CrossMark

CAMBRIDGE
UNIVERSITY PRESS

Review Article

When models fail: An introduction to posterior predictive checks and model misspecification in gravitational-wave astronomy

Isobel M. Romero-Shaw[1,2] , Eric Thrane[1,2] , and Paul D. Lasky[1,2]

[1]Monash Astrophysics, School of Physics and Astronomy, Monash University, Clayton, VIC 3800, Australia and [2]OzGrav: The ARC Centre of Excellence for Gravitational-Wave Discovery, Clayton, VIC 3800, Australia

Abstract

Bayesian inference is a powerful tool in gravitational-wave astronomy. It enables us to deduce the properties of merging compact-object binaries and to determine how these mergers are distributed as a population according to mass, spin, and redshift. As key results are increasingly derived using Bayesian inference, there is increasing scrutiny on Bayesian methods. In this review, we discuss the phenomenon of *model misspecification*, in which results obtained with Bayesian inference are misleading because of deficiencies in the assumed model(s). Such deficiencies can impede our inferences of the true parameters describing physical systems. They can also reduce our ability to distinguish the 'best fitting' model: it can be misleading to say that Model A is preferred over Model B if both models are manifestly poor descriptions of reality. Broadly speaking, there are two ways in which models fail. Firstly, models that fail to adequately describe the data (either the signal or the noise) have misspecified likelihoods. Secondly, population models—designed, for example, to describe the distribution of black hole masses—may fail to adequately describe the true population due to a misspecified prior. We recommend tests and checks that are useful for spotting misspecified models using examples inspired by gravitational-wave astronomy. We include companion python notebooks to illustrate essential concepts.

Keywords: gravitational waves – Bayesian inference – posterior predictive checking – model misspecification

1. Introduction

Bayesian inference and parameter estimation are the cornerstones of gravitational-wave astronomy. The Bayesian framework is used to derive posterior distributions for parameters such as the masses and spins of merging pairs of neutron stars and black holes. Examples range from the progenitor of the first detected gravitational wave, GW150914 (Abbott et al. 2016b) and the binary neutron star merger, GW170817 (Abbott et al. 2017), to the many sources now presented in the catalogues of gravitational-wave transients published by the LIGO–Virgo–KAGRA Collaboration (Abbott et al. 2019a; Abbott et al. 2021d,a,b) and independent authors (e.g., Venumadhav et al. 2019, 2020; Olsen et al. 2022). Using Bayesian inference, we obtain values for the marginal likelihood (also known as the evidence), which are used for model selection. Model selection has been used, for example, to compare models of the neutron star equation of state (Abbott et al. 2020b), assess the most likely progenitors of intermediate-mass black hole merger GW190521 (Abbott et al. 2020c; Romero-Shaw et al. 2020; Bustillo et al. 2021), and study gravitational-wave events for evidence of modified gravity (e.g., Ghosh 2022).

A second layer of Bayesian analysis is built upon this foundation to study the population properties of merging

binaries. Hierarchical models are used to estimate population hyper-parameters describing how sources of gravitational waves are distributed according to mass, spin, redshift, and so on. This has been key in, for example, the discovery of features in the distribution of binary black hole masses (Abbott et al. 2020a, 2021c).

Results obtained with Bayesian inference are only as reliable as the underlying model. The compact-object binary parameters reported in gravitational-wave transient catalogues are derived using models that describe physical systems; only if these models are sufficient descriptors of the true system can these results be meaningful. Bayesian inference can tell us that one model is a better explanation for the data than another. For example, Bayesian techniques have been used to suggest that intermediate-mass black hole merger GW190521 shows signs of non-zero orbital eccentricity (Romero-Shaw et al. 2020). However, Bayesian inference *on its own* does not tell us if either the quasi-circular or eccentric gravitational waveforms considered provide an adequate fit to the GW190521 data. Similarly, Bayesian inference has been used to suggest that the distribution of primary black hole mass is better fit by a broken power-law distribution than a power-law distribution with no break (Abbott et al. 2020a). However, Bayesian inference *on its own* does not tell us if either of these models is adequate to describe the observed distribution of primary black hole masses.

As the gravitational-wave catalogue grows and gravitational-wave detector sensitivity improves, we begin to see more events that push the boundaries of our understanding of the Universe. This makes it ever more important to test the validity of our models. A signal model that is valid for systems with mass ratios $q \geq 0.125$ may be invalid for a mass ratio of $q = 0.001$. A detector noise

model adequate for an event with signal-to-noise ratio SNR $= 30$ may inadequately represent the detail contained within an SNR $= 100$ signal. Additionally, as the number of gravitational-wave signal detections grows, the resolving power of the combined dataset increases. This makes it ever more important to test the validity of population models. A population model for the distribution of binary black hole redshifts that works reasonably well for a dozen events may be unsuitable for a catalogue with hundreds of events.

In this Article, we describe different ways in which models can fail and lay out commonly used tests that can be carried out to reveal these failures, often beginning with visualisation. For a broader discussion of how visualisation can assist in solving Bayesian inference problems, see Gabry et al. (2017). We discuss workarounds for misspecified models, including model redesign and data coarsening (Miller & Dunson 2019; Thomas & Corander 2019). For an idea of how Bayesian inference problems may be solved through a careful workflow and iterative model redesign of the prior and likelihood, see both Betancourt (2020) and Gelman et al. (2020). While we cast many examples in the language of gravitational-wave astronomy, we endeavour to use sufficiently general language so that this review is useful to a broad audience. For additional resources, we refer the reader to Chapters 6-7 of Gelman et al. (2013), respectively devoted to 'Model checking' and 'Evaluating, comparing, and expanding models.' See also 'Model Checking and Sensitivity Analysis' in Andreon & Weaver (2015).

Almost all of the Subsections in this review follow the same formula. After introducing a concept, we provide a bullet list of *recommended tests*. This list is followed by a *demonstration*, which illustrates the tests with simple examples. This layout is designed to help researchers scan the Article to quickly find the misspecification tests they are looking for. Our recommended tests do not constitute an exhaustive list of the ways in which one may test for misspecification. In addition, an analysis that passes all of these tests may still suffer from misspecification—there is no silver bullet to detect all forms of misspecification! Furthermore, models of complex physical processes will always be somewhat deficient; as statistician George Box famously stated, 'all models are wrong, but some are useful' (Box 1976). Nonetheless, the tests recommended here provide a starting point for checking the suitability of models for the task at hand: that is, to provide a 'good enough' description of the data that meaningful inferences can be made. All of the demonstration code is available in jupyter notebook form here: tinyurl.com/bf4n9vw5. There is a dedicated notebook for each Section.

Broadly speaking, two different kinds of models are required to do an inference calculation. Every such calculation requires a model for the distribution of the data—the likelihood function—and a model for the distribution of the parameters—the prior. The remainder of this Article is organised as follows. In Section 2, we discuss the importance of data visualisation. In Section 3, we describe misspecification of the data: misspecified likelihood functions. In Section 4, we describe misspecification of population models: misspecified priors. In Section 5, we describe how apparent outliers may or may not be signs of model misspecification. We provide closing thoughts in Section 6.

2. Preface: The importance of visualisation

Below we describe many goodness-of-fit tests that can be used to determine the suitability of different models. However, even
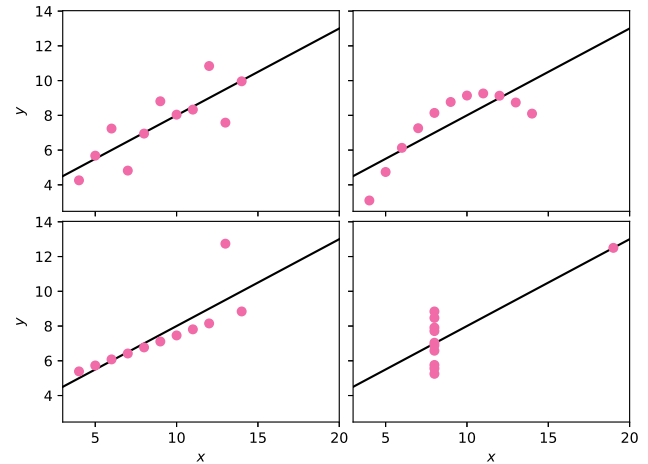

Figure 1. Anscombe's Quartet: a demonstration of the importance of data visualisation. While these datasets appear very different when plotted, they have identical summary statistics: mean $\bar{x} = 9$, $\bar{y} = 7.50$, sample variance $s_x^2 = 9$, $s_y^2 = 4.125 \pm 0.003$, $x - y$ correlation coefficient 0.816, linear regression line $y_R = 3.00 + 0.500 x_R$, and linear regression coefficient of determination $R^2 = 0.67$. An *Anscombe's Quartet* notebook is provided to demonstrate the calculation of these summary statistics for these datasets.

the most carefully crafted tests are no replacement for sanity checks with visualisation. Plotting data, we can sometimes see obvious model failures that we might not have thought to check *a priori*. The importance of visualisation is dramatically illustrated by Anscombe's Quartet (Anscombe 1973): four 11-point datasets with noticeably different trends that nonetheless have near-identical simple descriptive statistics.[a]

The four datasets that comprise Anscombe's Quartet are plotted in Figure 1. By studying these graphs, one can begin to diagnose anomalies: for example, two of the data sets each contain a single outlier that skews the correlation coefficient (lower left) or implies the existence of a relationship that is not supported by the rest of the data (lower right). We can also see that the dataset in the top right would be better-specified by a non-linear relationship between x and y. The Quartet is a cautionary tale to those who wish to establish the 'goodness' of their model: if one's model does not well-specify one's data, then the calculated 'goodness' metric is not to be trusted. The starting point for any exploration of misspecification, therefore, should be to visually compare the model and the data.

3. Likelihood misspecification

3.1. Basics

Models for the data are built on assumptions about the nature of both the noise and signal being measured. The data model is described by the likelihood function

$$\mathcal{L}(d|\theta), \qquad (1)$$

where d is the data and θ is a set of parameters describing the noise and/or signal.[b] The likelihood function is a normalised probability

---

[a]For a more recent and creative example, see the Datasaurus Dozen (Cairo 2016).
[b]Throughout, we follow the notation from Thrane & Talbot (2019).

density function *for the data*, not for the parameters θ (see Thrane & Talbot 2019, for more details):

$$\int d(d)\, \mathcal{L}(d|\theta) = 1, \tag{2}$$

$$\int d\theta\, \mathcal{L}(d|\theta) \neq 1. \tag{3}$$

It is useful to define a marginal likelihood, which is also known as the Bayesian evidence:

$$\mathcal{L}(d) = \int d\theta\, \mathcal{L}(d|\theta)\pi(\theta). \tag{4}$$

Here, $\pi(\theta)$ is the prior distribution for the parameters θ. The marginal likelihood is a model for the data, averaged over realisations of θ.

A well-known example of a model for gravitational-wave data is the Whittle likelihood model for Gaussian time-series noise, shown here for a single frequency bin:

$$\mathcal{L}(\tilde{d}|\theta) = \frac{1}{2\pi\sigma^2} e^{-|\tilde{d}|^2/2\sigma^2}. \tag{5}$$

Here, $\tilde{d}$ represents the frequency-domain gravitational-wave strain while σ^2 is related to the noise power spectral density (PSD) P and the frequency bin width Δf:

$$\sigma^2 = \frac{P}{4\Delta f}. \tag{6}$$

Equation (5) is an example of a *parameter-free* model of the data. If, for example, a gravitational-wave signal from a compact binary coalescence is present, then the likelihood depends on the $\gtrsim 15$ parameters associated with a compact binary coalescence (component masses, spins, etc.),

$$\mathcal{L}(\tilde{d}|\theta) = \prod_k \frac{1}{2\pi\sigma_k^2} e^{-\left|\tilde{d}_k - \tilde{h}_k(\theta)\right|^2/2\sigma_k^2}. \tag{7}$$

Here, $\tilde{h}_k(\theta)$ is a model for the frequency-domain strain from a gravitational-wave signal given binary parameters θ in frequency bin k. Sometimes, $h_k(\theta)$, which can be defined in either the time domain or the frequency domain, is referred to as 'the waveform model.'

Examining Equation (7), we can see various ways in which the likelihood can be misspecified, which we represent diagrammatically in Figure 2. First, the waveform $h(\theta)$ may be misspecified, which can lead to well-documented systematic error (Ohme et al. 2013; Wade et al. 2014; Ashton & Khan 2020; Gamba et al. 2021; Huang et al. 2021). This is an example of a misspecified signal model. Second, the noise model can be misspecified. There are typically two ways that this can happen. One possibility is that the functional form of the likelihood is correct, but the noise PSD is misspecified. A number of papers have proposed different methods for estimating the noise PSD in order to minimise this form of misspecification, for example, Littenberg & Cornish (2015), Cornish & Littenberg (2015), Chatziioannou et al. (2019).

The other possibility is that the functional form of the likelihood is itself misspecified. This may occur because of non-Gaussian noise artefacts (Röver, Meyer, & Christensen 2010) or uncertainty in the noise PSD (Talbot & Thrane 2020; Biscoveanu et al. 2020; Banagiri et al. 2020), both of which yield broader tails than the Whittle distribution. Likewise, marginalising over calibration uncertainty broadens the likelihood function (Sun et al.
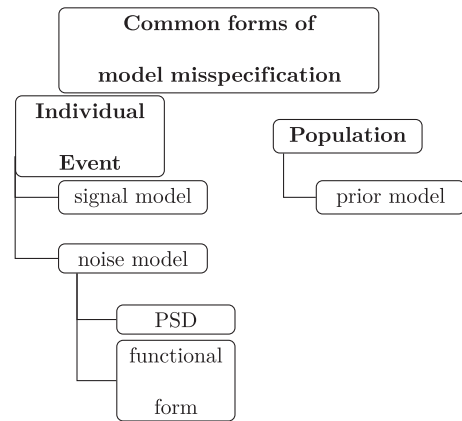

Figure 2. Forms of misspecification that we explore in this Article. Individual events can be misspecified if the model for the noise or the signal is not an adequate description of reality. The population of events may also be misspecified. This manifests itself as prior misspecification, which can impact both individual analyses (where the prior may be restricted to a limited portion of the true extent of the posterior) and population analyses (where the goal is to uncover the true distribution of the population).

2020; Payne et al. 2020; Vitale et al. 2021).[c] Covariance between neighbouring frequency bins induced from finite measurements of continuous noise processes can also lead to misspecification if not correctly accounted for (Talbot et al. 2021; Isi & Farr 2021). In the subsequent Subsections, we describe tests for these different forms of misspecification.

3.2. Testing for a misspecified signal model

In order to test for a misspecified waveform, it is sometimes useful to look at the whitened[d] residuals of the data in frequency

$$\tilde{r}(f|\theta) \equiv \frac{\tilde{d}(f) - \tilde{h}(f|\theta)}{\sigma(f)}, \tag{8}$$

and time

$$r(t|\theta) = \mathcal{F}^{-1}\left[\tilde{r}(f|\theta)\right], \tag{9}$$

where $\mathcal{F}^{-1}$ is the discrete inverse Fourier transform. Residuals can be useful to test for waveform misspecification because the differences between waveform models are clearly seen in the time and frequency domain.[e] Additionally, if there is a terrestrial noise artefact (glitch) present in the data, it is likely to appear clearly in the residuals. For example, in Abbott et al. (2016b), the best-fit, time-domain residuals for GW150914 were shown to be consistent with Gaussian noise (see Figure 1 of Abbott et al. 2016b), helping to show that the data are well explained by a gravitational waveform in Gaussian noise.

---

[c]Technically, calibration error is a form of *signal* misspecification in which the gravitational waveform $\lambda(f)\tilde{h}(f)$ includes a calibration correction $\lambda(f)$. However, marginalising over calibration uncertainty changes the functional form of the likelihood like the other examples in this list.

[d]Whitening is the process by which frequency-domain data $\tilde{d}(f)$ are normalised by the frequency-dependent noise $\tilde{d}(f) \to \tilde{d}(f)/\sigma(f)$.

[e]To see an example of residual analysis from optical astronomy, we direct the reader to the two-dimensional light intensity profiles in Weinzirl, Jogee, & Barazza (2008).
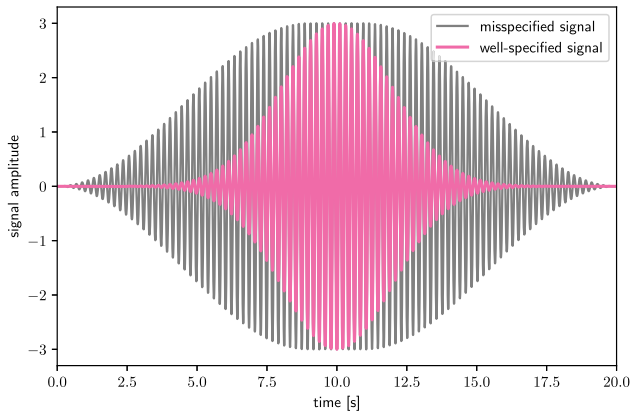
Figure 3. The correctly specified waveform (pink) and the misspecified waveform (grey) used in Section 3.2, plotted in the time domain.

Recommended tests:

- Plot the whitened residuals in both time $r(t)$ and frequency $|\tilde{r}(f)|$ and visually inspect for consistency with zero. In the frequency domain, the set of $\tilde{r}(f)$ are approximately independent measurements with Gaussian uncertainty equal to unity. Include residual curves for many posterior-distribution draws of θ in order to show theoretical uncertainty. A word of caution: the time-domain residuals are highly covariant in real data, and so it is less straightforward to interpret misspecification in the time domain than in the frequency domain.

- As a first step, look for consistency by eye. If there are signs of misspecification, one may quantify the inconsistency, for example, 'the residuals are five standard deviations away from zero at 500 Hz.' While post hoc analysis is helpful for catching egregious misspecification, in an ideal world, one should define the consistency tests *a priori* for unbiased tests. In practice, this is not always possible.

Demonstration:[f] Our signal model is a sine-Gaussian chirplet (i.e., a sine wave multiplied by a Gaussian function). We create two synthetic sets of data with Gaussian noise. The *correctly specified* data contains a signal that matches our model. The second dataset contains an intentionally *misspecified* signal: the same sine wave as before, but multiplied by a Tukey window. In both datasets, we assume Gaussian noise with a known power spectral density. In Figure 3, we plot these two simulated signals.

It is worth pausing to distinguish this discussion of model misspecification from the similar-but-different topic of model selection. If we were discussing model selection, we would keep the data fixed and compare it two different signal models. However, misspecification occurs when the analyst has not conceived of the correct model to test. Therefore, since we are discussing misspecification, we keep the model fixed and consider two hypothetical datasets: one correctly specified and one misspecified.

In the right-hand panel of Figure 4a, we display the time-domain residuals obtained when we subtract the sine-Gaussian waveform model from the misspecified data. In the left-hand panel, we show the time-domain residuals obtained from subtracting the same waveform model from the correctly specified data. If

the waveform is correctly specified, the residuals should be consistent with our Gaussian noise model as in the left panel. Although it is sometimes possible to see misspecification in the time-domain data (for example, when there is a short glitch), the misspecification may be more apparent in the frequency domain as is the case here.

In Figure 4b, we plot the amplitude spectral density of the residuals. Again, the left-hand panel shows the residuals for the correctly specified signal while the right-hand panel shows the residuals for the misspecified signal. While the correct waveform yields residuals consistent with the noise amplitude spectral density, the misspecified signal produces a peak in the amplitude spectral density that is conspicuously outside of the 90% range predicted by the model (shown in pink).

In order to assess the overall goodness of fit, it can be useful to plot the cumulative density function (CDF) of the residuals alongside the theoretical CDF predicted by the model. The CDF may be constructed from residuals in the time domain (to highlight transient phenomena), but more often, misspecification is most apparent using whitened frequency-domain residuals. An illustration of the CDF test is provided in Figure 4c, where we plot the CDF of the time-domain residuals. While there is only one realisation of the data (grey), we can generate arbitrarily many realisations of the theoretical CDF (pink). The thickness of the pink CDF shows the variability from 100 different realisations.(This is similar to, although not the same as, Gelman's posterior predictive checks (e.g., Gelman et al. 2013; Gelman & Rohilla Shalizi 2010). Gelman's checks involve drawing realisations from the histogram of the posterior probability distribution under the assumption of the model, and checking how probable it is for the model to produce realisations that are consistent with the observed data. This is something that we return to in Section 4. Here, we draw realisations from the noise model in order to establish the range over which it can credibly vary.)

To determine if the residuals agree with the model, one can employ any of the many established hypothesis-testing tools available to determine if measured samples are drawn from the theoretical distribution. For example, the Kolmogorov-Smirnov (KS) statistic,

$$D_n = \max_f \left| \mathrm{CDF}(f|\mathrm{measured}) - \mathrm{CDF}(f|\mathrm{predicted}) \right|, \quad (10)$$

is the maximum difference between the measured CDF of the data and the predicted CDF given n frequency bins. (The Anderson-Darling test is also commonly used to determine if measured samples are drawn from a predicted distribution.) The D_n statistic can be converted into a p-value with a look-up table that does not depend on the functional form of the CDF. The Kolmogorov-Smirnov p-values for the correctly specified data and misspecified dataset are provided in the legends of Figure 4c. The small p-value in the right panel suggests that our signal model is indeed misspecified.

Sometimes one may wish to inspect the data within a particular time or frequency interval. In this case, it can be enlightening to draw hundreds of distributions from the model, and count the number of times that the model CDF in this bin is above the CDF of the data. The fraction of draws above the CDF in this bin constitutes a p-value; if the data are correctly specified, the p-value will follow a uniform distribution. Thus, a p-value very close to 0 or 1 indicates that the data in this bin deviate significantly from the model. However, care must be taken if more than one p-value is

---

[f]*Misspecified signal model* notebook.
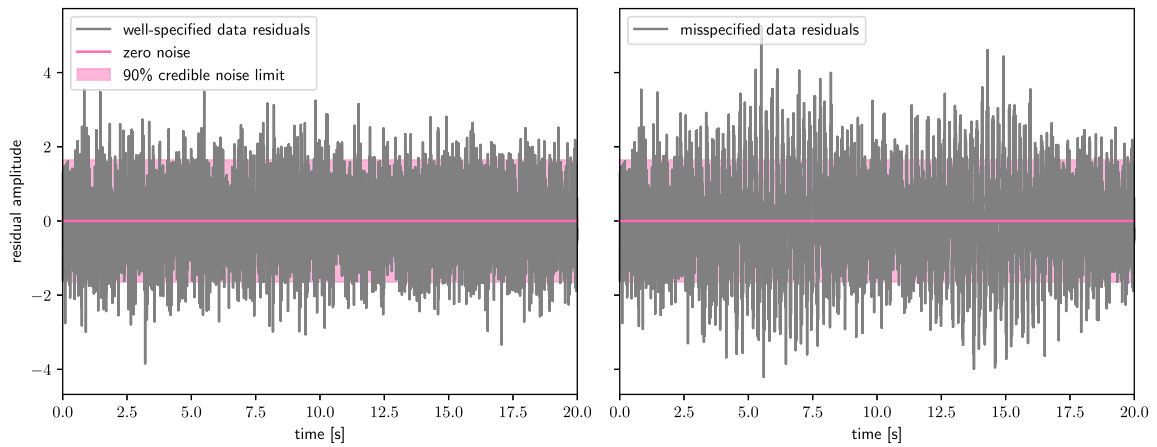
(a) Time series of the residuals calculated by subtracting two different waveform models from the simulated data. The plot on the left shows the residuals obtained by subtracting a correctly specified waveform that matches the signal hidden in the data, while those obtained by subtracting a misspecified waveform are on the right.
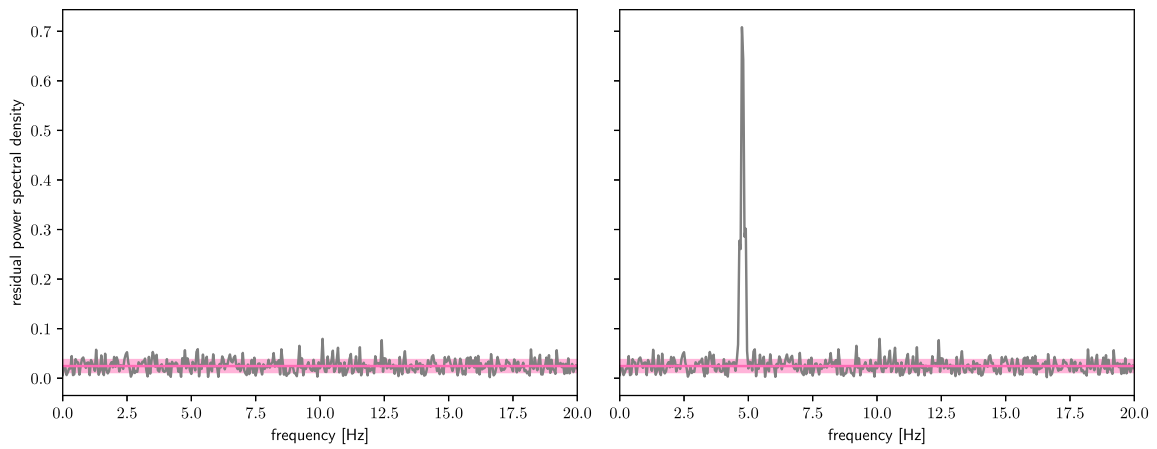
(b) Frequency-domain amplitude spectral densities of the residuals. It is not totally clear from the time domain data if one of the datasets is poorly specified by the model, but Fourier transforming the residuals reveals a suspicious peak inconsistent with Gaussian noise. On both rows, the residuals are plotted in grey while the pink band indicates the range where the model predicts 90% of the residuals will lie.
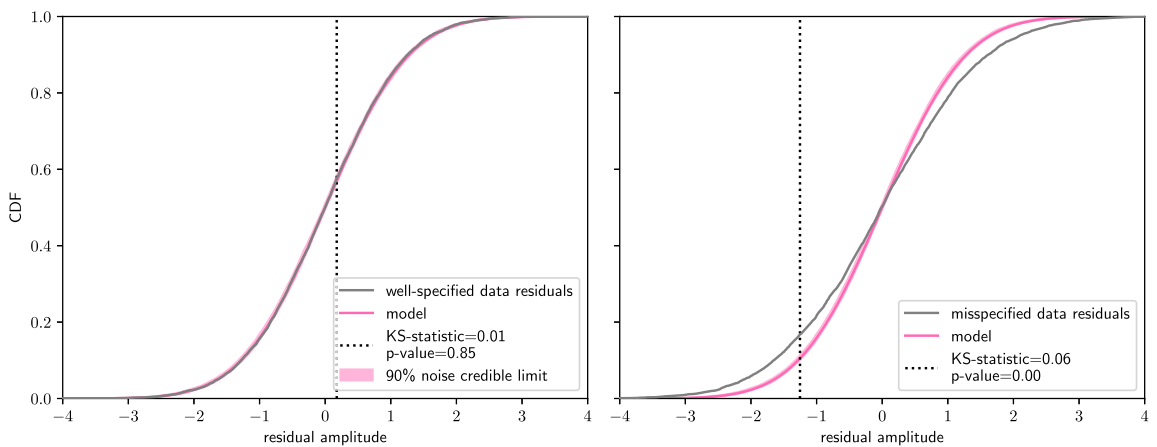
(c) Cumulative density functions (CDFs) for the residuals of the time-domain data (grey) and predicted range of the residuals (pink). We calculate the KS-statistic for the correctly specified and misspecified distributions. The location at which the KS test finds the maximum vertical distance between the model and the data is indicated with a dotted line. For this specific realisation of Gaussian noise, the KS-statistics are 0.01 and 0.06 for the correctly specified and misspecified data, with respective p-values of 0.85 and 0.00 (actually 8.25×10^{-17}).

Figure 4. Identifying a misspecified signal model. The left-hand column shows tests performed on data containing a signal consistent with the sine-Gaussian pulse model that we test against. The right-hand column shows the same tests performed on data containing a different signal.

calculated in the same set of data—while each *p*-value is uniformly distributed, the set of *p*-values from one dataset are correlated with each other.

Both of the tests described above assume that the predicted model is non-parametric. That is, the distribution of the residuals does not depend on any parameters. If the predicted distribution depends on one or more parameter θ, then we must rely on Monte Carlo methods to determine if the distributions agree. For example, we may calculate

$$D'_n = \max_f \left| \mathrm{CDF}(f|\mathrm{meas.}) - \mathrm{CDF}(f|\widehat{\theta}, \mathrm{pred.}) \right|, \qquad (11)$$

which is a KS-type statistic using the maximum-likelihood parameters $\widehat{\theta}$. Since we use the data to estimate $\widehat{\theta}$, D'_n is not distributed according to the Kolmogorov distribution. (It is easier to get smaller differences between the measured and predicted CDFs when the theoretical CDF varies depending on the parameters.) However, we can still calculate a *p*-value by generating synthetic data to empirically estimate the distribution of D'_n, which amounts to a generalisation of Lilliefors test (Lilliefors 1967). We demonstrate such a calculation below in Subsection 4.2 in the context of prior misspecification; see in particular Figure 7a.

3.3. Testing for a misspecified noise model

In order to observe noise misspecification and deviations from the Whittle likelihood, it is again useful to look at the distributions of residuals. By comparing the observed residuals to the theoretical likelihood distribution, it is possible to see, for example, if five-sigma deviations are more common than expected.

Recommended tests:

- Create a histogram representing the probability density function of the data; or, alternatively, plot the CDF of r, $|\tilde{r}|$, the whitened residuals. It is sometimes also useful to plot the distribution of the whitened residual power $|\tilde{r}|^2$; see Talbot & Thrane (2020), Chatziioannou et al. (2019). Include residual curves for many posterior distribution draws of θ in order to show theoretical uncertainty.

- For fixed models with no parameters, calculate a goodness-of-fit statistic using, for example, the KS test. When models have parameters, calculate the goodness of fit for the maximum-likelihood parameters.

- Plot the difference in the cumulative density functions (empirical - expected) as a function of the expected CDF as in Talbot & Thrane (2020), Chatziioannou et al. (2019). This is like a probability–probability ('*PP*') plot, in which the fraction of repeated measurements is plotted against the confidence level at which the known truth value exists, for data model checking.

- Bootstrapping methods—in which real data is used as a sampling distribution for synthetic data—can often be helpful for diagnosing noise misspecification. In gravitational-wave astronomy, new noise realisations can be created by bootstrapping data residuals (Cannon, Hanna, & Keppel 2013; Cannon, Hanna, & Peoples 2015; Ashton, Thrane, & Smith 2019) or using time slides (Dal Canton et al. 2014; Usman et al. 2016); both methods were integral to the first gravitational-wave detections (e.g., Abbott et al. 2016a). However, all bootstrap methods ultimately break down due to saturation effects

(Wąs et al. 2010); it is impossible to simulate all possible noise realisations using a finite dataset. In astro-particle physics, using instruments such as Super-Kamiokande, sidereal time scrambling is used to estimate typical fluctuations in signal strength due to noise (Thrane et al. 2009).

Demonstration:[g] We demonstrate how to diagnose noise misspecification. The true noise is Gaussian with a mean $\mu = 0$ and standard deviation $\sigma = 1$. The misspecified noise is distributed according to the Student's *t* distribution with $\nu = 5$ degrees of freedom. These parameters are chosen so that the noise profiles appear, at first glance, to be consistent with each other.

We display histograms of each noise dataset in Figure 5a. The region that the model predicts 90% of the data to lie within is also shown in these plots. The histograms both appear to largely lie within this 90% credible region, with the Student's *t*-distributed data only visibly deviating in the tails.

Next, we Fourier transform our datasets and compare the 90% range predicted by the noise model against histograms of the data in the frequency domain (Figure 5a). In the frequency domain, the misspecified data more clearly strays outside of the range predicted by the model. We then create a CDF of the frequency-domain data, comparing again to the 90% range predicted by the model (Figure 5b). We find a reasonable KS test *p*-value for the correctly specified data and an extreme *p*-value for the misspecified data, as expected. Finally, we compare the data to the model by plotting the data CDF against the difference between the data CDF and the model CDF (Figure 5c). In this final test, it is clear that the data is not well-represented by the model.

4. Prior misspecification

In the previous Section, we discussed various ways in which the likelihood can be misspecified and how we can detect this misspecification. Now we turn our attention to the misspecification of the *prior*, $\pi(\theta)$, a distribution that describes our prior knowledge of the parameters θ.

4.1. Priors with no hyper-parameters

In some cases, we can be very confident in our priors. For example, since there is no preferred direction in the Universe, the best prior for inclination angle ι (the angle between the orbital angular momentum and the line of sight) is uniform in $\cos \iota$. In other situations, we are not confident in the form of the prior distribution. In these cases, it can be useful to formalise this theoretical uncertainty using a conditional prior

$$\pi(\theta|\Lambda). \qquad (12)$$

Here, Λ is a 'hyper-parameter' we may vary to alter the shape of the prior for θ. In this Subsection, we focus on priors with no hyper-parameters while we cover parameterised priors in the next Subsection.

Recommended tests:

- Make a CDF plot comparing the reconstructed distribution of θ to the expected distribution of θ. In order to obtain a reconstructed distribution, obtain posterior samples for N different events, each *weighted with the population model to be tested.*

---

[g] *Misspecified noise model* notebook.
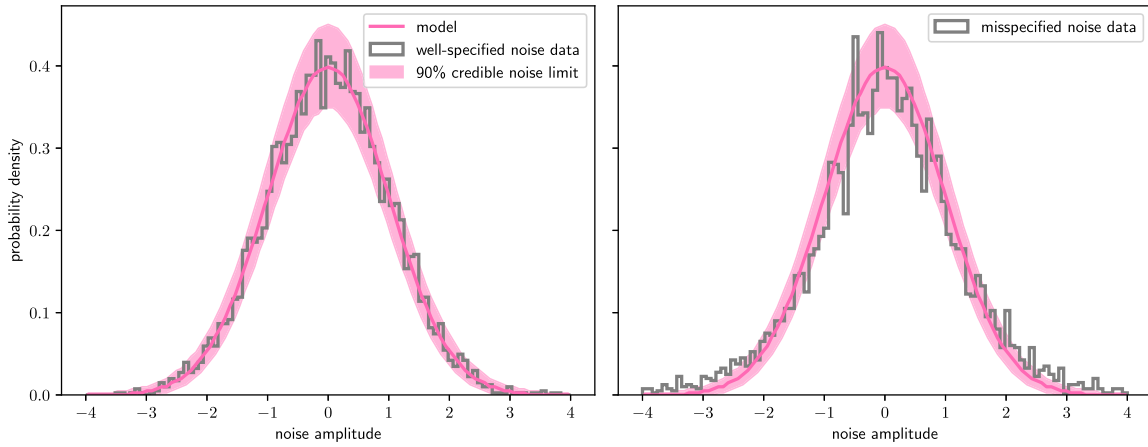
(a) Simulated noise distributed as a Gaussian in the frequency domain. The correctly specified Gaussian distribution (left) and the similar-but-misspecified Student's t distribution (right). The predicted (90% credible) range predicted by the model is shown by the pink band.
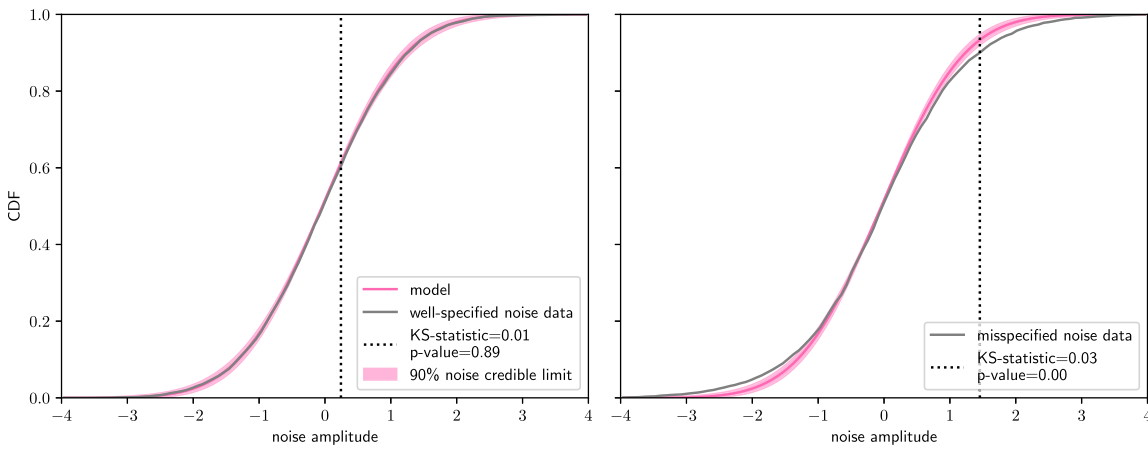
(b) Cumulative density functions (CDFs) for frequency-domain residuals (grey) and the range predicted by the model (pink). For these specific noise realisations, the KS-statistics are 0.01 and 0.03 for the correctly specified and misspecified models, with p-values of 0.89 and 3.03×10^{-6} respectively. As in Figure 4c, the location of the maximal KS distance is noted by a dotted line.
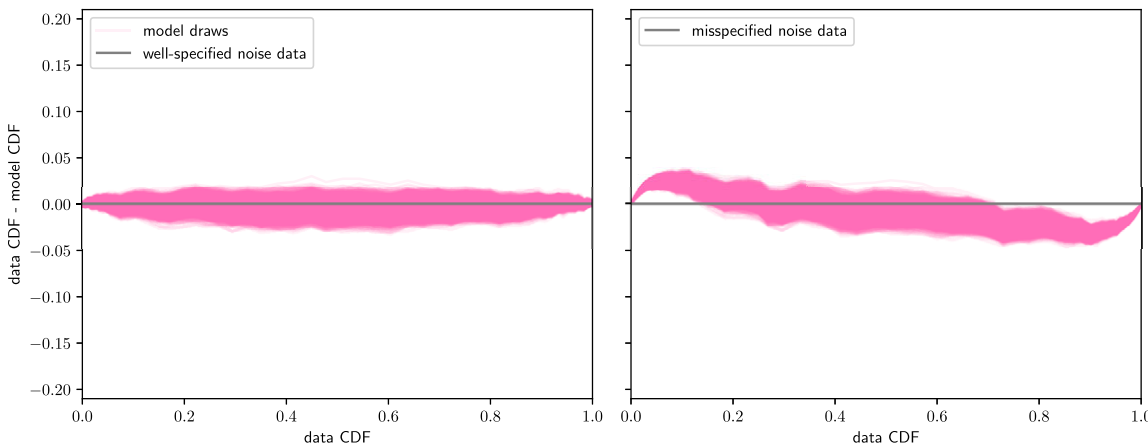
(c) Like the previous row, but the *difference* in data and model CDFs as a function of the data CDF.

Figure 5. Identifying a misspecified noise model. The left-hand column shows tests performed with a correctly specified Gaussian noise model while the right-hand column shows the same tests with the same Gaussian model, but performed against a misspecified Student's-t distribution.

Draw one posterior sample from each of the N events to make a realisation of the reconstructed distribution. Do this many times to make many realisations. Plot CDFs of the many realisations alongside the population model being tested. If the data agrees with the model, the CDFs should overlap. If the two CDFs do not overlap, the model is a poor description of the data. It is worth noting that a model can pass this test while still being quite badly misspecified. Abbott et al. (2021e) found evidence for negatively aligned spins in a population of compact-object binaries, but this was later shown to be a model-dependent feature (Roulet et al. 2021; Galaudage et al. 2021). This test may be particularly unreliable if there is a sharp feature in the data that is not captured by the prior.

- If a more quantitative test is desired, one may calculate the KS-statistic for each CDF. If the best-fitting reconstruction has a p-value below a certain threshold, for example p-value ≤ 0.00005 (Benjamin et al. 2017), then the model is not a good fit to the data.

- One may also identify problem areas of model under- or over-production by quantifying the discrepancies between the data draw CDFs and the model over the parameter space. This facilitates statements like '99% of the time, the model produces too many binary black hole events with $m_1 > 80\,\mathrm{M}_\odot$'.

Demonstration.[h] We simulate data d consisting of some physical parameter x and noise n:

$$d = x + n. \tag{13}$$

The noise is Gaussian distributed with zero-mean and unit variance. The values of x are drawn from the true prior distribution: a Gaussian distribution with mean $\mu = 10$ and width $\sigma = 2$. However, in order to demonstrate prior misspecification, we employ as our prior a Laplacian with the same mean and variance as the true distribution. The true prior distribution and the misspecified prior distribution are compared in Figure 6a.

We create a simulated dataset of $N = 1000$ events. In each case, we construct Gaussian likelihood curves of mean $\mu_i = d_i$. The posterior for each value of x_i is proportional to the likelihood multiplied by the prior. We draw 50 posterior samples from each of these simulated posteriors in order to create 50 CDF curves; see Figure 6b. We make two versions of this plot: one with the correctly specified prior and one with the misspecified prior.

It is also sometimes useful to plot the data CDF minus the model CDF. We include an example of this plot in Figure 6c. In this case, we see that the misspecified prior yields disagreements in the CDF in the distribution tails; the difference between pink model and grey data is, in general, inconsistent with zero when the model CDF is ≈ 0.1 and ≈ 0.9.

4.2. *Hyper-parameterised priors*

We are particularly concerned with population models (sometimes called 'hierarchical models'), where the prior for the parameters θ is conditional on a set of hyper-parameters Λ, describing the shape of the prior distribution:

$$\pi(\theta|\Lambda). \tag{14}$$

While data models are built on assumptions about the nature of noise, population models are built on assumptions about astrophysics. For example, if we assume that the distribution of primary black hole mass m_1 follows a power-law distribution with spectral index α,

$$\pi(m_1|\alpha) \propto m_1^\alpha, \tag{15}$$

then $\alpha \in \Lambda$ is a hyper-parameter for the distribution of $m_1 \in \theta$.

Hyper-parameters are frequently a convenient way of describing systematic theoretical uncertainty. For example, consider a parameter x drawn from a Gaussian distribution $\mathcal{N}(x|\mu, \sigma)$ with mean μ and width σ. If we do not know the precise value of μ or σ, our misspecification tests should take this uncertainty into account. We repeat the tests from the previous Subsection for the case of a hyper-parameterised prior.

Here is a summary of the differences that arise from the addition of a hyper-parameter. One must generate random draws of the hyper-parameter Λ. Then generate draws of the model CDF for each random draw of Λ, and weight posterior samples using the population predictive distribution—the conditional prior, marginalised over uncertainty in the hyper-parameters:

$$\mathrm{ppd}(\theta) = \int d\Lambda\, \pi(\theta|\Lambda) p(\Lambda|d). \tag{16}$$

Here $p(\Lambda|d)$ is the posterior for the hyper-parameters. Take care not to double-count; the posterior for Λ used to reweight event k should not be informed by event k; see, for example, Essick et al. (2022).

Demonstration:[i] We assume a Gaussian prior for parameter x with uncertain μ and width σ. We simulate $N = 1000$ true events, x_{true}, from two populations (priors): one Gaussian-distributed and one Laplacian-distributed, both with the same means μ_G, and with widths σ_G and σ_L, respectively. For each population, we calculate the maximum-likelihood detected value of each event, x_{meas} by offsetting it by a random number drawn from a Gaussian of the same width as the likelihood distribution, σ_{meas}. The mean average of these maximum-likelihood values is the maximum-likelihood estimate for the population prior mean, μ_E, and the measured population prior width is $\sigma_E = \sigma_{\mathrm{meas}}/\sqrt{N}$. The estimated mean of the population posterior is described by

$$\mu_P = \frac{\sigma_E^{-2}\mu_E + \sigma^{-2}\mu}{(\sigma_E^{-2} + \sigma^{-2})} = \frac{\mu_E}{\sigma_E^2(\sigma_E^{-2} + \sigma^{-2})}, \tag{17}$$

since $\mu = 0$. The estimated width is

$$\sigma_P = (\sigma_E^{-2} + \sigma^{-2})^{-\frac{1}{2}}. \tag{18}$$

The width of the posterior *predictive* distribution is

$$\sigma_{\mathrm{pp}} = \sqrt{\sigma_P^2 + \sigma^2}, \tag{19}$$

the width of a cross product of the uncertainty distribution of width σ and the posterior width σ_P. The population-weighted width for each individual event's posterior is

$$\sigma_x = (\sigma_E^{-2} + \sigma_{\mathrm{pp}}^{-2})^{-\frac{1}{2}}. \tag{20}$$

We draw $M = 100$ samples from each of these posteriors of mean x_{meas} and width σ_x. In practice, we do this by drawing N samples

---

[h] *Misspecified prior: unparameterised case* notebook.

[i] *Misspecified prior: parameterised case* notebook.
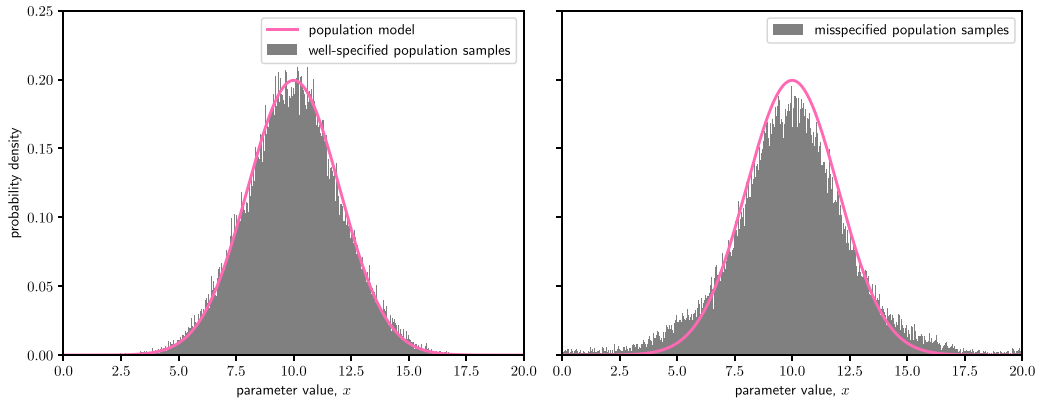
(a) The distribution of the parameter x. Simulated posterior samples are grey while the prior distribution used for the construction of the posterior is shown in pink. On the left, the (Gaussian) prior is correctly specified, and on the right, the Laplacian prior is misspecified.
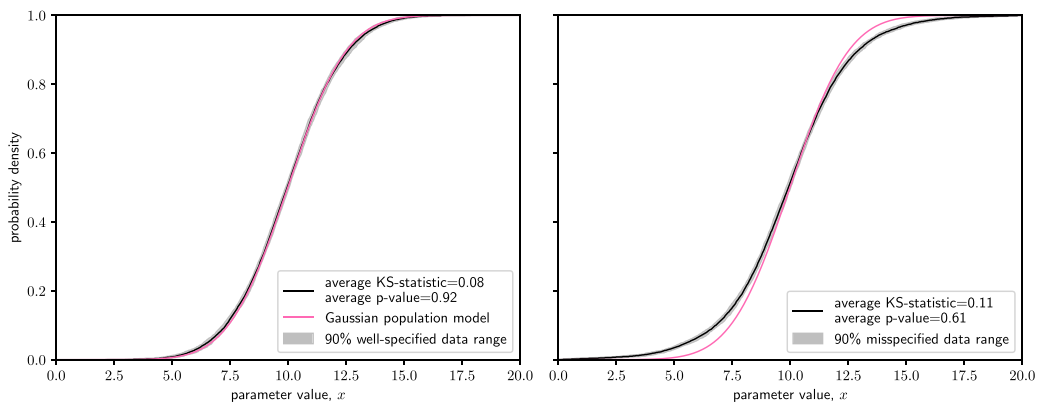
(b) Cumulative distribution of the parameter x. Pink (observed) shows the 90% interval for CDFs using draws from the posterior samples. Grey (predicted) is 90% predicted range of the CDF of the prior distribution used to generate the posterior samples. The misspecified data (right-hand panel) visibly disagrees with the model. The p-value for the average of the draws from the CDF is relatively low and the corresponding KS-statistic of the average CDF is relatively high; together with the visible deviation of the model from the data, these values indicate that the model does not well-specify this data.
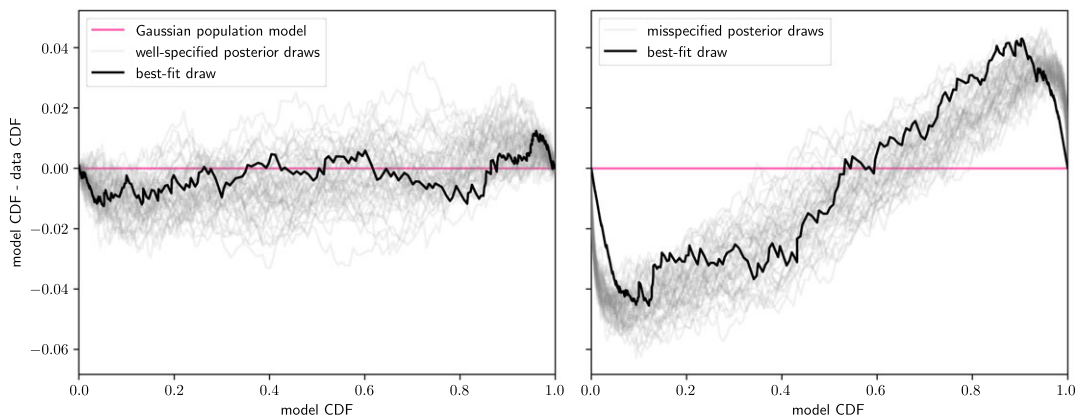
(c) The difference between the observed and predicted CDFs as a function of the predicted CDF. The difference is inconsistent with zero when the predicted CDF is ≈ 0.1 and ≈ 0.9

Figure 6. Plots illustrating how model misspecification may manifest for an unparameterised population model, with a correctly specified dataset shown in the left-hand plots, and the misspecified case in the right-hand plots.

M times from the population posterior. The CDFs of these draws can then be compared to the model CDF.

Since we have used the data to estimate the hyper-parameters (μ, σ), we cannot use the standard KS test to ascertain the degree of misspecification. Instead, we must define our own KS-like test for the p-value using the distance between the model and data. We calculate 100 KS distances between the model and draws from the model. We can histogram this data to show the distribution of KS

distances expected if the model is a good fit to the data; see the pink histogram in Figure 7a. We measure the KS distance between the average data realisation and the data-conditioned model, which is shown by a grey line in Figure 7a. The *p*-value equates to the fraction of the model KS distance distribution above the data KS distance. In Figure 7b, we plot the 90% credible bands of the parameterised model and compare against the 90% credible range of the data. Despite being conditioned on the same data, the misspecified model strays outside of the credible band of the data and achieves a *p*-value of only 0.01. Finally, in Figure, we show the same CDFs with the CDF of the model subtracted, and plot against the model CDF. This has the effect of emphasising the extent of the mismatch between the data and the misspecified model, while the well-specified model contains the median data draw over the entire range.

4.3. Models with more than one dimension

It is difficult to look for model misspecification in more than one or two dimensions. Tests like those described in this Article so far can be extended, but the tell-tale signs of misspecification (e.g., detectable structure in residuals) can be difficult to see in large-dimensional spaces. However, there are some tools available.

- Make two-dimensional scatter plots. If the model does not have any parameters, it can be represented by contours while the data can be represented with two-dimensional error bars or credible intervals.
- If the model is subject to theoretical uncertainty (i.e., it is described by hyper-parameters), it may be necessary to draw multiple contours. However, the plot can be difficult to read if there are more than two variables.
- There is no standardised test for goodness of fit in two or more dimensions. Boutique tests designed to identify particular forms of misspecification can work well, but one must beware trial-factor penalties when designing the test after looking at the data.

5. More symptoms of model misspecification

5.1. Posterior stability

In the previous Sections we describe tests to find misspecification. However, misspecification is sometimes manifest from surprising results. One such example is a phenomenon we refer to as 'posterior instability.' Let us consider the posterior for some parameter (call it θ) and imagine how this posterior changes as we accumulate data. On average, we expect the posterior to narrow as we include more information. It also tends to shift around, but only within the bounds of the previous credible regions. It would be surprising if the addition of data produced a posterior favouring $\theta = 3$ when an earlier posterior (calculated with less data) disfavoured $\theta = 3$ with high credibility.

In such cases we say that the posterior is not stable to the addition of data. This can be indicative of model misspecification. An example from gravitational-wave astronomy is the maximum black hole mass parameter, which proved to be unstable moving from GWTC-1 (Abbott et al. 2019b) to GWTC-2 (Abbott et al. 2020a), under the assumption that the primary black hole mass distribution is a power law with a sharp cut-off. When the mass model was improved to allow for additional features (deviations

from a power law Talbot & Thrane 2018), the maximum-mass parameter stabilised.[j] For another example of posterior instability from optical astronomy, see Liu et al. (2018), Zhu & Thrane (2020).

5.2. Outliers

Outliers are symptom of misspecification closely related to posterior stability (Fishbach, Farr, & Holz 2020; Essick et al. 2022). An outlier is an event with a parameter value appearing inconsistent with the rest of the distribution—in the context of a particular population model. Let us imagine that the distribution of events in our dataset (characterised by parameter θ) seem well-described by a normal prior with mean zero and unit variance: $\mathcal{N}(\theta|\mu = 0, \sigma = 1)$. If we subsequently observed an event with $\theta \gtrsim 10$, this could indicate that our prior model (a normal distribution) is inadequate.

Commonly, population outliers are identified using a 'leave-one-out' analysis method that compares an inferred distribution with and without the potentially anomalous data point; see Fishbach et al. (2020). However, care must be taken to take into account trial factors[k] since, in any catalogue, *some* event has to be the most extreme. The statistics required for a careful leave-one-out analysis are too complicated for us to summarise here. Instead we refer the reader to Essick et al. (2022), which describes a 'coarse-graining' method to identify outliers.

6. Discussion: Living with misspecification

Misspecified models can lead to flawed inferences. While it is not always practical, models should, when possible, be subjected to an array of visualisations and checks for misspecification. In an ideal world, models found to be misspecified should be improved so that they pass these checks. In particular, one may *refine* one's model in order to better capture features of the data (e.g., Gao & Ho 2017; Gabry et al. 2017), although this kind of post hoc data fitting can cause obvious biases in inferred results. Iteratively adding complexity can also cause computational costs to skyrocket, and each addition has the potential to add further misspecification.

Thus, in practice, model refinement is not always possible. For example, as data becomes more informative at higher signal-to-noise ratios (SNR), even subtle imperfections can lead to signs of misspecification. For example, if one attempts to fit a template to an optical image of a distant (but clearly resolved) galaxy, there will *always* be non-negligible residuals because our best templates are no match for the high signal-to-noise ratio of optical astronomy data.

One option in such cases is to apply *coarsening* to blur the data. For example, one may employ a coarsened posterior, conditional on the distance between the model and the data being below some threshold, but not zero (see Miller & Dunson 2019, for a detailed description of posterior coarsening). An alternative form of coarsening is to add a non-parameteric *error term* to the revised model to account for unknown or independent influences in the data; for example, Bhatt et al. (2017) use a Gaussian random field to generate multiplicative factors to the terms in their parameterised model for malaria mapping.

---

[j]Another example: many inferences in Abbott et al. (2020a) appear to be unstable with respect to the inclusion of the extreme mass ratio event GW190814, suggesting that the models in that work are not adequate to accommodate this event.

[k]Some literature refers to the 'look elsewhere effect,' for example, Gross & Vitells (2010).
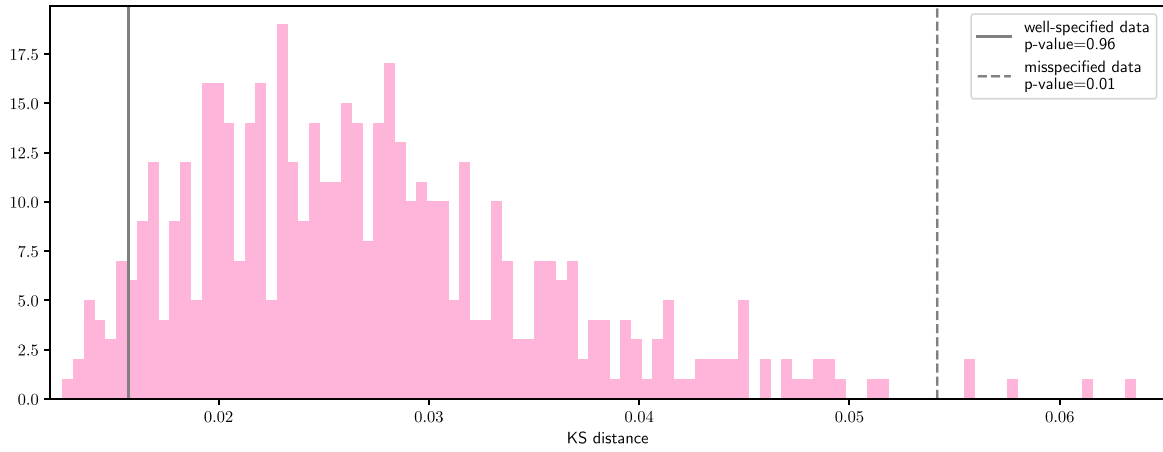
(a) Distribution of KS distances drawn from realisations of the data-conditioned model (pink). We use these to compute a KS-like *p*-value by calculating the fraction of the distribution above the KS distance calculated for the data (grey).
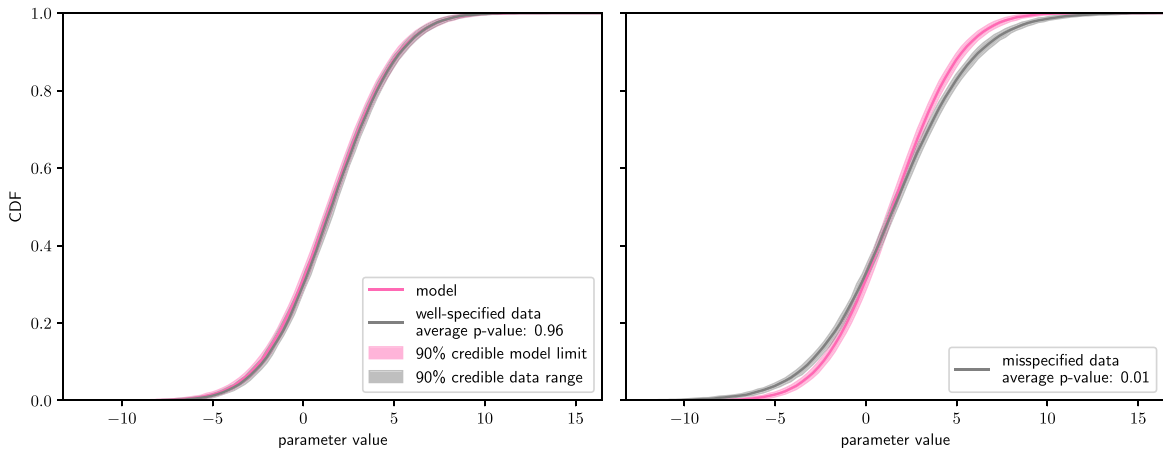
(b) Parameterised model (pink) and the data (grey); the width of both translucent bands represents the 90% credible interval around that quantity, with the uncertainty coming from the unknown (but estimated) hyper-parameter μ. The misspecified data on the right-hand side clearly strays outside of the 90% credible bands of the model, even though the model was conditioned on the same data, because the model does not describe the data well.
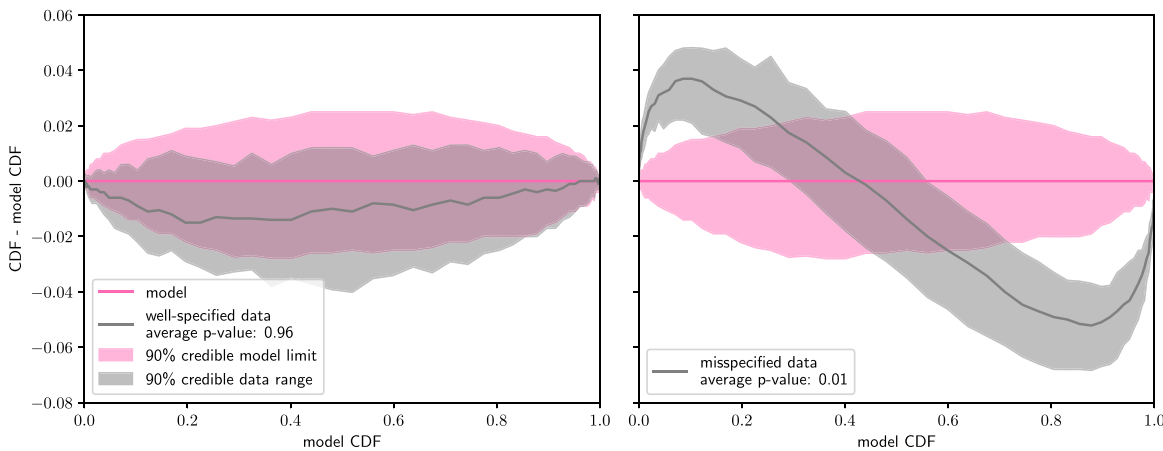
(c) The difference in data and model CDFs as a function of the model CDF. The median of the correctly specified data (left) stays within the band of 90% model uncertainty, but the misspecified data clearly deviates.

Figure 7. Model misspecification for a parameterised population model, with the correctly specified case demonstrated in the left-hand side of the lower two plots, and the misspecified case demonstrated on the right.

While the low-SNR regime has its own inference challenges, it can be more forgiving when confronted with misspecification. For example, the inclusion of nuisance parameters has helped to understand high-energy gamma-ray emission from the galactic centre (Storm, Weniger, & Calore 2017; Armand & Calore 2021; Bartels et al. 2018), resolve temporally varying structure in the brightness profile of the accretion disc around a supermassive black hole (Arras et al. 2022) and to assess the existence of a correlation between COVID-19 viral load and age (Guardiani et al. 2021).

Although we have described a number of ways in which models can be tested for misspecification, one must ultimately accept that all physical models are—to some degree—misspecified (Box 1976). The question, therefore, is not whether a model is wrong, but whether it is *adequate*. If a model does a 'good enough' job of describing a signal (that is, it is not *obviously misspecified*), then we may still able to use it to inform us about the Universe.

References

Abbott, B. P., et al. 2016a, PhRvD, 93, 122003
Abbott, B. P., et al. 2016b, PhRvL, 116, 061102
Abbott, B. P., et al. 2017, PhRvL, 119, 161101
Abbott, B. P., et al. 2019a, PhRvX, 9, 031040
Abbott, B. P., et al. 2019b, ApJL, 882, L24
Abbott, R., et al. 2020a, Accepted in ApJL.
Abbott, B. P., et al. 2020b, CQG, 37, 045006
Abbott, R., et al. 2020c, PhRvL, 125, 101102
Abbott, R., et al. 2021a, arXiv e-prints, p. arXiv:2108.01045
Abbott, R., et al. 2021b, arXiv e-prints, p. arXiv:2111.03606
Abbott, R., et al. 2021c, arXiv e-prints, p. arXiv:2111.03634
Abbott, R., et al. 2021d, PhRvX, 11, 021053
Abbott, R., et al. 2021e, ApJ, 913, L7
Andreon, S., & Weaver, B. 2015 (1st edn.; Springer, Switzerland), p. 191–205, DOI: https://doi.org/10.1007/978-3-319-15287-5
Anscombe, F. J. 1973, AS, 27, 17
Armand, C., & Calore, F. 2021, PhRvD, 103, 083023
Arras, P., Frank, P., Haim, P., Knollmüller, J., Leike, R., Reinecke, M., & Enßlin, T. 2022, NatAs, 6, 259
Ashton, G., & Khan, S. 2020, PhRvD, 101, 064037
Ashton, G., Thrane, E., & Smith R. J. E. 2019, PhRvD, 100, 123018
Banagiri, S., Coughlin, M. W., Clark, J., Lasky, P. D., Bizouard, M. A., Talbot, C., Thrane, E., & Mandic V. 2020, MNRAS, 492, 4945
Bartels, R., Storm, E., Weniger, C., & Calore, F. 2018, NatAs, 2, 819
Benjamin, D., et al. 2017, NatHB, 2
Betancourt, M. 2020, Towards A Principled Bayesian Workflow, https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html
Bhatt, S., Cameron, E., Flaxman, S., Weiss, D. J., Smith, D. L., & Gething, P. W. 2017, JRSI, 14
Biscoveanu, S., Haster, C.-J., Vitale, S., & Davies, J. 2020, PhRvD, 102, 023008
Box, G. E. P. 1976, JASA, 71, 791
Bustillo, J. C., et al. 2021, PhRvL, 126, 081101
Cairo, A. 2016, Download the Datasaurus: Never trust summary statistics alone; always visualize your data, http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

Cannon, K., Hanna, C., & Keppel, D. 2013, PhRvD, 88, 024025
Cannon, K., Hanna, C., & Peoples, J. 2015, arXiv e-prints, p. arXiv:1504.04632
Chatziioannou, K., Haster, C.-J., Littenberg, T. B., Farr, W. M., Ghonge, S., Millhouse, M., Clark, J. A., & Cornish, N. 2019, PhRvD, 100, 104004
Cornish, N. J., & Littenberg, T. B. 2015, CQG, 32, 135012
Dal Canton, T., et al. 2014, PhRvD, 90, 082004
Essick, R., Farah, A., Galaudage, S., Talbot, C., Fishbach, M., Thrane, E., & Holz, D. E., 2022, ApJ, 926, 34
Fishbach, M., Farr, W. M., & Holz, D. E. 2020, ApJ, 891, L31
Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. 2017, arXiv e-prints, p. arXiv:1709.01449
Galaudage, S., Talbot, C., Nagar, T., Jain, D., Thrane, E., & Mandel, I. 2021, ApJL, 921, L15
Gamba, R., Breschi, M., Bernuzzi, S., Agathos, M., & Nagar, A. 2021, PhRvD, 103, 124015
Gao, H., & Ho, L. C. 2017, ApJ, 845, 114
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. 2013, Bayesian Data Analysis (3rd edn.; Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis), https://books.google.com.au/books?id=ZXL6AQAAQBAJ
Gelman, A., & Rohilla Shalizi, C. 2010, arXiv e-prints, p. arXiv:1006.3868
Gelman, A., et al. 2020, arXiv e-prints, p. arXiv:2011.01808
Ghosh, A. 2022, arXiv e-prints, p. arXiv:2204.00662
Gross, E., & Vitells O. 2010, EPhJC, 70, 525
Guardiani, M., Frank, P., Kostić, A., Edenhofer, G., Roth, J., Uhlmann, B., & Enßlin, T. 2021, arXiv e-prints, p. arXiv:2105.13483
Huang, Y., Haster, C.-J., Vitale, S., Varma, V., Foucart, F., & Biscoveanu, S., 2021, PhRvD, 103, 083001
Isi, M., & Farr, W. M. 2021, arXiv e-prints, p. arXiv:2107.05609
Lilliefors, H. W. 1967, JASA, 318, 399
Littenberg, T. B., & Cornish, N. J. 2015, PhRvD, 91, 084034
Liu, T., Gezari, S., & Miller, M. C. 2018, ApJL, 859, L12
Miller, J. W., & Dunson, D. B. 2019, JASA, 114, 1113
Ohme, F., Nielsen, A. B., Keppel, D., & Lundgren, A. 2013, PhRvD, 88, 042002
Olsen, S., Venumadhav, T., Mushkin, J., Roulet, J., Zackay, B., & Zaldarriaga, M. 2022, arXiv e-prints, p. arXiv:2201.02252
Payne, E., Talbot, C., Lasky, P. D., Thrane, E., & Kissel, J. S. 2020, PhRvD, 102, 122004
Romero-Shaw, I. M., Lasky, P. D., Thrane, E., & Calderón Bustillo, J. 2020, ApJL, 903, L5
Roulet, J., Chia, H. S., Olsen, S., Dai, L., Venumadhav, T., Zackay, B., & Zaldarriaga, M. 2021, PhRvD, 104, 083010
Röver, C., Meyer, R., & Christensen, N. 2010, CQG, 28, 015010
Storm, E., Weniger, C., & Calore, F. 2017, JCAP, 2017, 022
Sun, L., et al. 2020, CQG, 37, 225008
Talbot, C., & Thrane, E. 2018, ApJ, 856, 173
Talbot, C., & Thrane, E. 2020, PhRvR, 2, 043298
Talbot, C., Thrane, E., Biscoveanu, S., & Smith, R. 2021, PhRvR, 3, 043049
Thomas, O., & Corander, J. 2019, arXiv e-prints, p. arXiv:1912.05810
Thrane, E., & Talbot, C. 2019, PASA, 36, E010
Thrane, E., et al. 2009, ApJ, 704, 503
Usman, S. A., et al. 2016, CQG, 33, 215004
Venumadhav, T., Zackay, B., Roulet, J., Dai, L., & Zaldarriaga, M. 2019, PhRvD, 100, 023011
Venumadhav, T., Zackay, B., Roulet, J., Dai, L., & Zaldarriaga, M. 2020, PhRvD, 101, 083030
Vitale, S., Haster, C.-J., Sun, L., Farr, B., Goetz, E., Kissel, J., & Cahillane, C. 2021, PhRvD, 103, 063016
Wade, L., Creighton, J. D. E., Ochsner, E., Lackey, B. D., Farr, B. F., Littenberg, T. B., & Raymond, V. 2014, Phys. Rev. D, 89, 103012
Wąs, M., et al. 2010, CQG, 27, 015005
Weinzirl, T., Jogee, S., & Barazza, F. D. 2008, in Astronomical Society of the Pacific Conference Series, Vol. 393, New Horizons in Astronomy, ed. A. Frebel, J. R. Maund, J. Shen, & M. H. Siegel, 279 (arXiv:0802.3903)
Zhu, X.-J., & Thrane, E. 2020, ApJ, 900, 117