From Figure 1, the area of the shaded region for the predictive error bar is reduced after the Negative Log Marginal Likelihood is minimized from 92.91 to 11.90. The region (i.e. [-2, 2]) with more training data has a narrower error bar, since the prediction is more certain. The characteristic length-scale, signal standard deviation and noise standard deviation are also optimized as shown in Table 1.

By initialising the hyperparameters differently, we can see how they affect the minimization. In short, standard deviations of both noise and signal do not affect the result a lot, in comparison of length-scale. The ones initialized with the same length-scale value will always be minimized to the same nlml value. Therefore, it is sensible to only plot a graph of length-scale against negative log marginal likelihood, and to compare which initial setting of length scale gives the lowest nlml.

The second least nlml is 78.22 at ell = -0.25. The first fit with ell = -1 is better since it has the least nlml and highest marginal likelihood.

In comparison to the fit from a , the error bar has a narrower width even for the region with less training data. I do think the data generating mechanism was not strictly periodic even though both plots showed periodic structures, especially for the interval [-1.5, 2]. It is possible that the data was generated with a covariance function that has several changepoints.

When comparing the marginal likelihood, this periodic model has better fit to the data since it has a higher marginal likelihood and also has an extra hyperparameter which increases the capacity to fit the data.

To apply the Cholesky decomposition, the covariance matrix must be symmetric and positive definite. Therefore, adding the diagonal matrix helps to retain the symmetry and change it to be positive definite.

Since the covariance function is a product of periodic and squared-exponential kernels, both the predictive mean and bars oscillate in a periodic exponential pattern. I can also see there are many datapoints positioning within the predictive error bars but further away from the predictive mean, which is sensible since the data is generated with noise.

In the 3d plot of both models, they look similarly. The predictive means of both models show the same peaks as also shown in the pure data. But their nlmls are different.

The ml ratio between model 1 and model 2 is 3.27x10-10. It shows the model is a better fit to the data since model 2 is built upon a sum of 2 squared-exponential kernels which give a higher degree of freedom to fit the data.

After the Negative Log Marginal Likelihood (nlml) is minimized from 92.91 to 11.90, the area of the shaded region for the predictive error bar is greatly reduced, especially for the region with more data points since the more data there is, more certain the prediction is.

The corresponding hyperparameters are then optimized as shown in the Table 1. Both signal and noise standard deviations are reduced so that the complexity penalty is reduced, and the data is more fit. The characteristic length-scale is also reduced, so that the value slightly exceeds the average period of the x-data in which the average separation between x-data points is 0.0721 < 0.13, which makes the signal variance less dominant to the covariance function.

By initializing the hyperparameters differently, different local optimums are found. From the top left plot in Figure 2, different colours in the colour bar represent different values of nlml, but only two colours appear in the colormap. This means nlml is always minimized to a limited number of local minima, the minimization of nlml and optimization of hyperparameters result depend on the initial settings of the hyperparameters. From the range of values have been tested, there are three nlml minima found, which are 11.90, 78.22 and 106.35 (Table 2).

So to find next minimized value of nlml (78.22), the gaussian process is initialized with the hyperparameters (log(length-scale) = 0, log(variance of signal) = 0, log(variance of noise)= 0). After nlml reaches the local minimum, the hyperparameters are also optimized to the following values as shown in Table 3 which produce the plot shown in Figure 3(b).

Compared with the nlml values (a: 11.90, b: 78.22), the optimized hyperparameters from part (a) is a better description to the data. The plot (Figure 1) in part (a) also shows smaller area of predictive error bars.

In comparison to the fit from part (a), the error bar has a narrower width even for the region with less training data. When comparing the marginal likelihood (a: nlml = 11.9, c: -1.6), this periodic model has better fit to the data since it has a higher marginal likelihood and has an extra hyperparameter which increases the capacity to fit the data.

I do think the data generating mechanism was not strictly periodic even though both plots showed periodic structures, especially for the interval [-1.5, 2]. It is possible that the data was generated with a covariance function that has several changepoints. Also, there is not much data existing outside the interval [-1.5, 2] so it does not provide enough evidence to prove ‘the mechanism was strictly periodic’ outside the region mentioned.

To apply the Cholesky decomposition, the covariance matrix must be symmetric and positive definite. Adding the diagonal matrix helps to retain the symmetry. It also increases the values of the diagonal elements of the original covariance matrix so that the eigenvalues can all be positive, and the adjusted covariance matrix can be positive definite.

By changing the magnitude of the diagonal matrix, the noise variance is also changed. From the Figure 5, the area of the predictive error bars has increased when the magnitude of the diagonal matrix is increased. When the noise variance is increased, the nlml value is increased since it is penalized by the complexity as shown in Figure 6.

Since the covariance function is a product of periodic and squared-exponential kernels, both the predictive mean and bars oscillate in a periodic exponential pattern. Some sample functions are also plotted by changing each hyperparameter at a time based on the initial hyperparameters of log[l1, p1, stdf1, l2, stdf2] = [-0.5, 0, 0, 2, 0].

In terms of characteristic length-scales, the length-scale hyperparameter in the periodic function can alter the shape of the prediction (Figure 7-Row 1). At a small length scale, the periodic function still retains some shape of exponential function. When it grows, the periodic pattern is smoother. Later, the shape becomes almost flat.

Refer to Figure 7-Row 4, the increases in length-scale in the squared-exponential function tends to make the periodic function more stable. The periodic pattern with lower length-scale tends to decay in amplitude when the x-data is further away.

Refer to Figure 7-Row 2, when the period value of the periodic function increases, it increases the width of the periodic pattern.

For the variance of the signal, it changes the amplitude of the repeating patterns and reduces the size of the predictive error bars as shown in Figure 7-Row 3 and 5.

Predictions made by both model 1 and 2 show similar predictive mean shape to the data (as shown in Figure 8 (a)). By subtracting predictive means of both models, an insignificant difference is found in Figure 8 (b). Refer to Table 5, when comparing nlml values, model 2 has a better marginal likelihood so that it is a better fit to the data.

This is because model 2 is built using a sum of squared-exponential kernels which have 6 optimizable hyperparameters giving a higher degree of freedom to fit the data.