Bayesianism and Gravitational Waves

Yixuan Dang
Balliol College
Mphysphil Philosophy Thesis

Abstract

Gravitational waves (GW) have attracted probably the most attention in astrophysics since its first detection in 2016. With GW observations, astrophysicists are not only able to detect gravitation events that are invisible through traditional electromagnetic means, but also carry out tests of General Relativity (GR) and extract constraints for other cosmological models. The entire research of GW is conducted based Bayesian statistics, which makes it a good subject for a concrete case study of philosophy of probability. This thesis aims to examine the application of Bayesian inference in GW research, specifically on parameter estimation (i.e. how masses of black holes are extracted from the signal), hypothesis comparison (i.e. how GW is tested to be our best theory), and an interesting example of GW astronomy. I attempt to show how data analysis in GW research is faithfully Bayesian and how the Bayesian methodologies should be in general preferred but with some existing philosophical concerns. Some key conclusions are the following:

1. Bayesianism allows us to extract more useful information from singular events like GW, comapared to Frequentism. Fisher information, as a mutual mathematical tool used in both frameworks, provides a measure of accuracy in Bayesianism but only a measure of error in Frequentism.

2. Bayes factor (i.e. a ratio of likelihoods) is a good metric of Bayesian hypothesis comparison as it rewards both simplicity and goodness of fit.

3. Hypothesis comparison by Bayes factor does not fall into Likelihoodism even if the prior odds ratio is taken as 1 (as it has been practiced in GW research).

4. No meaningful objective interpretation of Bayes factor is justified. The number says all there is to say.

5. Selection effects (i.e. only 'loud' signals can be detected and therefore participate in the inference) only occur in inference about population parameters (e.g. the distribution of mass of black holes in the universe), but not inference about individual properties (e.g. the mass of a particular black hole).

Table of Contents

1 | Introduction

As worrying as Bayesian inference might be to philosophers, it has been persistently used in sciences, notably in the Nobel-prize-winning discovery and analysis of gravitational wave (GW) signals. This thesis aims to give a detailed presentation of how Bayesianism is applied in GW research, and a justification of why the current application of Bayesianism deserves both preference and trust in these studies from the perspective of philosophy of science and probability. We aim to identify and articulate the initial challenges faced by learners in gravitational wave research, tracing these issues back to their philosophical roots.

The thesis is organised as follows:

> **Chapter 2** — Parameter Estimation. This chapter examines the Bayesian parameter estimation used in GW research, which is the most fundamental building block of Bayesian inference in GW. This is the process where physical properties (e.g. chrip mass) of the merger postulated from the signal are estimated. The methods of parameter estimation discussed in this chapter describes how current GW researchers extract information (i.e. values of each parameter) from the received signal, and this information is crucial to complete any of the 'more important' tasks in later chapters. In this chapter Bayesian methods will be persistently contrasted with its frequentist alternative, and we aim to argue that Bayesianism should be preferred for analyses of GW events, which are singular events by nature.
>
> **Chapter 3** — Hypothesis comparison. This chapter focuses on the Bayesian hypothesis comparison scheme by Bayes factor. Tests of General Relativity (GR) are one concrete example of Bayesian hypothesis comparison, which yield possibly the most significant results in GW studies. I will give arguments on how Bayes factor (1) does not lead the inference into Likelihoodism; (2) acts a good metric as it values both simplicity and goodness of fit; (3) cannot be further interpreted objectively.
>
> **Chapter 4** — Gravitational-wave cosmology. This chapter describes how GW data allow us to constrain the value of Hubble constant. In particular we will focus on the selection effects introduced in hierarchical Bayesian models.
>
> **Chapter 5** — Conclusions.

2 | Parameter estimation

2.1 An overview

One of the crucial goals in gravitational-wave (GW) research is to determine the GW source parameters given a detection. Parameter estimation is the process of inferring values of parameters of a model given measured data. Two major frameworks for parameter estimation are Frequentism and Bayesianism. Before going into the schematic details of how parameter estimation is done in the frequentist and Bayesian framework, we will first discuss how in general frequentists and Bayesians think about the nature of probability, which to some extent affect how they design methods of parameter estimation. Notice that this chapter mainly aims at comparison between the frequentist and Bayesian **methodologies** and endorsing Bayesianism only because the results come out from Bayesian method is more fruitful for singular events like GW. The below discussion on how frequentist and Bayesian interpret of probability is only providing an intuition for their complicated statistical schemes introduced later.

The frequentists consider probability as a property of a collection of events. However there exists debate about the size and nature of the collection. *Finite frequentists* believe that the probability of an attribute A in a finite reference class B is the relative frequency of actual occurrences of A within B, while *hypothetical frequentists* believe that the probability should be the limiting relative frequency of A if B were infinite (Hájek, 2023), and also that this limit exists. This conceptual difference does not induce a difference in mathematics, but only the interpretation of the same mathematical result. In practical frequentist algorithm, only observed data enter into the probability calculations, adhering to the principles of finite Frequentism. However with the assumption that that the relative frequency calculated from the actual data has already (almost) converged to the limiting relative frequency in the infinite collection, the same mathematics can also be employed by hypothetical frequentists. If this assumption isn't met, then hypothetical frequentist would still agree that the obtained percentage represents the relative frequency of current data, but only disagree with interpreting that frequency as 'probability'. However in real cases where frequentist statistics is applied, it is almost always the case that the sample size is large enough such that the probability converges. Hence later we won't choose which version Frequentism to take when the sample size is large enough: the percentage would be regarded as probability happily by both parties.

On the other hand, Bayesian regards probability as a property of singular events. We do not need a collection of coin tosses, or a hypothetical collection of coin

tosses to define or decide the probability of a single toss. Bayesians disagree among themselves on the nature of this single-event probability, and diverge into two branches. Most importantly, unlike Frequentism, the difference between these two branches results in difference on the mathmatical level. *Objective Bayesians* regard probabilities as the credence that rational agents ought to subscribe, which is an unique, objective number; conversely, *subjective Bayesians* regard probabilities as the subject credence that vary among different people and contexts (Chalmers, 1976). In later GW examples, we will see how GW research has been mainly carried on under subjective Bayesianism. The priors of competing hypotheses (e.g. GR and alternative gravity theory) are often set at 1 by agreement among the scientific community. The nature of setting-up the prior is still subjective, in the sense that it is decided based on the subjective belief of a group of people (no matter how intelligent they are!). In fact, this agreement has now been re-discussed in the GW community as one research group is working on how the priors for GR should be altered after it has passed through so many tests.

Now there is a quick argument that can be made for subjective Bayesianism. Objective Bayesians regard priors $P(\theta)$ as *a priori* probabilities. The biggest obstacle for objective Bayesians is how this *a priori* $P(\theta)$ can ever be determined. One attempt is to apply the principle of indifference, which treats each hypothesis as equally likely (believable) when no evidence is given. The consequence of doing this immediately stops us from doing any further mathematics: suppose we have $\theta \in (-\infty, \infty)$, for all θ, $P(\theta) = 0$ would be true for if there is no evidence. The situation in which $P(\theta) = 0$ is clearly unwanted because posteriors vanish with the priors no matter how strong the evidence is.

On the other hand subjective Bayesians avoid this consequence by setting-up priors subjectively, so they don't have to try so hard finding a unique, correct number for prior. No subjective beliefs would be false if you say that they are subjective! They do face the objection that prior as credence reduces any science using Bayesian statistics to a personal matter. One response to this objection is that with more updates of evidence, any choice of priors will eventually be washed out. However, with fewer cases of events, the choice of prior remains important. In most research papers of GW, priors are usually taken as a constant (or Jeffery's prior, see later) to undermine the influence of personal choice to parameter estimation.

We will now give a sketch of how parameter estimation is principally done in both frameworks. The next subsections introduce the concept of Fisher information (matrix), which is a crucial tool used in Bayesian parameter estimation in GW research. However it was originally derived from the Frequentism framework. The use of this tool will be thoroughly examined to ensure that under the context

of GW research, 1) it is correctly used under reasonable assumptions and 2) it exercises much more power in the Bayesian framework.

2.1.1 Frequentist method

To 'cook' a good parameter estimation in the frequentist framework, let's first look at the recipe book for a frequentist:

Ingredients Gather independent and identically distributed (i.i.d.) data.
Set up a likelihood density function according to how the experiment is designed.

Cooking instructions Choose a point estimator (the 'cooker'), and apply the point estimator to the i.i.d. data and the likelihood density function.

Main meal Enjoy a single guess of the value of the true parameter.

It is worth explaining that normally data from experimental observations are not i.i.d.. The assumption of them being i.i.d. is only an approximation of reality.

Samaniego (2010) nicely phrased the process parameter estimation as a decision problem. Player 1 is nature, whose action space represents all possible states of nature (so all possible values of θ); player 2 is the statistician, whose aim is to guess the card of state played by nature (so to guess the true value θ_0). The two players share the same action space in the scenario of parameter estimation. Suppose the game is to guess the chance of heads of flipping a coin based on the results of 10 flips, and nature 'chooses' $\theta_0 = 0.6$ from 0 to 1. Then player 2 has to make a single guess from 0 to 1 to minimise her loss according to a decision rule. **A point estimator acts like a such decision rule**, which is a map $\hat{\theta} : \mathscr{X}^n \longrightarrow A$ between the data space $\mathscr{X}^n$ and the action space A for player 2. The point estimator decides which value to report given input data. In the coin flipping case, $\mathscr{X}^n$ is all possible sequences of heads or tails in n tosses, and the action space A will be $(0,1)$, in which lies the values of all possible guesses from player 2.

We have understood how a point estimator processes data. Now here comes the most difficult and important step: **choosing the best cooker/point estimator!** We will embark a long and thorough argument on why the maximum likelihood estimator (MLE), under ideal assumptions, is the best point estimator. Below we will go through the argument slowly and introduce several crucial technicalities in frequentist statistics when moving along: risk function, bias, and a complete, sufficient statistic. Then at the end of this subsection, there will be concise organised form of this long argument.

Let's stick to Samaniego (2010)'s decision-theoretic way of understanding the frequentist framework. A good point estimator (i.e. a good decision rule) will aim to minimise the risk function, which is the expected value of loss function $L : \mathscr{X}^n \times A \longrightarrow R$. The loss function quantifies the loss of player 2 by taking the action of reporting $\hat{\theta}(x^n)$ as the true parameter, given that the true parameter is θ_0. One example is the sqaured-error loss function, which is

$$L(\theta_0, \hat{\theta}(x^n)) = (\theta_0 - \hat{\theta}(x^n))^2,$$

where θ_0 is the true parameter, x^n is the results of n flips and $\hat{\theta}(x^n)$ is a guess from player 2. Hence the further the guess $\hat{\theta}(x^n)$ is away from θ_0, the greater the loss of player 2 is going to be. The risk function is defined by the expected loss , averaged over all possible outcomes of the experiment, weighted by their appropriate likelihood,

$$R(\theta_0, \hat{\theta}(x^n)) = \int_{\mathscr{X}^n} L(\theta_0, \hat{\theta}(x^n)) f(x^n|\theta_0) dx^n. \tag{2.1}$$

Hence the choice of point estimator depends on both the choice of loss function L and the probability density model $f(x^n|\theta_0)$. The criterion of selecting between point estimators is called *admissibility*. Let $\hat{\theta}_1$, $\hat{\theta}_2$ be two point estimator, $\hat{\theta}_1$ dominates $\hat{\theta}_2$ iff $R(\theta_0, \hat{\theta}_1(x^n)) \leq R(\theta_0, \hat{\theta}_2(x^n))$ for all true values of parameters θ_0 and there exists at least one value of θ where $R(\theta_0, \hat{\theta}_1(x^n)) < R(\theta_0, \hat{\theta}_2(x^n))$. If a point estimator is not dominated by any other point estimators, we call this point estimator *admissible*. However it is almost never guaranteed that an admissible estimator exists or can be found so normally we only try to rank estimators within a group of choices, based on how close they are to perfect admissibility. The most popular group of estimators is the *unbiased* estimators $\hat{\theta}$, defined as

$$\theta_0 = E(\hat{\theta}) = \int_{\mathscr{X}^n} \hat{\theta}(x^n) f(x^n|\theta_0) dx^n. \tag{2.2}$$

We can spot the 'unbiasedness' from this definitive equation. In the case of flipping a coin for 10 times, if the data contains 6 heads given the true $\theta_0 = 0.6$, then $f(x^n|\theta_0)$ is large ; and if the data contains 1 heads given the true $\theta_0 = 0.6$, then then its corresponding $f(x^n|\theta_0)$ is very small. In other words, the decision rule corresponds to an unbiased estimator ensures that the expectation value $\hat{\theta}$ gets to the true θ_0 without introducing bias (i.e.$\theta_0 = E(\hat{\theta}) + b$ where b is bias) . The most admissible unbiased estimator is called the *uniformly minimum variance unbiased estimator* (UMVUE). UMVUE has the minimum risk function owing to the fact that the risk function is the variance: $R(\theta_0, \hat{\theta}) = E((\theta_0 - \hat{\theta})^2) = \sigma_\theta(\hat{\theta})$, where σ_θ represents the variance of $\hat{\theta}$ given θ_0.

But what does it take to be an UMVUE? From now we will give an argument of why the unbiased maximum likelihood estimator (MLE) is an UMVUE and therefore the ideal point estimator to take. Lehmann–Scheffé theorem states that **when the estimator can be written as a function of a complete sufficient statistic T such that $\hat{\theta}(T)$, then such $\hat{\theta}(T)$ is an UMVUE** [1]. A statistic functions like a summary of the data. Take the coin tossing case for example, instead of recording the exact sequence of heads and tails, we could choose to only do statistics about T, which could be the total number of heads. A statistic T is sufficient iff $P(x^n|T=t,\theta)$ has no dependence on θ. In other words, a specification of T is a sufficient to convey all information there is in the data about parameter θ. In our example the total number of heads would be a sufficient statistic of estimating the chance of heads θ. A sufficient statistic T is *complete* for parameter θ iff for every θ, $E_\theta g(T)=0$ holds for every measurable function g, then $g(T)$ is a trivial function of constant zero. In other words, there's no function of T (other than the trivial zero function) that will give you an expected value of zero across all values of θ, **which implies that any distribution of T corresponds to some information about θ**. When a statistic is both sufficient and complete, it means that it summarizes all the information in the sample that is relevant for estimating the parameter, and any information of the statistic must carry information about the parameter.

If a complete sufficient statistic exists, then it must be an unbiased MLE. MLE gives out the value of θ where the likelihood is at maximum: $\theta^{\hat{\mathrm{ML}}} = \arg(\frac{\partial}{\partial\theta}\mathrm{P}(\mathrm{x}^\mathrm{n}|\theta)) = 0$. Most importantly, MLE is a function of every sufficient statistic $\theta^{\hat{\mathrm{ML}}} = \theta^{\hat{\mathrm{ML}}}(T)$[2]. Hence later we will be justified to adopt MLE as the best point estimator in the coin tossing scenario.

With everything defined and explained, I have pieced together the argument for why the unbised MLE is the point estimator to go for in a concise form:

- The preferred estimator should be the best at admissibility (minimising risk), and unbiasedness.
- UMVUE is the most admissible, unbiased estimator
- When an estimator can be written as a function of a complete sufficient statistic T, it is UMVUE

[1] Notice the difference between $\hat{\theta}(T)$ and $\hat{\theta}(x^n)$. The former means that $\hat{\theta}$ is a function of T, and the latter notation used in previous paragraphs only means $\hat{\theta}$ takes in the data x^n but how $\hat{\theta}$ processes the data remains undefined (it might ignore the data at all)

[2] The main idea of the proof is that maximising $P(x^n|\theta)$ would be the same as maximising $P(T=t|\theta)$

- When a complete sufficient statistic exists, it must be an MLE.

Conclusion 1 Therefore, when a complete sufficient statistic exists, the unbiased MLE must be an UMVUE.

Conclusion 2 Therefore, when a complete sufficient statistic exists, the unbiased MLE is the preferred estimator.

Unfortunately, it is again, not always possible to find a complete sufficient statistic, especially for complicated probability density models. Hence we will conclude that it is a challenge for frequentists to find a best cooker (point estimator) to use.

2.1.2 Bayesian method

In comparison to the frequentist approach, here we present a Bayesian's recipe book of parameter estimation for a single experiment:

Ingredients Get data $\mathscr{D}$
Set up the likelihood density function $p(\mathscr{D}|\theta)$ according to the experiment design
Determine a prior distribution $p(\theta)$ reflecting our initial beliefs about θ

Cooking instruction Apply the Bayes theorem: $P(\theta|\mathscr{D}) = \frac{P(\mathscr{D}|\theta)*P(\theta)}{P(\mathscr{D})} = \frac{P(\mathscr{D}|\theta)*P(\theta)}{\int_\Theta p(\mathscr{D}|\theta)p(\theta)d\theta}$

Main meal Enjoy the posterior distribution of θ: $P(\theta|\mathscr{D})$

Cooking suggestions Choose a point estimator if you want to enjoy a single guess of θ! For example, one can choose the maximum posterior estimator, which outputs the value of θ where $P(\theta|\mathscr{D})$ reaches maximum
Give a 90% (or other percentage) *credible interval* of where the true value of θ lies. For example, $74.0^{+33.0}_{-12.0}$ at 90% MAP interval means that 74.0 is the value at the maximum posterior and $(74-12, 74+33)$ gives the smallest region enclosing 90% of the posterior (this comes from a concrete example where Hubble constant is estimated in the last chapter).

Cooking suggestions are not necessary as $P(\theta|\mathscr{D})$ already includes all information about θ that is available through Bayesian inference, but most Bayesian data analyses include at least one of these two steps as a descriptive conclusion drawn from the posterior distribution.

2.1.3 So why Bayesian?

The major difference between the frequentist and Bayesian recipe books originates from how they think about the nature of probability as discussed at the

beginning of this subsection. The major statistical property of GW events is that they happen once only. As frequentists think of probability as a property derived from a collection of occurrences, they find it hard to apply parameter estimation to one single event. Although you can say that a single event is still a collection of a single event, it is another stretch to make hypothetical frequentists believe that any probabilistic conclusion drawn from this one-event collection has converged to its limit. On the other hand, Bayesians come with a prior distribution, and it is principally unproblematic to update prior beliefs by the evidence from a single event. Therefore, it is more natural to fit gravitational-wave events into the Bayesian framework as they are non-repeatable single events.

Finite frequentists might argue that the fit between data and statistical principle is just as good for them as well, since they are happy to accept collections of any size. However, the difference in principle also results in difference in the interpretation of the mathematics, which will be clearly recognised in the next section when we compare the interpretation of Fisher information matrix in each framework. But for now, we can get the general idea of why the Bayesian meal would be more 'nutritious'. If we look at the empirical ingredients going to the frequentist recipe book, it consists of data, a likelihood density function (given according to the experiment setup and hence empirical), while the Bayesian recipe book contains one more prior distribution. The fact that the Bayesian have richer ingredients simply imply that they will also get a richer product, which we will very soon spot in the mathematics.

2.2 Fisher information

Despite disagreement on the principles of statistics, surprisingly both frequentists and Bayesians obtain the same crucial matrix in the process of parameter estimation. Fisher matrix (FM) is quick tool to work out errors of the single estimated value(in the frequentist framework) or the shape of the posterior distribution (in the Bayesian framework) as its inverse approximates the covariance matrix of the posterior. We will begin with the simplest case where the model is controlled by one parameter only, FM reduces to a 1D matrix, which is just a single value called the Fisher information, denoted by $I_T(\theta)$.

Fisher information $I_T(\theta)$ aims to quantity the amount of information we have about the value of a parameter θ from a statistic T in a given probability density model $f(T|\theta)$ (Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2017). The concept of a statistic is defined in the previous section where the frequentist paradigm is introduced for the first time. In short a statistic T is the the variable that we keep record of as a useful summary of the data. In the coin flipping case,

if we want to estimate the chance of heads, the statistic could be X^n, which is just the raw data of sequence of outcomes of n flips; or more concisely we can use Y, which is the total number of heads obtained in n throws. $I_T(\theta)$ is mathematically defined such that a higher $I_T(\theta = \theta_0)$ entails that the statistical model is designed such that it is more sensitive to whether $\theta = \theta_0$ when any data come. In other words, the Fisher information is a property of the statistical model (equivalently the likelihood density function or the experimental setup), which reflects how informative the model would be at each particular value of θ. For example, the experimental setup of coin tossing is really sensitive to whether $\theta = 0$ where θ is the chance of heads, in the sense that within very few throws, we can conclude that $\theta \neq 0$. While $I_T(\theta)$ is a function of θ, we should also be aware of how this function is constructed from the choice of statistic. Different statistics might have different function of Fisher information, which implies that how you take notes of the data is important.

There should be no more wait to introduce the actual mathematical definition of Fisher information of a random variable X about the parameter θ at θ_0:

$$I_X(\theta_0) = \int_{\mathscr{X}} \Big(\frac{d \log p(x|\theta)}{d\theta}|_{\theta_0}\Big)^2 p(x|\theta) dx, \tag{2.3}$$

where $\mathscr{X}$ is the space of outcomes recorded according to the chosen statistic. When X takes discrete values (like the number of heads in a row of throws), the integral is changed into a sum. We have also assumed regularity conditions for Fisher information (matrix) to mathematically exist: for example, $p(x|\theta)$ is always differentiable at every θ_0. The most important part of this definition is $\frac{d}{d\theta} \log p(x|\theta)|_{\theta_0}$, which describes how sensitive $p(x|\theta)$ is to a change in θ at a particular value θ_0. The more sensitive $p(x|\theta)$ is at θ_0, the higher the square of the derivative is, and hence the more information we have about whether $\theta = \theta_0$.

For an intuitive understanding of Fisher information, we can look at the simple scenario of flipping a coin. This example will not only convince us of how Fisher information encodes how informative the model/experiment setup is about each value parameter, but also allow us to understand how taking different statistics might or might not affect this. We begin by assuming that the outcomes of coin flipping are controlled by and only by the chance of heads. Suppose we have a coin with unknown chance $\theta_0 \in [0,1]$. The goal is to estimate the chance parameter θ. First let's choose the raw data X^n to be the statistic. Our statistical model $f(x_i|\theta)$ represents how θ is related to the outcome x_i of the ith throw X_i. Taking Bernoulli

distribution as our model, which is defined as:

$$f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}, \tag{2.4}$$

where x_i is the outcome of the ith flip, and $x_i = 1$ when landing on heads, $x_i = 0$ when landing on tails. If each throw is an individual event, then we have:

$$f(x^n|\theta) = \prod_i^n \theta^{x_i}(1-\theta)^{1-x_i}. \tag{2.5}$$

However, it's too much effort to memorise the exact sequence of outcomes. Instead of recording every outcome in sequence, we can just record the sum of the outcomes of n flips $Y \in [0,n]$, and use Y as the statistic. For the potential outcome y of Y, we have:

$$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \binom{n}{y}\prod_i^n \theta^{x_i}(1-\theta)^{1-x_i}. \tag{2.6}$$

Now through one simple calculation, it is straightforward to see that there is no information about θ left in X^n after observing Y. The probability of $X^n = x^n$ after observing $Y = y$ is:

$$P(X^n = x^n|Y = y, \theta = \theta_0) = 1/\binom{n}{y}. \tag{2.7}$$

We conclude that X^n and Y both contain the same volume of information that is useful to the parameter estimation of θ because $P(X^n = x^n|Y = y, \theta = \theta_0)$ has no dependence on θ. One more thing we can say is that since X^n is raw data, Y is a sufficient statistic as it extracts all the information about θ from the model[3].

Substituting likelihood density function Eq.2.5 and Eq.2.6 into the definition of Fisher information Eq.2.3, we have:

$$I_{X^n}(\theta_0) = I_Y(\theta_0) = \frac{1}{\theta_0(1-\theta_0)}. \tag{2.8}$$

This is a function that tends to infinity at $\theta_0 = 0$ and $\theta_0 = 1$, which means for a single toss, the model gives the most information about whether the chance of coin is 0 or 1. This agrees with our intuition: obtaining a head in a single toss should gives the most preference to $\theta = 1$, and the most rejection to $\theta = 0$. This implies that the experimental setup of tossing a coin is really sensitive at testing

[3] A sufficient statistic T is such that $P(x^n|T = t, \theta)$ has no dependence on θ.

extreme values of chance of heads.

In estimation of multiple parameters, the Fisher information extends to the Fisher matrix:

$$F_{il} =< \partial_i(\log p(s|\vec{\theta}))|_{\vec{\theta}_0}, \partial_l(\log p(s|\vec{\theta})|_{\vec{\theta}_0} >, \tag{2.9}$$

where s stands for signal data, and the ensemble product is defined by

$$< u, v >= \int u(s)v(s)p(s|\vec{\theta}_0)ds. \tag{2.10}$$

Vector symbols are used to represent multi-dimensional parameters.

The frequentists and Bayesians have not diverged at this point. Everything about Fisher information (matrix) at this point remains as the property of informative power of the statistical model/experimental setup. Now we are ready to extend the use of FM in both the frequentist and Bayesian paradigm, and it will shown that, with appropriate assumptions, it is a more powerful estimation tool in the Bayesian framework. FM, in the frequentist framework, only shows how precise the measurement is, but while it shows accuracy in the Bayesian framework.

2.2.1 Frequentist

The link between Fisher information and the frequentist parameter estimation is established through the Cramer-Rao bound (CRB):

> Suppose i.i.d. data x^n and the unbiased estimator $\hat{\theta}(x^n)$, then
>
> $$\text{var}(\hat{\theta}) \geq \frac{1}{\text{nI}_{\text{X}^\text{n}}} \tag{2.11}$$
>
> where $\text{var}(\hat{\theta})$ is the variance of $\hat{\theta}(x^n)$, and X^n is the outcome of observation. And in the case of multi-parameter estimation, we further have:
>
> $$C(\hat{\theta}_i, \hat{\theta}_l) \geq F^{-1}_{mj}, \tag{2.12}$$
>
> where $C(\hat{\theta}_i, \hat{\theta}_l)$ is the covariance matrix of estimated variables $\hat{\theta}_i, \hat{\theta}_l$.

When an unbiased point estimator actually achieves the lower bound (i.e. Covariance matrix = inverse of FM), we call this point estimator the *efficient* estimator. Recall from the last section that UMVUE has the least variance, and that if UMVUE exists, it is an unbiased MLE. Hence an unbiased MLE is the efficient estimator that actually achieves this bound. Now let's calculate the CRB in the simple coin-tossing scenario. The result is that, in n number of tosses, the

distribution of variable θ^{ML} has the following form (θ^{ML} is recentred to zero for simpler expression):

$$\theta^{ML} \sim \mathcal{N}(0, 1/nI_Y(\theta^{ML} = 0)) \tag{2.13}$$

where $\mathcal{N}$ stands for normal distribution, with $\sigma = \sqrt{1/nI_Y(\theta^{ML} = 0)}$. So the larger the Fisher information is at the centre, the less the standard deviation is, which corresponds to obtaining a more precise value of the unbiased MLE.

2.2.2 Bayesian

Fisher information plays two roles in Bayesianism. Our main focus is on how the inverse of FM approximates the covariance matrix of parameters, which contains information about the posterior distribution. Then we will also discuss how FM helps setting up priors in the uninformative cases.

Bernstein–von Mises (BVM) theorem concerns about the asymptotic behavior of Bayesian inference towards accumulation of events. It essentially states that the Bayesian posterior distribution of a parameter converges to a normal distribution the sample size increases, regardless of the form of the prior distribution. The part of the theorem that is useful for our discussion can be formulated as follows (Samaniego, 2010):

> For any prior distribution G with density $g(\theta) > 0$ for all $\theta \in \Theta$, and $\hat{\theta}_G$ is the Bayes estimator (will define later), when the number of samples n tends to infinity, under regularity conditions, we have
>
> $$\sqrt{n}(\hat{\theta_G} - \theta_0) \sim \mathcal{N}(0, I_X^{-1}(\theta_0)) \tag{2.14}$$

where θ_0 represents the true value of parameter θ. Bayes estimator $\hat{\theta}_G$ minimise the posterior expected loss called the Bayes risk $E_{\theta|X=x}L(\theta, \hat{\theta}(x))$. This is different from the risk function in the frequentist framework, in which the loss function is weighted by likelihood instead of posterior probability. Eq. 2.14 takes a similar form as Eq. 2.13 but with different interpretations. While Eq. 2.13 represents the measurement uncertainties of the guess generated from MLE, Eq. 2.14 represents how guesses generated from bayes estimator is distributed around the true value of parameter θ_0.

We are curious about whether this is the motivation of using FM in the Bayesian framework in GW studies. It seems that BVM theorem is not applicable in GW events because they are single, non-repeatable events, and the theorem describes the asymptotic behaviour of the posterior distribution. However the reason BVM

theorem requires a large number of experiments is to update the posterior frequent enough such that it washes out effects of priors. As we will see later, in data analysis of GW, prior is assumed as a constant so it does not contribute anything to the shape of the posterior and hence $\hat{\theta}_G$ ignores priors completely. Further more, the posterior distribution is assumed as a normal distribution because the signal is only analysed when it is very loud (such that we expect clean 'pointy' distribution about its parameters), which already assumes some of the BVM theorem. This could possibly be the reason why FM can still be derived to approximate the variance of GW parameters from the Bayesian approach.

Another role of Fisher information in the Bayesian framework is to define the Jeffery's prior, which will then solve the problem of constant prior. The problem of constant prior is that a different parameterisation of θ gives a different posterior of θ after being updated by observation. In the coin flipping case, suppose we have two parameterisation θ and $\phi = \theta^2$, a uniform prior of θ is equivalent to a non-uniform prior of ϕ.

$$\begin{aligned} p(\theta)d\theta &= p(\phi)d\phi \\ &= p(\phi)d(\theta^2) \\ &= p(\phi)2\theta d\theta \end{aligned} \tag{2.15}$$

Now suppose we take the constant prior. $P(\theta|x^n)$ would not be the same under these different parameterisation despite updating with same observation. As there exists many ways for parameterisation, it is underdetermined which parameter has the privilege to be uniform.

The Jeffery's prior would solve this problem, as it is proved for the coin flipping case and in general that a different parameterisation of ϕ does not change the posterior of ϕ. It is defined as

$$p_J(\theta) = \sqrt{I_X(\theta)} / \int_\Theta \sqrt{I_X(\theta)} d\theta \tag{2.16}$$

where Θ is the space of θ.
After some tedious algebra, the resulting posteriors have the following features:

$$p(\theta|x^n) = p(\phi|x^n). \tag{2.17}$$

Nevertheless I have not found any GW papers using Jeffery's prior in their data analysis. This is strange in the sense that a faithful Bayesian would use Jeffery's

prior instead of constant priors in the uninformative case. This does not constitute as a fatal problem for GW research since subjective Bayesians after all have freedom to set their priors however they like. It could be that, in most cases, the choice of parameterisation is rather unique (how can you parameterise the mass of a star differently?), and therefore the problem of constant priors does not actually occur.

2.3 FM applied to GW

We will first introduce the basics properties of GW waveform and noise, and then give a simplified version of derivation of FM within each framework.

2.3.1 The likelihood function

Both frequentist and Bayesian statistics require a likelihood density function. In the case of GW signals, the likelihood of observing signal s given the GW wave $h(\theta_0)$ is derived in (Creighton & Anderson, 2011). Here we will present some key steps of their derivation. Only the last three derived equations will be used for later philosophical discussion.

We first determine P(0), the prior probability of pure noise. Suppose now we have N samples of noise x_j at interval Δt, each being independence Gaussian variables (white noise), the probability of obtaining collection $\{x_j\}$ is :

$$p(\{x_j\}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\{-\frac{1}{2\sigma^2}\sum_{j=0}^{N-1} x_j^2\} \tag{2.18}$$

where we have assumed zero mean and variance of σ^2. Taking the limit $\Delta t \to 0$ and large T :

$$\lim_{\Delta t\to 0}\sum_{j=0}^{N-1} x_j^2 = \int_0^T x^2(t)dt \approx \int_{-\infty}^{\infty} |\widetilde{x}(f)|^2 df \tag{2.19}$$

For white noise, we have the power spectral density (PSD) $S_x(f) = \lim_{\Delta t\to 0} 2\sigma^2\Delta t$ (power density in the frequency domain) such that:

$$-\frac{1}{2\sigma^2}\lim_{\Delta t\to 0}\sum_{j=0}^{N-1} x_j^2 \approx \int_0^{\infty} 2\frac{|\widetilde{x}(f)|^2}{S_x} df \tag{2.20}$$

Hence we happily arrive at the continuum limit probability density of time-series

of noise x(t) :

$$p(x(t)) \propto \exp\left\{-2\int_0^\infty \frac{|\widetilde{x}(f)|^2}{S_x} df\right\} = e^{-(x,x)/2} \tag{2.21}$$

where the PSD S_x are determined as a detector parameter which is published by LIGO.

Now we are close to see the final result. let our null hypothesis H_0 be that no GW signal is present in the detector output. We notice that $P(s|H_0)$is just the $p(x(t))$ above, because the probability of obtaining the output s(t) with no GW signal is the same as obtaining a noise shaped like s(t). Therefore,

$$P(s|H0) \propto e^{-(s,s)/2}$$

Very conveniently, if a GW signal of $h(t,\theta)$ is present in the output signal s(t), we have the noise shaped like $s(t) - h(t,\theta)$. Hence,

$$P(s|H1) = P(s - h(\theta)|H0) \propto e^{-(s-h,s-h)/2}, \tag{2.22}$$

where H_1 is the hypothesis that a GW signal $h(\theta)$ is present in s. This is a very pretty result and will be used in later sections for many times.

Another important definition is the signal-to-noise ratio, which measures how strong a signal is compared to its noise. We will use SNR to normalise our waveform later and take the high SNR limit (A tends to infinity). Its expression is:

$$A = (h,h)^{1/2}.$$

2.3.2 The frequentist path

In Section 3.1, we proved that the unbiased MLE achieves CRB. Hence if we are to carry out GW parameter estimation in the frequentist framework, it follows that the inverse of Fisher matrix is the covariance matrix corresponds to the frequentist GW parameter distribution. However this is only true under the assumption that a complete sufficient statistics exists. We have not proven that this condition is met (it's also difficult!). Vallisneri (2008) carried out a GW-specific derivation of the covariance matrix of $\theta^{\hat{\mathrm{ML}}}$ without direct referencing CRB theorem, confirming that the lower bound is actually met. One might question how a frequentist method can be ever applied to a singular GW event as we said before that principally singular events does not fit naturally to Frequentism. This is an example

of how a singular event can still enter the frequentist mathematics, but with an interpretation of limited significance. The inverse of FM only gives the frequentist error in the MLE, but not a probability distribution of where the true parameter lies as Bayesians could supply (Porter & Cornish, 2015). **In other words, no matter how good the data is (i.e. how small the inverse of FM is), the frequentist framework can only achieve precision, but not accuracy.**

In this GW-specific derivation, we add two assumptions: the limit of high SNR is achieved, and also that the noise is Gaussian. First we expand the waveform $h(\theta)$ around the true signal that we received $h_0 = s - n$ (noise is known), and normalised it by its SNR amplitude $A = \sqrt{(h_0, h_0)}$:

$$h(\theta) = h_0 + \theta_k h_{,k} + \theta_j \theta_k h_{,j,k}/2 + ... = A(\bar{h}_0 + \theta_k \bar{h}_{,k} + \theta_j \theta_k \bar{h}_{,j,k}/2 + ...) \quad (2.23)$$

where $h_{,i} = \partial_i h|_{\theta=0}$ after adjusting the source parameters such that it sets $\boldsymbol{\theta_0} = 0$.

MLE is defined by $\theta^{\hat{\mathrm{ML}}} = \arg \frac{\partial}{\partial \theta} \mathrm{P}(\mathrm{s}|\theta) = 0$ such that the estimated parameters are found at the peak of the likelihood. Substituting the expanded waveform 2.23 into the GW likelihood formula 2.22, we obtain at this solution of MLE:

$$\hat{\theta}_j^{\mathrm{ML}} = \frac{1}{A} \left(\bar{h}_{,j}, \bar{h}_{,k}\right)^{-1} \left(\bar{h}_{,k}, n\right) + \frac{1}{A^2}\{\cdots\} + \frac{1}{A^3}\{\cdots\} + \cdots \quad (2.24)$$

We omit writing out the terms with A factor at orders lower than -1, because these terms vanish in the high SNR limit (i.e. large A). From this expression it is also clear that taking the high SNR limit is the same as taking linearised-signal approximation (LSA). After taking the limit, we are only left with the first term in 2.24, which is simply $\hat{\theta}_j^{\mathrm{ML}} = \frac{1}{A} \left(\bar{h}_{,j}, \bar{h}_{,k}\right)^{-1} \left(\bar{h}_{,k}, n\right)$. We can also directly calculate the variance:

$$\begin{aligned} \left\langle \hat{\theta}_j^{\mathrm{ML}} \hat{\theta}_k^{\mathrm{ML}} \right\rangle_n &= \frac{1}{A^2} \left(\bar{h}_{,j}, \bar{h}_{,l}\right)^{-1} \left\langle \left(\bar{h}_{,l}, n\right) \left(n, \bar{h}_{,m}\right) \right\rangle_n \left(\bar{h}_{,m}, \bar{h}_{,k}\right)^{-1} = \\ &= \frac{1}{A^2} \left(\bar{h}_{,j}, \bar{h}_{,l}\right)^{-1} \left(\bar{h}_{,l}, \bar{h}_{,m}\right) \left(\bar{h}_{,m}, \bar{h}_{,k}\right)^{-1} = \frac{1}{A^2} \left(\bar{h}_{,j}, \bar{h}_{,k}\right)^{-1} \end{aligned} \quad (2.25)$$

The Fisher matrix defined in Eq. 2.9, once substituted with GW likelihoods Eq. 2.22, reduces to $F_{jk} = (h_{,j}, h_{,k}) = A^2 \left(\bar{h}_{,j}, \bar{h}_{,k}\right)$. Therefore, result 2.25 confirms that $\hat{\theta}^{\mathrm{ML}}$ actually achieves the CRB. Hence we conclude that, under the assumptions of high SNR and Gaussian noise, the frequentist error of the maximum likelihood estimator is obtained as the inverse of the Fisher matrix.

Immediately later, Vallisneri (2008) made an odd move in proving frequentist's Fisher matrix encounters another failure in GW signal analysis. He made an attempt of adding prior knowledge of the parameter distribution to our PE process, and showed that, despite its mathematical feasibility, the results disagree with our physical intuition. The motivation itself is strange in the sense that frequentist would not try to add prior knowledge to their statistics. So we may question the necessity of his proof below.

He first swtiched from MLE to maximum posterior estimator $\hat{\theta}^{MP}$ such that it contains prior information, which is a Bayesian move that should be questioned:

$$\frac{\partial}{\partial\theta_i}P(\hat{\theta}^{MP}|s) = \frac{\partial}{\partial\theta_i}P(s|\hat{\theta}^{MP})p(\hat{\theta}^{MP}) = 0$$

Suppose a Gaussian prior distribution $p(\theta) \propto \exp\left(-P_{ij}(\theta_i - \theta_i^P)(\theta_j - \theta_j^P)\right)$ centring at θ^P, we find modified $\hat{\theta}^{MP}$ to be biased (bias is the expected value of $\hat{\theta}^{MP}$ and because we have adjusted the parameters such that in the unbiased case the expected value is zero):

$$b_i^{\mathrm{MP}} = \left\langle\theta_i^{\mathrm{MP}}\right\rangle = \left[\left(\bar{h}_{,i},\bar{h}_{,j}\right) + P_{ij}/A^2\right]^{-1}\left(P_{jk}/A^2\right)\theta_k^P. \tag{2.26}$$

We can also write out the covariance matrix, whose shape agrees with CRB:

$$C_{ij} = \frac{1}{A^2}\left[\left(\bar{h}_{,i},\bar{h}_{,k}\right) + P_{ik}/A^2\right]^{-1}\left(\bar{h}_{,k},\bar{h}_{,l}\right)\left[\left(\bar{h}_{,l},\bar{h}_{,j}\right) + P_{lj}/A^2\right]^{-1} \tag{2.27}$$

The result is correct in the sense that in the high SNR limit (very large A), we recover the unbiased ML estimator (i.e. $b_i^{\mathrm{MP}} = 0$). However, the behavior is not desirable in the low SNR limit (when there's almost no signal), as we expect the error converging to the effective width of the prior. Instead, in Eq.2.27 the covariance matrix tends to zero in the low SNR limit when A is very small. This implies that in the frequentist analysis, the statistics does not recover prior information even when no signal is present. However we might question whether this inconsistency comes from the inconsistency of boldly merging two opposite paradigms, but not within Frequentism itself. We will later not use this argument against applying Frequentism in GW, but stick to the idea of how computing accuracy of one measurement is a priority over computing precision of one measurement. If we want to remain faithful to one school of probability only, Bayesianism is more 'useful' in the context of GW research, or similar research studying singular events.

2.3.3 The Bayesian path

In the high SNR limit where the signal is strong and pure, we assume that the best-fit (posterior) parameter distribution will be a near Gaussian distribution centred on the true value (Cutler & Flanagan, 1994). The Guassian-posterior assumption is phrased as an assumption with weak necessity as the high SNR limit has already implied it to some extent. But in case the high SNR limit isn't sufficient to 'Gaussianised' the posterior distribution, it is safer to introduce it as a separate assumption. It turned out that, with the assumption that the posterior distribution is near Gaussian (which I supposed is what makes the result coincides with BVM theorem), we can describe the shape of such distribution by a covariance matrix, which is given by the inverse of FM. In other words, GW researchers on the Bayesian path aims at fast approximation of the real posterior probability distribution by a Gaussian distribution with its shape described by the inverse of FM.

In the following section of Vallisneri (2008), it is proven that, in a single experiment satisfying the high SNR limit, the fast approximation of posterior distribution using FM is valid. The exact derivation is long and tedious. Hence we will only give a summary of the crucial steps in the derivation. Notice also that the Bernstein–von Mises theorem is not mentioned at all in his derivation due to the fact that GW emission is a single event. We can either conclude that this is simply a mathematical coincidence or that by setting the prior as constant, the results have already converged to how they should be after the infinite number of observations because of uniform priors and the Gaussian-posterior assumption.

Starting from substituting the expanded waveform 2.23 into the GW-specific likelihood formula 2.22. Now, with this substituted $P(s|\vec{\theta})$, we will be able to compute the following integrals, which gives the covariance matrix of the estimated values of the parameters given that **any prior probabilities for the parameters are considered as constant over the parameter range**:

$$< \theta_i >= \int \theta_i P(s|\vec{\theta}) d\vec{\theta} / \int P(s|\vec{\theta}) d\vec{\theta} \tag{2.28}$$

$$< \theta_i \theta_j >= \int \theta_i \theta_j P(s|\vec{\theta}) d\vec{\theta} / \int P(s|\vec{\theta}) d\vec{\theta}. \tag{2.29}$$

The two integrals are not easy to calculate due to the complicated form of $P(s|\theta)$. The crucial step here is that we take the high SNR limit, expand $P(s|\theta)$ around

$P(s|\theta_0)$, leaving only the the first derivative. Then we are left with:

$$< \theta_i >= (\bar{h}_i, \bar{h}_j)^{-1}(n, \bar{h}_j) * A, \tag{2.30}$$

$$C_{ij} =< \theta_i \theta_j >= (h_{,i}, h_{,j})^{-1}, \tag{2.31}$$

whose inverse is the Fisher matrix $F_{ij} = (h_{,i}, h_{,j})$. Here FM gives a measure of uncertainty instead of the frequentist error.

One clarification on his derivation is that prior distribution $p(\theta)$ is neither explicitly shown in Eq. 2.28 or Eq. 2.29, because $p(\theta)$ is considered as constants and therefore cancels out in the fraction, leaving likelihoods alone in the integral. Many (or most) implementations of FM in GW studies use the uninformative, constant-valued prior. Some examples of prior distributions in GW research are: the prior of the sky location of a binary black hole merger is chosen such that it weighs each patch of sky as equally probable; the prior of the primary black hole mass is uniform over a reasonable region; the prior distribution the energy density of primordial gravitational waves is log-uniform as we do not know the order of magnitude of some quantity (Thrane & Talbot, 2019). In the end, it often results in priors being completely cancelled out like we see in Eq. 2.28 or Eq. 2.29 where only likelihoods are effectively present in the calculation. If using FM for fast approximation of posterior distribution relies on this cancellation, we might question if FM methods fall into likelihoodism, where the degree of evidential support for each hypothesis should be analysed and measured in terms of likelihoods alone. When being concerned about likelihoodism, we should be aware that parameter estimation is intrinsically a hypothesis testing process, in which the hypothesis of $\theta = \theta_0$ is compared among all the other hypotheses where θ takes a different value. Hence the danger of falling into likelihoodism should be a common problem in all kinds of Bayesian hypothesis testing scenarios where the priors are 'uneffective' in the calculation. We will postpone discussing whether the act of setting uniform priors disqualifies the statistician from being faithfully Bayesian until the next chapter when both priors of parameter-value hypotheses, and priors of physical-principle hypotheses have made their appearance. The general idea is that a constant prior distribution still carries prior beliefs and is not actually 'uneffective' in the analysis.

As a comparison to Frequentism, Vallisneri (2008) then showed that a non-constant prior (so an informative prior) might be added to the computation without unphysical implication, unlike the frequentist case. Multiplying the likelihood $P(s|\theta)$

with $p(\theta) \propto e^{-P_{ij}\theta_i\theta_j/2}$ in the integral, we can get:

$$\begin{aligned} \left\langle \bar{\theta}_i \right\rangle_p &= \left[\left(\bar{h}_i, \bar{h}_j \right) + P_{ij}/A^2 \right]^{-1} \left(n, \bar{h}_j \right), \\ \left\langle \Delta\bar{\theta}_i \Delta\bar{\theta}_j \right\rangle_p &= \left[\left(\bar{h}_i, \bar{h}_j \right) + P_{ij}/A^2 \right]^{-1}. \end{aligned} \tag{2.32}$$

The results do make sense in the low SNR limit: variance of the posterior tends to the variance of the prior. Informative prior seamlessly makes modification to the Bayesian FM, as they should do.

2.4 Conclusion

In this chapter, we introduced how parameter estimation is done in the Bayesian framework, in comparison to the frequentist framework. For singular events like GW emissions, it is more natural to carry out statistical inference in Bayesianism as the prior distribution can be uncontroversially updated by one piece of evidence. In comparison a singular observation can still enter the frequentist mathematical scheme but with questionable significance of results. We also conclude that as the Bayesian recipe book requires richer ingredient (i.e. the priors), the main meal's statistical content is also richer. Fisher information (matrix) is a metric of 'how informative is our statistical setup about the parameter being at this particular value'. Fisher information in frequentist framework provides a measure of precision (i.e. error of MLE) of a single measurement while in Bayesian framework it gives a measure of accuracy (i.e. the shape of the posterior distribution).

3 | Hypothesis comparison

One of the main goal of GW research is to test whether General Relativity (GR) is correct. This process involves model comparison, in which we compare GR model against other non-GR models. One clarification that I want to make here is that though in many published work this process is called 'model selection', we should be aware that only model comparison is possible. We are not selecting the correct model, but only giving preferences to the model with a higher posterior belief (or higher evidential support, see later discussion). It is very likely that neither model involved in the comparison is true.

We will first introduce the general template of Bayesian model comparison, then see how tests of GR are carried out based on this template. The majority of this chapter then will focus on a philosophical evaluation of this model comparison scheme.

Bayesian model comparison is ultimately a comparison between the posterior beliefs for two models. For this we need to compute the **posterior odds ratio**, which is defined by the ratio of the posteriors of seeing the signal s under model M_1 and M_2:

$$O_{\mathrm{M}_1,\mathrm{M}_2} = \frac{p(\mathrm{M}_1|s)}{p(\mathrm{M}_2|s)} = \frac{p(s|\mathrm{M}_1)p(\mathrm{M1})}{p(s|\mathrm{M}_2)p(\mathrm{M2})} \tag{3.1}$$

The posterior odds ratio relies on the crucial ratio called **Bayes factor** K, which is defined by the ratio of the likelihoods of seeing the signal s under model M_1 and M_2

$$K_{\mathrm{M}_1,\mathrm{M}_2} = \frac{p(s|\mathrm{M}_1)}{p(s|\mathrm{M}_2)} \tag{3.2}$$

The Bayes factor compares how well each model predicts the signal data. Evidence presented by data is given by the impact of the data on the evaluation of a theory. Bayes factor is one of the relative measures of evidence when it comes to model comparison. In fact, $p(s|\mathrm{M}_1)$ is sometimes called *evidence* in statistics textbooks. Our prior opinion about each model is updated simply by multiplying $K_{\mathrm{M}_1,\mathrm{M}_2}$ with the prior odds ratio $\frac{p(\mathrm{M}_1)}{p(\mathrm{M}_2)}$. Of course the prior odds ratio also affects value of O_{12}. However we normally set the prior odd ratio to 1 for two models that appear equally likely before observation (Thrane & Talbot, 2019), and this is almost always the case in GW research. [1]

In the case of flat prior ratio, Dudbridge (2023) summaries the following interpretation of Bayes factor from a series of statistical works:

[1] The value of 1 is usually chosen for single event analysis. For multiple-event analysis, we can take the prior odds ratio to be the posterior odds ratio calculated from the previous event.

K_{M_1,M_2}	Jeffreys (1998)	Kass and Raftery (1995)	Royall (1997)
0-3	Bare mention	Bare mention	
8			Fairly strong
10	Substantial		
20		Positive	
32	Strong		Strong
>100	Decisive		
150		Very strong	

Table 3.1: "Bare mention" means no preference should be given to any models. "Strong" means the evidence strongly prefers M_1 and so on.

There is no agreement on an objective quantitative interpretation of Bayes factor, but most agree at the level of order of magnitude. Here is one case of hypothesis comparison in B. P. Abbott et al. (2019):

> H_0: the emission immediately after GW170817 is pure noise
> H_1: the emission immediately after GW170817 is another GW signal
> Bayes factor: 256.79
> Conclusion: H_0 is preferred.

3.1 Tests of GR

One type of tests of GR is the parameterised tests of GR, in which we examine whether the signal deviates from GR by having a different value of fixed parameter in the waveform predicted by GR. Therefore, we are comparing the model GR with models allowing a deviation from GR. Below is the template of parameterised tests of GR.

We begin with the Fourier-domain waveform of early inspiral phase in post-Newtonian (PN) approximation (details of physics are not important for later discussion), which is derived from GR Yunes, Yagi, and Pretorius (2016):

$$\Phi_{\rm GR}(f) = 2\pi f t_c - \phi_c - \frac{\pi}{4} + \sum_{i=0}^{7} \Phi_i(f), \tag{3.3}$$

where

$$\Phi_i(f) = \frac{3}{128\eta} \mathfrak{u}^{i-5} \phi_i \tag{3.4}$$

where $\mathfrak{u} = (\pi M f)^{1/3}$, M is the red-shifted total mass, η is the symmetry mass

ratio. We call each $\Phi_i(f)$ the $i/2$-th PN order term. For the deviation model at $i/2$-th PN order, we add a deviation parameter $\delta\phi_i$, which leads to a phase correction in the waveform of R. Abbott et al. (2021)

$$\Delta\Phi_i(f) = \frac{3}{128\eta}\mathfrak{u}^{i-5}\delta\phi_i. \tag{3.5}$$

Therefore, for the deviation model of $i/2$-th PN order, we have the waveform:

$$\Phi_{i,\mathrm{non-GR}}(f) = 2\pi f t_c - \phi_c - \frac{\pi}{4} + \sum_{j=0}^{7}\Phi_j(f) + \Delta\Phi_i(f), \tag{3.6}$$

There is no need to explain the physics for the scope of this discussion. All we need to know about these deviation models is that they now include $\Delta\Phi_i(f)$ as a free variable with values not fixed at zero (as GR model always has $\Delta\Phi_i(f)$ as a constant of zero).

3.2 Bayes factor

Time to calculate the Bayes factor. Now suppose we observed the GW signal $s(t)$, and we are interested in whether the signal passes tests of GR at the 1PN order. In other words, we are comparing the hypothesis GR with the hypothesis that there is a deviation at 1PN order from GR. M_1 represents each model allowing the waveform deviations from GR at 1st PN order coefficients, denoted as $\mathrm{M}_{1\mathrm{PN}}$, and M_2 represents GR. It is a common practice to assume uniform priors $p(\vec{\theta}|\mathrm{M}_i) = \prod_\alpha p_{i,\alpha}$ under each model M_i, where α runs through every parameter θ_α in model M_i, and $p_{i,\alpha}$ are assumed to be **constants** of 1 divided by the length/volume of the allowed domain (this assumption will be reviewed in the next section). The likelihood is evaluated for every possible parameter value, weighted by its prior probability. With GW likelihood model derived in the last section, we can obtain:

$$\frac{P(x\mid \mathrm{M}_1)}{P(x\mid \mathrm{M}_2)} = \frac{\int p(\theta|\mathrm{M}_1)P(x\mid\theta,\mathrm{M}_1)\mathrm{d}\theta}{\int p(\theta|\mathrm{M}_2)P(x\mid\theta,\mathrm{M}_2)\mathrm{d}\theta} = \frac{(2\pi)^{N_{\mathrm{p}_1}/2}e^{-\hat{\chi}^2_{\mathrm{M}_1}/2}\prod_\alpha p_{1,\alpha}}{(2\pi)^{N_{\mathrm{p}_2}/2}e^{-\hat{\chi}^2_{\mathrm{M}_2}/2}\prod_\alpha p_{2,\alpha}}\sqrt{\frac{|\Sigma_1|}{|\Sigma_2|}} \tag{3.7}$$

where $\hat{\chi}^2_{\mathrm{M}_i}$ and $\Sigma_{\mathrm{M_i}}$ are numbers calculated from complicated functions of $g(t)$, noise, priors of parameter $p(\theta)$ etc., and there is no need going into details of these two functions as they do not contribute to the philosophical argument.

Simplifying Eq.3.7, I have derived the Bayes factor for $\mathrm{M}_{1\mathrm{PN}}$ against M_{GR} in a

recent publication ([Author is me.], 2023) (citation anonymised).

$$\begin{aligned} K_{\mathrm{M_{1PN}},\mathrm{M_{GR}}} &= \frac{P(g(t)|\mathrm{M_{1PN}})}{P(g(t)|\mathrm{M_{GR}})} \\ &= p_{\delta\phi_2}\sqrt{2\pi}\frac{e^{-\hat{\chi}^2_{\mathrm{M_{1PN}}}/2}}{e^{-\hat{\chi}^2_{\mathrm{M_{GR}}}/2}}\sqrt{\frac{|\Sigma_{\mathrm{M_{1PN}}}|}{|\Sigma_{\mathrm{M_{GR}}}|}}, \end{aligned} \tag{3.8}$$

where $p_{\delta\phi_2}$ is the prior probability for the 1PN deviation parameter $\delta\phi_2$. This extra factor comes from the ratio $\prod_\alpha p_{1,\alpha}/\prod_\alpha p_{2,\alpha}$ as a result of $\mathrm{M_{1PN}}$ having one extra free parameter representing deviation at 1PN order compared to $\mathrm{M_{GR}}$. We notice that, only when the domain of $\delta\phi_2$ contains one signal value, $p_{\delta\phi_2} = 1$. Otherwise, $p_{\delta\phi_2} < 1$. The more flexible we allow $\delta\phi_2$ to be, the lower the value $p_{\delta\phi_2}$ is. This is how **Occam's razor** manifests in Bayes factor, and we will discuss it in detail in later sections.

Now we are going to examine some worries of using Bayes factor for model comparison.

3.2.1 Likelihoodism?

The choice of flat prior ratio inevitably leads the Bayesian inference into *Likelihoodism*. Sober (2008) puts Likelihoodism in this way: you don't ask if the evidence raises, lowers or leaves unchanged the hypothesis' probability, but only compare the determinate likelihoods. In our case of GW research, when hypothesis testing is based on the Bayes factor, which ultimately compares the likelihood of data under different models, the philosophy of probability might be thought to deviate from Bayesianism into Likelihoodism. Likelihoodism is founded on two claims: the likelihood principle (LP) and the law of likelihood (LL).

Roughly speaking, LP is the claim that all the evidence about the model is contained in the likelihood. As we can see, in the calculation of Bayes factor (or even the entire Bayesian inference), data s only enters the statistics through $p(s|\theta)$ (θ is the model parameter(s)), by $p(s) = \int p(s|\theta)p(\theta)d\theta$ and $p(s|H) = \int p(s|\theta)p(\theta|H)d\theta$. In fact, Bayesians have actually been among the strongest advocates for LP (Royall, 1997).

Now we are certain that model comparison based on Bayes factor is at least likelihoodist on the LP level. This is not a striking conclusion as LP is where Bayesianism and Likelihoodism overlap because it is simply an objective fact that data is only mathematically present in the likelihood. The more interesting question is

whether model comparison by Bayes factor also follows LL. LL is a rule to interpret data and compare models, and it is not acknowledged by many Bayesians (Lindley, n.d.):

> Law of likelihood: The observations O favor hypothesis H_1 over hypothesis H_2 if and only if $P(O|H_1) > P(O|H_2)$. And the degree to which O favours H_1 over H_2 is given by the likelihood ratio $P(O|H_1)/P(O|H_2)$.

This 'likelihood ratio' in law of likelihood is simply the Bayes factor. With only LP and LL, a strict likelihoodist has to neglect priors in their inference. The result is that a strict likelihoodist can only conclude how each model is favoured by the observation, but not by themselves as statisticians. So is model comparison by Bayes factor an Likelihoodist act?

Two attitudes towards likelihoodism need to be distinguished, and I will argue that Bayesians with the second attitude cannot be said to be strictly likelihoodist. On top of LP and LL, they also endorse adding prior information (even though the prior distribution is uniform!), which results in a whole different interpretation of the same mathematics. GW researchers have chosen the Bayesian interpretation (i.e. GR is favoured by us, the physicists!), but not the likelihoodist interpretation (i.e. GR is favoured by the data).

The first attitude is to reject Bayesianism completely, and only carry out model comparison by law of likelihood even when priors are available. Likelihoodist inference is fundamentally different from Bayesianism as what you get out of it is only the likelihood function but not posterior function which is a combination of evidence (computed by likelihood function) and prior beliefs. Statisticians with this attitude believe that scientific inference should not include anyone's unconditional beliefs, and hence they get rid of subjectivity, which is the Achilles heel's of Bayesian inference. This strict likelihoodist interpretation left us with: the model with higher likelihoods should be preferred because it is favoured by the observation.

The second attitude is to maintain the strict Bayesian position that 'the result of my model comparison will be settled by the posterior probability', but accept that it is mathematically equivalent to the results of endorsing only the law of likelihood. For example, prior odds ratio $\frac{p(\mathrm{M}_1)}{p(\mathrm{M}_2)}$ is taken as 1 in GW astronomy as it is difficult and meaningless to agree on an exact, unfair number to represent the prior beliefs within the community. Hence inevitably researchers are left with the likelihood ratio (Bayes factor) and nothing else, and do later inference with this only. I believe that researchers with attitude of the second kind cannot be said

to be unfaithful to Bayesianism. They did go through every step to be taken by Bayesians. In fact, only with the second attitude, GW researchers would be able to compute Bayes factor, especially in the case where the inference is multi-leveled. In this case, we have (1) prior of each model $p(\mathrm{M}_1)$ and (2) the prior distribution of parameter given each model $p(\vec{\theta}|\mathrm{M}_i)$. Despite seemingly using LL as the decision rule for GR vs. non-GR model comparison after obtaining the Bayes factor, prior information about parameter distribution was folded in during the calculation of Bayes factor. This is evidence that GW researchers are still Bayesians. Recall that in Eq. 3.7, we have assumed a prior distribution $p(\vec{\theta}|\mathrm{M}_i) = \prod_\alpha p_{i,\alpha}$ (they are conditionalised but are still priors since a waveform model does not tell you about the distribution of free parameters), which cannot be removed from the Bayes factor calculation. This assumption wouldn't be available to strict likelihoodists who wouldn't accept any use of priors and they have to stop at the early stage of parameter estimation. As Joyce (2021) agrees, cases where hypotheses deductively entails a definite likelihood of the data are rare (which excludes parameterised tests of GR), so strict likelihoodists of the first kind have to adopt a theory of probability with very limited applications.

However in section 2.3.3 we have seen that the majority prior distributions in GW research are uniform. In those cases, likelihoods start to be the only mathematical determinant at a more fundamental level. The likelihoodism worry pointed out in this section might be stretched as from setting indifferent priors and having them cancelled out, physicists lose the important subjective element of Bayesianism, leaving us with only likelihoods in the calculation. **To completely resolve this worry, we now argue that there does not exist any indifferent, uninformative prior, and also that the essence of Bayesianism does not lie in the apparent form of mathematics, but in how we interpret the mathematics.** Back to the example of [Author is me.] (2023), $p(\theta|\mathrm{M}_1)$ is taken to be constant in Eq. 3.7 in the hope of setting up a"neutral" prior distribution to eradicate the subjectivity of prior opinions. But is this prior distribution really neutral? It might seem like a uniform prior distribution is the best choice we should go for when there is no information about the value of θ. However as Royall (1997) points out, the reason that any Bayesians efforts to effort pursue non-informative priors ultimately fails is that "pure ignorance cannot be represented by a probability distribution":

> Every probability distribution represents a particular state of uncertain knowledge; none represents the absence of knowledge. (Royall, 1997)

Therefore, a uniform prior does not say "we are ignorant of how likely each value of θ is". Instead, it says "we have the same degrees of belief about each

value of θ . Any distribution that can be written out carries information. The worry of priors does not stop at $p(\theta|M_1)$, and it continues to exist in $\frac{p(M_1)}{p(M_2)}$. We do not actually have any non-subjective ground to set the prior odds ratio to 1 instead of 1e-09, and it will greatly affect the value of posterior ratio. In fact, I consulted a LIGO group and it turned out that they were not happy about other setting prior ratio to 1, and are currently setting their own priors based on how well GR has been doing in the past. **The fact that the freedom to set this prior ratio proves that GW researchers still interpret the mathematical value of Bayes factor as the comparative degrees of posterior belief we have for two competing models, but not just simply the comparative evidential support the data gives to each model. This interpretation but not the mathematics is what makes GW data analysis Bayesia**n. We have seen similar coincidence of mathematics but difference in interpretation earlier in previous chapter where FM can be mathematically derived from both Frequentism and Bayesianism with different interpretations.

So despite the appearance of Bayes factor in LL, model comparison in the case of tests of GR is only partially likelihoodist as both Bayesians and likelihoodists endorse LP and LL. Bayes factor (given the prior ratio being 1) cannot be interpreted as 'the comparative posterior beliefs in one hypothesis over the other hypothesis' from solely LP and LL without adding a prior distribution reflecting your subjective beliefs.

3.2.2 An objective interpretation of Bayes factor

The title of this subsection might be misleading in the way it does not refer to the debate between objective and subjective Bayesianism. This section discusses the practice of objectively interpreting the value of Bayes factor, under subjective Bayesianism.

Acknowledging that the Bayes factor is sensitive to prior of parameters (Morey, Romeijn, & Rouder, 2016), it brings an element of subjectivity into any follow-up inference. We should be happy about the existence of subjectivity as it is the inevitable feature of Bayesian statistics that completes Bayes theorem equation. Bayes factor does not introduce any more subjectivity than what is principally needed in Bayesianism.

While setting up priors does not introduce more subjectivity than what is already there in Bayes theorem, the interpretation of the values of Bayes factor seems to do so (but wait for the twist in the next subsection). The second worry arises from Table. 3.1. Bayes theorem itself only computes the posterior distribu-

tion, and it does not offer any criteria on when we decide on which model beats the other. If there is no grounding for the numbers derived in Table. (3.1), then we can say 200 is still too small to indicate strong evidence preferring one model. The disagreement among statisticians originates from the vagueness of natural English: what does it mean for evidence to be "strong" anyways? Is "strong" evidence enough for us to claim that we have detected a GW signal rather than pure noise, or do we need "decisive" evidence for that?

I believe that any attempts to find a best, objective interpretation of Bayes factor would fail because natural language isn't defined according to Bayes theorem. **Bayes factor can stay as what it is as the number already says everything it contains.** The interpretation is only needed when making an announcement to the public that says "we have detected a gravitational wave!". Asking how high the Bayes factor should be to justify this announcement is like asking how low the plane crash rate should be for me to take a flight. Is a Bayes factor of 200 high enough to make a detection claim? Is a crash rate of 0.001% low enough to travel by flight? You can push the limit indefinitely and it will never be enough.

If we stick to the idea that the value of Bayes factor has already said everything it could possibly say, then we have the third worry of whether Bayes factor can be used for tests of GR, or it can allow us to believe in any theory. As Morey et al. (2016) puts it: a model having the highest Bayes factor means nothing more than that the model had the highest amount of evidence in favor of it out of the models currently under consideration. All Bayes factor can do is model comparison, and it could be that all available models fit badly to the data (Gelman & Rubin, 1995). In the case of parameterised tests of GR, if we say that we have proven GR is true, then we must have already assumed one of the models (GR and all GR deviation models) is true. It could be the case that none of the these tests can lead to the conclusion that GR is the correct theory, but only the best theory we can currently think of. Nevertheless we can argue that this is the limit of science, rather than exclusively the limit of model comparison using Bayes factor.

Royall (1997) suggests that an objective, quantitative understanding of Bayes factor[2] can be based on a simple scenario in which the intuition about strength of evidence is strong. We are given two urns containing either all white balls, or half white balls and half black balls. One of these two urns is randomly picked, and the

[2]In his likelihoodist paradigm, the likelihood ratio. But they share the same mathematical definition at this point, it is only later likelihoodists disagree with Bayesians on how the likelihood ratio/Bayes factor is calculated.

competing hypotheses are H_0: this urn is the white-ball-only urn; and H_1: this urn is the half-white-ball urn. Now we draw x balls from the urn, and found out that these x balls are all white balls. The Bayes factor of H_0 over H_1 is $1/(\frac{1}{2})^x = 2^x$. Now Royall's intuition drawing three white balls in a row is fairly strong evidence that the urn contains white balls only. Three white balls $x = 3$ corresponds to a Bayes factor of 8.

This way of objectively interpreting Bayes factor fails in three ways. First, different people have different intuitions on how many white balls can convince them that the urn contains only white balls. For example, I am only convinced by a row of at least 5 white balls, which corresponds to a Bayes factor of 32. So subjectivity does not just affect the interpretation of exact value of Bayes factor, but also the interpretation of the order of magnitude (32 does not have the same order of magnitude as 8). Secondly, suppose that the punishment is getting killed if the guess is wrong, and there is no limit on how many draws I can have before making the guess. Then I will at least take 50 draws, which corresponds to a Bayes factor of 2^{50} to believe that there is 'strong' evidence for H_0 against H_1, if 'strong' means that I can almost make the guess. Lastly and most importantly, why should the context of drawing balls from urns be the template of interpretation of the numerical value of Bayes factor in other contexts? Suppose we have a very noisy GW signal, Bayes factor for GR-deviation models over GR might easily get to 8 under such big noise just because it's easier to fit the data with more degrees of freedom, but it doesn't mean that we have strong evidence for GR-deviation models[3]. This signal is so noisy such that if GR-deviation models are winning, they should win by a lot! Royall (1997)'s response to this problem is that, BTU (unit of energy) in physics is defined as the amount of energy required to raise the temperature of one pound of water by 1 degree. And it doesn't mean that it is meaningless to rate air conditioners using BTU where no water is involved. Likewise, a likelihood ratio of k corresponds to evidence strong enough to cause a k-fold increase in the prior odds ratio $\frac{p(\mathrm{M}_1|s)}{p(\mathrm{M}_2|s)} = \frac{p(s|\mathrm{M}_1)p(\mathrm{M1})}{p(s|\mathrm{M}_2)p(\mathrm{M2})}$. But this analogy hardly supports how a Bayes factor of 8 is always 'strong' evidence in whichever scenario. Suppose an AC which cools down a huge office space using 20000 BTU is rated as 'high efficiency', then it is definitely ridiculous to rate an AC which cools down a small room using 20000 BTU as 'high efficiency' as well. It is indeed meaningless to interpret a value of BTU/Bayes factor without specific context of where it is used.

Therefore, we still stand by the point that Bayes factor says all it can say by the value (**for now!**). Any unique, 'objective' interpretation of it would fail (**for**

[3] In practice the noise will be carefully dealt with and reduced.

now!). In the next subsection Royall (1997) is ready to point out what he thinks goes wrong in our reasoning and we might change our mind after his counter-argument.

3.2.3 Evidence $\neq$ confidence

In Royall (1997), he talks about one seeming disaster of model comparison using Bayes factor, and he attributes this apparent disaster to the distinction between evidence and degree of belief, but not misusing Bayes factor to quantify evidence. The main idea is that a Bayes factor of 8 implies strong evidence in whatever cases, but strong evidence does not lead to strong confidence in the winning hypothesis. All our previous worries are towards 'is a Bayes factor of 8 enough to make me believe in this, or act this way?', but not 'is a Bayes factor of 8 represents that the relative evidential support is strong?'.

The disastrous scenario looks like this: we shuffle an deck of 52 playing cards and turn over the top card, which turns out to be the ace of diamonds. The Bayes factor of 'this deck is a trick deck only containing ace of diamonds' over 'this is a normal deck' is 52. According to Royall's map from numerical values of Bayes factor to strength of evidence (and almost all the others), a Bayes factor of 52 is more than strong evidence that this is a trick deck. But this doesn't feel right. It is completely normal for the top card to be the ace of diamonds. Whatever the top card turns out to be, there is always a corresponding trick deck that wins over the normal deck hypothesis. Royall's response to this is simply that **the observation might be strong evidence in favour of a trick deck, but it is not strong enough to over come the prior probability of a normal deck**[4]. We are only shocked by a Bayes factor of 52 because we (unconsciously) believe that a trick deck is rarer while making the calculation. As Edwards (1970) commented that a Martian faced with this problem would find the first hypothesis most appealing when seeing that all playing cards share the same pattern on the back, and wouldn't be surprised to find them sharing the same pattern on the front. There isn't anything counterintuitive in saying that the evidence for trick-deck hypothesis is stronger than the evidence for normal-deck hypothesis.

Our argument in the before subsection ignores this distinction between degrees of beliefs and strong evidence. Ignoring the distinction between subjective probability (prior and posterior probability) and objective probability (likelihood) is also

[4] Although Royall himself is an likelihoodist and the book is defending an likelihood paradigm, this attack seems to be only available to Bayesians. Later we will refer to this objection as 'Royall's' but it doesn't mean we agree with the likelihoodists

not what a faithful subjective Bayesian should do. People with different intuitions, under different contexts are subject to evidence of the same strength when Bayes factor is the same. It is our fault to wrongly take evidence as confidence in the first hypothesis.

Let's re-describe our previous counterexamples against his objective interpretation template.

1. Given that two urns look the same (so rationally my prior ratio of H_0 over H_1 is 1), I won't have good confidence in H_0 even if I draw 3 white balls in a row because I am hard to convince, in which case **there is 'Royall-strong' evidence** for H_0 over H_1.
2. Given that two urns look the same (so rationally my prior ratio of H_0 over H_1 is 1), I won't have good confidence in making a guess to determine my life unless I have determining confidence in H_0. I also won't have determining confidence in H_0 unless I draw 50 white balls in a row in which case there's determining evidence for H_0 over H_1. 'Royall-strong' evidence is not strong enough.
3. Given my prior belief that GR is more likely[5] and , I won't have confidence in GR-deviation models **even if there is 'Royall-strong' evidence (Bayes factor>8)** for GR-deviation models against GR model.

What would Royall say about the revised counterexamples?

1. Your confidence, which is the posterior probability would be 1*8=8. But you might not call a confidence of 8 'good' confidence.
2. Suppose you now had 49 draws of white balls. Your confidence in H_0 is 2^{49}, but you might not call a confidence of 2^{49} 'determining' confidence, which then does not give you 'good' confidence in taking a guess to determine your life.
3. Suppose your prior is set to 1/8, by multiplying it with evidence of 8, your posterior probability is only 1, which should correctly not give you any confidence.

There are three different process going on in our examples: (1) map the numerical value of Bayes factor to strength of evidence, (2) map the numerical value of posterior probability to strength of posterior belief, and (3) map the strength of posterior belief to actions. When we say that no objective interpretation of Bayes

[5]Because I anticipate that non-GR might just win by goodness of fit to the noise so to balance this out (some unconscious preference of GR here) I set up a high prior probability for GR.

factor is justified, it should refer to the interpretation maps (2) and (3), but not (1) which is what Royall and other's object template of interpretation of Bayes factor is targeted on.

Lastly we might still raise an objection to Royall's objective interpretation scheme. **It does not seem like the strength of evidence affect either our belief or decisions. Isn't it meaningless to speak of 'strong' evidence when 'strong' does not result in a strong belief or a direction to act?** Then we might agree that his interpretation scheme does not guarantee any counter-intuitive consequences, but it also leads to **no consequences at all**.

3.3 Occam's penalty

It seems that introducing more free parameters to the original model will always better fit the data (or at least as well as the original model). Why can't we just add as many free parameters as possible to the original model? Consider the following example in Jefferys and Berger (1992): suppose we are doing an experiment determining the object's equation of motion during free fall, and the hypotheses to be compared is (1) $s = a + ut + \frac{1}{2}gt^2$ and (2) $s = a + ut + \frac{1}{2}gt^2 + bt^3$. Knowing that the data won't fit the true theory perfectly because of the real motion is only approximately in free fall due to air friction, the cubic law (2) should be a better fit of the data than the quadratic law (1). In fact equations which go to higher-degree polynomials will always reduce the total error. So why do we stop at the quadratic term? Returning to tests of GR, we can always construct a model such that it fits the signal better than GR just because it has more free parameters. We can imagine how, a signal contaminated by noise, can fit better into this example model compared to GR because of these extra degrees of freedom. So why do we still go for GR instead of its more flexible counterparts?

Occam's razor is usually the answer to this kind of questions. Extra free parameters should not be introduced unless they are 'necessary'. It debatable what does it mean for the extra structure to be 'necessary' (necessary for what?), but the general idea is that apart from goodness of fit, simplicity is also a key criterion of model comparison. A qualitative answer to why simplicity matters can be the following: the more complicated the model is, the less it can say in general. In the case of determining law of motion during free fall, cubic law might be able to better fit the current set of data. This fit comes at the price of fixing a, u, g, b at $a = 1, u = 1, g = 10, b = 1$ for example. While the quadratic law, despite more residual errors to the real data, gives $a = 1, u = 1, g = 9.8$ as its best-fitted solution.

Suppose we then repeat the same free falling experiment under different environments (e.g. in air of different density, or some very tiny wind blowing in different directions). We expect the quadratic law to fit diverse data a lot better than the cubic law. So the cubic law is only preferred by the data of some specific free falls (e.g air of some specific density or wind blowing at some particular directions), and it will be a poor predictor of new data compared to the non-excessive quadratic law.

As both simplicity and goodness of fit are crucial values of a good hypothesis, the question becomes if they are both manifested in Bayes factor. Later we will find out that the answer is yes. However we are not saying that the correct hypothesis testing process has to take into account of both features, especially simplicity. It could be that fit is much more important than simplicity because scientific laws are contingent description of the world. With careful expectation, we will be only showing that Bayes factor penalises complexity of the hypothesis model and rewards goodness of fit. It is also not saying that this specific quantification of Occam's penalty and goodness of fit given by Bayes factor is the correct, unique way of quantifying the balance between the two, but only that it is the quantification you get once choosing Bayesian statistics.

It is often said that Bayes factor encodes the notion of Occam's razor. In Eq. (3.8), $p_{\delta\phi_2}$ is always less 1 if the domain contains more than a single value. This is not the unique case. Generally if we are to test whether a parameter δ takes a particular value, we can compute the Bayes factor of $H_0 : (\delta = \delta_0, \theta)$, where δ is fixed to a single value, against $H_1 : (\delta, \theta)$ which allows any values of δ. θ is the set of parameters that characterise the system and not been tested. This Bayes factor is called the **Savage-Dickey ratio**. In the case of parameterised tests of GR, H_0 is the GR model as the deviation parameter $\delta\phi_i$ is fixed to zero, and H_1 is the non-GR model allowing deviation at i-th PN order as it assumes $\delta\phi_i$ to be a free parameter of the waveform just like other real θ (e.g. chirp mass, distance, symmetry mass ratio etc.). Hence we have:

$$H_0 : (\delta\phi_i = 0, \theta), H_1 : (\delta\phi_i, \theta) \tag{3.9}$$

The Savage-Dickey ratio is computed as:

$$K_{H_0,H_1} = \frac{p(s|H_0)}{p(s|H_1)} = \frac{p(s \mid \delta\phi_i = 0, H_1)}{p(s \mid H_1)} = \frac{p(\delta\phi_i = 0 \mid s, H_1)\, p(s \mid H_1)}{p(\delta\phi_i = 0 \mid H_1)\, p(s \mid H_1)} = \frac{p(\delta\phi_i = 0 \mid s, H_1)}{p(\delta\phi_i = 0 \mid H_1)} \tag{3.10}$$

Now suppose that GR is the true theory of this universe, which means that s is composed of some noise, and a signal of GW where $\delta\phi_i$ takes a fixed value of 0. This will encourage $p(\delta\phi_i = 0 \mid s, H_1)/p(\delta\phi_i = 0 \mid H_1)$ to increase because the signal s which obeys GR, together H_1 is going to better predict $\delta\phi_i = 0$ than what H_1 can do by itself. Remember that s is a real signal, and we have assumed that the true theory is GR, so s carries the restriction $\delta\phi_i = 0$ to some extent. The restriction will not be at full strength (i.e. $p(\delta\phi_i = 0|s) < 1$) because s is the pure GW signal contaminated with noise. The presence of the noise will discourage $p(\delta\phi_i = 0 \mid s, H_1)/p(\delta\phi_i = 0 \mid H_1)$ to increase because $\delta\phi_i = 0$ rejects impurity and therefore $\delta\phi_i = 0$ will reduce the fit between model and noisy signal. Savage-Dickey ratio reveals the balance between Occam's penalty and improvement of goodness of fit in Bayes factor. This should be seen as an advantage of model comparison by Bayes factor. In summary:

model comparison criterion	effects on K_{H_0,H_1} given impure s
Goodness of fit	decrease
Simplicity	increase

As smoothly as everything agrees mathematically, there are few important philosophical questions that are left unanswered:

1. What is our notion of simplicity[6] exactly, if we take it as the opposite of Occam's penalty in Bayes factor?
2. Should a good account of model comparison consider both simplicity and goodness of fit? Can't it be that one of them is always superior?
3. Suppose that these two features should be components of a good account of model comparison, does Bayes factor provide the correct quantification?

We will try to give a response to these questions respectively in each subsection below.

3.3.1 Simplicity and precision

In the last section we take simplicity as a *mathematical* attribute of models such that **a model with fewer degrees of freedom (i.e. fewer free parameters), or smaller domain[7] of free parameters (i.e. less free parameters) is simpler.** But

[6]Here we only talk about parametric simplicity as it is what's relevant to Bayes factor model comparison. For other kinds of simplicity (e.g. computational simplicity), see Rochefort-Maranda (2016)

[7]More precisely, the region of values of the parameter where the prior probability isn't negligible.

what is it qualitatively? We first compare it with Lewis' definition of simplicity, in which simplicity is an attribute of a deductive system with few, short sentences. Does our mathematical capture of simplicity agree with this? A model with fewer parameters (suppose that it can be rewritten into Lewsian deductive system) does correspond to fewer or shorter sentences, as the system would not say anything about the extra parameters (e.g. a shorter expansion of GW waveform that does not contain deviation terms). However it is unclear how a smaller domain corresponds to a shorter axiom. The sentence 'The domain of parameter ϕ is (0,1)' is as short as the sentence 'The domain of parameter ϕ is (0,2)'.

Our mathematical definition of simplicity neither fits perfectly to Popper's characterisation (Karl, 2002), in which simplicity equates falsifiability. Simpler models have more potential falsifiers. It might seem that models with a smaller domain allowed for free parameters are indeed easier to be falsified as there are more 'forbidden' values of parameter that can lead to rejection of the model. However there are no 'forbidden' values in some or even most models, but only values assigned with a tiny possibility in the prior probability distribution. Hence definite falsification would not be possible. Models with fewer free parameters are also not more falsifiable for the same reason. We might correctly think that the mean square error between data and model's prediction is larger when there are fewer degrees of freedom, but this does not lead to more falsifiability. As models in Bayesian inference would not be rejected or accepted *by principle*, no mean square error, no matter how large, can lead to falsification of a model. However, undeniably our simplicity is somehow Popperian if we replace falsification with a less intense concept of disconfirmation. If a theory is more likely to be disconfirmed (i.e. more 'forbidden' values), it is simpler. Our notion of simplicity is closer to falsificationism than Lewis' account in the sense that the mathematical simplicity captures some ontological facts about the model relating to how sensitive the model is to falsification or disconfirmation. Lewis' account, on the other hand, treats simplicity as a measure of how elegant the model is in terms of its representation (e.g. fewer, shorter sentences in a deductive system).

What can be directly derived from our notion of simplicity is the volume of parameter space of a model, and then link this volume with the predictive power possessed by this model . We can construct the parameter space of a model, of which the number of dimensions equates the number of free parameters, and the extension on each dimension corresponds to how spread-out the prior distribution of each value is. The spread-out of the prior distribution generally aims to characterise how free the parameter is in its own dimension. It might be difficult to quantify this, but one idea could be using the 'effective length' of the prior distribution, which is the length of the region where the prior probability isn't negligible (and hence a model with the parameter wandering between (0,1) would be

more spread-out than a model with parameter that takes value of either 0 or 100). The notion of simplicity embedded in Bayes factor corresponds to the volume of parameter space for different models. We have added Fig. 3.1 to show how this might seem like: each circle represents a free parameter, and each free parameter is free to slide within the region corresponds to the 'effective' length.

We hope to make the claim that **the simpler the model is (i.e. the smaller the volume of parameter space is), the higher the precision the model possesses (i.e. the more precise the prediction yielded by the model is).** It is more accurate to replace the name of 'simplicity' with 'parametric parsimony', but for simplicity (pun intended), we will stick to referring to it as 'simplicity'. In the illustration, the simpler the model is, *with the same input*, the more certain we are about the output. Suppose we are predicting the waveform of GW from known chirp mass, symmetric mass ratio, and all rest of the parameters that are required to determine a GR-GW waveform (these are not drawn on the illustration for simplicity). Take M3 as GR, M3 will predict exactly one single solution of waveform. On the other hand, M1: a GR-deviation model allowing a narrow range of deviations, and M2: a GR-deviation model allowing a wide range of deviations, can only predict the space in which the solution lies **because the value of parameter 3 (i.e. $\delta\phi$ deviation parameter at i-th PN order) is unphysical (therefore unknown) and can be adjusted**. The difference between predictions of M1 and M2 is that the solution space for M1 is smaller. The last step is to compare each model's **precision**, and we will use the following rule: **given the same input information, the model with smaller solution space has higher precision** This rule makes sense because precision is linked to error of results, and mathematically a wider solution space[8] means higher error of where the solution actually lies. Notice that at this stage we are only discussing the precision (how precise the model is about its prediction), but not accuracy (how correctly the model predicts). We can now infer that these models are ranked as following by precision: M3>M1>M2, which is the same as the order of our notion of simplicity. To wrap up: higher simplicity (as manifested in Bayes factor) implies higher precision of solution that can be achieved by the model.

3.3.2 Simplicty vs. Goodness of fit

Simplicity and goodness of fit are two good virtues of models. But maybe one of them is superior to the other? Now we will try to argue that neither is superior to the other, and hence both must be included in a good account of model comparison scheme. Bayes factor is a good metric for hypothesis comparison as it embeds

[8]Like the domain, the solution space is the region where the likelihood isn't negligible. Smaller solution space gives each solution more certainty if the prior distribution of parameter 3 is uniform.

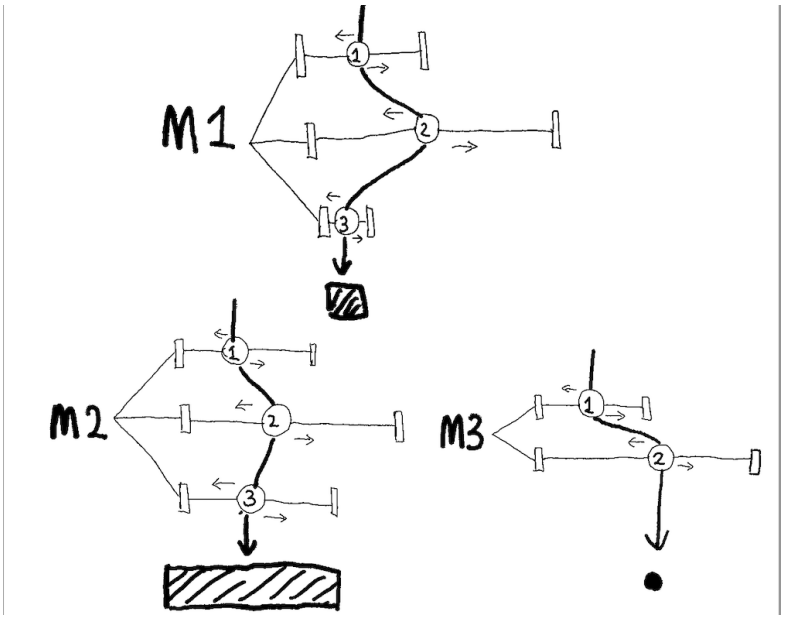

Figure 3.1: An illustration of models with different simplicity. Suppose we give the same input to these three models, the precision of prediction is less when the parameter space is larger. The illustration has assumed all priors of parameters to be flat for simplicity.

both aspects mathematically.

The battle between simplicity and goodness of fit might be equivalent to the tension between the model's predictive precision and accuracy. While simplicity offers a small solution space, goodness of fit makes sure that the solution is correct in the sense that it matches the reality. However we should also notice how goodness of fit is only approximately aiming at accuracy, but not exactly. The important fact is that real data never perfectly represent truth, but only an approximation of truth. Real data is always to be contaminated by some noise, so in fact, the correct model should never perfectly fit the data. The problem of overfitting occurs when the model fits to the noise, rather than the pure signal. Take an extreme example. Suppose we have n data points, then we should always choose the model that is a polynomial of degree $(n-1)$ (in the form of $\sum_{i=0}^{n-1} a_i x^i$), because mathematically for n data points this model is bound to achieve perfect fit. Suppose we live in a windy Newtonian world, and we collected 100 data points of the trajectory of a particle is in uniform motion, then without dispute the best model should be the one that requires inputting values of two parameters (speed and location) only to determine its trajectory. However if accommodation to data is the superior criterion then the best model should be the one that requires inputting values of 99 parameters. To avoid this ridiculous consequence accommodation to data cannot be superior to simplicity. This is why we seem to contradict to the previous claim that accuracy is superior to precision as the goal of science because science should aim to track to the truth, when comparing how Fisher matrix is useful in the frequentist and Bayesian framework. But now goodness of fit does not stand on the higher ground, as improvement of goodness of fit does not lead to improvement of accuracy.

Consider what would happen if simplicity is always superior to goodness of fit. The simplest (i.e. most predictively precise) model could just be: there is only one GW waveform which is of this shape, regardless of anything (e.g. chirp mass, distance from the merger to earth etc.). The waveform of next merger can be predicted with no input information, but the prediction will just be wrong in terms of having a negligible likelihood given the model. If part of the aim of science is to yield predictive success, then priority shouldn't be given to models just because they have less measurement uncertainty in their results. Simplicity has to be sacrificed to some extent such that the model still give approximately correct predictions.

To sum up, the counterbalance between simplicity and goodness of fit is exactly what's needed for a good account of model comparison.

3.3.3 Quantification problem

Though we said that both simplicity and goodness of fit matters in terms of selecting the best model, this does not justify Bayes factor as giving the best (or a good) quantification of these two properties in a model. One popular choice is the Akaike Information Criterion (AIC) (Akaike, 1974; Hitchcock & Sober, 2004):

> AIC $= 2k - 2\ln(L)$, where k is the number of freely adjustable parameters in the model, and L is the maximum of likelihood. The model with minimum AIC is preferred.

The model with minimum AIC is expected to have the best combination simplicity (so number of free parameters is as small as possible) and goodness of fit (so the maximum of likelihood is as high as possible). The maximum of likelihood is taken as a metric of goodness of fit as it implies how well the hypothesis can possibly fit the data. Bandyopadhyay and Forster (2011) tried to justify that AIC is applicable in both frequentist and Bayesian paradigm:

> **AIC in Frequentism** AIC is a frequentist construct in the sense that AIC provides a criterion, or a rule of inference, that is evaluated according to the characteristics of its long-run performance in repeated instances.
> **AIC in Bayesianism** What Forster and Sober have done in their chapter is to show that Bayesians can regard AIC scores as providing evidence for hypotheses about the predictive accuracies of models. A key point in their paper is to point out that this difference of evidential strength between hypotheses about predictive accuracy can be interpreted in terms of the law of likelihood (LL), which is something that Bayesians can accept. According to the LL, observation O favors H1 over H2 if and only if P(O|H1) > P(O|H2). (Bandyopadhyay & Forster, 2011)

Then we won't worry about whether AIC is not better than Bayes factor as a model comparison metric on the principle level. It is clear that the choice between AIC and Bayes factor can be put as a pure quantification problem under Bayesianism. Model comparison by AIC is different from model comparison by Bayes factor in the sense that simplicity gained due to a smaller number of free parameters is doubly rewarded. A comparison of AIC already contains a comparison between likelihoods such that Bayes factor appears in the difference between AICs of two models. While Bayes factor already takes care of simplicity once, it is counted twice in the difference between k. Can we justify for this double counting? Remember we discussed that how Bayesian model comparison by Bayes factor only is justifiable if the prior ratio of competing hypotheses can be set to 1. The double

counting might be understood as setting the prior ratio according to the simplicity of the model in terms of how many free parameters it has. If we are faithful subjectivist about Bayesianism, then AIC and Bayes factor function are equally good model comparison metrics because we accept that priors can be whatever subjective beliefs you have.

3.4 Is GR intrinsically better by intention?

In the calculation of Bayes factor, we have treated GR and GR-deviation models as equal competitors. However there might be reasons to prefer GR model just because it is generated to predict GW rather than to accomodate existing GW data.

Consider two scientists, the Predictor and the Accommodator. They have acquired the same sets of data and develop theories based of this data. The predictor formulates her theory based on the first set of data D1, and use her theory Tp to make predictions about the remaining data D2, which turned out to be successful. On the other hand, the accommodator formulates her theory Ta based on all data sets deliberately just to accommodate all the data. In parameterised tests of GR, GR would be the model formulated by the predictor (i.e. Einstein) and GR-deviation models would be formulated by the accommodator. The deviation parameters are solely introduced to better fit the data but do not represent any extra physical consideration. Now Hitchcock and Sober (2004) asks this important question: Does the fact that Tp predicted D2 whereas Ta was designed to accommodate this data give us reason to believe that Tp, is the better theory? Hitchcock and Sober (2004) argues that Tp is indeed the better theory as " there is some tendency for those (accommodator's) theories to be defective in ways that can be assessed independently of the intentions of theorists". The argument goes like this: suppose we draw more samples from the same population, with resepect to the new data D3, Tp is going to do better than (e.g. gives a higher likelihood to D3) than Ta because Tp has been predictively successful before about D2. This is what the tendency from formulating theories with the intention of prediction has brought us. Tp is preferred because it indeed has better predictive success, but not just because it's formulated from higher scientific moral grounds. This argument is also more than induction. Tp tends to fit a more diverse group of data because formulating theories for predictive success but not for accommodation inhibits it from over-fitting. As noise is random, Ta would not do well when new samples are drawn because the noise it fits to has changed.

But now the question becomes: if we acknowledge that GR is intrinsically better than GR-deviation theories by its construction, do we need to do anything extra

on the calculation show our preference? Not really, because Bayes factor has already penalised over-fitting, and the prior odds ratio (if not taken as one) can take consideration of GR's previous predictive success.

3.5 Conclusion

In this chapter, we have introduced parametric tests of GR, a key motivation for observing gravitational waves. These tests involve comparing hypotheses using the Bayes factor, specifically contrasting the GR waveform model against GR-deviation models. We addressed the worry that hypothesis comparison by Bayes factor might fall into Likelihoodism. We concluded that, while the mathematics is shared with Likelihoodism, it remains fundamentally Bayesian in in the sense that the relative probability is still interpreted as relative degree of beliefs, and a model is still favoured according to statistician's posterior beliefs but not likelihoods. Further, we argued that any objective interpretation of the Bayes factor—as either a measure of relative belief in one hypothesis over another or as relative evidential support given to one hypothesis over another–would be meaningless. We then looked into why the Bayes factor is a good metric for hypothesis comparison, as it rewards both goodness of fit and simplicity. We defined the concept of simplicity within the Bayes factor as "parametric parsimony" which is closely linked to a model's predictive precision, while goodness of fit relates to predictive accuracy but only approximately, due to the fate of over-fitting. Additionally, we explored the relative importance of simplicity versus goodness of fit, claiming that neither is inherently superior. We also introduced an alternative metric AIC for hypothesis comparison that differently quantifies these two critical aspects, and claimed that both AIC and Bayes factor are justified metric to be used in Bayesian inference. The last section briefly discussed the popular idea of how theories could be intrinsically better by its motivation, and concluded that Bayes factor has already folded in this knowledge.

4 | Gravitational-wave cosmology

On August 17th, 2017, the LIGO-Virgo GW detector network observed a GW signal (GW180817) consistent with a binary neutron star merger (Abbott et al., 2017b). This was the loudest signal at that time (with a signal-to-noise ratio of 32.4), and results in a precise sky position estimation. Most importantly, a γ ray burst (GRB 170817A) was immediately observed 1.7 seconds after the coalescence time estimated for this GW signal, which is located at the sky position estimated by the GW data. This further supports the identification of this GW event as a neutron star merger. As the first multi-messenger observation event which includes GW signal, it opened up a new world of multi-messenger astrophysics. As one of its contribution to the broader astrophysics, we are going to see how GW180817 and GRB 170817A together will help to constrain the Hubble constant (Abbott et al., 2017a). This is equivalent to treating H_0 as a free variable in the Hubble-law cosmological model, and do parameter estimation of H_0.

Suppose v is the recessional velocity due to the expansion of the Universe ($v = cz$ where z is the redshift of the source, inferred from the γ ray observation), d is the luminosity distance of the source inferred from the GW observation, the Hubble constant H_0 is defined by the Hubble law as:

$$v = cz = H_0 d. \tag{4.1}$$

However when the redshift is small (i.e. v is small), the peculiar velocity of individual astronomical object adds on to the observed recessional velocity by a non-negligible amount. If v_t is the true recessional velocity of a particular galaxy, we have the relation $v_t = v + v_p$. Hence more usefully:

$$v_t = H_0 d + v_p. \tag{4.2}$$

Now the task is to calculate the posterior distribution of the value of H_0 given observations: (1) x_{GR} the GW signal, (2) v_o the observed recessional velocity of the source directly calculated from the γ ray data, and (3) $< v_p >= 150 \mathrm{km} s^{-1}$ which is the standard error of the peculiar velocity[1] at the location of galaxy NGC4993, identified by the γ ray. Our goal is to calculate $\mathbf{p}(\mathbf{H_0}|\mathbf{x}_{\mathrm{GR}}, \mathbf{v_r}, < \mathbf{v_p} >)$, which involves computing $p(x_{\mathrm{GR}}, v_r, < v_p > |H_0)$. However any likelihood of x_{GR} depends on marginalising over the prior of parameters creating the waveform like d. Therefore we need to build a hierarchical model, in which the likelihood for data depends on parameters whose depends on other unknown parameters (called hy-

[1] The measured smoothed peculiar velocity field

perparameter, or population parameter as it is usully the parameter of a population rather than an individual source), for which we write down another prior.

4.1 Hierarchical models

The hierarchical model specific to this case is hierarchical in the following way:

(a) The GW data $x_{\rm GW}$ (observable) depends on the waveform of the source, which depends on the source parameters like d, inclination etc. (hyperparameter/population parameter),

(b) The observed recessional velocity v_o (observable) depends on the true recessional velocity v_t (individual parameter; v_o does not trivially equal to v_t as there exists uncertainty of electromagnetic measurement), which then depends on the true peculiar velocity v_p (hyperparameter/population parameter) and the Hubble recessional velocity $H_0 d$ (hyperparameter/population parameter),

(c) The measured smoothed peculiar velocity field $< v_p >$ (observable) at the host galaxy depends on the true peculiar velocity v_p (hyperparameter/population parameter).

As an example, we can write out the corresponding likelihood expression for (a) (Abbott et al., 2017b):

$$p\left(x_{\rm GW} \mid d, \cos\iota\right) = \int p\left(x_{\rm GW} \mid d, \cos\iota, \vec{\lambda}\right) p(\vec{\lambda}) \mathrm{d}\vec{\lambda} \tag{4.3}$$

where ι is the inclination angle of the source to our detector, d is the luminosity distance from the source of GW signal to the detector, and $\vec{\lambda}$ are other source parameters that we choose the marginalise over. Why are d and ι special among other source parameters? This is a very technical detail but it's worth explaining to avoid confusion. Luminosity distance d is singled out because the likelihood of v_o, $p(v_o|d, v_p, H_0)$ in (b) is conditionalised on d. In this slightly complicated hierarchical model, the likelihoods of two observables $x_{\rm GW}$ and v_o, depends on the same, unknown parameter d. Hence if we leave this shared unknown parameter conditionsed (not marginalised) here, then later it will be natural to combine the likelihoods of two observables into one, as[2]

$$p(v_o, x_{\rm GW}|d, v_p, H_0, \cos\iota) = p(v_o|d, v_p, H_0) * p(x_{\rm GW}|d, \cos\iota). \tag{4.4}$$

[2]Assuming $x_{\rm GW}$ does not depend on H_0, v_o does not depend on inclination

Now why is inclination separated from other parameters? Because there is still debates about what the priors for ι should be within the community, so it is easier to leave it in the equation such that it can be replaced with alternative priors. The debate on what the prior should be will be discussed in a later section. Another reason for why $\cos\iota$ is treated differently from other source parameters is that the estimation of d from the GW signal is significantly affected by the estimation of $\cos\iota$ (which makes sense because they are both description of location in space). Hence d and $\cos\iota$ always stick with each other on the same side of the equation, either both being the condition, or being conditionalised.

4.2 Selection effects

One of the important thing to account for in hierarchical modelling is the **selection effect**. Suppose we are to calculate likelihood $p(x|\lambda,\text{detectbility})$ where x stands for data, λ stands for a hyperparameter, and detectable stands for the decision to include only *detectable* events is going to affect the normalisation of $p(x|\lambda,\text{detection})$.

$$p(x|\lambda,\text{detectble}) = p(x|\lambda)/\mathscr{N}(\lambda), \mathscr{N}(\lambda) = \int_{\text{detectable}} p(x|\lambda)dx \tag{4.5}$$

Why are we making this selection? Because many combination of parameters might produce pretty GW waveforms, but they are just not detectable due to the sensitivity limit of current detectors. These waveforms' amplitudes will be buried in noise and therefore not recognized confidently from the raw data.Therefore given that the current GW signal at hand is *detected*, the integral does not need to include the undetectable events. This selection changes the normalisation factor, and therefore affect value of likelihood.

Now I am going to argue that, the selection effect only manifest when the inference is about the hyperparameters (e.g. the distance distribution for all galaxies as it is linked to the galaxy density of the universe), and it does not kick in when the inference is about parameters of an individual source (e.g. the chirp mass of a particular merger event).

If inference is done about the hyperparameters, then $p(\lambda|x,\text{detectable})$ would be the object to compute:

$$p(\lambda|x,\text{detectable}) \propto p(x|\lambda,\text{detectble})p(\lambda) = p(x|\lambda) * p(\lambda)/\mathscr{N}(\lambda). \tag{4.6}$$

On the other hand:

$$p(\theta|x,\text{detectable}) = \int p(\theta|x,\lambda,\text{detectable})d\lambda \tag{4.7}$$

$$\begin{aligned} p(\theta|x,\lambda,\text{detectable}) &= p(x|\theta,\text{detectble})p(\theta|\lambda,\text{detectble})p(x)/\mathcal{N}(\lambda) \\ &\propto \{p(x|\theta)/\mathcal{N}(\theta)\} * \{p(\theta,\text{detectable}|\lambda)/\mathcal{N}(\lambda)\} \\ &\propto \{p(x|\theta)/\mathcal{N}(\theta)\} * \{p(\text{detectable}|\theta,\lambda)p(\theta|\lambda)\} \\ &\propto \{p(x|\theta)/\mathcal{N}(\theta)\} * \{p(\text{detectable}|\theta)p(\theta|\lambda)\} \\ &\propto \{p(x|\theta)/\mathcal{N}(\theta)\} * \{\mathcal{N}(\theta)p(\theta|\lambda)\} \\ &\propto p(x|\theta)p(\theta|\lambda)\text{Nice!} \end{aligned} \tag{4.8}$$

Notice that in the second inference, $N(\lambda)$ are ignored because it does not contain information about θ so it is bound to be marginalise over. The result is nice in the sense that no selection effects factor $N(\theta)$ appears in the posterior distribution of θ, which means selection effects get cancelled out in inference about θ.

Specific to the context of GW research, an event is detectable if the signal-to-noise (SNR) ratio is larger than 8 (Abbott et al., 2017a). SNR ratio ρ characterise the strength of the GW signal compared to noise if the GW signal were to happen *in theory*, so it's a description of detectability of individual GW signal given knowledge of noise of current detectors. However this cutoff at 8 does not seem to originate from further assumptions but is only an agreement within the GW community. This is very similar to the scenario where a Bayes factor of 8 is interpreted as 'strong evidence', as discussed in the last section. Interestingly SNR ratio is present in the Bayes factor of the hypothesis H_0: This signal s is pure noise such that $s = n$; over the hypothesis H_1: This signal s contains a gravitational wave signal h such that $s = n + h$. Recall some GW statistical results from the first section that (Creighton & Anderson, 2011):

$$p(s|\mathrm{H}_0) \propto e^{-(s,s)/2}, \tag{4.9}$$

$$p(s|\mathrm{H}_1) \propto e^{-(s-h,s-h)/2}. \tag{4.10}$$

The Bayes factor is:

$$\frac{p(s|\mathrm{H}_1)}{p(s|\mathrm{H}_0)} = e^{-(s,h)/2}e^{(h,h)/2} \tag{4.11}$$

where we have omitted the detail of how constants before two likelihoods cancel

out. In previous section we have obtained that $\rho = (h,h)^{1/2}$[3], which happens to be the exponent on one of the factors.

Now it is very tempting to say that this detection cutoff is set at 8 because in Bayesian hypothesis comparison, 8 is interpreted as 'strong evidence', but it cannot be so as SNR ratio is in the exponent and $e^8 \approx 2981$. Additionally there is also the other factor $e^{-(s,h)/2}$ in the Bayes factor. Hence I tend to believe that there is no good explanation of this SNR ratio cutoff at 8 in terms of Bayes factor, since a GW signal h with SNR ratio larger than 8 is not guaranteed to have a Bayes factor of detection larger than 8. SNR ratio is only a measure of the GW signal's detectability, but not whether the real signal s (which contains the same GW signal) is preferred as a detection. Then what should we say about this SNR ratio threshold set at 8? There are two comments I want to make:

Indeed cutoff of SNR ratio instead of cutoff of Bayes factor should be used We might question why shouldn't a cutoff of Bayes factor preferring detection be used in the selection process. It shouldn't be used because what the integral is trying to filter out is waveforms with low amplitude compared to noise (i.e. low detectbility) but not waveforms that are not preferred as a detection given the current s. No data should participate in this selection process like they do in the calculation of Bayes factor (see Eq. 4.11, because this selection is in the integral of likelihood of this very data.

Does this SNR ratio face the same interpretation difficulty as Bayes factor? Yes and No. The answer is yes when inference is made about the hyperparameters of the population (e.g. luminosity distance d), and no when the inference is about parameters of individual source. The cut-off being at 8, instead of at 10, or 6 is set by convention without much justification. This might bring us back to the discussion of hypothesis comparison by Bayes factor in the previous section, in which Bayes factor struggles to find a non-subjective interpretation of numerical values in terms of strength of belief, but at least the request for this interpretation is not as urgent. We can either only care about the order of magnitude of the value, or just state the value and leave it there. However, in the selection of detectable signals, an immediate, precise threshold of SNR ratio is required to carry out the integral to obtain likelihood. Worse than Bayes factor, as the detection threshold is introduced during the integral, even the mathematics comes out differently. However as we see above, inference about individual properties is unaffected by the selection effects, and therefore escapes the problem.

[3] You might ask where is noise in signal-to-noise ratio. Noise term is embedded in the definition of the inner product.

Finally let's look at how prominent the selection effect is in the case of multi-messenger astronomy. The SNR ratio cutoff for GW signal sets a cutoff for how large d could be for the GW event to be detectable, as d affects the amplitude of the signal a lot more than any other parameters. This SNR ratio then, defines a horizon for detection of LIGO/Virgo detectors, which is around 190Mpc (Abbott et al., 2017a). Nearly no GW events further than 190Mpc from the earth can be detectable and hence we simply integrate over the domain of d extending to 190Mpc. One might ask why there is no selection effect on the γ ray detection side. There is but it is much less restrictive: the horizon of EM detection is around 400Mpc. Therefore it can be ignored as events happened at further than 400Mpc is already selected out on the GW side.

4.3 Results

Figure 4.1 shows the resulting constraint on H_0 (Abbott et al., 2017a). We can see that the constraints are not particularly strong and good. Hopefully it can be improved once there are more multi-messenger detections!

Parameter	68.3% Symm.	68.3% MAP	90% Symm.	90% MAP
$H_0/\left(\mathrm{km\,s^{-1}\,Mpc^{-1}}\right)$	$74.0^{+16.0}_{-8.0}$	$70.0^{+12.0}_{-8.0}$	74.0^{+33}_{-12}	70.0^{+28}_{-11}

Figure 4.1: "Symm." refers to a symmetric interval (e.g. median and 5% to 95% range), while "MAP" refers to maximum a posteriori intervals (e.g. MAP value and smallest range enclosing 90% of the posterior)

5 | Conclusions

This thesis has carefully looked into the most crucial applications of Bayesianism in GW research: parameter estimation and hypothesis comparison.

For parameter estimation, I have provided and compared the frequentist and Bayesian recipe books. The advantages of using Bayesian methods in GW research is (1) the posterior distribution is principally significant even by updating with a singular observation unlike the case in Frequentism; (2) the result produced by Bayesian method is more fruitful as the statistical content in the Bayesian ingredient is more abundant with the addition of a prior distribution. In particular, I have looked into the application of Fisher matrix in both frameworks. As a commonly used tool for model evaluation, it gives an approximation of error of MLE in Frequentism, and an approximation of posterior distribution in Bayesianism. I have argued that the latter is a much more useful result as error of MLE from a single measurement is hardly significant.

Tests of GR is carried out as standard Bayesian hypothesis comparison. Bayes factor (i.e. ratio of likelihoods of observation given each of the competing models) is the metric commonly used for hypothesis comparison. Despite the sole appearance of likelihoods, the use of Bayes factor remains faithfully Bayesian instead of Likelihoodist. Priors getting cancelled out in the mathematics does not mean they have not participated in the inference because any priors carry information (even if they can be cancelled out!). Bayes factor used by GW researchers is still interpreted as relative posterior degrees of belief. Bayes factor is argued as a good metric for hypothesis comparison as it rewards simplicity (or parametric parsimony) and goodness of fit. Both of these two properties are crucial for the model's predictive success, as they together aim at predictive precision and accuracy. I have also criticised any attempt of giving an objective interpretation of Bayes factor. As any values of Bayes factor do not lead to a definite 'strong' belief of one hypothesis over the other.

Lastly I briefly discussed Bayesian hierarchical model and the selection effects that come along in gravitational-wave astronomy. In the case of constraining the Hubble constant with GW, I showed that selection effects only kick in in inference about population parameters, but not inference about individual properties.

Overall, I conclude that Bayesianism provides a satisfactory template for GW research from the aspect of philosophy of probability.

6 | List of Abbreviations

GW . . . . . Gravitational wave

GR . . . . . General Relativity

SNR . . . . . Signal-to-noise (ratio)

FM . . . . . Fisher matrix

MLE . . . . . Maxmium-likelihood estimator

AIC . . . . . Akaike information criterion

PN . . . . . Post-Newtonian (approximation waveform)

UMVUE . . . . . Uniformly minimum variance unbiased estimator

BVM . . . . . Bernstein–von Mises (theorem)

CRB . . . . . Cramer-Rao bound

LL . . . . . Law of likelihood

LP . . . . . Likelihood principle

Bibliography

Abbott, B. P., et al. (2019). Properties of the binary neutron star merger GW170817. *Phys. Rev. X*, *9*(1), 011001. doi: 10.1103/PhysRevX.9.011001

Abbott, R., et al. (2021, 12). Tests of General Relativity with GWTC-3.

Abbott et al. (2017a, November). A gravitational-wave standard siren measurement of the Hubble constant. *Nature*, *551*(7678), 85–88. Retrieved 2024-03-01, from http://arxiv.org/abs/1710.05835 (arXiv:1710.05835 [astro-ph]) doi: 10.1038/nature24471

Abbott et al. (2017b, October). GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral. *Physical Review Letters*, *119*(16), 161101. Retrieved 2024-03-01, from https://link.aps.org/doi/10.1103/PhysRevLett.119.161101 doi: 10.1103/PhysRevLett.119.161101

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. doi: 10.1109/TAC.1974.1100705

[Author is me.]. (2023, 11). Impact of overlapping signals on parameterized post-Newtonian coefficients in tests of gravity.

Bandyopadhyay, P. S., & Forster, M. (Eds.). (2011). *Handbook of the philosophy of science, vol. 7: Philosophy of statistics*. Elsevier B.V.

Chalmers, A. F. (1976). *What is this thing called science?: An assessment of the nature and status of science and its methods*. St. Lucia, Q.: Univ. Of Queensland Press.

Creighton, J. D. E., & Anderson, W. G. (2011). *Gravitational-wave physics and astronomy: An introduction to theory, experiment and data analysis*.

Cutler, C., & Flanagan, E. E. (1994, March). Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral waveform? *Physical Review D*, *49*(6), 2658–2697. Retrieved from http://dx.doi.org/10.1103/PhysRevD.49.2658 doi: 10.1103/physrevd.49.2658

Dudbridge, F. (2023, February). *A scale of interpretation for likelihood ratios and Bayes factors*. arXiv. Retrieved 2024-01-08, from http://arxiv.org/abs/2212.06669 (arXiv:2212.06669 [stat])

Edwards, A. W. (1970). Likelihood (letter to the editor). *Nature*, *227(5253):92*(1). Retrieved 2024-01-17, from https://www.jstor.org/stable/29774559

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in bayesian social research. *Sociological Methodology*, *25*, 165–173. Retrieved 2024-01-17, from http://www.jstor.org/stable/271064

Hitchcock, C., & Sober, E. (2004). Prediction versus Accommodation and the Risk of Overfitting. *The British Journal for the Philosophy of Science*, *55*(1), 1–34. Retrieved 2024-02-06, from https://www.jstor.org/stable/3541832 (Publisher: [Oxford University Press, The British Society for the Philosophy of Science])

Hájek, A. (2023). Interpretations of Probability. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2023 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2023/entries/probability-interpret/.

Jefferys, W. H., & Berger, J. O. (1992). Ockham's Razor and Bayesian Analysis. *American Scientist*, *80*(1), 64–72. Retrieved 2024-01-17, from https://www.jstor.org/stable/29774559 (Publisher: Sigma Xi, The Scientific Research Society)

Jeffreys, H. (1998). *Theory of probability*. Clarendon Press.

Joyce, J. (2021). Bayes' Theorem. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2021 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/bayes-theorem/.

Karl, P. (2002). *The Logic of Scientific Discovery*. Routledge.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *J. Am. Statist. Assoc.*, *90*(430), 773–795. doi: 10.1080/01621459.1995.10476572

Lindley, D. (n.d.). *Discussion of royall (1992) the elusive concept of statistical evidence* (Vol. 4). Oxford University Press. Retrieved from http://arxiv.org/abs/2212.06669

Ly, A., Marsman, M., Verhagen, J., Grasman, R., & Wagenmakers, E.-J. (2017). *A tutorial on fisher information*.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016, June). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. Retrieved 2024-01-08, from https://www.sciencedirect.com/science/article/pii/S0022249615000723 doi: 10.1016/j.jmp.2015.11.001

Porter, E. K., & Cornish, N. J. (2015, May). Fisher vs. Bayes : A comparison of parameter estimation techniques for massive black hole binaries to high redshifts with eLISA. *Physical Review D*, *91*(10), 104001. Retrieved 2023-11-19, from http://arxiv.org/abs/1502.05735 (arXiv:1502.05735 [astro-ph, physics:gr-qc]) doi: 10.1103/PhysRevD.91.104001

Rochefort-Maranda, G. (2016, May). Simplicity and model selection. *European Journal for Philosophy of Science*, *6*(2), 261–279. Retrieved 2024-02-16, from http://link.springer.com/10.1007/s13194-016

-0137-1 doi: 10.1007/s13194-016-0137-1

Royall, R. M. (1997). *Statistical evidence : a likelihood paradigm*. London: Chapman and Hall.

Samaniego, F. J. (2010). *A comparison of the bayesian and frequentist approaches to estimation*. Springer New York.

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science* (1st ed.). Cambridge University Press. Retrieved 2023-12-27, from https://www.cambridge.org/core/product/identifier/9780511806285/type/book doi: 10.1017/CBO9780511806285

Thrane, E., & Talbot, C. (2019). An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models. *Publications of the Astronomical Society of Australia*, *36*, e010. Retrieved 2024-01-16, from http://arxiv.org/abs/1809.02293 (arXiv:1809.02293 [astro-ph]) doi: 10.1017/pasa.2019.2

Vallisneri, M. (2008, feb). Use and abuse of the fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects. *Physical Review D*, *77*(4).

Yunes, N., Yagi, K., & Pretorius, F. (2016, October). Theoretical physics implications of the binary black-hole mergers gw150914 and gw151226. *Physical Review D*, *94*(8). Retrieved from http://dx.doi.org/10.1103/PhysRevD.94.084002 doi: 10.1103/physrevd.94.084002