Optimal future waveform placement guided by Gaussian processes

Daniel Williams

February 16, 2017

The detection of gravitational wave signals in data collected by facilities such as the Advanced LIGO detectors in the USA rely on the comparison of noisy data with pre-determined models of the signals being searched for. In order to increase the efficiency of this process it is desirable to have access to a wide variety of models in order to conduct this process, known as matched filtering, through a large volume of the parameter space for a physical process which generates gravitational waves, for example, a binary black hole coalescence (BBH) event.

Models for these events must be produced by computationally expensive numerical relativity simulations, and as a result of this expense determining the most efficient exploration of the entire BBH parameter space is desirable.

Gaussian processes provide a means of interpolating over a parameter space, while also providing a measure of the uncertainty in the interpolated values in the form of a probability distribution. By training a Gaussian process to interpolate over the parameter space of numerical relativity waveforms for BBH events it is possible to produce a surrogate model which is capable of producing interpolated waveforms in regions of parameter space between sampled points. As this model is capable of assigning an error to these interpolated waveforms it is possible to identify areas of the parameter space where future sampling is desirable in order to reduce the error in the interpolation.

In a very sparsely-sampled parameter space, such as an BBH waveform catalogue, the interpolation error will reach a "saturation" point, and the vast majority of the error in the parameter space will be equal to this value; as a result a means of choosing the location for future samples within this saturated region is required. The training of the Gaussian process surrogate produces a measure of the scale-length of covariance within the parameter space, which can be used to produce a sampling lattice.

1 A Gaussian process surrogate

In order to determine the locations of parameter space which are most in need of a future simulation, we require some means to determine the areas of parameter space which *are* well explained by the current catalogue. To do this we trained a Gaussian process on the time-domain waveforms from a catalogue of BBH simulations (Georgia Tech waveform catalogue), using the squared-exponential covariance function, summing over a, b,

$$k(\vec{x}_i, \vec{x}_j) = \exp\left(-\lambda_{ab}\frac{1}{2}(\vec{x}_i - \vec{x}_j)^a(\vec{x}_i, \vec{x}_j)^b\right) \tag{1}$$

where $\lambda_{a,b}$ is a metric representing the scale-lengths of each dimension of the parameter space, and the various $x_i \in X$ are the coordinates of the training-points in parameter space.

The distribution of training points is displayed in figure 1.

The values for the elements λ_{ab} were determined by finding the values of λ_{ab} which optimise the marginalised likelihood (the evidence) of the Gaussian process trained off the data X, that is, by finding the λ_{ab} such that the quantity

$$\log p(\vec{y}|X, \lambda_{ab}) = -\frac{1}{2}\vec{y}^T K_y^{-1}\vec{y} - \frac{1}{2}\log|K_y| - \frac{n}{2}\log 2\pi \tag{2}$$

is maximised.

The output of a Gaussian process trained using the parameter-space training-data X, and the corresponding strain values, $y_i \in Y$ is then capable of interpolating waveform outputs at parameter-space coordinates which do not exist in the original training set. The full output of the Gaussian process is not a single interpolated function, however, but a distribution of plausible functions, and this provides a measure of the uncertainty in the interpolated function.

The magnitude of this error provides a means of detecting regions of the parameter-space which are poorly understood, and regions of high uncertainty should then be targeted for future simulations in order to improve the validity of the surrogate function across the entire parameter space. The values of λ_{ab} provide a suggested spacing for these new waveforms, which should be sampled at intervals of $\log(\lambda_{ab})$ in regions with high uncertainty or regions outwith the parameter-space region defined by the original training data. Figure 3 shows an 9-dimensional slice in the parameter space of the BBH waveforms from a Gaussian process trained off ten waveforms. We can see that even with a small number of waveforms it is possible to provide some estimate of the correct grid sampling scale-length required to complete the model[1].

In order to demonstrate that the parameter spacing is reasonably independent of the number of waveforms used in training the GP the model was generated with a differing number of training waveforms, and trained. The scale length for each parameter is found to be consistently similar, as can be seen in figure 2.

[1] For simplicity a hypercube lattice is illustrated in figure 3, however, more efficient lattice patterns exist to which the scale-length can be applied.

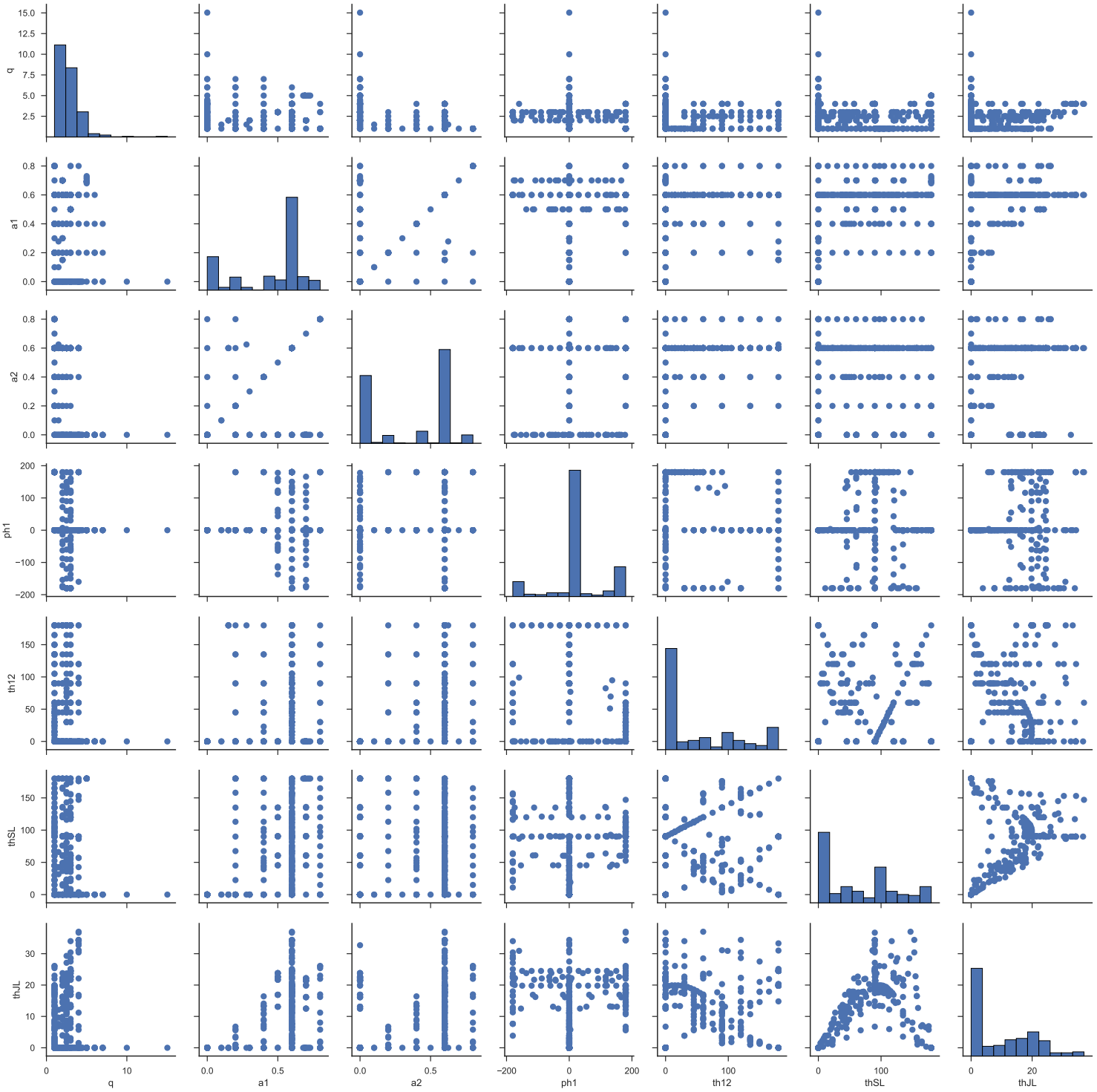

Figure 1: The distribution of training points within the numerical relativity BBH parameter space.

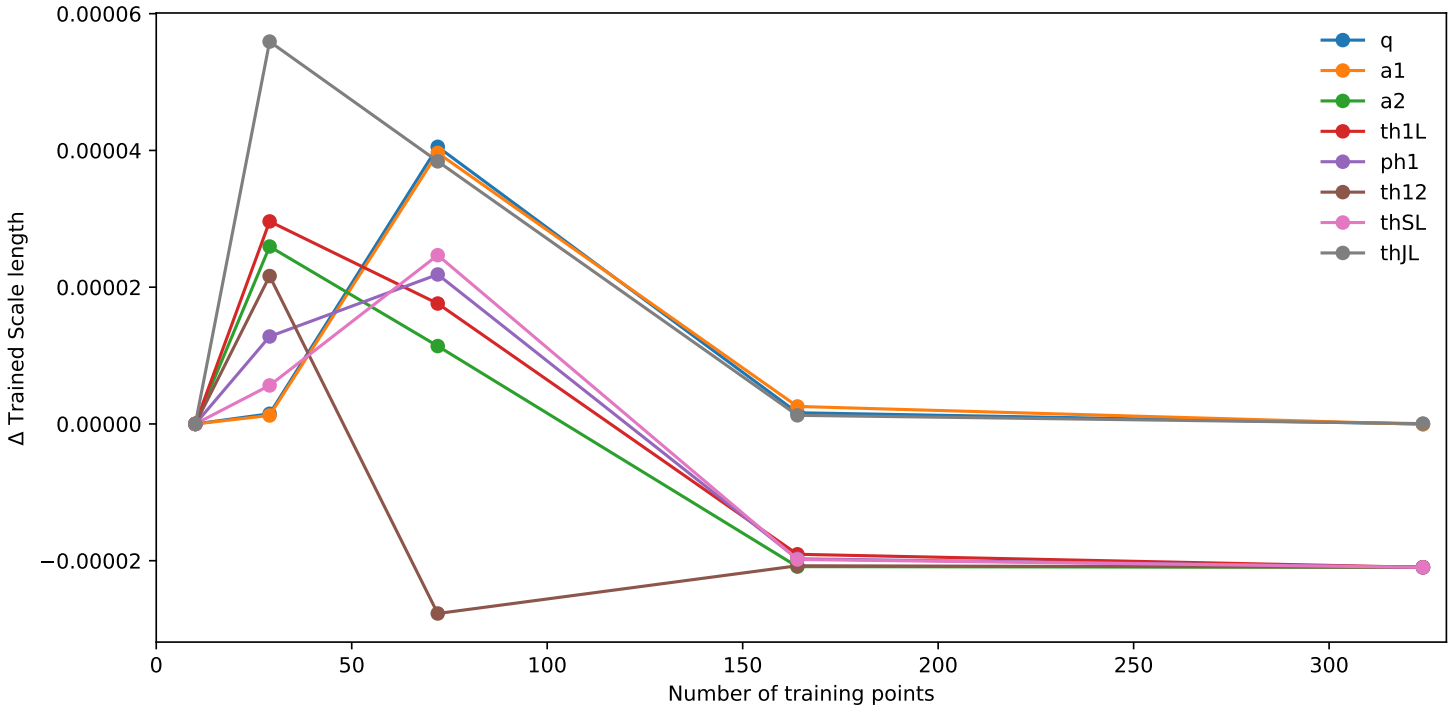

Figure 2: The difference in the scale-lengths of the trained Gaussian process given 10, 29, 72, 164, and 324 waveforms, compared to the GP trained from 10 waveforms.

2 Setting Priorities

Having established the optimal spacing of future waveforms we must then turn to the question of the optimal order in which they should be created. Sampling the entire parameter space at the suggested density would require in excess of 250,000 waveforms, which is an unachievably large quantity of data.

In other fields which exploit Gaussian processes the choice of future sample locations can be made using techniques from Bayesian optimisation. In these fields the Gaussian process is often emulating the behaviour of a complicated function surface, and the desired outcome is finding an optimum on this surface, for example, if a GP is used to model the distribution of pollutants in a lake, with the aim of identifying the source of the pollution we may want to find the region with the maximum amount of pollutant.

In the case of gravitational waveform modelling we are not terribly interested in the location of the function's maximum, but instead have the objective of achieving the greatest quantity of knowledge possible in the shortest possible period of time (or equivalently, with the smallest number of function evaluations). In this case the choice of future samples might be made to minimise the mean-squared error of the model.

By setting-aside some quantity of the available training data, to be used to test the Gaussian process, it is possible to calculate the mean-squared error of the GP's prediction compared to the known waveform. The mean-squared error is defined

$$\mathrm{MSE} = \sum_i (y_i^* - f(x_i^*))^2, \tag{3}$$

with $(x^*, y^*)_i$ the testing data.

Using methods based-on the use of testing data requires some of the available training data to be set-aside, and not used to train the Gaussian process. Given the small number of samples available, and the large volume of the parameter space this is likely to have a considerable impact on the predictive capability of the model. Other options are available, such as comparing the output to an analytical model, for example `IMRPhenomP`, but these approaches suffer from the incomplete understanding of the uncertainties in these models.

One possible approach is to simply use the metric defined on the parameter space by the Gaussian process to determine the location which is geometrically furthest from any pre-existing training point in the parameter space.

3 Open Questions

While this GP-informed grid approach provides an "optimal" grid spacing, a number of questions remain to be answered. First is the question of which covariance model should be used. While the squared-exponential function is a simple and fairly effective model, there is no immediate reason to assume it is the best model. We might expect to achieve broader grid spacings if a more informative model was used, thus reducing the number of required waveforms to provide a dense sampling across the parameter space.

Second is the question of how we identify the regions of parameter space most in need of future sampling to prioritise the exploration of the parameter space.

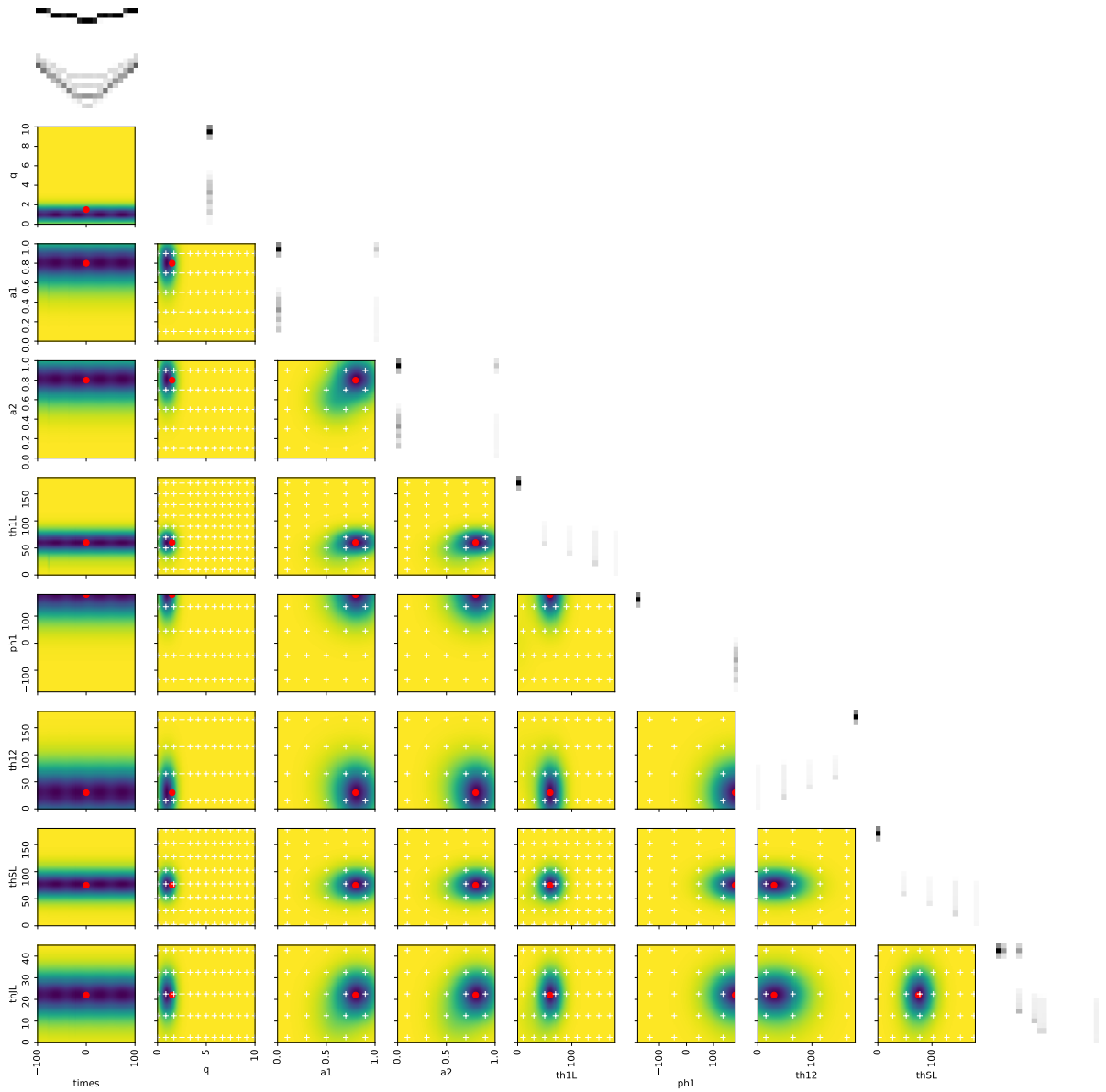

Figure 3: A corner plot of a 'hyperslice' the parameter space of the numerical relativity waveforms centred on $(t = 0, q = 1.5, a_1 = 0.8, a_2 = 0.8, \theta_{1L} = 60., \phi_1 = 180., \theta_{12} = 30., \theta_{SL} = 75., \theta_{JL} = 22.)$, showing the variation in the magnitude of the Gaussian process uncertainty over the parameter space in the colorplot, and the optimal spacing as implied by the width of the covariance function. The plots above each column are designed to provide a guide to the density of samples throughout the parameter space in that dimension, while the red point represents the point of intersection of the various planes. The Gaussian process used to produce these uncertainty estimates was trained off entire waveforms, including time domain information, but represents a series of slices which intersect at $(t = 0, q = 1.5, a_1 = 0.8, a_2 = 0.8, \theta_{1L} = 60., \phi_1 = 180., \theta_{12} = 30., \theta_{SL} = 75., \theta_{JL} = 22.)$