When Models Fail: An Introduction to posterior predictive checks and model misspecification in GW astronomy

Romero Shaw, Thrane & Lasky (2022)

Pinaki Roy[1]

[1]Ph.D. Astronomy, University of Warsaw

Statistics Journal Club

Outline

Bayesian inference (BI)

- BI is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Essentially BI uses prior knowledge, in the form of a prior distribution in order to estimate posterior probabilities.

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

$H \rightarrow$ hypothesis / model; $\quad P(H) \rightarrow$ prior probability
$E \rightarrow$ data / evidence (not used in computing prior)
$P(H|E) \rightarrow$ posterior probability; $\quad P(E|H) \rightarrow$ likelihood

- BI is a powerful tool in GW astronomy. It helps to deduce the properties of merging compact-object binaries and determine how these mergers are distributed as a population according to mass, spin and redshift.

Article purpose (abstract)

It discusses the phenomenon of **model misspecification**, in which results obtained with BI are misleading due to deficiencies in the assumed model.

Such deficiencies can impede our inferences of the true parameters describing physical systems. They can reduce our ability to distinguish the 'best fitting' model. There are broadly two ways in which models fail.

Firstly, models that fail to adequately describe the data (either the signal or the noise) have **misspecified likelihoods**.

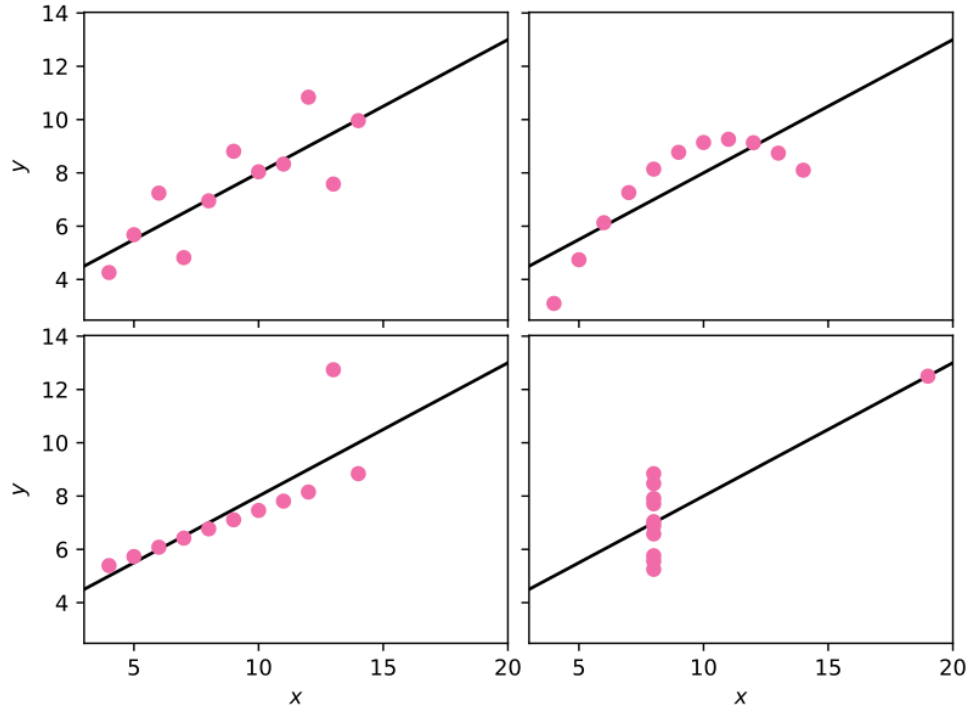
Secondly, population models—designed, for example, to describe the distribution of black hole masses—may fail to adequately describe the true population due to a **misspecified prior**.

It recommends **tests and checks** that are useful for spotting misspecified models using examples inspired by GW astronomy.

Article motivation

- With the increase in the detector sensitivity and growth of the GW event catalog, more events are seen that challenge the existing 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 SNR $= 30$ may be inadequate for an SNR $= 100$ signal.
- 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.

Goodness of a model (visualisation)



Anscombe's Quartet (Anscombe 1973): same mean ($\bar{x}$, $\bar{y}$), variance (s_x^2, s_y^2), linear regression line and linear regression coefficient. Bottom two contain outliers that disrupts the model established by rest of the data. The top right suggests a non-linear relation between x and y. Only the first fit is justified.

Likelihood function

Two different kinds of models are required to do an inference calculation: a model for the distribution of the data – the **likelihood** function – and a model for the distribution of the parameters – the **prior**.

Likelihood function: $\mathcal{L}(d|\theta)$

where d is the data and θ is a set of parameters describing the noise and/or signal. The likelihood function is a normalised probability density function for the data, not for the parameters θ.

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

Marginal likelihood is defined as

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

where $\pi(d|\theta)$ is the prior distribution for the parameters θ.

Likelihood function (contd.)

A common model for gravitational-wave data is the Whittle likelihood model for Gaussian time-series noise (parameter-free):

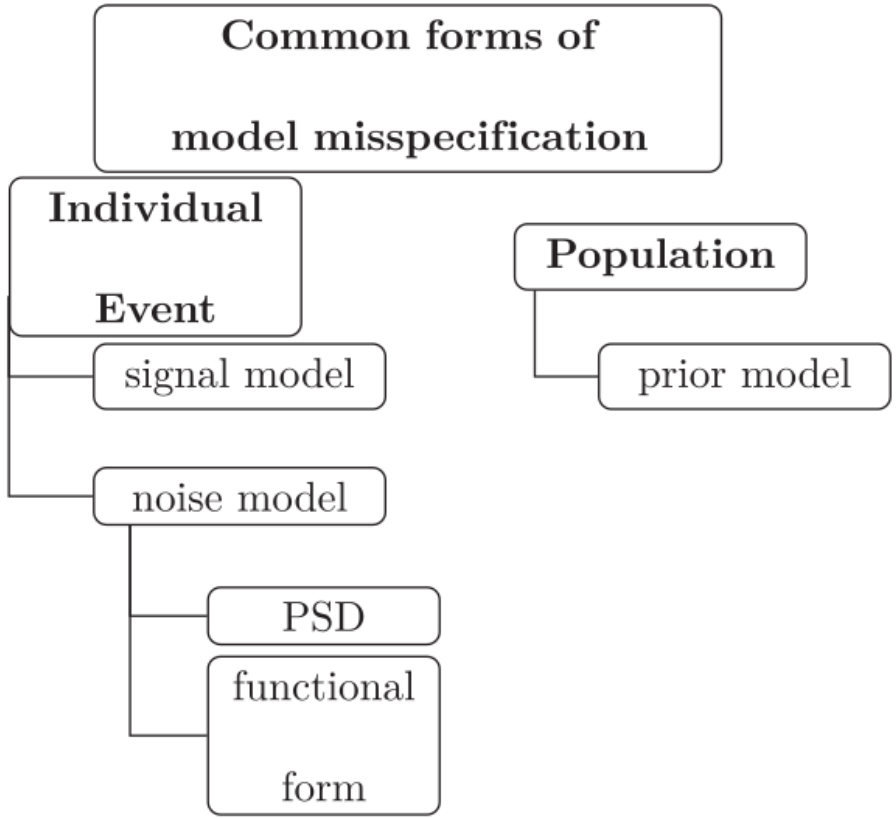
$$\mathcal{L}(\tilde{d}|\theta) = \frac{1}{2\pi\sigma^2} e^{-|\tilde{d}|^2/2\sigma^2}$$

where $\tilde{d}$ 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 as

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

For compact binary coalescence, the likelihood depends on $\gtrsim 15$ parameters (component masses, spins, etc.) and is given by

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



Forms of misspecification explored in the 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 and population analyses (where the goal is to uncover the true distribution of the population).

Testing for a misspecified signal model

To test for a misspecified waveform, it is useful to look at the whitened residuals of the data in frequency

$$\tilde{r}(f|\theta) = \frac{\tilde{d}(f) - \tilde{h}(f|\theta)}{\sigma(f)}$$

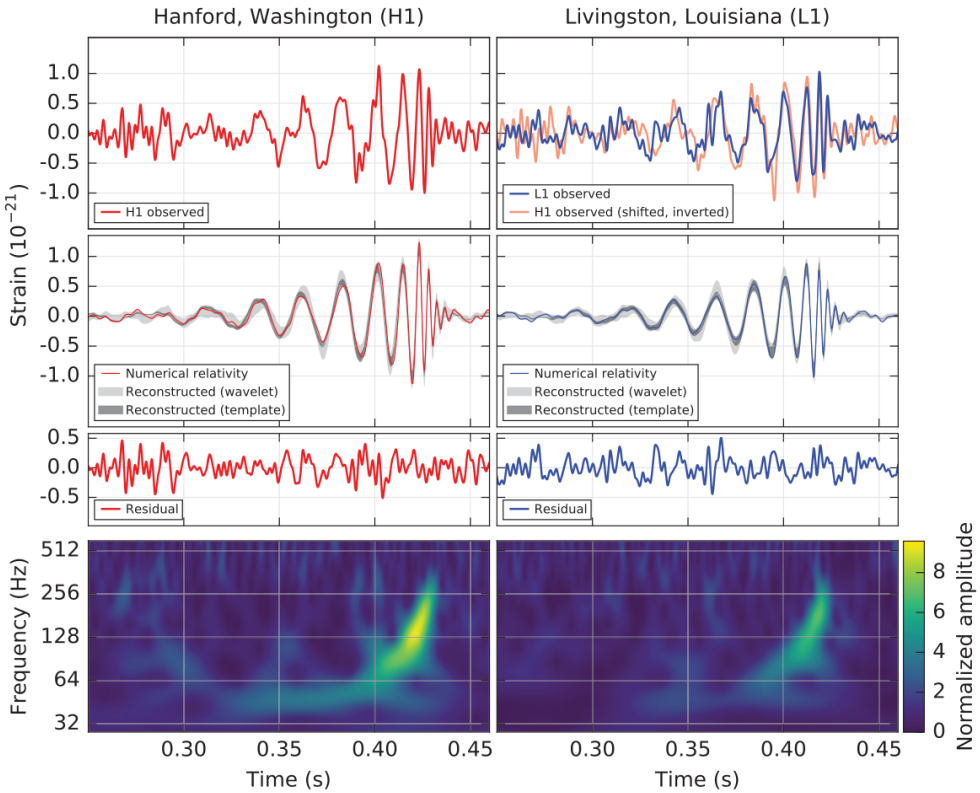
and in time

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

where $\mathcal{F}^{-1}$ is the discrete inverse Fourier transform.

Residuals are useful in testing waveform misspecification because the differences between waveform models are clearly seen in the time and frequency domain. Also, terrestrial noise artefact (glitch) in the data can be seen in the residuals.

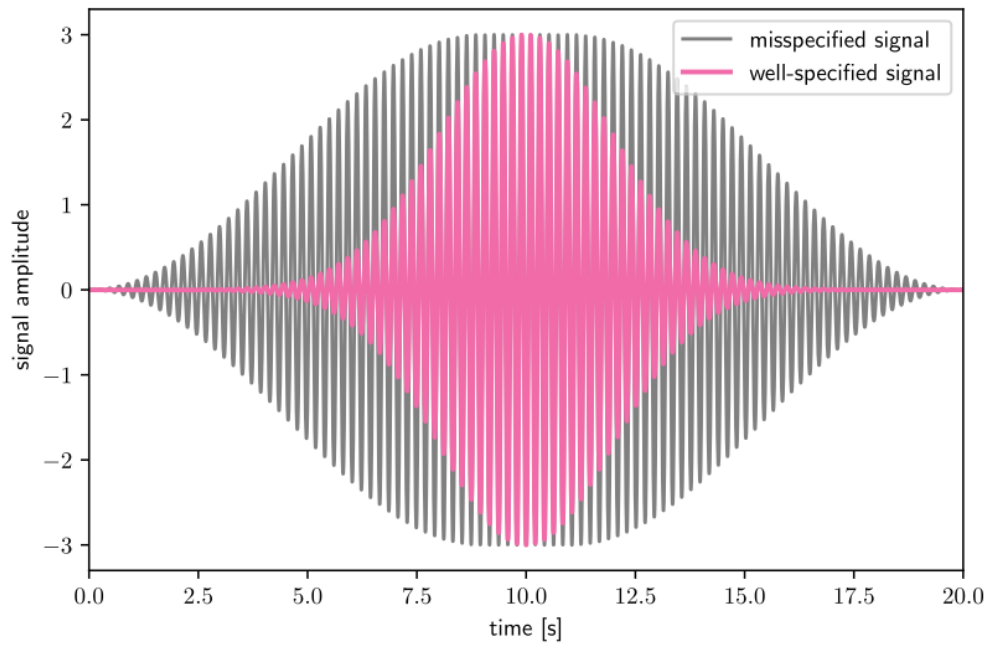
Residuals from best-fit for GW150914



For example, in Abbott et al. (2016), the best-fit, time-domain residuals for GW150914 were shown to be consistent with Gaussian noise showing that the data are well explained by a gravitational waveform in Gaussian noise at both the LIGO detectors.

Model misspecifiation test: Demonstration

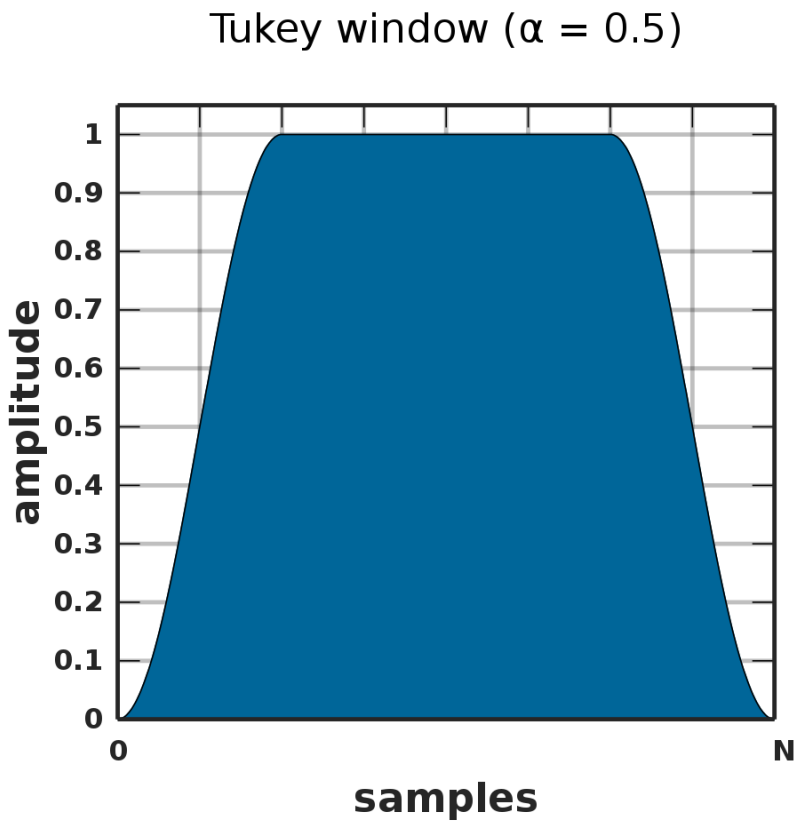
- Consider a signal model of sine-Gaussian chirplet (i.e., a sine wave multiplied by a Gaussian function). Create two synthetic datasets with Gaussian noise. The correctly specified data contains a signal that matches the model. The second dataset contains an intentionally misspecified signal: the same sine wave as before, but multiplied by a Tukey window. In both datasets, assume Gaussian noise with a known PSD.



Supplement 1 - Tukey window

- A window function is a mathematical function that is zero-valued outside of some chosen interval. Tukey window, or cosine-tapered window, can be regarded as a cosine lobe of width $N\alpha/2$ (spanning $N\alpha/2 + 1$ observations) that is convolved with a rectangular window of width $N(1 - \alpha/2)$.

$$
\left.
\begin{aligned}
w[n] &= \tfrac{1}{2}\left[1 - \cos\left(\tfrac{2\pi n}{\alpha N}\right)\right], && 0 \le n < \tfrac{\alpha N}{2} \\
w[n] &= 1, && \tfrac{\alpha N}{2} \le n \le \tfrac{N}{2} \\
w[N-n] &= w[n], && 0 \le n \le \tfrac{N}{2}
\end{aligned}
\right\}
$$



Model misspecifiation test: Demonstration - plot 1

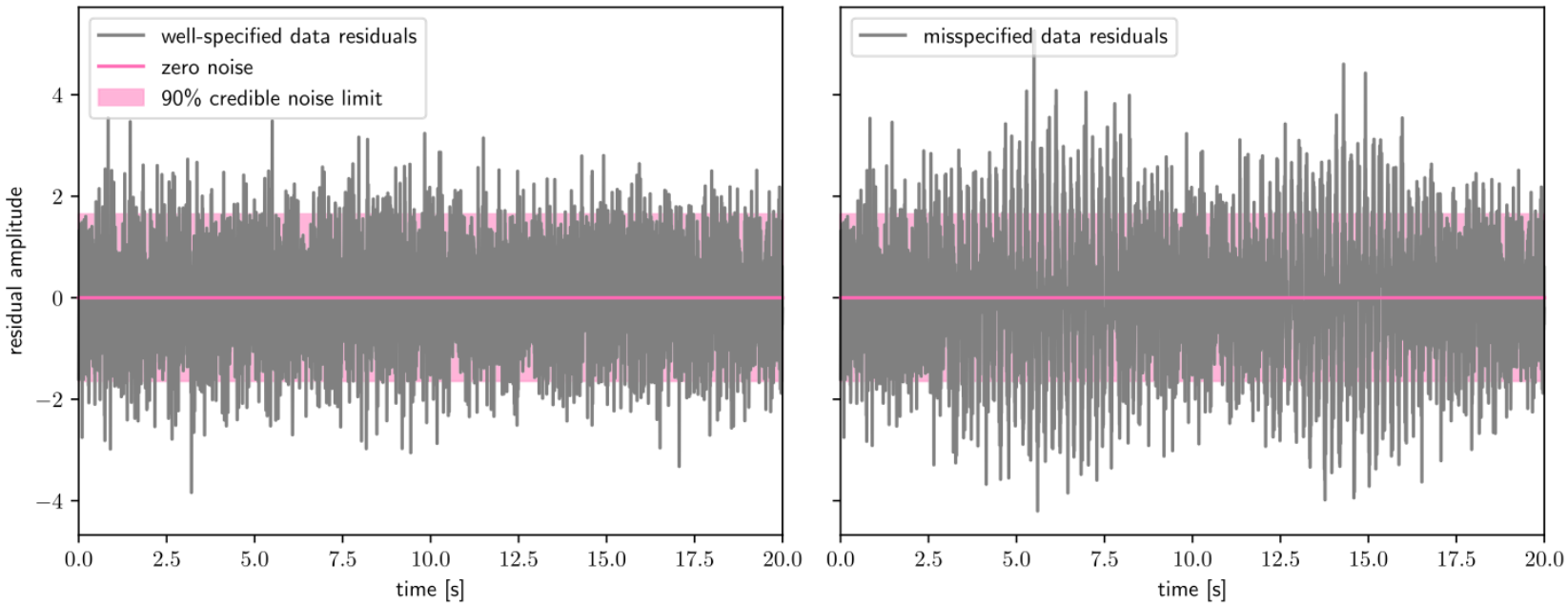


Figure 4: Time series of the residuals calculated by subtracting two different waveform models from the simulated data. The plot on the left shows the residuals obtained by subtracting a correctly specified waveform that matches the signal hidden in the data, while those obtained by subtracting a misspeified waveform are on the right.

Model misspecifiation test: Demonstration - plot 2

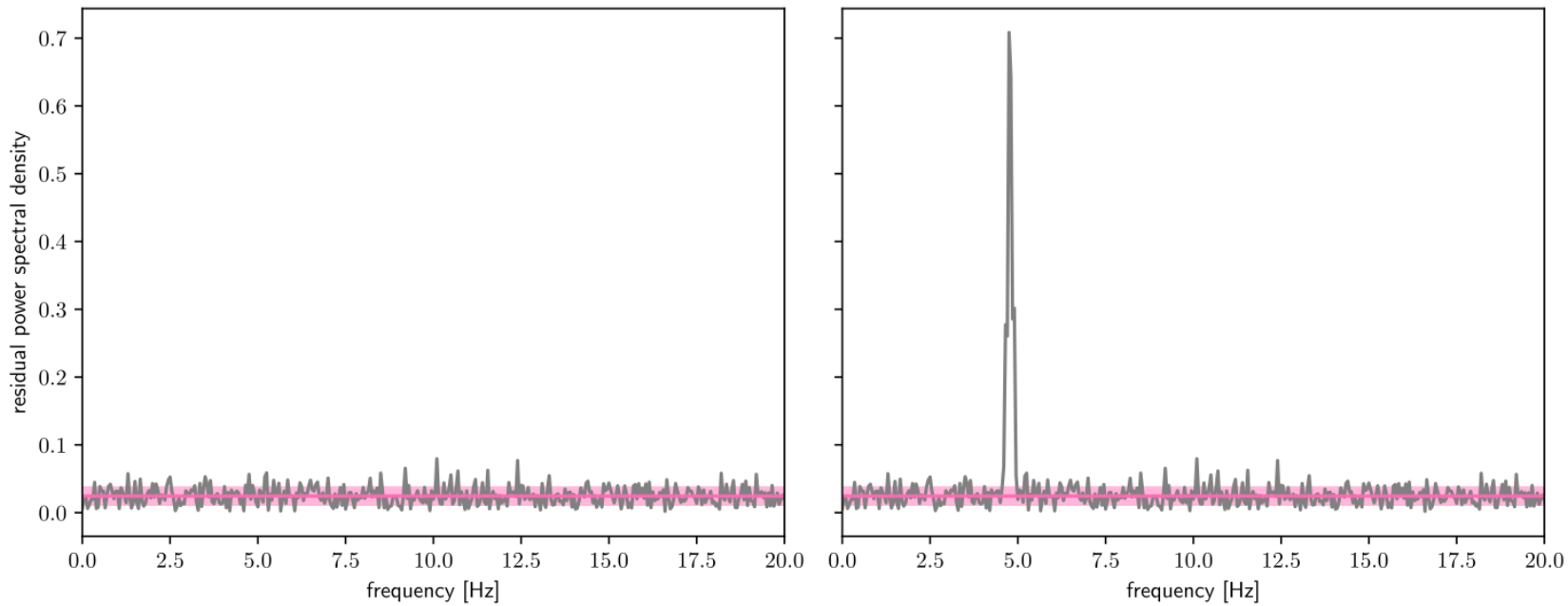


Figure 5: Frequency-domain amplitude spectral densities of the residuals. It is not totally clear from the time domain data if one of the datasets is poorly specified by the model, but Fourier transforming the residuals reveals a suspicious peak inconsistent with Gaussian noise. On both rows, the residuals are plotted in grey while the pink band indicates the range where the model predicts 90% of the residuals will lie.

Model misspecifiation test: Demonstration - plot 3

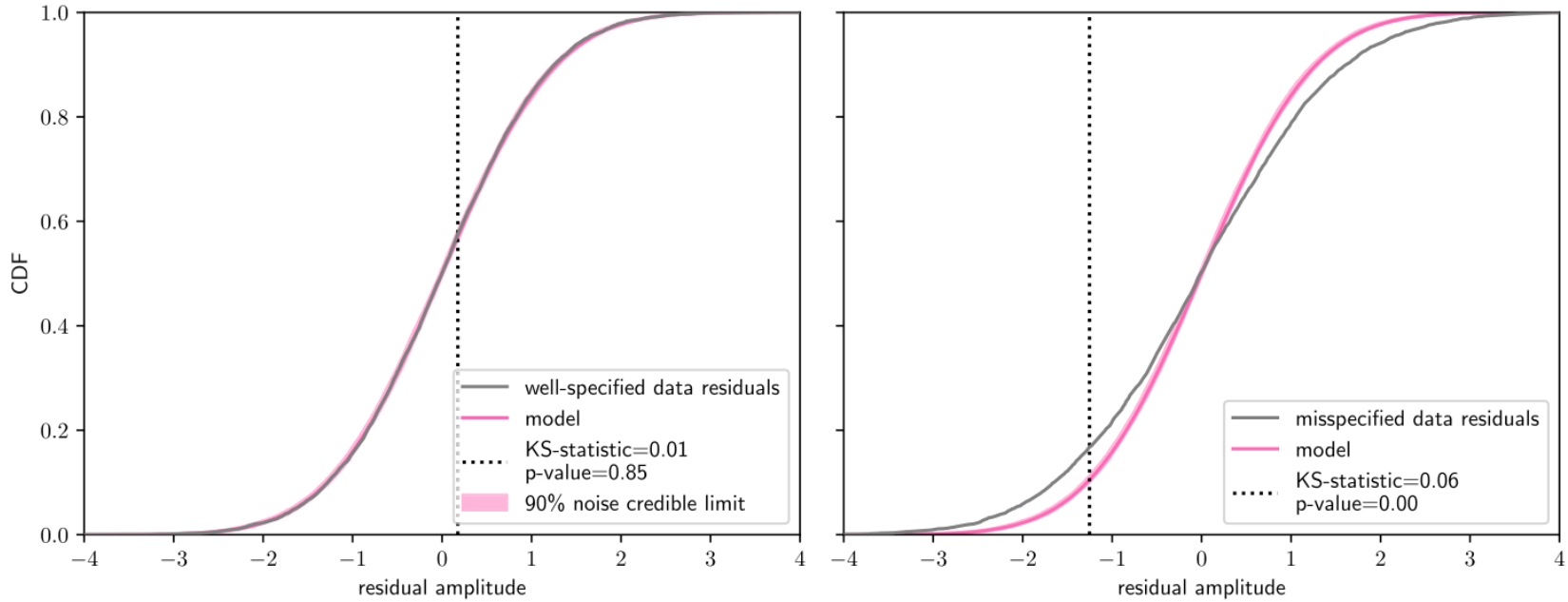


Figure 6: CDFs for the residuals of the time-domain data (grey) and predicted range of the residuals (pink). KS-statistic for the correctly specified and misspecified distributions is calculated. The location at which the KS test finds the maximum vertical distance b/w the model and the data is indicated with a dotted line. For this 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

Noise misspecifiation test: Demonstration

- Consider that 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 $\nu = 5$. These parameters are chosen so that the noise profiles appear, at first glance, to be consistent with each other.

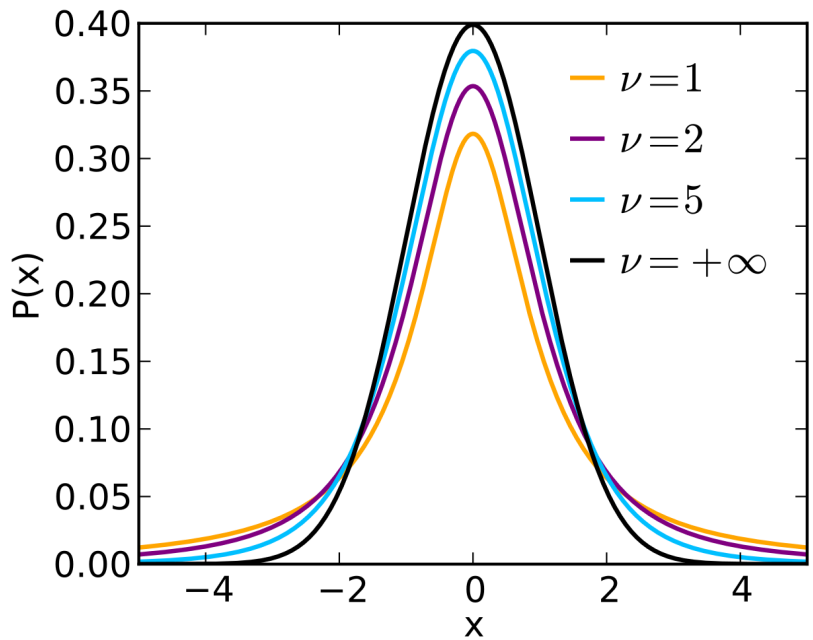
- Take the Fourier transform of the datasets and compare the 90% range predicted by the noise model against histograms of the data in the frequency domain. In the frequency domain, the misspecified data more clearly strays outside of the range predicted by the model.

- Create a CDF of the frequency-domain data and perform KS-test. An extreme p-value suggests misspecified noise data.

Supplement 2 - Student's t-distribution

- Student's t distribution is a continuous probability distribution that generalizes the standard normal distribution. Like the latter, it is symmetric around zero and bell-shaped. t_ν has heavier tails and the amount of probability mass in the tails is controlled by the parameter ν. For $\nu \to \infty$, it becomes the standard normal distribution which has very "thin" tails.

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\,\nu}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$



Noise misspecifiation test: Demonstration - plot 1

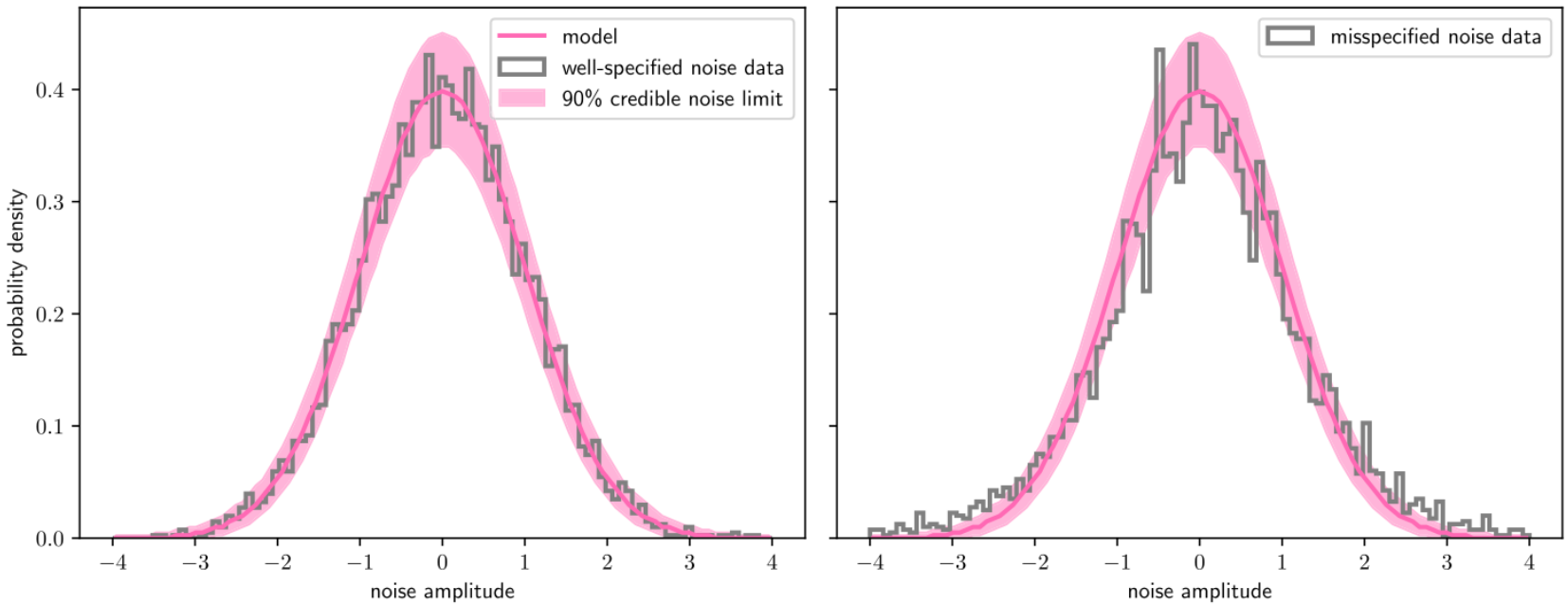


Figure 7: Simulated noise distributed as a Gaussian in the frequency domain. The correctly specified Gaussian distribution (left) and the similar-but misspecified Student's t distribution (right). The predicted (90% credible) range predicted by the model is shown by the pink band.

Noise misspecifiation test: Demonstration - plot 2

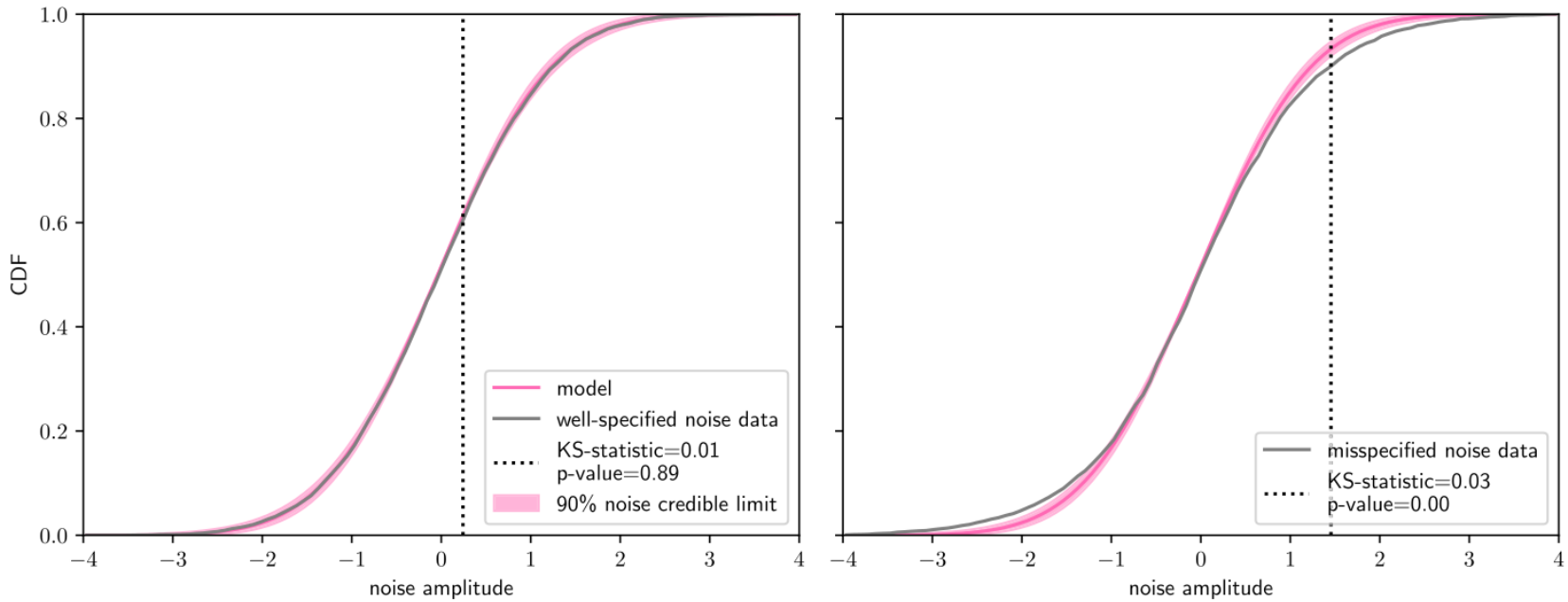


Figure 8: CDFs for frequency-domain residuals (grey) and the range predicted by the model (pink). For these specific noise realisations, the KS-statistics are 0.01 and 0.03 for the correctly specified and misspecified models, with p-values of 0.89 and 0.00 respectively. The location of the maximal KS distance is noted by a dotted line.

Noise misspecifiation test: Demonstration - plot 3

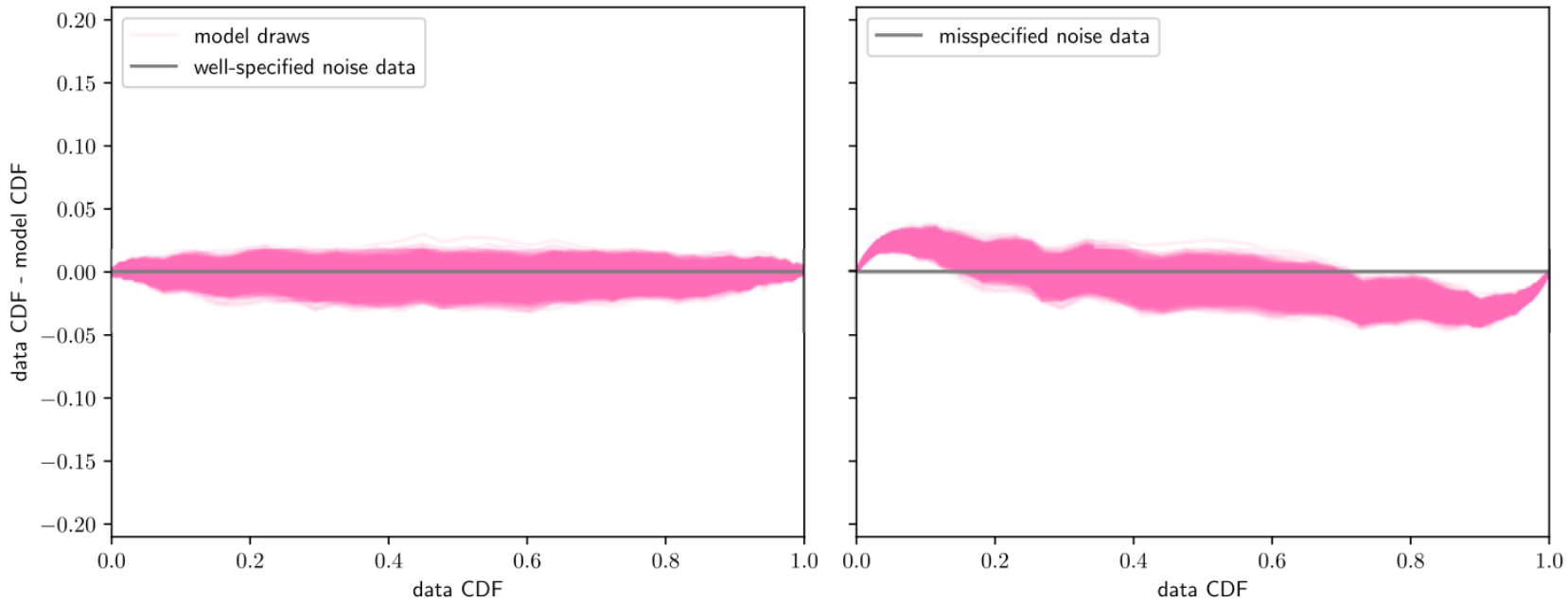


Figure 9: Like the previous plot, but the difference in data and model CDFs as a function of the data CDF.

CONCLUSION

- The article discusses various ways in which models can be tested for misspecification.

- Since all physical models are to some degree – misspecified, The question is not whether a model is wrong, but whether it is adequate or good-enough to describe a signal.

- The article employs statistical tools: Tukey window, Student's t-distribution, Gausssian distribution, KS-test and cumulative distribution function (CDF).

- The article uses the term cumulative density function (CDF) which is inappropriate. "The two words cumulative and density contradict each other. The value of a density function in an interval about a point depends only on probabities of sets in arbitrarily small neighborhoods of that point, so it is not cumulative." en.wikipedia.org/wiki/Cumulative_density_function