Leibniz Universität Hannover and Max Planck Institute for Gravitational Physics (Albert Einstein Institute)

Bachelor's Thesis

Systematic Differences in the Source Properties of the Third Gravitational-Wave Catalog

Max Melching

Supervisor: Dr. Frank Ohme
Co-Supervisor: Angela Borchers

Contents

List of Figures

List of Tables

Acronyms

GW gravitational wave

GWTC gravitational-wave transient catalog

JSD Jensen-Shannon divergence

LVK LIGO-Virgo-KAGRA collaboration

NR numerical relativity

PC principal component

PCA principal component analysis

PDF probability density function

SNR signal-to-noise ratio

Mathematical Notation

a thin letters (lower and upper case) are used to denote scalar values and variables with scalar values

$\mathbf{a}$ bold, lower case letters are used to denote vectors

$\mathbf{A}$ bold, capital letters are used to denote matrices

$a^\mu b_\mu$ the Einstein summation convention is adopted, where repeated indices denote sums:

$$a^\mu b_\mu = \sum_\mu a^\mu b_\mu$$

Declaration of Authorship

I hereby assure that the thesis at hand has been constituted independently and without the use of any other than the cited sources. I furthermore assure that all passages taken textually or analogously from other sources are marked as such. This thesis, in its current or a similar form, has not been submitted to any other examination office.

Hiermit erkläre ich, die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Wörtliche oder sinngemäße Übernahmen aus anderen Quellen wurden als solche kenntlich gemacht. Auch wurde diese Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt.

.....	.....
Datum	Unterschrift

1. Introduction

Gravitational waves (GWs) have attracted much attention in the past years and decades, especially since their first detection on September 14, 2015 made by the GW detectors of the LIGO-Virgo collaboration (which is now called LIGO-Virgo-KAGRA collaboration or LVK). By the time this work is written (2022), almost one hundred events have been detected. Their detection is not publicly announced one-by-one, but collectively in gravitational-wave transient catalogs (GWTCs), where analyses of them are released as well. As the thesis title already implies, we will predominantly deal with the 36 events published in the third catalog GWTC-3 [1].

The process of finding such events involves comparing simulated signals or waveforms to the data and looking for the best-fitting one (we will see later in which sense "best-fitting" can be understood). There are multiple models to produce such waveforms and this is for example useful to ensure consistency of the whole inference process. Besides asking "how well waveforms fit the data" (a very relevant question deciding if a detection is claimed or not), the use of more than one model motivates and in fact requires to also ask the question "how well different models fit each other". There are several ways to quantify this waveform agreement and some of them were also used in previous GWTCs [2, 3]. A straightforward idea is to compare the signal morphologies themselves to see how well they match (which we will refer to as mismatch-studies), but it is also possible to use a strictly mathematical approach and measure distances between waveforms, e.g. using the Fisher-matrix formalism. However, we will not use any of those methods directly and instead take a different approach. With the number of detected signals approaching one hundred throughout three catalogs, it also becomes possible to make meaningful and systematic studies using experimental data. To be more precise, this work will focus on waveform agreement and general systematics found in the source properties, which are inferred from the signals themselves using different waveform models.

The reason that investigating such systematics is relevant at all is that due to the increased number of detections during the last LVK observing run O3 (and a further increase expected from the next run O4, which will start soon), the number of events with inconsistencies in source properties also increased. In principle, that could point to inconsistencies in the whole inference process because waveform models and their predictions should not differ too much when being used to analyse the same data (which contains one fixed signal). However, this is not necessarily true because there are known errors and biases that could also cause such inconsistencies, so investigating them and assessing their significance is important. While there are sections on the analysis of waveform systematics and consistency

in GWTC-3, they contain only brief discussions of these topics (and to our knowledge, no separate discussion either by the LVK or other authors was published, too).

However, it is certainly possible to discuss these topics in great detail and this forms one of the main motivations of this work. Consequently, the main objective of is to study waveform systematics. That includes identifying events where significant differences in the source properties can be observed, finding potential patterns in the population of events, quantify and assess how severe they are and look for potential reasons of them. Similar analyses can be conducted for individual parameters instead of events. This process will involve finding and developing proper criteria to measure differences in probability distributions and subsequently, finding ways to represent results obtained from applying these criteria to the source properties computed using different waveform models. Furthermore, a well-known technique called principal component analysis (PCA) will be applied to see if it enables us to find new, interesting statements in the context of waveform systematics.

2. Gravitational Waves

To understand what posterior samples are, the most relevant data for this work, some basic concepts have to be reviewed. That includes what GWs are and how they are detected.

2.1. Basics

2.1.1. Theory

The majority of equations from this section are taken from chapter 12 of [4] or from [5].

Although the primary focus of this work does not require broad knowledge of the derivation of GWs or other related theory, it is nonetheless helpful to have a basic idea of it (especially to understand how they are detected). GWs arise as a prediction from our current model of gravity given by Albert Einstein's general theory of relativity (GR). There are many great sources on this (e.g. [4], which has a great discussion of GWs), so rather than explaining every idea behind it, a basic understanding of GR is assumed. In GR, gravity is associated with the curvature of spacetime and as such, it can be described by the metric $g_{\mu\nu}$ of the 4D Minkowski space that represents spacetime. A metric is the mathematical tool that allows to define geometric quantities like distances by introducing an inner product on abstract vector spaces like the Minkowski space. GWs are local perturbations of this metric generated by astrophysical sources with a time dependent quadrupole moment. Because these sources have distances of astrophysical scale from Earth, the effect of GWs on Earth is expected to be very small. That corresponds to a small change in the metric $g_{\mu\nu}$, which means it can be approximated well by first order deviations from the flat Minkowski metric $\eta_{\mu\nu}$:

$$g_{\mu\nu} \approx \eta_{\mu\nu} + h_{\mu\nu} \,. \tag{2.1}$$

$h_{\mu\nu}$ is the influence of the GW and "small" perturbations correspond to $|h_{\mu\nu}| \ll 1$. By solving the Einstein vacuum equations for $h_{\mu\nu}$, one obtains in the transverse traceless gauge and for a propagation in z-direction:

$$h_{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & h_+ & h_\times & 0 \\ 0 & h_\times & -h_+ & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} e^{ik^\mu x_\mu} \,. \tag{2.2}$$

Here, $x^\mu = (t, x, y, z)$ are the coordinates chosen and $k^\mu = (\omega/c, k^1, k^2, k^3)$ is the wave vector. From now on, we will adopt the commonly used convention of discarding the imaginary part and use $\mathrm{Re}\left(e^{ik^\mu x_\mu}\right) = \cos\left(k^\mu x_\mu\right) = \cos(\omega(t - z/c))$. This form shows that GWs have, to first order, two polarizations with respective amplitudes $h_+, h_\times$.

GWs do not have an effect on the coordinates of resting particles (which is an effect caused by coordinates, this does not mean there is no physical effect). But it was already mentioned that GWs do have an effect on the metric, so they do affect the distance L of two test masses (point-like particles) sitting at $(L/2, 0, 0)$ and $(-L/2, 0, 0)$. This can be seen from

$$\tilde{L} = \int_{-L/2}^{L/2} \sqrt{g_{xx}}\, dx = \int_{-L/2}^{L/2} \sqrt{1 + h_{xx}}\, dx \approx \int_{-L/2}^{L/2} 1 + \frac{h_{xx}}{2}\, dx = L\left(1 + \frac{h_+ \cos(\phi)}{2}\right) \tag{2.3}$$

where we defined $\phi := \omega(t - z/c)$ for a clearer notation. Likewise, the distance between to particles at $(0, L/2, 0)$ and $(0, -L/2, 0)$ is

$$\tilde{L} \approx L\left(1 - \frac{h_+ \cos(\phi)}{2}\right). \tag{2.4}$$

Because the component h_+ acts on the x- and y-components, whose corresponding coordinate axes form a +, it is clear where the index comes from.

Writing out the effect of $h_\times$ components is slightly more complicated as these act on the xy-, yx-components (which again explains the index $\times$). Therefore, we do not give an explicit calculation and only the result. A quantity which summarizes the effect of GWs on the distance of test particles is the line element ds^2. When a GW is present and using components in the basis (t, x, y, z) (such that $g_{ij} := g_{x_i x_j}$ is the coefficient for $dx_i dx_j$), it reads

$$ds^2 = -c^2 dt^2 + (1 + h_+ \cos(\phi))\, dx^2 + (1 - h_+ \cos(\phi))\, dy^2 + 2h_\times \cos(\phi)\, dxdy + dz^2\,. \tag{2.5}$$

For comparison, the line element for flat Minkowski space reads

$$ds^2 = -c^2 dt^2 + dx^2 + dy^2 + dz^2\,, \tag{2.6}$$

so a GW causes non-zero off-diagonal terms $dxdy$ and also affects the diagonal terms dx_i^2. Although these effects are very small on Earth due to small GW amplitudes, it is this effect that is measured by current, ground-based GW detectors.

But before we focus on that, there are some more details worth mentioning. Actually, it is not only possible compute the effect of GWs or how they propagate, but also how GW signals from astrophysical systems look like according to GR.[1] Systems consisting of two compact objects (black holes or neutron stars) are particularly interesting because their coalescence (a *compact binary coalescence* or CBC) produces GWs to which GW detectors

[1] The last part is one reason why GW observations are relevant. If detected signals do look like the simulated ones, this is a confirmation of GR, which is by no means unimportant.

are highly sensitive. While it is possible to simulate such systems and extract GWs by solving the Einstein equations using numerical relativity (NR), this is computationally very demanding/ expensive. To be able to generate many templates faster, approximations like weak-field or low-velocity limits are used to model GW signals. There are several of these approaches, each leading to a *waveform model* (for details see e.g. section 6.5 of [6]). To ensure the approximations do not change the resulting signals too much, some models can be calibrated to NR simulations, i.e. fitting the signal shapes to reproduce them. How well this works highly depends on how many NR simulations are available in the respective region of the parameter space.

All models have in common that the waveforms generated from them (*templates*) are governed by a set of parameters θ, which characterize the signal source (for instance the component masses and spins of a binary system). Thus, there is a useful alternative description (parametrization) of the templates $h = h_\theta$ as a point $\theta \in \Theta$ where Θ is the corresponding *parameter space* of the source, which is the space spanned by all possible combinations of parameter values. By generating templates for many points all over this parameter space (many parameter values), it is then possible to build up *template banks* for waveform models. Furthermore, it is typical to have more than one waveform model from a single approach, some of them taking more complicated effects like precession into account, and together they form *waveform families*.

2.1.2. Detection

This subsection only gives a rough overview on the topic of experimental detection since other topics are more relevant for the objective of this work. A more detailed description is given in [7] or the GWTCs [2, 3, 8, 1].

The last subsection made it clear how we want to measure GWs: they change the distance between objects, so the idea is to measure the distance between two test masses and compare it to a reference which will enable us to see differences possibly caused by GWs.

The tool which turns out to be appropriate for that is the Michelson-Interferometer, which has two arms where lasers is sent through. Any change in the arm length is picked up by the light while it propagates through the arms and stored in its phase. This information can be read out from a photo diode by combining the light from the two arms, which allows to measure differential changes in the arm lengths. It is important to emphasize the *differential* because if both arms are affected equally, the same phase will be picked up and this will only show up as a phase shift. These are not measurable since we are free to add phases without changing physics. An important consequence is that not all GWs are measured equally well. Mathematically, this shows up as response functions which introduce a difference between emitted signals and actually measured signals

$$h = F_+ \, h_+ + F_\times \, h_\times \, . \tag{2.7}$$

The best sensitivity is achieved when the polarization corresponds to the orientation of the arms and when the signal comes in perpendicular to the plane spanned by the arms. Thus, F_+ and $F_\times$ are affected by polarization and location of the GW source in the sky (for a more thorough discussion, see e.g. 8.3 of [7] II.B. of [9] or III of [10]).

Each GW detector produces a constant data stream d consisting of noise n and (possibly) a signal h, so we either have $d = n$ or $d = n + h$. Since the output comes from a photo diode, it is a voltage. Although one could also analyse this data, quantities with a more intuitive interpretation are preferred for convenience reasons. Thus, voltages are converted into lengths or, even more commonly used in GW analysis, relative length change or *strain*

$$\frac{\Delta L}{L} = \frac{\tilde{L} - L}{L} \,. \tag{2.8}$$

However, this conversion is not the only processing step done. As there is a multitude of noise sources, many of which are known and also monitored, great effort is put into calibrating the data such that it is not corrupted by noise. At first, these steps are applied directly to data coming out of the detector (low-latency processing), but there is also an offline calibration after the respective period of data taking (an *observing run*) is over. For the latter, it is possible to include more noise sources which surfaced during the measurements and thus further improve data quality. Analyses in GWTCs therefore only use offline data, although the quality of low-latency data would suffice as well.

2.2. Data Analysis

Collecting data is very important in science, but an equally important aspect of many science applications is how the measured data/ observations are analysed after being taken. That is the objective of data analysis. The goal of such an analysis in the specific case of data from GW detectors is obvious: finding GWs.

Besides finding GW signals to prove their existence, they are interesting because they allow the inference of many properties of the source that emitted them (because these govern the morphology/ shape of the signal). Very prominent sources of GWs are compact binary systems, consisting for example of two black holes, two neutron stars or one black hole and one neutron star. The reason why these binary systems are interesting is that they produce GWs which are very likely to be detected by the GW detectors and have already been detected numerous times.

To validate the quality of not only processing but also inference steps, simulated signals are injected into the data stream and then compared to the recovered signals and properties. Therefore, the processes described in the following subsections have a solid foundation.

2.2.1. Search Pipelines

The sources used for this subsection are the GWTCs [2, 3, 8, 1] and [11].

As mentioned earlier, during the data taking process there is a constant data stream. This can be described as a vector $\mathbf{d}$ with the components being data streams from each detector. However, in the following subsections only one data stream d will be used because it turns out that the formalism developed for d easily transfers to $\mathbf{d}$.

Similar to the processing steps, besides the more reliable offline searches, there are also online search methods. These use essentially the same algorithms with only slight modifications to reduce latency. To find signals, basically two approaches are used. The first one is to transform the data into Fourier/ frequency domain (which will be denoted by a $\tilde{}$ over the respective quantity) and search for unusual behaviour, that is unexpected excess power of $\tilde{d} = \tilde{d}(f)$ in some frequencies. Although some signals are visible when looking at the spectrogram of $\tilde{d}$ (plot of signal power in time-frequency-plane) and therefore at the signal amplitudes themselves, this is not the general way to identify excess power.[2] Instead, we use that the GW detector noise is approximately a stationary (properties do not change over time) and Gaussian random process[3], which means it follows the distribution

$$p(n) = p(n(t)) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(n(t)-\mu)^2}{2\sigma^2}} \tag{2.9}$$

[2] One problem is the frequent occurrence of *glitches*, short periods of very high noise caused e.g. by environmental sources (which is why noise sources are studied carefully and the known ones are monitored).

[3] This is useful to construct a formalism for detection, but real detector noise has non-Gaussian features. We have to account for non-Gaussian nature later on, e.g. by assigning false alarm rates (details in 7.5 of [11]).

at every time t.[4] Since measuring signals/ time series in reality involves measuring its value at different times t_j, there are many measurements/ samples $n_j := n(t_j)$ of the noise. The subsections 7.1.1 - 7.1.2 of [11] show in a very elegant way how the joint probability of these samples ($p(\{n_j\}) = \prod_j p(n_j)$, assuming they are independent) can be rewritten in frequency domain and then how (assuming that, on average, there is no noise and thus $\mu = \langle n \rangle = 0$) the variance σ^2 is related to the (one-sided) *power spectral density* (PSD) defined via

$$\langle \tilde{n}(f)\tilde{n}^*(f')\rangle = \frac{1}{2} S_n(f)\,\delta(f - f')\,. \tag{2.10}$$

Besides the fact that S_n gives us the signal power for each frequency, this equation also tells us that different frequencies are uncorrelated since $\delta(f - f') = 0$ if $f \neq f'$. In practice, S_n is not calculated by using this exact formula but instead by averaging over $|\tilde{n}(f)|^2$ for a certain time (which has to be large enough, in particular larger than the duration of expected signals; for details, see 4.1 of [7]) as an replacement of the expectation. The idea behind this is that different noise realizations are present at different times, so time averaging "$\overline{\cdot}$" is equivalent to taking an ensemble average (expectation) $\langle\cdot\rangle$ over all noise realizations. The reason that this replacement is necessary at all is that the real experiment/ data taking process conducted by the LVK is not repeatable which means only data from one run is available (for obvious reasons, we cannot control GWs passing by and most of them will appear just once), corresponding to one noise realization for the whole run.

Knowing (calculating) the PSD is important because it can be used to rewrite

$$p(n) \propto \exp\left(-2\int_0^\infty \frac{|\tilde{n}(f)|^2}{S_n(f)}\,df\right) =: \exp\left(-\frac{q(n)}{2}\right) \tag{2.11}$$

and this form shows that from the noise signal n one can compute a quantity $q(n)$ which is distributed according to a standard normal distribution, i.e. a Gaussian with zero mean and unit variance, standard deviation. On first glance one might wonder why the reweighted or *whitened* noise $\frac{\tilde{n}(f)}{\sqrt{S_n(f)}}$ instead of the Fourier spectrum $\tilde{n}(f)$ of the measured time series n appears here. The reason is that reweighting with the average noise or *amplitude spectral density* (ASD) $\sqrt{S_n(f)}$ ensures that each frequency has the same relevance/ response to excess power (high amplitudes for frequencies where much noise is expected from S_n are less relevant than for ones with less noise). This leads to a standard normal distribution.

By computing this distribution for measured data d (which means we have to whiten it), we can then compare it to the distribution expected if only noise was present. The presence of a GW signal would result in a shifted mean compared to (2.11), which could then be observed. Since there are also other potential causes for such shifts and because we do not know when GWs arrive, it is not possible to make a certain statement about the presence of a signal, but we can assign a probability of that being the case (or a probability of the data not being noise, a subtle but important difference; for each pipeline, this is encoded in a *ranking statistic*, a quantity equivalent to such probabilities).

[4] It is important to note that $p(n)$ is not meant to be the PDF of the whole time series $n(t)$, but rather for the realization $n(t)$ at a specific, fixed time t.

This excess power methods is, however, not without any caveats. To get statistically more meaningful statements, one combines the results from multiple times and frequencies[5] to obtain the noise distribution (a *chi-squared distribution*, which is essentially the sum of multiple squared standard normal distributions). The problem is that the signal strength required for it to be detected increases with the number of frequencies and samples used, but that reduces the number of signals which can potentially be detected because it diminishes the sensitivity to weaker signals. To overcome this, one can look for other ideas to detect signals, which leads to the second and probably more well known approach presented here. It uses that one can simulate the GW signals predicted by general relativity, so we can specifically check if such a signal is contained in the data. Now, the probability of that being the case will be assessed from the field of statistics using a bit more sophisticated math than the first approach (therefore, it is explained over several subsections).

2.2.2. Bayesian Inference

There is a variety of sources on statistics, but particularly useful are [12, 13] as they discuss the topic in the context of GW data analysis (more mathematical is [14]).

To use the mathematical methods of statistics, we first have to give a mathematical formulation of our goal. This is done by specifying two hypothesis which will be compared: the *null hypothesis* $\mathcal{H}_0$, which says that the data contains no signal and only noise, and the *signal hypothesis* $\mathcal{H}_1 = \mathcal{H}_{1,h}$, which says that the data stream contains a GW signal template h. The goal is to assign probabilities to these hypotheses and then infer statements based on that. For a confident statement "there is a signal in the data", the null hypothesis has to be rejected sufficiently (if that is not the case, no certain statement can be made).

A very popular and successful approach to infer such statements (not only in physics but also general data analysis) is to use *Bayes' theorem*. Using *probability density functions* (PDFs)[6] p, it reads

$$p(H|E) = \frac{p(E|H)\,p(H)}{p(E)}\,. \tag{2.12}$$

A derivation can be found in appendix A, along with further information on notation and more. The most important takeaways are that "|" marks a dependence (e.g. on models or parameters) and that "," marks "and". It is now important to explain each quantity:

- H: the hypothesis being tested.
- E: the data taken from a measurement, sometimes called *evidence*.
- $p(H)$: the probability that H is true from before E was measured. In the Bayesian approach to probability theory, this is the belief one has in H being true. Because it represents the belief *before* the new evidence E was measured, it is termed a *prior probability* or simply *prior*. Usually, it is denoted by $\pi(H) := p(H)$ and one way to obtain

[5] Also frequencies despite the integration over $[0, \infty]$ in (2.11) because we can split up the signal into many signals coming from each frequency bins we measure.

[6] The "density function" part will often be omitted when referring to quantities.

it is using theoretical models of what is measured (no prior knowledge can also be represented by choosing a uniform distribution).[7]

- $p(E)$: this prior of E does not play as big of a role as the other quantities because it mainly acts as a normalization constant which ensures that $\int_H p(H|E)\,dH = 1$. It is termed *model evidence* and often denoted by $\mathcal{Z}(E) := p(E)$.
- $p(H|E)$: the probability/ belief that H is true knowing the evidence E. This is the desired quantity as it uses the maximum amount of information available and thus provides a more confident statement than the prior $p(H)$. To emphasize the difference between the prior $p(H)$ and $p(H|E)$ which is obtained *after* measuring E, the latter is termed *posterior probability* or simply *posterior* (it is a function of H, not E).
- $p(E|H)$: represents how probable/ likely it would be to measure E if H was true (how well they fit together) and is thus termed *likelihood*. This works by assigning a probability distribution $p_H(E) = p(E|H)$ to every H, which can then be evaluated at the measured value E to see how well this observation can be described using H or in other words, to see how likely H is correct. To emphasize that it is a function of H rather than E (the measurement is taken to be fixed), it makes sense to switch the arguments in $p(E|H)$ and to avoid notational confusion with the posterior $p(H|E)$, the likelihood will be denoted by $\mathcal{L}(H|E) := p_H(E) = p(E|H)$.

Bayes' theorem therefore shows how beliefs can be updated systematically using new information: the prior belief is weighted with the likelihood of measuring E assuming H is true (classification of E) and then normalized to obtain the updated, posterior belief. In the context of GW data analysis, this allows us to go from certain assumptions about the distribution and properties of black holes/ neutron stars (obtained from cosmological models) to a probability that data streams contain GW signals with these properties.

Of course, we are interested specifically in a formulation of Bayes' theorem for its application to GW data analysis. In this context, it is very common to not write the posterior as a function $p(\mathcal{H}_1|d)$, but instead directly as a function $p(h|d)$ of the signal h tested for. Because template signals always depend on a certain set of parameters, i.e. $h = h(\theta) = h_\theta$ for $\theta \in \Theta$, and are even fully determined by them (when a waveform model is specified), another equivalent way to write the posterior is $p(\theta|d)$. This is a function of the parameters θ in the parameter space Θ and omitting further dependencies on models etc.[8], we can finally write:

$$p(h|d) = \frac{\mathcal{L}(h|d)\,\pi(h)}{\mathcal{Z}(d)} = p(\theta|d) = \frac{\mathcal{L}(\theta|d)\,\pi(\theta)}{\mathcal{Z}(d)}\,. \tag{2.13}$$

We can interpret this equation rather intuitively: it tells us that the probability $p(\theta|d)$ that a measurement d contains a signal h parametrized by a set of parameters θ depends on how well d and $h = h(\theta)$ fit each other (measured by likelihood), but also on how likely it is that the parameters θ are actually detected (measured by prior). The evidence is not a third

[7] Information on the priors used for GW data analysis can be found in section V.C. of [8], appendix E of [1].

[8] They would only make the notation more complicated. An additional model M, like for example some assumed cosmology, would be denoted by $\pi(h|M)$ and $p(h|d, M)$ following the employed notation.

parameter and thus really just a normalization constant because it can be obtained from the other quantities by essentially marginalizing the likelihood (see (A.11)), i.e.

$$\mathcal{Z}(d) = p(d) = \int_\Theta p(d,\theta)\, d\theta = \int_\Theta \mathcal{L}(\theta|d)\, \pi(\theta)\, d\theta\,. \tag{2.14}$$

This equation also justifies the given name "model evidence" because $\mathcal{Z}(d)$ is a measure of how appropriate the waveform model is to describe d at all (the integral checks h_θ for every parameter in $\theta \in \Theta$ for its fit and then averages over this fit, which gives us the probability of detecting d using a particular waveform model). That it is still a normalization constant despite this important interpretation is obvious from a second, equivalent derivation, which exploits the normalization constraint of $p(\theta|d)$ as a PDF:

$$1 = \int_\Theta p(\theta|d)\, d\theta = \frac{1}{\mathcal{Z}(d)} \int_\Theta \mathcal{L}(\theta|d)\, \pi(\theta)\, d\theta \quad \Leftrightarrow \quad \mathcal{Z}(d) = \int_\Theta \mathcal{L}(\theta|d)\, \pi(\theta)\, d\theta\,. \tag{2.15}$$

The next subsection will show how to find an optimal template h for given data d (under the assumptions made about the noise, i.e. n is stationary and Gaussian).

2.2.3. Candidate Identification

This section follows [7] and elements are taken e.g. from section 5.3.4 of [6] (5.1.1. contains an alternative approach), section II of [15] and subsection 2.3.3 - 2.3.4 of [5].

Similarly to how (A.15) was derived, we can consider the joint PDF $p(d,H)$ where $H \in \{\mathcal{H}_0, \mathcal{H}_1\}$ and then marginalize to obtain a new expression for $p(d)$. Because integrals in finite spaces are nothing but sums, this reads $p(d) = p(d,\mathcal{H}_0) + p(d,\mathcal{H}_1)$, so we obtain

$$p(h|d) = \frac{\mathcal{L}(h|d)\, \pi(h)}{p(d,\mathcal{H}_0) + p(d,\mathcal{H}_1)} = \frac{\mathcal{L}(h|d)\, \pi(h)}{\mathcal{L}(h|d)\, \pi(h) + \mathcal{L}(n|d)\, \pi(n)} = 1 \Big/ \left(1 + \frac{\mathcal{L}(n|d)\, \pi(n)}{\mathcal{L}(h|d)\, \pi(h)}\right) \tag{2.16}$$

where we express the evidence in terms of probabilities for $\mathcal{H}_1$ being true and not being true ($\mathcal{H}_0$ being true). This is not important to actually calculate the posterior, but it introduces an important quantity which will be very useful: $p(h|d) = p(h_\theta|d)$ is a monotonic function of

$$\Lambda = \Lambda(\theta|d) := \frac{\mathcal{L}(h_\theta|d)}{\mathcal{L}(n|d)}\,, \tag{2.17}$$

the *likelihood ratio*. Λ is the first example of an *optimal test/ ranking statistic*, a notion used to describe quantities which reduce the information contained in measured data to one number and then allow to assess whether or not the hypothesis is true. Optimality then refers to the described relationship with $p(h_\theta|d)$, high likelihood ratios will lead to high posterior probabilities (ignoring the priors for now). This should also make sense intuitively, since it means we want to find the value/ point θ which is most likely to have generated the observed data d.

Many probability distributions (and then also likelihoods) are exponential functions and in this case, also the likelihood ratio is of such form. Therefore, it is common to also look at log-likelihoods and log-likelihood ratios, i.e.

$$\log \Lambda(\theta|d) = \log \frac{\mathcal{L}(h_\theta|d)}{\mathcal{L}(n|d)} = \log \mathcal{L}(h_\theta|d) - \log \mathcal{L}(n|d)\,. \tag{2.18}$$

Because log is a strictly increasing function, the value maximizing $\log \Lambda$ maximizes Λ as well. For Gaussian noise, we can derive a more explicit form of $\log \Lambda$. Since $\mathcal{H}_0$ being true means $d = n$ and $\mathcal{H}_1$ being true means $r = d - h = n$, the respective expected distributions for each $\mathcal{H}_i$ can be expressed using $p(n)$ by replacing n with $d, d - h$. Since the expected distributions are nothing but the likelihoods of each hypothesis, we can use (2.11) to obtain

$$\begin{aligned}\log \Lambda(\theta|d) &= -\frac{1}{2}\langle d - h_\theta | d - h_\theta\rangle + \frac{1}{2}\langle d|d\rangle &(2.19)\\ &= -\frac{1}{2}\langle d|d\rangle + \langle d|h_\theta\rangle - \frac{1}{2}\langle h_\theta|h_\theta\rangle + \frac{1}{2}\langle d|d\rangle \\ &= \langle d|h_\theta\rangle - \frac{1}{2}\langle h_\theta|h_\theta\rangle\ . &(2.20)\end{aligned}$$

Because $\langle h_\theta|h_\theta\rangle = \|h_\theta\|^2$ is constant, $\log \Lambda$ is monotonic in $\langle d|h_\theta\rangle$, which is therefore another optimal test statistic, the *matched filter*. For convenience and brevity, we defined

$$\langle a|b\rangle = \int_{-\infty}^{\infty} \frac{\tilde{a}(f)\tilde{b}^*(f) + \tilde{a}^*(f)\tilde{b}(f)}{S_n(f)}\,df = 4\,\mathrm{Re}\int_0^{\infty} \frac{\tilde{a}(f)\tilde{b}^*(f)}{S_n(f)}\,df\,, \tag{2.21}$$

a noise-weighted inner product computed for the Fourier transformed functions $\tilde{a}, \tilde{b}$ and their complex conjugates $\tilde{a}^*, \tilde{b}^*$ (for sampled data, replace $\int$ with $\sum$). It essentially compares the functions $\frac{\tilde{a}}{\sqrt{S_n}}, \frac{\tilde{b}}{\sqrt{S_n}}$ over the whole Fourier domain. On first glance, one might wonder why reweighting with the ASD $\sqrt{S_n(f)}$ is done instead of using a, b or $\tilde{a}, \tilde{b}$ themselves. We can easily see the reason by thinking of an example signal $d = n + h$. In this case, high amplitudes of $\tilde{d}$ have different implications on the amplitude of $\tilde{h}$, depending on how much noise is present, i.e. the amplitude of $\tilde{n}$ for the respective frequency (only after reweighting with the expected/ average noise $S_n(f)$, each frequency has the same relevance when talking e.g. about excess power for it; these are same considerations as for whitening).

Using this inner product, (2.11) can be rewritten to $p(n) \propto \exp\left(-\frac{1}{2}\langle n|n\rangle\right)$ (which means $q(n)$ used there is $\langle n|n\rangle$) and (2.10) tells us that $\overline{\langle n|n\rangle} = 1$. Furthermore, it gives both equalities (2.19), (2.20) very intuitive interpretations: the first one shows that the likelihood ratio is maximized for the signal h which has the best overlap with the data d; the second one shows that an equivalent statement would be to demand minimization of the residual signal $r = d - h$. Here, minimization is meant with respect to the inner product (2.21) and means that r, the measured data without the signal, shall be very similar to/ indistinguishable from noise. Similarly, maximization means that d and h shall be as equal as possible and at the same time as unequal to noise as possible.

Suppose now that there really is a signal $h_\theta = h_{\hat{\theta}}$ present in detector data, i.e. $d = n + h_{\hat{\theta}}$. The value of $\log\Lambda$ itself is not sufficient to tell us whether or not the signal is contained because there are too many free parameters. One can roughly categorize them into overall shape of the signal (i.e. the normalized template $g_\theta := h_\theta/\|h_\theta\|$), amplitude (which we characterize using $\|h_\theta\|$, a scaling factor for the shape) and location of $h_{\hat{\theta}}$ in the time series d (in principle it could be buried anywhere). The location is not determined by the source properties, but rather by *extrinsic parameters* like time, phase shifts t_0, ϕ_0 (which describe e.g. location, orientation of the source relative to the detectors and are not *intrinsic parameters* like masses, spin). This is why we choose to eliminate them by optimizing $\log\Lambda$ with respect to them (one can think of this as a "fitting-process", where d and h_θ are matched). This is achieved by taking the absolute value of the integral in matched filter and maximizing over time shifts by hand (no analytical way since this depends on d). Moreover, we maximize with respect to the norm $\|h_\theta\|$ by rearranging $\frac{\partial\log\Lambda}{\partial\|h_\theta\|} = 0$ from (2.20) to yield $\|h_\theta\|_{\max} = \langle d|g_\theta\rangle$ and substituting that into $\log\Lambda$. For simplicity, we will first write down the result assuming the signal shape $g_{\hat{\theta}}$ is known. Using this assumption, the result reads

$$\max_{\|h_\theta\|,\phi_0,t_0} \log\Lambda = \frac{1}{2}\frac{\langle d|h_{\hat{\theta}}\rangle^2}{\langle h_{\hat{\theta}}|h_{\hat{\theta}}\rangle} = \frac{1}{2}\frac{\langle d|h_{\hat{\theta}}\rangle^2}{\|h_{\hat{\theta}}\|^2} = \frac{1}{2}\left\langle d\left|\frac{h_{\hat{\theta}}}{\|h_{\hat{\theta}}\|}\right.\right\rangle^2 =: \frac{1}{2}\rho^2 \tag{2.22}$$

where ρ is essentially a normalized matched filter (or equivalently the matched filter computed for data and the normalized template $g_{\hat{\theta}} = h_{\hat{\theta}}/\|h_{\hat{\theta}}\|$). It is important to note that the matched filter in ρ is not a function of the extrinsic parameters anymore, but for brevity, we do not write out the evaluation $\cdot\big|_{\|h_\theta\|_{\max},\phi_{0,\max},t_{0,\max}}$ explicitly in every term.[9]

An equivalent but still interesting form of ρ can be obtained by writing out $d = n + h_{\hat{\theta}}$ and taking the expectation with respect to n, i.e. averaging over all possible noise realizations (or over time). Since noise is a random process, the expectation $\overline{\langle n|h_{\hat{\theta}}\rangle}$ vanishes, yielding

$$\overline{\rho} = \frac{\overline{\langle d|h_{\hat{\theta}}\rangle}}{\|h_{\hat{\theta}}\|} = \frac{\overline{\langle n|h_{\hat{\theta}}\rangle}}{\|h_{\hat{\theta}}\|} + \frac{\overline{\langle h_{\hat{\theta}}|h_{\hat{\theta}}\rangle}}{\|h_{\hat{\theta}}\|} = \frac{\langle h_{\hat{\theta}}|h_{\hat{\theta}}\rangle}{\|h_{\hat{\theta}}\|} = \|h_{\hat{\theta}}\| = \sqrt{\langle h_{\hat{\theta}}|h_{\hat{\theta}}\rangle} =: \hat{\rho}\,. \tag{2.23}$$

Following the *Neyman-Pearson criterion*, a signal candidate h_θ is marked significant when $\Lambda\big|_\theta$ surpasses a certain threshold. This is equivalent to $\log\Lambda$ and hence ρ surpassing a certain threshold and the latter has a very helpful interpretation: to detect a signal $h_{\hat{\theta}}$, its norm $\|h_{\hat{\theta}}\|$ has to be sufficiently high, it has to be loud enough. Because this norm essentially measures the average ratio of $\tilde{h}_{\hat{\theta}}$ and S_n in frequency domain, the detectability of a signal depends strongly on how well the detectors can distinguish it from noise and therefore on the *signal-to-noise ratio* (SNR).[10] Both quantities $\rho, \hat{\rho}$ are a SNR, but there is a subtle distinction between them: ρ is the *detector SNR* obtained in real measurements and $\hat{\rho}$ is the *optimal SNR*

[9] There are many other ways and conventions to define the quantity ρ, we could e.g. omit maximization over t_0 and continue working with time series'. Similarly, we adopt the technique used in [5] to maximize over norm $\|h_\theta\|$ instead of amplitude A (these two are not equal; for a discussion using A, see example 7.2 [11]). This does not change the results because ρ is a normalized quantity anyway.

[10] A more general definition would be the ratio of expectation and standard deviation of a signal. Subsection 7.2.4 of [11] or section IV of [10] show the equivalence using $\langle d|h_\theta\rangle$ as the signal.

for the template $h_{\hat{\theta}}$, a special case of the averaged detector SNR $\overline{\rho}$ that often allows better interpretations since noise is averaged out.

In the case of real search pipelines however, the assumption that the waveform $g_{\hat{\theta}}$ is known has to be dropped, so the process of *maximum likelihood estimation* (MLE) has to be continued. To find our best guess θ' (the *maximum likelihood estimator*) for $\hat{\theta}$, one could in principle maximize over the source properties by solving $\frac{\partial \log \Lambda}{\partial \theta_i} = 0$ for each component θ_i. However, to save computational cost and gain speed, search pipelines currently use a different approach: they compute the detector SNR ρ for each template h_θ from their respective template bank (which can be thought of as a discretization of Θ, a lattice exploring it) and the template that maximizes ρ is, by definition, $h_{\theta'}$ (in general, $h_{\theta'} \neq h_{\hat{\theta}}$). The corresponding SNR value is used as a ranking statistic of each pipeline[11] and decides if a data stretch is marked to potentially contain a signal candidate. For GWTC-3, a peak SNR between $\rho = 4$ and $\rho = 4.8$ was demanded, depending on the pipeline (to interpret these values, the property $\overline{\langle n|n\rangle} = \overline{\|n\|} = 1$ is very useful).

After dropping the assumption of known $g_{\hat{\theta}}$, we also have to rewrite the expressions for the SNRs $\rho, \overline{\rho}$.[12] Replacing $\hat{\theta}$ with θ in every term but d yields

$$\rho = \frac{\langle d|h_\theta\rangle}{\|h_\theta\|} \quad , \quad \overline{\rho} = \frac{\langle h_{\hat{\theta}}|h_\theta\rangle}{\|h_\theta\|} = \|h_{\hat{\theta}}\| \frac{\langle h_{\hat{\theta}}|h_\theta\rangle}{\|h_{\hat{\theta}}\|\|h_\theta\|} = \|h_{\hat{\theta}}\| \left\langle \frac{h_{\hat{\theta}}}{\|h_{\hat{\theta}}\|} \middle| \frac{h_\theta}{\|h_\theta\|} \right\rangle =: \hat{\rho}\, O(h_{\hat{\theta}}, h_\theta)\,. \tag{2.24}$$

This tells us that even a loud signal will not produce high SNRs if the template h_θ has a small *overlap* or *match* $-1 \leq O(h_{\hat{\theta}}, h_\theta) = \langle g_{\hat{\theta}}|g_\theta\rangle \leq 1$ with it (where 1 encodes perfect correlation, 0 no correlation and -1 perfect anti-correlation).[13] In simple words, this quantity measures how similar h_θ and $h_{\hat{\theta}}$ are by giving the percentage of (optimal) SNR of $h_{\hat{\theta}}$ retained by h_θ. Because very loud signals are not likely to occur frequently, an important objective in GW data analysis is to have template banks which are able to produce good matches for any GW signal that can potentially be detected. That means one demands that the loss of SNR from the optimal value $\hat{\rho}$ to the value for the maximum likelihood estimator θ' is small and (2.24) tells us that this is equivalent to demanding a high optimized overlap

$$\mathrm{FF} = \max_{\theta\in\Theta} \frac{\langle h_\theta|h_{\hat{\theta}}\rangle}{\|h_\theta\|\|h_{\hat{\theta}}\|} = \max_{\theta\in\Theta} O(h_\theta, h_{\hat{\theta}}) = O(h_{\theta'}, h_{\hat{\theta}}) = \frac{\hat{\rho}}{\overline{\rho}\big|_{\theta'}} \tag{2.25}$$

which is called the *fitting factor* (perhaps a more common way to express this is to demand small *mismatches* $1 - O$ between all possible signals and their maximum likelihood estimator from the template bank). It sets a lower boundary value for the SNR threshold since for $\mathrm{FF} < 1$ an optimal SNR of $1/\mathrm{FF}$ times this threshold is necessary for detection.[14]

[11] That changes for multiple detectors, where SNRs from different data streams are combined.

[12] In contrast, the optimal SNR $\hat{\rho}$ does not change since it is defined as $\overline{\rho}$ evaluated in $\hat{\theta}$, not $\overline{\rho}$ itself.

[13] It is important to remember that for the SNR, maximization over amplitude, time and phase shifts was carried out, so this is implicitly done for the overlap as well (at least in our definition)!

[14] To reduce computational cost and thus save time, the size of template banks used in online searches is typically smaller than in offline searches (FF smaller). This can be accounted for e.g. by selecting a lower thresholds.

Another remark has to do with the fact that the analysis described here used data from a single detector. This is not realistic because for the analysis of real strain data, there are three detectors (in case of GWTC-3) and hence, three data streams. Fortunately, the transition to more than one detector is relatively easy when treating noise as uncorrelated between the different detectors. In this case, the noise sources influencing the data stream are of local origin for each instrument, so the joint likelihood is just the product of the likelihoods from each detector (sum instead of product for $\log \Lambda$). This assumption is however, again, not entirely appropriate for the real LVK detectors since they have common noise sources, which might produce similar noise (correlations). To account for uncertainties introduced by the approximations made throughout this section (stationary, Gaussian, uncorrelated noise), we can test the developed methods using *injections*, artificial signals put into detector data, and see how well the methods recover them. From that, quantities like *false alarm rates* or a probability of astrophysical origin p_{astro} are assigned, which are essentially telling us if some signal candidate indeed was a signal (this is not equal to the posterior value of the maximum likelihood estimator θ', PDF values are not probabilities). Only candidates which meet certain conditions like $p_{\text{astro}} \geq 0.5$ are considered for further, in-depth analyses.

2.2.4. Parameter Estimation

The most important sources for this subsection are [6] (particularly section 5.3.4), section V of [1] and [12, 13].

One might ask now why the procedure described in the previous subsection was titled "Candidate Identification" and not "Signal Identification". Are frequency domain searches, matched filtering and SNR not sufficient for finding GW signals and their parameters? The answer is no. While they are reliable for the identification part, there is a need for parameter estimation for each signal candidate. Reasons will be discussed in the following.

The first reason is that the resolution in the parameter space is not high enough, which means the distance between points in the parameter space is too big. Each of these points has a corresponding waveform (which is used in the matched filtering process), so in the end, the number of waveforms used by search pipelines is simply too small to infer sufficiently accurate statements about parameters. It is not possible to solve this problem by simply selecting more waveforms because the limiting factor is computational power. Search pipelines continuously analyse the strain data coming from detectors and have a delay of only a few seconds in order to be able to send out alerts about as part of multi-messenger astronomy. Including more waveforms would slow down the pipelines, which is not an option as one of their main objectives is to have high speed.

This reason justifies that a second analysis is carried out, where a higher resolution in the parameter space can be handled simply because the amount of data to process is much smaller (only relevant data stretches that exceed the thresholds). The term "candidate" makes sense because vetoes are only applied after computing the maximum SNR throughout the respective template bank. However, there is an additional and in fact more important reason why the parameters of the template with maximum SNR are not necessarily "correct"

(our best guesses). To see why this is the case, we have to look at equation (2.16) again. While it is true that the posterior probability is a monotonic function of the likelihood ratio, the previous subsection ignored the factor $\frac{\pi(n)}{\pi(h)}$. This is also a function of the template h, but it represents knowledge and is therefore not optimizable. As a consequence, high Λ do not necessarily lead to high posterior values. This should not be surprising because this factor is a ratio of priors and prior knowledge is a very important aspect of the Bayesian interpretation of probability. It takes into account that, despite a good fit between template and data, it is also relevant how likely a detection of the specific template is at all. Logically, this tells us that the meaning of a high matched filter or SNR value for the posterior is limited in case of an exotic astrophysical signal. Exotic here refers to systems which are either relatively unlikely to exist or relatively unlikely to be detected by LVK instruments and both cases should make us question whether a signal from this source was really measured (mathematically, this is accounted for by the prior ratio).

Therefore, instead of using methods to maximize the likelihood, the goal is now to maximize the posterior itself (maximum a posteriori estimation instead of maximum likelihood estimation), which corresponds to finding a point $\theta' \in \Theta$ such that

$$p(\theta'|d) = \max_{\theta} p(\theta|d)\,. \tag{2.26}$$

θ' approximates the real parameter $\hat{\theta}$ as accurately as possible.

Since an analytical evaluation of the posterior is computationally too expensive, the method currently used by parameter estimation algorithms like `Bilby` [13], `LALInference` [12] or `RIFT` [16] is Markov chain Monte Carlo (MCMC) sampling and variations of that. The basic idea is to randomly select points from the parameter space and generate templates, evaluate the likelihood etc. only in these points. There are sophisticated methods to determine which points to select based on the current values of the posterior (which is how Markov chains work) and those are described in the corresponding papers [13, 12]. Having good methods for that is important because otherwise, the resolution for regions with high slopes might be too small or the resolution in uninteresting regions might be unnecessarily high (in any case, this would have the effect that approximations of posteriors computed from these samples become inaccurate, so results inferred from them would be flawed).

2.2.5. Posterior Samples

The result of random sampling processes (possibly after some additional processing steps) is a set of points in the parameter space Θ which can be used to compute/ approximate the posterior PDF they have been generated from. Because the number of parameters necessary to describe binaries is typically bigger than 15[15] and therefore too big for a proper visualization, marginalized posteriors $p(\theta_i|d)$ for each parameter θ_i, $i = 1,\ldots,n = \dim(\Theta)$ are commonly used instead of the full-dimensional one $p(\theta|d)$. If we look for example at θ_1,

[15] That does not even include combinations of parameters or calibration parameters of the instruments.

the marginalized posterior is

$$p(\theta_1|d) = \int_\Theta p(\theta|d)\, d\theta_2 \dots \theta_n = \int_\Theta p(h_\theta|d)\, d\theta_2 \dots \theta_n \,. \tag{2.27}$$

Equivalently (see appendix A.1.1 for mathematical details), one can take only the first component θ_1 from each sample point θ and then compute the corresponding PDF of this one-dimensional dataset. That means $p(\theta_1|d)$ is something like the projection of $p(\theta|d)$ onto the θ_1-axis (and accordingly for the other θ_i). This method will be used extensively throughout this work. Fortunately, we can apply it easily because the LVK has published samples from each posterior in GWTC-3 online [17]. The data format used has the suffix `.hdf5/ .h5` (both describe the same one) and the files are organized similarly to folders on a computer: using keywords, it is possible to navigate through the file to find datasets or other information. An introduction on how to work with `.hdf5`-files in Python (the language of choice for this work) is given in the documentation of the h5py package (available here).

However, even working with 1D-posteriors $p(\theta_i|d)$ is not always convenient, for instance to summarize the parameter estimation results for 36 events like it was required in GWTC-3. There are several potential quantities suited to summarize the information contained in PDFs/ data sets in general and from the widely used ones *mean* (average value), *median* (middle value, roughly speaking), *mode* (maximum), the LVK chose the median. This choice is made for stability reasons that include resilience to outliers (which the mean is vulnerable to) and the shape of the distributions (in particular peak location, i.e. mode) having some dependencies on e.g. the prior which should not affect the summarizing quantity too much. Median values are then paired with 90% *credible intervals* to provide information about the range parameters most likely lie in[16], which can be thought of as a measure of (un-)certain the estimation is. They are computed by measuring the distance between the median as the 50th percentile and the 5th, 95th percentiles. To calculate these quantities in practice for GWTCs, not posteriors from each waveform are used. Instead, an equal number of samples from each waveform is taken and combined to a `Mixed` posterior and the median, credible interval from this distribution are cited in most publications. The idea behind this is that the results from each waveform should be rather equal (at least consistent) as they are inferred from identical data. Chapter 4 will show how appropriate this idea is.

[16] We cannot assign probabilities to some set of parameters and thus a point in Θ being the "true" parameters because points are null sets, so it only makes sense to assign them to intervals. In the Bayesian interpretation of probabilities, this represents a state of certainty that the "true" value lies in the interval.

3. Principal Component Analysis

One aspect of Data Analysis is the representation of data. While this might be trivial for data sets consisting of two or three variables (one can plot them), it gets increasingly difficult for higher dimensional ones. Due to the convenience that plotting provides, we might look for ways to modify the data and then reduce its dimensionality such that it becomes possible again.

Of course, this reduction should not discard relevant information contained in the data as that is what we are really interested in. Hence, the objective is to find a way to optimally (in the sense of retaining information) perform this modification. To get the theoretical framework which will be capable of achieving this goal, it is necessary to introduce some mathematical definitions.

3.1. Definitions

Every GW signal can be parametrized by one exact, "true" value for each of its parameters. However, the methods described in section 2.2 cannot find this parameter exactly, but only estimate it by essentially assigning probabilities to every possible value. As further argued at the end of section 2.2, the result of LVK inference processes is a collection of points drawn from the posterior probability distribution in the parameter space, a *sample*. There are several quantities from statistics/ probability theory which can be used to characterize probability distributions/ PDFs and there also exist analogous quantities for samples from them.[1]

Suppose now the parameter space has n dimensions, i.e. the data is described by n variables x_i, $i = 1, \dots, n$. The values measured for them can be collected in a data vector $\mathbf{x} = [x_1, \dots, x_n]$ where each component holds the values of one variable. Because samples usually contain more than one element/ data vector (let us assume there are m of them), their data vectors are also indexed and denoted by $\mathbf{x}_i$ (note that $\mathbf{x}_i \neq x_i$). All data vectors can then be collected in a *data matrix* $\mathbf{X}$. There are two possible choices how to construct it: write every measurement as a row vector or as a column vector. Here, the former is chosen, because the components of the matrix are easier to remember. Thus, we deal with $(m \times n)$ matrices

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{bmatrix} = \left[x_{ij}\right]_{i,j=1}^{m,n} := \left[(\mathbf{x}_i)_j\right]_{i,j=1}^{m,n} . \tag{3.1}$$

A better visualization is provided in figure 3.1.

[1] These sample versions of the quantities are the *maximum likelihood estimators* of their exact pendants.

3.1.1. Sample Statistics

There are several quantities from probability theory which can be used to characterize such a sample. These are called *(sample) statistics*. All of them are estimators of the corresponding exact quantities for the underlying statistical population/ distribution, like the expectation value or the variance (more precisely, the maximum likelihood estimators).

The first, most basic one is the *(sample) mean* of the measurements, which is simply defined as the average over all of them. In the general case with n variables, the mean is also a vector and its components are the variable means:

$$\overline{\mathbf{x}} = \frac{\mathbf{x}_1 + \cdots + \mathbf{x}_m}{m} = \frac{1}{m}(\overline{x}_1, \ldots, \overline{x}_n) = \left[\frac{1}{m}\overline{x}_j\right]_{j=1}^{n} := \left[\frac{1}{m}\overline{x_j}\right]_{j=1}^{n} = \left[\frac{1}{m}\sum_{i=1}^{m} x_{ij}\right]_{j=1}^{n}. \tag{3.2}$$

For the case of only one variable, $n = 1$, this becomes the well known formula

$$\overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{3.3}$$

for the arithmetic mean/ average.

When dealing with more than one variable, it might also be of interest to find out if two of the variables have relations of some kind. A quantity which measures linear versions of such (cor-)relations is the *(sample) covariance*. For n variables, there are $\left(n^2 + n\right)/2$ possible combinations of variables and for convenience, they are arranged in an $(n \times n)$ matrix. This *covariance matrix* is defined by

$$\mathbf{S} = \left[\frac{1}{m-1}\sum_{k=1}^{m}\left(x_{ki} - \overline{x}_i\right)\left(x_{kj} - \overline{x}_j\right)\right]_{i,j=1}^{n}. \tag{3.4}$$

Intuitively, every component $S_{ij} = \mathrm{cov}(x_i, x_j)$ contains information about how simultaneously the variables x_i, x_j deviate linearly from their mean. Because there is more than one measurement of each variable, it then essentially takes the average over all of them.[2] For uncorrelated variables x_i, x_j, this averaging will result in $S_{ij} = 0$. Furthermore, $\mathbf{S}$ is symmetric by definition, i.e. $S_{ij} = S_{ji}$. This should also be intuitive since a correlation always involves both variables, different values for $\mathrm{cov}(x_i, x_j)$ and $\mathrm{cov}(x_j, x_i)$ would not make sense (it is also a consequence of the fact that a $n \times n$ matrix has n^2 entries, but there are only $\left(n^2 + n\right)/2$ different combinations of variables).

It is also possible to express the definition of $\mathbf{S}$ in a more convenient way by writing it as a matrix product. Because not a value x_{ij} itself is relevant for the covariance, but instead its deviation from the mean $\overline{x}_j$, the matrix $\mathbf{X}$ cannot be used for that. $\mathbf{X}$ has to be mean-centered

[2] The $m - 1$ in the denominator comes from Bessel's correction to account for a bias in the data (details are not important here). Only these unbiased versions are the maximum likelihood estimators, so they are used.

first (then, subtracting the mean has no effect as it is zero) and this results in a new matrix

$$\tilde{\mathbf{X}} = \left[\tilde{x}_{ij}\right]_{i,j=1}^{m,n} = \left[x_{ij} - \overline{x}_j\right]_{i,j=1}^{m,n} = \begin{bmatrix} \mathbf{x}_1 - \overline{\mathbf{x}} \\ \vdots \\ \mathbf{x}_m - \overline{\mathbf{x}} \end{bmatrix} \tag{3.5}$$

which is sometimes called the deviation matrix (we do not adopt that, the distinction is made by using different symbols). Mean-centering can be thought of as a change of coordinates. The information contained in $\mathbf{X}$ and $\tilde{\mathbf{X}}$ is in principle the same, but for the latter the origin of the coordinate system is shifted to the mean.[3] Using this new matrix, it is easy to verify

$$\mathbf{S} = \frac{1}{m-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\,. \tag{3.6}$$

Besides measuring how two variables are (cor-)related, the diagonal entries of the covariance matrix measure the relation of a variable x_i with itself. This turns out to be a measure of the variability of x_i, the *(sample) variance*. We can see this equivalence by thinking of a simple alternative idea to quantify this variability: measuring the Euclidean distance of the variable from its mean and then average over the results for every measurement. By writing out the square of the result, we obtain

$$\operatorname{var}(x_i) := \frac{1}{m-1}\sum_{k=1}^{m}(x_{ki} - \overline{x}_i)^2 = \mathbf{S}_{ii} = \operatorname{cov}(x_i, x_i)\,. \tag{3.7}$$

It is now straightforward that

$$\operatorname{tr}(\mathbf{S}) = \sum_{i=1}^{n}\operatorname{var}(x_i) = \sum_{i=1}^{m}\frac{\|\mathbf{x}_i - \overline{\mathbf{x}}\|^2}{m-1}\,, \tag{3.8}$$

the total variance (= trace of covariance matrix) is the (square of the) average Euclidean distance between each data vector $\mathbf{x}_i$ and the mean vector $\overline{\mathbf{x}}$. Despite the square, linear deviations are measured by var (because it is only the distance that is squared, not the quantities; this square has the advantage that positive and negative deviations for different measurements do not cancel out).

A quantity closely related to the variance is the *standard deviation*

$$\sigma_i := \sigma_{x_i} = \sqrt{\operatorname{var}(x_i)}\,, \tag{3.9}$$

which can be interpreted very similarly to the variance (but we can omit the "square of" part when describing it as a Euclidean distance).

[3] $\tilde{\mathbf{X}}$ and $\mathbf{X}$ produce the same covariance matrices, they contain the same information in this regard (because the deviations from their respective mean are the same). The main reason to introduce $\tilde{\mathbf{X}}$ here is to get (3.6).

3.2. Theory

The two sources used for this section are [18, 19, 20][4]. The proofs of most of the properties can be found in [18] (but careful: PCA with sample data is important here).

3.2.1. Principal Component Analysis using Covariance Matrices

After knowing some basic notions of statistics, we will now be able to understand how to achieve the goal set in the introduction (although some knowledge of linear algebra will be required, too). The basic idea is to find a transformation to replace the variables x_i with some new ones z_i. The word transformation does not mean the information contained in the data will change under it, we simply look at it differently after the transformation.

The issue with the variables x_i is that interesting patterns in the data may not be visible in the x_i because they are caused by combinations of them. This is the case when variables are correlated. Such correlations show up in the covariance matrix $\mathbf{S}$ as defined in equation (3.4). Hence, it is clear what to do: eliminate correlations between variables, i.e. $S_{ij} = 0, \forall i \neq j$. For everybody with some familiarity in linear algebra, it is clear that this diagonalization of $\mathbf{S}$ can be achieved by finding a special basis made up of the eigenvectors $\mathbf{a}_i$ of $\mathbf{S}$ and compute it in this basis. These eigenvectors are nothing but representations of the z_i and their j-th component tells us what the contribution of the j-th variable x_j to z_i is.[5]

Notationally, we will continue to use the definitions introduced in the previous section. The n-dimensional data vectors $\mathbf{x}_i$ obtained from every of the m measurements (which one is specified in the index i) are used as row vectors to form an $(m \times n)$-matrix $\mathbf{X}$ (see (3.1)) and the covariance matrix is $\mathbf{S} = \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}/(m-1)$ where $\tilde{\mathbf{X}}$ is the mean-centered version (3.5) of $\mathbf{X}$. A new quantity is the matrix

$$\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n] = [\mathbf{a}_i]_{i=1}^n \tag{3.10}$$

which consists of the eigenvectors $\mathbf{a}_i$ of $\mathbf{S}$ written next to each other in columns. Due to a general property of symmetric matrices like $\mathbf{S}$, we can be sure that eigenvalues λ_i and eigenvectors are real and also that the eigenvectors are orthogonal (they are also assumed to be normalized, i.e. $\|\mathbf{a}_i\|^2 = 1$, which makes them orthonormal).

It is a general property that permuting basis vectors in an arbitrary way does not change the fact that they form a basis (it could happen that the orientation is changed, but this can be corrected by multiplying one of the $\mathbf{a}_i$ with -1). Using this, henceforth we will assume the eigenvectors to be ordered in a way that the corresponding eigenvalues λ_i have descending order $\lambda_i \geq \lambda_{i+1}$ (that will turn out to be very convenient).

[4] [18, 18] follow the convention used here, [19] does not (data vectors are columns there).

[5] It might be a good idea to elaborate a bit on this because talking about coordinates often causes confusion: in the coordinates used initially, the value of x_i determines the component on axis i (belonging to the standard basis vector $\mathbf{e}_i$ which is 1 in component i and 0 in every other component). In these coordinates, the value of z_j determines the component of the eigenvector $\mathbf{a}_j$. By rotating (or mathematically speaking: transforming) the coordinate system, one can change the appearance of vectors such that the component of the i-th axis $\mathbf{e}_i$ is now determined by z_i. In these new coordinates, the axes are uncorrelated and the matrix $\mathbf{S}$ diagonal.

To have a clear distinction between the different representations, new symbols will be used to denote the transformed quantities. Just like the data vectors containing values of the x_i were denoted by $\mathbf{x}_j$, the transformed data vectors containing values of the z_i will be denoted by $\mathbf{z}_j$. Computing them is possible by using the general law to transform row vectors:

$$\mathbf{z}_i = \mathbf{x}_i\mathbf{A} = [\mathbf{x}_i\mathbf{a}_1, \ldots, \mathbf{x}_i\mathbf{a}_n] = \left[\mathbf{x}_i\mathbf{a}_j\right]_{j=1}^{n} = \left[\sum_{k=1}^{n} (\mathbf{x}_i)_k \left(\mathbf{a}_j\right)_k\right]_{j=1}^{n} = \left[\sum_{k=1}^{n} x_{ik}a_{kj}\right]_{j=1}^{n}. \tag{3.11}$$

It should be pointed out that because $\mathbf{x}_i$ is a row vector, $\mathbf{x}_i\mathbf{a}_j$ is the inner product between them and thus a scalar. Although $\mathbf{X}$ is a matrix, it is not transformed like a "typical" matrix because it is not associated with a linear transformation. Instead, it is a collection of row vectors and each row vector shall be transformed. As only the representation of the data is changed, this basic shape should stay the same for the transformed data matrix $\mathbf{Z}$ and by writing everything out explicitly, we indeed obtain

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_m \end{bmatrix} = [\mathbf{z}_i]_{i=1}^{m} = [\mathbf{x}_i\mathbf{A}]_{i=1}^{m} = \left[\sum_{k=1}^{n} x_{ik}a_{kj}\right]_{i,j=1}^{m,n} = \mathbf{X}\mathbf{A} \tag{3.12}$$

using the general formula for the multiplication of two matrices. Based on this, some new terminology will be employed: the columns of $\mathbf{Z}$ are of particular importance since they contain the values for each z_i and are called *principal components (PCs)*.[6] Each entry z_{ij} of the matrix $\mathbf{Z}$ (and therefore, of the PCs) is a *PC score* containing the value of the j-th PC for the i-th observation. It is important to note that the PCs are not the same as the $\mathbf{a}_i$ (which are sometimes referred to as *principal axes*). Furthermore, the components a_{ij} of $\mathbf{A}$ and $\mathbf{a}_i$ do have a name: they are termed *PC loadings* and determine the influence of each original variable x_i on the transformed ones $z_j = \sum_{i=1}^{n} x_i(\mathbf{a}_j)_i = \sum_{i=1}^{m} x_i a_{ij}$. A visualization of most of these new quantities along with a comparison with $\mathbf{X}$ is given in figure 3.1.

To see if the objective of no covariance is fulfilled for the new representation $\tilde{\mathbf{Z}}$ of the data, we have to compute the corresponding covariance matrix. Because that is much easier when using mean-centered data matrices, we first introduce

$$\tilde{\mathbf{Z}} = \begin{bmatrix} \mathbf{z}_1 - \overline{\mathbf{z}} \\ \vdots \\ \mathbf{z}_m - \overline{\mathbf{z}} \end{bmatrix} = \tilde{\mathbf{X}}\mathbf{A}\,. \tag{3.13}$$

That the property of zero mean is preserved under the transformation $\mathbf{A}$ can be calculated, but one can also look at a geometric reason: mean-centering was nothing but a shift of the coordinate origin. But the origin does not change under a rotation (transformations between bases with the same orientation are nothing but rotations) and thus the mean stays zero.

[6] As it is also stated in some sources [19, 20], the notation surrounding PCA is very inconsistent and that starts with the definition of PCs. Sometimes, these are defined as the z_i themselves and not as the measured values for the z_i (so admittedly, the distinction is subtle).

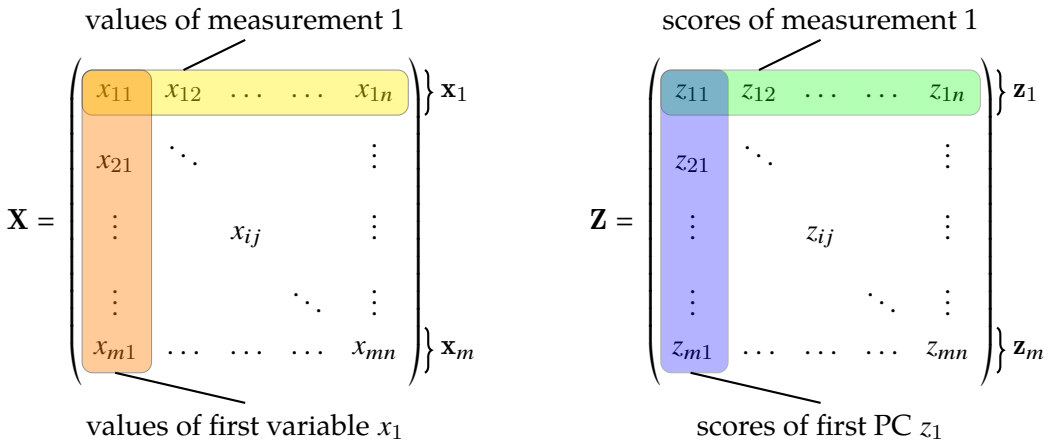


Figure 3.1.: Visualization of data matrices and notions surrounding them. Of course, this would be the same for $\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}$.

Following the formula from equation (3.6), the covariance matrix for $\tilde{\mathbf{Z}}$ is

$$\mathbf{S}_a = \frac{\tilde{\mathbf{Z}}^T \tilde{\mathbf{Z}}}{m-1} = \frac{\left(\tilde{\mathbf{X}}\mathbf{A}\right)^T \tilde{\mathbf{X}}\mathbf{A}}{m-1} = \frac{\mathbf{A}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}\mathbf{A}}{m-1} = \mathbf{A}^T \mathbf{S}\mathbf{A}\,, \tag{3.14}$$

which means it is indeed the matrix obtained by transforming $\mathbf{S}$ and therefore diagonal by definition: $(S_a)_{ij} = \lambda_i \delta_{ij}$. That also makes it very easy to compute the variances of the z_i:

$$\operatorname{var}(z_i) = \operatorname{cov}(z_i, z_i) = (S_a)_{ii} = \lambda_i\,. \tag{3.15}$$

Because the i-th PC contains the values used to compute $\operatorname{var}(z_i)$, there is another interpretation of this: the variance of z_i is the square of the Euclidean norm of the i-th PC (when the mean $\bar{\mathbf{z}}_i$ is subtracted) when interpreted as a vector in the m-dimensional Euclidean space $\mathbb{R}^m$ (equivalent: square of the Euclidean norm of the i-th column of $\tilde{\mathbf{Z}}$).

An important part of the analysis of PCs is to compare their variances (which are sample estimators of the variances of the z_i) to those of the original variables. If some of them are bigger than the original variances, it is very likely that a dimensionality reduction can be applied without losing too much information. That the variances of x_i and z_i can be compared in this way should not be taken for granted, it is only possible because

$$\sum_{i=1}^{n} \operatorname{var}(x_i) = \operatorname{tr}(\mathbf{S}) = \operatorname{tr}(\mathbf{S}_a) = \sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} \operatorname{var}(z_i)\,. \tag{3.16}$$

Hence, the total variance is preserved under the applied transformation, which validates the comparison of $\operatorname{var}(x_i)$ and $\operatorname{var}(z_j)$. If (3.16) did not hold, such statements comparing the magnitude of variances in different bases would not make sense. That is because we can imagine a scenario where the relative proportion of variance (which corresponds to its importance) that two variables x_i, z_i account for is much bigger for x_i than for z_j, but the absolute value $\operatorname{var}(z_j)$ is bigger than $\operatorname{var}(x_i)$. However, since $\operatorname{tr}(\mathbf{S}) = \operatorname{tr}(\mathbf{S}_a)$, absolute

statements are equivalent to relative ones on the proportion, i.e.

$$\operatorname{var}(x_i) \geq \operatorname{var}(z_i) \quad \Leftrightarrow \quad \frac{\operatorname{var}(x_i)}{\operatorname{tr}(\mathbf{S})} \geq \frac{\operatorname{var}(z_i)}{\operatorname{tr}(\mathbf{S}_a)}\,. \tag{3.17}$$

That the eigenvalues are ordered implies $(S_a)_{11} \geq \cdots \geq (S_a)_{nn}$ and we will now explain how this is useful. In the motivation of PCA, one objective was to reduce the dimension of the data from n to q while retaining as much information as possible. It turns out that the transformation which is optimal in this sense is $\mathbf{A}_q = \left[\mathbf{a}_1, \ldots, \mathbf{a}_q\right]$, i.e. the projected data $\mathbf{X}\mathbf{A}_q$ contains as much of the original variability as possible (it is the maximum likelihood estimator of all possible projections). The exact proportion of the variance retained is

$$\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{n} \lambda_i} = \frac{\sum_{i=1}^{q} \lambda_i}{\operatorname{tr}(\mathbf{S})}\,. \tag{3.18}$$

3.2.2. Principal Component Analysis using Correlation Matrices

Although the objective of the PCA is to examine the data contained in $\mathbf{X}$, it is often useful to work with modified versions of this data. A first example of that was $\tilde{\mathbf{X}}$, which is obtained from $\mathbf{X}$ by subtracting the mean of the respective variables. Another modification, which is especially important for physical data, is introduced here.

It is very easy to imagine data sets where the different parameters take values on very different scales. An example from physics would be the difference between wavelengths of light and the distances between mirrors in an interferometer, which is several orders of magnitude. Suppose now that the relative variability of the data around its mean is the same for all parameters (something like 10%). Then, a PCA of the data would not give the expected results because the variance and hence absolute variability of the data is evaluated there. But this absolute value will of course be dominated by the parameter with the highest mean value as its absolute deviation from this mean will be the highest. That means the values in the covariance matrix will be dominated by this parameter regardless of any information (about the variability) contained in it.

To deal with the problem described here, one has to correct the unwanted influence of the variance in the original variables x_i. One straightforward way is to divide the values for each variable by its variance. The problem here is that variances and covariances are proportional to the square of one/ product of two variables, so these new data would suffer from a similar problem as before. The correct "weighting factor" turns out to be the standard deviation because the variables x_i/σ_i obtained by this have normalized variance, i.e. $\operatorname{var}(x_i/\sigma_i) = 1$ ($\Rightarrow \operatorname{tr}(\mathbf{S}) = n$). The covariance matrix of this weighted data is called *correlation matrix* and can equivalently be obtained by replacing S_{ij} with $S_{ij}/\sigma_i\sigma_j$.

The case for correlation matrices is not only made because it eliminates scaling issues, but also because it solves another problem that occurs very frequently in data analysis: units. The new variables z_i are linear combinations of the old variables x_i. But what to do when

one variable is measured in metres and another in kilograms? Does distance plus mass even make sense? Fortunately, these problems do not occur for correlation matrices because the standard deviation has the same units as the respective variables, so x_i/σ_i is dimensionless. This is particularly useful for GW data, where many different quantities are used.

It should be pointed out that there is no general relation between the PCs of covariance matrices and those of the corresponding correlation matrix, so their analysis will produce different results for the majority of data sets.

3.2.3. Geometrical Viewpoint

The approach to PCA shown until now is not the only possible viewpoint on it, there is also a more geometrical one. We will now briefly look at this because it provides some interesting new insights and actually, Karl Pearson used it for the first known derivation of PCs.

The idea is to find the line (vector) that fits the data points optimally. The fit can be described mathematically by the distance between the points and their projection onto the line, which means the distance is measured from line to points and orthogonally to the line. Finding the "best fit" then means finding the line that minimizes these orthogonal distances, which is equivalent to retaining the maximum amount of correlation/ covariance (this correspondence of distance and variance was also mentioned at the end of 3.1.1).

After having found the best fitting line, the objective is to look for the next line of best fit, but with the additional constraint that it has to be orthogonal to the first line (or more generally: all previous lines). This is continued inductively.

The way this connects to PCA is very intuitive. The "lines of best fit containing as much covariance as possible" are nothing but the eigenvectors of the covariance matrix, the projected data points are the PCs and the constraint of orthogonality ensures that the variables are uncorrelated. In this geometrical approach, it is also clear from the beginning why the eigenvalues are ordered and it does not appear as some arbitrary choice, which is a validation of why it was done in the algebraic approach chosen before.

Using this interpretation, we can make the correspondence between variance and information contained in the data. The first PC captures the most variance and at the same time is the most important one to describe the data. Normalizing data before the PCA then ensures that each of the original parameters contains the same amount of information and in this case, a principal axis having an eigenvalue greater than 1 means that it captures more information than the original parameters, which makes it more useful to describe the data.

4. Systematic Differences in Gravitational Wave Event Data

As it was already stated, several pipelines are used separately in parameter estimation to make results more reliable. Although their approach of using MCMC sampling is mostly the same, some use different waveform models to compute posteriors. This may introduce certain differences due to errors caused e.g. by imperfect calibration to NR and there are also biases caused e.g. by noise. While certain differences are therefore inevitable, it does not explain arbitrarily high ones which still have to be explained.

In this chapter, we will develop methods to find such differences and assess their significance Moreover, we will try to explain where they might come from and how they fit in with current knowledge about waveform systematics.

4.1. Criteria

To be able to find events with bad agreement (and also those with good agreement since a reference is necessary for meaningful statements), we need criteria to assess the similarity of two probability distributions. Of course, one could also look at every single event and analyse the posteriors by hand, but with an increased number of events this approach becomes more and more impractical. Hence, we are looking for systematic ways to objectively analyse these differences and that means we need mathematical criteria.

The first and main criterion is the Jensen-Shannon divergence, a well known method to compare different distributions which was also used in the first two GWTCs [2, 3]. It is very reliable for finding differences, but it does not tell us where the differences come from (possible reasons are location of peaks, width of distribution or general shape). Therefore, we will use some additional criteria in order to get statements which are more specific. Their definitions will be much simpler and in case any differences occur, it is almost immediately clear where they come from. Although we cannot guarantee for their stability or provide mathematical proofs of properties, the ideas leading to them should be rather straightforward and they will also show in which cases they are suited well (i.e. produce reliable results).

In order to explain each criterion, the normal distribution will serve as a standard example. It can be characterized by two parameters, mean μ and standard deviation σ, and reads

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{\sigma^2}} . \tag{4.1}$$

4.1.1. Jensen-Shannon Divergence

This subsection was mainly inspired by appendices in the GWTCs and [21].

The first and also main "tool" chosen here is the *Jensen-Shannon divergence* (JSD). For two discrete probability distributions p, q it is defined as

$$D_{\mathrm{JS}}(p,q) = \frac{1}{2}\left(D_{\mathrm{KL}}\left(p, \frac{p+q}{2}\right) + D_{\mathrm{KL}}\left(q, \frac{p+q}{2}\right)\right) \tag{4.2}$$

where D_{KL} denotes the *Kullback-Leibler divergence* (KLD)

$$D_{\mathrm{KL}}(p,q) = \sum_x p(x)\,\log\left(\frac{p(x)}{q(x)}\right) = E_p\left[\log\left(\frac{p(x)}{q(x)}\right)\right] = E_p[\log(p(x)) - \log(q(x))]\,. \tag{4.3}$$

For continuous p, q one has to replace $\sum_x$ with $\int_{\mathbb{R}} dx$. It might no be immediately clear why this measures how different p and q are, but the second formula in (4.3) gives a glance at why. It shows that the KLD is the expectation (taken with respect to p) of the log-difference between p, q and it is useful to think of the expectation as an average over all x here. But still, the question remains what this log-difference tells us. To answer that, we rewrite it as

$$\log(p(x)) - \log(q(x)) = \log\left(\frac{1}{q(x)}\right) - \log\left(\frac{1}{p(x)}\right) =: I_q(x) - I_p(x) \tag{4.4}$$

where I is the *Shannon-information*. One can think of the term information as how surprising the occurrence of a certain event is (for high probabilities, the surprise is low and vice versa). Adding the log is convenient here because then, $p(x) = 1 \implies I_p(x) = 0$. If the surprises of p and q are very equal for all x in the sample space, then the distributions are probably very equal and the KLD will be small. A prominent usage example is to compare models and measurements, for instance the prior and posterior in Bayesian statistics. The result then assesses how much information an observation contains.

However, the KLD has some undesirable properties, e.g.

$$D_{\mathrm{KL}}(p,q) \neq D_{\mathrm{KL}}(q,p)\,. \tag{4.5}$$

One could build a symmetric divergence measure from it by using $D_{\mathrm{KL}}(p,q) + D_{\mathrm{KL}}(q,p)$ or $\frac{D_{\mathrm{KL}}(p,q)+D_{\mathrm{KL}}(q,p)}{2}$, but the measure used here is the average deviation of each distribution p, q from the average of the distributions $\frac{p+q}{2}$ or mathematically, (4.2).[1] This is nothing but the JSD and one of the main reasons for this choice is that it satisfies

$$0 \leq D_{\mathrm{JS}} \leq 1 \tag{4.6}$$

[1] By a straightforward calculation, one can show that equivalent definition is $D_{\mathrm{JS}}(p,q) = H\left(\frac{p+q}{2}\right) - \frac{H(p)+H(q)}{2}$ where $H(p) = -\sum_x p(x)\,\log p(x)$ is the *Shannon entropy*, which is nothing but the average surprise mentioned earlier. Thus, the JSD computes the difference of information/ surprise contained in the average distribution and the average information/ surprise contained in the distributions.

where $D_{\mathrm{JS}}(p,q) = 0$ when $p = q$ and $D_{\mathrm{JS}}(p,q) = 1$ when p, q are disjoint (the KLD is only bounded by $D_{\mathrm{KL}} \geq 0$ and having an upper boundary is very convenient). However, the evolution between these values is highly nonlinear, so $D_{\mathrm{JS}}(p,q) = 0.5$ does not correspond to 50% overlap (in a way one would have to specify further, but this is not the point).

Luckily, there is a way to quantify and visualize values of the JSD by expressing it in terms of two normal distributions. For example, if p, q are two normal distributions with the same standard deviation σ, $D_{\mathrm{JS}}(p,q) \approx 0.007$ means that their means differ by 20% of σ (we refer to that as a "mean difference of 20% in units of σ").

When it comes to actually computing the JSD of two posterior probabilities from different waveforms, there are some technical difficulties (mainly the different sample size and different range of values). These are discussed in the appendix, see A.2. Moreover, although it will be omitted for the most part, it should be noted that the JSD has a unit. Which one it is exactly differs with the $\log$ chosen and for our choice of base 2 (which is also the one used in GWTCs), this unit is bit, a general measure of information. For the way we use the JSD, this unit is not of particular importance because the interpretations will also be clear from the values alone due to its boundaries.

Besides using the JSD to compare p, q directly, we can also use it to compare their shape and thus infer a more specific statement about possible differences between them. In order to do that, we transform the data by normalizing and centering it. That means we divide the data by its standard deviation and then subtract its mean, which results in distributions $\tilde{p}, \tilde{q}$ with mean $\mu_{\tilde{p}} = \mu_{\tilde{q}} = 0$ and standard deviation $\sigma_{\tilde{p}} = \sigma_{\tilde{q}} = 1$. If $D_{\mathrm{JS}}(\tilde{p}, \tilde{q})$ is small, we can infer that p, q have similar shapes and potential big values of $D_{\mathrm{JS}}(p,q)$ would then be caused by very different means or standard deviations.[2]

4.1.2. Mean and Median Differences

The end of the last subsection showed a way to effectively measure shapes of distributions p, q, which means we still need criteria for other potential causes like location or width. It should be intuitively clear what the width refers to (e.g. how long the interval is in which some proportion of the total probability lies) and location can also be visualized very easily as it refers to how much the distributions overlap. We are not directly interested in the location of peaks because in the context of GW posteriors and Bayesian probability in general, the peak location may vary depending on our choice of priors as $p(h|d) \propto p(h)$. The location (mathematically, this corresponds to the support of the distribution) however will not change so much for different priors, making it more stable and thus more desirable to measure.

For the standard example of a normal distribution, properties influencing the overlap are its parameters, namely mean for location and standard deviation for width. Different locations intuitively correspond to different predictions for most probable values, so it does make a lot

[2] Shape here refers to the very basic shape of the distribution, e.g. the number of peaks or if it is symmetric. If one distribution is just a stretched version of the other, the difference is caused by different standard deviations. Such differences, and likewise different locations of the distributions, are not present in these transformed distributions and shape does not refer to them in this context.

of sense to compare μ_p and μ_q. A first idea might be to look at $|\mu_p - \mu_q|$ and then choose a threshold to mark significance. However, this difference does not necessarily produce results which can be interpreted directly because it has different meanings for sharply peaked distributions, where even small differences can be significant, and very wide distributions, where small differences are not very significant. A way to take this into account is to look at the mean difference and compare it to the average standard deviation of p, q, which we will refer to as a "mean difference in units of the average standard deviation", i.e.

$$|\mu_p - \mu_q| \leq c \, \frac{\sigma_p + \sigma_q}{2} \, . \tag{4.7}$$

Here, c is determined the threshold we choose and $c = 0.2$ would correspond to a 20% mean shift (in units of the average standard deviation). The quantity which will be computed is

$$\frac{|\mu_p - \mu_q|}{(\sigma_p + \sigma_q)/2} \tag{4.8}$$

and p, q are marked significantly different if it exceeds a certain threshold (like $c = 0.2$). This criterion should have the additional advantage of not marking parameters with relatively big differences but high uncertainty significant (because this uncertainty should affect in the distribution by widening it, which should then appear in the standard deviation).

However, although posteriors will often have shapes similar to normal distributions, they will not be exactly of this type. In this case, it was already mentioned that the median is often a more stable quantity than the mean and thus, we employ a criterion analogous to (4.8), just replacing the μ_i with medians $p_{50,i}$ and the σ_i with $t\%$ credible intervals $p_{50,i} - p_{50-t/2,i}, p_{50+t/2,i} - p_{50,i}$. There are a few more things to specify when dealing with credible intervals over standard deviations: they are not symmetric and also can have different lengths depending on the percentage of probability included. The first one is relatively easy to account for, we simply take the ones between both medians, so assuming $p_{50,1} \geq p_{50,2}$ (which is possible without loss of generality) the final criterion reads

$$\frac{p_{50,1} - p_{50,2}}{\left((p_{50+t/2,1} - p_{50,1}) + (p_{50,2} - p_{50-t/2,2})\right)/2} \tag{4.9}$$

Selecting the percentage t is a bit more difficult since this a somehow arbitrary choice, but we choose 68.27%. We do that because for a normal distribution, this is the interval of $\pm\sigma$ around the median (and $p_{50} = \mu$ in this case), so this criterion and (4.8) will match exactly. For non-normal distributions (e.g. with longer tail on one side), there will be certain differences, which is why using both criteria will be useful.

4.2. Results

After having introduced several criteria to assess differences between two probability distributions, we will now apply them to GW posteriors. Throughout this work, we will use the data with suffix `nocosmo` over `cosmo` because their sample size is typically higher, which will make the majority of our statements more reliable. Additionally, the statements on waveform differences inferred from each data should not differ much between them because the same cosmological model is applied to both going from `nocosmo` to `cosmo`.[3]

To actually examine the posteriors generated using different waveform models, we need two more things to make the criteria developed in the previous section applicable: (i) thresholds to mark significant differences and (ii) a set of parameters to examine. Both will be introduced in the first subsections of this section. The explicit waveforms we will deal with in this work are `IMRPhenomXPHM` [22] and `SEOBNRv4PHM` [9] because they were used for parameter estimation in GWTC-3.[4] They are two state-of-the-art models including precession effects as well as higher harmonics. A summary of the respective approaches and also more generally on topics related to waveforms is given in [23].

4.2.1. Thresholds

One method to find significant differences using criteria is to set thresholds for their values and essentially define significance as exceeding this threshold. We will use two of them, one being rather conservative and one more strict. Data failing the first threshold will surely not have good agreement and data passing the second threshold will surely have very good agreement. The explicit choices for each criterion are:

- For the JSD and JSD 2, we choose $D_{\mathrm{JS}} = 0.05$ and $D_{\mathrm{JS}} = 0.01$. The former is motivated by the catalogs since in appendix B.2. of GWTC-1 [2], there is the following statement: "The JSD values are, in general, smaller than approximately 0.05 bits, which indicates that the posteriors from the two BBH waveform models agree well", which further justifies our first threshold. The latter is motivated by a threshold of $D_{\mathrm{JS}} = 0.007$ applied in GWTC-2 [3], but is set a bit higher to take sampling variations into account and ensure that no event is marked falsely by the JSD (which appear during the computation of the JSD, as figure 4.1 shows).
- Further motivation for the thresholds for D_{JS} is given by the ones following now. For normal distributions, they correspond to a 50%, 20% mean shift in units of the standard deviation in case of $\sigma_1 = \sigma_2$ (at least roughly, depending on the sample size, as figure 4.1 shows). These are chosen as the thresholds, which means we use 0.5 and 0.2 as thresholds for the mean criterion (4.8).

[3] As it is stated in [1] and the abstract of [17], the prior of the luminosity distance is changed. This might have slightly different effects on different waveforms because the likelihoods $\mathcal{L}(h|d)$ are reweighted differently.

[4] Sometimes, samples from `IMRPhenomXPHM:HighSpin` and `IMRPhenomXPHM:LowSpin` were provided instead of those from `IMRPhenomXPHM`. In these cases, we chose to use samples from the `HighSpin` model. In any case, BBH waveforms are used.

In just the same manner, we choose thresholds of 0.5 and 0.2 for the median difference in units of the average credible interval (4.9). This statement will produce the exact same results like the one for mean, standard deviation if the compared distributions are normal. For non-normal distributions however, they may produce slightly different results, i.e. the means of two distributions might show significant differences according to this threshold, while the medians do not.

Figure 4.2 shows the chosen thresholds for normal and Rayleigh distributions as reference. One can see that the 20% threshold is indeed very strict as even for the non-normal Rayleigh distribution the credible intervals overlap to a very high degree, but would be marked as significantly different. The 50% threshold, on the other hand, produces more notable differences (but still not too severe ones), so it will suffice as a first filter.

4.2.2. Choice of Parameters

The goal of this subsection is to specify a set of parameters capturing the most important information from the relevant parameter space. This parameter space is high-dimensional, although black holes are, in principle, relatively simple objects having only a mass and spin (charges are neglected). More complexity is added when looking at binary systems, there are more intrinsic parameters and many combinations of these are interesting, too. Besides that, there are also extrinsic parameters describing properties of the GW detectors such as the SNR for every instrument and those are independent of waveform models. All in all, there are up to 172 parameters per posterior and it should be clear that we cannot to look at all of them (and that it also is not necessary).

The first intuition might be to choose the component masses m_i (where the convention $m_1 \geq m_2$ is adopted) and spin vectors $\mathbf{S}_i$ (or the dimensionless spin vectors $\chi_i = \frac{c\mathbf{S}_i}{Gm_i^2}$ which have magnitudes $\in [0,1]$). But it turns out that other parameters are better suited in the context of GWs, e.g. because they appear in important relations. To describe masses, we choose the chirp mass $\mathcal{M}$ and mass ratio q (measured in source frame, not detector), where

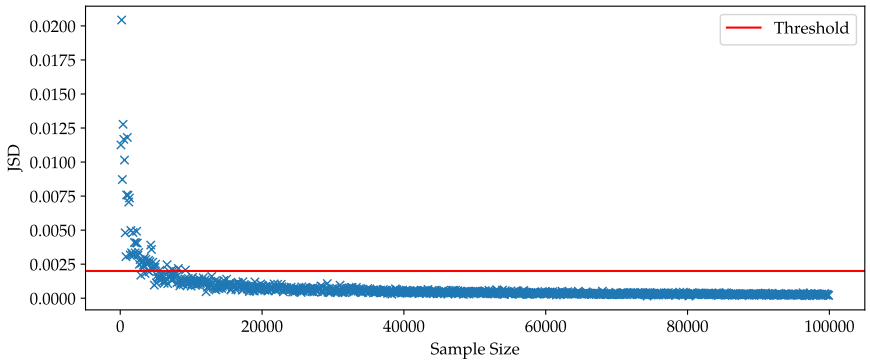
$$\mathcal{M} = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}}, \quad q = \frac{m_2}{m_1}\,. \tag{4.10}$$

For spins, we will use effective spin χ_{eff} and precession spin χ_p, where

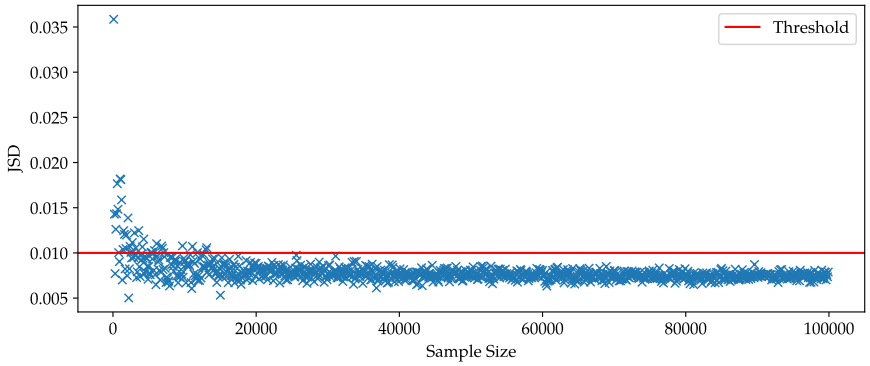
$$\chi_{\text{eff}} = \frac{(m_1\chi_1 + m_2\chi_2)\cdot\hat{\mathbf{L}}_N}{m_1 + m_2}, \quad \chi_p = \max\left\{\chi_{1,\perp}, \frac{q(4q+3)}{4+3q}\chi_{2,\perp}\right\}. \tag{4.11}$$

Here, $\hat{\mathbf{L}}_N$ is the Newtonian angular momentum perpendicular to the orbital plane and $\chi_{i,\perp}$ is the component of χ_i perpendicular to $\hat{\mathbf{L}}_N$. Other important parameters used are luminosity distance D_L and inclination θ_{jn}. Furthermore, the total mass $M = m_1 + m_2$ and the component masses m_1, m_2 are added in this section because the additional effort it requires is fairly small. In the end, we therefore use the following set:

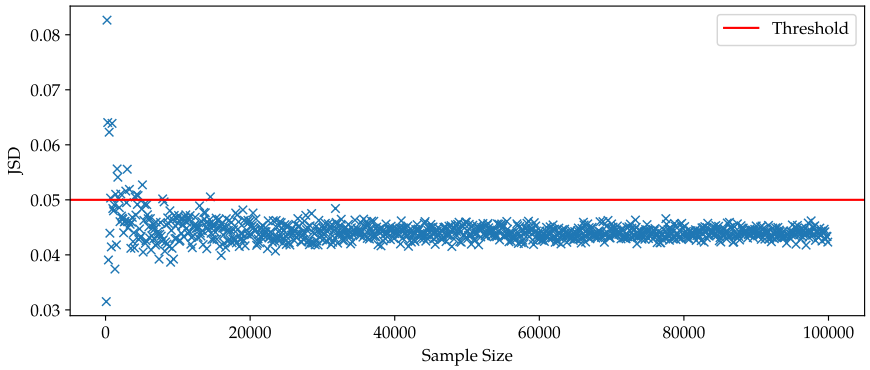
$$[\mathcal{M}, q, \chi_{\text{eff}}, \chi_p, D_L, \theta_{jn}, M, m_1, m_2]\,. \tag{4.12}$$



(a) Sampling error using two samples from the same normal distribution (0% shift)

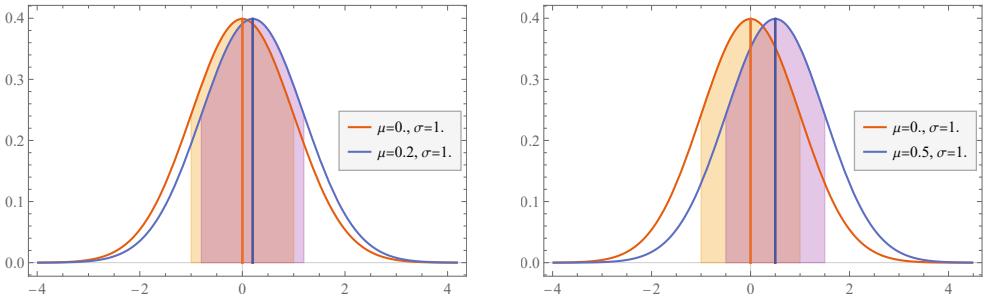


(b) Sampling error using samples from two normal distributions with 20% shift



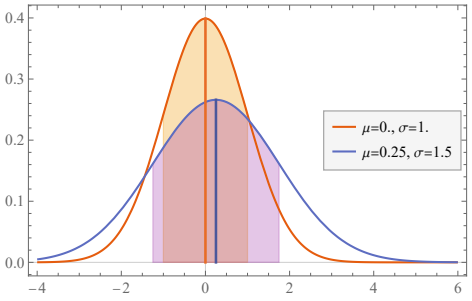
(c) Sampling error using samples from two normal distributions with 50% shift

Figure 4.1.: Sampling error for the JSDs of two normal distributions as a function of the sample size m (all have the same standard deviation $\sigma = 1$ and means are shifted by a certain amount in units of σ). These plots show how the finiteness of samples we work with leads to an uncertainty in the JSD values. With increasing m, these values converge against the analytical ones of $D_{\mathrm{JS}} = 0, 0.0072, 0.044$. Since the majority of posteriors samples have $m \gtrsim 10000$ for `SEOBNRv4PHM` and $m \gtrsim 100000$ for `IMRPhenomXPHM` (confer figure 4.5 (b) or 4.11 (d)), our thresholds $D_{\mathrm{JS}} = 0.002$ for sampling deviation ($\equiv 0\%$ shift; also mentioned in [3]) and $D_{\mathrm{JS}} = 0.01, 0.05$ for the 20%, 50% shifts should be more robust against sampling deviations than the analytic values (although we can never make completely sure to have eliminated them).

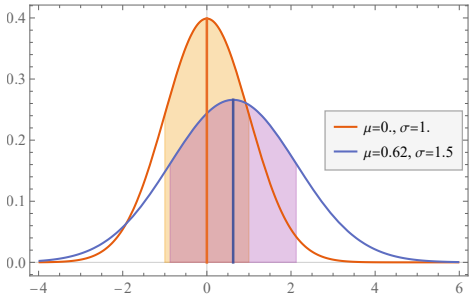


(a) $\sigma_1 = \sigma_2,\ |\mu_1 - \mu_2| = 0.2\ |\sigma_1 + \sigma_2|/2$

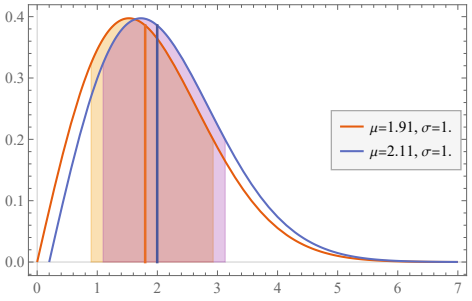
(b) $\sigma_1 = \sigma_2,\ |\mu_1 - \mu_2| = 0.5\ |\sigma_1 + \sigma_2|/2$



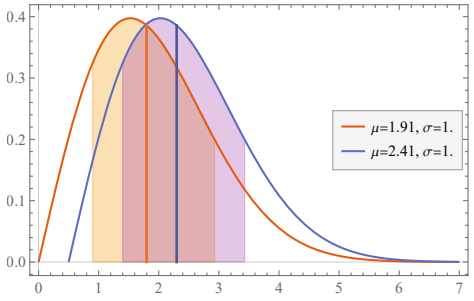
(c) $\sigma_1 \neq \sigma_2,\ |\mu_1 - \mu_2| = 0.2\ |\sigma_1 + \sigma_2|/2$



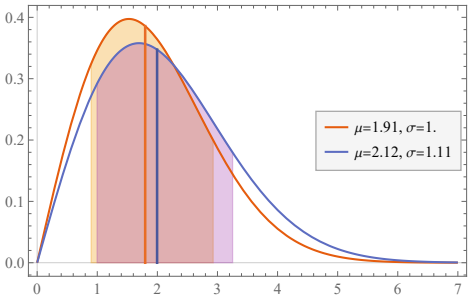
(d) $\sigma_1 \neq \sigma_2,\ |\mu_1 - \mu_2| = 0.5\ |\sigma_1 + \sigma_2|/2$



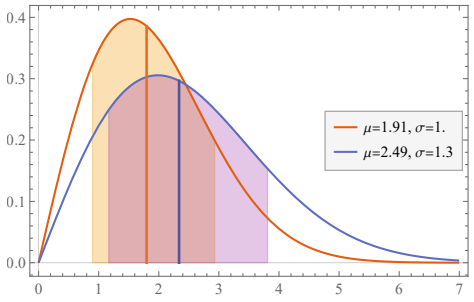
(e) $\sigma_1 = \sigma_2,\ |\mu_1 - \mu_2| = 0.2\ |\sigma_1 + \sigma_2|/2$



(f) $\sigma_1 = \sigma_2,\ |\mu_1 - \mu_2| = 0.5\ |\sigma_1 + \sigma_2|/2$



(g) $\sigma_1 \neq \sigma_2,\ |\mu_1 - \mu_2| = 0.2\ |\sigma_1 + \sigma_2|/2$



(h) $\sigma_1 \neq \sigma_2,\ |\mu_1 - \mu_2| = 0.5\ |\sigma_1 + \sigma_2|/2$

Figure 4.2.: Comparison of normal ((a) - (d)) and Rayleigh distributions ((e) - (h)). Every distribution is drawn along with the median and 68.27% credible interval. Although there are slight deviations between median, credible interval and mean, standard deviation in case of the Rayleigh distribution, μ, σ are given in the legend. That is because there are general formulas $\mu = c\sqrt{\pi/2}, \sigma = c\sqrt{(4-\pi)/2}$ for them, which allow to infer the parameter c of the distribution and thus reproduce these plots (for the median, to our knowledge, there is no such general formula).

4.2.3. On Agreement of All Events

Now that criteria, thresholds and parameters are all defined, we can start analysing events from GWTC-3 by comparing posteriors generated from different waveform models. The first way to do this is to look at each event separately and monitor which parameters show waveform differences that exceed our thresholds.

To further assess potential differences between `cosmo` and `nocosmo` data sets, results for both are calculated and they are shown in tables 4.1, 4.2, 4.3, 4.4, where each column has the following meaning: 1. contains the name for each event, for 2. the JSD is calculated, for 3. the JSD of normalized and centered data is calculated (the procedure described at the end of 4.1.1), 3. is the mean difference in units of average standard deviation and 4. is the median difference in units of the average credible interval. Each criterion is calculated for the selected set of parameters and only the parameters exceeding the respective criterion are given in the corresponding cell, while all that pass are not. To assess the results, we set a lower significance on shape differences than on mean/ median ones as it is more vulnerable against sampling deviations and depends more strongly on the chosen prior. The JSD measures all possible differences, so it is used as a first and main criterion.

The color encoding is as follows: green means that no parameter exceeds the given threshold, yellow means one did and red means two or more did. We do not mark an event significant immediately when the waveform difference for one of the parameters exceeds the respective threshold because that might just be due to the criterion chosen. The discussion in the previous section showed that there is no perfect criterion, they might produce false positives in some cases (event has no significant differences, but criterion says so).[5] That however could be just because the criterion is vulnerable to the specific (and often typical) shape of the posterior of a parameter and not because there really are significant differences. Therefore, if one criterion marks an event yellow and the others green, we will not prioritize taking a look at it. If, on the other hand, the others also mark it red, the priority will be high.

The following systematics can be extracted from the tables:

- Overall, no significant differences between `cosmo` and `nocosmo` data are present, so choosing to work with only one is reasonable. As already stated, the `nocosmo` data has a higher sample size (sometimes by factors of $3 - 10$) and the discussion of the influence of sample size on JSD values motivates this choice.
- Often, the JSD marks many parameters while median and mean criterion do not. One explanation of that could be shape differences, which would be recognized by the JSD, but not by mean and median criterion. Another reason might also be the JSD's sensitivity, i.e. even small shifts or small differences in peak height (which mostly comes from different widths and is especially present for narrow distributions) might lead to it marking significance. Another potential reason which could explain this observation is that there are differences in location, width and shape, which are all not

[5] We can live with that since we look at the significant event data anyway, which would will reveal the truth, and it was more important to find criteria which do not produce false negatives (event has significant differences, but criterion does not recognize them) because these would not be found despite being relevant.

Event	JSD	JSD 2	Mean difference	Median difference
GW191103_012549				
GW191105_143521				
GW191109_010717	$\chi_{\rm eff}$ χ_p D_L θ_{jn}	θ_{jn}	χ_p	χ_p D_L
GW191113_071753		χ_p		
GW191126_115259				
GW191127_050227	q M m_1	θ_{jn}	M m_1	M m_1
GW191129_134029				
GW191204_110529				
GW191204_171526		θ_{jn}		
GW191215_223052				
GW191216_213338	q $\chi_{\rm eff}$ M m_1 m_2	θ_{jn} M m_1	q M m_1 m_2	q M m_1 m_2
GW191219_163120	χ_p θ_{jn} M m_1	χ_p D_L θ_{jn}	χ_p	$\mathcal{M}$ q χ_p D_L θ_{jn} M m_1
GW191222_033537				
GW191230_180458		θ_{jn} m_1		
GW200105_162426	q $\chi_{\rm eff}$ χ_p M m_1 m_2	$\chi_{\rm eff}$	χ_p	χ_p
GW200112_155838	$\mathcal{M}$ q θ_{jn} m_2	θ_{jn}	$\mathcal{M}$ q θ_{jn} m_2	$\mathcal{M}$ q θ_{jn} m_2
GW200115_042309	χ_p	θ_{jn}	χ_p	χ_p
GW200128_022011	$\chi_{\rm eff}$		$\chi_{\rm eff}$	$\chi_{\rm eff}$
GW200129_065458	$\mathcal{M}$ q $\chi_{\rm eff}$ χ_p D_L θ_{jn} m_1 m_2	q χ_p θ_{jn}	$\mathcal{M}$ q χ_p θ_{jn} m_1 m_2	$\mathcal{M}$ q χ_p θ_{jn} m_1 m_2
GW200202_154313				
GW200208_130117		θ_{jn}		
GW200208_222617	$\mathcal{M}$ q $\chi_{\rm eff}$ M m_1	$\mathcal{M}$ q $\chi_{\rm eff}$ M m_1	$\mathcal{M}$ q $\chi_{\rm eff}$ M m_1	$\mathcal{M}$ q $\chi_{\rm eff}$ M m_1
GW200209_085452				
GW200210_092254	$\chi_{\rm eff}$ θ_{jn} M m_1	θ_{jn}		
GW200216_220804		θ_{jn}		
GW200219_094415				
GW200220_061928				
GW200220_124850				
GW200224_222234				
GW200225_060421				
GW200302_015811			m_2	m_2
GW200306_093714				
GW200308_173609	$\mathcal{M}$ M m_1	q M	$\mathcal{M}$ M m_1	M m_1
GW200311_115853				
GW200316_215756	q $\chi_{\rm eff}$ M m_1 m_2	q θ_{jn}	q $\chi_{\rm eff}$ M m_1 m_2	q $\chi_{\rm eff}$ M m_1 m_2
GW200322_091133	D_L M m_1	M m_1	D_L M m_1	D_L

Table 4.1.: Results for agreement of `IMRPhenomXPHM` and `SEOBNRv4PHM` for GWTC-3 events using 50% thresholds and `nocosmo` data

Event	JSD	JSD 2	Mean difference	Median difference
GW191103_012549				
GW191105_143521				
GW191109_010717	$\chi_{\rm eff}$ χ_p D_L θ_{jn}	θ_{jn}	χ_p D_L	χ_p D_L
GW191113_071753		q χ_p		χ_p
GW191126_115259				
GW191127_050227	q θ_{jn} M m_1	q θ_{jn}	q M m_1	q M m_1
GW191129_134029				
GW191204_110529				
GW191204_171526		θ_{jn}		
GW191215_223052				
GW191216_213338	q $\chi_{\rm eff}$ M m_1 m_2	θ_{jn} M m_1	q m_1 m_2	q M m_1 m_2
GW191219_163120	χ_p θ_{jn} M m_1	θ_{jn}	χ_p M m_1	q χ_p D_L θ_{jn} M m_1
GW191222_033537				
GW191230_180458		m_1		
GW200105_162426	q $\chi_{\rm eff}$ χ_p M m_1 m_2	$\chi_{\rm eff}$	χ_p	χ_p
GW200112_155838	$\mathcal{M}$ q θ_{jn} m_2	θ_{jn}	$\mathcal{M}$ q θ_{jn} m_2	$\mathcal{M}$ q θ_{jn} m_2
GW200115_042309	χ_p	θ_{jn}	χ_p	χ_p
GW200128_022011	$\chi_{\rm eff}$		$\chi_{\rm eff}$	$\chi_{\rm eff}$
GW200129_065458	$\mathcal{M}$ q $\chi_{\rm eff}$ χ_p D_L θ_{jn} m_1 m_2	q χ_p θ_{jn}	$\mathcal{M}$ q χ_p θ_{jn} m_1 m_2	$\mathcal{M}$ q χ_p D_L θ_{jn} m_1 m_2
GW200202_154313				
GW200208_130117				
GW200208_222617	$\mathcal{M}$ q $\chi_{\rm eff}$ M m_1 m_2	$\mathcal{M}$ q $\chi_{\rm eff}$ M m_1	$\mathcal{M}$ q $\chi_{\rm eff}$ M m_1	$\mathcal{M}$ q $\chi_{\rm eff}$ M m_1 m_2
GW200209_085452				
GW200210_092254	q $\chi_{\rm eff}$ θ_{jn} M m_1	θ_{jn}		
GW200216_220804				
GW200219_094415				
GW200220_061928		M m_1		
GW200220_124850				
GW200224_222234				
GW200225_060421				
GW200302_015811				
GW200306_093714				
GW200308_173609	$\mathcal{M}$ M m_1	m_1	M m_1	
GW200311_115853				
GW200316_215756	q $\chi_{\rm eff}$ M m_1 m_2	q θ_{jn}	q $\chi_{\rm eff}$ M m_1 m_2	q $\chi_{\rm eff}$ M m_1 m_2
GW200322_091133	$\mathcal{M}$ $\chi_{\rm eff}$ χ_p D_L M m_1	D_L M m_1	D_L	D_L

Table 4.2.: Results for agreement of `IMRPhenomXPHM` and `SEOBNRv4PHM` for GWTC-3 events using 50% thresholds and `cosmo` data

Event	JSD	JSD 2	Mean difference	Median difference
GW191103_012549	$q\ \chi_{\rm eff}\ \chi_p\ m_1\ m_2$	$\theta_{jn}\ M$	$\chi_p\ m_1$	χ_p
GW191105_143521	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$\theta_{jn}\ M\ m_1$	$q\ \chi_p\ M\ m_1\ m_2$	$q\ m_1\ m_2$
GW191109_010717	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_2$	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_2$	$\chi_p\ D_L$	$\chi_p\ D_L\ \theta_{jn}$
GW191113_071753	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1\ m_2$	$\chi_p\ D_L$	$q\ \chi_p\ D_L\ M\ m_1\ m_2$
GW191126_115259	θ_{jn}	θ_{jn}		θ_{jn}
GW191127_050227	$\mathcal{M}\ q\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_p\ \theta_{jn}$	$q\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \theta_{jn}\ M\ m_1\ m_2$
GW191129_134029	$q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \theta_{jn}\ m_1\ m_2$	$q\ M\ m_1\ m_2$	$q\ \theta_{jn}\ M\ m_1\ m_2$
GW191204_110529	$\chi_{\rm eff}\ \chi_p\ \theta_{jn}$	$\chi_p\ \theta_{jn}\ M\ m_1$	$\chi_{\rm eff}\ \chi_p$	$\chi_{\rm eff}\ \chi_p\ \theta_{jn}$
GW191204_171526	$\mathcal{M}\ q\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \theta_{jn}\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_p\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$
GW191215_223052	$\chi_{\rm eff}\ \theta_{jn}$	θ_{jn}	$\chi_{\rm eff}$	
GW191216_213338	$q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$
GW191219_163120	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1$	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1$
GW191222_033537	$\chi_{\rm eff}\ D_L$		$\chi_{\rm eff}\ D_L$	$\chi_{\rm eff}\ D_L$
GW191230_180458	M	$\theta_{jn}\ M\ m_1$		$\mathcal{M}\ \theta_{jn}\ M$
GW200105_162426	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	χ_p	χ_p
GW200112_155838	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_p\ \theta_{jn}$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$
GW200115_042309	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \theta_{jn}\ M\ m_1\ m_2$	$\chi_p\ \theta_{jn}$	$\chi_p\ \theta_{jn}$
GW200128_022011	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_2$	χ_p	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_2$	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_2$
GW200129_065458	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ m_1\ m_2$	$q\ \chi_p\ D_L\ \theta_{jn}\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ m_1\ m_2$
GW200202_154313	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$\chi_{\rm eff}\ M\ m_1$		
GW200208_130117	$\chi_{\rm eff}$	θ_{jn}	$\chi_{\rm eff}$	$\chi_{\rm eff}$
GW200208_222617	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ M\ m_1$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$
GW200209_085452	$\mathcal{M}\ M$	χ_p	$\mathcal{M}\ M$	$\mathcal{M}\ M$
GW200210_092254	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ \chi_p\ D_L\ \theta_{jn}$	$\chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$
GW200216_220804	M	$\theta_{jn}\ M\ m_1$		$\chi_{\rm eff}$
GW200219_094415	$\chi_{\rm eff}$	θ_{jn}	$\chi_{\rm eff}\ \chi_p$	$\chi_{\rm eff}\ \chi_p$
GW200220_061928	$\mathcal{M}\ M\ m_1\ m_2$	$\theta_{jn}\ M\ m_1$	$\mathcal{M}\ M\ m_2$	$\mathcal{M}\ M\ m_1\ m_2$
GW200220_124850		θ_{jn}	m_1	
GW200224_222234	$\mathcal{M}\ q\ \theta_{jn}\ m_2$	θ_{jn}	$\mathcal{M}\ q\ \theta_{jn}\ m_2$	$\mathcal{M}\ q\ \theta_{jn}\ m_2$
GW200225_060421	$\chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}$	θ_{jn}	$\chi_{\rm eff}\ \chi_p\ D_L$	$\chi_{\rm eff}\ D_L\ \theta_{jn}$
GW200302_015811	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \theta_{jn}$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ M\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ m_1\ m_2$
GW200306_093714		$\theta_{jn}\ M\ m_1$		θ_{jn}
GW200308_173609	$\mathcal{M}\ q\ M\ m_1\ m_2$	$\mathcal{M}\ q\ M\ m_1\ m_2$	$\mathcal{M}\ q\ M\ m_1$	$\mathcal{M}\ q\ M\ m_1$
GW200311_115853	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ M\ m_2$		$\mathcal{M}\ q\ \chi_{\rm eff}\ M\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ M\ m_2$
GW200316_215756	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1\ m_2$
GW200322_091133	$\mathcal{M}\ q\ D_L\ M\ m_1$	$\mathcal{M}\ q\ D_L\ M\ m_1$	$\mathcal{M}\ D_L\ M\ m_1$	$\mathcal{M}\ D_L\ M\ m_1$

Table 4.3.: Results for agreement of `IMRPhenomXPHM` and `SEOBNRv4PHM` for GWTC-3 events using 20% thresholds and `nocosmo` data

Event	JSD	JSD 2	Mean difference	Median difference
GW191103_012549	$q\ \chi_p\ \theta_{jn}\ m_1\ m_2$	$\theta_{jn}\ M$	χ_p	χ_p
GW191105_143521	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$\chi_{\rm eff}\ \theta_{jn}\ M$	$q\ \chi_p\ m_1\ m_2$	$q\ m_1\ m_2$
GW191109_010717	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_2$	$\chi_p\ D_L$	$\chi_p\ D_L\ \theta_{jn}$
GW191113_071753	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_p\ D_L\ M\ m_1\ m_2$	$q\ \chi_p\ D_L\ M\ m_1\ m_2$
GW191126_115259	$\chi_p\ \theta_{jn}$	θ_{jn}	χ_p	
GW191127_050227	$\mathcal{M}\ q\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_p\ \theta_{jn}\ m_2$	$q\ M\ m_1\ m_2$	$q\ \theta_{jn}\ M\ m_1\ m_2$
GW191129_134029	$q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \theta_{jn}\ m_1\ m_2$	$q\ M\ m_1\ m_2$	$q\ \theta_{jn}\ M\ m_1\ m_2$
GW191204_110529	$\chi_{\rm eff}\ \chi_p\ \theta_{jn}$	$\chi_p\ \theta_{jn}\ M\ m_1$	χ_p	$\chi_{\rm eff}\ \chi_p\ \theta_{jn}$
GW191204_171526	$\mathcal{M}\ q\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \theta_{jn}\ m_1\ m_2$	$q\ \chi_p\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$
GW191215_223052	$\chi_{\rm eff}\ \theta_{jn}$	θ_{jn}	$\chi_{\rm eff}$	$\chi_{\rm eff}$
GW191216_213338	$q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$
GW191219_163120	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1$	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1$
GW191222_033537	$\chi_{\rm eff}\ D_L$		$\chi_{\rm eff}\ D_L$	$\chi_{\rm eff}\ D_L$
GW191230_180458		$\theta_{jn}\ M\ m_1$	θ_{jn}	θ_{jn}
GW200105_162426	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	χ_p	χ_p
GW200112_155838	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_p\ \theta_{jn}$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$
GW200115_042309	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \theta_{jn}\ M\ m_1\ m_2$	$\chi_p\ \theta_{jn}$	$\chi_p\ \theta_{jn}$
GW200128_022011	$\mathcal{M}\ \chi_{\rm eff}\ D_L\ M$		$\mathcal{M}\ \chi_{\rm eff}\ D_L$	$\mathcal{M}\ \chi_{\rm eff}\ D_L\ M$
GW200129_065458	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ m_1\ m_2$	$q\ \chi_p\ D_L\ \theta_{jn}\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_p\ D_L\ \theta_{jn}\ m_1\ m_2$
GW200202_154313	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$\chi_{\rm eff}\ M\ m_1$		
GW200208_130117	$\chi_{\rm eff}$	θ_{jn}	$\chi_{\rm eff}$	$\chi_{\rm eff}$
GW200208_222617	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1\ m_2$
GW200209_085452	D_L	χ_p		
GW200210_092254	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_p\ \theta_{jn}$	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$
GW200216_220804	M	$\theta_{jn}\ M\ m_1$		
GW200219_094415	$\chi_{\rm eff}\ D_L$	θ_{jn}	$\chi_{\rm eff}\ D_L$	$\chi_{\rm eff}\ D_L$
GW200220_061928	$\mathcal{M}\ q\ \chi_{\rm eff}\ M\ m_1\ m_2$	$\mathcal{M}\ M\ m_1$	$\chi_{\rm eff}\ m_2$	$\mathcal{M}\ \chi_{\rm eff}\ M\ m_2$
GW200220_124850		θ_{jn}	m_1	
GW200224_222234	$\mathcal{M}\ q\ \theta_{jn}\ m_2$	θ_{jn}	$q\ \theta_{jn}\ m_2$	$q\ \theta_{jn}$
GW200225_060421	$\chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}$	θ_{jn}	$\chi_p\ D_L$	$\chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}$
GW200302_015811	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \theta_{jn}$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ m_1\ m_2$
GW200306_093714		$\theta_{jn}\ M\ m_1$		θ_{jn}
GW200308_173609	$\mathcal{M}\ q\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ q\ M\ m_1$	$\mathcal{M}\ q\ M\ m_1$
GW200311_115853	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ M\ m_2$	χ_p	$\mathcal{M}\ q\ \chi_{\rm eff}\ M\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ M\ m_2$
GW200316_215756	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1\ m_2$
GW200322_091133	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_p\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ D_L\ m_2$

Table 4.4.: Results for agreement of `IMRPhenomXPHM` and `SEOBNRv4PHM` for GWTC-3 events using 20% thresholds and `cosmo` data

significant enough to be marked by the dedicated criteria, but add up such that the overall differences lead to the JSD recognizing them.

This will appear again in the next subsection when events with bad agreement are discussed and shows that, although the JSD is a very useful tool, it still remains valuable to use other criteria in addition to it or even look at the posteriors.

- The 20% thresholds are not very useful or reliable since many parameters for almost every event are marked to have significant difference. That does not mean the overall data quality is bad, it simply comes from the threshold being very strict (figure 4.2 already displayed this), especially when considering sampling effects. It was motivated from GWTC-2.1, where its purpose was to decide on the use of higher harmonics and whether to use mixed samples, so the intention was a bit different and it makes sense that differences also had to be assessed differently by choosing a stricter threshold.

  The 50% threshold on the other hand produces results that can be used much better (some events are marked significant, but not too much).

4.2.4. On Agreement of Specific Events

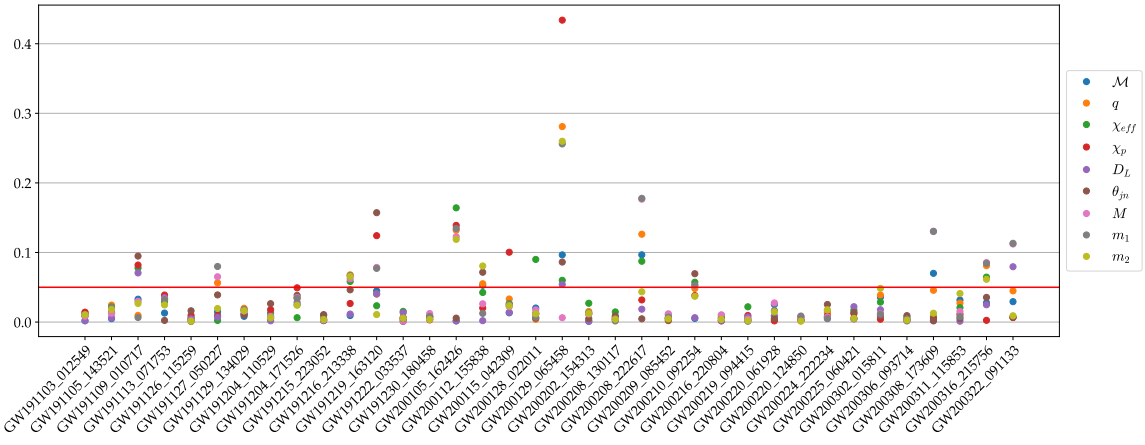
It is the purpose of this subsection to assess which events exhibit extended deviations. Those will be collected and referred to as "bad data". Since we do not know how GW posteriors should look like or perform under application of our criteria, reference is needed (also useful for extraction of potential systematics when the agreement is bad/ good). This will be provided by events with particularly good agreement ("good data").

For the actual classifications, the tables from the previous subsection will be used. Events marked red for 50% thresholds (and `nocosmo` data) are taken as bad data, while events which are green for 50% thresholds and perform well for 20% thresholds are taken as good data. These two lists will then be verified by looking at the posteriors to compare them (this will be interesting to find false positives, i.e. events in bad data which are not really bad; for good data this will very likely not be the case because the JSD is so sensitive). Additionally, it is also possible to visualize the actual, explicit values of the criteria that were used to obtain tables 4.1, 4.2, 4.3, 4.4 (their purpose was to give an overview, which is why they do not display these values) and therefore presenting more details from the information used to create them (figure 4.3).

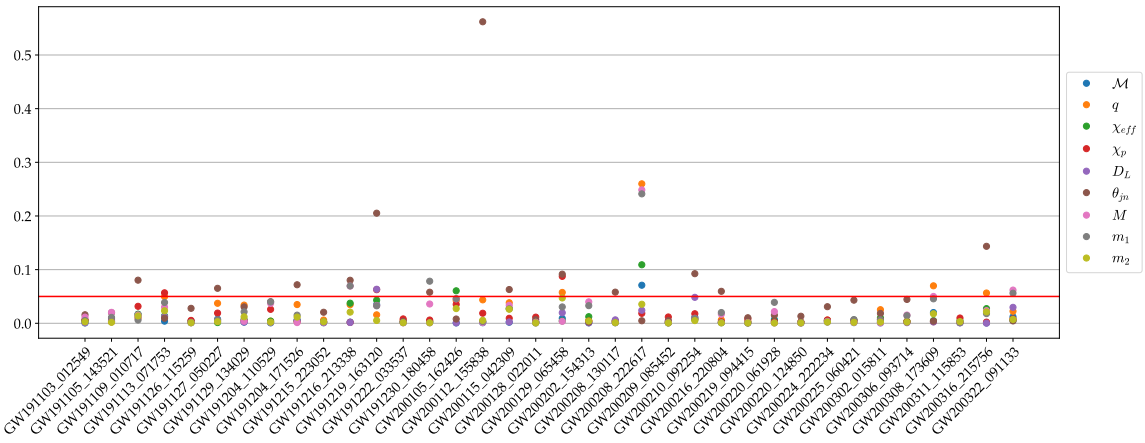
We start by listing good data:

GW191126_115259, GW191215_223052, GW191222_033537, GW200208_13011, GW200209_085452, GW200216_220804, GW200220_124850, GW200306_093714 .

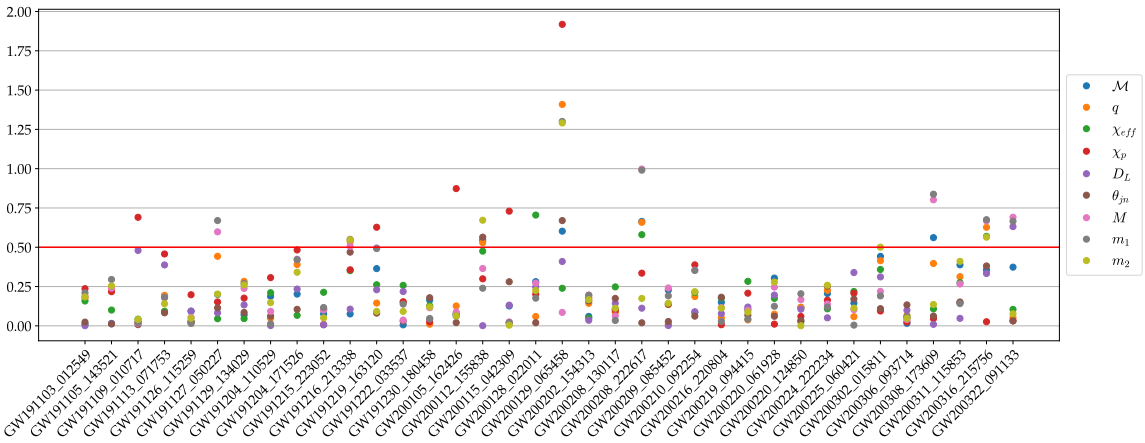
Although this work often focuses on the negative aspects and examples of GW data, this is a good place to note that bad agreement is not the regular case. Many events show good data quality and some of them (from good data) even show incredible agreement.



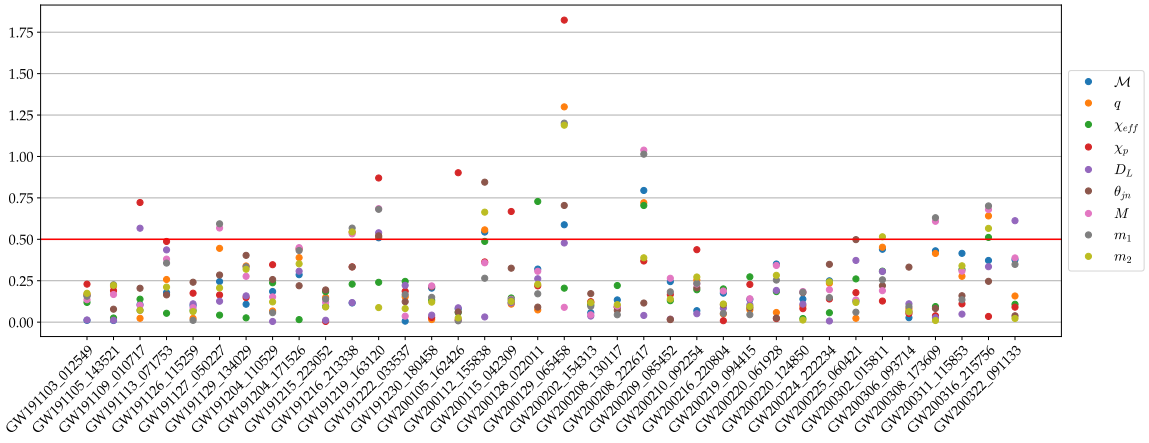
(a) JSD



(b) JSD 2



(c) Mean criterion



(d) Median criterion

Figure 4.3.: Agreement for all events from GWTC-3 and set of parameters specified in plot legend. Each dot represents the criterion value for a parameter and the 50% threshold for the respective criterion is visualized as a red line.

Events where that is certainly not the case are collected now in the list of bad data:

- GW191109_010717: q, χ_{eff} have exotic peak shapes for the Phenom model, which almost look like multimodality and probably cause bad agreement. Furthermore, χ_p, D_L, θ_{jn} do not agree well.
- GW191127_050227: q has bad agreement, for χ_p it is also not too good (but not too bad either). Compared to other events, the component masses and total mass have bad agreement as well.
- GW191216_213338: a slight shift is present in $\mathcal{M}$ and there might be multimodality in q for the Phenom model (at very low values) and in the component masses.
- GW191219_163120: all distributions are very narrow, so significance might be marked fairly fast. However, χ_p still looks bad and so do component masses, θ_{jn}. An interesting fact is that both waveform models have a sample size of exactly 20000. This event is mentioned in GWTC-3 because it has significant support for mass ratios in non-calibrated regions plus potential uncertainties in p_{astro}.
- *GW200105_162426*: this event is very interesting. First of all, GWTC-3 mentions it because it has $p_{\text{astro}} < 0.5$ (but is still included in the analyses). It is then often the case that the JSD marks significant differences because the Phenom model has slightly wider distributions, so the peak height differs and makes distributions look more different than they really are. Since that is not really what one would consider a significant difference and because mean, median criterion do not mark this event red, we exclude it from the list bad data. Also interesting is that almost all distributions have a Gaussian distribution, even for parameters that usually have different shapes (like q, M, m_1, m_2).

- GW200112_155838: $\mathcal{M}$ has a slight shift (relatively much for this parameter). q, χ_p, the component masses and to some degree χ_{eff} also do not agree very well.
- GW200129_065458: has clearly bad agreement (even for $\mathcal{M}$; M on the other hand is not affected by overall bad agreement), especially for the mass parameters due to multimodality in the Phenom posterior. χ_p also has exceptionally bad agreement along with an usual shape for this parameter (that there is almost no overlap also shows up in our criteria, the value for χ_p is the maximum throughout the whole catalog and all parameters for JSD, mean criterion and median criterion) and D_L, θ_{jn} do not look too good either (and their shapes are unusual, too). Again, it is not surprising that we find this event, (i) because it has a very high SNR of 26.8 where worse agreement is known/ expected and (ii) it caused some controversy in the past and is in fact still discussed in papers ([24] is a very recent example), partially due to the presence of a glitch. Besides that, its sample size of the EOB posterior is fairly small, so sampling deviations might play a big role (especially for the JSDs).
- GW200208_222617: multimodality occurs for both waveform models and shows up in $\mathcal{M}$, q (which also has a very small value), M, m_1 and χ_{eff}. Also, despite similar shapes, shifts are present in χ_p, D_L.
- *GW200210_092254*: only χ_p and maybe θ_{jn} show deviations, but overall, the agreement should not really be marked significant (mean, median criterion do not) and this event is excluded from bad data.
- GW200308_173609: both models show multimodality in $\mathcal{M}, m_2$ and also have remarkably different q. In fact, q looks almost uniform for SEOB and D_L looks uniform for both, which is very unusual. Lastly, θ_{jn} also does not have as deep of a valley between the double peaks as it is usual.
- GW200316_215756: shift in $\mathcal{M}$ despite very similar shapes. q has remarkably bad agreement and it looks like multimodality occurs in both waveform models (though it is not clear to see). Other parameters which show signs of multimodality are χ_{eff}, m_2.
- GW200322_091133: this event is a very interesting and exotic case because the posterior for every parameter are very similar to the prior (so much that it seems like almost no information comes from the data). Despite that, there still are some differences, but they might also be caused by the exceptionally small sample sizes for the EOB model (981 for nocosmo and 46 for cosmo) and the big uncertainties in our results which follow from that.

Those events indeed show above average deviations, although they at least have similar supports for the most part (especially for the masses where priors cover a wide range, this at least shows consistency to some degree). However, during the manual posterior check it turned out that bad data indeed contained false positives (events which do not actually have bad agreement) and they are highlighted in italics. The resulting list contains ten elements, which we can now compare to the results obtained in section V.E. of GWTC-3, where waveform systematics are explored. Events found to have significant differences here are GW191109_010717, GW191219_163120, GW200129_065458, GW200208_222617 and they are also all recognized by our criteria, i.e. they are in bad data. However, our criteria expand this

list by adding GW191127_050227, GW191216_213338, GW200112_155838, GW200308_173609, GW200316_215756, GW200322_091133. Apart from GWTC-3, there is also a recent analysis [25] on how even Gaussian noise might affect results and they find that it is very likely that GW200308_173609 and in particular GW200322_091133 are not astrophysical signals, but instead Gaussian noise. This matches our findings in a sense that these events show significant waveform inconsistencies (which would make sense if there is no underlying astrophysical signal to agree on; moreover, it would explain the lack of prior differences for GW200322_091133, which we observed when looking at the posteriors).

Besides events with significant posterior differences, section V of GWTC-3 also lists events with multimodality in the posterior of certain parameters. This effect is most significant for GW200208_222617, GW200308_173609, GW200322_091133, but it also occurs for GW200129_065458, GW200225_060421, GW200306_093714. There is a certain overlap with the previous list (which is not surprising since such a complicated posterior can very likely lead to waveform inconsistencies) and the significant ones are in bad data as well. A particularly interesting event in this context is GW200306_093714, which shows multimodality in the redshifted chirp mass, but this disappears in the source frame and it is one of the few events to show exceptionally good agreement even for the 20% thresholds (so it is in good data).

One might now argue that validation by hand (eye) is not an appropriate way to confirm agreement and it is indeed subjective to some degree (plus it is convenient only for a relatively small number of events). A more objective way is to look at a different representation of the information contained in the big tables by looking at how many events out of all, bad and good data fail the thresholds (table 4.5; table 4.6 provides the corresponding numbers for the `cosmo` data to show that no significant differences occur here as well), which shows that bad data indeed has below average agreement and good data above average agreement (compared to corresponding numbers for all events/ all data as an average).

4.2.5. On Parameter Agreement

Besides showing that good and bad data deserve their respective names, the rows for all data in table 4.5 also reveal other interesting systematics. For a better visualization of them, it is useful to rank the parameters according to their agreement (for each criterion; table 4.7).[6] To give some context for these values, we further provide the difference between prior and (`Mixed`) posterior as measured by KLD and JSD, average over all events from GWTC-3. This is potentially interesting because for prior-dominated posteriors (where the signal and therefore likelihood as a data-, waveform-dependent term does not change as a function of the parameter, i.e. the likelihood only has a small influence on the result), less waveform

[6] A quick remark on the way we quantify agreement: one might argue that violation of a threshold is no good criterion because the thresholds are arbitrary. Although there were reasons to select them, this arbitrariness is still correct. However, other statistics like the average JSD for each parameter do not reveal very different systematics and they also have some caveats (like vulnerability to outliers for the average, which e.g. plays a huge role for χ_p as it has a JSD value of 0.4 for one event, the maximum value we find out of all parameters).

	Parameter	JSD		JSD 2		Mean difference		Median difference	
		50%	20%	50%	20%	50%	20%	50%	20%
Bad data	$\mathcal{M}$	4/ 10 (40%)	9/ 10 (90%)	1/ 10 (10%)	5/ 10 (50%)	4/ 10 (40%)	7/ 10 (70%)	4/ 10 (40%)	8/ 10 (80%)
	q	6/ 10 (60%)	9/ 10 (90%)	4/ 10 (40%)	9/ 10 (90%)	5/ 10 (50%)	7/ 10 (70%)	6/ 10 (60%)	8/ 10 (80%)
	χ_{eff}	5/ 10 (50%)	7/ 10 (70%)	1/ 10 (10%)	5/ 10 (50%)	2/ 10 (20%)	6/ 10 (60%)	2/ 10 (20%)	6/ 10 (60%)
	χ_p	3/ 10 (30%)	6/ 10 (60%)	2/ 10 (20%)	6/ 10 (60%)	3/ 10 (30%)	6/ 10 (60%)	3/ 10 (30%)	6/ 10 (60%)
	D_L	3/ 10 (30%)	7/ 10 (70%)	1/ 10 (10%)	5/ 10 (50%)	1/ 10 (10%)	5/ 10 (50%)	3/ 10 (30%)	5/ 10 (50%)
	θ_{jn}	4/ 10 (40%)	7/ 10 (70%)	7/ 10 (70%)	7/ 10 (70%)	2/ 10 (20%)	4/ 10 (40%)	3/ 10 (30%)	7/ 10 (70%)
	M	7/ 10 (70%)	9/ 10 (90%)	4/ 10 (40%)	7/ 10 (70%)	6/ 10 (60%)	8/ 10 (80%)	6/ 10 (60%)	8/ 10 (80%)
	m_1	8/ 10 (80%)	9/ 10 (90%)	3/ 10 (30%)	7/ 10 (70%)	7/ 10 (70%)	9/ 10 (90%)	7/ 10 (70%)	9/ 10 (90%)
	m_2	4/ 10 (40%)	9/ 10 (90%)	0/ 10 (0%)	6/ 10 (60%)	4/ 10 (40%)	5/ 10 (50%)	4/ 10 (40%)	6/ 10 (60%)
All data	$\mathcal{M}$	4/ 36 (11%)	18/ 36 (50%)	1/ 36 (3%)	6/ 36 (17%)	4/ 36 (11%)	14/ 36 (39%)	4/ 36 (11%)	16/ 36 (44%)
	q	7/ 36 (19%)	21/ 36 (58%)	4/ 36 (11%)	15/ 36 (42%)	5/ 36 (14%)	13/ 36 (36%)	6/ 36 (17%)	16/ 36 (44%)
	χ_{eff}	8/ 36 (22%)	24/ 36 (67%)	2/ 36 (6%)	10/ 36 (28%)	3/ 36 (8%)	16/ 36 (44%)	3/ 36 (8%)	15/ 36 (42%)
	χ_p	5/ 36 (14%)	19/ 36 (53%)	3/ 36 (8%)	12/ 36 (33%)	5/ 36 (14%)	17/ 36 (47%)	5/ 36 (14%)	15/ 36 (42%)
	D_L	3/ 36 (8%)	15/ 36 (42%)	1/ 36 (3%)	7/ 36 (19%)	1/ 36 (3%)	11/ 36 (31%)	3/ 36 (8%)	11/ 36 (31%)
	θ_{jn}	5/ 36 (14%)	17/ 36 (47%)	13/ 36 (36%)	26/ 36 (72%)	2/ 36 (6%)	6/ 36 (17%)	3/ 36 (8%)	18/ 36 (50%)
	M	9/ 36 (25%)	24/ 36 (67%)	4/ 36 (11%)	18/ 36 (50%)	6/ 36 (17%)	17/ 36 (47%)	6/ 36 (17%)	17/ 36 (47%)
	m_1	10/ 36 (28%)	20/ 36 (56%)	4/ 36 (11%)	19/ 36 (53%)	7/ 36 (19%)	15/ 36 (42%)	7/ 36 (19%)	16/ 36 (44%)
	m_2	5/ 36 (14%)	23/ 36 (64%)	0/ 36 (0%)	11/ 36 (31%)	5/ 36 (14%)	14/ 36 (39%)	5/ 36 (14%)	16/ 36 (44%)
Good data	$\mathcal{M}$	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	1/ 8 (12%)
	q	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)
	χ_{eff}	0/ 8 (0%)	3/ 8 (38%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	3/ 8 (38%)	0/ 8 (0%)	3/ 8 (38%)
	χ_p	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)
	D_L	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	1/ 8 (12%)
	θ_{jn}	0/ 8 (0%)	2/ 8 (25%)	2/ 8 (25%)	6/ 8 (75%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	2/ 8 (25%)
	M	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	1/ 8 (12%)
	m_1	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)
	m_2	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)

Table 4.5.: Event statistics for GWTC-3 data. The numbers and percentages show how often each parameter exceeds the respective threshold of the column.
All data provides an average value of how many events from GWTC-3 fail the chosen thresholds (≡ summary of tables 4.1, 4.3), which can then be compared to these respective numbers for our list of bad, good data (`nocosmo` data used).

Eventlist	Parameter	JSD		JSD 2		Mean difference		Median difference	
		50%	20%	50%	20%	50%	20%	50%	20%
Bad data	$\mathcal{M}$	5/ 10 (50%)	9/ 10 (90%)	1/ 10 (10%)	5/ 10 (50%)	3/ 10 (30%)	7/ 10 (70%)	3/ 10 (30%)	6/ 10 (60%)
	q	6/ 10 (60%)	10/ 10 (100%)	4/ 10 (40%)	9/ 10 (90%)	6/ 10 (60%)	7/ 10 (70%)	7/ 10 (70%)	9/ 10 (90%)
	χ_{eff}	6/ 10 (60%)	8/ 10 (80%)	1/ 10 (10%)	5/ 10 (50%)	2/ 10 (20%)	7/ 10 (70%)	2/ 10 (20%)	6/ 10 (60%)
	χ_p	4/ 10 (40%)	8/ 10 (80%)	1/ 10 (10%)	6/ 10 (60%)	3/ 10 (30%)	7/ 10 (70%)	3/ 10 (30%)	7/ 10 (70%)
	D_L	3/ 10 (30%)	8/ 10 (80%)	1/ 10 (10%)	6/ 10 (60%)	2/ 10 (20%)	5/ 10 (50%)	4/ 10 (40%)	6/ 10 (60%)
	θ_{jn}	5/ 10 (50%)	9/ 10 (90%)	7/ 10 (70%)	7/ 10 (70%)	2/ 10 (20%)	4/ 10 (40%)	3/ 10 (30%)	7/ 10 (70%)
	M	7/ 10 (70%)	9/ 10 (90%)	3/ 10 (30%)	7/ 10 (70%)	5/ 10 (50%)	8/ 10 (80%)	5/ 10 (50%)	7/ 10 (70%)
	m_1	8/ 10 (80%)	9/ 10 (90%)	4/ 10 (40%)	7/ 10 (70%)	7/ 10 (70%)	9/ 10 (90%)	6/ 10 (60%)	8/ 10 (80%)
	m_2	5/ 10 (50%)	10/ 10 (100%)	0/ 10 (0%)	8/ 10 (80%)	4/ 10 (40%)	7/ 10 (70%)	5/ 10 (50%)	7/ 10 (70%)
All data	$\mathcal{M}$	5/ 36 (14%)	17/ 36 (47%)	1/ 36 (3%)	6/ 36 (17%)	3/ 36 (8%)	10/ 36 (28%)	3/ 36 (8%)	11/ 36 (31%)
	q	8/ 36 (22%)	23/ 36 (64%)	5/ 36 (14%)	16/ 36 (44%)	6/ 36 (17%)	15/ 36 (42%)	7/ 36 (19%)	17/ 36 (47%)
	χ_{eff}	9/ 36 (25%)	25/ 36 (69%)	2/ 36 (6%)	11/ 36 (31%)	3/ 36 (8%)	16/ 36 (44%)	3/ 36 (8%)	17/ 36 (47%)
	χ_p	6/ 36 (17%)	21/ 36 (58%)	2/ 36 (6%)	12/ 36 (33%)	5/ 36 (14%)	17/ 36 (47%)	6/ 36 (17%)	15/ 36 (42%)
	D_L	3/ 36 (8%)	18/ 36 (50%)	1/ 36 (3%)	6/ 36 (17%)	2/ 36 (6%)	11/ 36 (31%)	4/ 36 (11%)	13/ 36 (36%)
	θ_{jn}	6/ 36 (17%)	20/ 36 (56%)	10/ 36 (28%)	25/ 36 (69%)	2/ 36 (6%)	7/ 36 (19%)	3/ 36 (8%)	16/ 36 (44%)
	M	9/ 36 (25%)	22/ 36 (61%)	4/ 36 (11%)	18/ 36 (50%)	5/ 36 (14%)	13/ 36 (36%)	5/ 36 (14%)	14/ 36 (39%)
	m_1	10/ 36 (28%)	20/ 36 (56%)	6/ 36 (17%)	18/ 36 (50%)	7/ 36 (19%)	16/ 36 (44%)	6/ 36 (17%)	14/ 36 (39%)
	m_2	6/ 36 (17%)	23/ 36 (64%)	0/ 36 (0%)	13/ 36 (36%)	4/ 36 (11%)	16/ 36 (44%)	5/ 36 (14%)	15/ 36 (42%)
Good data	$\mathcal{M}$	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)
	q	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)
	χ_{eff}	0/ 8 (0%)	3/ 8 (38%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	3/ 8 (38%)	0/ 8 (0%)	3/ 8 (38%)
	χ_p	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)
	D_L	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	1/ 8 (12%)
	θ_{jn}	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	6/ 8 (75%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)
	M	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)
	m_1	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)
	m_2	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)

Table 4.6.: Event statistics for GWTC-3 data. The only difference between this table and table 4.5 is that the `cosmo` data has been used here to again show the broad overall agreement of both models (which means it is a summary of tables 4.2, 4.4).

differences are expected since the waveforms have a smaller effect on the resulting posterior. Hence, we should take this effect into account when assessing the results.

Several interesting features appear in both tables (we focus on results for `nocosmo` data):

- Masses are known to have a significant impact on the morphology of GW signals, especially chirp mass $\mathcal{M}$ and total mass M. Spins on the other hand leave a smaller imprint and therefore their values harder to infer from the signal (signals do not change much as a function of this parameter, so the likelihood is potentially not sharply peaked). Therefore, it is surprising that the worst agreement occurs for mass parameters and not spins. Moreover, not only does this relative behaviour seem unusual, the absolute values do as well: percentages of $25\%, 28\%$ (JSD) for M, m_1 are remarkably high.

  The respective values for the spins on the other hand are relatively small, for instance the median value for χ_{eff} is even smaller than the one for $\mathcal{M}$, which is generally regarded to be one of the or even the best measured parameter. The only real exception to these small criterion values is the JSD value for χ_{eff}, but this is very likely caused by the high JSD values for M, m_1, which are a part of the definition of χ_{eff}.

- Besides the remarkably high disagreement of some mass parameters, it is also remarkable that not all of them show this behaviour. After all, they are highly correlated and can even be computed from each other. While we can make sense of that for $\mathcal{M}$ as a very well-measured parameter and q (which can be measured reasonably well and does not show too big differences), the behaviour of m_2 is unusual. It has a much better overall agreement than other masses (measured by JSD), especially m_1 and M where we would in principle expect very similar values due to $M = m_1 + m_2$.

  There were several hypotheses to explain this, including the way component masses are calculated from other mass parameters (is possible e.g. from $\mathcal{M}, q$ or M, q; the idea was that different relative errors in the component masses could lead to different expected uncertainties) or a correlation between waveform agreement and the value of the component masses (if there was a correlation which would e.g. show worse agreement for higher masses, this could be an explanation because of the convention $m_1 \geq m_2$). However, none of these hypotheses could be verified to be true. Instead, the behaviour of M could turn out to be the reason. This is because one notable difference we could identify is how the component masses are correlated with other parameters (figure 4.4). While the values for q are really similar and just have opposing signs (which makes sense since m_2 is in the denominator and m_1 in the numerator), m_1 is less strongly correlated with $\mathcal{M}$ than m_2 (on average and for the majority of events) and the correlations with M differ significantly, too. On average it looks like m_2 is not correlated with M at all, but that turns out to be an effect of averaging. When looking at the individual events, we can see that there are correlations. However, they are much weaker than the ones between m_1, M and often also have a negative sign, which is not necessarily expected from $M = m_1 + m_2$. This negative sign and weaker

[7] The KLD is commonly used to assess differences between measured distributions and a reference distribution (e.g. used in appendix C of GWTC-1), so we use it here as well (although the JSD as a measure of difference is suited very well to do this as well).

		Order of parameters (Minimum to Maximum)								
Posterior	JSD	D_L (8%)	$\mathcal{M}$ (11%)	χ_p (14%)	θ_{jn} (14%)	m_2 (14%)	q (19%)	χ_{eff} (22%)	M (25%)	m_1 (28%)
	JSD 2	m_2 (0%)	$\mathcal{M}$ (3%)	D_L (3%)	χ_{eff} (6%)	χ_p (8%)	q (11%)	M (11%)	m_1 (11%)	θ_{jn} (36%)
	Mean difference	D_L (3%)	θ_{jn} (6%)	χ_{eff} (8%)	$\mathcal{M}$ (11%)	q (14%)	χ_p (14%)	m_2 (14%)	M (17%)	m_1 (19%)
	Median difference	χ_{eff} (8%)	D_L (8%)	θ_{jn} (8%)	$\mathcal{M}$ (11%)	χ_p (14%)	m_2 (14%)	q (17%)	M (17%)	m_1 (19%)
Prior	JSD	χ_p (0.076)	θ_{jn} (0.190)	χ_{eff} (0.212)	m_1 (0.402)	q (0.415)	M (0.466)	m_2 (0.584)	$\mathcal{M}$ (0.586)	D_L (0.768)
	KLD	χ_p (0.29)	θ_{jn} (0.76)	χ_{eff} (0.79)	m_1 (1.60)	q (1.61)	M (1.87)	m_2 (1.89)	$\mathcal{M}$ (2.58)	D_L (3.74)

Table 4.7.: Comparison of posterior agreement and prior difference. Values for the former are simply taken from 50% columns in table 4.5, while the latter is computed between prior and `Mixed` posterior (obtained by mixing samples from both waveform models). We provide the explicit JSD, KLD[7] values (averaged over all events) and not percentages with respect to some threshold because for the prior difference, we are interested specifically in *how* different they are and not statements which assess the agreement using a threshold. Values where the KLD is ∞ were excluded because they corrupt the average over all events.

correlation are actually prior-induced as figure 4.4 shows. Since it makes sense that for some events this correlation will be prior-dominated, while some do have positive values (more in accordance with what we expect), this explains why the value is zero on average. Independently from potential explanations for the behaviour of this correlation[8], since we noticed that M has much higher disagreement than $\mathcal{M}$ (again, we have no reason for that yet, but the systematic still exists), the different correlations could lead to a systematic transfer of inconsistencies into m_1, but not m_2, and could hence be a potential explanation for the discrepancy between m_1, m_2.

Despite having found a potential reason, we also have to take into account the possibility that this behaviour does not point to any systematic pattern, but is some kind of statistical outlier. For instance, it could in principle be the case that many criterion values for m_2 are very similar to the ones of m_1, but those for m_1 lie slightly above the threshold, so it is the way we analyse the data which causes the different behaviour. However, this is not true for the JSD values (where the systematic is observed). 22 events have a higher JSD for m_1 compared to the value of m_2 by on average 0.026, while the JSD of m_2 is higher than the value of m_1 for only 14 events and by an average of 0.013. For the mean (median) criterion, the respective numbers are 21 (19) where m_1 has a higher JSD than m_2 by an average of 0.18 (0.17) and 15 (17) events where m_2 has a higher value than m_1 by an average of 0.13 (0.09). As we can see, in one case the number of events where m_1 has worse agreement than m_2 is higher (with similar average differences) while in the other case the average difference is higher when m_1 has worse agreement (with a similar number of events where this happens). Therefore, mean and median criterion cannot confirm the hypothesis unambiguously nor can they disprove it (which would confirm the outlier hypothesis).

[8] The prior value indicates that there is a reason and that it is not accidental, priors are chosen carefully. However, we did not look into this in more detail (so we did not find a reason) since the observed values are sufficient here and we do not necessarily need an explanation to be satisfied. This might be interesting for follow-ups, though.

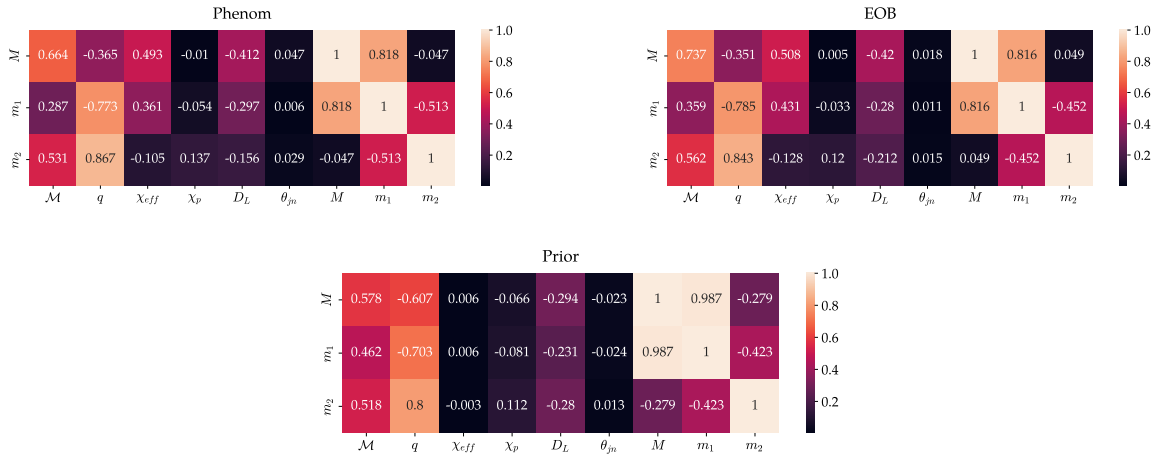


Figure 4.4.: Correlations of mass parameters M, m_1, m_2 with all other parameters, averaged over all events from GWTC-3. Since correlations depend on the distribution of the respective parameter, there are values for the posterior of each waveform model and also the prior.

- θ_{jn} often fails the JSD 2 criterion, which is interesting since the JSD does not mark nearly as much events. One reason is that its posteriors often has two peaks (inclinations which differ by exactly $\pi/2$ are hard to tell apart) because although the peak location and therefore rough shape often agrees well between the posteriors, the height of one/ both peaks or of the valley between the peaks is often slightly different. This may result in shape differences which are intended to be recognized by the JSD 2 criterion.

  This very typical shape with two peaks also explains why the prior-posterior difference is among the smallest, there simply is a natural high overlap between them since both prior and (in most cases) posterior are non-zero on almost the whole interval $[0, \pi]$, despite different shapes.

- That D_L differs a lot from its prior can be explained rather easily: with higher distances, the volume where potential sources are in grows, so it becomes more and more likely to have sources of detectable GWs the further we go away. However, the posteriors typically only span a much smaller fraction of the range the prior is defined on and also have a different shape (more equal to a normal distribution than to the prior). For these reasons, a high prior-posterior difference are not surprising.

- Mean and median criterion agree very well in their overall results. This means that the posteriors shapes do not generally lead to significantly different mean, standard deviation and median, credible interval combinations (at least the differences are not significant in a sense that they change the behaviour regarding our 50% threshold), which is interesting to know.

  Moreover, the results of these criteria should be meaningful despite the fact that many distributions have a limited range of values (notably, we have $\chi_{\text{eff}} \in [-1, 1]$, $\chi_p \in [0, 1]$, $\theta_{jn} \in [0, \pi]$). This is because we do not measure absolute distances of means or medians, but relative ones since they are always compared to the standard deviation or credible

interval. Therefore, the range of values for which the distribution is allowed to exist does not influence how well mean and median criterion can recognize differences.[9] What does play a role on the other hand is which fraction of this allowed interval the distributions span, i.e. their width compared to the interval length. In case of e.g. χ_p or θ_{jn} where the distributions span the whole range for the majority of events, it might be harder for the criteria to recognize differences in them than for other parameters with distributions that are very narrow compared to their possible range of values (particularly the masses, but also χ_{eff} belongs to this category). Therefore, criterion values in this case might underestimate the true differences.[10]

- The order of values for prior-posterior differences obtained from JSD and KLD broadly agree, which is reasonable. Nonetheless, it is good to see since it means that comparing the (potentially better suited) KLD values yields basically the same results as comparing the more widely used and better interpretable JSD values.

  The explicit values also meet our intuition, masses are among the best-informed parameters while spins are among the worst-informed. These results also help to asses the systematics found for posterior differences: that the prior difference for χ_p is only 0.076 (barely above our 50% threshold of 0.05 to recognize significant differences) means this parameter is highly prior-dominated. One can actually see this when looking at the posteriors, where many of them have a characteristic wide shape which is also how the prior looks like. Knowing that, values of 14% for JSD, mean difference, median difference do point to substantial inconsistencies between the two waveform models because it means that despite smaller impact of the data-, waveform-dependent terms, it is relatively often the case that significant waveform differences occur (particularly striking for mean, median values of χ_p, for wide distributions on a small interval we would expect really good overall agreement).

  For χ_{eff}, things look a bit different. While the prior-posterior difference is not as high as for the mass parameters, a JSD of 0.212 is still significant and much higher than our threshold (especially when taking non-linear behaviour of the JSD into account).

  It should be noted that it is possible that the true source parameters and posteriors are already very well described by the prior, which then causes a smaller prior-posterior difference. This might for instance be the case for χ_{eff} if the signal belongs to a non-spinning binary, which would produce a posterior centered around zero, just like the prior is. However, it does not apply to χ_p where the prior is very wide because well-informed posteriors are expected to be more narrow. In any case, our interpretations should remain valid because they are obtained from averaging over 36 events.

While the information about prior-posterior differences helped assessing the values for spin agreement, they do not really explain the high mass disagreement. Assuming consistent

[9] In subsection 5.1.2 it will even be shown mathematically that their criterion value does not change when changing the argument of the distributions according to $x \mapsto x/c$, so a change in the interval length $l \mapsto l/c$ as a difference $l = x_{\text{end}} - x_{\text{start}}$ does not affect them as well.

[10] As we will see in the next subsection where events from other catalogs are examined, fortunately, this problem is not so severe that no significant differences can be detected at all.

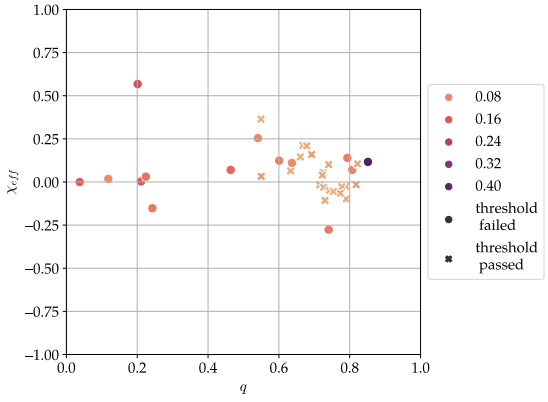
waveform models, similar posteriors should be inferred especially for well-informed parameters (which M certainly belongs to). This intuition is confirmed by the numbers for $\mathcal{M}, D_L$ (both are the best informed parameters and at the same time among the ones with best agreement), so it is urgent to explain these values for mass parameters. If no explanation can be found, they might point to significant waveform inconsistencies.

To assess these values properly, we can use some known relations from theoretical studies on waveform agreement, in particular mismatch studies. While statements from such studies are not guaranteed to transfer into the source properties of experimental data, the likelihood still compares data and template to assign a probability of this signal being present, so its morphology has an impact on the posterior and we can imagine that certain systematics indeed transfer. These known relations are explored for example in section IV of [9] and figure 9 of [26][11] and appear when looking at the events as points in planes spanned by certain parameters (slices in the parameter space). There it is found that the agreement between waveform models as measured by the mismatch/ unfaithfulness is best for equal mass, non-spinning binaries and diminishes with smaller q, growing $\chi_{\text{eff}}, \chi_p$, which motivates us to also look at this (figure 4.5).[12]

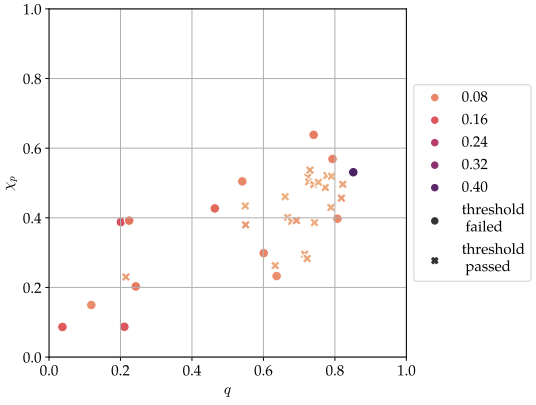
First of all, we can observe a clear tendency that the agreement becomes worse as q gets smaller. An indicator for that is not the absolute number of events with significant differences, but the relative number (i.e. the absolute one compared to the total number of events in the respective region). When dividing the planes in figure 4.5 for instance at the $q = 0.5$ axis, $7/8$ (87.5%) events on the left half with more extreme mass ratios exceed our chosen limit, while only $7/28$ (25%) on the right half with less extreme mass ratios do so. That means we can confirm at least a tendency that the behaviour of q observed in mismatch studies also appears in experimentally detected GW signals. Wwhile we only show plots for where the JSD is used as a criterion, the same tendency shows up for mean and median criterion. That means it is not caused by the JSD being too sensitive and thus further strengthens the significance of the statement. It is also not the case that this is only caused by M, m_1, m_2, excluding them even does not change the systematics significantly.

However, there is one caveat which slightly diminishes our confidence in this statement and is also the reason that we are not able to confirm the statements that higher spin values also produce worse waveform agreement. This reason is the relatively small number of events in GWTC-3. Although 36 GW signals is a lot when thinking about it from the GW detection and analysis perspective, it is not really a statistically significant amount. In addition to that, the events are not distributed evenly across each plane. The spins in particular are clustered around the median values of their priors (0 for χ_{eff}, 0.38 for χ_p) while events in extreme spin regions are yet to be detected, so it is not possible to infer meaningful statements on them. For q, this clustering is not as extreme as for the spins, so statements on this parameter remain significant.

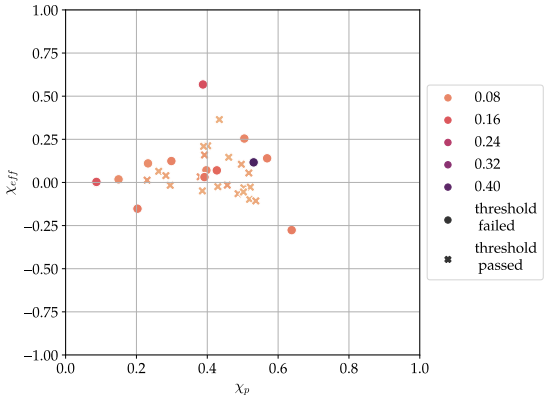
[11] A quick remark: like GWTC-3, we use the convention $q \leq 1$ whereas in these papers, $\tilde{q} = 1/q \Rightarrow \tilde{q} \geq 1$ is used.
[12] This behaviour is no surprise since the agreement of waveforms is partially ensured by calibrating them to NR simulations and there are more of those for non-extreme parameter combinations.



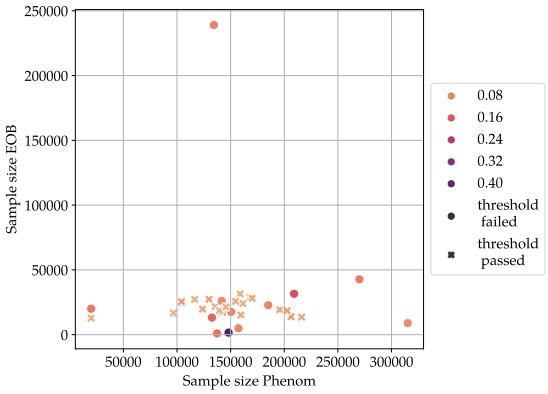
(a) Correlation plot in q-χ_{eff}-plane for all events from GWTC-3



(b) Correlation plot in q-χ_p-plane for all events from GWTC-3



(c) Correlation plot in χ_p-χ_{eff}-plane for all events from GWTC-3

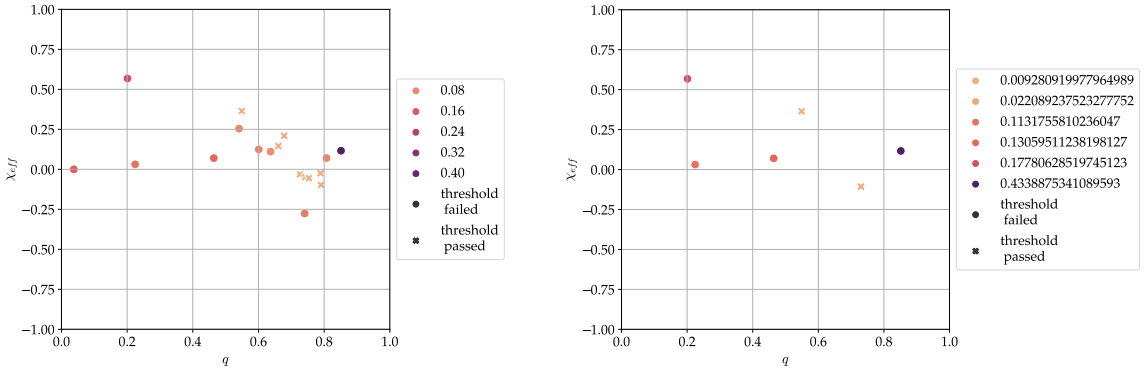


(d) Correlation plot in plane spanned by sample sizes for all events from GWTC-3

Figure 4.5.: Visualization of all events as points in different planes. The coordinates are determined by (a) - (c) the medians of the respective parameters or (d) the values themselves and there is a coloring based on the maximum JSD value between the marginalized `IMRPhenomXPHM` and `SEOBNRv4PHM` posteriors for the set of parameters used throughout this section (although results do not change significantly when excluding M, m_1, m_2, in particular not if an event fails/ passes). To make it more apparent which value surpasses the 50% threshold, two different symbols are used (see the respective plot legend).

Besides plots which allow the comparison with mismatch studies, there are many other potential correlations we can think of. One of them regards a notion which played a role throughout this work for a few times already, the sample size. All statements on the agreement of posteriors in this work are based on histograms computed from samples of the posteriors, so our results are naturally affected by sampling deviations. The relevant question is how severe these deviations are. We can see that by looking at the sample sizes of each waveform model and event (figure 4.5 (d)). Again, we can observe a tendency of some correlation being present. The (relative) amount of events exceeding the threshold in the small cluster around the point $(150000, 25000)$ and outside of it differ a lot (roughly $7/8$ (87.5%) outside vs. $5/28$ (17.8%) inside or $9/14$ (64.3%) outside vs. $3/22$ (13.6%) inside, depending on which events are included in the cluster). That means if the sample size of one of the posteriors is much smaller or larger than usual, it is very likely that the agreement will suffer from that. Explaining this for small sample sizes is relatively straightforward, statements here are very vulnerable to statistical errors, which might lead to bad agreement (although any statement would not be very certain, also good agreement; this is especially the case for events like GW200322_091133 with a sample size of 981 for `SEOBNRv4PHM`). Very high sample sizes on the other hand might point to problems with sampling the posterior (e.g. due to a very complicated shape), which could also cause disagreement. However, we did not investigate the latter in more detail and hence have no proof for this statement. For us, it is sufficient to know that disagreement caused by sampling effects should be subdominant since the thresholds are robust for the sample sizes of the majority of events.

Until now, the maximum disagreement of all parameters in the q-χ_{eff}-plane was used, but there might also be patterns in the (dis-)agreement of a single parameter (figure 4.7). This brings us back to the initial motivation for looking at such a representation in the first place, explaining what causes the excess disagreement for mass parameters, in particular M, m_1 (where we already identified that the behaviour of m_1 are probably caused by the one of M). It turns out that there is a very clear tendency specifically for those parameters to disagree in extreme-mass-ratio regions. To be more explicit, for $6/8$ (75%) events on the left half of the q-χ_{eff}-plane the differences in M and m_1 exceed the JSD's 50% threshold, but it is only $1/8$ (12.5%) for m_2, $2/8$ (25%) for $\mathcal{M}, q$ and $3/8$ (37.5%) for $\chi_{\text{eff}}, \chi_p$. While this argument relies on the use of thresholds, there seems to be an underlying correlation that is independent of them, which supports this claim (figure 4.8). Although it is not perfectly linear, there definitely is a relation between the agreement of M, m_1 and the value of q (which is not present or at least not very clear for $\mathcal{M}, q, \chi_{\text{eff}}, m_2$). Just like we when discussing the discrepancy between m_1, m_2, we also have to take into account the possibility that the values for M are a statistical outlier, which is e.g. caused by the JSD being too sensitive and marking non-significant differences as significant (which would again be related to the use of a threshold-based visualization). However, while the statement would be more solid if the behaviour also showed up in other criteria (which is not the case; extracting a correlation from them is not possible, other than for the overall tendency in extreme-mass-ratio regions), the JSD is the only criterion we use that is sensitive to all differences in the posteriors. Therefore, its values and behaviour alone should be sufficient to point to a real systematic, especially because there is also a general correlation and not just one related to thresholds (although we still cannot fully disprove the outlier hypothesis).



(a) Correlation plot in q-χ_{eff}-plane for events from good, bad data (equivalent to threshold passed, failed here)

(b) Correlation plot in q-χ_{eff}-plane for all events from GWTC-3 with multimodality

Figure 4.6.: Visualization of specific events as points in q-χ_{eff}-plane (lists are given in respective caption). The coordinates are determined by the medians of the respective parameters and there is a coloring based on the maximum JSD value between the marginalized `IMRPhenomXPHM` and `SEOBNRv4PHM` posteriors for the set of parameters used throughout this section (although results do not change significantly when excluding M, m_1, m_2, in particular not if an event fails/ passes). To make it more apparent which value surpasses the 50% threshold, two different symbols are used (see the respective plot legend).

Having identified a potential correlation, it is natural to look what it is caused by. A potential explanation for the masses being so affected by the the values of χ_{eff} and particularly q is that they can be measured relatively accurately and due to this accuracy, mismatch differences known for these extreme values transfer into the source properties as inconsistencies. It would then be clear why such a systematic does not appear e.g. in the spin parameters, they simply cannot be measured accurately enough to pick up the waveform differences. While this explanation would not contradict what we currently know about waveforms and parameters, it is still not perfectly certain for several reasons. First of all, the inconsistencies do not appear in $\mathcal{M}, q$. This could be due to the fact that they can be measured so accurately that waveform inconsistencies do not affect them, but we have no proof for that and it seems unlikely that q can be measured that much more accurately than M (for $\mathcal{M}$, it would be more reasonable). On the other hand, it also seems unlikely that such differences affect M and not q. While q might have a smaller impact on signals, it still has an important one as it determines the phase evolution at only one order lower than $\mathcal{M}$ (for information on mass parameters see e.g. section V.A.1. of [1]). Secondly, it does not appear in other criteria than the JSD, which was already discussed. The final conclusion for this topic is that while we could identify potential reasons for the systematics surrounding m_1, m_2, this was not possible for M, at least not very confidently. To fully assess these statements, many effects like mismatches and priors have to be taken into account and doing that would involve many methods and thoughts which are beyond the scope of this work, so this is left open for future research.

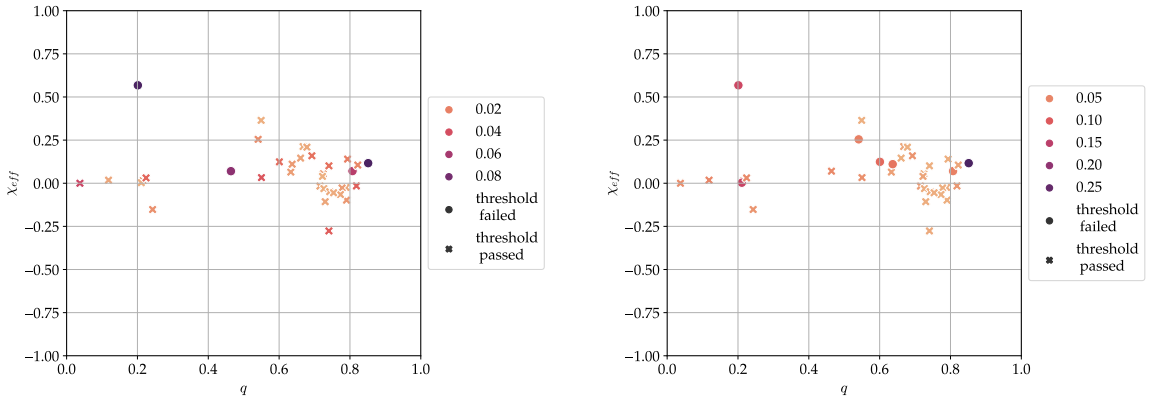
Two further interesting relations we are also able to examine are (i) that the chirp mass $\mathcal{M}$ can be measured more precisely in lower-mass systems and (ii) that the total mass M has a higher influence for higher-mass systems, i.e. it can be measured better (both of these claims are made in section V.A. of [1]). It would not be surprising if these systematics also appear in the agreement of waveform models for these parameters. However, neither the JSD nor the median criterion reproduce this behaviour (figure 4.9). Instead, there is no clear correlation (in fact, this is true for the whole set of parameters we deal with and the SNR).

At the end of this subsection, we provide some concluding thoughts and arguments on the general treatment of waveform disagreement (inconsistencies). As we can see in almost every plot shown in this subsection and also when looking at the events from bad data (shown in the q-χ_{eff}-plane in figure 4.6 (a)), not all events with bad agreement lie in extreme regions of the parameter space. However, this is not a problem in a sense that it points to inconsistencies in statements inferred here because there are several other reasons for waveform disagreement, which are completely independent from the ones we dealt with. These reasons are for example sampling deviations from very small sample sizes, multimodality or simply noise. As already discussed, sampling effects should be subdominant for GWTC-3 (but that does not mean they do not occur at all). Multimodality also does not occur too often (we refer to the six events which have this effect according to GWTC-3) and also with no regular pattern with respect to masses or spins (figure 4.6 (b)). However, despite this small absolute number, for some parameters they make up a significant fraction of all events, where waveform differences for this parameter exceed the 50% threshold. Notable are $3/4$ for the chirp mass, $2/3$ for the luminosity distance, $3/9$ for the total mass, $4/10$ for mass 1 (on the other hand, only $2/8$ for the effective spin, $1/5$ for the precession spin). The last reason is probably the most intuitive one, there is noise present in the data of GW detectors. While it is of course the case that the noise n in $d = n + h$ disturbs our analyses, it is non-Gaussian and non-stationary features which are really problematic because the likelihood and quantities defined based on it do not take them into account. In fact, [27] finds seven events that overlap with glitches, which have to be subtracted before the actual analyses, and four of them are in bad data (GW191109_010717, GW191127_050227, GW191219_163120, GW2000129_065458; GW200105_162426 is also mentioned and indeed shows disagreement, but it was excluded from bad data). Moreover, it is even expected that some events are purely noise because false alarm rates or similar quantities assigned to assess astrophysical origin are only probabilities, some false assessments are inevitable. Therefore, it is clear that noise introduces some uncertainty and is a potential reason for waveform disagreement.

Another systematic supporting this fact is that all events from good data lie in regions where we would expect them, for example the the non-extreme regions of the q-χ_{eff}-plane.

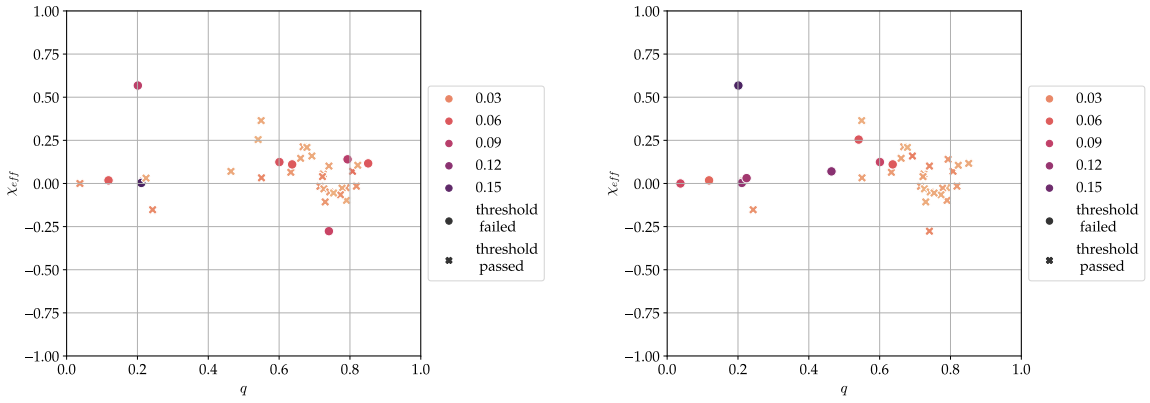
4.2.6. Comparison with previous Catalogs

After the detailed analysis of GWTC-3 data in the previous subsections, we are confident enough to assess the quality of this data. However, such statements on "good" or "bad"



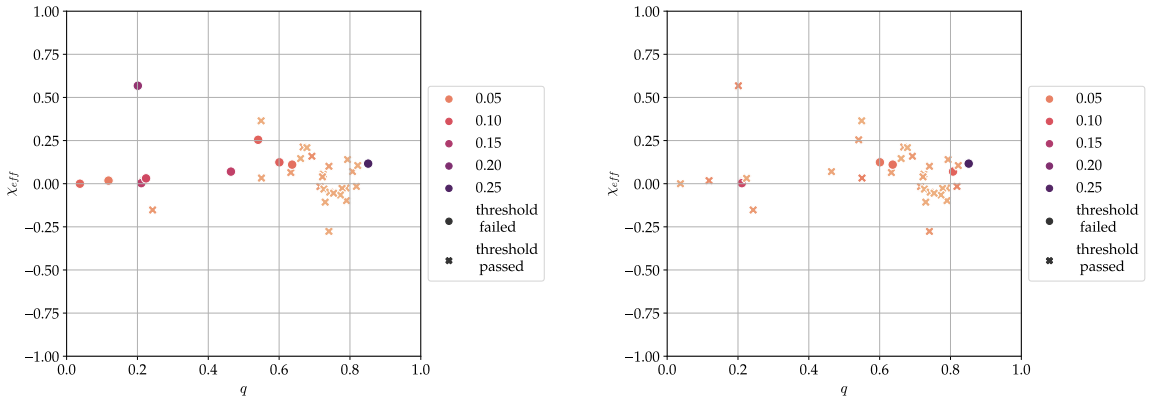
(a) Chirp mass $\mathcal{M}$

(b) Mass ratio q



(c) Effective spin χ_{eff}

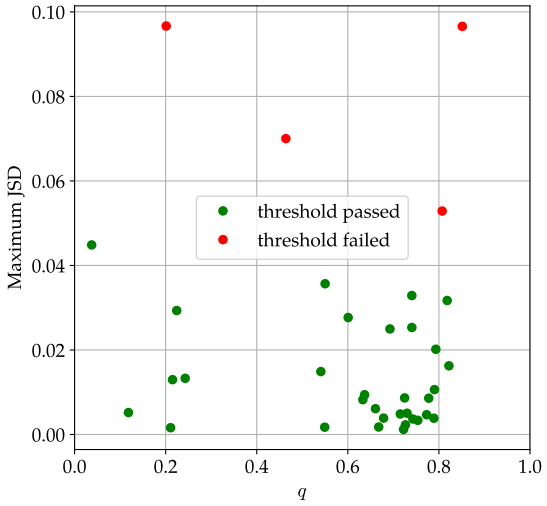
(d) Total mass M



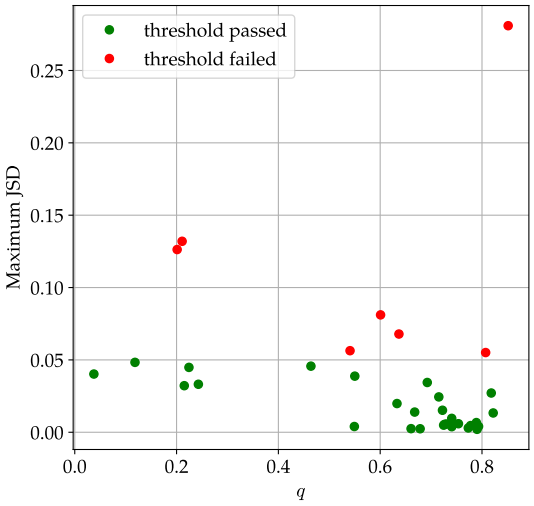
(e) Component mass 1 m_1

(f) Component mass 2 m_2

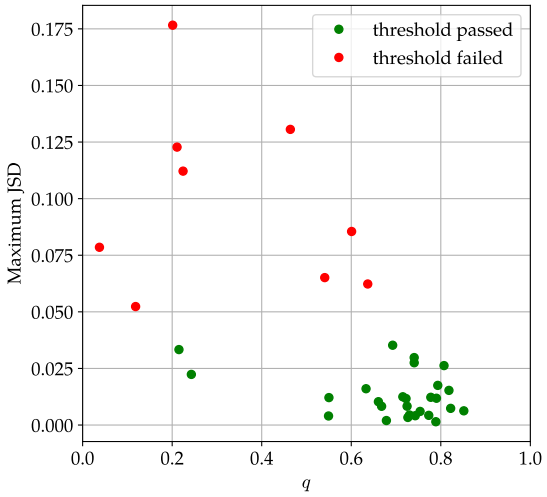
Figure 4.7.: Visualization of all events from GWTC-3 as points in q-χ_{eff}-plane. The coordinates are determined by the medians of the respective parameters and there is a coloring based on the JSD value between the marginalized `IMRPhenomXPHM` and `SEOBNRv4PHM` posteriors for the parameters given in the respective caption. To make it more apparent which value surpasses the 50% threshold, two different symbols are used (see the respective plot legend).



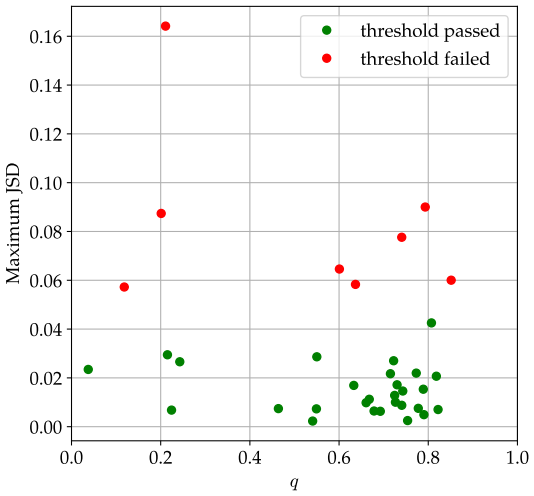
(a) Chirp mass $\mathcal{M}$



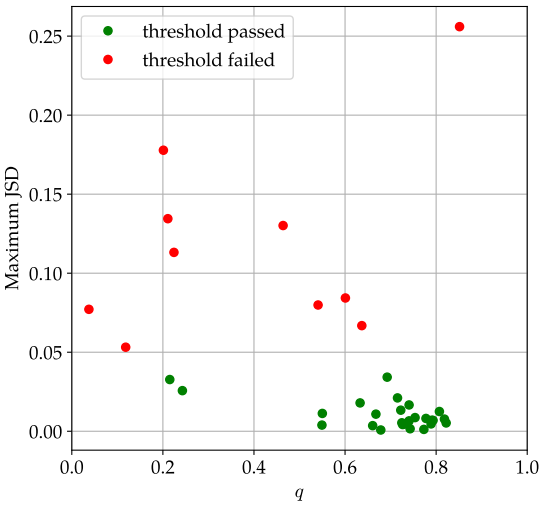
(b) Mass ratio q



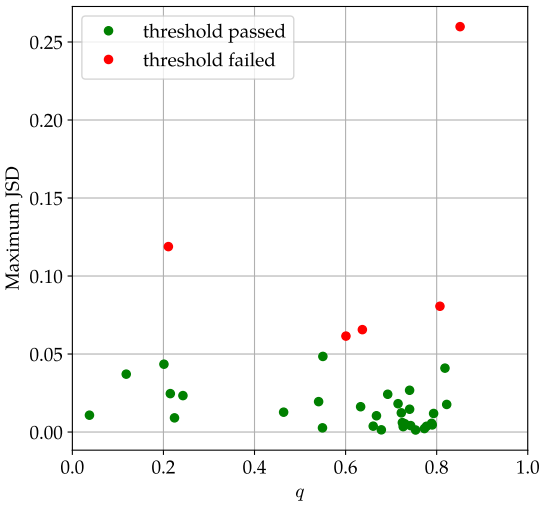
(c) Total mass M



(d) Effective spin χ_{eff}

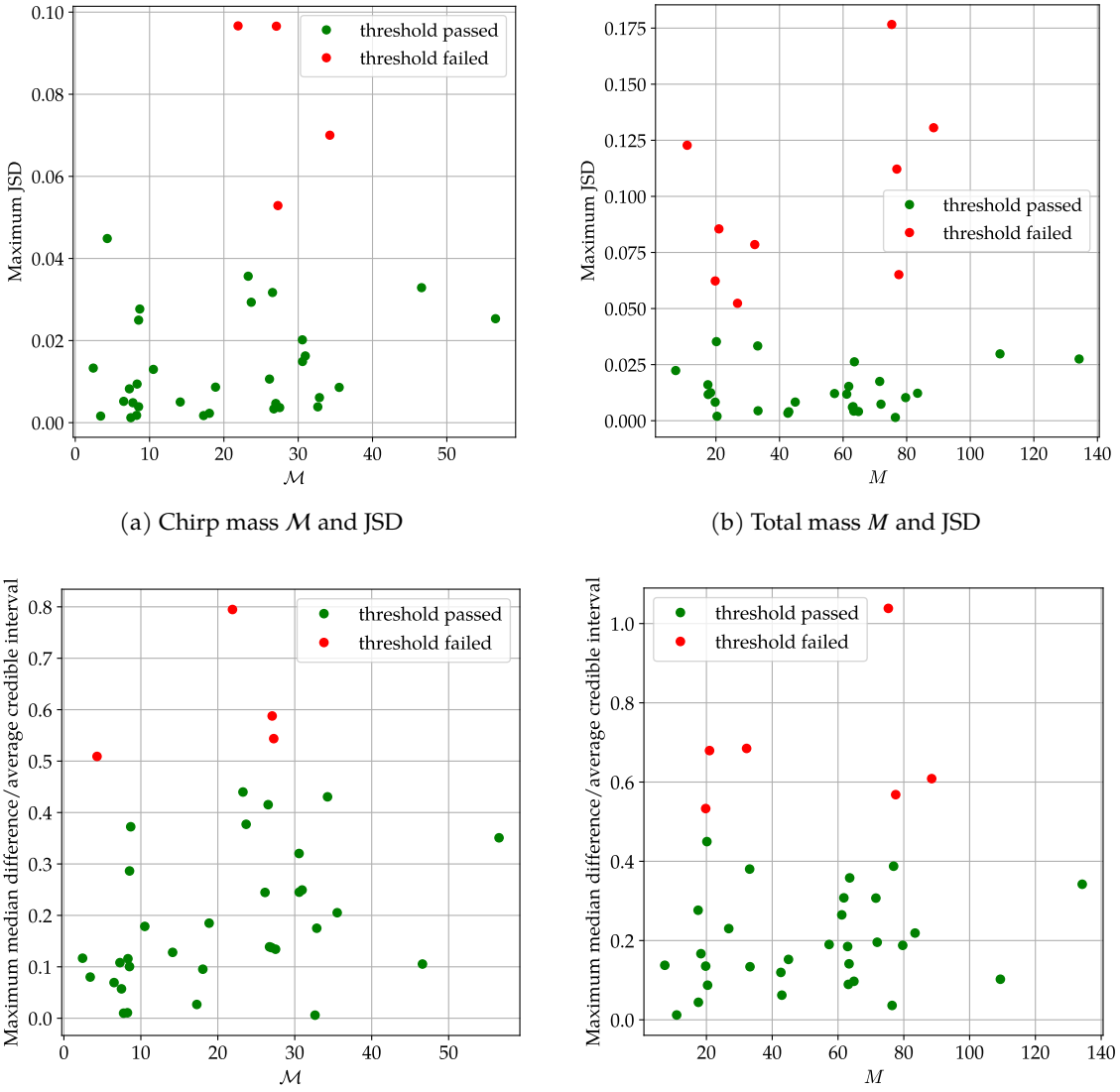


(e) Component mass 1 m_1



(f) Component mass 2 m_2

Figure 4.8.: Visualization of of all events from GWTC-3 as points on the q-axis. The y-values are determined by the JSD value between the marginalized `IMRPhenomXPHM` and `SEOBNRv4PHM` posteriors for the parameters specified in the respective caption (which is the maximum criterion value in this case because only one parameter is used). To make it more apparent which value surpasses the 50% threshold, two different colors are used (see the respective plot legend).



(a) Chirp mass $\mathcal{M}$ and JSD

(b) Total mass M and JSD

(c) Chirp mass $\mathcal{M}$ and median criterion

(d) Total mass M and median criterion

Figure 4.9.: Visualization of all events from GWTC-3 as points on different axes given in the respective caption. The y-values are determined by the median of this parameter and the criterion value (specified on y-axis) between the marginalized `IMRPhenomXPHM` and `SEOBNRv4PHM` posteriors for this parameter (which is the maximum criterion value in this case because only one parameter is used). To make it more apparent which value surpasses the 50% threshold, two different colors are used (see the respective plot legend).

overall data quality will always remain subjective to some degree. A more objective statement is to repeat the analysis from the previous subsections for the other GWTCs. That would enable us to assess how the quality in GWTC-3 compares to them, so we do it in this subsection (the examination will not be as detailed as for GWTC-3 since the goal is to compare the catalogs, not to extract statements from them). To make results for the different catalogs comparable, it is important that the same waveform models have been used for parameter estimation (otherwise, differences in the waveform models might corrupt the results). Hence, we are lucky that there was a very recent data release [28], where posterior samples for all significant (i.e. $p_{\text{astro}} > 0.5$) events from GWTC-1 and GWTC-2.1 were (re-)released.[13] Not all of them contained samples from `IMRPhenomXPHM` *and* `SEOBNRv4PHM` posteriors (more precisely $9/10$ for GWTC-1, $30/44$ for GWTC-2.1 did), but our methods are applicable to the majority of 72.2% of them, so any statements comparing the different catalogs is statistically meaningful. There is no point in analyzing the other events here because only one waveform is used, so it is impossible to compare agreement.

We start by applying our criteria to the events from GWTC-1 and GWTC-2.1 (where it is possible, i.e. where results for `IMRPhenomXPHM` and `SEOBNRv4PHM` are available), to look at the agreement of each parameter (tables 4.8, 4.9). Immediately, we notice some differences:

- The agreement varies significantly between the catalogs. The first one has remarkably good agreement (compared to GWTC-3, but also in general when only looking at the absolute values), while the second one has much worse agreement than the others (which leads to the combined data from GWTC-1 and GWTC-2.1 having a level of agreement comparable to the one of GWTC-3). That pattern persists throughout each criterion and hence we are confident in saying that it is no statistical outlier caused e.g. by the JSD being very sensitive.
- Some of the unexpected features of the GWTC-3 data do not appear for the other catalogs: the spins are indeed among the parameters with the worst agreement and in general have higher disagreement than in GWTC-3, even for GWTC-1 which otherwise has better agreement (plus the mean, median criterion mark higher fraction of events for them as well, so the relatively small number for GWTC-3 points to actual good agreement and does not exclusively come from insensitivity of the criteria). Also, the discrepancy between the agreement of m_2 and the other mass parameters is not noticeable and, when taking into account the overall higher disagreement, there is no substantial excess disagreement for the mass parameters in GWTC-2.1.

  This is all despite very similar average correlation values (overall and for individual events) for the posteriors and also the prior. However, this does not disprove the reasoning we built up to explain these values in the previous subsections. As we will

[13] There is no consistent use of the GWTC names, which is why we clarify the use here. If we refer to GWTC-3, we only refer to events which are new in GWTC-3 (i.e. detected in observing run O3b) and *not* to those already published in GWTC-1 (summary of O1, O2) and -2.1 (second version of summary of O3a).

[14] Not every event with posteriors for both waveforms has prior samples, but it is again a significant proportion: $8/9$ (88.89%) events in GWTC-1 and $26/30$ (86.67%) in GWTC-2.1. We cannot just take samples from another event because although the overall shape is the same all the time, scaling differs for each event (which is also the case in each catalog, so the same priors were used).

Eventlist	Parameter	JSD		JSD 2		Mean difference		Median difference	
		50%	20%	50%	20%	50%	20%	50%	20%
GWTC-1	$\mathcal{M}$	0/ 9 (0%)	4/ 9 (44%)	0/ 9 (0%)	0/ 9 (0%)	0/ 9 (0%)	4/ 9 (44%)	0/ 9 (0%)	4/ 9 (44%)
	q	1/ 9 (11%)	4/ 9 (44%)	0/ 9 (0%)	1/ 9 (11%)	1/ 9 (11%)	2/ 9 (22%)	0/ 9 (0%)	3/ 9 (33%)
	$\chi_{\rm eff}$	2/ 9 (22%)	5/ 9 (56%)	0/ 9 (0%)	1/ 9 (11%)	2/ 9 (22%)	3/ 9 (33%)	1/ 9 (11%)	4/ 9 (44%)
	χ_p	2/ 9 (22%)	7/ 9 (78%)	0/ 9 (0%)	6/ 9 (67%)	1/ 9 (11%)	5/ 9 (56%)	1/ 9 (11%)	5/ 9 (56%)
	D_L	0/ 9 (0%)	7/ 9 (78%)	0/ 9 (0%)	1/ 9 (11%)	0/ 9 (0%)	7/ 9 (78%)	0/ 9 (0%)	7/ 9 (78%)
	θ_{jn}	0/ 9 (0%)	3/ 9 (33%)	1/ 9 (11%)	6/ 9 (67%)	0/ 9 (0%)	0/ 9 (0%)	0/ 9 (0%)	2/ 9 (22%)
	M	0/ 9 (0%)	5/ 9 (56%)	0/ 9 (0%)	1/ 9 (11%)	0/ 9 (0%)	5/ 9 (56%)	0/ 9 (0%)	5/ 9 (56%)
	m_1	0/ 9 (0%)	4/ 9 (44%)	0/ 9 (0%)	1/ 9 (11%)	0/ 9 (0%)	4/ 9 (44%)	0/ 9 (0%)	3/ 9 (33%)
	m_2	1/ 9 (11%)	3/ 9 (33%)	0/ 9 (0%)	1/ 9 (11%)	1/ 9 (11%)	2/ 9 (22%)	1/ 9 (11%)	1/ 9 (11%)
GWTC-2.1	$\mathcal{M}$	6/ 30 (20%)	18/ 30 (60%)	2/ 30 (7%)	3/ 30 (10%)	4/ 30 (13%)	18/ 30 (60%)	4/ 30 (13%)	17/ 30 (57%)
	q	8/ 30 (27%)	20/ 30 (67%)	1/ 30 (3%)	14/ 30 (47%)	6/ 30 (20%)	16/ 30 (53%)	6/ 30 (20%)	17/ 30 (57%)
	$\chi_{\rm eff}$	11/ 30 (37%)	21/ 30 (70%)	2/ 30 (7%)	9/ 30 (30%)	8/ 30 (27%)	17/ 30 (57%)	8/ 30 (27%)	16/ 30 (53%)
	χ_p	9/ 30 (30%)	19/ 30 (63%)	1/ 30 (3%)	13/ 30 (43%)	8/ 30 (27%)	19/ 30 (63%)	7/ 30 (23%)	19/ 30 (63%)
	D_L	5/ 30 (17%)	16/ 30 (53%)	1/ 30 (3%)	5/ 30 (17%)	4/ 30 (13%)	13/ 30 (43%)	3/ 30 (10%)	15/ 30 (50%)
	θ_{jn}	4/ 30 (13%)	17/ 30 (57%)	12/ 30 (40%)	25/ 30 (83%)	1/ 30 (3%)	7/ 30 (23%)	3/ 30 (10%)	13/ 30 (43%)
	M	7/ 30 (23%)	19/ 30 (63%)	5/ 30 (17%)	9/ 30 (30%)	6/ 30 (20%)	19/ 30 (63%)	8/ 30 (27%)	18/ 30 (60%)
	m_1	8/ 30 (27%)	20/ 30 (67%)	3/ 30 (10%)	10/ 30 (33%)	7/ 30 (23%)	18/ 30 (60%)	7/ 30 (23%)	18/ 30 (60%)
	m_2	9/ 30 (30%)	20/ 30 (67%)	0/ 30 (0%)	6/ 30 (20%)	7/ 30 (23%)	19/ 30 (63%)	7/ 30 (23%)	18/ 30 (60%)
GWTC-1 & GWTC-2.1	$\mathcal{M}$	6/ 39 (15%)	22/ 39 (56%)	2/ 39 (5%)	3/ 39 (8%)	4/ 39 (10%)	22/ 39 (56%)	4/ 39 (10%)	21/ 39 (54%)
	q	9/ 39 (23%)	24/ 39 (62%)	1/ 39 (3%)	15/ 39 (38%)	7/ 39 (18%)	18/ 39 (46%)	6/ 39 (15%)	20/ 39 (51%)
	$\chi_{\rm eff}$	13/ 39 (33%)	26/ 39 (67%)	2/ 39 (5%)	10/ 39 (26%)	10/ 39 (26%)	20/ 39 (51%)	9/ 39 (23%)	20/ 39 (51%)
	χ_p	11/ 39 (28%)	26/ 39 (67%)	1/ 39 (3%)	19/ 39 (49%)	9/ 39 (23%)	24/ 39 (62%)	8/ 39 (21%)	24/ 39 (62%)
	D_L	5/ 39 (13%)	23/ 39 (59%)	1/ 39 (3%)	6/ 39 (15%)	4/ 39 (10%)	20/ 39 (51%)	3/ 39 (8%)	22/ 39 (56%)
	θ_{jn}	4/ 39 (10%)	20/ 39 (51%)	13/ 39 (33%)	31/ 39 (79%)	1/ 39 (3%)	7/ 39 (18%)	3/ 39 (8%)	15/ 39 (38%)
	M	7/ 39 (18%)	24/ 39 (62%)	5/ 39 (13%)	10/ 39 (26%)	6/ 39 (15%)	24/ 39 (62%)	8/ 39 (21%)	23/ 39 (59%)
	m_1	8/ 39 (21%)	24/ 39 (62%)	3/ 39 (8%)	11/ 39 (28%)	7/ 39 (18%)	22/ 39 (56%)	7/ 39 (18%)	21/ 39 (54%)
	m_2	10/ 39 (26%)	23/ 39 (59%)	0/ 39 (0%)	7/ 39 (18%)	8/ 39 (21%)	21/ 39 (54%)	8/ 39 (21%)	19/ 39 (49%)
GWTC-3	$\mathcal{M}$	4/ 36 (11%)	18/ 36 (50%)	1/ 36 (3%)	6/ 36 (17%)	4/ 36 (11%)	14/ 36 (39%)	4/ 36 (11%)	16/ 36 (44%)
	q	7/ 36 (19%)	21/ 36 (58%)	4/ 36 (11%)	15/ 36 (42%)	5/ 36 (14%)	13/ 36 (36%)	6/ 36 (17%)	16/ 36 (44%)
	$\chi_{\rm eff}$	8/ 36 (22%)	24/ 36 (67%)	2/ 36 (6%)	10/ 36 (28%)	3/ 36 (8%)	16/ 36 (44%)	3/ 36 (8%)	15/ 36 (42%)
	χ_p	5/ 36 (14%)	19/ 36 (53%)	3/ 36 (8%)	12/ 36 (33%)	5/ 36 (14%)	17/ 36 (47%)	5/ 36 (14%)	15/ 36 (42%)
	D_L	3/ 36 (8%)	15/ 36 (42%)	1/ 36 (3%)	7/ 36 (19%)	1/ 36 (3%)	11/ 36 (31%)	3/ 36 (8%)	11/ 36 (31%)
	θ_{jn}	5/ 36 (14%)	17/ 36 (47%)	13/ 36 (36%)	26/ 36 (72%)	2/ 36 (6%)	6/ 36 (17%)	3/ 36 (8%)	18/ 36 (50%)
	M	9/ 36 (25%)	24/ 36 (67%)	4/ 36 (11%)	18/ 36 (50%)	6/ 36 (17%)	17/ 36 (47%)	6/ 36 (17%)	17/ 36 (47%)
	m_1	10/ 36 (28%)	20/ 36 (56%)	4/ 36 (11%)	19/ 36 (53%)	7/ 36 (19%)	15/ 36 (42%)	7/ 36 (19%)	16/ 36 (44%)
	m_2	5/ 36 (14%)	23/ 36 (64%)	0/ 36 (0%)	11/ 36 (31%)	5/ 36 (14%)	14/ 36 (39%)	5/ 36 (14%)	16/ 36 (44%)

Table 4.8.: Comparison of event statistics, based on the same principle as table 4.5, but for different catalogs instead of different lists of events from GWTC-3.

			Order of parameters (Minimum to Maximum)								
GWTC-1	Posterior	JSD	$\mathcal{M}$ (0%)	D_L (0%)	θ_{jn} (0%)	M (0%)	m_1 (0%)	q (11%)	m_2 (11%)	$\chi_{\rm eff}$ (22%)	χ_p (22%)
		JSD 2	$\mathcal{M}$ (0%)	q (0%)	$\chi_{\rm eff}$ (0%)	χ_p (0%)	D_L (0%)	M (0%)	m_1 (0%)	m_2 (0%)	θ_{jn} (11%)
		Mean difference	$\mathcal{M}$ (0%)	D_L (0%)	θ_{jn} (0%)	M (0%)	m_1 (0%)	q (11%)	χ_p (11%)	m_2 (11%)	$\chi_{\rm eff}$ (22%)
		Median difference	$\mathcal{M}$ (0%)	q (0%)	D_L (0%)	θ_{jn} (0%)	M (0%)	m_1 (0%)	$\chi_{\rm eff}$ (11%)	χ_p (11%)	m_2 (11%)
	Prior	JSD	χ_p (0.032)	$\chi_{\rm eff}$ (0.215)	θ_{jn} (0.288)	m_1 (0.463)	q (0.512)	M (0.544)	$\mathcal{M}$ (0.716)	m_2 (0.766)	D_L (0.934)
		KLD	χ_p (0.12)	$\chi_{\rm eff}$ (0.77)	θ_{jn} (1.18)	m_1 (1.74)	q (2.03)	M (2.22)	m_2 (2.41)	D_L (3.42)	$\mathcal{M}$ (3.95)
GWTC-2.1	Posterior	JSD	θ_{jn} (13%)	D_L (17%)	$\mathcal{M}$ (20%)	M (23%)	q (27%)	m_1 (27%)	χ_p (30%)	m_2 (30%)	$\chi_{\rm eff}$ (37%)
		JSD 2	m_2 (0%)	q (3%)	χ_p (3%)	D_L (3%)	$\mathcal{M}$ (7%)	$\chi_{\rm eff}$ (7%)	m_1 (10%)	M (17%)	θ_{jn} (40%)
		Mean difference	θ_{jn} (3%)	$\mathcal{M}$ (13%)	D_L (13%)	q (20%)	M (20%)	m_1 (23%)	m_2 (23%)	$\chi_{\rm eff}$ (27%)	χ_p (27%)
		Median difference	D_L (10%)	θ_{jn} (10%)	$\mathcal{M}$ (13%)	q (20%)	χ_p (23%)	m_1 (23%)	m_2 (23%)	$\chi_{\rm eff}$ (27%)	M (27%)
	Prior	JSD	χ_p (0.050)	θ_{jn} (0.185)	$\chi_{\rm eff}$ (0.270)	m_1 (0.435)	q (0.445)	M (0.485)	$\mathcal{M}$ (0.525)	m_2 (0.583)	D_L (0.768)
		KLD	χ_p (0.18)	θ_{jn} (0.74)	$\chi_{\rm eff}$ (0.97)	m_1 (1.65)	q (1.69)	M (1.82)	m_2 (2.10)	$\mathcal{M}$ (2.31)	D_L (3.30)
GWTC-3	Posterior	JSD	D_L (8%)	$\mathcal{M}$ (11%)	χ_p (14%)	θ_{jn} (14%)	m_2 (14%)	q (19%)	$\chi_{\rm eff}$ (22%)	M (25%)	m_1 (28%)
		JSD 2	m_2 (0%)	$\mathcal{M}$ (3%)	D_L (3%)	$\chi_{\rm eff}$ (6%)	χ_p (8%)	q (11%)	M (11%)	m_1 (11%)	θ_{jn} (36%)
		Mean difference	D_L (3%)	θ_{jn} (6%)	$\chi_{\rm eff}$ (8%)	$\mathcal{M}$ (11%)	q (14%)	χ_p (14%)	m_2 (14%)	M (17%)	m_1 (19%)
		Median difference	$\chi_{\rm eff}$ (8%)	D_L (8%)	θ_{jn} (8%)	$\mathcal{M}$ (11%)	χ_p (14%)	m_2 (14%)	q (17%)	M (17%)	m_1 (19%)
	Prior	JSD	χ_p (0.076)	θ_{jn} (0.190)	$\chi_{\rm eff}$ (0.212)	m_1 (0.402)	q (0.415)	M (0.466)	m_2 (0.584)	$\mathcal{M}$ (0.586)	D_L (0.768)
		KLD	χ_p (0.29)	θ_{jn} (0.76)	$\chi_{\rm eff}$ (0.79)	m_1 (1.60)	q (1.61)	M (1.87)	m_2 (1.89)	$\mathcal{M}$ (2.58)	D_L (3.74)

Table 4.9.: Comparison of posterior agreement and prior difference for events from all catalogs (this table is to table 4.8 what table 4.7 is to table 4.5).[14]

see later in this subsection, the previous catalogs have almost no events in extreme-mass-ratio regions, which means they are not suited to assess whether or not the hypothesis is true. To confirm the observations that M tends to disagree for smaller q and that this transfers into m_1 but not m_2, more data will be very helpful and in case of a confirmation, it would make sense to search for explanations of this.

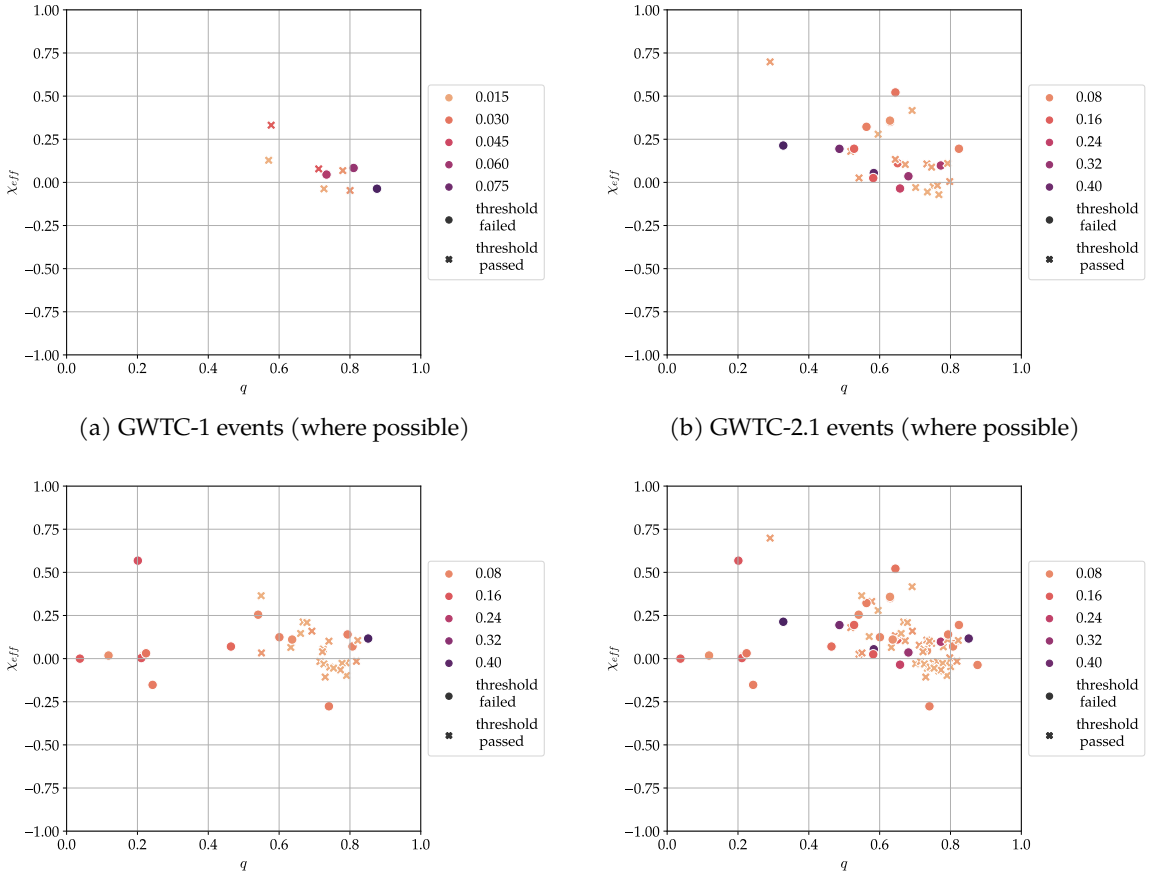
- In general, the values for the prior-posterior difference undergo only minor changes, especially when comparing GWTC-2.1 and GWTC-3. GWTC-1 does have have higher values for almost all parameters, but this increase is not so big that any interpretation would change from good informed to bad informed or vice versa. Moreover, KLD and JSD again broadly agree in their predictions.

  The only smaller exception from the systematics described here is the value for χ_p, where the value grows from 0.032 to 0.05 and finally 0.076 for the JSD (and thus from below the 50% threshold to slightly above). This is partially caused by few events (one in GWTC-2.1 and two in GWTC-3) with very high values up to 0.5, which can be regarded as outliers in some sense because these values are not representative for all events. Excluding them still shows a slight growth, but even less significant.

Differences between catalogs have to come from the data $d = n + h$ because the same waveforms, priors etc. were used for the analyses, so either the signals or the noise have to be different. We will not focus on the noise aspect in this work, but note that significant changes in the detectors occur in the time between different observing runs. They are therefore expected from GWTC-1 to GWTC-2.1, but not necessarily from GWTC-2.1 to GWTC-3, which summarize events from different periods of the same observing run O3 (between these catalogs, there should not be a significant impact on how much and how accurately information about the signals could be extracted from the data).

To see how the signals might affect the overall agreement, we can look at different representations of all events again, for instance in the q-χ_{eff}-plane (figure 4.10) or as a function of parameters like the chirp mass $\mathcal{M}$ (figure 4.12). Both approaches reveal the same systematics: the agreement for GWTC-1 is best, but it also covers the smallest region in the parameter space compared to GWTC-2.1 and GWTC-3. This is also one major reason for the remarkable agreement, all events have very similar properties in the equal-mass, non-spinning region, where better agreement is expected. That changes slightly for GWTC-2.1 and notably for GWTC-3, where a much wider region of the parameter space is covered by the data (to emphasize this, all plots in figure 4.10 show the whole range of possible values). Similarly, GWTC-1 covers a much smaller fraction of the $\mathcal{M}$-axis than the other catalogs. Additionally, the number of events is much smaller for GWTC-1. That makes sense since the detectors were less sensitive in O1 and O2 compared to O3a and O3b, so less detections are expected, but it also means statements on the whole catalog have a higher statistical uncertainty (the number of events corresponds to a sample size for these statements).

Speaking of sample size, it is also possible to examine how this quantity changed between



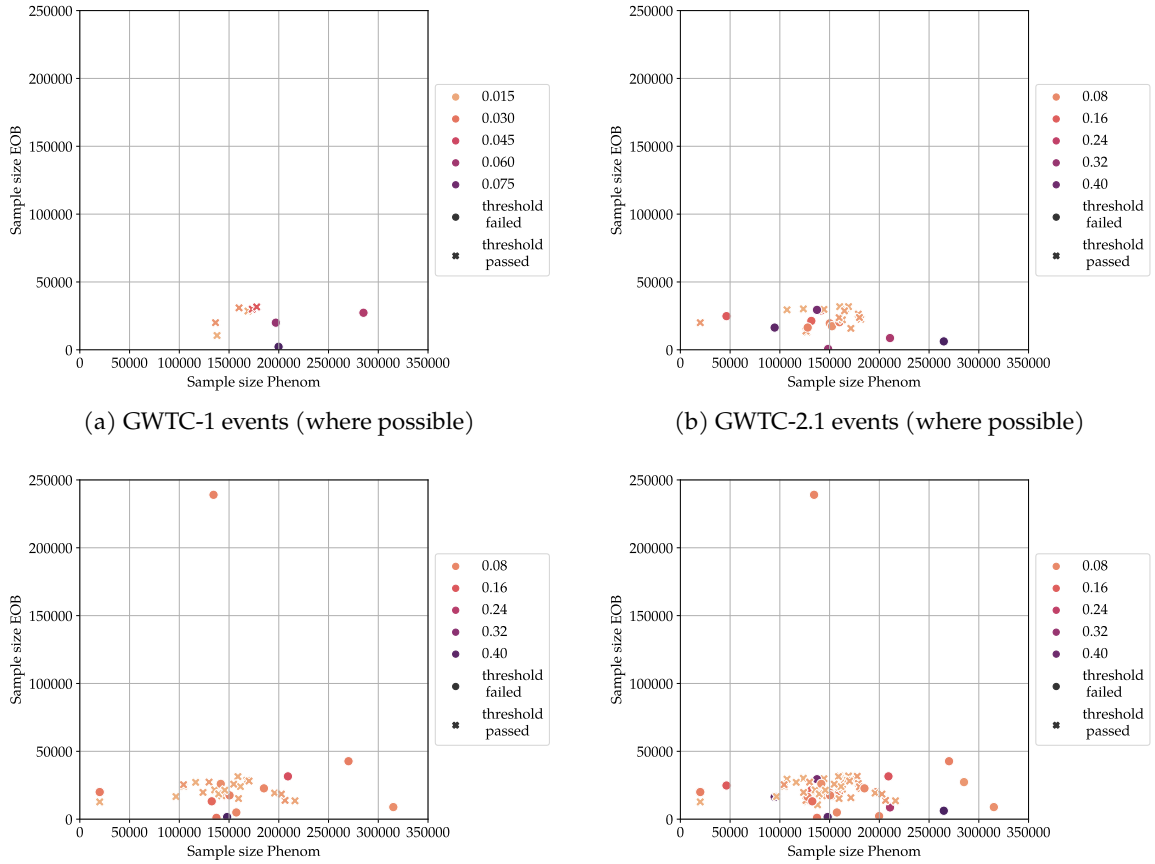
(a) GWTC-1 events (where possible)

(b) GWTC-2.1 events (where possible)

(c) GWTC-3 events

(d) All events (where possible)

Figure 4.10.: Visualization of events from different catalogs a point in the q-χ_{eff}-plane. The coordinates are determined by the medians of the respective parameters and there is a coloring based on the maximum JSD value between the marginalized `IMRPhenomXPHM` and `SEOBNRv4PHM` posteriors for the set of parameters used throughout this section (results do not change significantly when excluding M, m_1, m_2, in particular not if an event fails/ passes). To make it more apparent which value surpasses the 50% threshold, two different symbols are used (see the respective plot legend).



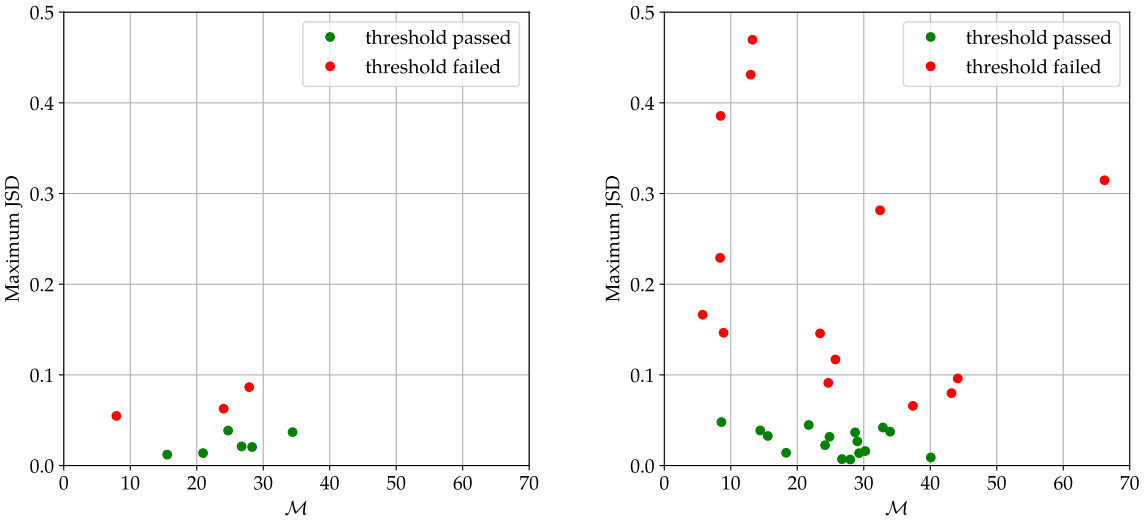
(a) GWTC-1 events (where possible)

(b) GWTC-2.1 events (where possible)

(c) GWTC-3 events

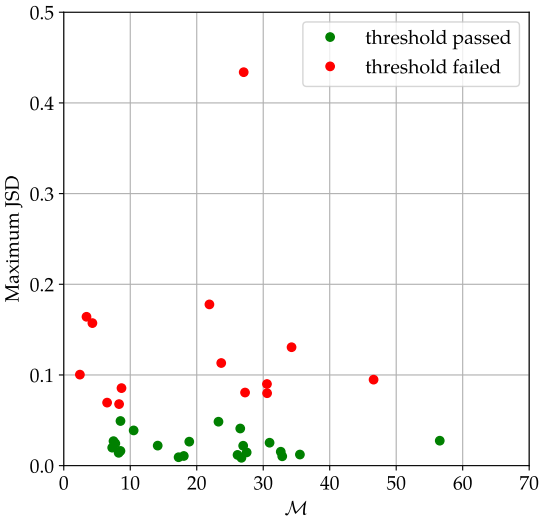
(d) All events (where possible)

Figure 4.11.: Visualization of events from different catalogs as points in the sample size plane. The coordinates are determined by the medians of the respective parameters and there is a coloring based on the maximum JSD value between the marginalized `IMRPhenomXPHM` and `SEOBNRv4PHM` posteriors for the set of parameters used throughout this section (results do not change significantly when excluding M, m_1, m_2). To make it more apparent which value surpasses the 50% threshold, two different symbols are used (see the respective plot legend).

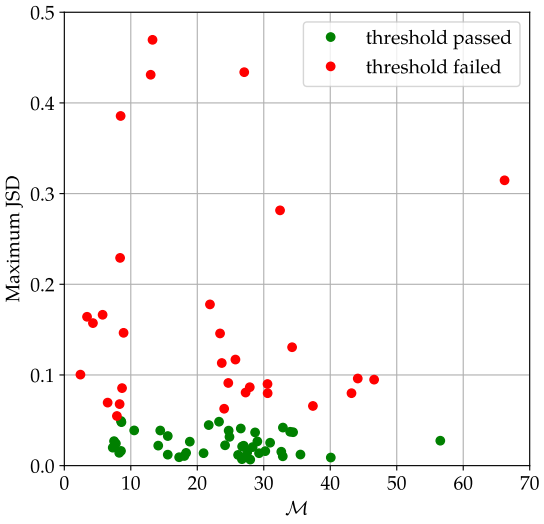


(a) GWTC-1 events (where possible)

(b) GWTC-2.1 events (where possible)



(c) GWTC-3 events



(d) All events (where possible)

Figure 4.12.: Visualization of events from different catalogs as points on the $\mathcal{M}$-axis. The coordinates are determined by the median of $\mathcal{M}$ and the maximum JSD value between the marginalized `IMRPhenomXPHM` and `SEOBNRv4PHM` posteriors for the set of parameters used throughout this section (results do not change significantly when excluding M, m_1, m_2). To make it more apparent which value surpasses the 50% threshold, two different colors are used (see the respective plot legend).

the catalogs (figure 4.11[15]). Again, GWTC-1 tends to have events in the non-extreme region while the other catalogs are more spread out. However, sampling effects should still not be dominant when inferring statements on whole catalogs. Hence, these plots can only partially explain why the agreement got much worse from GWTC-1 to GWTC-2.1 and not at all why it is worse for GWTC-2.1 than for GWTC-3 (which covers even wider regions of the parameter space). The plots in the q-χ_{eff}-plane and $\mathcal{M}$-axis examined before also do not explain that because other than for GWTC-3, where many of the events with bad agreement were in extreme regions of the parameter space, the ones with bad agreement in GWTC-2.1 to a very significant fraction lie in the non-extreme regions.

After seeing how the waveform agreement compares against the ones from previous catalogs, we now provide some thoughts and reasons on whether the inference quality is good or not. With what we have looked at, it is hard to say if the quality is good, simply because that highly depends on the expected uncertainty from known biases and errors. However, our results show that it is almost certainly not bad. Some events have bad agreement, but a great portion of that can be explained using existing knowledge about the waveform models themselves and their agreement (e.g. extreme mass ratios, multimodality) or other features of the data (e.g. noise, especially non-Gaussian one). To prove this with numbers, out of the $10/36$ events with agreement concerns, $7-8$ of them lie in regions where certain differences are expected. Comparing GWTC-3 results to the ones for previous catalogs then showed that the amount of inconsistencies we find fits in well. In fact, it may even be slightly better than one might expect when considering which kind of signals were detected (except for M, m_1, which show excess disagreement; however, the agreement for them is still better than for GWTC-2.1). Of course, this classification also depends on how severe one assesses the differences marked by the 50% thresholds to be. In our opinion, which is for instance based on figure 4.2, not exceeding this threshold may not necessarily be sufficient to infer good quality, but at least, it ensures the quality is not too bad. If we instead follow the statement from GWTC-1 mentioned in subsection 4.2.1, it indeed points to good data quality.

For the sake of completeness, we also provide the results for the agreement of GWTC-1, -2.1 events (tables 4.10 and 4.11, figure 4.13), but they do not reveal any new statements. The overall pattern looks fairly similar to the results from GWTC-3, although more events seem to be marked red (which corresponds to the worse agreement found before). This is confirmed by number of events marked for bad data quality: following the method used before (choose red events from 50% thresholds), this would be $11/39$ events (28.2%) or $14/39$ (35.9%) at most, depending on how strict the thresholds are applied. For comparison, the number for GWTC-3 was $10/36$ (27.8%) $(12/36$ (33.3%) initially), which corresponds to a very similar proportion and means that in this regard, not much has changed compared to GWTC-3. When considering the previous catalogs separately, things change because $0/9$ (0%) or at most $2/9$ (22.2%) events from GWTC-1 are marked for bad quality, while $9/30$ (30%) or at most $12/30$ (40%) of events from GWTC-2 are. This is again in accordance with the results presented earlier in this subsection.

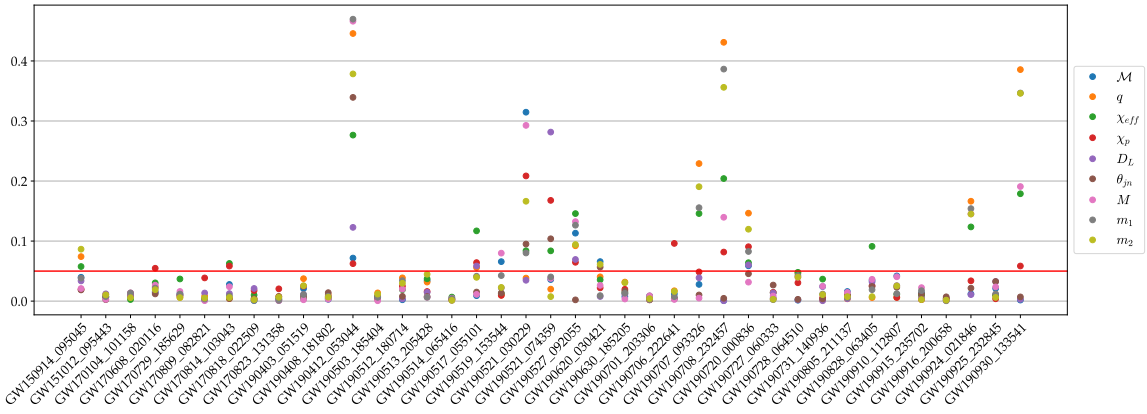
[15] This figure justifies the thresholds chosen because sampling error will not play a role for the majority of events (see 4.1), so our inference should be robust against sampling error.

Event	JSD	JSD 2	Mean diff	Median diff
GW150914_095045	$q\ \chi_{\rm eff}\ m_2$	θ_{jn}	$q\ \chi_{\rm eff}\ m_2$	m_2
GW151012_095443				
GW170104_101158				
GW170608_020116	χ_p			
GW170729_185629				
GW170809_082821				
GW170814_103043	$\chi_{\rm eff}\ \chi_p$		$\chi_{\rm eff}\ \chi_p$	$\chi_{\rm eff}\ \chi_p$
GW170818_022509				
GW170823_131358				
GW190403_051519				
GW190408_181802		θ_{jn}		
GW190412_053044	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\chi_{\rm eff}\ \theta_{jn}$	$q\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1\ m_2$
GW190503_185404				
GW190512_180714				
GW190513_205428				
GW190514_065416				
GW190517_055101	$q\ \chi_{\rm eff}\ \chi_p\ D_L$		$q\ \chi_{\rm eff}\ \chi_p\ D_L$	$q\ \chi_{\rm eff}\ \chi_p\ D_L$
GW190519_153544	$\mathcal{M}\ M$		$\mathcal{M}\ M$	$\mathcal{M}\ M$
GW190521_030229	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ \chi_p\ \theta_{jn}\ M$	$\mathcal{M}\ \chi_p\ M\ m_1\ m_2$	$\mathcal{M}\ \chi_p\ M\ m_1\ m_2$
GW190521_074359	$\chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}$	θ_{jn}	$\chi_{\rm eff}\ \chi_p\ D_L$	$\chi_{\rm eff}\ \chi_p\ D_L$
GW190527_092055	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ M\ m_1$	$\mathcal{M}\ \chi_p\ M\ m_1$	$\chi_p\ M\ m_1$
GW190620_030421	$\mathcal{M}\ \theta_{jn}\ m_2$	θ_{jn}	$\mathcal{M}\ m_2$	$\mathcal{M}\ m_2$
GW190630_185205				
GW190701_203306				
GW190706_222641	χ_p		χ_p	χ_p
GW190707_093326	$q\ \chi_{\rm eff}\ m_1\ m_2$		$q\ \chi_{\rm eff}\ m_1\ m_2$	$q\ \chi_{\rm eff}\ m_1\ m_2$
GW190708_232457	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	θ_{jn}	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$
GW190720_000836	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ m_1\ m_2$	$\mathcal{M}\ D_L\ \theta_{jn}$	$\chi_p\ D_L$	χ_p
GW190727_060333		θ_{jn}		θ_{jn}
GW190728_064510		$M\ m_1$		
GW190731_140936				
GW190805_211137				
GW190828_063405	$\chi_{\rm eff}$	θ_{jn}	$\chi_{\rm eff}$	$\chi_{\rm eff}$
GW190910_112807				$\mathcal{M}\ M$
GW190915_235702				
GW190916_200658				
GW190924_021846	$q\ \chi_{\rm eff}\ M\ m_1\ m_2$	$\theta_{jn}\ M\ m_1$	$q\ \chi_{\rm eff}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ M\ m_1\ m_2$
GW190925_232845		θ_{jn}		
GW190930_133541	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$\theta_{jn}\ M$	$q\ \chi_{\rm eff}\ \chi_p\ m_1\ m_2$	$q\ \chi_{\rm eff}\ M\ m_1\ m_2$

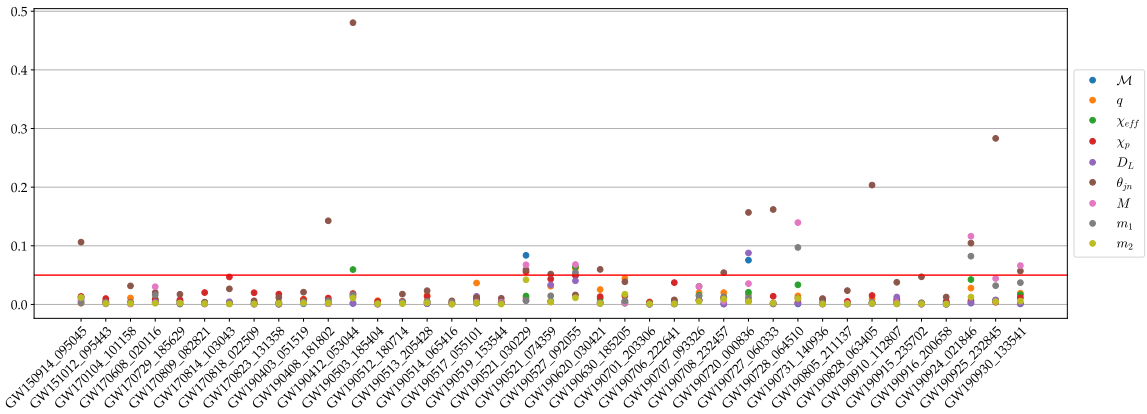
Table 4.10.: Results for agreement of `IMRPhenomXPHM` and `SEOBNRv4PHM` for GWTC-1, -2.1 events (where runs with both are available) using 50% thresholds and `nocosmo` data

Event	JSD	JSD 2	Mean diff	Median diff
GW150914_095045	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ M\ m_1\ m_2$
GW151012_095443	$q\ D_L\ m_2$	χ_p	D_L	D_L
GW170104_101158	$q\ \theta_{jn}\ M\ m_1$	$q\ \theta_{jn}$	$D_L\ M\ m_1$	$q\ D_L\ \theta_{jn}\ M\ m_1$
GW170608_020116	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\theta_{jn}\ M\ m_1$	$\mathcal{M}\ q\ \chi_p\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_p\ D_L\ M\ m_1$
GW170729_185629	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1$	θ_{jn}	$\mathcal{M}\ \chi_{\rm eff}\ M\ m_1$	$\mathcal{M}\ \chi_{\rm eff}\ \theta_{jn}\ M$
GW170809_082821	$\chi_p\ D_L$	χ_p	$\chi_p\ D_L$	$\chi_p\ D_L$
GW170814_103043	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ M$	$\chi_p\ \theta_{jn}$	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ M$	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ M$
GW170818_022509	$\chi_{\rm eff}\ \chi_p\ D_L$	χ_p	$\chi_p\ D_L$	$\chi_{\rm eff}\ \chi_p\ D_L$
GW170823_131358	χ_p	$\chi_p\ \theta_{jn}$	χ_p	χ_p
GW190403_051519	$\mathcal{M}\ q\ \chi_p\ D_L\ m_1\ m_2$	θ_{jn}	$\mathcal{M}\ \chi_p\ D_L\ m_2$	$\mathcal{M}\ \chi_p\ D_L\ \theta_{jn}\ m_2$
GW190408_181802	$q\ \theta_{jn}$	$\chi_p\ \theta_{jn}$	θ_{jn}	$\theta_{jn}\ m_1$
GW190412_053044	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1\ m_2$
GW190503_185404	$q\ m_2$		$q\ m_2$	$q\ \theta_{jn}$
GW190512_180714	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	θ_{jn}	$q\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_p\ M\ m_1\ m_2$
GW190513_205428	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ m_2$	$q\ \chi_p\ \theta_{jn}$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ m_2$
GW190514_065416				
GW190517_055101	$q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ D_L\ \theta_{jn}$	$q\ \chi_{\rm eff}\ \chi_p\ D_L\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1\ m_2$
GW190519_153544	$\mathcal{M}\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	θ_{jn}	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$
GW190521_030229	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$
GW190521_074359	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1$	$q\ \chi_p\ D_L\ \theta_{jn}\ m_1$	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1$
GW190527_092055	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_p\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_p\ D_L\ M\ m_1$
GW190620_030421	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_2$	$q\ \chi_p\ \theta_{jn}$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ M\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_2$
GW190630_185205	$q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ m_1\ m_2$	$q\ \chi_p\ \theta_{jn}\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ m_1$	$q\ \chi_{\rm eff}\ \chi_p\ D_L\ m_1\ m_2$
GW190701_203306			$\mathcal{M}\ M$	$\mathcal{M}\ M$
GW190706_222641	$\mathcal{M}\ q\ \chi_p\ m_2$	χ_p	$\mathcal{M}\ q\ \chi_p\ m_2$	$\mathcal{M}\ q\ \chi_p\ m_2$
GW190707_093326	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \theta_{jn}\ M\ m_1$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ m_1\ m_2$
GW190708_232457	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_p\ \theta_{jn}\ m_1$	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$
GW190720_000836	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_2$	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ m_2$
GW190727_060333	$\chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}$	$\chi_p\ \theta_{jn}$	$\chi_p\ D_L\ \theta_{jn}$	$\chi_p\ D_L\ \theta_{jn}$
GW190728_064510	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_p\ M\ m_1\ m_2$
GW190731_140936	$\mathcal{M}\ \chi_{\rm eff}\ M\ m_1\ m_2$		$\mathcal{M}\ \chi_{\rm eff}\ M\ m_1\ m_2$	$\mathcal{M}\ \chi_{\rm eff}\ M\ m_1\ m_2$
GW190805_211137	$\mathcal{M}\ \theta_{jn}\ M$	θ_{jn}	$\mathcal{M}\ \chi_p\ M$	$\mathcal{M}\ \chi_p\ M$
GW190828_063405	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1$	$\chi_p\ \theta_{jn}$	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1$	$\mathcal{M}\ \chi_{\rm eff}\ \chi_p\ D_L\ M\ m_1$
GW190910_112807	$\mathcal{M}\ \chi_{\rm eff}\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$D_L\ \theta_{jn}$	$\mathcal{M}\ \chi_{\rm eff}\ D_L\ M\ m_1\ m_2$	$\mathcal{M}\ D_L\ M\ m_1\ m_2$
GW190915_235702	$\mathcal{M}\ \chi_{\rm eff}\ \theta_{jn}\ M\ m_1$	θ_{jn}	$\mathcal{M}\ M\ m_1$	$\mathcal{M}\ \chi_{\rm eff}\ \theta_{jn}\ M\ m_1$
GW190916_200658		θ_{jn}		
GW190924_021846	$\mathcal{M}\ q\ \chi_{\rm eff}\ \chi_p\ D_L\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1\ m_2$
GW190925_232845	$\mathcal{M}\ q\ D_L\ \theta_{jn}\ M\ m_1$	$\theta_{jn}\ M\ m_1$	$\mathcal{M}\ D_L\ \theta_{jn}\ M\ m_1$	$D_L\ \theta_{jn}$
GW190930_133541	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ \theta_{jn}\ M\ m_1$	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$	$q\ \chi_{\rm eff}\ \chi_p\ M\ m_1\ m_2$

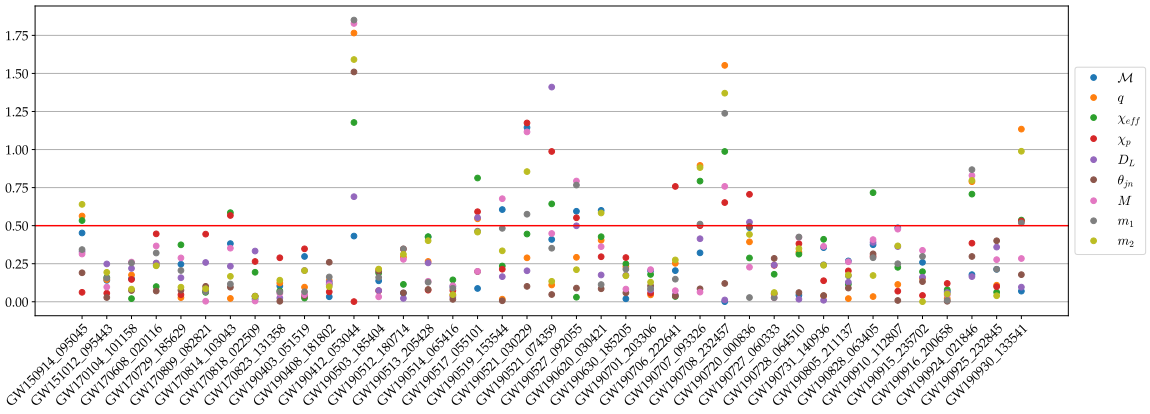
Table 4.11.: Results for agreement of `IMRPhenomXPHM` and `SEOBNRv4PHM` for GWTC-1, -2.1 events (where runs with both are available) using 20% thresholds and `nocosmo` data


0.0
0.1
0.2
0.3
0.4
$\mathcal{M}$
q
χ_{eff}
χ_p
D_L
θ_{jn}
M
m_1
m_2
GW150914_095045
GW151012_095443
GW170104_101158
GW170608_020116
GW170729_185629
GW170809_082821
GW170814_103043
GW170818_022509
GW170823_131358
GW190403_051519
GW190408_181802
GW190412_053044
GW190503_185404
GW190512_180714
GW190513_205428
GW190514_065416
GW190517_055101
GW190519_153544
GW190521_030229
GW190521_074359
GW190527_092055
GW190620_030421
GW190630_185205
GW190701_203306
GW190706_222641
GW190707_093326
GW190708_232457
GW190720_000836
GW190727_060333
GW190728_064510
GW190731_140936
GW190805_211137
GW190828_063405
GW190910_112807
GW190915_235702
GW190916_200658
GW190924_021846
GW190925_232845
GW190930_133541

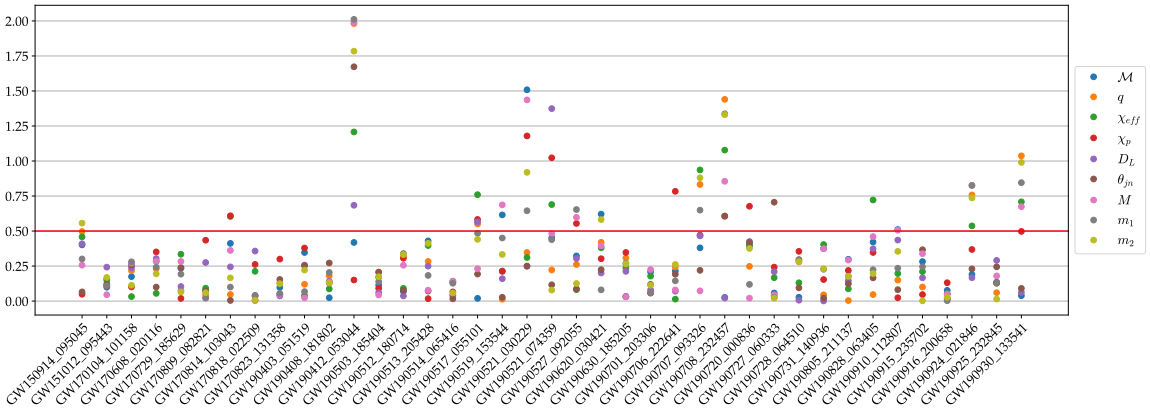
(a) JSD


0.0
0.1
0.2
0.3
0.4
0.5
$\mathcal{M}$
q
χ_{eff}
χ_p
D_L
θ_{jn}
M
m_1
m_2
GW150914_095045
GW151012_095443
GW170104_101158
GW170608_020116
GW170729_185629
GW170809_082821
GW170814_103043
GW170818_022509
GW170823_131358
GW190403_051519
GW190408_181802
GW190412_053044
GW190503_185404
GW190512_180714
GW190513_205428
GW190514_065416
GW190517_055101
GW190519_153544
GW190521_030229
GW190521_074359
GW190527_092055
GW190620_030421
GW190630_185205
GW190701_203306
GW190706_222641
GW190707_093326
GW190708_232457
GW190720_000836
GW190727_060333
GW190728_064510
GW190731_140936
GW190805_211137
GW190828_063405
GW190910_112807
GW190915_235702
GW190916_200658
GW190924_021846
GW190925_232845
GW190930_133541

(b) JSD 2



(c) Mean criterion



(d) Median criterion

Figure 4.13.: Agreement for all events from GWTC-1, -2.1 and set of parameters specified in plot legend. Each dot represents the criterion value for a parameter and the 50% threshold for the respective criterion is visualized as a red line.

5. Principal Components of Gravitational Wave Event Data

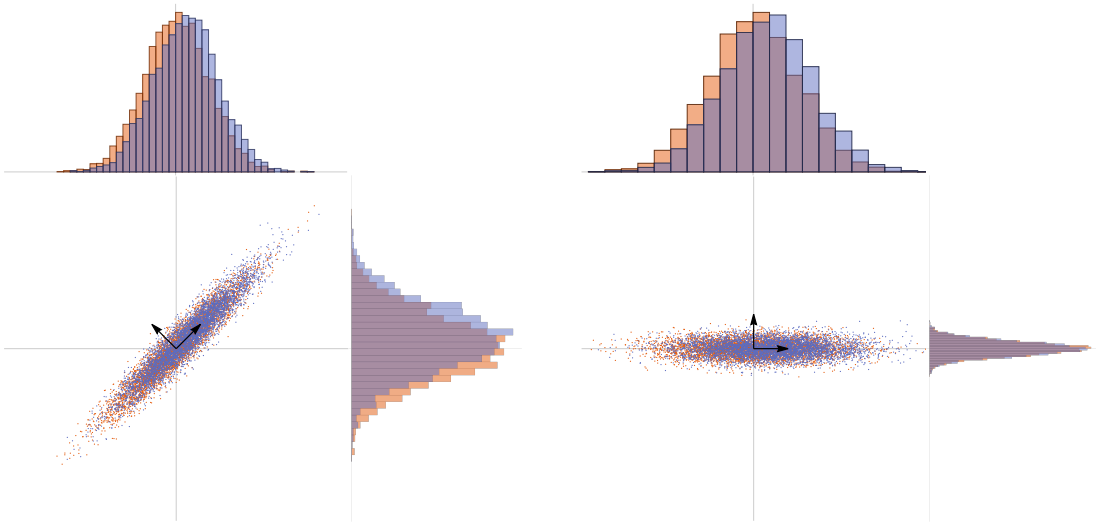
After reviewing the technique of PCA in the chapter 3, the following question arises: how can it be related to the topic of GW data? The answer is that it allows another viewpoint on the high-dimensional posteriors, of which we can only analyze 1D representations. PCA can be used to find a new coordinate basis with special, potentially useful properties (our goal is not to reduce the dimensionality of posteriors). It will be interesting to see if the corresponding 1D representations reveal new systematics, which we are not able to see in chapter 4 when using "ordinary" parameters as the basis of the parameter space.

5.1. Preliminary Considerations

5.1.1. Idea

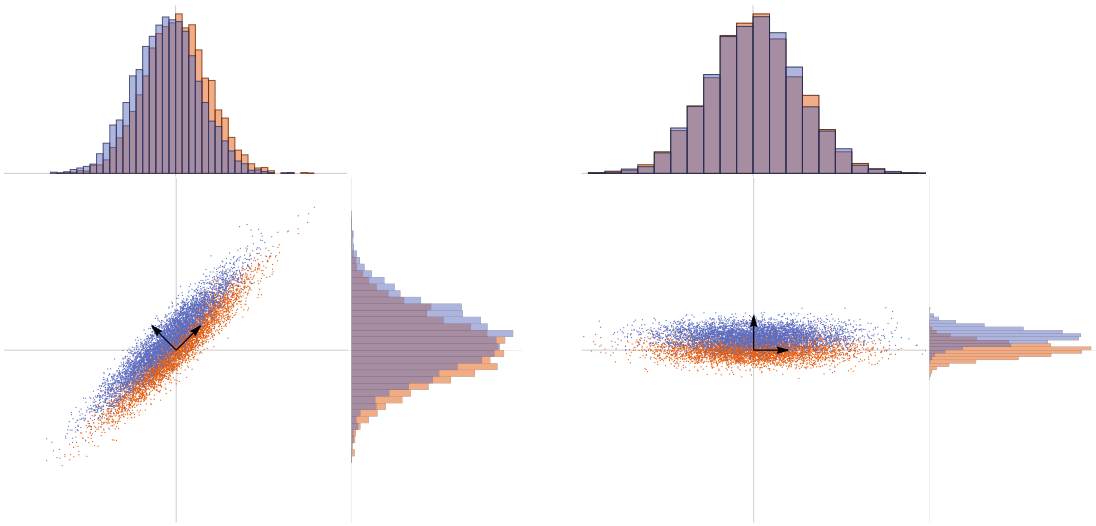
The way we intend to use PCA is not how this technique is usually applied, so it is necessary to explain it. The overall goal in this work is to compare posteriors and that has not changed for this chapter. However, it is not our goal to compare PCs of different waveform posteriors as it does not really make sense. This has several reasons, the most striking one being that even in case of very similar distributions, the eigenvectors might differ by a factor of -1 or their order might differ, so despite in principle similar distributions, one could potentially find big differences when comparing the PCs. Instead, we compute the eigenvectors determining the transformation into the PCs of each of them and then transform both posteriors into the same basis to compare them with respect to the coordinate axes given by this basis. This ensures that differences found really correspond to differences in the posteriors.

In general, the last PC is often viewed as the least important one since it contains the least amount of variance (information). For the way we intend to use PCA, however, this statement is not necessarily true and a priori, each PC is equally important to us. This is because we want to compare two data sets in the same basis and for such a comparison, not only the variances are important, but also the location (as measured by means, medians). Such differences can appear in the first PC (as figure 5.1 shows). But it is also possible to construct a very similar example, that requires looking at the second PC in order to be able to tell the distributions apart (figure 5.2). Telling apart is meant in a visual sense, i.e. from looking at the histograms as 1D representations of the data, but also applies for the JSD using the 50% threshold of 0.05. For this reason, we will not restrict ourselves to PCs containing a certain amount of variance (at least not a priori, there might be exclusions).



(a) Original data ($D_{\mathrm{JS},x} = 0.018 = D_{\mathrm{JS},y}$) (b) PCs ($D_{\mathrm{JS},x} = 0.020,\ D_{\mathrm{JS},y} = 0.002$)

Figure 5.1.: Samples from two normal distributions ($\mu_1 = (0,0),\ \mu_2 = (0.3,0.3),\ \sigma_1 = (1,1) = \sigma_2$) and corresponding PCs. While the maximum JSD does not change much, minimum and average JSD decrease.



(a) Original data ($D_{\mathrm{JS},x} = 0.018 = D_{\mathrm{JS},y}$) (b) PCs ($D_{\mathrm{JS},x} = 0.002,\ D_{\mathrm{JS},y} = 0.46$)

Figure 5.2.: Samples from two normal distributions ($\mu_1 = (0,0),\ \mu_2 = (-0.3,0.3),\ \sigma_1 = (1,1) = \sigma_2$), which is not the same as in the previous figure) and corresponding PCs. In this case, the maximum and average JSD increase while the minimum decreases.

5.1.2. Formalism

For the parameters of common interest in GW data analysis, posteriors have to be normalized before applying a PCA. That is because their ranges differ over several orders of magnitude and so do the variances (D_L has values and variances $\gtrsim 10^7$, $\mathcal{M}$ around 10 or even 100, while $\chi_{\text{eff}}, \chi_p, q$ have a maximum absolute value of 1 and thus also variances of the order of 1). As already discussed in the subsection 3.2.2, that would lead to the PCA not serving our purpose of finding the "most important" axes of the data. One could now normalize each posterior by dividing it by its corresponding standard deviation, such that every parameter has $\sigma = 1$. However, in general, the standard deviations for a parameter will differ from one waveform to the other, which means that their posteriors will not change in the same way when using this normalization method. This is a problem because want to compare the transformed posteriors as well and if the normalization already changes their criterion value, then we cannot know whether this normalization process or actual differences in the transformed data caused bad agreement indicated by this criterion.

A solution of this problem is to normalize the posteriors by dividing both by the same factor. We choose that factor to be the standard deviation of the posterior for which covariance matrix (= correlation matrix here, normalized data is used) and eigenvectors are computed. The parameters of the respective other posterior will not have unit standard deviation if we do this, but it also should not be that much smaller/ greater as for the vast majority of events, the posteriors have roughly the same shape and thus roughly the same variance. The motivation to use this workaround is that every criterion we apply to assess differences does not change when both distributions are normalized in this way. This statement will be proven now (where proofs are mostly calculations here).

Each criterion is applied to one-dimensional data (samples for one parameter or PC) consisting of many measurements x and in this case, normalization is nothing but a transformation using the diffeomorphism $x \mapsto \hat{x} := cx$ with $c = \frac{1}{\sigma}$ (we assume $c \neq 0$). We are now interested in how the PDF changes under normalization. That can be inferred from the fact that PDFs must integrate up to unity. Denoting the PDF of the original data with $p(x)$ and the one of the normalized data with $\hat{p}(\hat{x})$, this constraint yields (we directly use $c\mathbb{R} = \mathbb{R}$):

$$\begin{aligned} & \int_{\mathbb{R}} \hat{p}(cx)\, c dx = \int_{\mathbb{R}} \hat{p}(\hat{x})\, d\hat{x} = 1 = \int_{\mathbb{R}} p(x)\, dx \\ \Leftrightarrow \quad & 0 = \left(\int_{\mathbb{R}} \hat{p}(cx)\, c dx \right) - \left(\int_{\mathbb{R}} p(x)\, dx \right) = \int_{\mathbb{R}} (c\hat{p}(cx) - p(x))\, dx \end{aligned}$$

where $d\hat{x} = d(cx) = d(c)x + cd(x) = cdx$ because c is a constant. Since $\hat{p}(\hat{x}), p(x) \geq 0$, the integral being zero $\forall \hat{p}, p$ is equivalent to the integrand being zero, so we have proven:

$$\hat{p}(\hat{x}) = \hat{p}(cx) = \frac{p(x)}{c}\,. \tag{5.1}$$

This formula enables us to calculate how criteria behave under $x \mapsto \hat{x}$ since it e.g. directly tells us how the median behaves. However, it is not apparent how mean and standard

deviation behave, so that has to be calculated. Using the transformation rule yields:

$$\hat{\mu} = \int_{\mathbb{R}} \hat{x}\hat{p}(\hat{x})\, d\hat{x} = \int_{\mathbb{R}} cx\frac{p(x)}{c}\, cdx = c\int_{\mathbb{R}} xp(x)\, dx = c\mu \tag{5.2}$$

$$\hat{\sigma} = \sqrt{\int_{\mathbb{R}} (\hat{x}-\hat{\mu})^2\hat{p}(\hat{x})\, d\hat{x}} = \sqrt{\int_{\mathbb{R}} (cx-c\mu)^2\frac{p(x)}{c}\, cdx} = \sqrt{c^2\int_{\mathbb{R}} (x-\mu)^2 p(x)\, dx} = c\sigma\,. \tag{5.3}$$

This should not be too surprising since summation and integration are linear operations, so we would gather that $E[cX] = cE[X]$ holds. We can now use these results to see how the mean criterion (4.8) behaves:

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{(\hat{\sigma}_1 + \hat{\sigma}_2)/2} = \frac{c\mu_1 - c\mu_2}{(c\sigma_1 + c\sigma_2)/2} = \frac{\mu_1 - \mu_2}{(\sigma_1 + \sigma_2)/2}\,. \tag{5.4}$$

Therefore, it is invariant under the transformation $x \mapsto \hat{x}$.

Similarly, one can compute the behaviour of D_{KL}. To do that, we will use its integral version and not (4.3) because that makes it easier to apply the transformation:

$$D_{\mathrm{KL}}(\hat{p},\hat{q}) = \int_{\mathbb{R}} \hat{p}(\hat{x}) \log\left(\frac{\hat{p}(\hat{x})}{\hat{q}(\hat{x})}\right) d\hat{x} = \int_{\mathbb{R}} \frac{p(x)}{c} \log\left(\frac{p(x)/c}{q(x)/c}\right) cdx = \int_{\mathbb{R}} p(x)\log\left(\frac{p(x)}{q(x)}\right) dx\,. \tag{5.5}$$

Therefore, $D_{\mathrm{KL}}(\hat{p},\hat{q}) = D_{\mathrm{KL}}(p,q)$ and from (4.2), we immediately get $D_{\mathrm{JS}}(\hat{p},\hat{q}) = D_{\mathrm{JS}}(p,q)$. The second JSD criterion used is also invariant under normalization because it divides each distribution by its standard deviation anyway.

Although (5.1) already gives a hint at how the medians $\hat{p}_{50}, p_{50}$ and also credible intervals are related, it is also possible to show this for arbitrary percentiles p_t using their definition and the transformation rule again:

$$\int_{-\infty}^{p_t} p(x)\, dx = \int_{-\infty}^{p_t/c} p(\hat{x})\, d\hat{x} = \int_{-\infty}^{\hat{p}_t} p(\hat{x})\, d\hat{x}\,. \tag{5.6}$$

For the behaviour of the median criterion (4.9) we then obtain (assuming $p_{50} \geq q_{50}$ again and using the $t\%$ credible interval):

$$\begin{aligned}\frac{\hat{p}_{50} - \hat{q}_{50}}{\left(\left(\hat{p}_{50+t/2} - \hat{p}_{50}\right) + \left(\hat{q}_{50} - \hat{q}_{50-t/2}\right)\right)/2} &= \frac{\frac{p_{50}}{c} - \frac{q_{50}}{c}}{\left(\left(\frac{p_{50+t/2}}{c} - \frac{p_{50}}{c}\right) + \left(\frac{q_{50}}{c} - \frac{q_{50-t/2}}{c}\right)\right)/2} \\ &= \frac{p_{50} - q_{50}}{\left(\left(p_{50+t/2} - p_{50}\right) + \left(q_{50} - q_{50-t/2}\right)\right)/2}\,.\end{aligned} \tag{5.7}$$

Therefore, we finally have shown that all criteria used in this work are invariant under the chosen method of normalization (they cannot distinguish original, normalized data), so any

differences showing up for PCs can really be attributed to differences caused by the change of representation. A drawback of this normalization method is that the comparison of variances between the posteriors does not make sense anymore because their total variance will most likely differ (one has $\sum_i \mathrm{var}(x_i) = n = \#\text{parameters}$ and the other may have a similar value, but not the same). That means we have to compare relative variances instead of the absolute values when comparing PCs from different posteriors. Such a problem does not appear when comparing the variances of a fixed posterior in the original and principal axes basis because the total variance is still preserved under transformations.

While we do use normalization, mean-centering will not be applied, although it does not affect the transformation we get because covariance measures deviations around the mean. There are two reasons for that: firstly, it would affect the values of most of the criteria we use since the compared distributions will have different means. Secondly, we are specifically interested in differences of distributions, so by not mean-centering we retain the ability to extract information about the posteriors from their means (even if we somehow had criteria which are not affected by centering, it would therefore make no sense).[1]

[1] If we make the same change as for the normalization process and subtract the same amount from both distributions, it would only shift the distributions on the x-axis and therefore be somehow pointless (it does not provide an advantage in the analysis of the PCs like normalization did; the only change to the PCs is that they also have different means, zero in case of mean-centering).

5.2. Results

Before starting with the actual analysis of results, we have to specify the set of parameters used. In principle, this is the same as before, but we exclude M, m_1, m_2. Previously, including them did not require additional effort and even lead to interesting results. That changes in the context of PCA, where the fact that they are highly correlated with other parameters and among each other would cause many non-zero terms in the covariance matrix. This would greatly affect the eigenvectors. Additionally, PCA results are harder to examine for a higher number of parameters. Therefore, it does not make sense to include them here and the set of parameters used in this section (often referred to as "original parameters") is:

$$[\mathcal{M}, q, \chi_{\text{eff}}, \chi_p, D_L, \theta_{jn}]\,. \tag{5.8}$$

The criteria and thresholds we use do not change, so it is possible to compare statements obtained for the original parameters and for the principal axes. We can now see why the mean criterion was used in chapter 4 (although the median criterion is expected to be more stable), it ensures that the posteriors do not show exotic behaviour for the mean regularly. This would have a substantial and negative impact on the analyses in this section since the mean plays a very important role in how the transformations into the PCs are computed (covariances are measured around the mean). For this reason, the result that mean and median criterion agree broadly is important in this context.

Due to the normalization process we established, there will always be two comparisons, one in the basis given by the eigenvectors of the covariance matrix computed from the `IMRPhenomXPHM` posterior and one for the `SEOBNRv4PHM` posterior. To simplify wording throughout this section, we will refer to comparisons of PCs when comparing components of different posteriors in these new bases despite knowing it is not perfectly accurate (at least for one of them, it will indeed be the PCs which we look at, though).

5.2.1. Regular Principal Components

As already mentioned, analysing the transformed posteriors is not a completely new task, but very similar to what has been done in chapter 4. This section will thus use the methods, criteria, thresholds etc. developed there to ensure that statements inferred during the analysis are comparable to the ones for the original parameters. That means we look at the same representations of the agreement, i.e. separate analyses for each event (tables 5.1, 5.2 and figure 5.3) and each PC (table 5.3). An additional systematic which is new and very relevant in this subsection is the comparison of the agreement for each event when going from original parameters to PCs (figures 5.4, 5.5).

The first impression from the tables 5.1, 5.2 on all events is that both PC bases show less disagreement than the original parameters, there are significantly less events marked in red (almost half). The same is true for table 5.3, there seems to be a general decrease in how often events exceed the chosen thresholds for each PC compared to how often this

Event	JSD	JSD 2	Mean difference	Median difference
GW191103_012549				
GW191105_143521				
GW191109_010717	1 4 6		4 6	4 6
GW191113_071753				1
GW191126_115259				
GW191127_050227	6			6
GW191129_134029				
GW191204_110529				
GW191204_171526	3			
GW191215_223052				
GW191216_213338	3 5	4	3 5	3 5
GW191219_163120	3	4	3	1 3
GW191222_033537				
GW191230_180458				
GW200105_162426	4 5		4	4
GW200112_155838	1 4	4	1	1 4
GW200115_042309	4		4	4
GW200128_022011	2		2	2
GW200129_065458	1 3	5	1 3	1 3
GW200202_154313				
GW200208_130117				
GW200208_222617	1 4 6	1 6	1 4	1 4
GW200209_085452				
GW200210_092254				
GW200216_220804				
GW200219_094415				
GW200220_061928				
GW200220_124850				
GW200224_222234	5			5
GW200225_060421				
GW200302_015811	1		1	1
GW200306_093714		6		
GW200308_173609	1			
GW200311_115853	5		5	5
GW200316_215756	1	3	1	1
GW200322_091133	1			

(a) Phenom basis

Event	JSD	JSD 2	Mean difference	Median difference
GW191103_012549				
GW191105_143521				
GW191109_010717	1 2 4 5		5	2 5
GW191113_071753		3		3
GW191126_115259				
GW191127_050227	5			6
GW191129_134029				
GW191204_110529				
GW191204_171526	4	3	4	4
GW191215_223052				
GW191216_213338	4 5	4 5	3 4	3 4
GW191219_163120	3 4	3	4	3 4
GW191222_033537				
GW191230_180458				
GW200105_162426	1 4 5	6	4	4
GW200112_155838	3		1 3	1 3
GW200115_042309	4	3	4	4
GW200128_022011	2		2	2
GW200129_065458	1 2 3 4 5 6	4	1 3 4	1 2 3 4
GW200202_154313				
GW200208_130117				
GW200208_222617	1 4 5	1	1 4 5	1 4 5
GW200209_085452				
GW200210_092254				
GW200216_220804				
GW200219_094415				
GW200220_061928				
GW200220_124850				
GW200224_222234				
GW200225_060421				
GW200302_015811			1	
GW200306_093714		6		
GW200308_173609	6		6	
GW200311_115853	5		5	5
GW200316_215756	1	3	1	1
GW200322_091133	6			

(b) EOB basis

Table 5.1.: Results for agreement of `IMRPhenomXPHM` and `SEOBNRv4PHM` PCs for GWTC-3 events using 50% thresholds and normalized `nocosmo` data. Event names visualized in green/ red belong to the lists good/ bad data identified in the previous chapter.

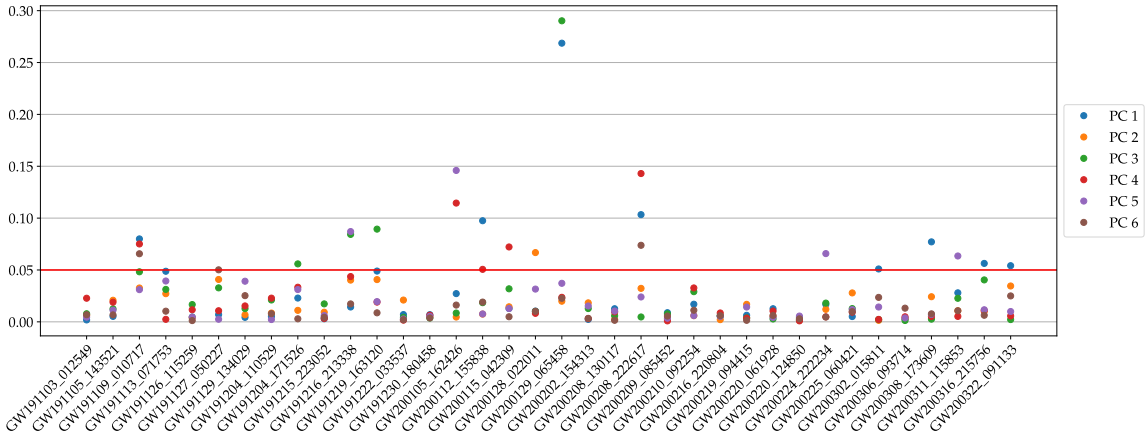
Event	JSD	JSD 2	Mean difference	Median difference
GW191103_012549	4		4	4
GW191105_143521	2 3 4 5	3	4	4 5
GW191109_010717	1 2 3 4 5 6	1 3	2 3 4 5 6	1 2 4 5 6
GW191113_071753	1 2 3 5 6	2 5 6	1 5	1 2
GW191126_115259	3 4	3 6	4	3
GW191127_050227	2 3 4 6	3	2 4 6	2 4 6
GW191129_134029	3 4 5 6	3	4 5 6	3 4 5 6
GW191204_110529	3 4	4	3	3 4
GW191204_171526	1 2 3 4 5		3 4 5	1 3 4 5
GW191215_223052	3	3		3
GW191216_213338	1 2 3 4 5 6	2 4 5	2 3 4 5 6	2 3 4 5 6
GW191219_163120	1 2 3 4 5	1 2 3 4 5	1 2 3 4	1 2 3 4 5 6
GW191222_033537	2		2	2
GW191230_180458				
GW200105_162426	1 4 5 6	1 4 5 6	1 4	1 4 5
GW200112_155838	1 3 4 6	4	1 4	1 3 4
GW200115_042309	1 2 3 4 5	1 3 5	3 4	1 3 4
GW200128_022011	1 2 3 5 6		2 5 6	1 2 5 6
GW200129_065458	1 2 3 4 5 6	3 5	1 2 3 4 5	1 2 3 4
GW200202_154313	2 3 5			
GW200208_130117	1 5	4 5	1	
GW200208_222617	1 2 4 5 6	1 2 6	1 4	1 4
GW200209_085452				1
GW200210_092254	1 3 4 6	1 2 4 6	3 4	4 6
GW200216_220804		3	4	
GW200219_094415	2 5		2 5	2 5
GW200220_061928	1 4	6	1 4	1 4
GW200220_124850				
GW200224_222234	1 2 3 5	5	3 5	3 5
GW200225_060421	2 3 5	3	2	2 3
GW200302_015811	1 5 6		1 6	1 6
GW200306_093714	6	6		4 6
GW200308_173609	1 2		1 2	1 2
GW200311_115853	1 2 3 5 6		1 2 3 5	1 2 5
GW200316_215756	1 2 3 4 5	3	1 2 3 5	1 2 3 5
GW200322_091133	1 2 6		1 2 6	1 2 6

(a) Phenom basis

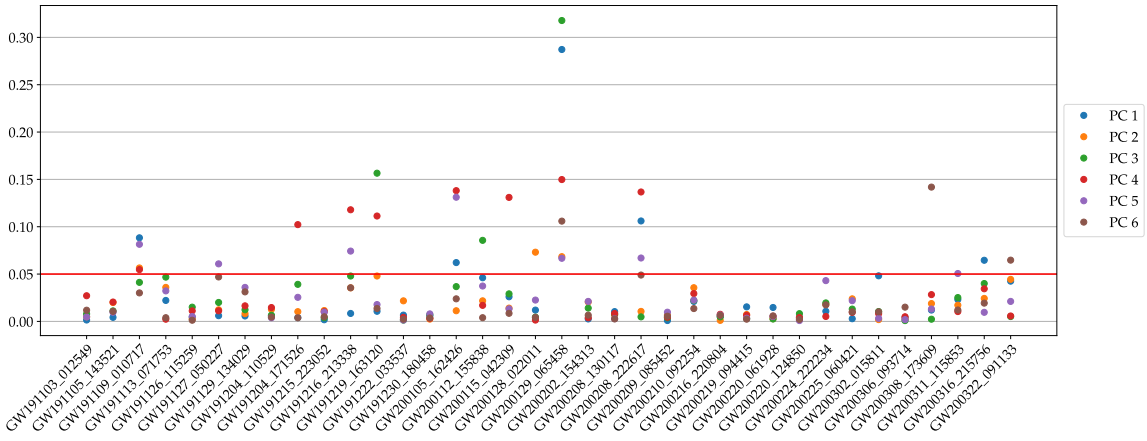
Event	JSD	JSD 2	Mean difference	Median difference
GW191103_012549	4 6		4 6	4 6
GW191105_143521	2 3 4 6	3 6	4	4
GW191109_010717	1 2 3 4 5 6	1 3	2 3 4 5 6	1 2 4 5 6
GW191113_071753	1 2 3 5	2 3 5	1 3	1 3
GW191126_115259	3 4	3 6		3
GW191127_050227	2 3 4 5 6	5	3 5 6	3 5 6
GW191129_134029	3 4 5 6	3	4 5 6	3 4 5 6
GW191204_110529	2 4		2 4	4
GW191204_171526	2 3 4 5	3	4 5	3 4 5
GW191215_223052	2 4	4	2	2
GW191216_213338	2 3 4 5 6	2 4 5	2 3 4 6	2 3 4 5 6
GW191219_163120	1 2 3 4 5 6	1 2 3 4 6	2 4	2 3 4
GW191222_033537	2		2	2
GW191230_180458		3		3
GW200105_162426	1 2 3 4 5 6	3 4 5 6	1 3 4	1 4
GW200112_155838	1 2 3 4 5	3	1 2 3 4 5	1 2 3 4 5
GW200115_042309	1 2 3 4 5	1 3 4 5	1 4	1 4
GW200128_022011	1 2 5		1 2 5	1 2 5
GW200129_065458	1 2 3 4 5 6	1 3 4 5	1 2 3 4 6	1 2 3 4 6
GW200202_154313	2 3 5		3	
GW200208_130117	1	2 4	1	
GW200208_222617	1 2 4 5 6	1 6	1 4 5	1 4 5 6
GW200209_085452				5
GW200210_092254	1 2 3 4 5 6	1	3 4 6	2 4
GW200216_220804		3		
GW200219_094415	1		1	1
GW200220_061928	1	6	1	1
GW200220_124850				
GW200224_222234	1 2 3 5 6	5	3 5	3 5
GW200225_060421	2 3 5 6	4	2 5	2 4 5
GW200302_015811	1 3 6	3	1	1 6
GW200306_093714	6	6		6
GW200308_173609	1 2 4 5 6	2 6	2 4 5 6	2 4 5 6
GW200311_115853	1 2 3 4 5 6		2 3 5	1 2 3 5
GW200316_215756	1 2 3 4 6	3 6	1 2 3 4	1 2 4
GW200322_091133	1 2 5 6	6	1 2 5	1 2 5

(b) EOB basis

Table 5.2.: Results for agreement of `IMRPhenomXPHM` and `SEOBNRv4PHM` PCs for GWTC-3 events using 20% thresholds and normalized `nocosmo` data. Event names visualized in green/ red belong to the lists good/ bad data identified in the previous chapter.



(a) Comparison in basis computed from `IMRPhenomXPHM` model



(b) Comparison in basis computed from `SEOBNRv4PHM` model

Figure 5.3.: Agreement for PCs from all events from GWTC-3. Each dot represents the JSD value for a PC and the 50% threshold is visualized as a red line.Each dot represents the JSD value for a PC and the 50% threshold is visualized as a red line.

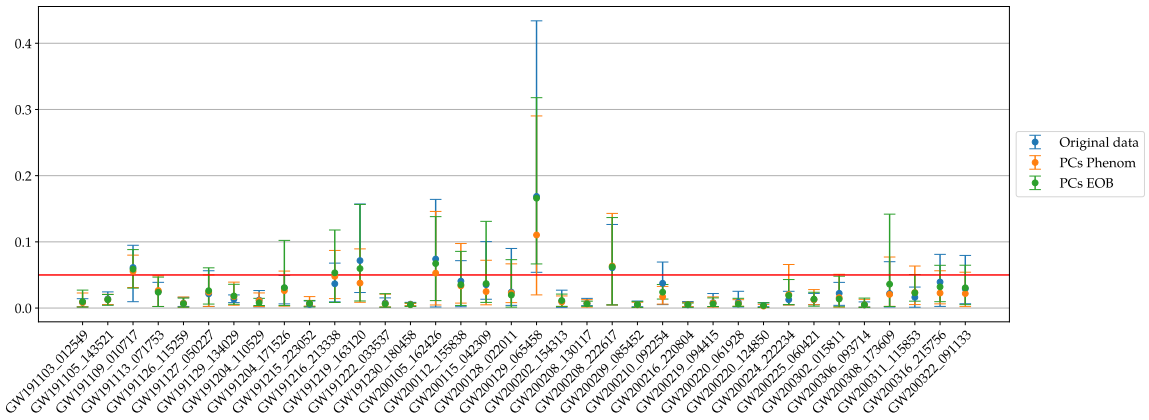
Eventlist	Parameter	JSD		JSD 2		Mean difference		Median difference	
		50%	20%	50%	20%	50%	20%	50%	20%
Bad data	PC 1	7/ 10 (70%)	9/ 10 (90%)	1/ 10 (10%)	3/ 10 (30%)	4/ 10 (40%)	7/ 10 (70%)	5/ 10 (50%)	8/ 10 (80%)
	PC 2	0/ 10 (0%)	9/ 10 (90%)	0/ 10 (0%)	3/ 10 (30%)	0/ 10 (0%)	8/ 10 (80%)	0/ 10 (0%)	8/ 10 (80%)
	PC 3	3/ 10 (30%)	7/ 10 (70%)	1/ 10 (10%)	5/ 10 (50%)	3/ 10 (30%)	5/ 10 (50%)	3/ 10 (30%)	5/ 10 (50%)
	PC 4	3/ 10 (30%)	8/ 10 (80%)	3/ 10 (30%)	3/ 10 (30%)	2/ 10 (20%)	7/ 10 (70%)	3/ 10 (30%)	7/ 10 (70%)
	PC 5	1/ 10 (10%)	6/ 10 (60%)	1/ 10 (10%)	3/ 10 (30%)	1/ 10 (10%)	4/ 10 (40%)	1/ 10 (10%)	4/ 10 (40%)
	PC 6	3/ 10 (30%)	7/ 10 (70%)	1/ 10 (10%)	1/ 10 (10%)	1/ 10 (10%)	4/ 10 (40%)	2/ 10 (20%)	5/ 10 (50%)
All data	PC 1	8/ 36 (22%)	20/ 36 (56%)	1/ 36 (3%)	6/ 36 (17%)	5/ 36 (14%)	13/ 36 (36%)	7/ 36 (19%)	17/ 36 (47%)
	PC 2	1/ 36 (3%)	20/ 36 (56%)	0/ 36 (0%)	5/ 36 (14%)	1/ 36 (3%)	13/ 36 (36%)	1/ 36 (3%)	14/ 36 (39%)
	PC 3	4/ 36 (11%)	21/ 36 (58%)	1/ 36 (3%)	12/ 36 (33%)	3/ 36 (8%)	11/ 36 (31%)	3/ 36 (8%)	13/ 36 (36%)
	PC 4	5/ 36 (14%)	18/ 36 (50%)	3/ 36 (8%)	7/ 36 (19%)	4/ 36 (11%)	17/ 36 (47%)	5/ 36 (14%)	17/ 36 (47%)
	PC 5	4/ 36 (11%)	20/ 36 (56%)	1/ 36 (3%)	8/ 36 (22%)	2/ 36 (6%)	11/ 36 (31%)	3/ 36 (8%)	12/ 36 (33%)
	PC 6	3/ 36 (8%)	15/ 36 (42%)	2/ 36 (6%)	7/ 36 (19%)	1/ 36 (3%)	7/ 36 (19%)	2/ 36 (6%)	10/ 36 (28%)
Good data	PC 1	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	1/ 8 (12%)
	PC 2	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	1/ 8 (12%)
	PC 3	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	3/ 8 (38%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	2/ 8 (25%)
	PC 4	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	1/ 8 (12%)
	PC 5	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)
	PC 6	0/ 8 (0%)	1/ 8 (12%)	1/ 8 (12%)	2/ 8 (25%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)

(a) Phenom basis

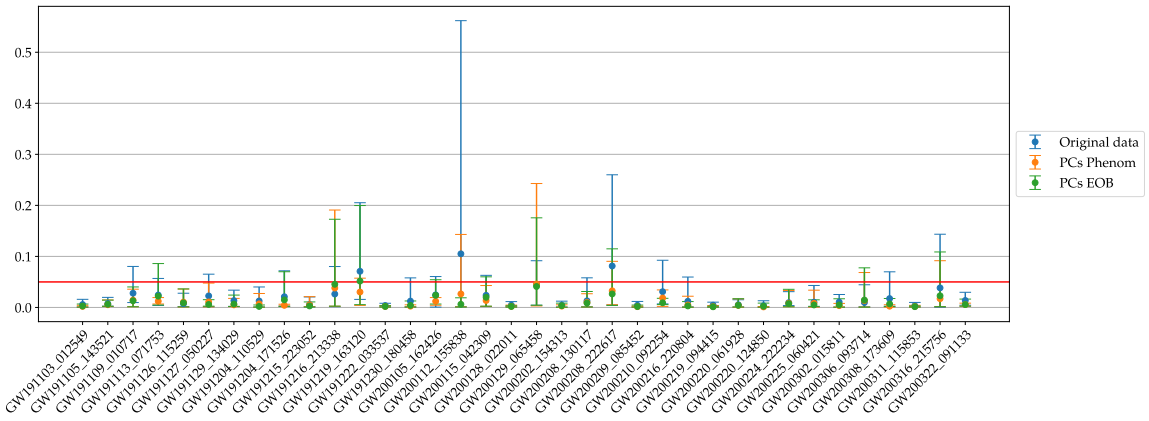
Eventlist	Parameter	JSD		JSD 2		Mean difference		Median difference	
		50%	20%	50%	20%	50%	20%	50%	20%
Bad data	PC 1	4/ 10 (40%)	8/ 10 (80%)	1/ 10 (10%)	4/ 10 (40%)	4/ 10 (40%)	5/ 10 (50%)	4/ 10 (40%)	6/ 10 (60%)
	PC 2	2/ 10 (20%)	10/ 10 (100%)	0/ 10 (0%)	3/ 10 (30%)	0/ 10 (0%)	8/ 10 (80%)	2/ 10 (20%)	8/ 10 (80%)
	PC 3	3/ 10 (30%)	7/ 10 (70%)	2/ 10 (20%)	5/ 10 (50%)	3/ 10 (30%)	6/ 10 (60%)	4/ 10 (40%)	5/ 10 (50%)
	PC 4	5/ 10 (50%)	9/ 10 (90%)	2/ 10 (20%)	3/ 10 (30%)	4/ 10 (40%)	8/ 10 (80%)	4/ 10 (40%)	8/ 10 (80%)
	PC 5	5/ 10 (50%)	9/ 10 (90%)	1/ 10 (10%)	3/ 10 (30%)	2/ 10 (20%)	6/ 10 (60%)	2/ 10 (20%)	7/ 10 (70%)
	PC 6	3/ 10 (30%)	9/ 10 (90%)	0/ 10 (0%)	5/ 10 (50%)	1/ 10 (10%)	5/ 10 (50%)	1/ 10 (10%)	6/ 10 (60%)
All data	PC 1	5/ 36 (14%)	19/ 36 (53%)	1/ 36 (3%)	6/ 36 (17%)	5/ 36 (14%)	13/ 36 (36%)	4/ 36 (11%)	14/ 36 (39%)
	PC 2	3/ 36 (8%)	24/ 36 (67%)	0/ 36 (0%)	5/ 36 (14%)	1/ 36 (3%)	14/ 36 (39%)	3/ 36 (8%)	14/ 36 (39%)
	PC 3	3/ 36 (8%)	20/ 36 (56%)	5/ 36 (14%)	15/ 36 (42%)	3/ 36 (8%)	12/ 36 (33%)	5/ 36 (14%)	12/ 36 (33%)
	PC 4	8/ 36 (22%)	20/ 36 (56%)	2/ 36 (6%)	8/ 36 (22%)	7/ 36 (19%)	16/ 36 (44%)	7/ 36 (19%)	17/ 36 (47%)
	PC 5	7/ 36 (19%)	20/ 36 (56%)	1/ 36 (3%)	7/ 36 (19%)	3/ 36 (8%)	12/ 36 (33%)	3/ 36 (8%)	14/ 36 (39%)
	PC 6	3/ 36 (8%)	19/ 36 (53%)	2/ 36 (6%)	10/ 36 (28%)	1/ 36 (3%)	8/ 36 (22%)	1/ 36 (3%)	10/ 36 (28%)
Good data	PC 1	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)
	PC 2	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	2/ 8 (25%)
	PC 3	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)
	PC 4	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)
	PC 5	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)
	PC 6	0/ 8 (0%)	1/ 8 (12%)	1/ 8 (12%)	2/ 8 (25%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)

(b) EOB basis

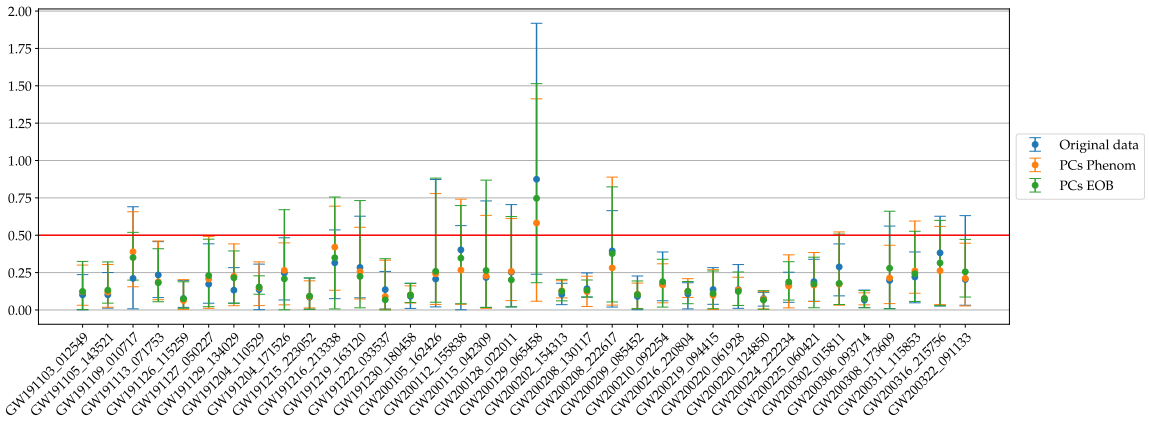
Table 5.3.: Event statistics for PCs of normalized GWTC-3 data. The numbers and percentages show how often each PC exceeds our thresholds.
All data provides an average value of how many events from GWTC-3 fail the chosen thresholds (≡ summary of tables 5.1, 5.2), which can then be compared to these respective numbers for our list of bad, good data (`nocosmo` data used).



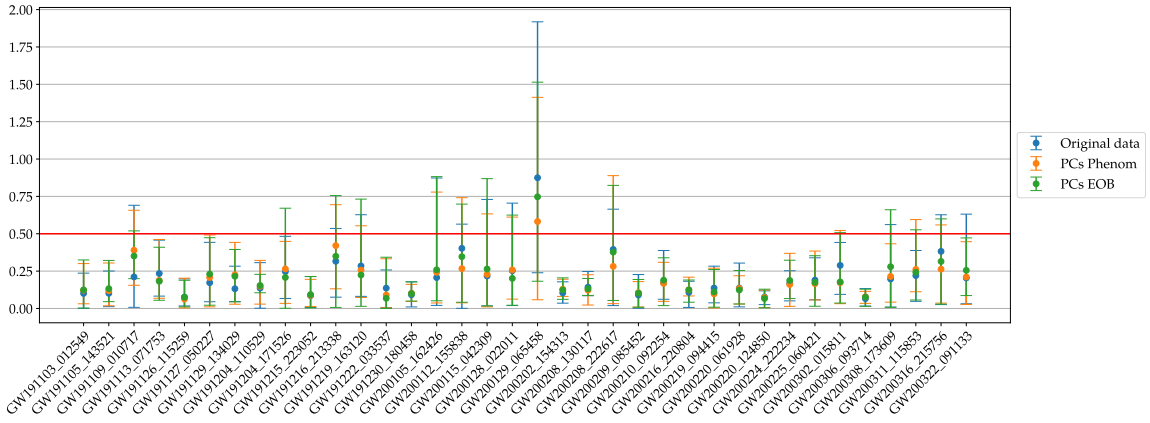
(a) JSD



(b) JSD 2

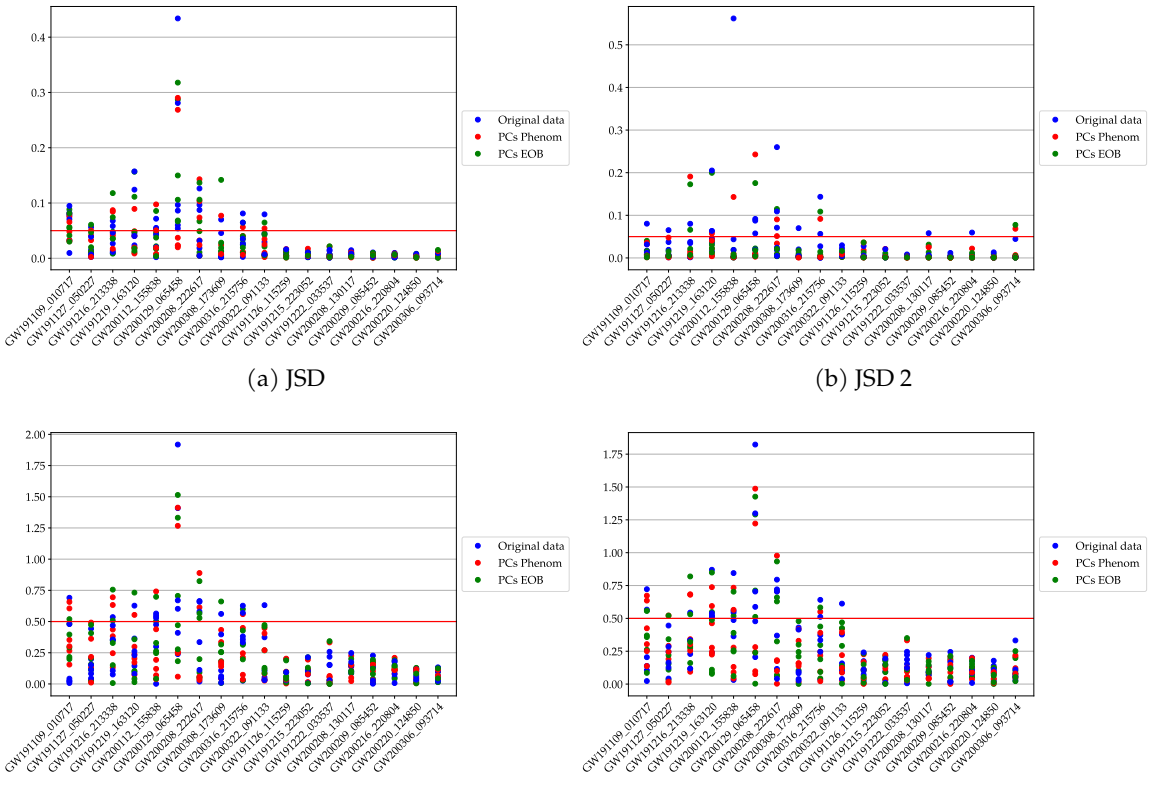


(c) Mean difference



(d) Median difference

Figure 5.4.: Comparison of agreement of parameters and PCs for all events from GWTC-3. Dots represent the average criterion value of all parameters or PCs for the event and the error bars visualize the maximum and minimum criterion value.



(a) JSD

(b) JSD 2

(c) Mean difference

(d) Median difference

Figure 5.5.: Comparison of agreement of parameters and PCs for good and bad data lists. Criterion values are shown for every parameter and PC from each basis (different bases are indicated by different colours) to enable a more detailed analysis. This was not done for all events because the resulting graphic is more confusing.

was the case for the original parameters. However, this pattern does not appear so clearly when directly comparing the agreement for original parameters and PCs for each event (like figure 5.4 does), in fact these plots reveal no systematic change. Similarly, while there are changes for individual events, there is no systematic change in the properties of good and bad data. As figure 5.5 shows for each criterion, bad agreement transfers from original parameters into PCs and good agreement persists as well. The only exception to this tendency is GW200105_162426 as it is the only event marked red while not being in bad data. However, the JSD is the only criterion to do so and among other reasons (see subsection 4.2.4 for the corresponding discussion), this is why it was manually excluded from bad data.

As of now, it looks like there are two inconsistent statements/ patterns and hence conclusions on how the agreement changes. Each one is extracted from different representations and when thinking about it in more detail, these representations are an important part of why this is the case. In plots like 5.4 that compare the agreement in original parameters and PCs, the error bars visualize the maximum and minimum agreement of each parameter/ PC and this means that even one high criterion value can cause the bars to stretch out. Looking back at the tables 5.1, 5.2 for each event, we can see that despite less red events, many are marked yellow, which means that precisely one PC exceeds the threshold. For the original parameters on the other hand, there are less events marked in yellow and instead more in red, which points to the following pattern: if there is a parameter in the original data exceeding the threshold for a criterion, then it is very likely that several others do that as well. This is not the case very often for the PCs, so we can see that the statements are indeed consistent: while the overall agreement does increase for the PCs, this change is not so significant for the individual properties of the data (for example whether they tend to have good/ bad agreement compared to the other events). An illustrative explanation of this is that the disagreement many of the original parameters exhibited is now concentrated, i.e. it appears in fewer components (similar to what happened in the figures 5.1, 5.2). Another reason for this decrease in the number of red cells is that three parameters in M, m_1, m_2 were excluded, which reduces the number of parameters that could potentially disagree. In fact, we saw that the differences for these parameters exceed the 50% threshold for a significant number of events, so instead of relying on statements based on thresholds, direct comparisons like figure 5.5 are better suited in this context. The conclusion of this discussion is that the agreement is better for the PCs than for the original parameters, but not as much as the tables 5.1, 5.2 make it look like.

One might now be tempted to compare the results obtained in the bases computed from `IMRPhenomXPHM` and `SEOBNRv4PHM` posteriors. This is indeed possible, but one has to be very careful with what meaning such a comparison has. For individual events for example, such a comparison does not make much sense. This is because the posteriors might (and do) have different correlations between the parameters, so the correlation matrix and subsequently the principal axes will be different for each event. Comparing the agreement in Phenom and EOB basis then corresponds to comparing the agreement along different axes and this is no statement of interest (at least in this work). Moreover, we should emphasize that such a comparison is not even necessary since the individual results from each basis are already statements on the waveform agreement. On the other hand, comparing how a certain PC agrees averaged over all events should make more sense because in principle, correlations

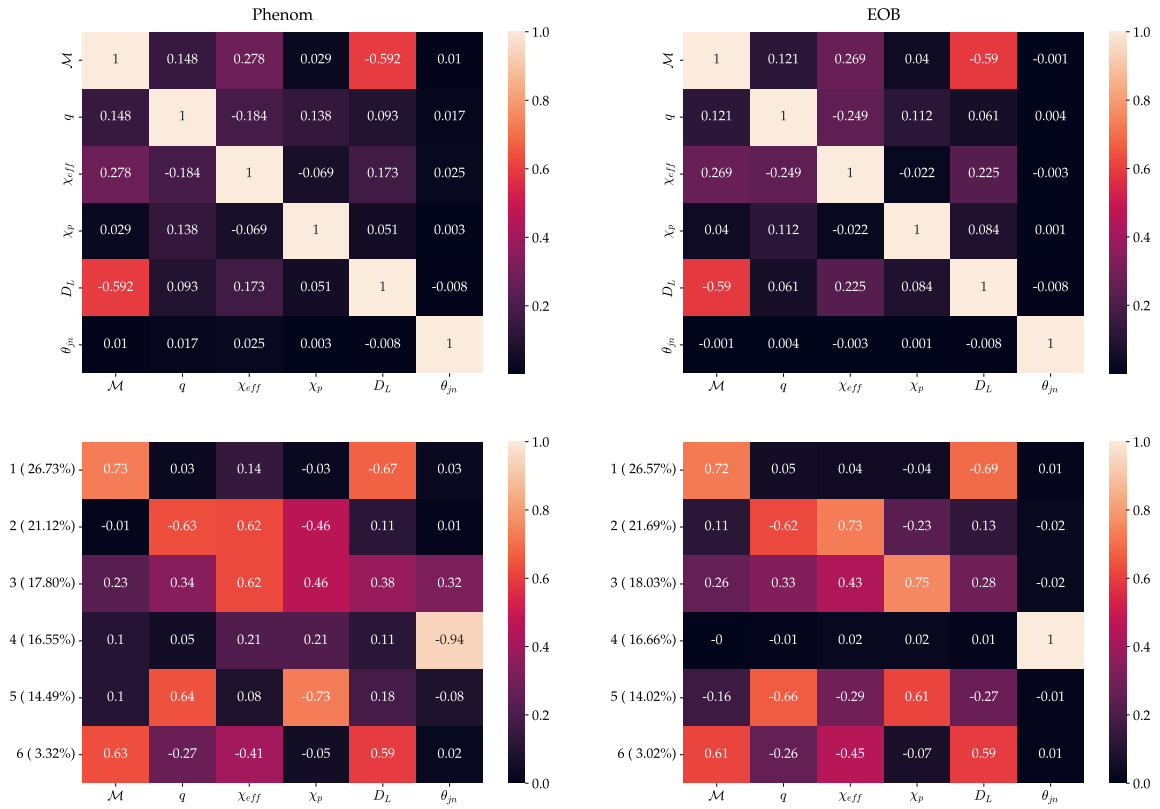
are similar for the waveform models and thus, differences might average out. Still, the meaning of such a statement taking into account all events may not have the meaning one would expect. PC k in table 5.3 having good agreement tells us something about the axis along which the k-th most variance is captured, so this is not a fixed axis for each event. Such a comparison does not reveal clear, systematic differences, so there does not seem to be a model exhibiting differences between the models better than the other.

5.2.2. Average Basis

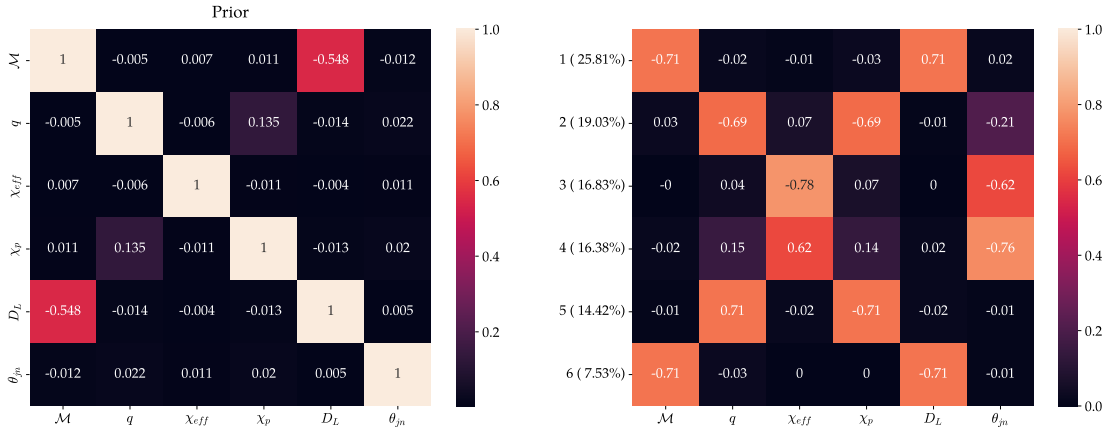
During the last subsection we encountered the fact that statements on the k-th PC of all events only have a very abstract meaning, they quantify the average agreement along the axis with the k-th highest variance. Discussing this is still valuable since it is a very individual analysis of each event and the results are also meaningful and interesting. Still, being able to examine the agreement along a specific set of fixed axes would allow us to asses if there are axes like in figure 5.2 and statements would have a very illustrative interpretation because each principal axis corresponds to a certain linear combination of parameters. Hence, it is worthwhile to look for a way to do that. Those axes should still be somehow related to the data, so we cannot simply take an arbitrary set. The first idea might be to take the average of each principal axis, but there are some issues with that. First of all, they are different for each event. While this is also the case for the majority of other potentially suited quantities, the principal axes have the disadvantage that it is hard to think of a method to average over them in a way that makes sense. For example, the principal axis k taken from two events could represent the same direction, but have a different sign (so their components would cancel out despite similar correlations). Or, it could be the case that identical principal axes account for different amounts of variances, so their ranking among all principal axes differs (in this case, it would make sense to still take their average and thus, we would have to tackle the complicated task of recognizing similar directions). While finding such a method may still be possible, we did not try to do so.

That does not mean we did not pursue this approach at all, we just found another method that should be suited equally well to achieve the goal formulated in the previous paragraph. To understand the idea behind this alternative method, we remind ourselves that principal axes are computed from a covariance or correlation matrix and correlations are something which we can certainly average over. Hence, the approach chosen in this subsection is to take the correlation matrices of all events, compute their average (i.e. the average of every component) and then diagonalize it to obtain a set of principal axes (results of this process are shown in figure 5.6). These axes still have a meaning for each event (though it might be limited), but also allow to examine the agreement along one fixed set of coordinate axes to assess how well certain parameter combinations are measurable.

The first things to notice when looking at the correlation matrices are that the waveform models broadly agree in terms of correlations present in their posteriors and that most of them are not prior-induced. The latter is important because it ensures that the systematics we may observe during the following analyses are either produced by the use of different waveforms or by the data (in any case, they are interesting for us). Most correlations present



(a) Results for posteriors



(b) Result for prior

Figure 5.6.: Average correlation matrix and corresponding eigenvectors for GWTC-3. There are results for each of the waveform models again and for the prior as well (waveform-independent, therefore only one of each matrices). The bottom matrices in (a) and right hand matrix in (b) give in each row an eigenvector/ principal axis of the corresponding correlation matrix, which are ordered according to the amount of variance they capture (given as percentages on the left hand side next to them).

are not surprising. Masses and particularly the chirp mass $\mathcal{M}$ determine the amplitude of a signal as well as the distance D_L of the source, so for a signal of given strength the value of one of them influences the value of the other. Similarly, it is not surprising that $\chi_{\text{eff}}, \chi_p$ show correlations with $\mathcal{M}, q$ since they are mass-weighted spin parameters (q even directly appears in χ_p and χ_{eff} can be rewritten to also contain it). Besides a smaller correlation between $\chi_{\text{eff}}, \chi_p$ and D_L induced by that, no other significant correlations appear. Because some of these reflect very fundamental properties of the parameters and GW signals, it also makes sense that the prior has non-zero values for $\text{corr}(\mathcal{M}, D_L)$ and $\text{corr}(q, \chi_p)$.

Turning to the eigenvectors of these matrices, one can see that the amount of variance each one of them accounts for and the contribution from each parameter (or equivalently, the directions they point along in the parameter space) are all very similar for both waveform models. Opposing signs like for eigenvectors 4 and 5 do not really matter in this context, in fact, we could switch the signs of both and get exactly the same results for eigenvalues and other quantities. This tells us that the waveforms infer (on average) posteriors with similar properties as measured by correlations between parameters and their distribution along specific axes. Since the correlation matrices already are similar, this is what we would expect and it is good to have such a confirmation.

The components of each eigenvector also have implications on the data and most of them correspond to ones already found from looking at the correlation matrices. The first eigenvector tells us that posteriors often have shapes similar to the distribution in figure 5.2 when viewed in the $\mathcal{M}$-D_L-plane because these parameters are highly correlated. It makes sense that this specific axis captures the most variance/ information of the data, after all $\mathcal{M}, D_L$ are among the best measurable parameters, which simply means one can extract a lot of information about them from the data (they have a high influence on where in the parameter space the posterior is non-zero). Principal axes two and five are related, they reflect the correlation between q and the spin parameters $\chi_{\text{eff}}, \chi_p$ (in number two; the components do vary quite a bit between the waveforms because the EOB model has shows a stronger correlation between q, χ_{eff} than the Phenom one), but that other quantities like the individual spins also contribute to them (appears in number five, which is something like the contribution to χ_p "orthogonal" to q, i.e. not determined by it). Eigenvector number three is harder to interpret since it has clearly non-zero components for almost all parameters. It also shows differences between the models (now Phenom has a higher component for χ_{eff} while EOB has a higher one for χ_p, which is the reverse pattern from eigenvector two). The fourth eigenvector is very clearly dominated by θ_{jn}, this parameter basically has a PC for itself because almost no correlations with any parameter are present (this is more apparent for the EOB model, where it also does not contribute to any other eigenvector; this is different for the Phenom model). The sixth eigenvector captures only a very small amount of variance and essentially collects all remaining correlations. The corresponding PC could be regarded as unimportant due to its small variance, the variable represented by this eigenvector is very likely not an independent variable. For almost degenerate variables like x, y in figure 5.2, the second PC is determined by their difference and thus only non-zero due to sampling deviations, i.e. the probabilistic nature of the distributions. This is not necessarily the case here (3% of the total variance just from sampling deviations would be very high, so part of the contribution comes from the data), but it is most likely not a fully independent variable. This also means that the "true" dimensionality of the six-dimensional set of parameters examined in this

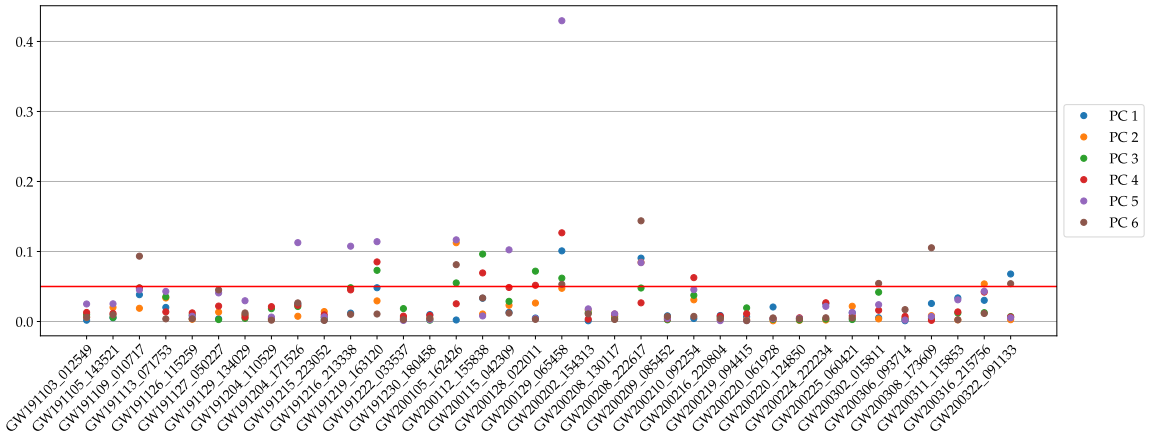
subsection seems to be five rather than six (degeneracies between the parameters lead to reduced dimensionality).

Another interesting pattern is that the eigenvectors for the EOB model look "cleaner" compared to the Phenom ones, i.e it has more dominant components. For instance, θ_{jn} really has a separate PC and neither do other parameters contribute to it nor does θ_{jn} contribute to other eigenvectors. Admittedly, though, there is no reference of how the "right" correlations look like, so this statement is not very certain (it could very well be the case that the small perturbations present in the Phenom model are more accurate). These doubts are supported by the fact that the Phenom model has a much higher sample size for the majority of events, so in principle, the results should be more reliable (especially since EOB has exceptionally small sample size sometimes, which leads to significant errors induced on the components of the eigenvectors). Despite these potential uncertainties of the axes, results on the agreement along them should not be affected much (which we ensured in chapter 4 already). As this is what we are really interested in, we also do not need a concluding answer on the "right" axes.

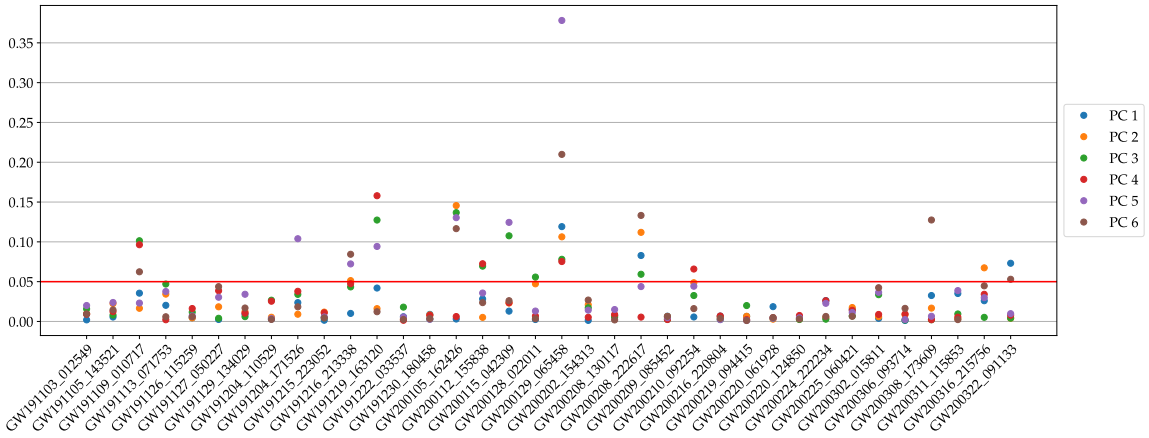
After a long preparation, we can now start looking into waveform systematics in the average PC basis. When analysing the events in these new bases, some differences to results from the previous subsections are expected. That is because the explicit correlation matrices do differ between the events, sometimes significantly (for example it is not unusual to have values of -0.9 for $\mathrm{corr}(\mathcal{M}, D_L)$, but values around 0 do appear as well). These fluctuations do not seem to follow any pattern, at least not related to waveform agreement (in particular not related to good and bad data), so we did not investigate them any further. What this tells us though is that the average principal axes will also differ from the regular ones for some events, so the agreement could in principle differ along these axes as well. To investigate if that is the case, we will look at the same representation and visualizations as before, i.e. figures and tables encoding the agreement for all events (figure 5.7, tables 5.4, 5.5) or the agreement for each PC (table 5.6). Additionally, it makes sense to compare the agreement for regular and average PCs (figure 5.8) as well as the agreement for original parameters and average PCs (figure 5.9). The most important features from each of these results will be highlighted now. One thing to keep in mind when assessing them is that, similarly to the previous subsection, it does not make much sense to compare statements on PC k from Phenom and EOB model as they already tell us something about the waveform agreement along a certain axis in the parameter space (so a comparison is not even necessary).

We will start by comparing the agreement for all events. Keeping in mind the results for the regular PCs from the previous subsection, we can see that not too much has changed. The most important takeaway is that patterns regarding agreement do not change significantly, for example bad data still has above average disagreement and good data below average. The overall agreement is slightly worse than for the regular PCs (as figure 5.8 or table 5.6 show), but still slightly better than for the original parameters (as figure 5.9 shows).

That a certain change from the results of the previous subsection occurs was expected. It is nonetheless interesting that this change is for the worse because that could mean we found axes like in figures 5.1, 5.2, which reveal more significant differences than the coordinates

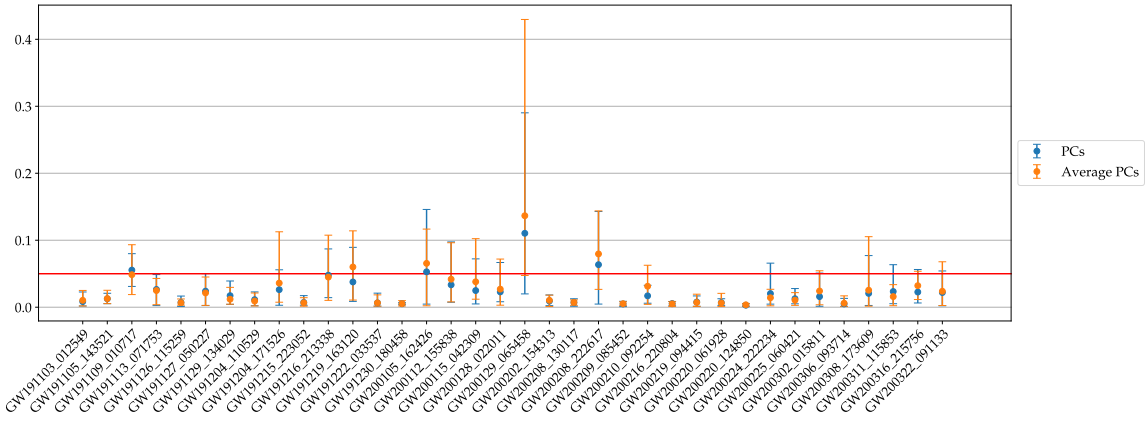


(a) Phenom

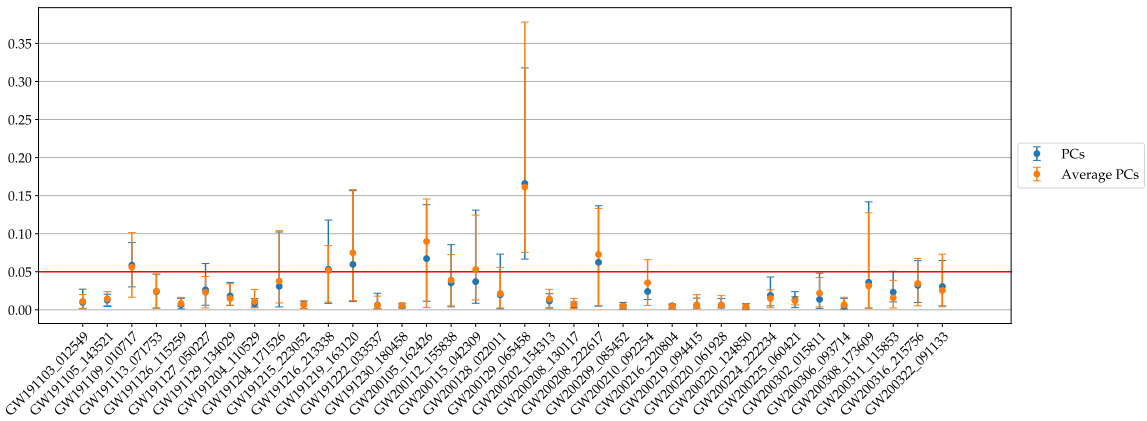


(b) EOB

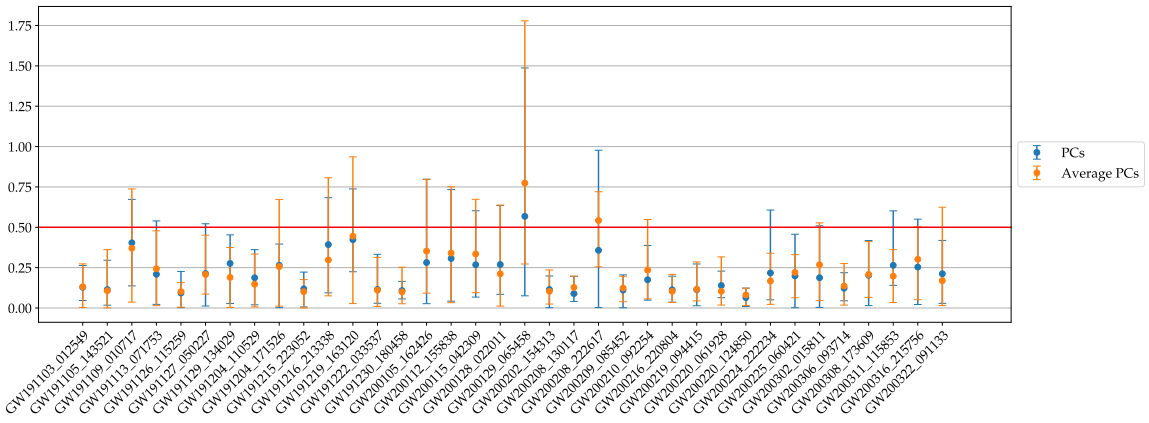
Figure 5.7.: Agreement in average PC basis for all events from GWTC-3, the corresponding basis vectors are given in figure 5.6. Each dot represents the JSD value for a PC and the 50% threshold is visualized as a red line.



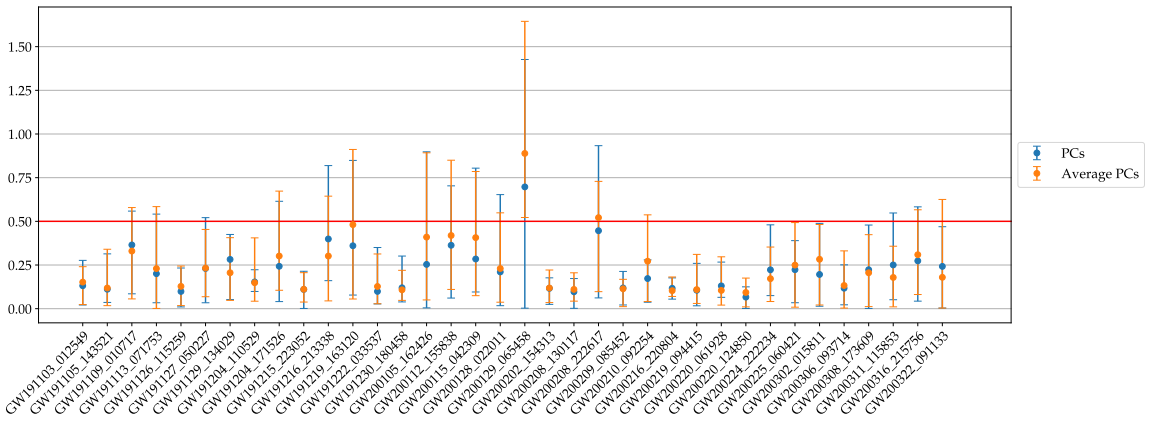
(a) JSD in Phenom basis



(b) JSD in EOB basis

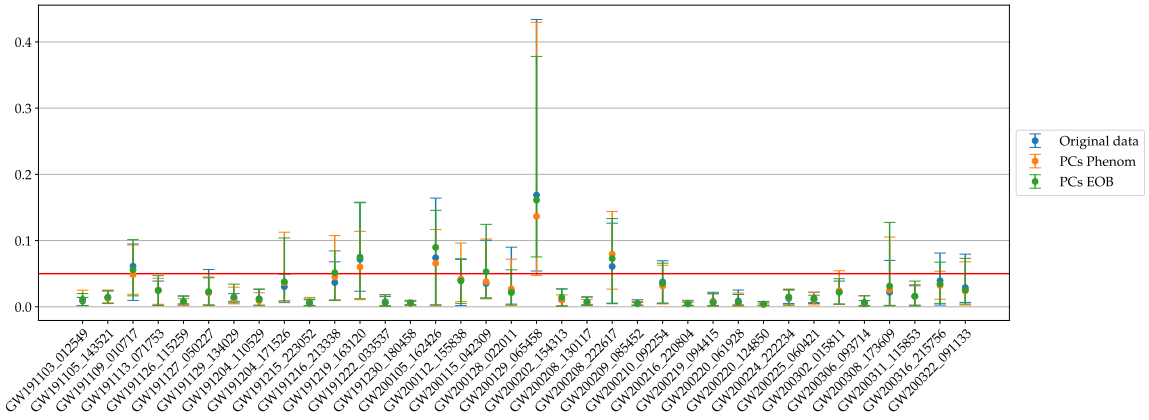


(c) Median difference in Phenom basis

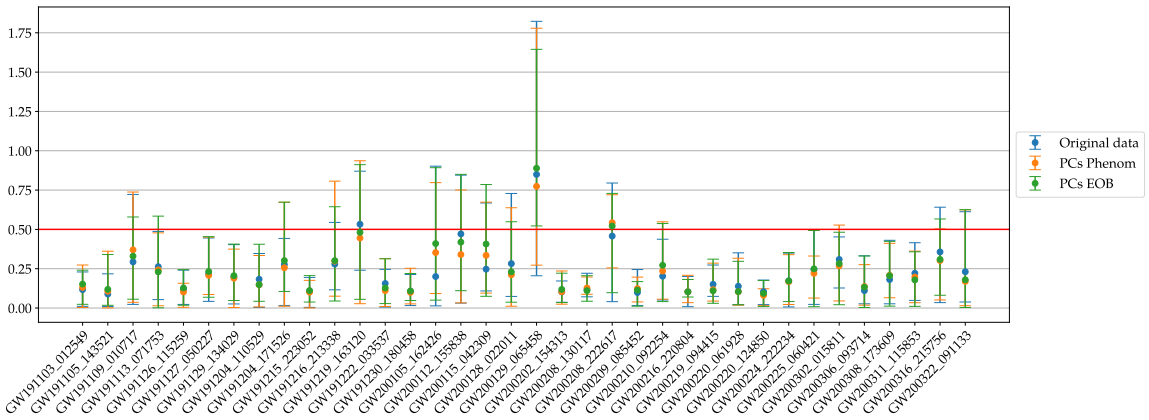


(d) Median difference in EOB basis

Figure 5.8.: Comparison of agreement of PCs and average principal axes in different bases for all events from GWTC-3. Dots represent the average criterion value of all parameters or PCs for the event and the error bars visualize the maximum and minimum criterion value.
This type of plot is provided because a comparison of criterion values for regular PCs and from the average basis in general does not make sense (and one might get confused really easily due to too many points), but comparing the range of agreement does. We only give results for JSD and median criterion as they produce the most relevant results and are used the most.



(a) JSD



(b) Median difference

Figure 5.9.: Comparison of agreement of original parameters and average principal axes for all events from GWTC-3. Dots represent the average criterion value of all parameters or PCs for the event and the error bars visualize the maximum and minimum criterion value. We only give results for JSD and median criterion as they produce the most relevant results and are used the most.

Event	JSD	JSD 2	Mean difference	Median difference
GW191103_012549				
GW191105_143521				
GW191109_010717	6		6	6
GW191113_071753				
GW191126_115259				
GW191127_050227				
GW191129_134029				
GW191204_110529				
GW191204_171526	5		5	5
GW191215_223052				
GW191216_213338	5	4	5	5
GW191219_163120	3 4 5	4	5	1 3 5
GW191222_033537				
GW191230_180458				
GW200105_162426	2 3 5 6		3 5	5
GW200112_155838	3 4	4	3	3 4
GW200115_042309	5		5	5
GW200128_022011	3 4		3	3
GW200129_065458	1 3 4 5 6	4	1 4 5 6	1 4 5 6
GW200202_154313				
GW200208_130117				
GW200208_222617	1 2 5 6	2 6	1 5	1 2 5 6
GW200209_085452				
GW200210_092254	4			5
GW200216_220804				
GW200219_094415				
GW200220_061928				
GW200220_124850				
GW200224_222234				
GW200225_060421				
GW200302_015811	6			6
GW200306_093714		6		
GW200308_173609	6			
GW200311_115853				
GW200316_215756	2	4		5
GW200322_091133	1 6		1	1

(a) Phenom basis

Event	JSD	JSD 2	Mean difference	Median difference
GW191103_012549				
GW191105_143521				
GW191109_010717	3 4 6	4	6	3 6
GW191113_071753		3		3
GW191126_115259				
GW191127_050227		4		
GW191129_134029				
GW191204_110529				
GW191204_171526	5	4	5	5
GW191215_223052				
GW191216_213338	2 5 6	4 6	5	5
GW191219_163120	3 4 5	1 4	3 5	3 4 5
GW191222_033537				
GW191230_180458		4		
GW200105_162426	2 3 5 6		3 5	3 5
GW200112_155838	3 4	4	3 4	3 4
GW200115_042309	3 5	4	3 5	3 5
GW200128_022011	3		2 3	3
GW200129_065458	1 2 3 4 5 6	4	1 2 4 5 6	1 2 3 4 5 6
GW200202_154313				
GW200208_130117		4		
GW200208_222617	1 2 3 6	2 6	1 2 3	1 2 3 6
GW200209_085452				
GW200210_092254	4	4		5
GW200216_220804		4		
GW200219_094415				
GW200220_061928				
GW200220_124850				
GW200224_222234				
GW200225_060421				
GW200302_015811				
GW200306_093714		6		
GW200308_173609	6	6	6	
GW200311_115853				
GW200316_215756	2	4	2	2
GW200322_091133	1 6		1	1

(b) EOB basis

Table 5.4.: Results for agreement of `IMRPhenomXPHM` and `SEOBNRv4PHM` in average PC basis for GWTC-3 events using 50% thresholds and normalized `nocosmo` data. Event names visualized in green/ red belong to the lists good/ bad data identified in the previous chapter.

Event	JSD	JSD 2	Mean difference	Median difference
GW191103_012549	4 5		5	5
GW191105_143521	2 4 5		5	5
GW191109_010717	1 2 3 4 5 6	3 4	1 2 3 5 6	1 2 3 5 6
GW191113_071753	1 2 3 4 5	2 3 4 5	1 3 4	1 2 3 4
GW191126_115259	4			
GW191127_050227	2 4 5 6	4 5 6	2 5 6	2 5 6
GW191129_134029	5 6	4	5	4 5
GW191204_110529	3 4	4	3	3 4
GW191204_171526	1 3 4 5 6	4	1 5 6	1 5 6
GW191215_223052	2	4		
GW191216_213338	1 2 3 4 5 6	2 3 4	2 3 4 5	2 3 4 5
GW191219_163120	1 2 3 4 5 6	1 2 3 4 5 6	1 3 5	1 2 3 5
GW191222_033537	3		3	3
GW191230_180458		4		4
GW200105_162426	2 3 4 5 6	2 3 5 6	2 3 5 6	2 3 5 6
GW200112_155838	1 2 3 4 6	4	1 3 4	1 2 3 4
GW200115_042309	1 2 3 4 5 6	2 3 4 5	2 3 4 5	2 3 4 5
GW200128_022011	2 3 4		2 3 4	2 3
GW200129_065458	1 2 3 4 5 6	1 3 4 5	1 2 4 5 6	1 2 3 4 5 6
GW200202_154313	2 3 5 6		5	5
GW200208_130117	4 5	4	4 5	
GW200208_222617	1 2 3 4 5 6	1 2 3 6	1 2 3 4 5 6	1 2 3 4 5 6
GW200209_085452				
GW200210_092254	2 3 4 5	1 3 4 6	3 5	3 5
GW200216_220804		4		1
GW200219_094415	3 4		3	3
GW200220_061928	1	6	1	1
GW200220_124850				
GW200224_222234	1 4 5	4	4 5	4 5
GW200225_060421	1 2 5		1 2	1 2 4
GW200302_015811	3 4 5 6	6	3 5 6	3 5 6
GW200306_093714	6	4 6		4 6
GW200308_173609	1 6	1 6	1 6	1 2 6
GW200311_115853	1 3 4 5		1 4 5	1 4 5
GW200316_215756	1 2 3 4 5 6	2 4	1 2 3 4 5	1 2 3 5
GW200322_091133	1 6	6	1	1

(a) Phenom basis

Event	JSD	JSD 2	Mean difference	Median difference
GW191103_012549	3 5	4	3 5	3 5
GW191105_143521	2 4 5 6	4 6	5	5
GW191109_010717	1 2 3 4 5 6	3 4	1 3 5 6	1 3 4 5 6
GW191113_071753	1 2 3 5	2 3 5	1 3	1 3
GW191126_115259	3 4	4		4
GW191127_050227	2 4 5 6	4 6	2 5 6	2 4 6
GW191129_134029	2 4 5 6	4	5	4 5
GW191204_110529	3 4	4	3	3 4
GW191204_171526	1 3 4 5 6	4	3 5	1 3 4 5 6
GW191215_223052	2 4	4	2	2
GW191216_213338	1 2 3 4 5 6	2 4 6	2 3 4 5 6	2 3 4 5
GW191219_163120	1 2 3 4 5 6	1 2 3 4 5 6	1 3 5	1 3 4 5
GW191222_033537	3		3	3
GW191230_180458		4		4
GW200105_162426	2 3 5 6	2 3 5 6	2 3 5 6	2 3 5 6
GW200112_155838	1 3 4 5 6	4	1 3 4 5	1 3 4 5 6
GW200115_042309	1 2 3 4 5 6	2 4	3 4 5 6	2 3 4 5 6
GW200128_022011	2 3 5		2 3 5	2 3 5
GW200129_065458	1 2 3 4 5 6	1 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6
GW200202_154313	2 3 5 6	6	5	5
GW200208_130117	5	4	5	5
GW200208_222617	1 2 3 5 6	1 2 6	1 2 3 5 6	1 2 3 5 6
GW200209_085452				
GW200210_092254	2 3 4 5 6	1 3 4	3 5 6	3 4 5 6
GW200216_220804		4		
GW200219_094415	3		3	3
GW200220_061928	1	4 6	1	1
GW200220_124850		4		
GW200224_222234	1 4 5	4	4 5	4 5
GW200225_060421	1 2 4 5	4	1 2	1 2 4 5
GW200302_015811	3 5 6	4	3 5 6	3 4 5 6
GW200306_093714	6	4 6		4 6
GW200308_173609	1 2 6	1 6	1 2 6	1 2 6
GW200311_115853	1 5	5	1 5	1 5
GW200316_215756	1 2 4 5 6	2 4 6	1 2 4 5 6	1 2 4 5
GW200322_091133	1 6	6	1 6	1

(b) EOB basis

Table 5.5.: Results for agreement of `IMRPhenomXPHM` and `SEOBNRv4PHM` in average PC basis for GWTC-3 events using 20% thresholds and normalized `nocosmo` data. Event names visualized in green/ red belong to the lists good/ bad data identified in the previous chapter.

Eventlist	Parameter	JSD		JSD 2		Mean difference		Median difference	
		50%	20%	50%	20%	50%	20%	50%	20%
Bad data	PC 1	3/ 10 (30%)	8/ 10 (80%)	0/ 10 (0%)	4/ 10 (40%)	3/ 10 (30%)	8/ 10 (80%)	4/ 10 (40%)	8/ 10 (80%)
	PC 2	3/ 10 (30%)	8/ 10 (80%)	0/ 10 (0%)	2/ 10 (20%)	0/ 10 (0%)	6/ 10 (60%)	2/ 10 (20%)	6/ 10 (60%)
	PC 3	2/ 10 (20%)	8/ 10 (80%)	2/ 10 (20%)	5/ 10 (50%)	2/ 10 (20%)	5/ 10 (50%)	1/ 10 (10%)	5/ 10 (50%)
	PC 4	1/ 10 (10%)	7/ 10 (70%)	3/ 10 (30%)	7/ 10 (70%)	1/ 10 (10%)	5/ 10 (50%)	1/ 10 (10%)	5/ 10 (50%)
	PC 5	6/ 10 (60%)	7/ 10 (70%)	0/ 10 (0%)	3/ 10 (30%)	6/ 10 (60%)	7/ 10 (70%)	6/ 10 (60%)	7/ 10 (70%)
	PC 6	5/ 10 (50%)	10/ 10 (100%)	1/ 10 (10%)	6/ 10 (60%)	2/ 10 (20%)	5/ 10 (50%)	3/ 10 (30%)	5/ 10 (50%)
All data	PC 1	3/ 36 (8%)	16/ 36 (44%)	0/ 36 (0%)	5/ 36 (14%)	3/ 36 (8%)	13/ 36 (36%)	4/ 36 (11%)	14/ 36 (39%)
	PC 2	3/ 36 (8%)	17/ 36 (47%)	1/ 36 (3%)	7/ 36 (19%)	0/ 36 (0%)	10/ 36 (28%)	1/ 36 (3%)	14/ 36 (39%)
	PC 3	5/ 36 (14%)	19/ 36 (53%)	0/ 36 (0%)	9/ 36 (25%)	3/ 36 (8%)	15/ 36 (42%)	3/ 36 (8%)	16/ 36 (44%)
	PC 4	5/ 36 (14%)	23/ 36 (64%)	5/ 36 (14%)	19/ 36 (53%)	1/ 36 (3%)	11/ 36 (31%)	2/ 36 (6%)	13/ 36 (36%)
	PC 5	7/ 36 (19%)	21/ 36 (58%)	0/ 36 (0%)	6/ 36 (17%)	7/ 36 (19%)	19/ 36 (53%)	9/ 36 (25%)	18/ 36 (50%)
	PC 6	7/ 36 (19%)	17/ 36 (47%)	2/ 36 (6%)	10/ 36 (28%)	2/ 36 (6%)	8/ 36 (22%)	4/ 36 (11%)	9/ 36 (25%)
Good data	PC 1	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)
	PC 2	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	1/ 8 (12%)
	PC 3	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	1/ 8 (12%)
	PC 4	0/ 8 (0%)	2/ 8 (25%)	1/ 8 (12%)	6/ 8 (75%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	2/ 8 (25%)
	PC 5	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)
	PC 6	0/ 8 (0%)	1/ 8 (12%)	1/ 8 (12%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)

(a) Phenom basis

Eventlist	Parameter	JSD		JSD 2		Mean difference		Median difference	
		50%	20%	50%	20%	50%	20%	50%	20%
Bad data	PC 1	3/ 10 (30%)	8/ 10 (80%)	0/ 10 (0%)	4/ 10 (40%)	3/ 10 (30%)	8/ 10 (80%)	4/ 10 (40%)	8/ 10 (80%)
	PC 2	2/ 10 (20%)	8/ 10 (80%)	1/ 10 (10%)	4/ 10 (40%)	1/ 10 (10%)	6/ 10 (60%)	2/ 10 (20%)	7/ 10 (70%)
	PC 3	6/ 10 (60%)	7/ 10 (70%)	0/ 10 (0%)	3/ 10 (30%)	3/ 10 (30%)	7/ 10 (70%)	5/ 10 (50%)	7/ 10 (70%)
	PC 4	4/ 10 (40%)	7/ 10 (70%)	6/ 10 (60%)	7/ 10 (70%)	2/ 10 (20%)	4/ 10 (40%)	3/ 10 (30%)	6/ 10 (60%)
	PC 5	3/ 10 (30%)	9/ 10 (90%)	0/ 10 (0%)	1/ 10 (10%)	3/ 10 (30%)	8/ 10 (80%)	3/ 10 (30%)	9/ 10 (90%)
	PC 6	6/ 10 (60%)	10/ 10 (100%)	3/ 10 (30%)	8/ 10 (80%)	3/ 10 (30%)	8/ 10 (80%)	3/ 10 (30%)	6/ 10 (60%)
All data	PC 1	3/ 36 (8%)	16/ 36 (44%)	1/ 36 (3%)	5/ 36 (14%)	3/ 36 (8%)	12/ 36 (33%)	3/ 36 (8%)	13/ 36 (36%)
	PC 2	5/ 36 (14%)	18/ 36 (50%)	1/ 36 (3%)	7/ 36 (19%)	4/ 36 (11%)	10/ 36 (28%)	3/ 36 (8%)	11/ 36 (31%)
	PC 3	8/ 36 (22%)	19/ 36 (53%)	1/ 36 (3%)	6/ 36 (17%)	6/ 36 (17%)	17/ 36 (47%)	9/ 36 (25%)	17/ 36 (47%)
	PC 4	5/ 36 (14%)	17/ 36 (47%)	13/ 36 (36%)	25/ 36 (69%)	2/ 36 (6%)	6/ 36 (17%)	3/ 36 (8%)	18/ 36 (50%)
	PC 5	6/ 36 (17%)	23/ 36 (64%)	0/ 36 (0%)	5/ 36 (14%)	6/ 36 (17%)	21/ 36 (58%)	7/ 36 (19%)	21/ 36 (58%)
	PC 6	7/ 36 (19%)	19/ 36 (53%)	4/ 36 (11%)	13/ 36 (36%)	3/ 36 (8%)	12/ 36 (33%)	3/ 36 (8%)	12/ 36 (33%)
Good data	PC 1	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)
	PC 2	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	1/ 8 (12%)
	PC 3	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)
	PC 4	0/ 8 (0%)	2/ 8 (25%)	0/ 8 (0%)	4/ 8 (50%)	0/ 8 (0%)	1/ 8 (12%)	0/ 8 (0%)	3/ 8 (38%)
	PC 5	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)
	PC 6	0/ 8 (0%)	1/ 8 (12%)	1/ 8 (12%)	1/ 8 (12%)	0/ 8 (0%)	0/ 8 (0%)	0/ 8 (0%)	1/ 8 (12%)

(b) EOB basis

Table 5.6.: Event statistics for average PCs of normalized GWTC-3 data (see figure 5.6 for the axes which served as a basis). The numbers and percentages show how often each parameter exceeds the respective threshold of the column.
All data provides an average value of how many events from GWTC-3 fail the chosen thresholds (≡ summary of tables 5.4, 5.5), which can then be compared to these respective numbers for our list of bad, good data (`nocosmo` data used).

used before. Another scenario could be that one parameter with overall bad agreement contributes to multiple principal axes which would cause bad agreement in multiple PCs. From table 5.6 we can see which principal axes are responsible for bad agreement: the fifth one from the Phenom model and third one from the EOB model. Looking back at figure 5.6, we notice that both have contributions from χ_p. The fifth Phenom axis represents the direction "orthogonal" to q, i.e. not correlated with it and thus spin-dominated. It is very reasonable that we observe disagreement along this axis, spin components are known to be measured not very precisely. The third EOB axis on the other hand has contributions from $\chi_p, \chi_{\text{eff}}, q, D_L, \mathcal{M}$, so an interpretation of the direction is not so easy. We do see though, that the biggest contribution to the principal axis comes from χ_p, which means it now seems to be the main source for disagreement, in contrast to what was found in chapter 4. Part of that is the exclusion of M, m_1 as parameters with the highest disagreement, but the values in table 5.6 are already enough to support this claim as they are higher than the ones for each individual parameter in table 4.5. Regardless of interpretations of these axis, this shows that the disagreement along these axes is not caused by a single parameter. Instead, different parameters disagree for different events and this specific combination of them reveals axes where more events show disagreement along compared to the individual parameters. This is not only reflected in the rows for all data, but also the ones for bad data, where the same PCs have the highest disagreement. Thus, they seem to be a major source of why events are in bad data, not just a systematic which shows up independently from the overall agreement.

After this detailed discussion of two PCs, we also want to take a look at the agreement for the other principal axes. The first one is very similar for both models and so are the results. They indicate very good agreement, just like we expect from the results for its dominant components $\mathcal{M}, D_L$. The results for PC 2 may be a bit surprising because it shows exceptionally good agreement for both models, despite having non-negligible contributions from χ_p (especially for the Phenom model). This seems to be contradicting what we said in the previous paragraph, but can be explained by the fact that it is the part of χ_p proportional to q. Since q can be measured reasonably well, this part should be measurable more accurately (the contribution from q to PC 3 is much smaller than the contribution here, so the part of χ_p "parallel" to q seems to be measurable well after all). Here we can also see that more events exceed the JSD threshold for the EOB model, which is very likely due to the contribution of χ_{eff} being higher (which has a significant amount of events with high differences marked by the JSD). PC 4 probably has the least surprising results since it is dominated by θ_{jn} as a parameter with almost no correlations with other parameters and thus it is also not surprising that the behaviour of very high criterion values for JSD 2 transfers from the previous chapter. For both models, the agreement along the sixth principal axis is among the worst, which indicates that a situation similar to the one shown in figure 5.2 might be present. However, due to contributions from many parameters (similar to PC 3), the implications on the data cannot be visualized very easily. Additionally, the behaviour appears mostly for the JSD, but not so clearly for the other criteria (so differences are apparently not too severe in the end as they do not appear in mean/ median criterion).

The last systematic from table 5.6 we point out here is that the tendencies regarding good and bad agreement of certain axis not only appear in the rows where all events are analysed, but also in the ones where only good/ bad data are analysed. This shows us that while the overall

behaviour of parameters and combinations of them does change under transformations, the same is not necessarily true for each event (like the tables and plots at the beginning of this subsection already showed). This further validates our claim that patterns regarding good/bad agreement do not change significantly under a transformation, which is consistent with the observations from the regular PCs in the previous subsection. Again, GW200105_162426 appears as the only real outlier from the tendencies described and this time, it is even more apparent. Besides the JSD, mean and median criterion mark it red in the EOB basis (only mean criterion in Phenom basis), so perhaps excluding it from bad data was not the correct decision after all (although it was not clearly wrong at the point this decision was made, we now simply have more knowledge).

5.2.3. Comparison with previous Catalogs

In principle, one could do the exact same analyses that were done over the last subsections for the events from GWTC-1 and GWTC-2.1 as well. However, a detailed discussion would not be possible due to time constraints of this thesis. The only result we provide is the agreement of the previous catalogs in the basis given in figure 5.6 because this will allow us to see whether those axes are generally suited to reveal disagreement (which would support our claim from the previous subsection). From table 5.7 we can see that, apparently, these axes indeed show high disagreement. However, in case of GWTC-2.1 there are also other axes which exhibit comparable disagreement, so the pattern observed for GWTC-3 (especially in the median criterion) is not found there. For GWTC-1, a similar pattern does appear, but due to the overall very good agreement it is also not perfectly clear (small number where it occurs makes it vulnerable to outliers).

Eventlist	Parameter	JSD		JSD 2		Mean difference		Median difference	
		50%	20%	50%	20%	50%	20%	50%	20%
GWTC-1	PC 1	0/ 9 (0%)	4/ 9 (44%)	0/ 9 (0%)	0/ 9 (0%)	0/ 9 (0%)	1/ 9 (11%)	0/ 9 (0%)	4/ 9 (44%)
	PC 2	0/ 9 (0%)	3/ 9 (33%)	0/ 9 (0%)	0/ 9 (0%)	0/ 9 (0%)	1/ 9 (11%)	0/ 9 (0%)	1/ 9 (11%)
	PC 3	2/ 9 (22%)	7/ 9 (78%)	0/ 9 (0%)	0/ 9 (0%)	2/ 9 (22%)	7/ 9 (78%)	2/ 9 (22%)	6/ 9 (67%)
	PC 4	1/ 9 (11%)	3/ 9 (33%)	1/ 9 (11%)	3/ 9 (33%)	0/ 9 (0%)	1/ 9 (11%)	0/ 9 (0%)	3/ 9 (33%)
	PC 5	2/ 9 (22%)	6/ 9 (67%)	0/ 9 (0%)	1/ 9 (11%)	2/ 9 (22%)	6/ 9 (67%)	1/ 9 (11%)	6/ 9 (67%)
	PC 6	0/ 9 (0%)	7/ 9 (78%)	0/ 9 (0%)	2/ 9 (22%)	0/ 9 (0%)	6/ 9 (67%)	0/ 9 (0%)	6/ 9 (67%)
GWTC-2.1	PC 1	6/ 30 (20%)	17/ 30 (57%)	1/ 30 (3%)	3/ 30 (10%)	4/ 30 (13%)	13/ 30 (43%)	6/ 30 (20%)	12/ 30 (40%)
	PC 2	9/ 30 (30%)	17/ 30 (57%)	0/ 30 (0%)	3/ 30 (10%)	8/ 30 (27%)	12/ 30 (40%)	7/ 30 (23%)	13/ 30 (43%)
	PC 3	6/ 30 (20%)	15/ 30 (50%)	0/ 30 (0%)	0/ 30 (0%)	5/ 30 (17%)	13/ 30 (43%)	5/ 30 (17%)	13/ 30 (43%)
	PC 4	5/ 30 (17%)	22/ 30 (73%)	4/ 30 (13%)	18/ 30 (60%)	2/ 30 (7%)	12/ 30 (40%)	1/ 30 (3%)	15/ 30 (50%)
	PC 5	12/ 30 (40%)	20/ 30 (67%)	0/ 30 (0%)	6/ 30 (20%)	8/ 30 (27%)	18/ 30 (60%)	7/ 30 (23%)	17/ 30 (57%)
	PC 6	5/ 30 (17%)	19/ 30 (63%)	2/ 30 (7%)	5/ 30 (17%)	5/ 30 (17%)	17/ 30 (57%)	5/ 30 (17%)	16/ 30 (53%)
GWTC-3	PC 1	3/ 36 (8%)	16/ 36 (44%)	0/ 36 (0%)	5/ 36 (14%)	3/ 36 (8%)	13/ 36 (36%)	4/ 36 (11%)	14/ 36 (39%)
	PC 2	3/ 36 (8%)	17/ 36 (47%)	1/ 36 (3%)	7/ 36 (19%)	0/ 36 (0%)	10/ 36 (28%)	1/ 36 (3%)	14/ 36 (39%)
	PC 3	5/ 36 (14%)	19/ 36 (53%)	0/ 36 (0%)	9/ 36 (25%)	3/ 36 (8%)	15/ 36 (42%)	3/ 36 (8%)	16/ 36 (44%)
	PC 4	5/ 36 (14%)	23/ 36 (64%)	5/ 36 (14%)	19/ 36 (53%)	1/ 36 (3%)	11/ 36 (31%)	2/ 36 (6%)	13/ 36 (36%)
	PC 5	7/ 36 (19%)	21/ 36 (58%)	0/ 36 (0%)	6/ 36 (17%)	7/ 36 (19%)	19/ 36 (53%)	9/ 36 (25%)	18/ 36 (50%)
	PC 6	7/ 36 (19%)	17/ 36 (47%)	2/ 36 (6%)	10/ 36 (28%)	2/ 36 (6%)	8/ 36 (22%)	4/ 36 (11%)	9/ 36 (25%)

(a) Phenom

Eventlist	Parameter	JSD		JSD 2		Mean difference		Median difference	
		50%	20%	50%	20%	50%	20%	50%	20%
GWTC-1	PC 1	0/ 9 (0%)	3/ 9 (33%)	0/ 9 (0%)	0/ 9 (0%)	0/ 9 (0%)	3/ 9 (33%)	0/ 9 (0%)	4/ 9 (44%)
	PC 2	0/ 9 (0%)	4/ 9 (44%)	0/ 9 (0%)	0/ 9 (0%)	0/ 9 (0%)	2/ 9 (22%)	0/ 9 (0%)	2/ 9 (22%)
	PC 3	3/ 9 (33%)	6/ 9 (67%)	0/ 9 (0%)	1/ 9 (11%)	2/ 9 (22%)	6/ 9 (67%)	2/ 9 (22%)	6/ 9 (67%)
	PC 4	0/ 9 (0%)	3/ 9 (33%)	1/ 9 (11%)	6/ 9 (67%)	0/ 9 (0%)	0/ 9 (0%)	0/ 9 (0%)	2/ 9 (22%)
	PC 5	2/ 9 (22%)	7/ 9 (78%)	0/ 9 (0%)	1/ 9 (11%)	2/ 9 (22%)	6/ 9 (67%)	2/ 9 (22%)	4/ 9 (44%)
	PC 6	0/ 9 (0%)	5/ 9 (56%)	0/ 9 (0%)	2/ 9 (22%)	0/ 9 (0%)	3/ 9 (33%)	0/ 9 (0%)	4/ 9 (44%)
GWTC-2.1	PC 1	4/ 30 (13%)	16/ 30 (53%)	2/ 30 (7%)	5/ 30 (17%)	4/ 30 (13%)	12/ 30 (40%)	5/ 30 (17%)	14/ 30 (47%)
	PC 2	9/ 30 (30%)	20/ 30 (67%)	0/ 30 (0%)	4/ 30 (13%)	7/ 30 (23%)	18/ 30 (60%)	7/ 30 (23%)	13/ 30 (43%)
	PC 3	10/ 30 (33%)	17/ 30 (57%)	0/ 30 (0%)	4/ 30 (13%)	8/ 30 (27%)	17/ 30 (57%)	7/ 30 (23%)	18/ 30 (60%)
	PC 4	4/ 30 (13%)	17/ 30 (57%)	12/ 30 (40%)	24/ 30 (80%)	1/ 30 (3%)	9/ 30 (30%)	3/ 30 (10%)	14/ 30 (47%)
	PC 5	11/ 30 (37%)	20/ 30 (67%)	0/ 30 (0%)	6/ 30 (20%)	7/ 30 (23%)	16/ 30 (53%)	6/ 30 (20%)	16/ 30 (53%)
	PC 6	7/ 30 (23%)	20/ 30 (67%)	3/ 30 (10%)	6/ 30 (20%)	6/ 30 (20%)	17/ 30 (57%)	7/ 30 (23%)	17/ 30 (57%)
GWTC-3	PC 1	3/ 36 (8%)	16/ 36 (44%)	1/ 36 (3%)	5/ 36 (14%)	3/ 36 (8%)	12/ 36 (33%)	3/ 36 (8%)	13/ 36 (36%)
	PC 2	5/ 36 (14%)	18/ 36 (50%)	1/ 36 (3%)	7/ 36 (19%)	4/ 36 (11%)	10/ 36 (28%)	3/ 36 (8%)	11/ 36 (31%)
	PC 3	8/ 36 (22%)	19/ 36 (53%)	1/ 36 (3%)	6/ 36 (17%)	6/ 36 (17%)	17/ 36 (47%)	9/ 36 (25%)	17/ 36 (47%)
	PC 4	5/ 36 (14%)	17/ 36 (47%)	13/ 36 (36%)	25/ 36 (69%)	2/ 36 (6%)	6/ 36 (17%)	3/ 36 (8%)	18/ 36 (50%)
	PC 5	6/ 36 (17%)	23/ 36 (64%)	0/ 36 (0%)	5/ 36 (14%)	6/ 36 (17%)	21/ 36 (58%)	7/ 36 (19%)	21/ 36 (58%)
	PC 6	7/ 36 (19%)	19/ 36 (53%)	4/ 36 (11%)	13/ 36 (36%)	3/ 36 (8%)	12/ 36 (33%)	3/ 36 (8%)	12/ 36 (33%)

(b) EOB

Table 5.7.: Event statistics for average PCs of normalized data (see figure 5.6 for the axes which served as a basis). The numbers and percentages show how often each parameter exceeds the respective threshold of the column (`nocosmo` data used). The idea is to examine and compare the behaviour of all catalogs in this basis.

6. Conclusion

It is now time to review the goals formulated in the introduction and how each one could be achieved. The first important choice of this work was to select criteria, which can effectively find differences in probability distributions and allow to assess their significance. It turned out that there is no perfect criterion which can be used solely, but this was also not expected. Therefore, results from several criteria had to be used. The combination of JSD, mean criterion and median criterion introduced in chapter 4 turned out to be very useful to simultaneously find differences and assess how severe they are, considering the context of predicting the source properties of GW events. The JSD 2 criterion, which was also introduced in this chapter, was not used as much as the others. But that was mainly because we did not set the priority for studying shape differences too high since they are less severe differences compared to the ones found by the other criteria (it seemed to produce reasonable values, that is not an issue). Similarly, the 50% threshold chosen turned out to be very well suited to distinguish events with good and bad agreement since it did not mark too many or too few events (both would not allow for an analysis that makes sense), while the 20% threshold all in all did not produce very helpful results and was not used much.

The results inferred using this set of criteria then tell us something about a certain set of parameters, among which are widely used ones like the chirp mass $\mathcal{M}$, mass ratio q or effective and precession spin $\chi_{\text{eff}}, \chi_p$. All in all, those results tell us that we do not have to change the way we think about the parameters, waveform models and corresponding posteriors fundamentally. While there were some interesting and partially unexpected systematics, it was possible to find a potential explanation for most of them from already known results like e.g. mismatch-studies. One systematic we could not explain with previous knowledge was the remarkably high disagreement for mass parameters and in particular the total mass M. Although a potential reason in extreme mass ratios could be identified, we could not find reasons why this parameter should be affected by that much more than others. Moreover, certain differences in the behaviour of the two component masses m_1, m_2 did not match our expectations and we also could not find a satisfying explanation for them. But they seem to be prior-induced, so it is very likely that an explanation related to the choice of these priors exists, which we are simply not aware of. Apart from the last two patterns, we do not find evidence that the limitations of waveform models are not well understood or insufficient to explain GWTC-3 data.

To get a comparison for the results, we also conducted a similar, but not so detailed, analysis of the previous catalogs GWTC-1, GWTC-2.1. This revealed that it is not necessarily possible to transfer systematics between the catalogs because overall agreement and also that the source properties of the events in each catalog vary substantially.

The second step was analysing the posteriors using a PCA. Besides doing that in a more conventional way, i.e. compute a transformation for the data from each event, we also presented a way to analyse each event in terms of correlations of the whole population of events. Both analyses further confirmed that no fundamental change of the current understanding of waveform models and their posteriors is necessary because the overall agreement of an event does not seem to be highly dependent on the axes we use to represent its posterior (although the overall agreement increases in both PC bases and average PC bases). That is, however, not true for the agreement along individual axes, which does depend on the representation we choose. In the first basis, which is given by the set of parameters we use, it is often the case that many axes show a certain level of disagreement. In contrast, for the results in the average basis, it is more common that less axes show disagreement, but now it is more severe than before.

Most of these results can be extended very well by future research. The most straightforward application would be doing essentially the same analyses for data from future observing runs and comparing the results (although a comparison of posteriors makes most sense when they are generated from the same waveform models, so a re-analysis might be necessary). Further potential topics include more detailed investigations of the behaviour found for M, m_1 and if a correlation with the corresponding q-value can be further supported, but also how the choice of priors might affect waveform systematics (we saw that some of the unexpected results may be related to or even caused by the prior). Additionally, it would be very interesting to look at the theoretical counterpart of this work, in particular results from the Fisher-matrix formalism, to see how well predictions of uncertainties and inconsistencies match what was found here. This would probe how well the quality of LVK inferences is understood and subsequently may be predicted. Complementary to such an analysis would be an in-depth study of how glitches, noise in general or events with bad agreement affect the overall results of this work (like the behaviour in the q-χ_{eff}-plane or the average correlation matrix and consequently average PCs).

During this thesis, a lot of code was produced and used to obtain the results presented here. For the sake of transparency, there is a plan to make it publicly available on GitHub along with proper documentation and scripts to reproduce the figures and tables related to GW posteriors and waveform systematics (if time constraints allow to do so).

A. Probability Theory and Statistics

A.1. Probability

The main sources of this paragraph are [29] (focussed on the mathematics of probability theory; introduces important notions and shows proofs) and [14] (great introduction to Bayesian statistics). Additionally, information from GW-related papers cited throughout this work have been used as most of them include a short review of Bayesian statistics.

The mathematical methods of probability theory make heavy use of measure theory and Lebesgue integrals (although calculations can be carried out using regular Riemann integrals for the most part). Because this work does not require detailed knowledge of this mathematical aspects, we will only review them briefly, starting with a summary of the *Kolmogorov axioms* which the theory is built up on.

Describing an experiment takes place in a *probability space* (Ω, F, P), which is a measure space with some additional requirements.

(i) Ω is the *sample space* containing every possible outcome of the experiment. For a fixed experiment, there are many possible sample spaces and one has to choose the best one containing only relevant information.

(ii) F is the *event space* containing sets of outcomes from Ω. These sets are *events* and they can also represent very abstract objects like some hypothesis being true or false.

(iii) P is a *probability measure*/ *distribution*. It maps events $E \in F$ to their probabilities

$$P(E) = \int_E dP \tag{A.1}$$

which are restricted to the interval $[0, 1]$ by demanding $P(\emptyset) = 0$, $P(\Omega) = 1$.

$\Rightarrow$ using $E, E_1, E_2 \in F$, some basic properties of probability distributions are:

$$E_1 \subset E_2 \Rightarrow P(E_1) \leq P(E_2) \tag{A.2a}$$

$$P(E_1) + P(E_2) = P(E_1 \cup E_2) + P(E_1 \cap E_2) \tag{A.2b}$$

$$P(E^c) = P(F \backslash E) = 1 - P(E) \tag{A.2c}$$

$$P(E_1) = P((E_1 \cap E_2) \cup (E_1 \backslash E_2)) = P(E_1 \cap E_2) + (E_1 \backslash E_2) \tag{A.2d}$$

Suppose now that an event $E \in F$ depends on two other events $A, B \in F$ via $E = A \cap B$. Then the probability $P(E)$ is nothing but the probability of A *and* B, the *joint probability*

$$P(A, B) := P(A \cap B) = \int_{A \cap B} dP\,. \tag{A.3}$$

For *independent* events, where the outcome of one does not affect the other, it is simply

$$P(A, B) = P(A)\,P(B)\,. \tag{A.4}$$

A more general definition is to use the *conditional probability* $P(A|B)$ which assigns a probability to A occurring given that B occurred. Because the joint probability $P(A, B)$ simply measures the probability of A and B occurring, one can express it as A occurring and B occurring given A (or vice versa), which translates to

$$P(A, B) = P(A)\,P(B|A) = P(B)\,P(A|B)\,. \tag{A.5}$$

For independent events A, B, the occurrence of B does not influence the occurrence of A and thus, $P(A|B) = P(A)$ in this case, which is consistent with (A.4).

A.1.1. Random Variables and Vectors

The next step is to look at quantities depending on random events (like a probability to win a certain amount of money from a coin toss) and these can be described by a *random variable*[1] $X : \Omega \to A$ mapping events to an arbitrary measurable space $(A, \mathcal{A})$. Each value $x = X(s) \in A$, $s \in \Omega$ is a *realization* of X.[2] By using the concept of pushforward measure from measure theory, one can assign probabilities to subsets $S \subset A$:

$$\begin{aligned} P(X \in S) := P(X(s) \in S) &= \int_S d(X_* P) = \int_S d\Big(P \circ X^{-1}\Big) \\ &= \int_{X^{-1}(S)} dP = P\Big(X^{-1}(S)\Big) = P(\{s \in \Omega :\ X(s) \in S\})\,. \end{aligned} \tag{A.6}$$

This is the *(probability) distribution* of X. The idea is to measure the probability of all outcomes mapping to S (which form an event). An equivalent definition can be given by using a *probability density function* (PDF)[3] $p_X : A \to [0, \infty),\ x \mapsto p_X(x)$:

$$P(X \in S) = \int_S p_X\, d\mu \tag{A.7}$$

[1] In this review, we will ignore the difference between continuous and discrete random variables because the latter arises as a special case of the former when P is a counting measure.

[2] Assigning probabilities to a single x does not make sense because $P(\{x\}) = 0$, points are null sets. Instead, one can only make statements about probabilities that realizations close to x happen.

[3] The PDF of a discrete variables is termed *probability mass function* and to compute probabilities for them, $\int d\mu$ can be replaced with $\sum$ which is an effect of using a counting measure. Then, $p_X(x) = P(X = x)$, so density and probability are basically equivalent (which is *not* true in the continuous case).

where μ is a measure such that $(A, \mathcal{A}, \mu)$ is a measure space. $p_X(x)$ measures how likely a realization of X close to x would be. It is connected to (A.6) via $dX_*P = p_X d\mu \Leftrightarrow p_X = \frac{dX_*P}{d\mu}$.

These definitions can be extended to multiple random variables X, Y on the same probability space (Ω, F, P) by using a *multivariate random variable* / *random vector* $(X, Y) : \Omega \to A_X \times A_Y$, which has essentially the same properties as a random variable, but in more dimensions. The corresponding probability distribution

$$P(S \subset \Omega_X \times \Omega_Y) = \int_S p_{X,Y}\, d\mu_X d\mu_Y \tag{A.8}$$

then assigns probabilities to subsets $S = S_X \times S_Y$ of realizations x of X *and* y of Y, so it is nothing but a joint probability. Thus, $p_{X,Y} : A_X \times A_Y \to [0, \infty)$, $(x, y) \mapsto p_{X,Y}(x, y)$ is the joint PDF of X, Y. It turns out that joint PDFs encode the PDF of each random variable out of the random vector they belong to (which is (X, Y) in this case). This is due to Fubini's theorem, which tells us how to calculate integrals in product spaces and states:

$$\int_{S_X \times S_Y} p_{X,Y}\, d\mu_X d\mu_Y = \int_{S_Y} \left(\int_{S_X} p_{X,Y}\, d\mu_X \right) d\mu_Y = \int_{S_X} \left(\int_{S_Y} p_{X,Y}\, d\mu_Y \right) d\mu_X \,. \tag{A.9}$$

Comparing this formula to (A.7), one can already see similarities to the probability distributions of a random variable. Indeed, it can be shown (theorem 5.5.3 in [29]) that the PDF of X (or Y in just the same manner) can be expressed as

$$p_X(x) = \int_{A_Y} p_{X,Y}(x, y)\, d\mu_Y = \int_{A_Y} p_{X,Y}(x, y)\, dy \tag{A.10}$$

and in this context, it is also termed *marginal PDF* describing the *marginal distribution* $P(X \in S)$. The intuition behind this formula is that the variable Y is "averaged out" by integrating over all of its values in A_Y. This process is known as *marginalization*.

Analogously to the formulas for events, $p_{X,Y}(x, y) = p_X(x)\, p_Y(y)$ for independent X, Y and also $p_{X,Y}(x, y) = p_{X|Y=y}(x|y)\, p_Y(y)$. Because of the latter, marginalization can be written as

$$p_X(x) = \int_{A_Y} p_{X|Y=y}(x|y)\, p_Y(y)\, d\mu_Y = \int_{A_Y} p_{X|Y=y}(x|y)\, p_Y(y)\, dy \tag{A.11}$$

which is very useful in certain situations.

Although their mathematical properties will not be covered, it should be mentioned that one can also deal with sequences of random variables. An example of such a *random process* is a time-dependent random variable $X(t)$, i.e. a random variable $X = X_t$ at each time t.

A.1.2. Bayesian Probability

The notion of probability is often thought of as telling us which outcomes to expect when repeating an experiment n times. It turns out though that this *frequentist interpretation* of probability is not the only possible one. The same theory can also be built up using probability as a degree of belief in something and this is the *Bayesian interpretation*. The Bayesian equivalent to saying that a certain event will occur k times out of n repeated experiments is the statement "we are $p\%$ sure that the experiment will yield a certain result". The mathematical implications are the same for either way, so these approaches are equivalent.

We will now turn to equation (A.5), which can be rewritten to read

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \tag{A.12}$$

and this is *Bayes' theorem* (valid for PDFs as well). Although this theorem holds regardless of the interpretation chosen, it has a particularly interesting meaning in the Bayesian one. A detailed explanation of how to interpret each quantity is given in subsection 2.2.2, but the main point can also be understood without this knowledge: Bayes' theorem allows to update the belief in A from $P(A)$ to $P(A|B)$ by a computation and using B, which can e.g. be observed data. Also, $P(B)$ does *not* have to be known or estimated because using equation (A.11) this can be expressed using the quantities in the numerator as

$$P(B) = \int P(B|A)\,P(A)\,dA\,. \tag{A.13}$$

Alternatively (and a bit more explicitly), using properties (A.2d) and (A.5) we can rewrite

$$\begin{aligned} P(B) &= P((B \cap A) \cup (B \backslash A)) = P(B \cap A) + (B \backslash A) \\ &= P((B \cap A) \cup (B \backslash A)) = P(B \cap A) + (B \cap \neg A) \\ &= P(B|A)P(A) + P(B|\neg A)P(\neg A)\,. \end{aligned} \tag{A.14}$$

and that allows us to obtain another formulation of Bayes' theorem, i.e.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)} = \frac{1}{1 + \frac{P(B|\neg A)P(\neg A)}{P(B|A)P(A)}} \tag{A.15}$$

which shows interesting dependencies of the posterior probability.

A.2. Histograms

For many applications we want to work with the PDF p of a random variable, but only have a sample of the distribution. Such a sample is of course not equal to the PDF, although it contains information which can be used to estimate p. To understand how that is done, lets imagine how the values in the sample would look like if plotted on a line (we restrict to 1D values, which are points on e.g. the x-axis): there would be segments with lots of points, segments with only a few points and segments with no points. Intuitively, the number of points in a segment are a density and, when normalized with the total number of points, indeed an estimate of the PDF we are looking for. From now on, these segments are termed *bins* and the values in a bin form the *histogram*, an estimate for the PDF (figure A.1).

A.2.1. Optimal Bin Size

Now, we will make the description more mathematical. The PDF $p = p(x)$ is approximated by $p_i = \frac{f_i}{h_i}$ on k bins $[x_i, x_{i+1}]$, $1 \leq i \leq k$, which are usually chosen to have equal length $h = h_i = x_{i+1} - x_i$ (so are they here). $f_i = \frac{m_i}{m}$ denotes the frequency/ normalized number of points in the i-th bin and they fulfil $1 = \sum_{i=1}^{k} f_i \Leftrightarrow m = \sum_{i=1}^{k} m_i$. The length of the interval in which the sample lies is denoted by s.

How good of an estimator the histogram is mainly depends on the number of bins k chosen. On the one hand, it is clear that small k will lead to a bad approximation of the PDF (for the most extreme case of $k = 1$ it is only one big box, which in general is not the PDF). On the other hand, choosing too many bins will also lead to a bad approximation because some bins will contain zero values despite the PDF being non-zero (will certainly happen for $k > m$ and even well below because the elements of the sample are, in general, not equally distributed). Both cases are visualized in figure A.2 for a normal distribution.

Therefore, we have to find a balanced value for k and mathematically speaking, this is an optimization problem. As [30] argues, the total relative error caused by the two factors described above can (approximately) be written as

$$E(h) = \frac{h}{s} + \sqrt{\frac{s}{m\,h}}\,. \tag{A.16}$$

The corresponding graph is shown in figure A.3. Optimization corresponds to finding the h_{opt} which minimizes E. Differentiating E with respect to h and looking for its zeros yields

$$0 = \frac{1}{s} - \frac{1}{2}\sqrt{\frac{s}{m\,h_{\text{opt}}^3}} \qquad \Leftrightarrow \qquad h_{\text{opt}} = \frac{s}{2^{2/3} m^{1/3}} \tag{A.17}$$

and because the bins shall span the whole interval, i.e. $s = k\,h$, this optimal bin size gives us the optimal number of bins

$$k_{\text{opt}} = \frac{s}{h_{\text{opt}}} = 2^{2/3} m^{1/3} = (4m)^{1/3}\,. \tag{A.18}$$

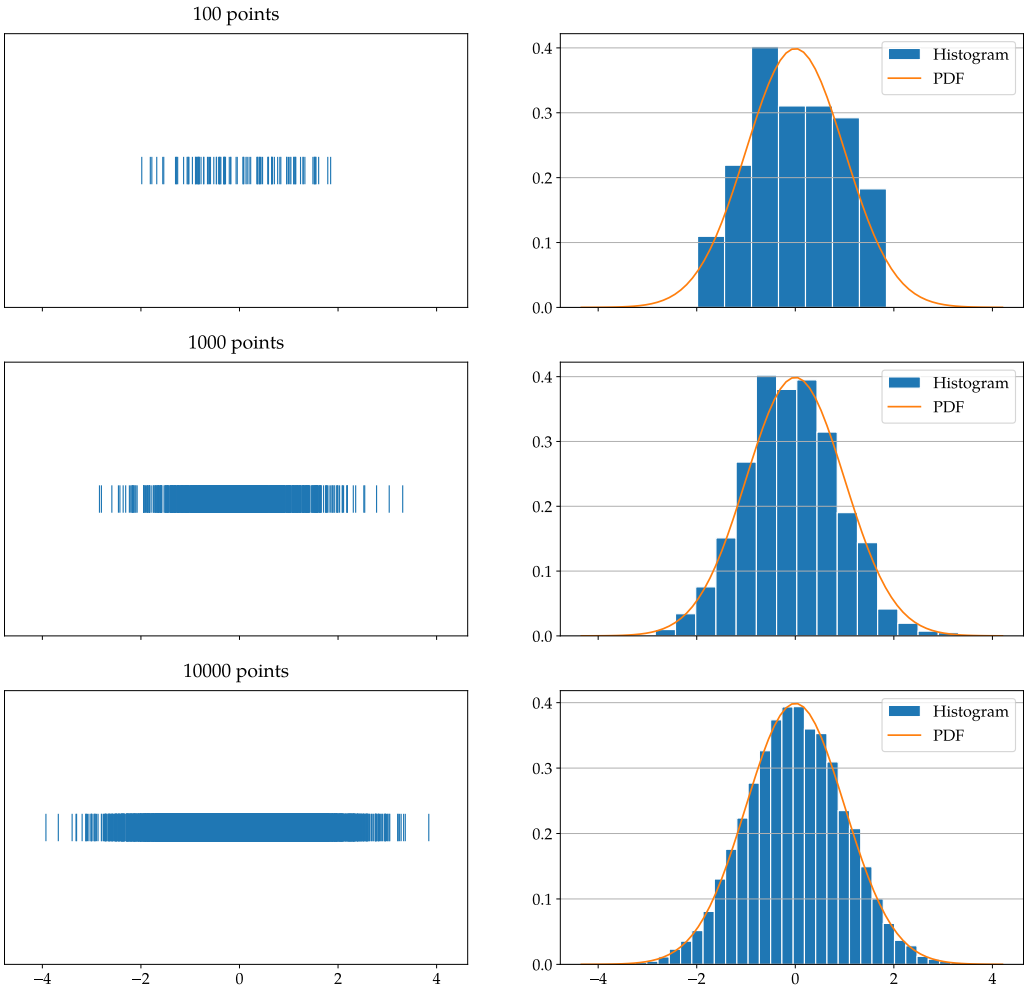


Figure A.1.: Distribution of data vs. histogram. This is a comparison of the data points themselves (samples from a normal distribution with zero mean, standard deviation of 1) and the histogram generated from them. For smaller sizes m, the points can be distinguished in the left plots, whereas for bigger m this is not the case anymore (density too high).
By counting the number of points in a certain interval, the histograms shown on the right are obtained and they approximate the PDF better for higher m.

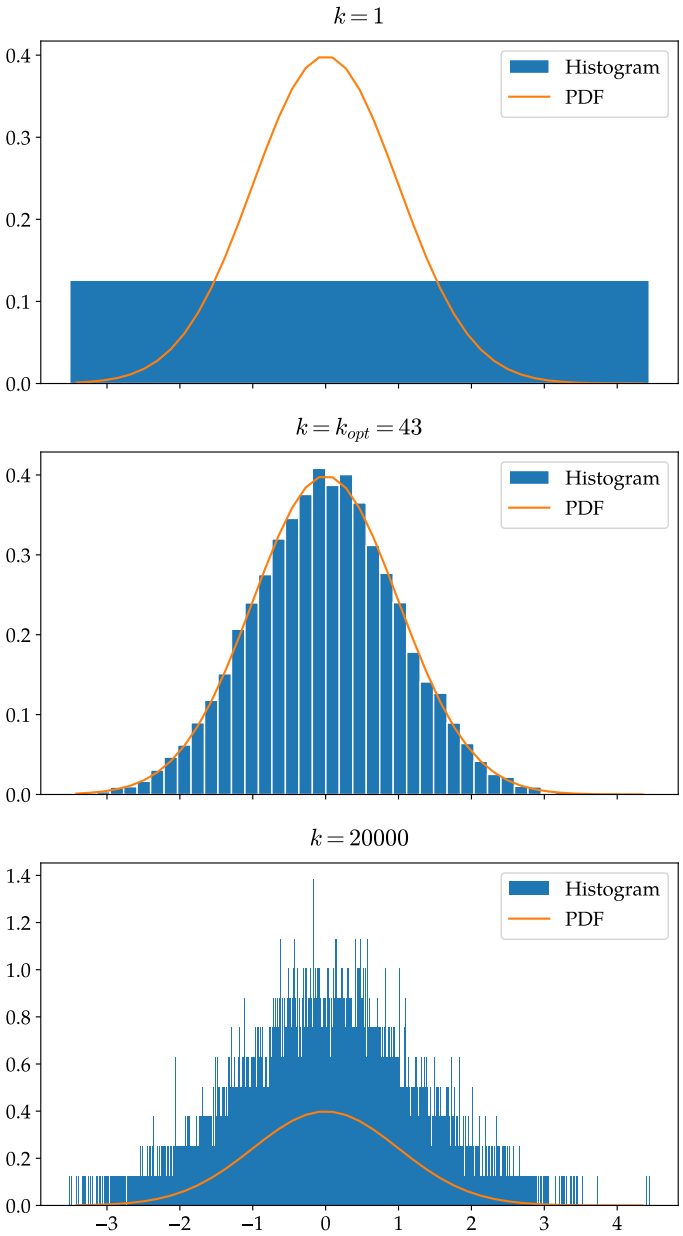


Figure A.2.: Histograms for different bin numbers k. Of course, taking only one bin is a bad approximation of the PDF (normal distribution with zero mean and standard deviation of 1). However, $k = m$ (third plot) is also a very bad approximation because many bins have $f_i = 0 = p_i$ despite the PDF being non-zero, which causes other ones to overestimate f_i (normalization constraint still has to be fulfilled), which is an effect of the finite number of samples m (this error would be minimized for the case of one bin, i.e. $h = s$, and then be equal to the well known relative error $1/\sqrt{m}$ because the information in form of samples we have is not split up into different bins). For comparison, the middle plot shows the approximation using k calculated according to (A.18).

Although this result does not take the variation of the data into account, it will suffice for the objectives of this work. It is also the asymptotically optimal bin size.

A.2.2. Extending Histogram Intervals

The last question we have to deal with in this context is: what to do if we want to calculate the histogram on an interval that is not s? This is indeed an interesting question for this work because we made use of the JSD of two posterior probabilities from different waveform models. Because the posteriors we use are samples and sampling is a stochastic process, the intervals s_1, s_2 will very likely not be equal. The problem then is that the histograms computed from the samples do not have equal bins. Equation (4.2) tells us that in this case, the JSD (which takes PDFs or histograms as their estimates as arguments) will not compare p_i from the same bins and functions in computer programs like the `scipy.distance.jensenhannon` function in Python (used in this work) will make this mistake as well.

To solve this problem, we will compute the histogram on the interval $s_{12} = s_1 \cup s_2$. That will only add bins with $p_i = 0$ to the histogram because outside of the respective interval s_1, s_2 there are no samples by construction, i.e. $f_i = 0 = p_i$.[4] If the sample size is sufficiently high, we can be sure that no samples in a certain interval actually tell us that the PDF is 0 or at least close to 0 there and so the values these bins still estimate the PDF well.

After extending the respective intervals s_1, s_2, we now have to be careful with bin sizes and numbers. The optimal number of bins (A.18) has been computed for the s_j, $j = 1, 2$, so it has to be adapted to s_{12}. The optimal bin size (A.17) remains the same because the extension contains no samples, so we can write

$$s_{12} = k_{j,\mathrm{opt}}\, h_{j,\mathrm{opt}} \quad \Leftrightarrow \quad k_{j,\mathrm{opt}} = \frac{s_{12}}{h_{j,\mathrm{opt}}} = \frac{s_{12}\,(4m_j)^{1/3}}{s_j} \tag{A.19}$$

where $j = 1, 2$. This means there are different bin numbers for the different histograms, which does make sense because the sample sizes will differ in general.

Because computing the JSD requires the same number of bins for both samples, we will have to choose one of the optimal $k_{j,\mathrm{opt}}$. As this will not be the optimal choice for the other one of the samples, we have to decide whether it is better to choose the bin size too large or too small. To assess that, it makes sense to look at the the behaviour of the error $E(h)$ in each direction and from figure A.3 we can see that the slope is steeper for smaller h. Therefore, we choose $\max\left(h_{1,\mathrm{opt}}, h_{2,\mathrm{opt}}\right)$ or, equivalently, $\min\left(k_{1,\mathrm{opt}}, k_{2,\mathrm{opt}}\right)$. The use of this method leads to consistent results for JSDs (which is important because computing it was one of the main motivations to use histograms in the first place), e.g. it reproduces the numbers cited in appendix A of [3] (for instance that the JSD of two samples from the same normal distribution which can be up to 0.002 bits due to sampling variations).

[4] Usually, bins like this are not interesting as they provide no new insights. Most PDFs are 0 or very close to 0 on the interval $\mathbb{R}\backslash s$ and the interesting property is where $p(x) \neq 0$. Mathematically speaking, we are only interested in the support of p, which is estimated by s.

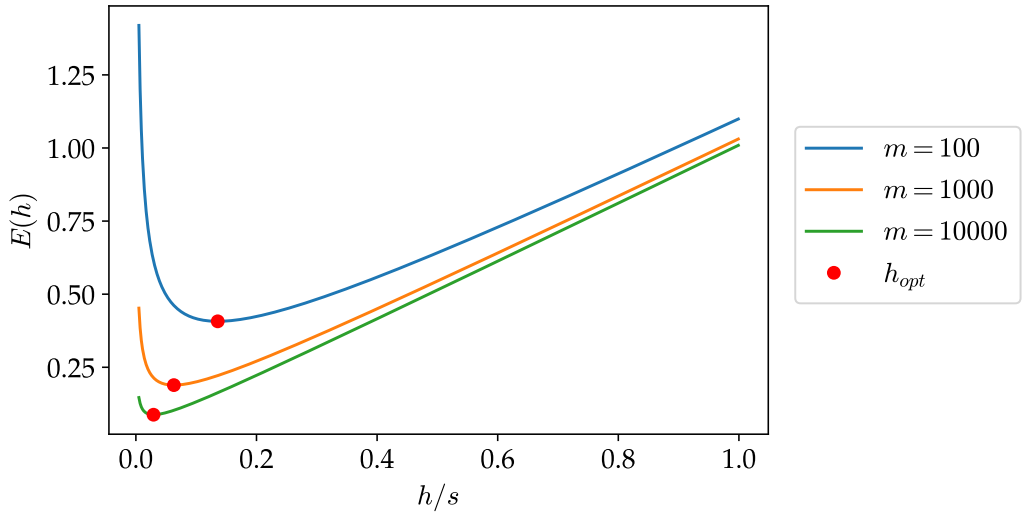


Figure A.3.: Error $E(h)$ for different sample sizes m. Important observations are that the overall error is smaller for bigger m (as we would expect) and that $E(h)$ it has a steeper slope for bin sizes smaller than h_{opt}.

Acknowledgements

Thanks to everybody who gave me a chance and put me in the position to be able to write this thesis. I especially want to thank Dr. Frank Ohme and Angela Borchers, who both spent a lot of their precious time on helping me with this project. Moreover, I want to thank all members of the Binary Mergers & Numerical Relativity group, but also every other members of the AEI I could meet, for being so nice and welcoming during the time I spent there.

I want to thank my parents, who always gave me the courage to do what I want and supported me unconditionally, as well as my whole family and friends for accepting me as the person I am and making life outside of work enjoyable.

Finally, I want to acknowledge that this project would not have been possible without the effort of many people to build open science projects and open source/ free software (like GWOSC, Python or LaTeX, to mention only a few of them).

References

[1] The LIGO Scientific Collaboration, The Virgo Collaboration, The KAGRA Collaboration, R. Abbott *et al.*, "GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run," 2021. [Online]. Available: https://arxiv.org/abs/2111.03606

[2] B. P. Abbott *et al.*, "GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs," *Physical Review X*, vol. 9, no. 3, Sep 2019. [Online]. Available: https://doi.org/10.1103%2Fphysrevx.9.031040

[3] R. Abbott *et al.*, "GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo during the First Half of the Third Observing Run," *Physical Review X*, vol. 11, no. 2, Jun 2021. [Online]. Available: https://doi.org/10.1103%2Fphysrevx.11.021053

[4] C. Bambi, *Introduction to General Relativity*, 1st ed. Springer Singapore, 2018.

[5] F. Ohme, "Bridging the gap between Post-Newtonian theory and numerical relativity in gravitational-wave data analysis," Ph.D. dissertation, Universität Potsdam, Mathematisch-Naturwissenschaftliche Fakultät, 2012.

[6] B. S. Sathyaprakash and Bernard F. Schutz, "Physics, Astrophysics and Cosmology with Gravitational Waves," *Living Reviews in Relativity*, vol. 12, no. 1, Mar 2009. [Online]. Available: https://doi.org/10.12942%2Flrr-2009-2

[7] B. P. Abbott, R. Abbott, T. D. Abbott *et al.*, "A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals," *Classical and Quantum Gravity*, vol. 37, no. 5, p. 055002, Feb 2020. [Online]. Available: https://doi.org/10.1088/1361-6382/ab685e

[8] The LIGO Scientific Collaboration, The Virgo Collaboration, R. Abbott *et al.*, "GWTC-2.1: Deep Extended Catalog of Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run," 2021. [Online]. Available: https://arxiv.org/abs/2108.01045

[9] S. Ossokine, A. Buonanno, S. Marsat, R. Cotesta, S. Babak, T. Dietrich, R. Haas, I. Hinder, H. P. Pfeiffer, M. Pürrer, C. J. Woodford, M. Boyle, L. E. Kidder, M. A. Scheel, and B. Szilágyi, "Multipolar effective-one-body waveforms for precessing binary black holes: Construction and validation," *Phys. Rev. D*, vol. 102, p. 044055, Aug 2020. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.102.044055

[10] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, "FINDCHIRP: An algorithm for detection of gravitational waves from inspiraling compact binaries," *Physical Review D*, vol. 85, no. 12, jun 2012. [Online]. Available: https://doi.org/10.1103%2Fphysrevd.85.122006

[11] J. D. E. Creighton and W. G. Anderson, *Gravitational-Wave Physics and Astronomy: An Introduction to Theory, Experiment and Data Analysis*. Wiley, 2011.

[12] J. Veitch, V. Raymond, B. Farr *et al.*, "Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library," *Phys. Rev. D*, vol. 91, p. 042003, Feb 2015. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.91.042003

[13] I. M. Romero-Shaw, C. Talbot, S. Biscoveanu *et al.*, "Bayesian inference for compact binary coalescences with bilby: validation and application to the first LIGO–Virgo gravitational-wave transient catalogue," *Monthly Notices of the Royal Astronomical Society*, vol. 499, no. 3, pp. 3295–3319, 09 2020. [Online]. Available: https://doi.org/10.1093/mnras/staa2850

[14] D. S. Silva, *Data Analysis: A Bayesian Tutorial*, 2nd ed. Oxford University Press, 2006.

[15] A. J. K. Chua, "A one-stop function for gravitational-wave detection, identification and inference," 2022. [Online]. Available: https://arxiv.org/abs/2205.08702

[16] C. Pankow, P. Brady, E. Ochsner, and R. O'Shaughnessy, "Novel scheme for rapid parallel parameter estimation of gravitational waves from compact binary coalescences," *Phys. Rev. D*, vol. 92, p. 023002, Jul 2015. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.92.023002

[17] L. S. Collaboration, V. Collaboration, and K. Collaboration, "GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run — Parameter estimation data release," Nov 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5546663

[18] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer Verlag, 2002.

[19] J. E. Jackson, *A User's Guide to Principal Components*. Wiley, 2003.

[20] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions Royal Society*, vol. 374, 2016. [Online]. Available: https://doi.org/10.1098/rsta.2015.0202

[21] J. Fischer. [Online]. Available: https://johfischer.com/2021/12/31/intuitive-explanation-of-the-kullback-leibler-divergence/

[22] G. Pratten, C. García-Quirós, M. Colleoni, A. Ramos-Buades, H. Estellés, M. Mateu-Lucena, R. Jaume, M. Haney, D. Keitel, J. E. Thompson, and S. Husa, "Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes," *Phys. Rev. D*, vol. 103, p. 104056, May 2021. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.103.104056

[23] F. Ohme, "Analytical meets numerical relativity: status of complete gravitational waveform models for binary black holes," *Classical and Quantum Gravity*, vol. 29, no. 12, p. 124002, jun 2012. [Online]. Available: https://doi.org/10.1088/0264-9381/29/12/124002

[24] E. Payne, S. Hourihane, J. Golomb, R. Udall, D. Davis, and K. Chatziioannou, "The curious case of gw200129: interplay between spin-precession inference and data-quality issues," 2022. [Online]. Available: https://arxiv.org/abs/2206.11932

[25] G. Morras, J. F. N. Siles, J. Garcia-Bellido, and E. R. Morales, "The false alarms induced by gaussian noise in gravitational wave detectors," 2022. [Online]. Available: https://arxiv.org/abs/2209.05475

[26] M. Colleoni, M. Mateu-Lucena, H. Estellés, C. García-Quirós, D. Keitel, G. Pratten, A. Ramos-Buades, and S. Husa, "Towards the routine use of subdominant harmonics in gravitational-wave inference: Reanalysis of GW190412 with generation X waveform models," *Phys. Rev. D*, vol. 103, p. 024029, Jan 2021. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.103.024029

[27] D. Davis, T. B. Littenberg, I. M. Romero-Shaw, M. Millhouse, J. McIver, F. Di Renzo, and G. Ashton, "Subtracting glitches from gravitational-wave detector data during the third observing run," 2022. [Online]. Available: https://arxiv.org/abs/2207.03429

[28] L. S. Collaboration and V. Collaboration, "GWTC-2.1: Deep Extended Catalog of Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run - Parameter Estimation Data Release," May 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6513631

[29] R. Meester, *A Natural Introduction to Probability Theory*, 1st ed. Birkhäuser Basel, 2003.

[30] A. Gholamy and V. Kreinovich, "What Is the Optimal Bin Size of a Histogram: An Informal Description"," *International Mathematical Forum*, vol. 12, no. 15, pp. 731–736, 2017. [Online]. Available: https://scholarworks.utep.edu/cs_techrep/1176/

Index