\frac{1}{2\Sigma_2^2}\mu_2^2\right]. \tag{79}$$

We are interested in comparing the two models

$$M_1 : \mu_2 = 0, \qquad M_2 : \mu_2 \in (-\infty, \infty).$$

The evidence for M_1 can be computed as

$$\mathcal{Z}_1 = \frac{1}{2\pi\sigma_2}\sqrt{\frac{1-\rho^2}{\sigma_1^2+\Sigma_1^2}} \exp\left[-\frac{x_2^2(\sigma_1^2-(1-\rho^2)\Sigma_1^2) + 2\rho x_1 x_2 \sigma_1\sigma_2 + \sigma_2^2 x_1^2}{2\sigma_2^2(\sigma_1^2+\Sigma_1^2)}\right]$$

and for M_2 it is

$$\begin{aligned}\mathcal{Z}_2 = \frac{1}{2\pi}\sqrt{\frac{1-\rho^2}{\sigma_1^2(\sigma_2^2+\Sigma_2^2)+\Sigma_1^2(\sigma_2^2+(1-\rho^2)\Sigma_2^2)}}\times \\ \times \exp\left[-\frac{x_2^2((1-\rho^2)\Sigma_1^2+\sigma_1^2) + 2\rho x_1 x_2\sigma_1\sigma_2 + x_1^2((1-\rho^2)\Sigma_2^2+\sigma_2^2)}{2\Sigma_1^2((1-\rho^2)\Sigma_2^2+\sigma_2^2)+2\sigma_1^2(\sigma_2^2+\Sigma_2^2)}\right]\end{aligned} \tag{80}$$

which gives the posterior odds ratio in favour of M_2, for equal prior odds (which is just the Bayes factor)

$$\begin{aligned}\mathcal{O}_{21} = \frac{\mathcal{Z}_2}{\mathcal{Z}_1} = \sigma_2\sqrt{\frac{\Sigma_1^2+\sigma_1^2}{\Sigma_1^2((1-\rho^2)\Sigma_2^2+\sigma_2^2)+\sigma_1^2(\Sigma_2^2+\sigma_2^2)}}\times \\ \times \exp\left[\frac{\Sigma_2^2(x_2((1-\rho^2)\Sigma_1^2+\sigma_1^2)+\rho x_1\sigma_1\sigma_2)^2}{2(\Sigma_1^2+\sigma_1^2)\sigma_2^2(\sigma_1^2(\Sigma_2^2+\sigma_2^2)+\Sigma_1^2((1-\rho^2)\Sigma_2^2+\sigma_2^2))}\right].\end{aligned} \tag{81}$$

This is difficult to interpret, but if we now assume that $\Sigma_1^2 \gg \sigma_1^2$, i.e., that the prior in μ_1 is much broader than the typical measurement uncertainty, the odds ratio simplifies to

$$\mathcal{O}_{21} \approx \sigma_2 \sqrt{\frac{1}{(1-\rho^2)\Sigma_2^2 + \sigma_2^2}} \exp\left[\frac{(1-\rho^2)x_2^2}{2\sigma_2^2}\right]$$

We see that there is a competition between the size of the additional variable dimension (characterised by Σ_2) in the first term and the weight of evidence for the additional effect in the data (characterised by the second term). Only if the addition of the extra dimension significantly improves the fit to the data (characterised by x_2 which is effectively the peak of the posterior in μ_2 when that parameter is allowed to vary) should the more complex model be favoured. If the fit does not improve, then the addition of the extra dimension is penalised by the first term and so the more complex model should not be preferred. It is often said that Bayesian posterior odds ratios automatically encode the notion of "Occam's razor", i.e., one should use the simplest model that adequately describes the data since adding extra degrees of freedom always improves a fit. This is the sense in which it is meant. Addition of extra dimensions typically includes a prior penalty, as we see here, which will lead to the disfavouring of an alternative model unless the likelihood shows a significantly great improvement when the extra degrees of freedom are included.

5.7 Predictive checking

In both a frequentist and a Bayesian context it is natural to ask whether the model is a good representation of the observed data. In the Bayesian context this is accomplished by using **predictive distributions**.

Definition: the **prior predictive distribution** is the probability distribution

$$p(\mathbf{x}) = \int_{\vec{\theta}\in\Theta} p(\mathbf{x}|\vec{\theta})p(\vec{\theta})\mathrm{d}\vec{\theta}.$$

This is the likelihood weighted by the assigned prior distribution and therefore represents our *a priori* belief about the distribution of data sets that would be observed. Similarly, we have the following

Definition: the **posterior predictive distribution** is the probability distribution

$$p(\mathbf{y}|\mathbf{x}) = \int_{\vec{\theta}\in\Theta} p(\mathbf{y}|\vec{\theta})p(\vec{\theta}|\mathbf{x})\mathrm{d}\vec{\theta}.$$

This is the likelihood weighted by the posterior probability based on the observed data $\mathbf{x}$ and is our expectation about the distribution of future data sets $\mathbf{y}$.

The posterior predictive distribution can be used to assess whether the observed data is unusual within the posterior distribution, which is an indicator about whether or not the model is a good fit. Based on the observed data $\mathbf{x}$ we generate a large number of new data sets $\{\mathbf{y}_1, \ldots \mathbf{y}_N\}$ that are similar to $\mathbf{x}$, i.e., they consist of the same number of observations. For each data set we compute a set of summary statistics, and hence obtain the distribution of the summary statistics over many realisations of the posterior predictive distribution. We can then assess the "p-value" of the observed data within these distributions. If it looks like an outlier in any one of these distributions this suggests the model is not a good fit. Suitable

summary statistics could include the maximum, minimum, median, skewness, kurtosis etc. Ideally we choose summary statistics that are orthogonal to the model parameters to increase sensitivity, since we are using the data twice (once to compute the posterior and once to compare to the predictive distribution). Statistics that are effectively tuned to the observed data will tend to lie in the middle of the predictive distributions by construction, even if the model is poor. We will see an example of this in the next section.

5.8 Example: regression

To illustrate some of the ideas discussed above we will present a Bayesian analysis of a regression problem. We suppose that we have made measurements of a set of values, $\{y_i\}$, corresponding to sets of p known explanatory variables, $\{\mathbf{x}_i\}$, and we believe that these follow a linear relationship with equal variance normally distributed errors

$$y_i \sim N(\mathbf{x}_i^T \vec{\beta}, \sigma^2), \quad i = 1, \ldots, N.$$

We want to infer the parameters of the linear relationship, $\vec{\beta}$, and the unknown precision $\tau = 1/\sigma^2$. We use a Bayesian framework and so must write down prior distributions on these parameters. We can assume a separable prior

$$p(\vec{\beta}, \tau) = p(\tau) \prod_{i=1}^{p} p(\beta_j)$$

and take Normal priors for the β_j's and a Gamma prior for τ as these are conjugate priors in the Normal-Gamma model

$$\beta_j \sim N(\mu_{\beta_j}, \sigma^2_{\beta_j}), \quad \tau \sim \text{Gamma}(a, b).$$

In the absence of prior information it is reasonable to set $\mu_{\beta_j} = 0$. Inferred values of the coefficients that are non-zero then provide evidence for the existence of a relationship between the observed data and those explanatory variables. Setting σ_j^2 to a large value, say 10^4, indicates large uncertainty in the parameter values and avoids strong prior dependence in the results. For the prior on τ, it is usual to take small values of a and b, for example $a = b = 0.1$ or $a = b = 0.01$. However, such priors lead to a preferred value (i.e., a peak) in the prior and so the use of such priors is somewhat controversial.

To illustrate fitting such a model, we can use a standard data set, the MTCARS data set, which is available in the R statistical software package and may also be found online. The data set contains observations, y_i, of the miles driven per gallon in the i'th of 32 different models of car, with explanatory variables x_{i1}, the rear axle ratio, x_{i2}, the weight of the i'th car and x_{i3}, the time to drive 0.25 miles from rest. We fit the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, 1/\tau), \quad i = 1, \ldots 32,$$

with $\beta_j \sim N(0, 1000)$ and $\tau \sim \text{Gamma}(0.1, 0.1)$. We can use statistical software (in this case R) to generate samples from the posterior. Techniques for doing this will be discussed in the next chapter, and in the associated practical. using these samples we can obtain a posterior mean and 95% symmetric credible interval for each parameter. These can be compared to the frequentist estimates of the same parameters and the frequentist 95% confidence interval (see problem sheet 1). This comparison is in Table 2.

	Bayesian results		Frequentist results	
Parameter	Posterior mean	95% credible interval	MLE	95% confidence interval
β_0	10.369	[-5.098,36.349]	11.395	[-5.134,27.922]
β_1	1.777	[-0.721,4.166]	1.750	[-0.857,4.169]
β_2	-4.335	[-5.702,-2.995]	-4.347	[-5.787,-3.009]
β_3	0.968	[0.449,1.493]	0.946	[0.410,1.482]
σ^2	6.978	[4.160,11.729]	6.554	—

Table 2: Comparison between Bayesian and frequentist estimates of the linear model fit to the MTCARS data set.

The results of the Bayesian fit are quite consistent between the two approaches, although there are some differences and the interpretation of the results is different. We now want to assess the quality of the results. In a frequentist setting, assessment of the quality of a linear model fit is done through the production of *studentised residuals* and *Q-Q plots*. A studentised residual is

$$\hat{\epsilon}_i = \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where $\hat{\beta}$ are the estimated parameters, $\hat{\sigma}$ is the estimated standard deviaiton and h_{ii} is the i'th diagonal element of the matrix $H = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$. These quantities follow a student-t distribution which is why they are called studentised residuals. A $Q - Q$ plot is a plot of the distribution of these values against the theoretical distribution, which should be approximately a straight line if the model is a good description of the data.

We can construct analogous quantities in the Bayesian case, but now the parameters are described by distributions rather than point estimates. A point estimate can be constructed in a number of different ways — using posterior mean values, using a single draw from the posterior, or averaging over the full posterior. The latter approach involves computing the studentised residual for a large number of draws from the posterior and averaging them, and is called the *posterior mean of the residual*. Studentised residuals are plotted in various ways in Figure 9.

We can also produce posterior predictive checks as described in section 5.7. We compute realisations of similar data sets and estimate the distribution of various summary statistics which we then compare to the values in the observed data sets. In this case we compute the distributions of the minimum, maximum, median and skewness in repeated data sets. These are shown in Figure 10, along with the values in the observed data set. We see that the observed values lie within the distributions in all cases, except for skewness. Seeing that the observed data lies in the tail of the distribution may indicate a failure of the model. In this case we might want to try varying the assumption of normally distributed errors and homoskedacity (equal error variance).

The issue with the posterior predictive checks could indicate a failure of the model, or the influence of an outlying data point. One way to tackle this is to modify the model so that the distribution of the errors ϵ_i is no longer assumed to be normal. The most common approach is to replace the normal distribution by a t_ν-distribution, as these have heavier tails. This is referred to as **robust regression**. The degrees of freedom, ν, in the t_ν-distribution can be fixed to some reasonable value, or allowed to vary in a hierarchical model (see next section). In that case the prior on ν is usually taken to be a Gamma distribution, $\nu \sim \text{Gamma}(c, d)$.

For the MTCARS dataset we try this, using prior values $c = d = 0.1$, and then look at the

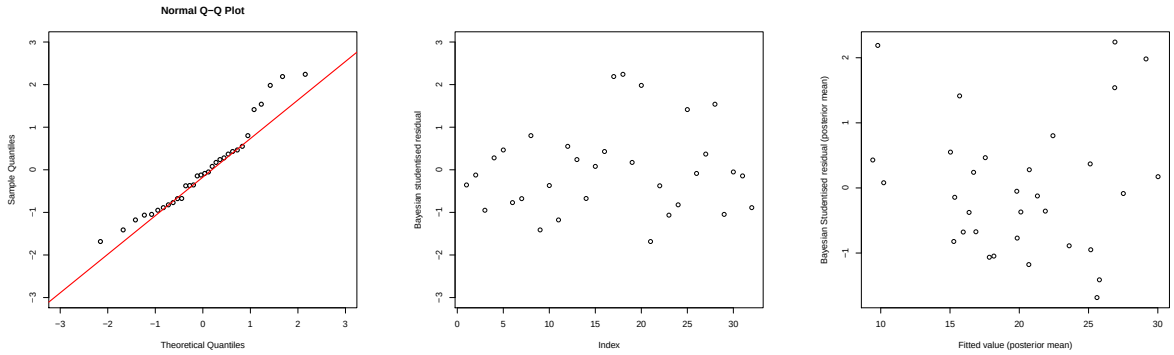


Figure 9: $Q-Q$ plot of the studentised residuals (left), studentised residual versus index of data point (middle) and studentised residual versus posterior mean of the predicted value, $\hat{y}_i$, for the Bayesian fit to the MTCARS data set. We look for the left hand plot to be on the diagonal line, for the middle and right hand plots we want the values to be randomly distributed (i.e., no trend with the x value) and in the range from minus a few to plus a few. These constraints are all satisfied here and so we see no cause for concern.

posterior predictive distribution again. The results for the skewness are shown in Figure 11. We that robustifying regression can help to improve the model fit in this case. The observed dat moves from lying at the 99.6% point of the distribution to lying at the 96.3%. So, it is still something of an outlier but it is not so much a cause for concern. It is perhaps not surprising that the use of robust regression only helped a small amount in this case, since we are trying to compensate for non-zero skew in the data and the t-distribution is also a symmetric distribution.

5.9 Hierarchical models

In many contexts, for example the observation of mergers of compact binary coalescences through gravitational wave observations, the likelihood describes the observation of a single event, and the prior describes the distribution of parameter values in the population from which the events are drawn. Often the parameters of the population prior are not themselves known but are of interest. For example, we do not know the distribution of masses of black holes in binaries and would like to learn about this from observations of the gravitational wave sources. This leads to the notion of a **hierarchical model**, in which the likelihood for data depends on parameters for which we write down a prior that in turn depends on unknown parameters (usually termed **hyperparameters**), for which we write down another prior (the **hyperprior**).

This hierarchy can be continued to more and more levels, but such models increase rapidly in complexity. Inference on complex hierarchical models can be simplified by imposing a *conditional independence* structure in the models, e.g., $p(x,y,z) = p(x|z)p(y|z)p(z)$. Conditional dependence structures can be compactly represented using *graphical models*. These are directed acyclic graphs that indicate dependencies between various components of the model. It is important that the graph has no cycles as only then can the joint probability be factorised. An example of a graphical model is shown in Figure 12. This model represents the following conditional dependence structure

$$p(p,q,r,s,t,u,v,w,x,y,z) = p(x|y,z)p(y|u,w)p(w|v)p(u)p(v)p(z|r)p(r|p,q)p(p)p(q) \quad (82)$$

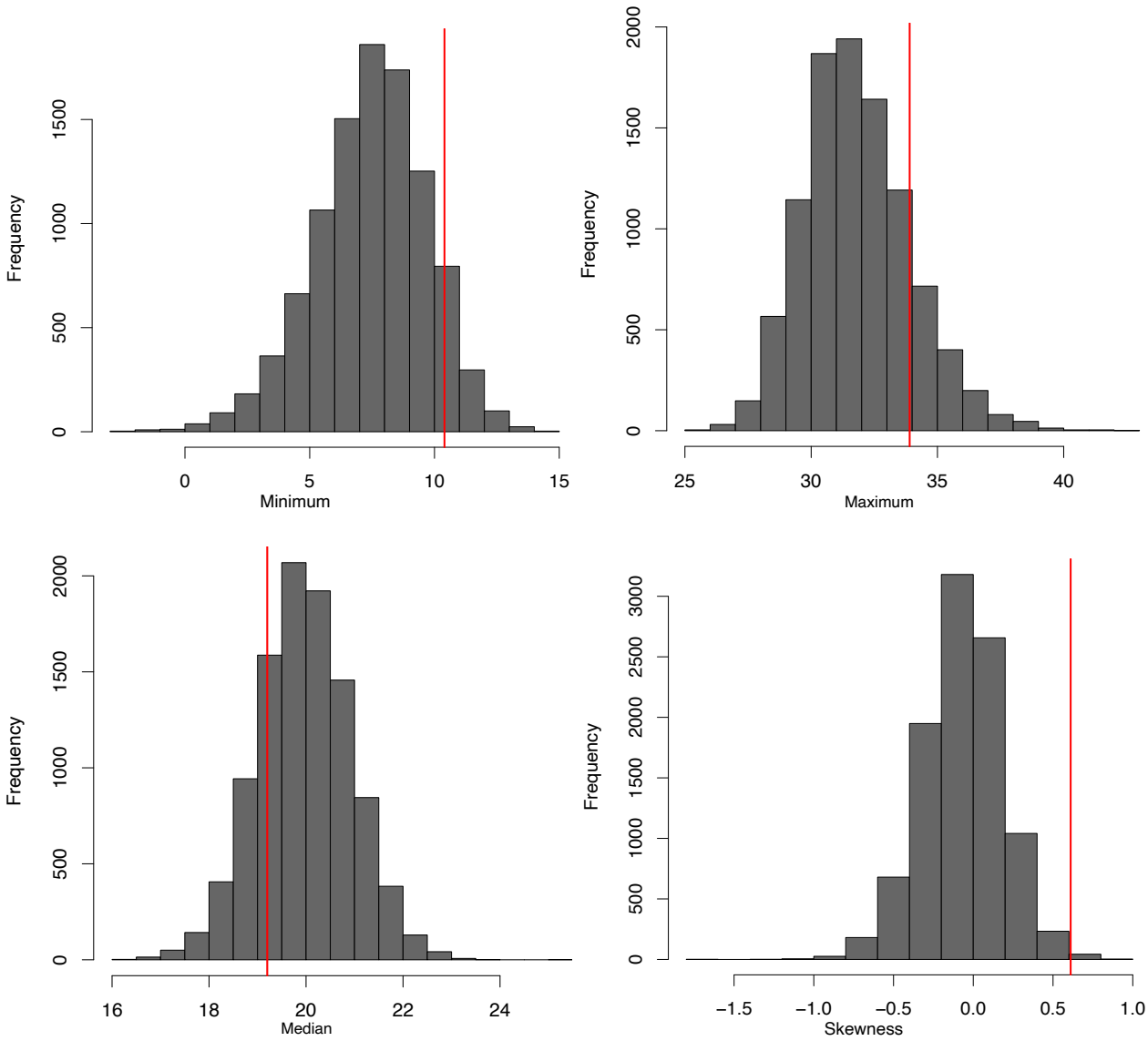


Figure 10: Predictive distributions for the maximum (top left), minimum (top right), median (bottom right) and skewness (bottom right) in replicated data sets of size 32, based on the posterior distribution from the MTCARS data set. The vertical red lines indicate the values in the data set form which the posterior was obtained. We see that this lies in the middle of the distribution in all cases, except skewness, in which it lies in the tail, which might indicate a failure to properly fit the data.

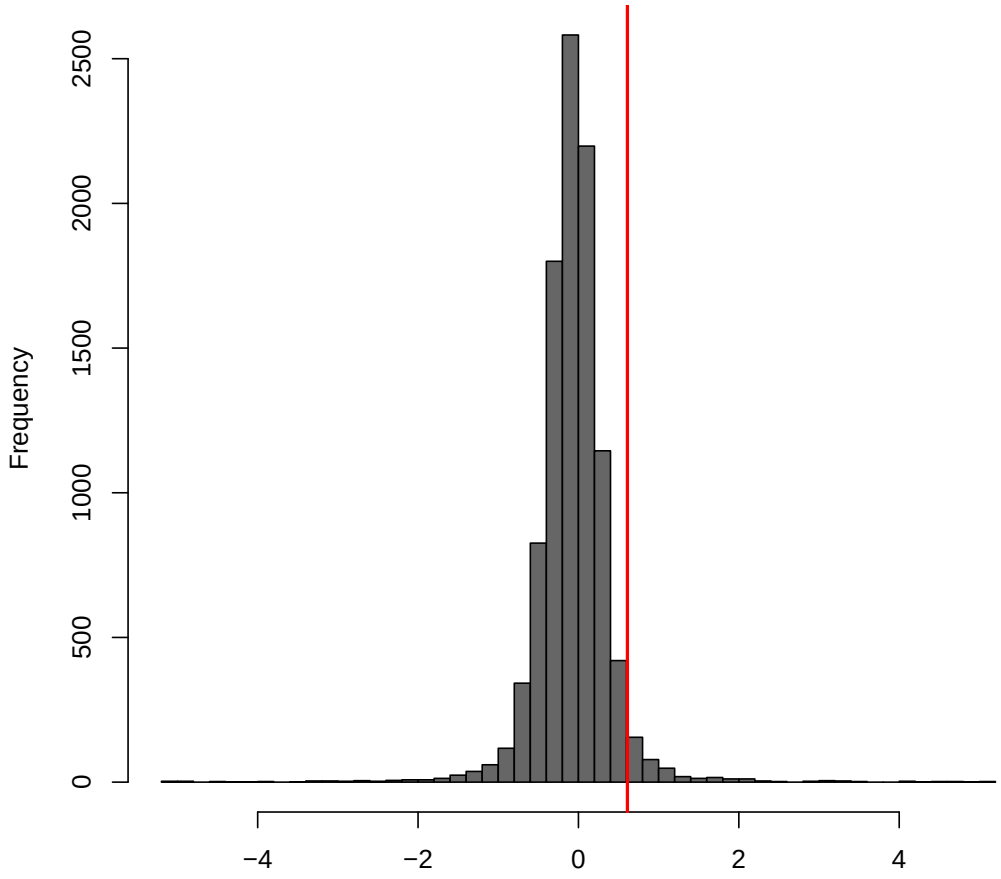


Figure 11: Posterior predictive distribution of skewness for the robustified regression model. The observed value of the skewness is indicated by a vertical red line as before.

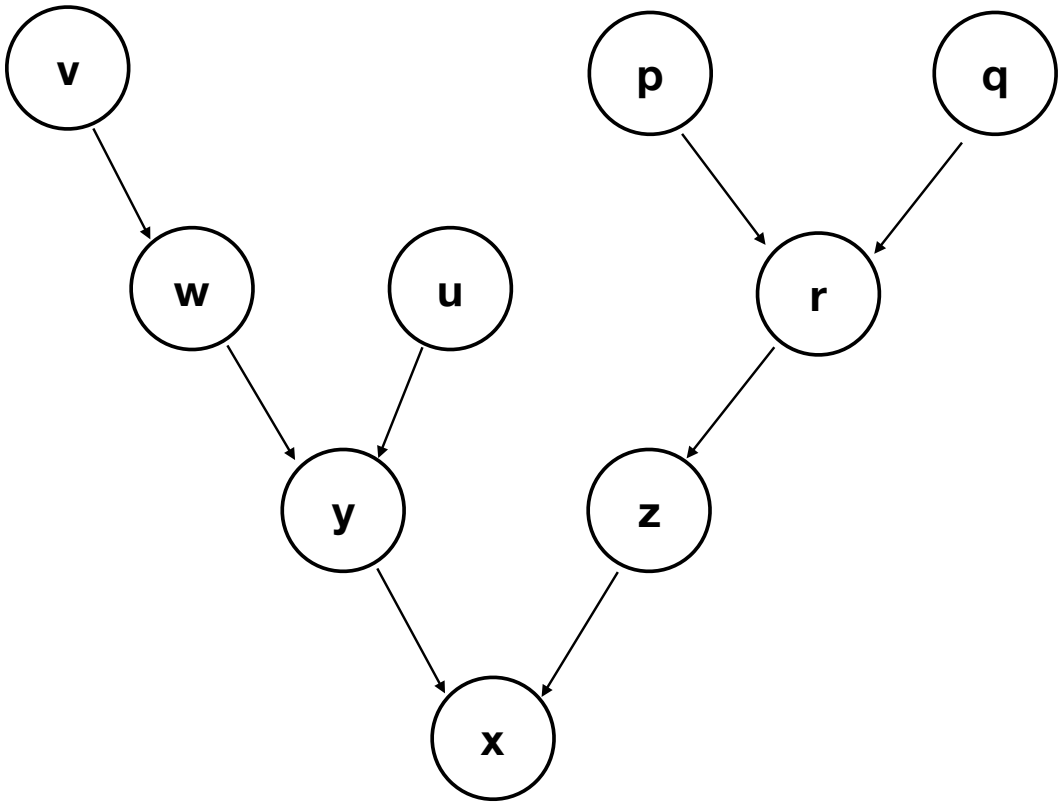


Figure 12: Illustration of a Bayesian graphical model. This is an acyclic directed graph that indicates conditional dependencies in complex Bayesian hierarchical models.

5.9.1 Selection effects

One thing that is important to account for in hierarchical modelling are selection effects. The decision about whether or not to include an event in a catalogue used for inference is based on whether or not the event is "detected", i.e., whether or not the observed data passes some pre-determined threshold criterion for inclusion. This is usually a property of the data only. Selection effects can be included by modifying the likelihood so that it represents the likelihood of "detected" data sets. If the un-corrected likelihood is $p(\mathbf{x}|\vec{\theta})$ then the likelihood for observed events is just

$$p(\mathbf{x}|\vec{\theta}, \text{obs}) = \frac{1}{p_s(\vec{\theta})} p(\mathbf{x}|\vec{\theta}), \quad \text{where } p_s(\vec{\theta}) = \int_{\mathbf{x}>\text{threshold}} p(\mathbf{x}|\vec{\theta}) \mathrm{d}\mathbf{x}.$$

The integral is over all data sets that would have been considered as "detections", i.e., passing the threshold for inclusion in inference. What we have done here is renormalise the likelihood so that it integrates to 1 over all above threshold data sets. Since the partition of the data into observed and unobserved is a property of $\mathbf{x}$ only, the relative probabilities of different above threshold data sets must be in proportion to their probabilities in the set of all data sets.

Usually, the likelihood will depend on parameters of the particular source, $\vec{\theta}$, that are

themselves determined by the priors, which depends on the hyperparameters of the population, $\vec{\lambda}$. Then the likelihood for observed events, marginalised over the source parameters is simply

$$p(\mathbf{x}|\vec{\lambda},\mathrm{obs}) = \frac{1}{p_s(\vec{\lambda})}\int p(\mathbf{x}|\vec{\theta})p(\vec{\theta}|\vec{\lambda})\mathrm{d}\vec{\theta}, \quad \text{where } p_s(\vec{\lambda}) = \int_{\mathbf{x}>\mathrm{threshold}}\int p(\mathbf{x}|\vec{\theta})p(\vec{\theta}|\vec{\lambda})\mathrm{d}\vec{\theta}\mathrm{d}\mathbf{x}. \tag{83}$$

Usually we are interested in the parameters of individual sources as well as the overall population parameters. The joint likelihood of $\mathbf{x}$ and $\vec{\theta}$, conditioned on detection, is

$$p(\mathbf{x},\vec{\theta}|\vec{\lambda},\mathrm{obs}) = p(\mathbf{x}|\vec{\theta},\mathrm{obs})p(\vec{\theta}|\vec{\lambda},\mathrm{obs}).$$

The first term is Eq. (5.9.1), but for the source parameters $\vec{\theta}$

$$p(\mathbf{x}|\vec{\theta},\mathrm{obs}) = \frac{p(\mathbf{x}|\vec{\theta})}{p(\mathrm{obs}|\vec{\theta})}, \quad \text{where } p(\mathrm{obs}|\vec{\theta}) = \int_{\mathbf{x}>\mathrm{threshold}} p(\mathbf{x}|\vec{\theta})\mathrm{d}\mathbf{x}.$$

The second term is the prior on $\vec{\theta}$ *for events above threshold*. However, this prior is modified from $p(\vec{\theta}|\vec{\lambda})$ by the conditioning on detection, namely

$$p(\vec{\theta}|\vec{\lambda},\mathrm{obs}) = \frac{p(\vec{\theta},\mathrm{obs}|\vec{\lambda})}{p(\mathrm{obs}|\vec{\lambda})} = \frac{p(\mathrm{obs}|\vec{\theta},\vec{\lambda})p(\vec{\theta}|\vec{\lambda})}{p(\mathrm{obs}|\vec{\lambda})} = \frac{p(\mathrm{obs}|\vec{\theta})p(\vec{\theta}|\vec{\lambda})}{p_s(\vec{\lambda})}.$$

Putting this together we see that the terms relating to selection on $\vec{\theta}$, $p(\mathrm{obs}|\vec{\theta})$, cancel and the joint likelihood is

$$p(\mathbf{x},\vec{\theta}|\vec{\lambda},\mathrm{obs}) = \frac{p(\mathbf{x}|\vec{\theta})p(\vec{\theta}|\vec{\lambda})}{p_s(\vec{\lambda})}$$

giving a posterior on $\vec{\theta}$

$$p(\vec{\theta}|\mathbf{x},\vec{\lambda},\mathrm{obs}) \propto p(\mathbf{x}|\vec{\theta})p(\vec{\theta}|\vec{\lambda})$$

which is unchanged from the posterior that would be written down if there is no selection. We see that the selection effects corrections do not change inference about the parameters of individual sources, only inference about the hyperparameters governing the population as a whole.

This approach implicitly assumes that the number of observed events contains no information about the unknown parameters. An alternative approach is to write down a joint likelihood for all events, both the N_{obs} events that are observed, $\{\mathbf{x}_i\}$, with parameters $\{\vec{\theta}_i\}$, and the N_{nobs} events that are unobserved, $\{\mathbf{x}_j\}$, with parameters $\{\vec{\theta}_j\}$. We model the number of events as a Poisson process with overall rate $N(\vec{\lambda})$, and rate density $\mathrm{d}N/\mathrm{d}\vec{\theta}$. The joint likelihood is

$$p\left(\left\{\vec{\theta}_i\right\},\left\{\vec{\theta}_j\right\},\{\mathbf{x}_i\},\{\mathbf{x}_j\} \mid \vec{\lambda}\right) \propto \left[\prod_{i=1}^{N\mathrm{obs}} p\left(\mathbf{x}_i \mid \vec{\theta}_i\right)\frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_i}\left(\vec{\lambda}\right)\right] \times \\ \times \left[\prod_{j=1}^{N_{\mathrm{nobs}}} p\left(\mathbf{x}_j \mid \vec{\theta}_j\right)\frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_j}\left(\vec{\lambda}\right)\right]\exp\left[-N\left(\vec{\lambda}\right)\right] \tag{84}$$

We can marginalise over the unobserved data to obtain

$$p\left(\left\{\vec{\theta}_i\right\}, \{\mathbf{x}_i\} \mid \vec{\lambda}\right) \propto \left[\prod_{i=1}^{N_{\rm obs}} p\left(\mathbf{x}_i \mid \vec{\theta}_i\right) \frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_i}\left(\vec{\lambda}\right)\right] \frac{N_{\rm ndet}^{N_{\rm nobs}}\left(\vec{\lambda}\right)}{N_{\rm nobs}!} \exp\left[-N\left(\vec{\lambda}\right)\right] \tag{85}$$

where

$$N_{\rm ndet}\left(\vec{\lambda}\right) \equiv \int_{\{\mathbf{x}<\text{threshold}\}} \mathrm{d}\mathbf{x}\, \mathrm{d}\vec{\theta}\, p\left(\mathbf{x} \mid \vec{\theta}\right) \frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}}\left(\vec{\lambda}\right). \tag{86}$$

We can then marginalise over the unknown number of unobserved events to obtain

$$p\left(\left\{\vec{\theta}_i\right\}, \{\mathbf{x}_i\} \mid \vec{\lambda}\right) \propto \left[\prod_{i=1}^{N_{\rm obs}} p\left(\mathbf{x}_i \mid \vec{\theta}_i\right) \frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_i}\left(\vec{\lambda}\right)\right] \exp\left[-N_{\rm det}\left(\vec{\lambda}\right)\right]. \tag{87}$$

We can now introduce the overall rate in the Universe, N, by writing $\mathrm{d}N/\mathrm{d}\vec{\theta} = Np(\vec{\theta}|\vec{\lambda})$. Then

$$N_{\rm det}(\vec{\lambda}) = N \int_{\mathbf{x}>\text{threshold}} \int p(\mathbf{x}|\vec{\theta}) p(\vec{\theta}|\vec{\lambda}) \mathrm{d}\vec{\theta}\mathrm{d}\mathbf{x} = N p_s(\vec{\lambda}). \tag{88}$$

Setting a scale-invariant prior on N (which states that the number of detected events does not convey information about the unknown parameters of the population), $p(N) \propto 1/N$ we can marginalise N out of the likelihood and recover Eq. (83).

5.9.2 Examples of hierarchical models

We finish this section with two examples of Bayesian hierarchical models.

Example 1: Salmon fishery In a given year, several fish hatcheries located along rivers in Washington state, USA raise coho salmon from eggs to a juvenile stage. Each hatchery releases a batch of juvenile fish into the rivers. The fish then travel to the ocean and some of them return to the hatchery 3 years later. The probability that a juvenile salmon returns varies between hatcheries due to different hatchery practices and river conditions at the point of release. We construct a hierarchical model for this as follows

- Suppose there are J fisheries and n_j salmon observed at fishery j.
- The data for an individual observation, x_{ji}, of the i'th salmon at fishery j is Bernoulli (salmon returned or did not return), with parameter p_j, where j labels the fishery. The data for the total number of returning salmon at site j, x_j, is Binomial with parameters (n_j, p_j).
- We assume that the p_j's are drawn from some common global distribution and use the conjugate prior of Beta(a, b).
- The parameters a and b are not known and fixed as in the usual case, but these are unknown quantities of interest as they characterise the variability in the population. These are the hyperparameters of the prior on p_j.
- We define a suitable hyperprior $p(a, b)$ on the hyperparameters, for example a Gamma prior.

- The joint posterior on the set $(\{p_j\}, a, b)$ is

$$p(\{p_j\}, a, b|\mathbf{x}) \propto p(\mathbf{x}|\{p_j\}) \left[\prod_{j=1}^{J} p(p_j|a, b)\right] p(a, b).$$

  Note that the hyperprior on the hyperparameters appears only once as these parameters are common to all of the individual observations of fisheries.

- The marginal distribution on the hyperparameters (a, b) can be found by marginalising over the $\{p_j\}$'s

$$p(a, b|\mathbf{x}) \propto p(a, b) \prod_{j=1}^{J} \frac{B\left(a + x_j, b + n_j - x_j\right)}{B(a, b)}.$$

- Marginals on individual p_j's can be found in a similar way.

Example 2: Gravitational wave cosmology In August 2017 the LIGO/Virgo gravitational wave detectors observed gravitational waves from the inspiral and merger of a binary neutron star for the first time, GW170817. There was both a short gamma ray burst and a kilonova associated with this event, which allowed the unique identification of the host galaxy, NGC 4993, and hence the recessional velocity (redshift) of the host. The gravitational waves provide a measurement of the luminosity distance of the source. The rate of expansion of the Universe as a function of distance is a key observable for constraining cosmological parameters. The relationship is linear at low distances and the constant of proportionality is called the *Hubble constant*,

$$v = cz = H_0 d,$$

where v is the recessional velocity due to the expansion of the Universe, z is the corresponding redshift, H_0 is the Hubble constant and d is the luminosity distance. At low distance/redshift, the *peculiar velocity* of individual galaxies, relative to the overall expansion of the Universe (the "Hubble flow") is significant and so the observed recessional velocity, v_r, must be corrected by writing $v_r = H_0 d + v_p$. Observations of galaxies provide an estimate of the smoothed peculiar velocity field, $\langle v_p \rangle$. We are interested in inferring the value of the Hubble constant and build a hierarchical model as follows.

- The observed gravitational wave data, $x_{\rm GW}$, depends on the waveform of the source, which in turn depends on the source parameters. Most of these are not of interest, denoted $\vec{\lambda}$, and so we can marginalise them out, but we treat distance d and inclination, ι, separately

$$p(x_{\rm GW} \mid d, \cos\iota) = \int p(x_{\rm GW} \mid d, \cos\iota, \vec{\lambda})\, p(\vec{\lambda}) \mathrm{d}\vec{\lambda}. \tag{89}$$

- The measured recessional velocity, v_r, depends on the true recessional velocity, which depends on the peculiar velocity, v_p, and the Hubble redshift, $H_0 d$. Representing the electromagnetic measurement uncertainty as a Normal distribution we have

$$p\left(v_r \mid d, v_p, H_0\right) = N\left[v_p + H_0 d, \sigma_{v_r}^2\right](v_r) \tag{90}$$

- The measured smoothed peculiar velocity field at the location of the host galaxy depends on the true peculiar velocity there (and perhaps also on other quantities, but we suppress other dependencies here)

$$p\left(\langle v_p\rangle \mid v_p\right) = N\left[v_p, \sigma^2_{v_p}\right]\left(\langle v_p\rangle\right). \tag{91}$$

- The combined likelihood for the observations of $x_{\rm GW}$, $\langle v_p\rangle$ and v_r is

$$p(x_{\rm GW}, v_r, \langle v_p\rangle \mid d, \cos\iota, v_p, H_0) = \frac{1}{\mathcal{N}_s(H_0)} p(x_{\rm GW} \mid d, \cos\iota)\, p(v_r \mid d, v_p, H_0)\, p(\langle v_p\rangle \mid v_p). \tag{92}$$

Here the factor $\mathcal{N}_s(H_0)$ is the selection effects factor discussed earlier, which corrects for the fact that we only analyse events that exceed some threshold in the gravitational wave detector

$$\mathcal{N}_{\rm s}(H_0) = \int_{\rm detectable} \mathrm{d}\vec{\lambda}\,\mathrm{d}d\,\mathrm{d}v_p\,\mathrm{d}\cos\iota\,\mathrm{d}x_{\rm GW}\,\mathrm{d}v_r\,\mathrm{d}\langle v_p\rangle \times \Big[p(x_{\rm GW} \mid d, \cos\iota, \vec{\lambda})\, p(v_r \mid d, v_p, H_0) \times p(\langle v_p\rangle \mid v_p)\, p(\vec{\lambda})\, p(d)\, p(v_p)\, p(\cos\iota)\Big], \tag{93}$$

At the time of GW170817 the horizon for detection of binary neutron stars by the LIGO/Virgo detectors was much smaller ($\sim$ 100Mpc) than the distance to which the kilonova radiation could have been confidently observed ($\sim$ 400Mpc). This means that gravitational wave selection effects were dominant. As these depend directly on the luminosity distance, the dependence on H_0 is a higher order correction and so the selection function was approximately independent of H_0. A correct treatment of election effects will become increasingly important as the LIGO horizon increases in the future.

- We define priors on H_0, d, v_p and $\cos\iota$. These are independent and so we write down a product prior

$$p(d, \cos\iota, v_p, H_0) = p(d)p(\cos\iota)p(v_p)p(H_0).$$

We use flat priors on $\cos\iota$ and v_p, a volumetric prior on d, $p(d) \propto \mathrm{d}V_c/\mathrm{d}d$, where V_c is the comoving volume. We leave $p(H_0)$ unspecified, but note that the analysis in Abbott et al. (2017) used a scale-invariant prior $p(H_0) \propto 1/H_0$.

- We have now fully specified the hierarchical model. A graphical representation of this model is given in Figure 13. The posterior can now be found as

$$p(H_0, d, \cos\iota, v_p \mid x_{\rm GW}, v_r, \langle v_p\rangle) \propto \frac{p(H_0)}{\mathcal{N}_{\rm s}(H_0)}\, p(x_{\rm GW} \mid d, \cos\iota)\, p(v_r \mid d, v_p, H_0) \times\, p(\langle v_p\rangle \mid v_p)\, p(d)\, p(v_p)\, p(\cos\iota), \tag{94}$$

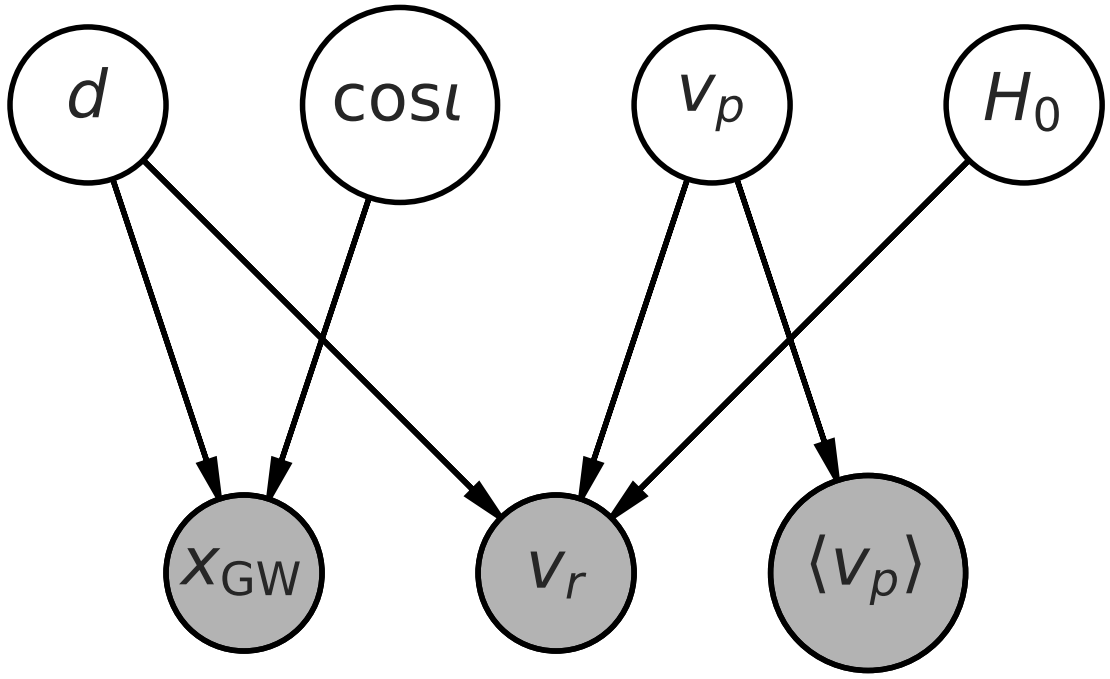


Figure 13: Graphical model for the Hubble constant measurement with gravitational wave observations of binary neutron stars. Figure reproduced from Abbott et al., *Nature Lett.* **551** 85 (2017).

- This posterior can be marginalised over d, $\cos\iota$ and v_p to give

$$p(H_0 \mid x_{\rm GW}, v_r, \langle v_p \rangle) \propto \frac{p(H_0)}{\mathcal{N}_s(H_0)} \int {\rm d}d\, {\rm d}v_p\, {\rm d}\cos\iota \times p(x_{\rm GW} \mid d, \cos\iota)\, p(v_r \mid d, v_p, H_0) \times p(\langle v_p \rangle \mid v_p)\, p(d)\, p(v_p)\, p(\cos\iota)\,. \quad (95)$$

  This marginalised posterior is shown in Figure 14.

- If we make subsequent observations of binary neutron star mergers with counterparts, indexed by a superscript $i = 1, \ldots, N$, we can combine these

$$p(H_0 \mid \{x^i_{\rm GW}, v^i_r, \langle v_p \rangle^i\}) \propto \frac{p(H_0)}{\mathcal{N}^N_s(H_0)} \prod_{i=1}^{N} \Bigg[\int {\rm d}d\, {\rm d}v_p\, {\rm d}\cos\iota \times p(x^i_{\rm GW} \mid d, \cos\iota)\, p(v^i_r \mid d, v_p, H_0) \times p(\langle v_p \rangle^i \mid v_p)\, p(d)\, p(v_p)\, p(\cos\iota) \Bigg]\,. \quad (96)$$

  Note that, as in the previous example, the prior on the common hyperparameters, $p(H_0)$, occurs only once. The selection effect correction appears once for every observation.

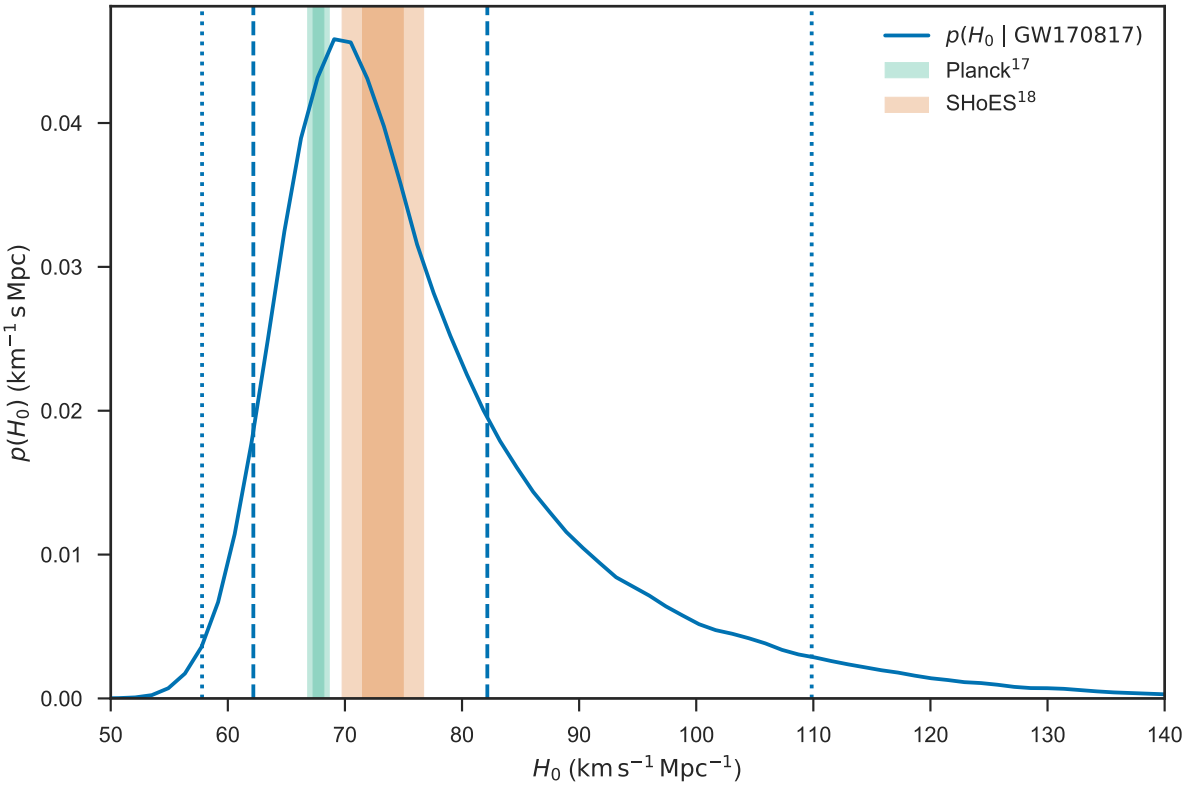


Figure 14: Posterior on the Hubble constant derived from GW170817. Figure reproduced from Abbott et al., *Nature Lett.* **551** 85 (2017).

6 Bayesian Sampling

As emphasised before, the output of Bayesian inference is a posterior probability distribution that encodes our state of knowledge about the parameters of the model based on the observed data and prior information. In certain contexts, for example when using conjugate models, the posterior can be written down in a closed analytic form and used directly for subsequent computation of derived quantities of interest. However, most often the posterior is not known in closed form. There are three approaches to inference in such situations. One is to use a Normal approximation to the posterior, the second is to use brute force integration methods and the third is to draw a set of representative samples from the posterior for us in Monte Carlo integration over the posterior.

6.1 Posterior computation: Bayesian Central Limit Theorem

The **Bayesian Central Limit Theorem** can be used to approximate posteriors, in the limit that the number of observations, $n \to \infty$. Suppose that we have samples $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p(x \mid \boldsymbol{\theta})$ and that the prior, $p(\boldsymbol{\theta})$, and likelihood, $p(x|\boldsymbol{\theta})$, are both twice differentiable near $\widehat{\boldsymbol{\theta}}_{\text{post}}$, the location of the peak of the posterior distribution. Then, for $n \to \infty$, we can approximate

$$p(\boldsymbol{\theta} \mid \mathbf{x}) \sim \mathrm{N}\left(\widehat{\boldsymbol{\theta}}_{\text{post}}, [I^{\text{post}}(\boldsymbol{\theta}, \mathbf{x})]^{-1}\right)$$

where

$$I^{\text{post}}(\boldsymbol{\theta}, \mathbf{x}) = -\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log p(\boldsymbol{\theta} \mid \mathbf{x})\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{post}}}.$$

The Bayesian central limit theorem follows from the usual central limit theorem. It used to be widely used due to the computational cost of generating posterior samples. However, it relies on the number of observations being large, which is often difficult to ensure in practice. Therefore, its use is no longer so widespread since computers are now sufficiently powerful to enable the generation of large numbers of posterior samples relatively cheaply.

6.2 Posterior computation: numerical integration

In low numbers of dimensions, posterior integrals can be computed using standard numerical integration techniques. There is a large literature on approximating integrals in various ways. The simplest is a grid approach, where the posterior is evaluated at a set of regularly spaced points in the space of waveform parameters. This can be thought of as a type of sampling approximation, where the samples are on a uniform grid. Direct integration rapidly becomes prohibitively expensive as the dimensionality of the model parameter space increases. In addition, it can be inefficient, if the posterior has relatively compact support within the space of allowed values, since many of the grid points will be in regions with low posterior weight.

6.3 Posterior computation: direct sampling methods

As discussed before, sampling methods attempt to generate a set $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$, from the posterior, which can be used to approximate integrals over the posterior

$$\int f(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{x})\, \mathrm{d}\boldsymbol{\theta} \approx \frac{1}{M} \sum_{i=1}^{M} f(\boldsymbol{\theta}_i).$$

Sampling methods can be **direct** or **stochastic**. Direct methods draw samples directly (or nearly directly) from the target probability distribution. Stochastic methods use **Markov chain Monte Carlo** methods to generate a sequence of samples that are drawn form the target distribution.

6.3.1 Method of inversion

The method of inversion is a simple application of the probability integral transformation. If we denote by F the cumulative distribution function of some random variable X, then the random variable $F(X)$ follows a $U[0,1]$ distribution. Therefore, if we can analytically compute the inverse of the cumulative distribution function, we can generate samples form X by generating samples from a uniform distribution. If

$$F(x) = \mathbb{P}(X \leq x)$$

and it has inverse F^{-1} then the algorithm is simply

1. Generate $u \sim U[0,1]$.
2. Compute $x = F^{-1}(u)$.

Example: exponential distribution with parameter r Suppose we want to draw $X \sim \mathrm{Exp}(r)$. The pdf of the exponential distribution is

$$p(x|r) = r \exp(-rx)$$

which has cumulative density function

$$F(X) = \int_0^X r \exp(-rx)\mathrm{d}x = 1 - \exp(-rX).$$

The inverse can be found as

$$u = F(x) \quad \Rightarrow \quad x = F^{-1}(u) = -\frac{1}{r} \ln(1-u).$$

Samples generated by applying this inverse to $U[0,1]$ samples are shown in Figure 15.

6.3.2 Rejection sampling

Rejection sampling draws samples from a distribution that can be directly sampled and then discards a subset of them that do not match the desired distribution. The simplest rejection sampling algorithm draws uniform samples from a box that encloses the distribution. Suppose that we want to draw samples $\theta_1, \ldots, \theta_n$ from a probability distribution with pdf $p(\theta)$ and that the pdf has compact support, so $p(\theta) = 0$ if $\theta \notin [a,b]$. Suppose additionally that the pdf at the mode of the probability distribution is $M = \max[p(\theta)]$. Rejection sampling proceeds as follows

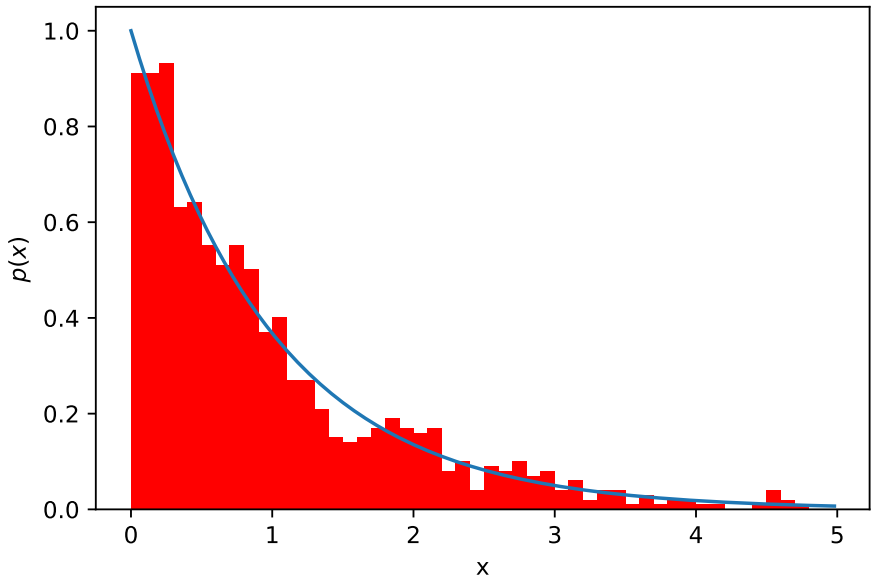


Figure 15: Histogram of samples drawn from the Exp(1) distribution using the method of inversion. The pdf of the exponential distribution is shown as a line for comparison.

1. Draw $\theta \sim U[a, b]$.
2. Draw $y \sim U[0, M]$.
3. If $y \leq p(\theta)$, accept θ as a sample from $p(\theta)$. Otherwise return to step 1.

Example: beta distribution We want to draw samples from a Beta$(3, 2)$ distribution. This has compact support on the interval $[0, 1]$ and the maximum value of the pdf is $M = 16/9$ (EXERCISE). In Figure 16 we illustrate this procedure by indicating which of the first 50 samples drawn in this way are rejected or accepted. In Figure 19 we show a histogram of the accepted samples in the first 1000 draws, which illustrates that the distribution of samples does follow the Beta$(3, 2)$ distribution as desired.

The box rejection sampling procedure does not work at all when the support of the target distribution is unbounded. In addition, it can be very inefficient for compact distributions with long tails. An alternative approach is to draw samples from an easy-to-sample distribution, $g(\theta)$, that is similar to the target distribution $p(\theta)$. First we find a number M such that $Mg(\theta) \geq p(\theta)\ \forall\theta$, i.e., we require $Mg(\theta)$ to contain the target distribution. The algorithm is then

1. Draw $\theta \sim g(\theta)$.
2. Draw $y \sim U[0, 1]$.
3. If $y \leq p(\theta)/(Mg(\theta))$, accept θ as a sample from $p(\theta)$. Otherwise return to step 1.

Trial samples are taken uniformly from within the region between the curve $Mg(\theta)$ and the θ axis. Samples that fall in the region between $p(\theta)$ and $Mg(\theta)$ are rejected. Therefore we

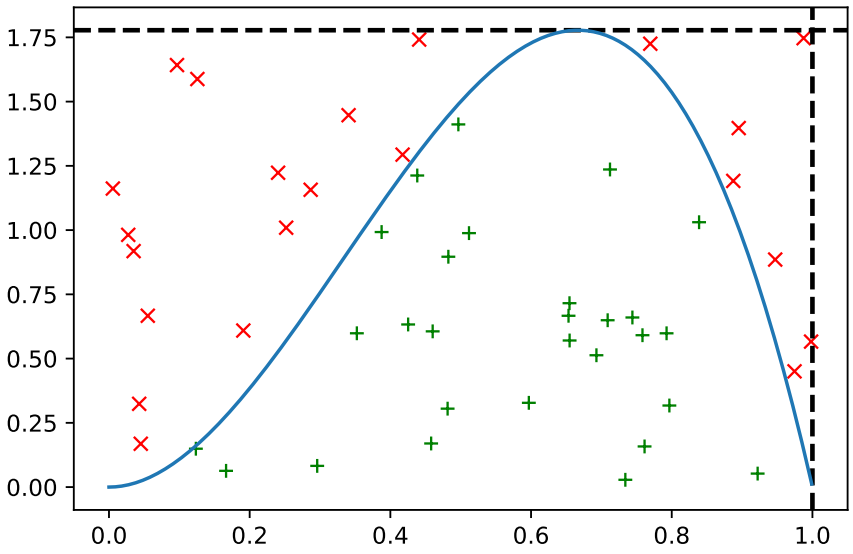


Figure 16: Accepted (green plusses) and rejected (red crosses) samples in the first 50 draws of the rejection sampling algorithm used to simulate the Beta$(3, 2)$ distribution. Only samples that lie within the target pdf are accepted.

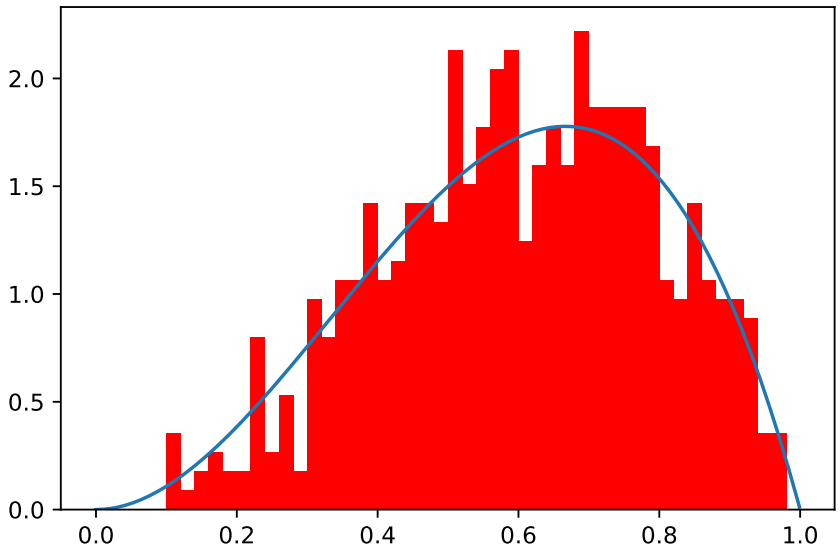


Figure 17: Histogram of the accepted samples in 1000 iterations of the rejection sampling algorithm. We compare the distribution to Beta$(3, 2)$, which is the target distribution.

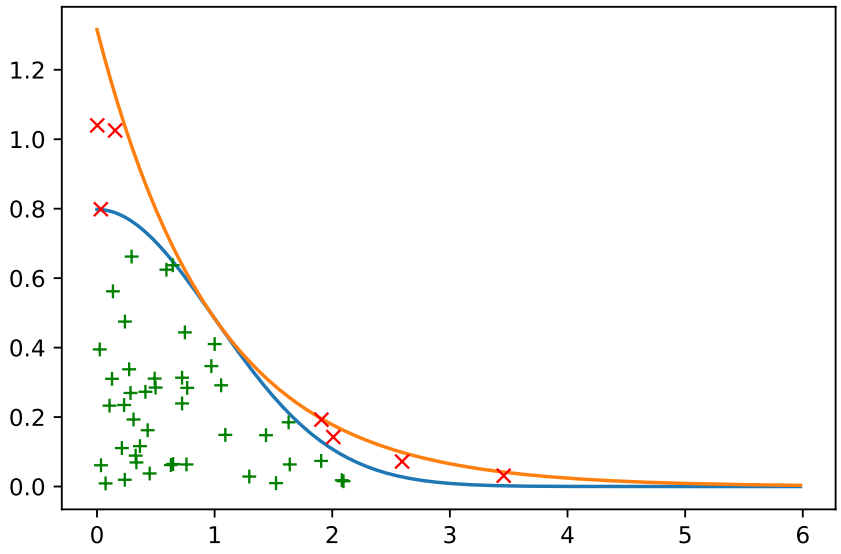


Figure 18: Accepted (green plusses) and rejected (red crosses) samples in the first 50 draws of the rejection sampling algorithm to simulate draws from a half-Normal distribution. Only samples that lie within the target pdf are accepted.

make the efficiency (i.e., the fraction of samples that are accepted) as large as possible by making the choice

$$M = \sup_{\theta} \left(\frac{p(\theta)}{g(\theta)} \right).$$

Example: half-Normal distribution We want to draw samples from the half-Normal distribution with pdf

$$p(\theta) = \begin{cases} \sqrt{\frac{2}{\pi}} \mathrm{e}^{-\frac{\theta^2}{2}} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

We will take $g(\theta) = \exp(-\theta)$, i.e., the exponential distribution with rate 1. We find M from

$$M = \sup_{\theta} \left(\frac{p(\theta)}{g(\theta)} \right) = \sup_{\theta > 0} \left(\sqrt{\frac{2}{\pi}} \exp\left[-\frac{1}{2}(\theta - 1)^2 + \frac{1}{2} \right] \right) = \sqrt{\frac{2}{\pi}} \mathrm{e}^{\frac{1}{2}}.$$

In Figure 18 we show the samples accepted and rejected during the first 50 iterations of the algorithm, and in Figure **??** we show a histogram of the accepted samples during 1000 iterations of the algorithm. We see that the histogram is correctly approximating the desired distribution.

6.3.3 Importance sampling

Rejection sampling can be effective and easy to implement, but it is not always possible to find an easy-to-sample target distribution that closely matches the target distribution. Additionally effort is wasted drawing samples and evaluating the posterior at points which

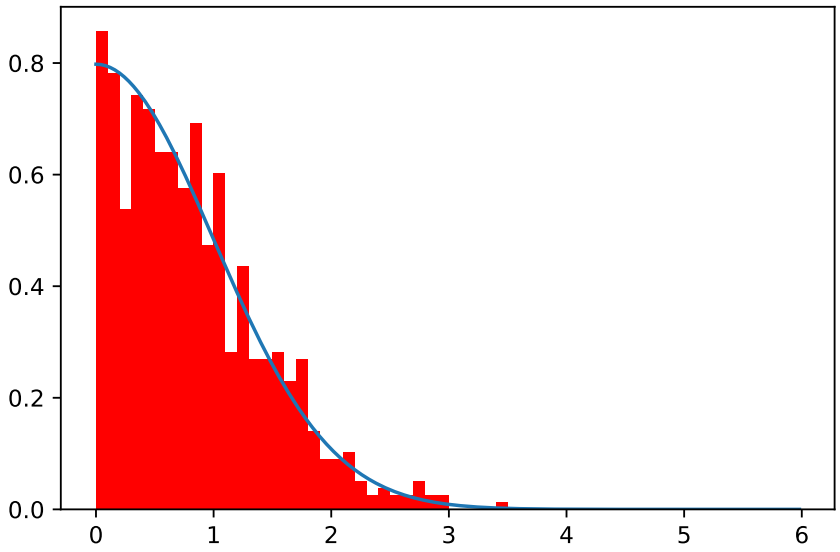


Figure 19: Histogram of the accepted samples in 1000 iterations of the rejection sampling algorithm. We compare the distribution to the target distribution, which in this case is a half-Normal distribution with mean 0 and variance 1.

are subsequently discarded as rejected samples. **Importance sampling** attempts to address the latter problem by using all samples.

Importance sampling uses an easy-to-sample reference distribution $g(\theta)$ as before, but now this is not rescaled, the only stipulation is that the support is common to that of the target distribution, i.e., if $p(\theta) > 0$ then $g(\theta) > 0$. No samples are discarded. Instead the samples are defined **importance weights** via

$$w_i = \frac{p(\theta)}{g(\theta)}$$

and integrals over the target distribution are approximated by weighted averages over the samples

$$\int f(\theta)p(\theta)\mathrm{d}\theta \approx \frac{1}{M}\sum_{i=1}^{M} w_i f(\theta_i).$$

It is straightforward to see that

$$\mathbb{E}_g(w_i f(\theta_i)) = \int w(\theta)f(\theta)g(\theta)\mathrm{d}\theta = \int \frac{p(\theta)}{g(\theta)}f(\theta)g(\theta)\mathrm{d}\theta = \int f(\theta)p(\theta)\mathrm{d}\theta = \mathbb{E}_p(f(\theta))$$

so the importance sampling estimate is unbiased. However

$$\begin{aligned}\mathrm{var}_g(w_i f(\theta_i)) &= \int w^2(\theta)f^2(\theta)g(\theta)\mathrm{d}\theta - [\mathbb{E}_p(f(\theta))]^2 = \int \frac{p(\theta)}{g(\theta)}f^2(\theta)p(\theta)\mathrm{d}\theta - [\mathbb{E}_p(f(\theta))]^2 \\ &= \mathbb{E}_p\left(\frac{p(\theta)}{g(\theta)}f^2(\theta)\right) - [\mathbb{E}_p(f(\theta))]^2. \end{aligned} \tag{97}$$

We see that the importance sampling estimate can suffer from high variance if $g(\theta)$ is much smaller than $p(\theta)$ in regions where the function of interest has significant support.

Note that the above assumes that the normalisation of the target distribution is known, but this is not always the case when sampling from posterior distributions due to the difficulty of computing the Bayesian evidence. If the posterior is not normalised the weights can be renormalised as

$$\tilde{w}_i = \frac{w_i}{\sum_{j=1}^{M} w_j}.$$

The results on the mean and variance are now only approximate, but are valid asymptotically.

Example: Cauchy distribution Suppose we have a standard Cauchy distribution with pdf

$$p(\theta) = \frac{1}{\pi(1+\theta^2)}$$

and want to compute $\mathbb{P}(\theta > 2)$. We can sample from the distribution $g(\theta) = 2/\theta^2 \mathbb{I}(\theta > 2)$ using the method of inversion. This has the same support as the portion of $p(\theta)$ of interest. We define the importance weights

$$w_i = \frac{\theta_i^2}{2\pi(1+\theta_i^2)}$$

and then compute

$$\hat{p}_{>2} = \frac{1}{M}\sum_{i=1}^{M} w_i$$

since we are interested in $\mathbb{P}(\theta > 2)$ which is the integral of $\mathbb{I}(\theta > 2)$, but this equal to 1 throughout the region where $g(\theta)$ has support. Note that in this case it would be wrong to renormalise the weights since then we would compute the probability as 1. As an exercise, verify that using the above weights in the usual sampling estimate gives the expected result.

In Figure 20 we show the convergence of the importance sampling estimate of $\mathbb{P}(\theta > 2)$ as a function of the number of importance samples. We see that it converges much faster than if we used Monte Carlo draws from the Cauchy distribution itself. The correct probability is $\pi/2 - \tan^{-1}(2)/\pi = 0.14758$.

6.3.4 Sampling importance resampling

Sampling importance resampling is a simple extension of importance sampling that uses the importance samples to generate samples approximately from the target distribution. Given M importance samples, $\{\theta_1, \ldots, \theta_M\}$, the importance weights are computed and normalised as described above. Then M samples, $\{\phi_1, \ldots, \phi_M\}$ are drawn, with replacement, from the original set using the normalised weights as probabilities. Integrals over the target distribution can then be approximated by

$$\int f(\theta)p(\theta)\mathrm{d}\theta \approx \frac{1}{M}\sum_{i=1}^{M} f(\phi_i).$$

Sampling importance resampling is a form of **particle filtering**. One problem that it can suffer form is **particle depletion**, where a small number of samples carry the majority of

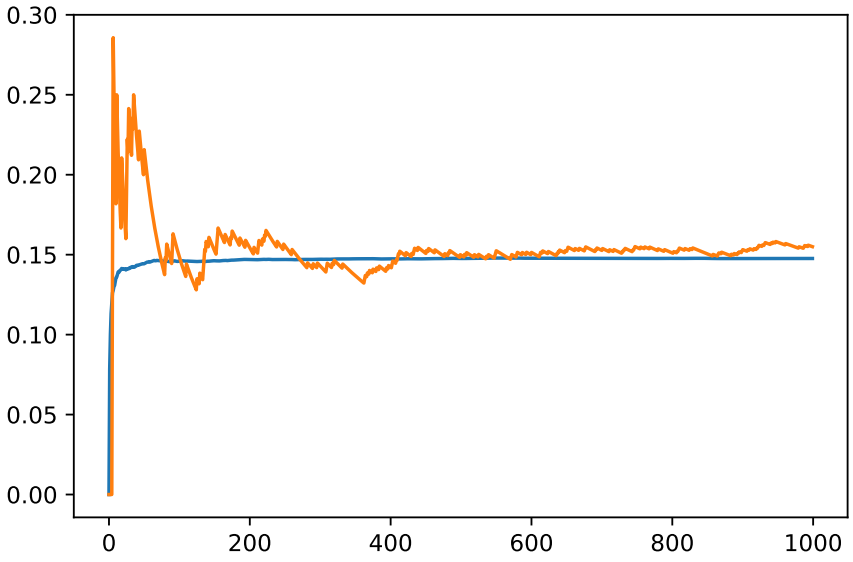


Figure 20: Importance sampling estimate of $\mathbb{P}(\theta > 2)$ for the standard Cauchy distribution as a function of number of samples (blue line), compared to Monte Carlo estimate using direct samples from the Cauchy distribution (yellow line).

the weight and therefore only a small number of points are represented repeatedly in the final data set. Particle depletion leads to poor estimates of derived quantities.

Example: Cauchy distribution We use sampling importance resampling to generate samples from the Cauchy distribution with $\theta > 2$ using the samples generated for the example in the previous section. A histogram of these values is shown in Figure 21, where they are compared to the target distribution, which is a truncated Cauchy distribution.

6.4 Posterior computation: Markov chain Monte Carlo

Direct sampling methods suffer form the problem of dimensionality. They are typically easy to implement in one dimension, but become increasingly challenging, inefficient or impossible to implement as the number of dimensions increases. In higher dimensions it is more common to use stochastic methods, in which a sequence of samples is constructed that has a distribution that follows the target distribution. Typically this is done using Markov chain Monte Carlo algorithms.

A **Markov Chain** is a sequence of random numbers, $\theta^1, \theta^2 \ldots$, such that the value of θ^{n+1} depends only on the previous values, θ^n, and not on earlier numbers in the sequence. A Markov chain can be simulated using a **transition kernel**, $\mathcal{K}(\theta^{n+1}|\theta^n)$, which is a conditional probability distribution for θ^{n+1} given the value of θ^n. The transition kernel uniquely defines the Markov chain. If we assume the Markov chain is **aperiodic** and **irreducible** then the distribution of samples in the Markov chain will converge to a **stationary distribution**, which is independent of the initial starting state of the chain. In Bayesian inference, the goal is to construct a Markov chain such that the stationary distribution is the posterior distribution, $p(\boldsymbol{\theta}|\mathbf{x})$.

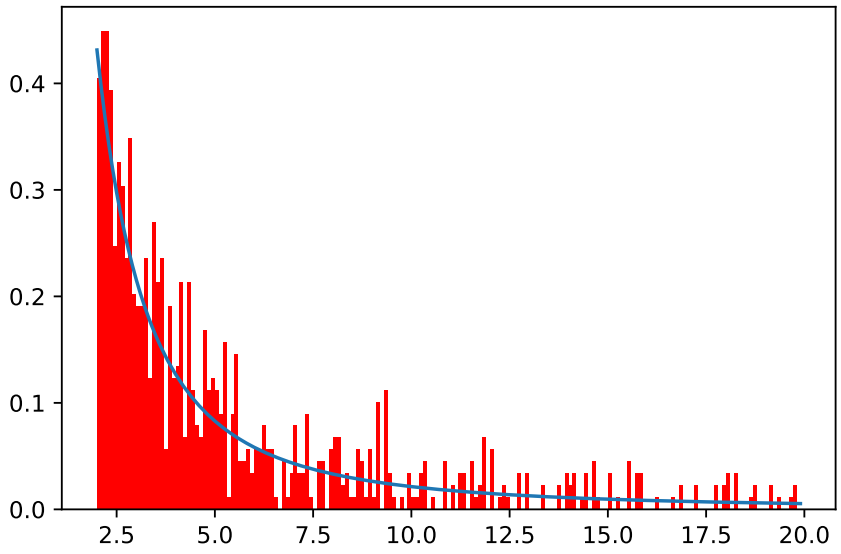


Figure 21: Histogram of 1000 sampling importance resampling samples for the Cauchy distribution in the region $\theta > 2$. These were generated from the importance samples constructed in the example in the last section. The line shows the expected distribution, which is the truncated standard Cauchy distribution.

A Markov chain with transition kernel $\mathcal{K}(\theta^{n+1}|\theta^n)$ is said to satisfy **detailed balance** for a distribution $\pi(\theta)$ if

$$\pi(\theta)\mathcal{K}(\phi|\theta) = \pi(\phi)\mathcal{K}(\theta|\phi) \; \forall \, \phi, \theta,$$

in which case $\pi(\theta)$ is the stationary distribution of the Markov chain. Enforcing detailed balance in the Markov chain, for $\pi(\theta) = p(\theta|\mathbf{x})$, will ensure we generate samples from the posterior distribution.

There are two widely used approaches to construct Markov chains satisfying detailed balance with a particular stationary distribution — Gibbs sampling and the Metropolis-Hastings algorithm.

6.4.1 Gibbs Sampling

Gibbs sampling for multi-variate probability distributions works by sampling sequentially from full conditional distributions on each parameter given the current state of the other parameters. Algorithmically it works as follows. We suppose that the distribution of interest, $p(\boldsymbol{\theta}|\mathbf{x})$, depends on a multi-dimensional parameter vector, $(\theta) = (\theta_1, \theta_2, \ldots, \theta_p)$. We use $(\theta)^k$, θ_i^k to denote the value of the full parameter vector and its i'th component at iteration k of the algorithm. We denote by $\boldsymbol{\theta}_{(i)}$ the vector of all parameter values except the i'th and use $p(\theta_i|\boldsymbol{\theta}_{(i)}, \mathbf{x})$ to denote the full conditional distribution of θ_i, given the values of all the other components and the data. If the value of the Markov chain at step t is $\boldsymbol{\theta}^t$, then the value at step $t+1$ is obtained via

- Sample θ_1^{t+1} from $p(\theta_1|\theta_2^t, \theta_3^t, \ldots, \theta_p^t, \mathbf{x})$.
- Sample θ_2^{t+1} from $p(\theta_2|\theta_1^{t+1}, \theta_3^t, \ldots, \theta_p^t, \mathbf{x})$.

- . . . . . .
- Sample θ_i^{t+1} from $p(\theta_i|\theta_j^{t+1}$ for $j < i$ and θ_j^t for $j > i, \mathbf{x})$.
- . . . . . .
- Sample θ_p^{t+1} from $p(\theta_p|\theta_1^{t+1}, \ldots, \theta_{p-1}^{t+1}, \mathbf{x})$.

This set of sequential updates is repeated at each iteration of the algorithm to generate a set of samples from the target distribution.

The transition kernel in Gibbs sampling is

$$\mathcal{K}_G(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t) = \prod_{i=1}^{k} p(\theta_i|\theta_j^{t+1} \text{ for } j < i \text{ and } \theta_j^t \text{ for } j > i, \mathbf{x})$$

which satisfies detailed balance with target distribution $p(\boldsymbol{\theta}|\mathbf{x})$.

6.4.2 Metropolis-Hastings algorithm

In the Metropolis-Hastings algorithm all the parameters of the model are typically updated simultaneously. This is achieved using a **proposal distribution**, $q(\boldsymbol{\phi}|\boldsymbol{\theta})$, to propose a new point $\boldsymbol{\phi}$, given the current parameter values $\boldsymbol{\theta}$. The algorithm is as follows

1. Initialise $\boldsymbol{\theta}^0$ by drawing from a distribution of starting vlaues (often the prior can be used for this).
2. At step t:
   (a) Propose a new point $\boldsymbol{\phi} \sim q(\boldsymbol{\phi}|\boldsymbol{\theta}^{t-1})$.
   (b) Compute the **acceptance probability**
   $$\alpha = \min\left(1, \frac{p(\boldsymbol{\phi}|\mathbf{x})q(\boldsymbol{\theta}^{t-1}|\boldsymbol{\phi})}{p(\boldsymbol{\theta^{t-1}}|\mathbf{x})q(\boldsymbol{\phi}|\boldsymbol{\theta}^{t-1})}\right).$$
   (c) Draw $u \sim U[0,1]$. If $u < \alpha$, set $\boldsymbol{\theta}^t = \boldsymbol{\phi}$, otherwise set $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$.
3. Repeat until the desired number of iterations, T, have been completed.

the initial version of this algorithm, due to Metropolis, used symmetric proposal distributions and so that factor cancels out of the acceptance probability. A subsequent paper by Metropolis and Hastings generalised the result to non-symmetric proposals.

It can be readily verified in this case as well that the Markov chain constructed in this way satisfies detailed balance with target distribution equal to the posterior $p(\boldsymbol{\theta}|\mathbf{x})$.

There are a few special cases of the Metropolis-Hastings algorithm

- **The Metropolis Algorithm** This is the case described above where the proposal is symmetric, $q(\boldsymbol{\phi}|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\phi})$, and the acceptance probability reduces to
  $$\alpha = \min\left(1, \frac{p(\boldsymbol{\phi}|\mathbf{x})}{p(\boldsymbol{\theta^{t-1}}|\mathbf{x})}\right).$$

- **Random Walk Metropolis** If we use $q(\boldsymbol{\phi}|\boldsymbol{\theta}) = f(\boldsymbol{\theta} - \boldsymbol{\phi})$, with f some function satisfying $f(\mathbf{y}) = f(-\mathbf{y})$, then the kernel driving the chain is a random walk. This is a symmetric proposal and so the accpetance probability is as in the Metropolis Algorithm above.

- **The Independence Sampler** If we take $q(\boldsymbol{\phi}|\boldsymbol{\theta}) = f(\boldsymbol{\phi})$, the candidate value is independent of the current value. The acceptance probability is

$$\alpha = \min\left(1, \frac{w(\boldsymbol{\phi})}{w(\boldsymbol{\theta})}\right)$$

  where $w(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{x})/f(\boldsymbol{\theta})$.

- **Single-updates** Individual parameters of the parameter vector can be updated sequentially in the Metropolis-Hastings algorithm in the same way they are during the Gibbs sampling algorithm. At step t we sequentially propose updates, ϕ_j, to each component, θ_j, of the parameter vector in turn. After updating parameter j, the new parameter vector is $(\theta_1^{t+1}, \ldots \theta_{j-1}^{t+1}, \theta_j, ^{t+1}\, \theta_{j+1}^t, \ldots, \theta_p^t)$. The new value, θ_j^{t+1} is chosen by the algorithm

  1. Propose a new candidate value $\phi_j \sim q(\phi_j|\boldsymbol{\theta}_j^t)$ and set $\boldsymbol{\phi}_j = (\theta_1^{t+1}, \ldots, \theta_{j-1}^{t+1}, \phi_j, \theta_{j+1}^t, \ldots, \theta_p^t)$.
  2. Evaluate the acceptance probability

$$\alpha = \min(1, A), \quad \text{where} \quad A = \frac{p(\boldsymbol{\phi}_j|\mathbf{x})q(\theta_j^t|\boldsymbol{\phi}_j)}{p(\boldsymbol{\theta}_j^t|\mathbf{x})q(\phi_j|\boldsymbol{\theta}_j^t)} = \frac{p(\phi_j|\boldsymbol{\theta}_{(j)}^t, \mathbf{x})q(\theta_j^t|\boldsymbol{\phi}_j)}{p(\theta_j^t|\boldsymbol{\theta}_{(j)}^t, \mathbf{x})q(\phi_j|\boldsymbol{\theta}_j^t)}$$

  3. Draw $u \sim U[0,1]$. If $u < \alpha$, set $\theta_j^{t+1} = \phi_j$, otherwise set $\theta_j^{t+1} = \theta_j^t$.

6.4.3 MCMC diagnostics

The Markov chain is only guaranteed to converge to the stationary distribution asymptotically so it is natural to ask how many samples are needed before the sample is representative of the posterior. The first issue to address is **burn-in**. A Markov chain retains some memory of its initial state for a number of iterations. If the initial sample is in a region of low probability in the stationary distribution, then the first samples will typically not be very characteristic of the stationary distribution. These initial samples should be discarded and samples only retained after the initial burin-in period used for inference. Typically between a few hundred and a few thousand burn-in samples are required and it can be diagnosed using a **trace plot**, which is a plot of the parameter value in the chain as a function of iteration number. Initially the trace plot will show a trend as the chain moves toward parameter values with high posterior support. Once the chain is sampling properly, the values will oscillate back and forth. This is illustrated in Figure 22. The trace plot allows the burn-in period to be identified and removed, and is also a useful diagnostic of the performance of the algorithm. Chains that are moving back and forth rapidly are sampling well from the posterior.

MCMC samples are used to produce Monte Carlo estimates of parameters of interest. If the samples were independent draws from the posterior then these estimates are unbiased and would have a variance that scales like σ^2/M, where σ^2 is the variance of a single sample and M is the number of samples. This could in principle be used to estimate how many

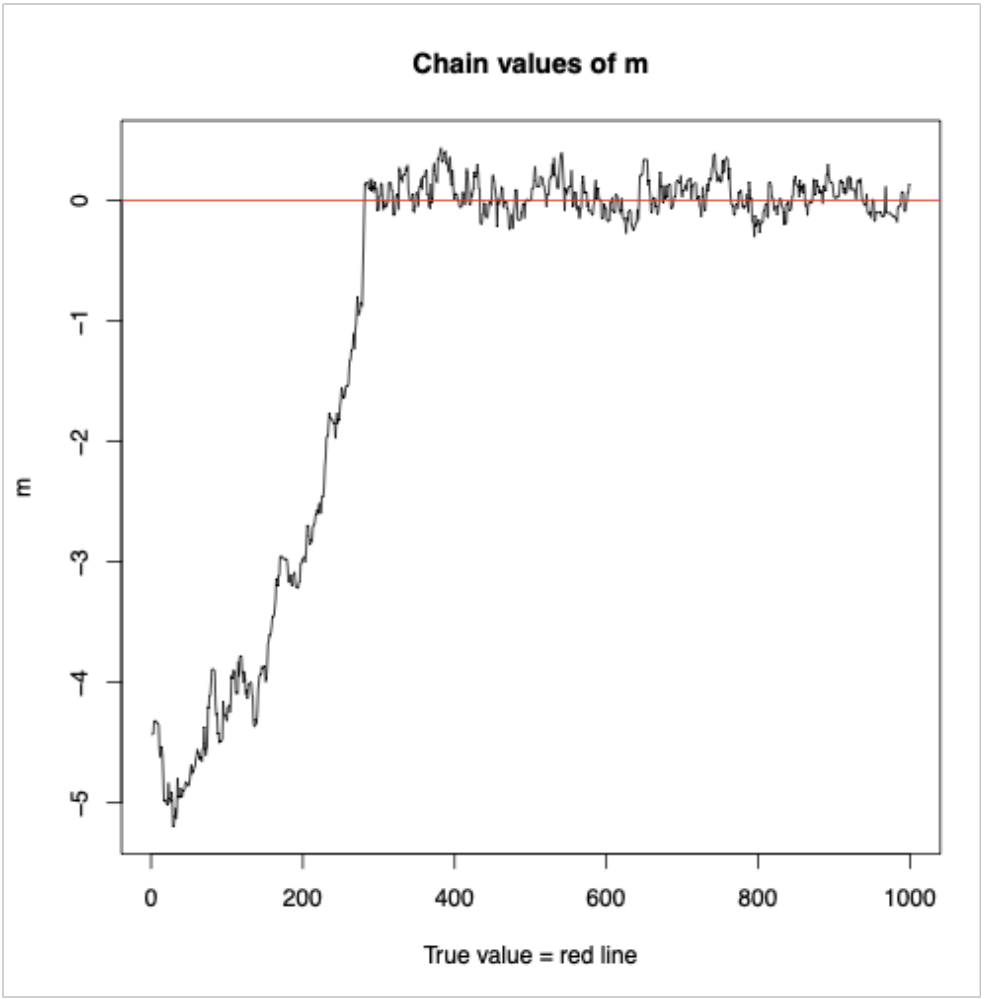


Figure 22: Trace plot for burn-in period of a chain. Initially the chain moves form the starting point to the region of high probability density, so there is a tendency to move in a particular direction. Once the chain reaches the correct region it oscillates back and forth in the region of high posterior support.

samples are needed to achieve a certain target precision on a quantity of interest. However, MCMC samples are not independent. This modifies the variance estimate to

$$\sigma^2 = \mathrm{var}_p(\theta) + 2\sum_{k=2}^{\infty} \mathrm{cov}(\theta^1, \theta^k).$$

This is difficult to compute in practice, so what is usually done is to generate m different chains of length M, estimate the value of the quantity of interest in each one, $\bar{f}_1, \ldots, \bar{f}_m$, compute the value using the pooled samples from all chains, $\bar{f}$, and then construct the *batch means estimate*

$$\hat{\sigma}^2 = \frac{T}{m-1}\sum_{i=1}^{m}(\bar{f}_i - \bar{f})^2.$$

The estimated Monte Carlo error in f is then $\hat{\sigma}^2/n$.

Correlation in MCMC samples can also be estimated using the **autocorrelation function** (ACF). The *lag-k autocorrelation coefficient* or *autocorrelation at lag-k* is $\mathrm{cov}(\theta^i, \theta^{i+k})$ and computed via

$$\rho_k = \frac{\sum_{i=1}^{N-k}(\theta^i - \bar{\theta})(\theta^{i+k} - \bar{\theta})}{\sum_{i=1}^{M}(\theta^i - \bar{\theta})^2}$$

where θ now denotes one parameter of the target distribution, and $\bar{\theta}$ is the mean of that parameter in the chain. Looking at ACF plots is another useful diagnostic of MCMC performance. Examples of good, bad and normal ACF plots are given in Figure 23.

If MCMC chains have very high lags, most likely they are not taking big enough jumps in parameter space and so the size of proposed jumps should be increased. It is typical to monitor **acceptance rates** when using the Metropolis-Hastings algorithm and a target acceptance rate is used to adjust proposed jump sizes. If proposed jumps are too small, the acceptance rate will be high but there will also be high autocorrelation between samples. If the proposed jumps are too large, the acceptance rate will be low, but those samples that are accepted will show very low autocorrelation.Ultimately we care about maximising the rate at which we obtain new independent samples. This can be estimated by tracking the **effective sample size**

$$\mathrm{ESS} = \frac{M}{1 + 2\sum_{k=1}^{\infty}\rho_k}$$

where M is the number of samples in the chain. It has been shown that, under certain assumptions, the optimal rate of obtaining new effective samples is achieved by aiming to have an acceptance rate around 23.4%.

The final diagnostic we will mention here is the use of multiple chains. For complex probability distributions that have many modes it is possible for Markov chains to get stuck sampling from only one of them. Chains starting from different points in parameter space may end up exploring different modes. As a diagnostic of this kind of behaviour, it is good practice to run a handful of runs, starting at different points in parameter space. We can be confident in the final results once the different chains are producing samples that are consistent with one another. This consistency can be quantified using the **Gelman-Rubin statistic**.

Suppose we have m independent chains and have discarded the initial burn-in samples

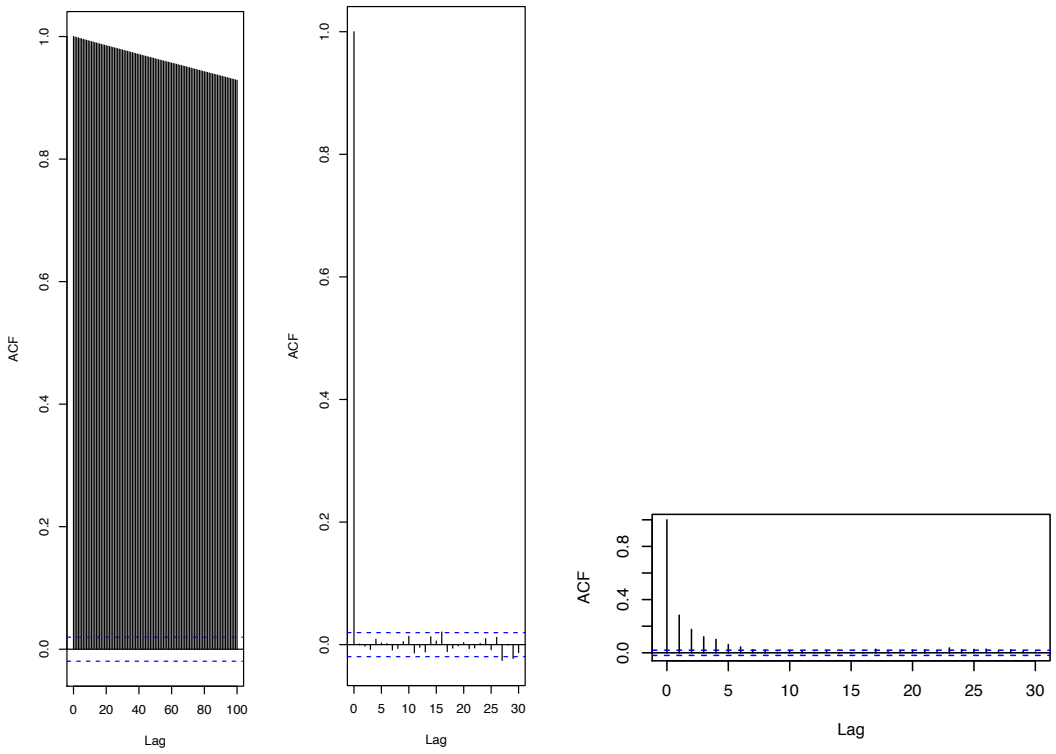


Figure 23: Examples of plots of the autocorrelation function. This should decline to numbers close to 0 for short lags. In the left hand plot, the ACF is still above 0.8 at a lag of 100, indicating highly correlated samples, which is not desirable. In the middle plot we show an ideal example where the ACF is already close to zero at lag of 1, indicating a high level of independence in the samples. The right hand plot is a typical example of MCMC chains that are sampling well. The ACF falls to low values for lags of a few.

to leave chains of length N. We calculate the *within chain variance*

$$W = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{N-1}\sum_{i=1}^{N}(\theta_{ij} - \bar{\theta}_j)^2$$

where θ_{ij} is the i'th sample in the j'th chain. We similarly define the *between chain variance*

$$B = \frac{N}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j - \bar{\bar{\theta}})^2, \quad \text{where } \bar{\bar{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j.$$

Note that we are assuming that θ is a one-dimensional parameter, which could be one component of a multi-dimensional parameter vector. The variance in this parameter can be computed as

$$\text{var}(\theta) = \left(1 - \frac{1}{N}\right)W + \frac{1}{N}B$$

from which the **potential scale-reduction factor** can be computed

$$\hat{R} = \sqrt{\frac{\text{var}(\theta)}{W}}.$$

Values of R greater than about 1.1 or 1.2 indicate that the chains are not yet converged.

6.4.4 Speeding up MCMC

MCMC can be made faster by a good choice of the proposal distribution. Proposal distributions that are well approximated to the form of the target distribution are to be preferred. As well as tuning the proposal distribution, accelerated convergence can be achieved using **annealing**. The idea of annealing is to transform the posterior surface as

$$p(\boldsymbol{\theta}|\mathbf{x}) \to [p(\boldsymbol{\theta}|\mathbf{x})]^{\beta}, \quad \text{where } \beta = \frac{1}{kT}.$$

As $T \to \infty$ the new distribution becomes flatter and flatter, so the contrast in probabilities between different points is reduced. This means that moves proposed in a Metropolis-Hastings algorithm are more likely to be accepted. Figure 24 shows the effect of the annealing transformation on the probability distribution being sampled as the temperature increases.

There are two common applications of annealing. In **simulated annealing** the temperature is gradually changed as the initial phase of the run progresses, according to some scheme, for example, a linear decrease with iteration number. The idea is that in the early phase the chain explores the parameter space widely and rapidly, identifying areas of higher posterior density. As the temperature decreases the chain gets trapped in a region of high posterior probability, hopefully the primary mode of the distribution. The simulated annealing phase does not produce useful samples, since detailed balance is satisfied, but after the simulated annealing phase, the chain will evolve as normal and return valid samples from the posterior.

The other use of annealing is **parallel tempering**. In parallel tempering, a number of chains are evolved simultaneously at different temperatures. At each iteration, a given chain will update its parameters as normal, but with a certain probability an interchange is proposed, in which the states of two chains (usually neighbouring in temperature) will

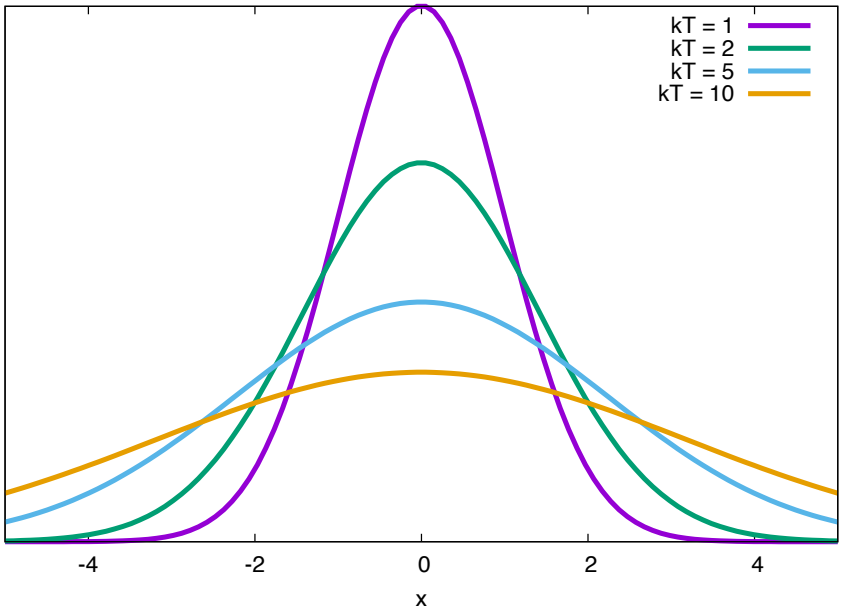


Figure 24: Effect of annealing on the target probability distribution.

be exchanged. If the two chains are labelled i and j, have temperatures T_i and T_j, and current parameter values $\boldsymbol{\theta}^i$ and $\boldsymbol{\theta}^j$, then the appropriate acceptance probability for the swap $\boldsymbol{\theta}^i \leftrightarrow \boldsymbol{\theta}^j$ is

$$\alpha = \min\left(1, \left[\frac{p(\boldsymbol{\theta}^j|\mathbf{x})}{p(\boldsymbol{\theta}^i|\mathbf{x})}\right]^{\frac{1}{T_i}} \left[\frac{p(\boldsymbol{\theta}^i|\mathbf{x})}{p(\boldsymbol{\theta}^j|\mathbf{x})}\right]^{\frac{1}{T_j}}\right).$$

The idea of parallel tempering is that higher posterior density regions of the parameter space that the widely-exploring high temperature chains identify, propagate down to lower temperature chains, which explore them thoroughly. Efficiency is dependent on the difference in the temperatures of neighbouring chains, so the number of chains and their spacing must be tuned for each given problem.

6.5 Posterior computation: variable model dimension

In some circumstances we might be interested in fitting multiple different models to the data simultaneously. the most common situation is when the total number of parameters needed to describe the data is unknown. In a gravitational wave context this arises when the total number of sources present in the data set is unknown, e.g., for the LISA gravitational wave detector. In these circumstances one can still construct Markov chains, but now these chains can move between different models. The fraction of samples that the chain spends in each model is proportional to the evidence for that model and, in the case of models that differ only in the total number of sources, the evidences give the relative probabilities for the unknown number of sources in the data.

The most widely used algorithm for fitting multiple models is **reversible jump Markov**

chain Monte Carlo (RJMCMC). RJMCMC generates a Markov chain such that at each step either an update within the model is proposed, or, with a certain probability, a jump to an alternative model is proposed. Usually the jumps are between models that differ by only one source if that is the type of model hierarchy being considered. When proposing a jump to a new model, with parameters $\boldsymbol{\theta}'$, the values of the parameters of that model must also be proposed. This is achieved by generating a set of random numbers $\mathbf{u}$ from some distribution $q(\mathbf{u})$. In order to ensure reversibility we imagine that these random numbers are part of the parameters of the model, but because they are random we only need to generate them when they are used in a between-model jump. Similarly we may need some random variables $\mathbf{u}'$ to propose jumps back from the new model space to the original model parameters $\boldsymbol{\theta}$. The dimensionality of the joint space $(\boldsymbol{\theta}, \mathbf{u})$ must equal that of $(\boldsymbol{\theta}', \mathbf{u}')$ and there will be a deterministic, invertible mapping between the two. In the case of nested models, the reverse jump might just delete a set of parameters and so the dimensionality of $\mathbf{u}'$ is 0. However, if the particular source is deleted at random rather than, say, the lowest SNR source always being deleted, a random variable that selects which source to delete is required. The generalisation of the acceptance probability for RJMCMC is

$$\alpha = \min\left(1, \frac{p(\boldsymbol{\theta}'|\mathbf{x})q(\boldsymbol{\theta}')}{p(\boldsymbol{\theta}|\mathbf{x})q(\boldsymbol{\theta})}\left|\frac{\partial(\boldsymbol{\theta}', \mathbf{u}')}{\partial(\boldsymbol{\theta}, \mathbf{u})}\right|\right)$$

where the last term is the Jacobian for the transformation between the two sets of variables.

Example: mixture of Gaussians Suppose that model M_1 is a single Gaussian with mean θ_1 and unit variance and model M_2 is a mixture of two Gaussians with means θ_1' and θ_2' and both of unit variance. We have random variables $\mathbf{u} = (u_1, u_2)$ with $u_1 \sim N(0, \sigma_0^2)$ and $u_2 \sim U[0,1]$ in the M_1 model space and $\mathbf{u}' = u_1' \sim U[0,1]$ in the M_2 model space. The random variable u_1 gives the value of the mean of the new Gaussian to be added, while u_1' selects which Gaussian to delete in the reverse step. The second random variable u_2 ensures the dimensionality is consistent. We can define the mapping between the parameter spaces via

$$\begin{aligned}
\theta_1' &= \begin{cases} \theta_1 & \text{if } u_2 < 0.5 \\ u_1 & \text{if } u_2 \geq 0.5 \end{cases} \\
\theta_2' &= \begin{cases} u_1 & \text{if } u_2 < 0.5 \\ \theta_1 & \text{if } u_2 \geq 0.5 \end{cases} \\
u_1' &= u_2.
\end{aligned}$$

and the reverse mapping

$$\begin{aligned}
\theta_1 &= \begin{cases} \theta_1' & \text{if } u_1' < 0.5 \\ \theta_2' & \text{if } u_1' \geq 0.5 \end{cases} \\
u_1 &= \begin{cases} \theta_2' & \text{if } u_1' < 0.5 \\ \theta_1' & \text{if } u_1' \geq 0.5 \end{cases} \\
u_2 &= u_1'.
\end{aligned}$$

The Jacobian for this transformation is 1 and so the acceptance probability is just

$$\alpha = \min\left(1, \frac{p(\boldsymbol{\theta}'|\mathbf{x})q(\boldsymbol{\theta}')}{p(\boldsymbol{\theta}|\mathbf{x})q(\boldsymbol{\theta})}\right).$$

6.6 Evidence computation

As described earlier, the Bayesian evidence is required for model comparison and Bayesian model selection, but it is difficult to compute accurately using standard MCMC methods. **Nested sampling** (Skilling 2004) was developed as an alternative approach, specifically tuned for evidence computation. It calculates the evidence by transforming the multi–dimensional evidence integral into a one–dimensional integral that is easy to evaluate numerically. This is accomplished by defining the prior volume X as $dX = \pi(\mathbf{\Theta})d^D\mathbf{\Theta}$, so that

$$X(\lambda) = \int_{\mathcal{L}(\mathbf{\Theta})>\lambda} \pi(\mathbf{\Theta})d^N\mathbf{\Theta}, \tag{98}$$

where the integral extends over the region(s) of parameter space contained within the iso-likelihood contour $\mathcal{L}(\mathbf{\Theta}) = \lambda$. The evidence integral, Eq. (77), can then be written as

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X)dX, \tag{99}$$

where $\mathcal{L}(X)$, the inverse of Eq. (98), is a monotonically decreasing function of X. Thus, if one can evaluate the likelihoods $\mathcal{L}_i = \mathcal{L}(X_i)$, where X_i is a sequence of decreasing values,

$$0 < X_M < \cdots < X_2 < X_1 < X_0 = 1, \tag{100}$$

as shown schematically in Fig. 25, the evidence can be approximated numerically using standard quadrature methods as a weighted sum

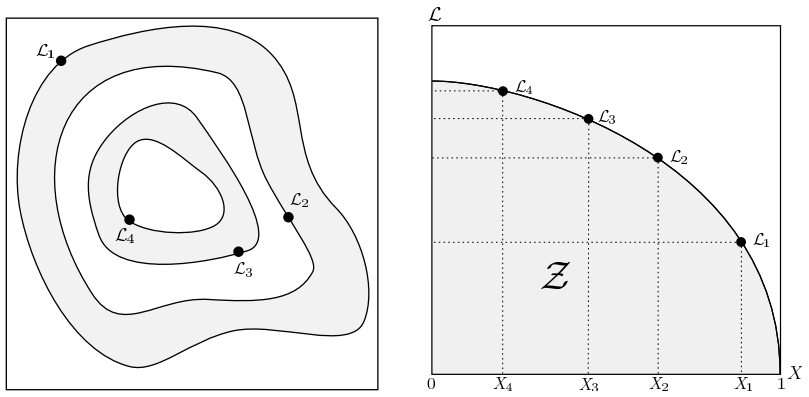
$$\mathcal{Z} = \sum_{i=1}^{M} \mathcal{L}_i w_i, \tag{101}$$

where the weights w_i for the simple trapezium rule are given by $w_i = \frac{1}{2}(X_{i-1} - X_{i+1})$. An example of a posterior in two dimensions and its associated function $\mathcal{L}(X)$ are also shown in Fig. 25.

6.6.1 Evidence Evaluation

The summation in Eq. (101) can be performed using a set of 'active' (or 'live') points. The algorithm proceeds as follows

- Set the iteration counter to $i = 0$. Draw N samples from the full prior, $\pi(\mathbf{\Theta})$. The initial prior volume is $X_0 = 1$.
- Sort the samples in order of likelihood. Denote the lowest likelihood by $\mathcal{L}_0$ and remove the corresponding point from the active set, hence becoming 'inactive'.
- Draw a new point uniformly from the prior, subject to the constraint that the point has likelihood $\mathcal{L} > \mathcal{L}_0$. The prior volume within this iso-likelihood contour is $X_1 = t_1 X_0$, where t_1 follows the distribution $\mathbb{P}(t) = Nt^{N-1}$ (i.e., the probability distribution for the largest of N samples drawn uniformly from the interval $[0, 1]$).
- For each subsequent iteration, i, repeat the procedure of finding the lowest likelihood, $\mathcal{L}_i$, in the active set and removing the corresponding point, drawing a replacement point uniformly from within the prior volume with $\mathcal{L} > \mathcal{L}_i$, and reducing the enclosed prior volume $X_i = t_i X_{i-1}$.



(a) (b)

Figure 25: Cartoon illustrating (a) the posterior of a two dimensional problem; and (b) the transformed $\mathcal{L}(X)$ function where the prior volumes X_i are associated with each likelihood $\mathcal{L}_i$.

- Continue until the entire prior volume has been traversed, assessed by some pre-specified stopping criterion (see section 6.6.2).

The algorithm travels through nested shells of likelihood as the prior volume is reduced. The prior volume at step i is

$$\log X_i = \sum_{j=1}^{i} \log t_j,$$

which is the sum of i independent, identically distributed random variables and so has a mean equal to $i\,\mathbb{E}[\log t]$ and a variance i var$[\log t]$. A simple calculation gives

$$\mathbb{E}[\log t] = -1/N, \quad \text{var}[\log t] = 1/N^2. \tag{102}$$

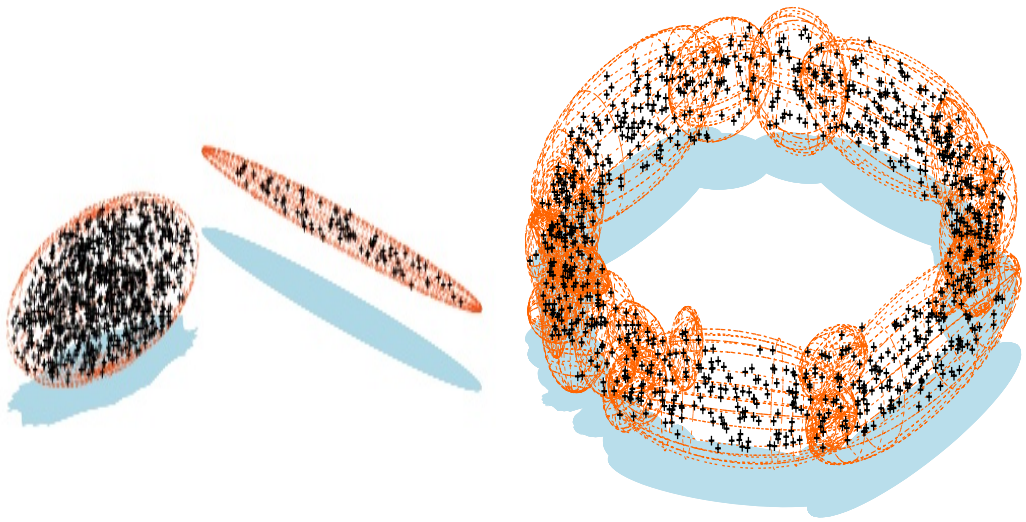
After i iterations the prior volume will have shrunk down such that $\log X_i \approx -(i \pm \sqrt{i})/N$. So, to evaluate the sum (101), we can take $X_i = \exp(-i/N)$.

6.6.2 Stopping Criterion

The nested sampling algorithm should be terminated once the evidence has been computed to a pre-specified precision. One way to ensure this is to proceed until the evidence estimated at each replacement changes by less than a specified tolerance. Skilling suggested an alternative, more robust, condition based on determining an upper limit on the missing portion of the evidence. By selecting the maximum–likelihood, $\mathcal{L}_{\max}$, in the set of active points, the largest evidence contribution that can be made by the remaining portion of the posterior is $\Delta\mathcal{Z}_{\rm i} \approx \mathcal{L}_{\max} X_{\rm i}$, i.e., the product of the remaining prior volume and maximum likelihood value. The algorithm should stop when this quantity drops below some user–defined value, e.g., 0.5 in log-evidence.

6.6.3 Posterior Inferences

Once the evidence $\mathcal{Z}$ is found, posterior inferences can be easily generated using the final live points and the full sequence of discarded points, i.e., the points with the lowest likelihood



(a) (b)

Figure 26: Illustrations of the ellipsoidal decompositions performed by MultiNest. The points given as input are overlaid on the resulting ellipsoids. 1000 points were sampled uniformly from: (a) two non-intersecting ellipsoids; and (b) a torus.

value at each iteration i of the algorithm. Each such point is simply assigned the probability weight

$$p_i = \frac{\mathcal{L}_i w_i}{\mathcal{Z}}. \tag{103}$$

These samples can then be used to calculate inferences of posterior parameters such as means, standard deviations, covariances and so on, or to construct marginalised posterior distributions.

6.6.4 MultiNest Algorithm

The most challenging task in implementing the nested sampling algorithm is drawing samples from the prior within the hard constraint $\mathcal{L} > \mathcal{L}_i$ at each iteration i. Employing a naive approach that draws blindly from the prior would result in a steady decrease in the acceptance rate of new samples with decreasing prior volume (and increasing likelihood). The MultiNest algorithm tackles this problem through an ellipsoidal rejection sampling scheme by enclosing the live point set within a set of (possibly overlapping) ellipsoids. A new point is then drawn uniformly from the region enclosed by these ellipsoids. The number of points in an individual ellipsoid and the total number of ellipsoids is decided by an 'expectation–maximization' algorithm so that the total sampling volume, which is equal to the sum of volumes of the ellipsoids, is minimized. This allows maximum flexibility and efficiency. Simple Gaussian-like modes are decomposed into a relatively small number of ellipsoids, but modes with complex curving degeneracies are broken up into a relatively large number of small 'overlapping' ellipsoids (see Fig. 26).

This ellipsoidal decomposition scheme also provides a mechanism for mode identification. Ellipsoids that overlap are regarded as part of the same ellipsoidal chain. The algorithm can then identify distinct modes with distinct ellipsoidal chains. In panel (a) of Fig. 26 the algorithm identifies two modes, while in panel (b) it correctly identifies the existence of just

one single mode, even though a large number of ellipsoids are needed to cover it. Once distinct modes have been identified, they are evolved independently.

Another feature of the MultiNest algorithm is the evaluation of the global as well as the 'local' evidence values associated with each mode. These evidence values can be used to calculate the probability that an identified 'local' peak in the posterior corresponds to a real feature.

There are many other nested sampling algorithms around today, including Polychord, which obtains samples from within the iso-likelihood surface through slice sampling, cpnest and dynesty. The latter two samplers form part of the Bilby parameter estimation software suite for LIGO.

7 Examples of Bayesian statistics in gravitational wave astronomy

In this section we will provide some examples of the application of Bayesian statistics in gravitational wave astronomy. In most cases we will briefly outline what is done, and provide references where further information can be obtained.

7.1 LIGO Parameter Estimation

Parameter estimation results for sources detected by the LIGO interferometers are obtained and summarised as posterior distributions using the Bayesian techniques described earlier in this course. Typically, LIGO parameter estimation results are quoted as posterior medians and symmetric credible intervals. In this section we will illustrate the ways that the LVC presents observational results using results from the first LVC *Gravitational Wave Transient Catalogue*, GWTC-1 (Abbott et al. (2019), *Phys. Rev.* X **9** 031040). This described the properties of all events (10 BBH and 1 BNS) observed during the O1 and O2 observing runs. Figure 27 shows the primary results table from GWTC-1, which summarises parameter estimation results for all of the events observed by LIGO and Virgo during O1 and O2.

The LVK collaboration has published two additional catalogues since GWTC-1. GWTC-2 (*Phys. Rev.* X **11**, 021053 (2021), arxiv:2010.14527) describes all the events observed in O3a, the first half of the O3 observing run. GWTC-3 (arxiv:2111.03606) describes all events observed up until the end of the O3 observing run in March 2019. The aim of this part of the notes is to illustrate how the LVLK presents results, and so we use GWTC-1 as the number of events is somewhat more manageable, but the ways in which results are presented are the same in all three catalogues.

LIGO/Virgo parameter estimation results in O1 and O2 were computed using the *LALInference* software suite, which includes two separate parameter estimation codes. *LALInferenceMCMC* is a Markov Chain Monte Carlo code, which generates posterior distributions using the Metropolis-Hastings algorithm and proposal distributions that are tuned to features expected in the likelihood for gravitational wave observations of compact binary inspirals. Further details can be found in

- Röver, C., Meyer, R., and Christensen, N., *Bayesian Inference on Compact Binary Inspiral Gravitational Radiation Signals in Interferometric Data*, *Class. Quantum Grav.* **23**, 4895 (2006).

- van der Sluys, M., Raymond, V., Mandel, I., Röver, C., Christensen, N., Kalogera, V., Meyer, R., and Vecchio, A., *Parameter Estimation of Spinning Binary Inspirals Using Markov-Chain Monte Carlo*, *Class. Quantum Grav.* **25**, 184011 (2008).

LALInferenceNest is a nested sampling algorithm, which obtains candidate values for updates to the live point set by carrying out short MCMC chains originating at the current lowest likelihood point in the live point set. Further details can be found in

- Veitch, J., and Vecchio, A., *Phys. Rev.* D **81**, 062003 (2010).

A summary of the *LALInference* package can be found in

Event	$m_1/M_\odot$	$m_2/M_\odot$	$\mathcal{M}/M_\odot$	$\chi_{\rm eff}$	$M_f/M_\odot$	a_f	$E_{\rm rad}/(M_\odot c^2)$	$\ell_{\rm peak}/({\rm erg\,s^{-1}})$	$d_L/{\rm Mpc}$	z	$\Delta\Omega/{\rm deg}^2$
GW150914	$35.6^{+4.7}_{-3.1}$	$30.6^{+3.0}_{-4.4}$	$28.6^{+1.7}_{-1.5}$	$-0.01^{+0.12}_{-0.13}$	$63.1^{+3.4}_{-3.0}$	$0.69^{+0.05}_{-0.04}$	$3.1^{+0.4}_{-0.4}$	$3.6^{+0.4}_{-0.4}\times10^{56}$	440^{+150}_{-170}	$0.09^{+0.03}_{-0.03}$	182
GW151012	$23.2^{+14.9}_{-5.5}$	$13.6^{+4.1}_{-4.8}$	$15.2^{+2.1}_{-1.2}$	$0.05^{+0.31}_{-0.20}$	$35.6^{+10.8}_{-3.8}$	$0.67^{+0.13}_{-0.11}$	$1.6^{+0.6}_{-0.5}$	$3.2^{+0.8}_{-1.7}\times10^{56}$	1080^{+550}_{-490}	$0.21^{+0.09}_{-0.09}$	1523
GW151226	$13.7^{+8.8}_{-3.2}$	$7.7^{+2.2}_{-2.5}$	$8.9^{+0.3}_{-0.3}$	$0.18^{+0.20}_{-0.12}$	$20.5^{+6.4}_{-1.5}$	$0.74^{+0.07}_{-0.05}$	$1.0^{+0.1}_{-0.2}$	$3.4^{+0.7}_{-1.7}\times10^{56}$	450^{+180}_{-190}	$0.09^{+0.04}_{-0.04}$	1033
GW170104	$30.8^{+7.3}_{-5.6}$	$20.0^{+4.9}_{-4.6}$	$21.4^{+2.2}_{-1.8}$	$-0.04^{+0.17}_{-0.21}$	$48.9^{+5.1}_{-4.0}$	$0.66^{+0.08}_{-0.11}$	$2.2^{+0.5}_{-0.5}$	$3.3^{+0.6}_{-1.0}\times10^{56}$	990^{+440}_{-430}	$0.20^{+0.08}_{-0.08}$	921
GW170608	$11.0^{+5.5}_{-1.7}$	$7.6^{+1.4}_{-2.2}$	$7.9^{+0.2}_{-0.2}$	$0.03^{+0.19}_{-0.07}$	$17.8^{+3.4}_{-0.7}$	$0.69^{+0.04}_{-0.04}$	$0.9^{+0.0}_{-0.1}$	$3.5^{+0.4}_{-1.3}\times10^{56}$	320^{+120}_{-110}	$0.07^{+0.02}_{-0.02}$	392
GW170729	$50.2^{+16.2}_{-10.2}$	$34.0^{+9.1}_{-10.1}$	$35.4^{+6.5}_{-4.8}$	$0.37^{+0.21}_{-0.25}$	$79.5^{+14.7}_{-10.2}$	$0.81^{+0.07}_{-0.13}$	$4.8^{+1.7}_{-1.7}$	$4.2^{+0.9}_{-1.5}\times10^{56}$	2840^{+1400}_{-1360}	$0.49^{+0.19}_{-0.21}$	1041
GW170809	$35.0^{+8.3}_{-5.9}$	$23.8^{+5.1}_{-5.2}$	$24.9^{+2.1}_{-1.7}$	$0.08^{+0.17}_{-0.17}$	$56.3^{+5.2}_{-3.8}$	$0.70^{+0.08}_{-0.09}$	$2.7^{+0.6}_{-0.6}$	$3.5^{+0.6}_{-0.9}\times10^{56}$	1030^{+320}_{-390}	$0.20^{+0.05}_{-0.07}$	308
GW170814	$30.6^{+5.6}_{-3.0}$	$25.2^{+2.8}_{-4.0}$	$24.1^{+1.4}_{-1.1}$	$0.07^{+0.12}_{-0.12}$	$53.2^{+3.2}_{-2.4}$	$0.72^{+0.07}_{-0.05}$	$2.7^{+0.4}_{-0.3}$	$3.7^{+0.4}_{-0.5}\times10^{56}$	600^{+150}_{-220}	$0.12^{+0.03}_{-0.04}$	87
GW170817	$1.46^{+0.12}_{-0.10}$	$1.27^{+0.09}_{-0.09}$	$1.186^{+0.001}_{-0.001}$	$0.00^{+0.02}_{-0.01}$	≤ 2.8	≤ 0.89	≥ 0.04	$\geq 0.1\times10^{56}$	40^{+7}_{-15}	$0.01^{+0.00}_{-0.00}$	16
GW170818	$35.4^{+7.5}_{-4.7}$	$26.7^{+4.3}_{-5.2}$	$26.5^{+2.1}_{-1.7}$	$-0.09^{+0.18}_{-0.21}$	$59.4^{+4.9}_{-3.8}$	$0.67^{+0.07}_{-0.08}$	$2.7^{+0.5}_{-0.5}$	$3.4^{+0.5}_{-0.7}\times10^{56}$	1060^{+420}_{-380}	$0.21^{+0.07}_{-0.07}$	39
GW170823	$39.5^{+11.2}_{-6.7}$	$29.0^{+6.7}_{-7.8}$	$29.2^{+4.6}_{-3.6}$	$0.09^{+0.22}_{-0.26}$	$65.4^{+10.1}_{-7.4}$	$0.72^{+0.09}_{-0.12}$	$3.3^{+1.0}_{-0.9}$	$3.6^{+0.7}_{-1.1}\times10^{56}$	1940^{+970}_{-900}	$0.35^{+0.15}_{-0.15}$	1666

Figure 27: Parameter estimation results summary from the first Gravitational Wave Transient Catalogue published by the LIGO/Virgo collaboration (*Phys. Rev.* X **9** 031040 (2019)). Results are presented as the median and 90% symmetric credible interval of the Bayesian posterior distribution.

- Veitch, J., et al., *Parameter Estimation for Compact Binaries with Ground-Based Gravitational-Wave Observations Using the LALInference Software Library*, *Phys. Rev.* D **91**, 042003 (2015).

and the version used in the analysis of the O2 events can be downloaded from

- https://git.ligo.org/lscsoft/lalsuite/tree/lalinference_o2 .

From O3 onwards, an additional parameter estimation code, *Bilby*, was used to obtain posterior distributions for LIGO/Virgo detections. This code uses generic freely available Bayesian sampling codes to draw samples from the posterior distribution, such as DYNESTY and PTMCMC. The rest of the code consists of wrappers and functions to compute the correct likelihood to feed to the sampling codes. The description of the software can be found in

- Ashton, G., et al. (2019), *Astrophys. J. Supp.* **241**, 27

and the software can be downloaded from

- https://git.ligo.org/lscsoft/bilby

As well as providing tables summarising the median and symmetric credible intervals for the observed sources, LIGO papers typically include plots of the full Bayesian posterior distributions. These take various forms. Two-dimensional joint posterior distributions are often given for pairs of parameters that are correlated, such as the chirp mass and mass ratio or the final mass and spin of the remnant black hole produced by the merger or the sky location of the merger event. Examples of two-dimensional posterior distributions are shown in Figure 28 and Figure 29. One dimensional posteriors are often plotted as "violin plots" to allow comparison between the results for multiple events. The violin plot plots the parameter value on the y-axis and the posterior density on the x-axis, which is opposite to the usual convention. Additionally, the posterior is reflected in the y-axis so that it is symmetric about that axis for each event. The width of the resulting violin plot is proportional to the

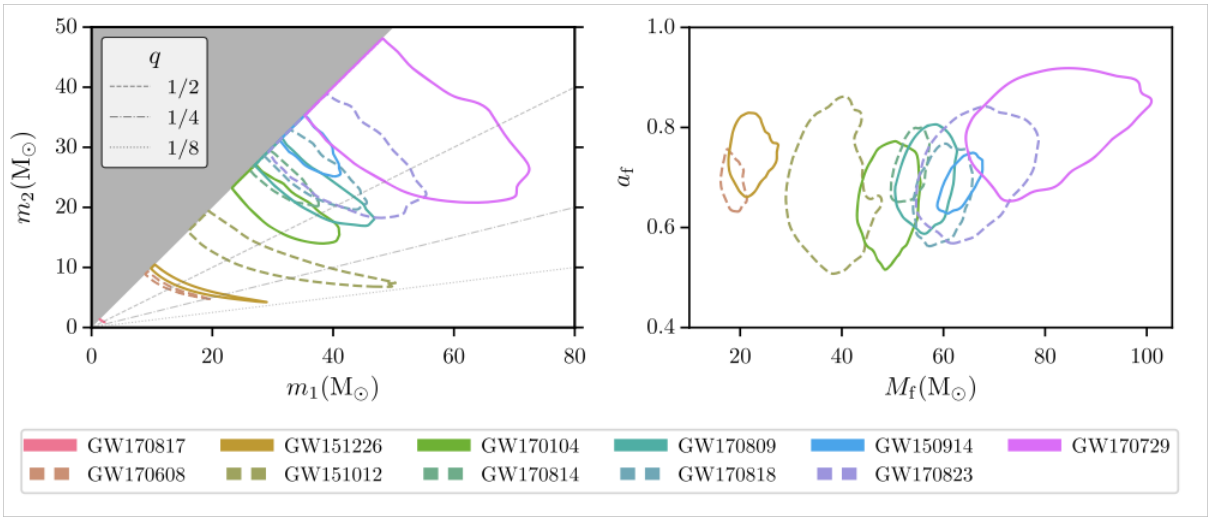


Figure 28: Joint two dimensional posterior on mass and mass ratio (left) and on final mass and spin (right) for all of the events observed by LIGO/Virgo during the O1 and O2 observing runs. Reproduced from Abbott et al. (2019), *Phys. Rev.* X **9** 031040.

posterior probability for the corresponding value of the parameter. An example is shown in Figure 30. Posteriors in the spins of the black holes, which is fundamentally a three-dimensional quantity, are typically represented by semi-circular density plots such as those shown in Figure 31. The full 3D posterior is marginalised over the (poorly constrained) azimuthal direction of the spin, and the resulting 2D posterior is represented on a semi-circle with the spin-magnitude as the radial direction and the angle between the spin vector and the orbital angular momentum as the angular direction. The density of the colour in these plots is proportional to the posterior density for the corresponding spin vector.

LALInference/*Bilby* are also used to obtain posterior deviations on parameters characterising deviations from general relativity, to facilitate tests of GR. More details can be found, along with results from analysis of the O1 and O2 events, in Abbott, B.P., et al., *Phys. Rev.* D **100**, 104036 (2019). Results for the analysis of O3a events can be found in Abbott, B.P., et al., *Phys. Rev.* D **103**, 122002 (2021).

7.2 Reduced order modelling*

LIGO parameter estimation codes are computationally expensive, primarily due to the cost of evaluating models of the gravitational waveforms to compute likelihoods. To make inference more efficient, it is advantageous to have models of the signals that are quicker to evaluate. This has been achieved by building **reduced order models** and **surrogate models**. The principle of both approaches is quite similar. First, a basis for the space of waveforms is found that has lower dimensionality than the number of samples in the original waveforms. Then either a fast interpolant is constructed to map physical parameters to the weights of the basis functions (in the case of some surrogate models, the interpolant is built directly for the waveform itself) or a **reduced order quadrature** representation of the likelihood is constructed.

In the latter approach, a projection of the target waveform onto the reduced basis is obtained not by using overlaps to find the best projection, but instead by requiring the

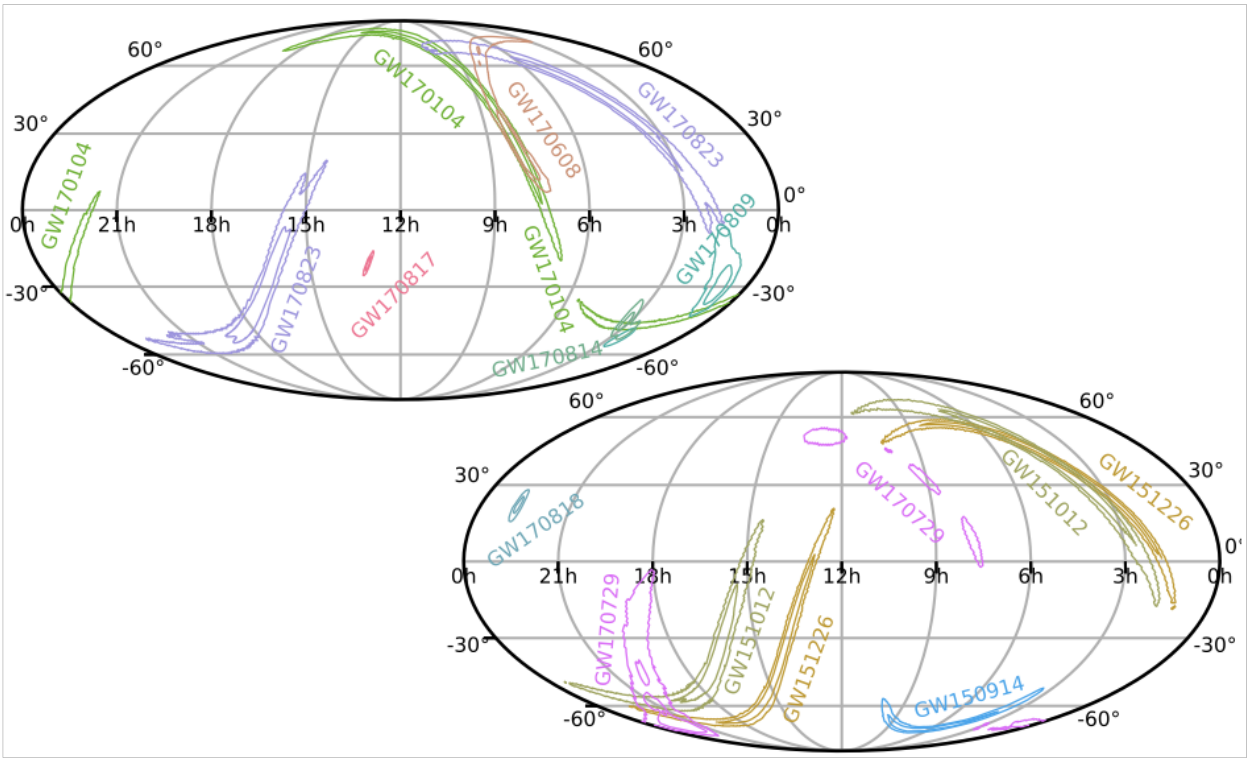


Figure 29: Sky location posterior distribution for all events observed by LIGO/Virgo during the O1 and O2 observing runs. Reproduced from Abbott et al. (2019), *Phys. Rev.* X **9** 031040.

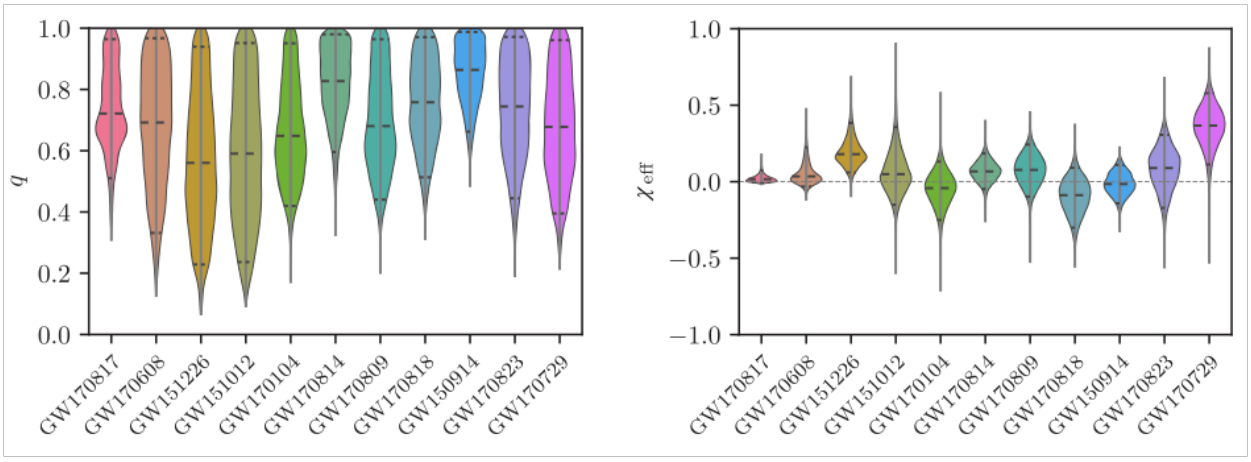


Figure 30: One-dimensional marginalised posteriors on the mass ratio (left) and effective spin (right) for all the events observed by LIGO/Virgo during the O1 and O2 observing runs. The one-dimensional posteriors are represented as "violin plots" as described in the text. Reproduced from Abbott et al. (2019), *Phys. Rev.* X **9** 031040.

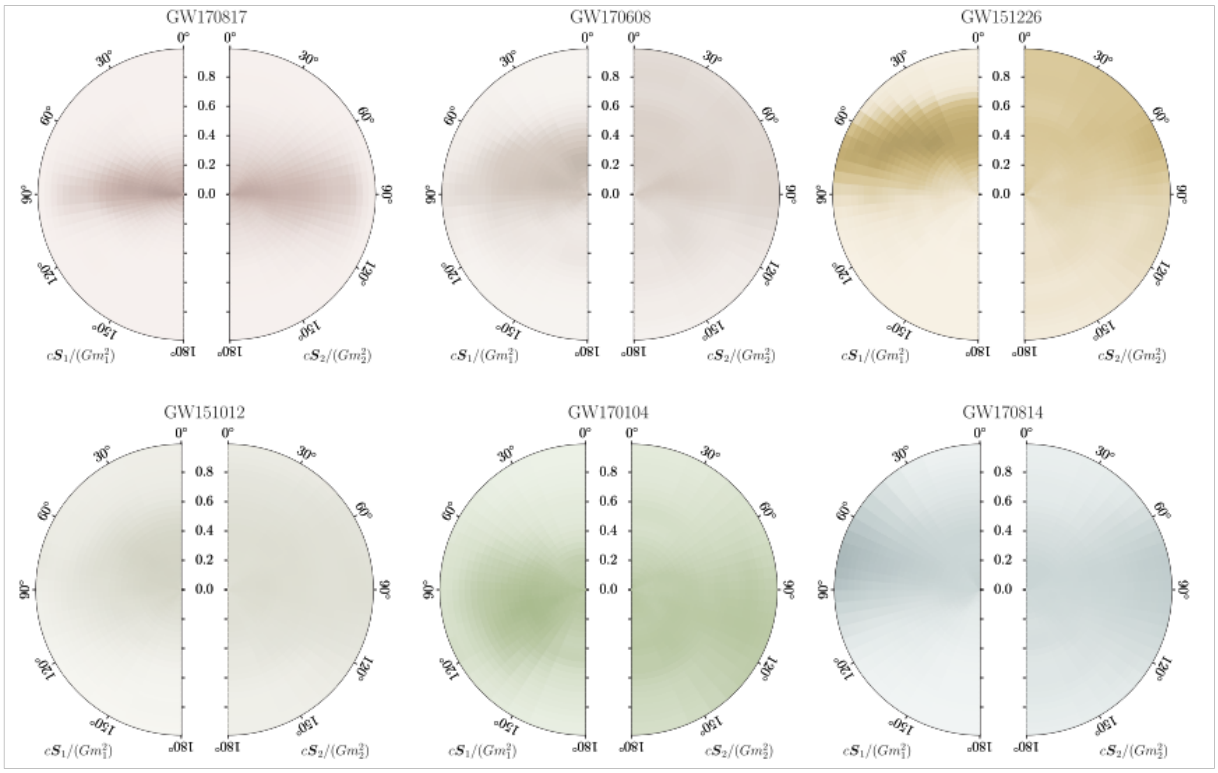


Figure 31: Posteriors on the spins of the two components in the binary for all of the events observed by LIGO/Virgo during the O1 and O2 observing runs. The distance from the origin represents the magnitude of the spin, and the angle represents the direction of the spin. The two halves of the plot are for the primary (left) and secondary (right) object in the binary. The density of colour is proportional to the posterior density for that spin value. Reproduced from Abbott et al. (2019), *Phys. Rev.* X **9** 031040.

target waveform to exactly match a linear combination of basis waveforms at a number of points, called **quadrature interpolation points**, equal to the number of functions in the basis. This allows the likelihood quadrature to be reduced to a sum over the target waveform evaluated at the quadrature points weighted by data-dependent constants that can be computed prior to running inference from overlaps of the basis functions with the data. The full procedure is as follows

- Find a (reduced) set of basis functions, $\{e_i(f)\}$, of size m, that can represent all waveforms in the training data set to a certain pre-specified precision. This is done using a greedy algorithm, sequentially selecting the least well represented waveform to add to the current reduced basis set, until the desired representation accuracy is reached.

- Identify m quadrature interpolation points, at which the reduced basis representation will be forced to match the target waveform. This is again done using a greedy algorithm, choosing at each stage to put an interpolation point at the point where the difference between the next basis function and the current interpolated representation is largest.

- Define the matrix $A_{ij} = e_j(F_i)$. For any given choice of the waveform parameters, λ, define the vector $\vec{h}(\lambda)$ by $h_i = h(F_i|\lambda)$. The interpolated representation of $h(f|\lambda)$ is $A^{-1}_{ij} h_j(\lambda) e_i(f)$.

- The overlap of the waveform $h(\lambda)$ with the data can then be represented via

$$\begin{aligned}\left(h(\vec{\lambda})|d\right) &= 4\Re \int_0^\infty \frac{\tilde{h}(\vec{\lambda})\tilde{d}^*(f)}{S_h(f)}\,\mathrm{d}f \\ &\approx 4\Re \left[\sum_{k=0}^{N/2} d^*(f_k) e_i(f_k) \Delta f \mathbf{A}^{-1}_{ij}\right] h_j(\vec{\lambda}) \\ &= 4\Re \sum_{k=1}^{m} \omega_k h_k(\vec{\lambda}). = 4\Re \sum_{k=1}^{m} \omega_k h(F_k;\vec{\lambda}). \end{aligned} \tag{104}$$

  where the weights

$$\omega_k = \left[\sum_{k=0}^{N/2} d^*(f_k) e_i(f_k) \Delta f \mathbf{A}^{-1}_{ik}\right]$$

  are independent of the parameters and can therefore be pre-computed prior to inference.

Evaluating this reduced order quadrature likelihood now requires only summing m terms, rather than the N required for the full likelihood, so it represents a considerable saving when $m \ll N$. Reduced order quadrature approximations to likelihoods are the state of the art in LIGO parameter estimation, but they require being able to evaluate the target waveform at certain frequencies quickly and so can only be directly used with frequency-domain waveform approximants. This problem is overcome in surrogate models by constructing an additional interpolant across parameter space at each of the quadrature interpolation points.

For further information on reduced basis and surrogate models, please consult the following papers and references therein

- Field, S., et al., *Reduced basis catalogs for gravitational wave templates*, *Phys. Rev. Lett.* **106** 221102 (2011).
- Canizares, P., et al., *Gravitational wave parameter estimation with compressed likelihood evaluations*, *Phys. Rev.* D **87** 124005 (2013).
- Field, S., et al., *Fast prediction and evaluation of gravitational waveforms using surrogate models*, *Phys. Rev.* X **4** 031006 (2014).
- Canizares, P., et al., *Accelerated gravitational-wave parameter estimation with reduced order modeling*, *Phys. Rev. Lett.* **114** 071104 (2015).
- Blackman, J., et al., *Fast and accurate prediction of numerical relativity waveforms from binary black hole coalescences using surrogate models*, *Phys. Rev. Lett.* **115** 121102 (2015).
- Varma, V., et al., *Surrogate models for precessing binary black hole simulations with unequal masses*, *Phys. Rev. Research* **1** 033015 (2019).

7.3 Population inference

Inference on the properties of the population of sources form which the observed LIGO events are drawn also uses Bayesian methods, specifically Bayesian hierarchical modelling. We encountered one example of this in Section 5.9, which is the inference of cosmological parameters using gravitational wave observations of binary neutron star mergers with counterparts. Other examples include inference on the rate of mergers of different types of source in the Universe, and on the distributions of masses and spins of black holes and neutron stars. Full details on the range of population analyses carried out for the O1 and O2 events can be found in Abbott, B.P., et al., *Astrophys. J. Lett.* **882**, L24 (2019) and for O3 events in Abbott, B.P., et al., arxiv:2111.03634 (2021), but we summarise some of the key analyses here.

7.3.1 Rate estimation

Accurate estimation of the rate of events in the Universe is complicated by confusion with detector noise, i.e., identifying which events are real gravitational wave events and which are instrumental artefacts, and by the need to make assumptions about the distribution of parameters of sources in the population. The first problem was tackled in Farr, W., Gair, J.R., Mandel, I., and Cutler, C., *Phys. Rev.* D **91**, 023005 (2014). If the output of the detector is represented by a sequence of values of a detection statistic, x, and any statistic value that exceeds some threshold, $x_{\rm min}$, is regarded as a detection, then the observed data is a set of detection statistic values above threshold, $\{x_i\}$. Some of these events correspond to real foreground events, while others arise due to noise fluctuations in the detector and are background. We introduce an (unobserved) parameter f_i for each event such that $f_i = 1$ is it is a foreground event and f_i=0 if it is background. The foreground and background events are assumed to be generated by independent Poisson processed with rates

$$\frac{{\rm d}N_f}{{\rm d}x} = R_f \hat{f}(x, \theta_f), \qquad \frac{{\rm d}N_b}{{\rm d}x} = R_b \hat{b}(x, \theta_b)$$

and corresponding cumulative distributions $\hat{F}(x,\theta_f)$, $\hat{B}(x,\theta_b)$. Here R_f and R_b are the foreground and background rates respectively and θ_f and θ_b represent any unknown parameters that characterise the foreground and background distributions. The combined posterior for the rates, event flags and distribution parameters is

$$p(f_i, R_f, R_b, \theta | d_{\rm to}, N) = \frac{\alpha}{p(d_{\rm to}, N) N!} \left[\prod_{i|f_i=1} R_f \hat{f}(x_i, \theta) \right] \left[\prod_{i|f_i=0} R_b \hat{b}(x_i, \theta) \right] \exp[-(R_f + R_b)] \frac{p(\theta)}{\sqrt{R_f R_b}}$$

where $p(\theta)$ is the prior on the posterior parameters and we are using a Jeffreys' prior $p(R) \propto 1/\sqrt{R}$ on the rates. The subscript on $d_{\rm to}$ indicates that we are using time-ordered data. The data could also be analysed ordered by ranking statistic. This posterior can be marginalised over the unknown flags to give posteriors on the rates, or over the rates to give posterior probabilities for $f_i = 1$ for each event.

One complication with this approach is that it relies on a model for the foreground and background distributions. These can be estimated by injections and time-slides, but, since LIGO is not equally sensitive to all types of CBC event, the former requires imposing some model of the astrophysical population from which the events are drawn. One approach to this is to assume that all events in the Universe are the same as the one that has been observed. This approach was used in Kim, Kalogera and Lorimer (*Astrophys. J.* **584**, 985 (2003)) to estimate the rate of double neutron star mergers and so is often referred to as the "KKL method". In the first LIGO detection paper, for GW150914, the combination of the rate estimation accounting for confusion (FGMC) and the KKL method was used to infer the rate of binary black hole mergers. The application of this "alphabet soup" method was complicated by the fact that the data being analysed to infer the background for GW150914 contained a second CBC trigger, LVT151012. The parameters of this event were completely different to GW150914, so the KKL method could still be applied, but generalising to the case where all events in the Universe were either like GW150914 or LVT151012. Further details can be found in Abbott, B.P., et al. *Astrophys. J. Lett.* **833**, 1 (2016) and Abbott, B.P., et al. *Astrophys. J. Supp.* **227**, 14 (2016).

One additional trigger, GW151226, was present in the LIGO O1 data, and that again had sufficiently distinct parameters that the KKL approach could be used. In O2, the events began to have much more posterior overlap and so this method could no longer be used. Now, a model of the population is assumed in event rate estimation. O2 analyses used both a power-law mass distribution or a flat in log-mass distribution in an attempt to bound the range of possible rate. Results in O3 were obtained by simultaneously fitting for the rate and the population model, using the models for the mass distribution described in the next section.

7.3.2 Black hole mass distribution

The mass distribution of stellar-origin black holes in binaries can be inferred from LIGO/Virgo observations in a hierarchical analysis by placing a prior on the mass of individual events that depends on some unknown parameters that can be constrained from analysing the full set of events. The GWTC-1 analysis using O1 and O2 events is described in Abbott, B.P., et al., *Astrophys. J. Lett.* **882**, L24 (2019), while the analysis of GWTC-3 using all events observed up until the end of O3, is described in Abbott, B.P., et al., arxiv:2111.03634. These analyses used a number of different models to describe the mass distribution. The GWTC-1

analysis employed two different parametric models, the **truncated mass model** and the **power law + peak** model.

Truncated mass model This model assumes a power law distribution on mass and mass ratio, with low and high mass cut-offs

$$p(m_1, m_2|m_{\rm min}, m_{\rm max}, \alpha, \beta_q) \propto \begin{cases} C(m_1) m_1^{-\alpha} q^{\beta_q} & \text{if } m_{\rm min} \leq m_2 \leq m_1 \leq m_{\rm max} \\ 0 & \text{otherwise} \end{cases} .$$

The GWTC-1 analysis considered two different versions of this model. In the first, the parameters $m_{\rm min} = 5M_\odot$, $\beta_q = 0$ were fixed, leaving only$m_{\rm max}$ and α free to vary. In the second variant, all four parameters were allowed to vary.

Power law + peak This model mixes a power-law component of the above form, with a Gaussian component, which was designed to fit any excess of events near the lower mass limit of the pair-instability supernova mass gap. The model is

$$p(m_1|\theta) = \left[(1-\lambda_m) A(\theta) m_1^{-\alpha} \Theta(m_{\rm max} - m_1) + \lambda_m B(\theta) \exp\left(-\frac{(m_1-\mu_m)^2}{2\sigma_m^2}\right)\right] S(m_1|m_{\rm min}, \delta m)$$
$$p(q = m_2/m_1|m_1, \theta) = C(m_1, \theta) q^{\beta_q} S(m_2|m_{\rm min}, \delta m). \tag{105}$$

Here, $A(\theta)$ and $B(\theta)$ are computed to normalize the truncated-power-law and Gaussian components of the model, respectively. The function $S(m_1|m_{\rm min}, \delta m)$ is a smoothing function that cuts off the distribution at lower masses, which is defined by

$$S(m|m_{\rm min}, \delta m) = \begin{cases} 0 & \text{if } m < m_{\rm min} \\ \left[1 + \exp\left(\frac{\delta_m}{m-m_{\rm min}} + \frac{\delta_m}{m-m_{\rm min}-\delta_m}\right)\right]^{-1} & \text{if } m_{\rm min} < m < m_{\rm min} + \delta_m \\ 1 & \text{if } m > m_{\rm min} + \delta_m \end{cases} \tag{106}$$

The mass distributions obtained by fitting these models to the O1 and O2 data are shown in Figure 32.

With the larger numbers of events observed in O3, it was possible to fit more complex population models. In particular, it was found that the truncated model was insufficient to describe the observed population, although the power law + peak model was still a reasonable approximation to the distribution of masses observed in the binary black hole systems. The LVK analysis of the GWTC-3 population introduced one additional parametric model, the **power law + dip + peak** model, and also performed three different non-parametric fits.

Power law + dip + peak This model was introduced to allow for the fitting of a single distribution that covered all compact objects, i.e., both neutron stars and black holes. The model takes the form

$$p(m|\theta) = n(m|M_{\rm low}^{\rm gap}, M_{\rm high}^{\rm gap}, A) \times l(m|m_{\rm max}, \eta) \times \begin{cases} m^{\alpha_1} & \text{if } m < M_{\rm high}^{\rm gap} \\ m^{\alpha_2} & \text{if } m > M_{\rm high}^{\rm gap} \\ 0 & \text{if } m > m_{\rm max} \text{ or } m < m_{\rm min} \end{cases} , \tag{107}$$

where $l(m|m_{\rm max}, \eta)$ is a low-pass filter with power-law fall-off η, applied at $m_{\rm max}$ and the function $n(m|M_{\rm low}^{\rm gap}, M_{\rm high}^{\rm gap}, A)$ is a notch filter, of depth A and applied between $M_{\rm low}^{\rm gap}$ and

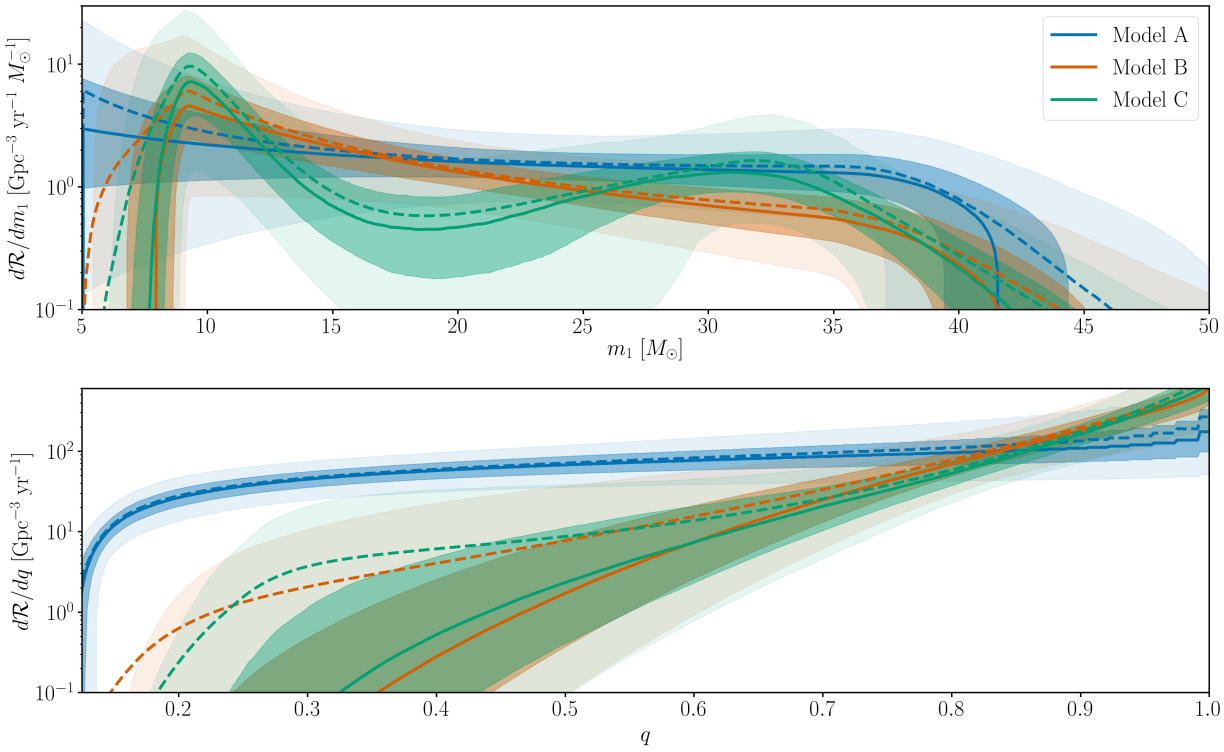


Figure 32: Black hole mass function inferred from LIGO/Virgo events observed in the O1 and O2 observing runs. Figure reproduced from Abbott, B.P., et al., *Astrophys. J. Lett.* **882**, L24 (2019).

$M_{\rm high}^{\rm gap}$. The purpose of the notch is to suppress objects in the mass range between neutron stars and black holes, as these should be distinct populations. The purpose of the low-pass filter is to make the distribution smoothly go to zero at the lower edge of the pair-instability mass gap and plays a similar role to the truncation in the **truncated** model.

This mass function is used to describe both components in the binary, and is combined with a pairing function that describes correlations between m_1 and m_2. The GWTC-3 analysis considered two different pairing models: (i) random pairing, in which m_1 and m_2 are independent draws from $p(m|\theta)$, constrained so that $m_2 < m_1$; and (ii) power-law-in-mass-ratio pairing, in which $p(m_1, m_2|\theta, \beta) \propto p(m_1|\theta)p(m_2|\theta)q^\beta$ if $m_2 < m_1$ or 0 otherwise.

Non-parametric models The LVK population analysis of GWTC-3 also used three non-parametric models to describe the astrophysical distribution of black hole masses. Non-parametric models typically have arbitrary numbers of degrees of freedom, giving them greater flexibility to fit observed data. In practice, a non-parametric model is fitted by gradually increasing the number of parameters until the additional model complexity is no longer supported by sufficient improvement in the fit to the data. The LVK analysis considered three different non-parametric approaches

- **Power law + spline** In this model the mass distribution is represented as a perturbed truncated power-law. The model is

$$p(m_1|\alpha, m_{\rm min}, m_{\rm max}, \delta_m, \{f_i\}) = k\, p(m_1|\alpha, m_{\rm min}, m_{\rm max}, \delta_m) \exp[f(m_1|\{f_i\})] \tag{108}$$

  where k is a normalising constant, $p(m_1|\alpha, m_{\rm min}, m_{\rm max}, \delta_m)$ is the power-law + peak model with $\lambda_m = 0$ so that the Gaussian peak is removed from the model and $f(m_1|\{f_i\})$ is a cubic spline with weights $\{f_i\}$ at n knots equally spaced in $\log(m_1)$ in the range 2—100$M_\odot$. The non-parametric part of this model is the flexibility in the choice of n. The optimal choice for n is chosen by comparing evidences for different choices of n and, in the GWTC-3 analysis, 20 knots was found to be optimal.

- **Flexible mixtures** This model aims to represent both the mass and spin distribution simultaneously, by representing the joint distribution as a sum of separable components. The model takes the form

$$p(\mathcal{M}, q, s_{1z}, s_{2z}|\theta) = \sum_{i=1}^{N} w_i G(\mathcal{M}|\mu_i^{\mathcal{M}}, \sigma_i^{\mathcal{M}})\, G(s_{1z}|\mu_i^{sz}, \sigma_i^{sz})\, G(s_{2z}|\mu_i^{sz}, \sigma_i^{sz})\, P(q|\alpha_i^q, q_i^{\rm min}, 1), \tag{109}$$

  where N is the number of components, $\mathcal{M}$ is the chirp mass and s_{iz} is the component of the spin of the i'th component aligned with the orbital angular momentum. The function $G(X|\mu, \sigma)$ is a Gaussian in the variable X, with mean μ and variance σ^2, while $P(X|\alpha, x_{\rm min}, x_{\rm max})$ denotes a power-law with slope α, truncated below $x_{\rm min}$ and above $x_{\rm max}$, The flexibility in this model again comes from varying the number of components, N. For a fixed N, the parameters of each of the distributions and the relative component weights, w_i, are fitted to the data. The optimal N is found from maximizing the evidence, and the optimal choice was found to be 11 in the analysis of GWTC-3.

- **Binned Gaussian process** In this model, the (m_1, m_2) parameter space is divided up into a set of bins, labelled i. The rate density of mergers in each bin, n_i, is represented

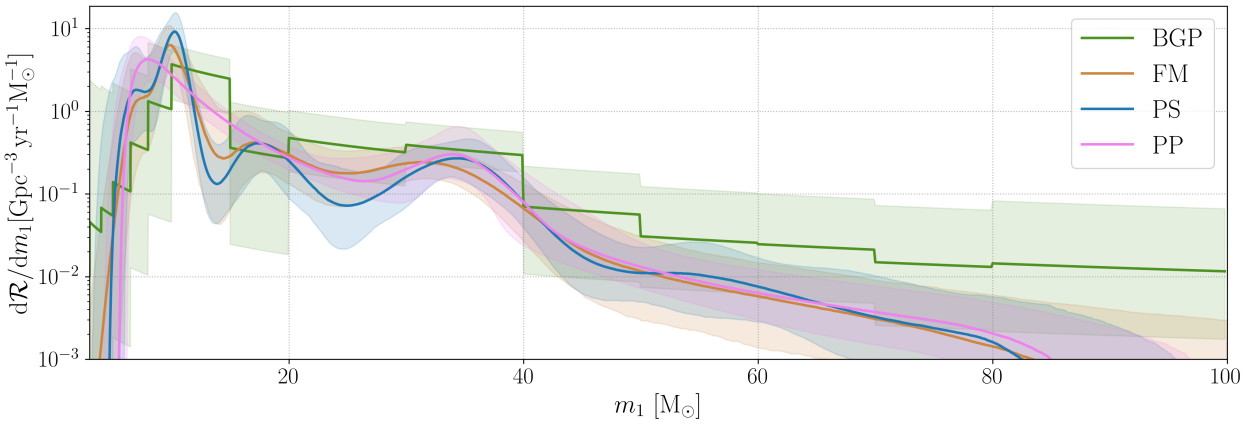


Figure 33: Black hole mass function inferred from LIGO/Virgo events in GWTC-3. Figure reproduced from Abbott, B.P., et al., arxiv:2111.03634 (2021). The various coloured curves show the result of fitting the power-law + peak (PP), power-law + spline (PS), flexible mixtures (FM) and binned Gaussian process (BGP) models to the data.

as a Gaussian process, characterized by a constant mean μ, and covariance function

$$\Sigma(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{|\mathbf{x} - \mathbf{x}'|^2}{2l^2}\right). \tag{110}$$

In the binned model, the vector $\mathbf{x}$ is taken to be the vector of mass values in the centre of the bin. The model is characterized by the three parameters μ, σ and l, but because the rates in each bin are a random draw from the Gaussian process, these rates can be regarded as additional model parameters. When fitting the model, the individual rates are inferred simultaneously with the parameters characterising the Gaussian process, resulting in a map of the rate density across the mass parameter space. For the GWTC-3 analysis, the mass distribution was fitted in this way while assuming a fixed evolution of the merger rate with redshift, and a fixed isotropic distribution of black hole spins.

Figure 33 shows the result of fitting these three non-parametric distributions to the observations reported in the GWTC-3 catalogue, compared to the result of fitting the power-law + peak model.

7.3.3 Black hole spin distribution

A hierarchical analysis of LIGO/Virgo events can also provide insight into the spin distribution. This can also be done either parametrically or non-parametrically. As spin magnitudes lie in the range $[0, 1]$ it is natural to model these with a Beta distribution, whose support is confined to that range. This was used to analyse the O1/O2 events in Abbott, B.P., et al., *Astrophys. J. Lett.* **882**, L24 (2019), and again as the default spin model to analyse O3 events in arxiv:2111.03634. The model is

$$p(\chi_i|\alpha_\chi, \beta_\chi) = \frac{\chi_i^{\alpha_\chi - 1}(1 - \chi_i)^{\beta_\chi - 1}}{B(\alpha_\chi, \beta_\chi)}.$$

A non-parametric analysis was also used for the analysis of GWTC-1, but not GWTC-3. This modelled the spin-magnitude distribution as a set of heights of a binned distribution,

with the bin heights free parameters to be determined by the observations. For example, a three-bin distribution (Farr, B., Holz, D., and Farr, W., *Astrophys. J.* **854**, L9 (2018))

$$p(\chi) = \begin{cases} A_1/3 & 0 \leq \chi \leq 1/3 \\ A_2/3 & 1/3 \leq \chi \leq 2/3 \\ 1-(A_1+A_2)/3 & 2/3 \leq \chi \leq 1 \end{cases}.$$

The posteriors obtained from applying these models to the O1 and O2 events are shown in Figure 34, and the posterior from applying the parametric model to GWTC-3 is shown in Figure 35.

LIGO observations measure the effective spin, $\chi_{\rm eff}$, better than individual spins. The recovered distribution has significant support for $\chi_{\rm eff} < 0$, which has significant implications for binary formation models, as negative effective spins are very hard to produce in standard isolated binary evolution and are therefore an indicator of dynamical formation channels. To assess the robustness of this result, a truncated-Gaussian model was fitted, with truncation in the range $[\chi_{\rm eff,min}, 1]$. This supported $\chi_{\rm eff,min} < 0$ at $> 99\%$ confidence. It was argued in Roulet et al. (*Phys. Rev.* D **104** 083010 (2021)) and Galaudage et al. (arxiv:2109.02424 (2021)) that this result might arise from the combination of a population with $\chi_{\rm eff} > 0$ and a smaller sub-population with vanishing spin, $\chi_{\rm eff} = 0$. To address this the LVK GWTC-3 analysis fitted a model of the form

$$p(\chi_{\rm eff}|\mu_{\rm eff}, \sigma_{\rm eff}, \chi_{\rm eff,min}) = \zeta_{\rm bulk}\mathcal{N}_{[\chi_{\rm eff,min},1]}(\chi_{\rm eff}|\mu_{\rm eff}, \sigma_{\rm eff}) + (1-\zeta_{\rm bulk})\mathcal{N}_{[-1,1]}(\chi_{\rm eff}|0, 0.01),$$

where $\mathcal{N}_{[a,b]}(x|\mu,\sigma)$ denotes a normal distribution with mean μ and variance σ^2, truncated to the range $[a,b]$. Fitting such a model reduced the probability that $\chi_{\rm eff,min} < 0$ to 88%.

To explore possible correlation structures in the spin distribution of the observed events, a model was also fitted to the joint distribution of $\chi_{\rm eff}$ and the precessional spin component, χ_p, of the form

$$p(\chi_{\rm eff}, \chi_p|\mu_{\rm eff}, \sigma_{\rm eff}, \mu_p, \sigma_p, \rho) \propto G\left(\chi_{\rm eff}, \chi_p \middle| \mu = \begin{pmatrix} \mu_{\rm eff} \\ \mu_p \end{pmatrix}, \mathbf{\Sigma} = \begin{pmatrix} \sigma_{\rm eff}^2 & \rho\sigma_{\rm eff}\sigma_p \\ \rho\sigma_{\rm eff}\sigma_p & \sigma_p^2 \end{pmatrix}\right)$$

where $G(x, y|\mu, \mathbf{\Sigma})$ is a two-dimensional Gaussian distribution for the parameters x and y, with mean vector μ and covariance matrix $\mathbf{\Sigma}$. The purpose of fitting a generic distribution like this is to try and quantify the covariance of the variables, which is indicated by non-zero values for the correlation parameter ρ.

The spin direction is also a parameter of interest astrophysically, as different formation scenarios predict either isotropically distributed spin directions, or a preference for spins to be aligned with the angular momentum of the binary. To capture this, we can use a mixture model

$$p(\cos t_1, \cos t_2|\sigma_t, \zeta) = \frac{(1-\zeta)}{4} + \frac{2\zeta}{\pi}\prod_{i\in\{1,2\}}\frac{\exp(-(1-\cos t_i)^2/2\sigma_t^2)}{\sigma_t \mathrm{erf}(\sqrt{2}/\sigma_t)}.$$

For the LVC analysis of the GWTC-1 catalogue the Gaussian component was modified so that the t_1 and t_2 components had independent variances, but in the analysis of GWTC-3, these were fixed to be equal. LIGO measurements in O1 and O2 were not sufficiently informative about spins to strongly constrain the parameters of the model, but later catalogues have shown increasing support for an isotropic distribution (see Figure 36).

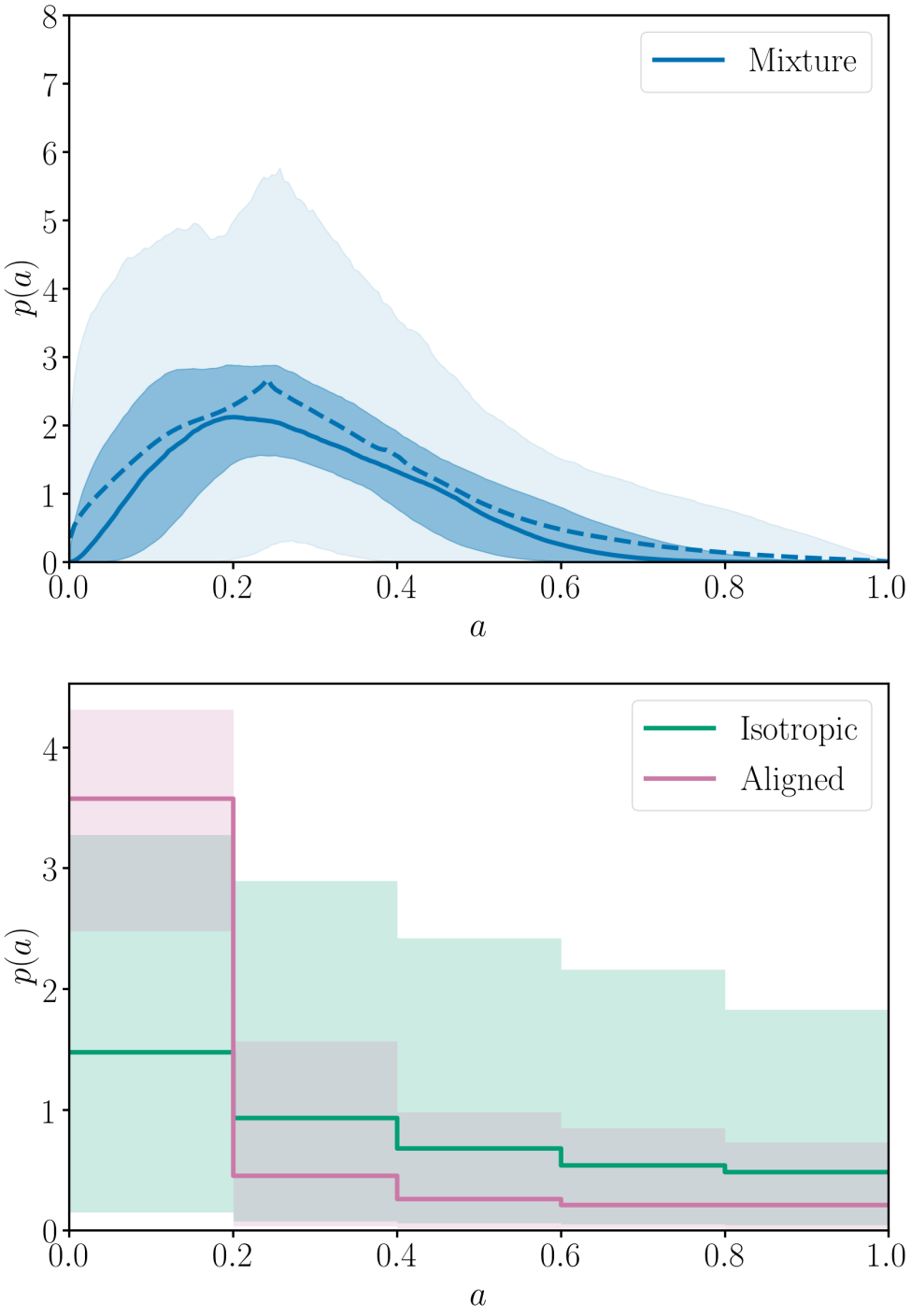


Figure 34: Black hole spin distribution inferred from LIGO/Virgo events observed in the O1 and O2 observing runs, using a parametric (top panel) or non-parametric (bottom panel) approach. Figures reproduced from Abbott, B.P., et al., *Astrophys. J. Lett.* **882**, L24 (2019).

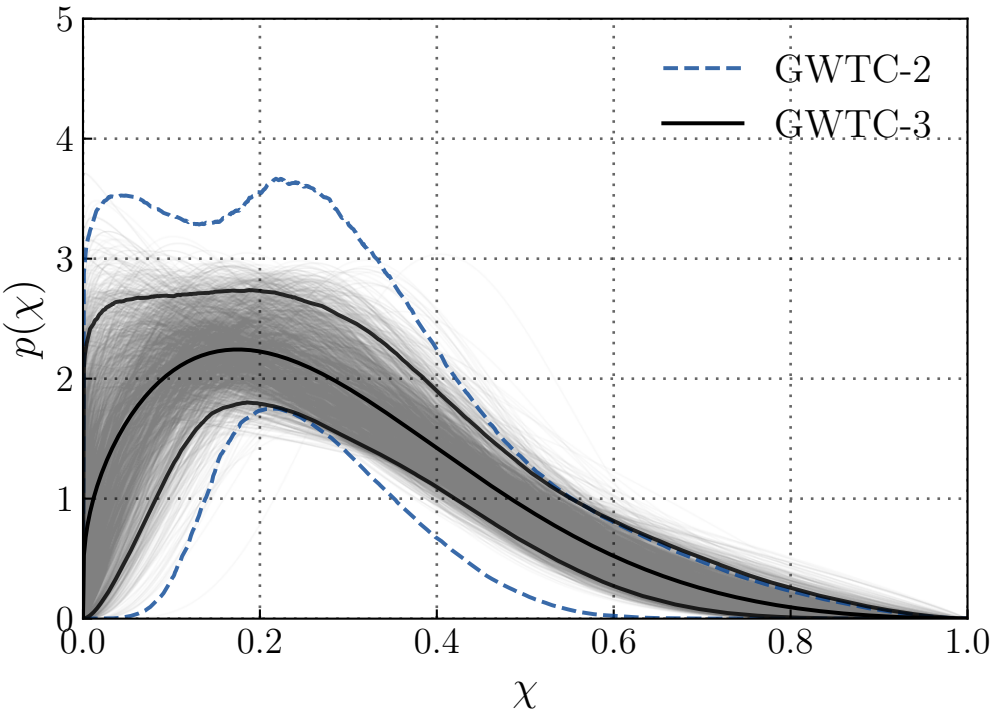


Figure 35: Black hole spin distribution inferred from LIGO/Virgo events observed up to the end of the O3 observing run, using the same parametric approach as in the upper panel of Figure 34. Figure reproduced from Abbott, B.P., et al., arxiv:2111.03634 (2021).

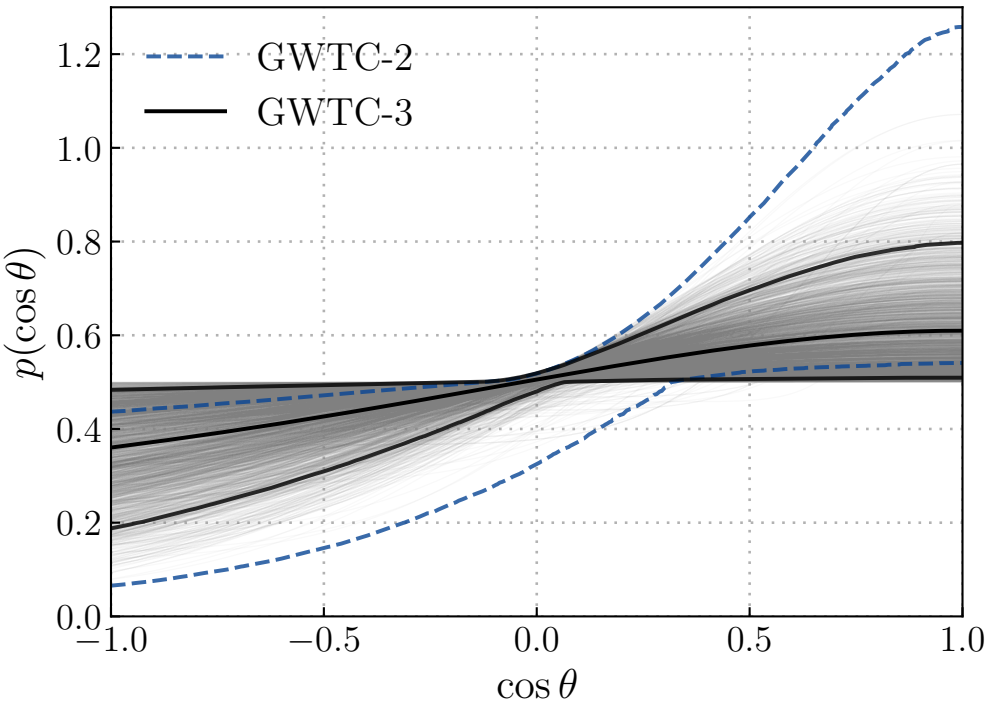


Figure 36: Black hole spin-tilt distribution inferred from LIGO/Virgo events observed up until the end of the O3 observing run. Figure reproduced from Abbott, B.P., et al., arxiv:2111.03634 (2021).

7.3.4 Mixed mass-spin distributions

As masses and spins are both indicators of the evolution path of a particular binary, it might be expected that there are correlations between mass and spin properties that reveal information about the underlying astrophysics. The **flexible mixtures** model can represent correlations, since while each component is separable, the sum is not. However, this model is not well adapted to extracting any correlations that are present and this is more easily achieved using suitably designed parametric models. In the GWTC-3 analysis, the LVK explored one particular possible correlation, between mass ratio, q, and the effective spin, $\chi_{\rm eff}$. The specific model used took the form

$$\begin{aligned} p(\chi_{\rm eff}|q) &\propto \exp\left[-\frac{(\chi_{\rm eff}-\mu(q))^2}{2\,\sigma^2(q)}\right] \\ \text{where } \mu(q) &= \mu_0 + \alpha(q-1) \\ \log_{10}\sigma(q) &= \log_{10}\sigma_0 + \beta(q-1). \end{aligned} \tag{111}$$

The parameters α and β represent simple linear evolution with mass ratio of the mean and variance of the effective spin distribution. The GWTC-3 data favoured $\alpha < 0$ with 98% credibility, i.e., that black holes with more equal mass ratios tend to have smaller spins. The GWTC-3 results are shown in Figure 37.

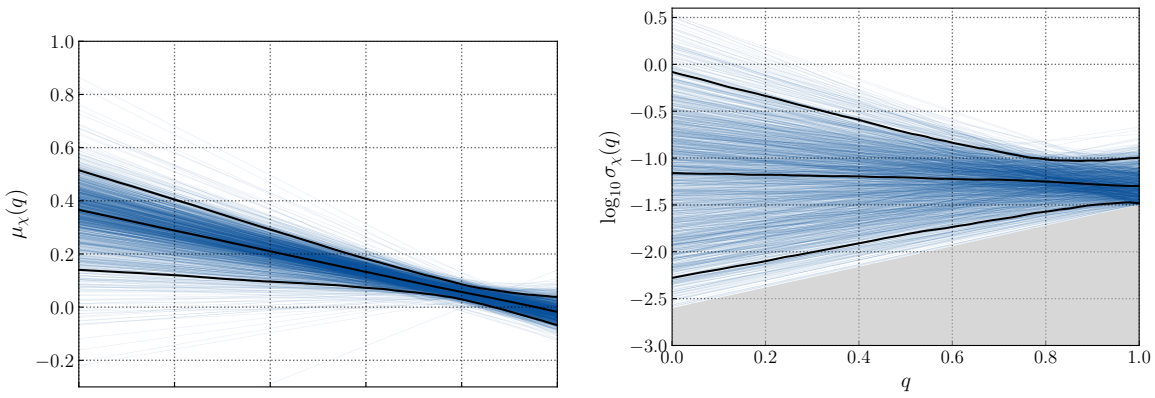


Figure 37: Variation in the mean (left) and log-standard deviation (right) of the effective spin distribution with mass ratio, q. Figure reproduced from Abbott, B.P., et al., arxiv:2111.03634 (2021).

7.3.5 Rate evolution

The FGMC+KKL method described earlier assumes that the rate of mergers is constant, but in principle this could evolve over cosmic history (the FGMC framework can handle this, but the interpretation of R_f is different, as the average rate over the sensitive volume of the detector). An evolution of the rate can be explicitly included and constrained by introducing an extra parameter into the rate density

$$\frac{\mathrm{d}R}{\mathrm{d}\xi}(z|\theta) = R_0 p(\xi|\theta)(1+z)^\lambda.$$

The analysis of the O1 and O2 events provided weak evidence for an evolution in rate with redshift, but this was mostly due to the event GW170729, which was the most marginal detection. Analysis of GWTC-3 showed stronger evidence for evolution of the rate, concluding that $\lambda > 0$ with 99.4% confidence, i.e., there is strong evidence that the rate of mergers was higher in the past.

7.4 Model selection

Bayesian methods are also applied to model selection using the LIGO/Virgo observations, through the evaluation of **evidence ratios** or **Bayes factors** for pairs of alternative hypotheses for the data. Some examples of applications to gravitational wave data are

- Test for the presence of a signal in the data after the end of the merger of the two neutron stars in GW170817. Such a signal might be evidence that the merger project was a hypermassive neutron star rather than a black hole. For GW170817 the Bayes factor for the noise model over the signal model was 256.79 (Abbott, B.P., et al., *Phys. Rev.* X **9** 011001 (2019)), providing strong evidence that no such signal was present.

- Test of the polarisation state of gravitational waves. Possible models are that the gravitational waves have tensor polarisation, as expected in GR, or have scalar polarisation or vector polarisation. The analysis of GW170818 gave Bayes factors of 12 for tensor versus vector polarisation and 407 for tensor versus scalar, while the analysis of

GW170814 gave Bayes' factors of 30 and 220 respectively (Abbott, B.P., et al., *Phys. Rev.* D **100** 104036 (2019)).

- Tests of the no-hair property of the remnant black hole formed in a merger, by comparing the properties of the observed ringdown radiation to that predicted by GR (Brito, Buonanno and Raymond, *Phys. Rev.* D **98**, 084038 (2018)).

- Probing alternative theories of gravity. For example, looking for evidence for dynamical gravity with the polarisation of continuous gravitational waves (Isi et al., *Phys. Rev.* D **96**, 042001 (2017)).

7.5 Source reconstruction

Although Bayesian inference relies on the existence of models, it is also possible to use these methods to recover "unmodelled" sources. One such implementation is the BayesWave algorithm. The method works by modelling the noise and signals in the data from the various detectors as a superposition of simple components. BayesWave represents the noise as a combination of a smooth PSD component, described by a cubic spline, lines represented by Lorentzians and glitches modelled by wavelets. Signals in the data are also modelled by wavelets, but with parameters that are common across the detectors, as opposed to the noise components which are independent in different detectors. Wavelets are simple functions that are compact in both time and frequency. We will encounter these again in the non-parametric regression section of this course. There are many different wavelet families, but the wavelets used in BayesWave are known as the Morley-Gabor basis.

BayesWave fits its model using reversible jump MCMC. The reversible jump element is required to add or remove wavelet or line components, as the number of these required is not known a priori. Further details on the BayesWave algorithm can be found in

- Cornish, N.J., and Littenberg, T.B., *Class. Quantum Grav.* **32**, 135012 (2015).

- Littenberg, T.B., and Cornish, N.J., *Phys. Rev.* D **91**, 084034 (2015).

BayeWave is used in LIGO analyses for PSD estimation, glitch removal and for non-parametric waveform reconstruction. The good agreement between the BayesWave reconstructed waveform and the best fit model found by parameter estimation for GW150914 (see Figure 38) provided extra support to the fact that this was a true signal.

7.6 Rapid localisation

Since the start of the O1 observing run, LIGO/Virgo have been sending out triggers to facilitate follow-up of gravitational wave events by electromagnetic telescopes. To avoid delays to these alerts, it is necessary to rapidly estimate the sky location of the triggers so that astronomers know where to point their telescopes. Bayesian techniques are also used for this purpose. Full Bayesian parameter estimation is not possible in low-latency, so the rapid localisation algorithms are not truly Bayesian, but make approximations in evaluating the posterior that allow it to be computed quickly.

The Bayestar algorithm replaces the full likelihood by the autocorrelation likelihood, which is the likelihood evaluated at the maximum likelihood parameter values, as returned

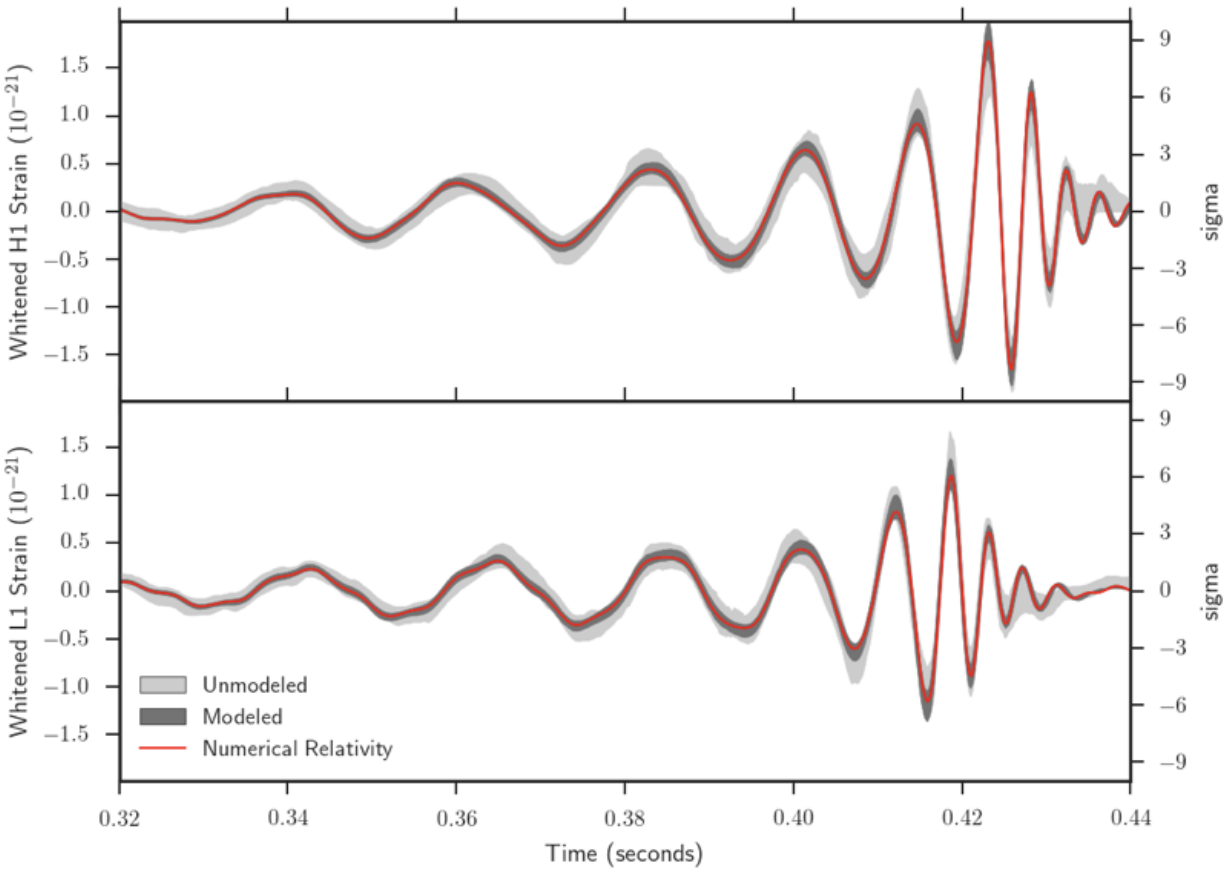


Figure 38: BayesWave reconstruction of GW150914 (labelled "unmodelled"), compared to the waveform corresponding to the maximum a posteriori parameters obtained by parameter estimation (labelled "modelled") and a numerical relativity waveform with consistent parameters. Figure reproduced from Abbott, B.P., et al., *Phys. Rev. Lett.* **116**, 061102 (2016).

by the online search algorithms. This autocorrelation likelihood takes the form

$$\exp\left[-\frac{1}{2}\sum_i \rho_i^2 + \sum_i \rho_i \Re\left\{e^{-i\gamma_i} z_i^*(\tau_i)\right\}\right]$$

where ρ_i denotes the signal to noise ratio in detector i, γ_i and τ_i are the phase and time of arrival of the trigger in detector i and $z_i(t)$ is the time-series of the matched filter overlap in detector i. The marginalisation of this integral over all parameters except sky location is accelerated using approximations to the marginalisation integrals and by employing look-up tables. The result of running the algorithm is a sky map probability density, i.e., a weighting of pixels on the sky by their relative probability of being the true location of the observed transient.

More details on the BAYESTAR algorithm can be found in

- Singer, L., and Price, L., *Phys. Rev.* D **93**, 024013 (2016).

Another rapid localisation algorithm used in LIGO is LALINFERENCEBURST or LIB. In this case, computational savings in the model are obtained by representing an arbitrary signal as a single sine-Gaussian

$$h_+(t) = \cos(\alpha)\frac{h_{\rm rss}}{\sqrt{Q(1+\cos(2\phi_0)e^{-Q^2})/4f_0\sqrt{\pi}}}\sin(2\pi f_0(t-t_0)+\phi_0)e^{-(t-t_0)^2/\tau^2}.$$

While this simple model cannot accurately describe all signals, it does represent the relative amplitudes of the signal in different detectors correctly and that is enough to obtain reasonable sky-localisation accuracies.

There is also an online version of LIB, called OLIB, that uses Bayesian evidences computed by LIB to assess triggers identified in a time-frequency analysis. The evidences for the triggers being noise versus signal and being coherent in different detectors versus incoherent are used to identify potentially interesting candidate events for follow-up. OLIB was running at the time of GW150914 and, along with CWB, was the first algorithm to identify this signal in the data.

More details on the LALINFERENCEBURST algorithm and on *oLIB*, can be found in

- Essick, R., Vitale, S., Katsavounidis, E., Vedovato, G., and Klimenko, S., *Astrophys. J.* **800**, 81 (2015).

- Lynch, R., Vitale, S., Essick, R., Katsavounidis, E., and Robinet, F., *Phys. Rev.* D **95**, 104046 (2017).

7.7 LISA parameter estimation*

Bayesian methods have also been used in the context of data analysis development for LISA, mostly in the framework of the sequence of Mock LISA Data Challenges (MLDCs) that took place between 2006 and 2010. Bayesian techniques, with some frequentist simplifications such as the use of the $\mathcal{F}$-statistic, were used not only to characterise the identified sources, but also to search for sources in the data set. A variety of techniques were employed, including Markov Chain Monte Carlo algorithms, genetic algorithms and nested sampling. These methods were successfully able to find and characterise sources in the sample data sets,

source ($\mathrm{SNR}_{\mathrm{true}}$)	group	$\Delta M_c/M_c$ $\times 10^{-5}$	$\Delta\eta/\eta$ $\times 10^{-4}$	Δt_c (sec)	Δsky (deg)	Δa_1 $\times 10^{-3}$	Δa_2 $\times 10^{-3}$	$\Delta D/D$ $\times 10^{-2}$	SNR	FF_A	FF_E
MBH-1	AEI	2.4	6.1	62.9	11.6	7.6	47.4	8.0	1657.71	0.9936	0.9914
(1670.58)	CambAEI	3.4	40.7	24.8	2.0	8.5	79.6	0.7	1657.19	0.9925	0.9917
	MTAPC	24.8	41.2	619.2	171.0	13.3	28.7	4.0	1669.97	0.9996	0.9997
	JPL	40.5	186.6	23.0	26.9	39.4	66.1	6.9	1664.87	0.9972	0.9981
	GSFC	1904.0	593.2	183.9	82.5	5.7	124.3	94.9	267.04	0.1827	0.1426
MBH-3	AEI	9.0	5.2	100.8	175.9	6.2	18.6	2.7	846.96	0.9995	0.9989
(847.61)	CambAEI	13.5	57.4	138.9	179.0	21.3	7.2	1.5	847.04	0.9993	0.9993
	MTAPC	333.0	234.1	615.7	80.2	71.6	177.2	16.1	842.96	0.9943	0.9945
	JPL	153.0	51.4	356.8	11.2	187.7	414.9	2.7	835.73	0.9826	0.9898
	GSFC	8168.4	2489.9	3276.9	77.9	316.3	69.9	95.6	218.05	0.2815	0.2314
MBH-4	AEI	4.5	75.2	31.4	0.1	47.1	173.6	9.1	160.05	0.9989	0.9994
(160.05)	CambAEI	3.2	171.9	30.7	0.2	52.9	346.1	21.6	160.02	0.9991	0.9992
	MTAPC	48.6	2861.0	5.8	7.3	33.1	321.1	33.0	149.98	0.8766	0.9352
	JPL	302.6	262.0	289.3	4.0	47.6	184.5	28.3	158.34	0.8895	0.9925
	GSFC	831.3	1589.2	1597.6	94.4	59.8	566.7	95.4	−45.53	−0.1725	−0.2937
MBH-2	AEI	1114.1	952.2	38160.8	171.1	331.7	409.0	15.3	20.54	0.9399	0.9469
(18.95)	CambAEI	88.7	386.6	6139.7	172.4	210.8	130.7	24.4	20.36	0.9592	0.9697
	MTAPC	128.6	45.8	16612.0	8.9	321.4	242.4	13.1	20.27	0.9228	0.9260
	JPL	287.0	597.7	11015.7	11.8	375.3	146.3	9.9	18.69	0.9661	0.9709
MBH-6	AEI	1042.3	1235.6	82343.2	2.1	258.2	191.6	26.0	13.69	0.9288	0.9293
(12.82)	CambAEI	5253.2	1598.8	953108.0	158.3	350.8	215.4	29.4	10.17	0.4018	0.4399
	MTAPC	56608.7	296.7	180458.8	119.7	369.2	297.6	25.1	11.34	-0.0004	0.0016

Figure 39: Summary of the fractional errors in the recovery of parameters of the supermassive black hole binary mergers in the third MLDC data challenge. The final two columns, labelled FF_A and FF_E, give the overlap (or "fitting factor") of the waveform corresponding to the recovered parameters with the true injected waveform. Each row represents a separate entry from one of the groups responding to the challenge. Table reproduced from Babak, S., et al., *Class. Quantum Grav.* **27**, 084009 (2010).

although these were somewhat simplified, containing only Gaussian instrumental noise with known PSD and a reduced number of astrophysical sources. In Figure 39 we show a table of parameter measurement precisions of supermassive black hole mergers for all submissions to the third round of the MLDC. The final two columns of the table show the fitting factor, i.e., overlap, of the submitted entry with the true source in each of the two independent LISA data channels, A and E.

The use of Bayesian techniques for searches as well as parameter estimation in the LISA context is motivated by the nature of the data. In the LIGO/Virgo context, most sources are of short duration relative to the time between signals, and so it is necessary to efficiently sift through large amounts of data to find candidate sources of interest. In the LISA context, the source duration is comparable to the length of the data stream and so the entire data stream is relevant for the analysis of all sources. It is natural therefore to find and characterise sources simultaneously.

While the MLDCs demonstrated the effectiveness of the use of Bayesian methods to find and characterise most source types, several open questions remain, in particular related to the impact of non-stationary noise and instrumental artefacts such as gaps, the full extent of source confusion and the detection and characterisation of extreme-mass-ratio inspirals (EM-

type[1]	ν (mHz)	$\mu/M_\odot$	$M/M_\odot$	e_0	θ_S	φ_S	λ	a/M^2	SNR
True	0.1920421	10.296	9517952	0.21438	1.018	4.910	0.4394	0.69816	120.5
Found	0.1920437	10.288	9520796	0.21411	1.027	4.932	0.4384	0.69823	118.1
True	0.34227777	9.771	5215577	0.20791	1.211	4.6826	1.4358	0.63796	132.9
Found	0.34227742	9.769	5214091	0.20818	1.172	4.6822	1.4364	0.63804	132.8
True	0.3425731	9.697	5219668	0.19927	0.589	0.710	0.9282	0.53326	79.5
Found	0.3425712	9.694	5216925	0.19979	0.573	0.713	0.9298	0.53337	79.7
True	0.8514396	10.105	955795	0.45058	2.551	0.979	1.6707	0.62514	101.6
Found	0.8514390	10.106	955544	0.45053	2.565	1.012	1.6719	0.62534	96.0
True	0.8321840	9.790	1033413	0.42691	2.680	1.088	2.3196	0.65829	55.3
Found	0.8321846	9.787	1034208	0.42701	2.687	1.053	2.3153	0.65770	55.6
Blind									
True	0.1674472	10.131	10397935	0.25240	2.985	4.894	1.2056	0.65101	52.0
Found	0.1674462	10.111	10375301	0.25419	3.023	4.857	1.2097	0.65148	51.7
True	0.9997627	9.7478	975650	0.360970	1.453	4.95326	0.5110	0.65005	122.9
Found	0.9997626	9.7479	975610	0.360966	1.422	4.95339	0.5113	0.65007	116.0

Figure 40: Maximum a posteriori parameter values (labelled "Found") recovered for all five EMRIs in the MLDC data set 1B (upper rows) and two additional random chosen sources. These are compared to the "Ture" parameters which were used ot generate the injected signals. Table reproduced from Babak, S., Gair, J.R., and Porter, E.K., *Class. Quantum Grav.***26**, 135004 (2009).

RIs). While the EMRI sources in the MLDC data sets were successfully characterised under simplified assumptions (see Figure 40), the likelihood for an EMRI is very complicated, with many secondary maxima in parameter space. The successful algorithms relied on knowledge of the structure of the likelihood surface, which was specific to the simplified model of the EMRI employed in the MLDC, and the fact that all identified secondaries were generated by the same EMRI signal. While the structure of the likelihood surface can probably be learned for more accurate waveform models, the correct grouping of secondary modes will be much more challenging for real LISA data which could contain many hundreds of EMRIs.

Nested sampling has also been used in the context of LISA data analysis. In fact, the first application of the MultiNest nested sampling algorithm in a gravitational wave context was to the characterisation of supermassive black hole mergers in LISA data (Feroz, F., Gair, J.R., Hobson, M.P., and Porter, E.K., *Class. Quantum Grav.* **26**, 215003). MultiNest was also used to find and characterise supermassive black hole mergers and gravitational wave bursts from cosmic string cusps in MLDC data. In the latter case, the computed Bayesian evidences were used to test the hypothesis that the burst signals were consistent with a cosmic string cusp as opposed to a generic sine-Gaussian burst model (see Figure 41 and Feroz, F., Gair, J.R., Graff, P., Hobson, M.P., and Lasenby, A., *Class. Quantum Grav.* **27**, 075010 (2010)).

Further details on LISA data analysis can be found in the MLDC papers, and references therein:

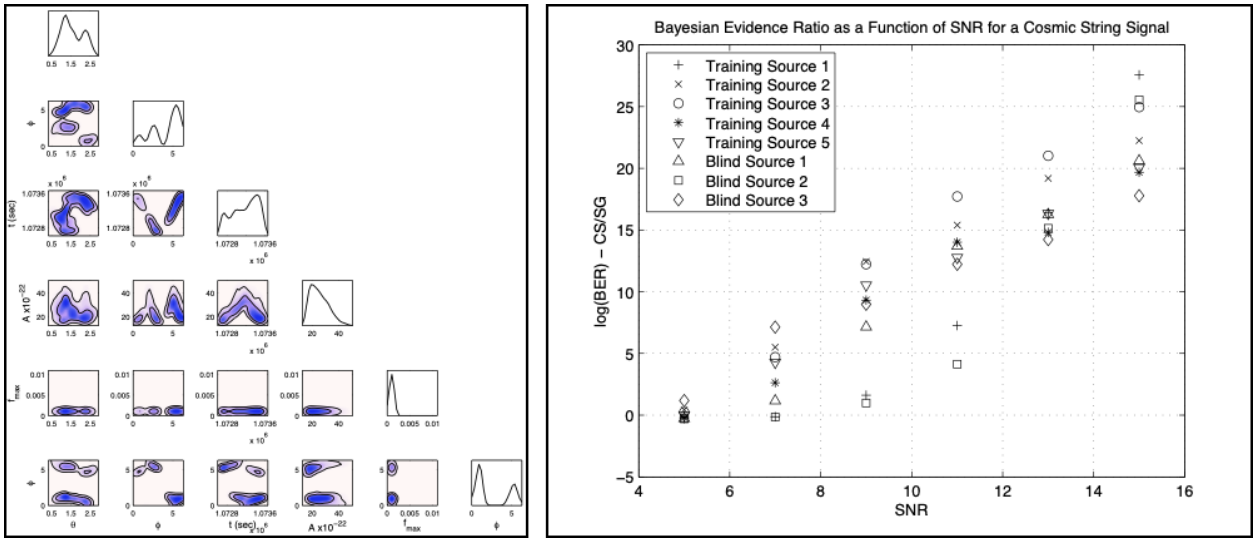


Figure 41: Left panel: posterior on the parameters characterising one of the cosmic string cusp gravitational wave bursts in the MLDC round 3 cosmic string data set. Right panel: evidence ratio in favour of the true (cosmic string cusp) model versus an alternative (sine-Gaussian) model for the burst, as a function of the burst signal-to-noise ratio. Figures reproduced from Feroz, F., Gair, J.R., Graff, P., Hobson, M.P., and Lasenby, A., *Class. Quantum Grav.* **27**, 075010 (2010).

- Arnaud, K.A., et al. *The Mock LISA Data Challenges: An overview*, *AIP Conf. Proc.* **873**, 619 (2006).
- Arnaud, K.A., et al., A How-To for the Mock LISA Data Challenges, *AIP Conf. Proc.* **873**, 625 (2006).
- Arnaud, K.A., et al., *Report on the first round of the Mock LISA Data Challenges*, *Class. Quantum Grav.* **24**, S529 (2007).
- Arnaud, K.A., et al., *An overview of the second round of the Mock LISA Data Challenges*, *Class. Quantum Grav.* **24**, S551 (2007).
- Babak, S., et al., *Report on the second Mock LISA Data Challenge*, *Class. Quantum Grav.* **25**, 114037 (2008).
- Babak, S., et al., *The Mock LISA Data Challenges: from Challenge 1B to Challenge 3*, *Class. Quantum Grav.* **25**, 184026 (2008).
- Babak, S., et al., *The Mock LISA Data Challenges: from Challenge 3 to Challenge 4*, *Class. Quantum Grav.* **27**, 084009 (2010).