Data, Estimation, and Inference Lab Report

AIMS CDT 2020

Dominik Kloepfer

October 19, 2020

1 Background

The task in this lab assignment was to explore the capabilities of Gaussian Processes (GP) applied to a dataset containing weather sensor data from the Port of Southampton. Before discussing my methods and results, I will first give a quick introduction to the mathematical background.

1.1 Gaussian Processes

A Gaussian Process is a collection of (infinitely many) random variables, one for each point in its domain, such that every finite collection of these random variables has a jointly Gaussian distribution. This means that a GP does not predict a single function given some data but rather a distribution of functions. To sample from this distribution, one effectively fixes a (large) number of (closely spaced) points in the domain and samples from the multivariate Gaussian distribution for this collection of points that the GP prescribes.

In more mathematical terms, if $\boldsymbol{f}$ is the vector of n function values at the n p-dimensional points in the domain X (so X is an $n \times p$-matrix), then we sample $\boldsymbol{f}$ by using

$$\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{\mu}(X), K(X, X)), \tag{1}$$

where $\boldsymbol{\mu}$ and K are the mean vector and the covariance matrix respectively.

As alluded to by the above notation, mean and covariance matrix are given by functions that return the mean of the function value at the given point and the covariance between the function values at two given points. The practitioner encodes any prior domain knowledge into tese functions, and they then get modified by the observation of training data.

Gaussian Processes incorporate training data by conditioning the multivariate Gaussian distribution on observing a given set of function values at training data points. If we let the training set consist of the values $\boldsymbol{y}$ (i.i.d with a normal distribution around the true function values $\boldsymbol{f}$ with variance σ_n^2) at points X, then the distribution for the function values $\boldsymbol{f}_\star$ at points $X_\star$ conditioned on the training data is

$$
\begin{aligned}
\boldsymbol{f}_\star | X_\star, \boldsymbol{y}, X \sim \mathcal{N}(\bar{\boldsymbol{f}}_\star, \operatorname{cov}(\boldsymbol{f}_\star)), \text{ where} \qquad & (2)\\
\bar{\boldsymbol{f}}_\star \equiv \mathbb{E}[\boldsymbol{f}_\star | X_\star, \boldsymbol{y}, X] = \boldsymbol{\mu}(X_\star) + K(X_\star, X)[K(X,X) + \sigma_n^2 I]^{-1}\boldsymbol{y}, \qquad & (3)\\
\operatorname{cov}(\boldsymbol{f}_\star) = K(X_\star, X_\star) - K(X_\star, X)[K(X,X) + \sigma_n^2 I]^{-1} K(X, X_\star). \qquad & (4)
\end{aligned}
$$

1.2 Marginal Log-Likelihood

Given measurements $\boldsymbol{y}$ at points X with variance σ_n^2, one can marginalise out different possible function draws to calculate the marginal likelihood:

$$p(\boldsymbol{y}|X) = \int p(\boldsymbol{y}|\boldsymbol{f}, X) p(\boldsymbol{f}|X) d\boldsymbol{f}. \qquad (5)$$

Using standard Gaussian identities, this becomes the log-marginal likelihood with the following closed form:

$$\log p(\boldsymbol{y}|X) = -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}(X))^\top (K + \sigma_n^2 I)^{-1} (\boldsymbol{y} - \boldsymbol{\mu}(X)) - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log 2\pi. \qquad (6)$$

This formula can be used for finding better prior mean and covariance functions by incorporating hyperparameters in their definitions and maximising the marginal log-likelihood of the training data with respect to these function parameters.

It can also be used as a metric for comparing different models by computing the marginal log-likelihood of ground truth data; for a good model that predicts the ground truth values well the marginal log-likelihood will be larger.

2 Predicting Missing Values

In the first step of the investigation, the GP model was used to interpolate missing values in a time series. This was done by conditioning the posterior distribution on the known datapoints; the posterior distribution was then visualised through its mean function and bands corresponding to function draws one and two standard deviations removed from the mean.

For lack of space I will illustrate the process and the results for the two variables of air temperature (in degrees Celsius) and tide height (in m). The time variable is measured in hours since the first measurement in the dataset. For both of these the injection of artificial noise (jitter) was necessary to ensure numerical stability of the operations involved in the optimisation.

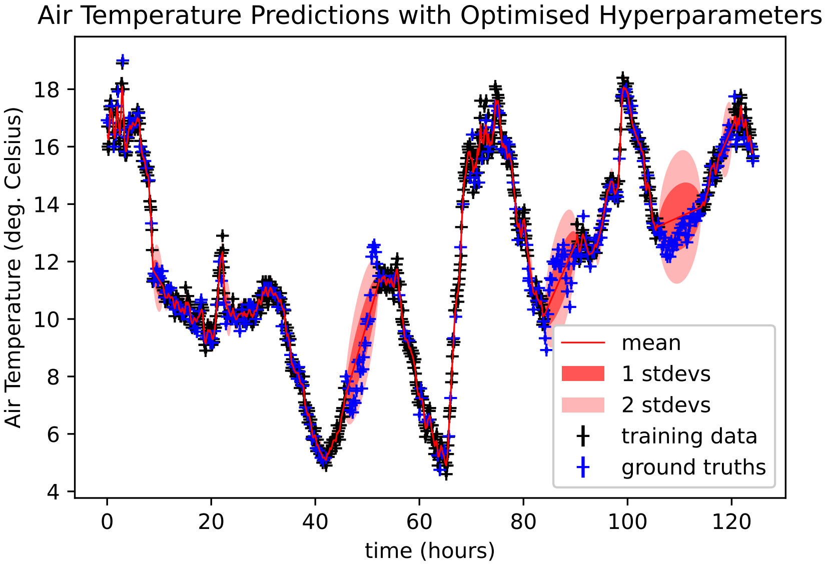

Figure 1: This plot shows the posterior mean function and the standard deviation that the posterior distribution assigns to each point when interpolating missing air temperature measurements. The variance of the posterior distribution increases quickly in the gaps between training datapoints.

2.1 Interpolation of Air Temperature

For the prior mean function I chose a constant function for simplicity; perhaps surprisingly, the optimisation algorithm chose for this constant a value that was higher than mean value of the dataset (it chose a value of 13.5 degrees Celsius instead of the mean of 11.9 degrees Celsius). For the covariance function I chose a Matérn-1/2 with two parameters based on the training data looking similar to the type of functions that Matérn-1/2 covariance function would encourage.

As one can see in the results in fig. 1, the posterior mean function tracks the training data very well and the variance increases quickly in the larger gaps between training datapoints. While the datapoints belonging to the ground truth are not always perfectly predicted, the uncertainty estimate seems reasonably well calibrated and roughly the expected number of ground truth datapoints fall within one and two standard deviations of the posterior mean function.

2.2 Interpolation of Tide Height

Here too I chose a constant prior mean function for simplicity, and here too the optimised value of the constant was very close to the mean value of the complete dataset. Since the tide height is known to have a period of twelve hours, I chose

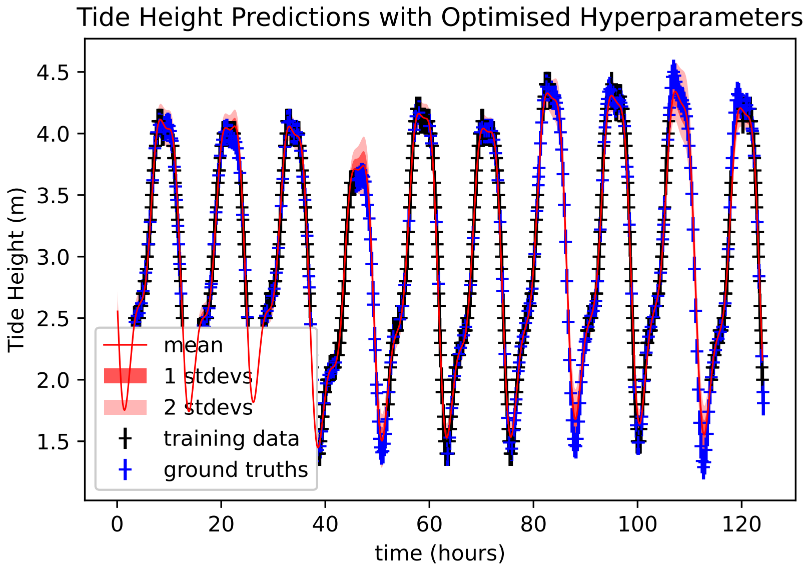

Figure 2: This plot shows the posterior mean function and the standard deviation that the posterior distribution assigns to each point when interpolating missing tide height measurements. The mean function shows a very good fit to the ground truth, so the relatively low variance of the posterior distribution is justified.

the linear combination of a periodic function and an exponentiated quadratic as prior covariance function; this combination means that the predicted value for the tide height is influenced both by other points shifted by a whole number of periods and also by points close to the point in question. As can be seen in fig. 2, this leads to a very good fit; the variance of the posterior distribution is quite low but this is justified since it fits the ground truth data very well.

3 Sequential Prediction

In this section, the predictions and uncertainty for each point were calculated individually using all the training points up to a fixed lookahead before the point in question; the function value at time t is predicted using the training data points up to time $t - \text{lookahead}$.

In implementing this, I used the same mean and covariance functions as in section 2 whose hyperparameters were optimised using all available data; while this leads to a better fit to the data than one would expect during actual deployment, it allows us to investigate the effect of varying the lookahead without any confounding effect from bad hyperparameters.

3.1 Sequential Prediction of Air Temperature

The Matérn-1/2 covariance function used leads to an excellent fit for small lookahead (fig. 3), but as the lookahead gets larger we start to see a more and more pronounced translation of the posterior mean function to the right compared to the ground truth values (fig. 4 and especially fig. 5). This can be explained by the fact that the Matérn-1/2 function is unable to anticipate rapid changes in function value because it is a local covariance function and so assumes that the closest training datapoint has the largest predictive power for the point that the GP is currently trying to predict. If the closest training datapoint is however quite far removed due to a large lookahead, it does not contain any real indication that in the ground truth data a rapid change in function value occurs at the point in question.

The uncertainty estimates however seems mostly well calibrated and as expected are larger for a larger lookahead and for regions with rapid changes in function value (though it is arguably too small in regions of rapid change in fig. 5).

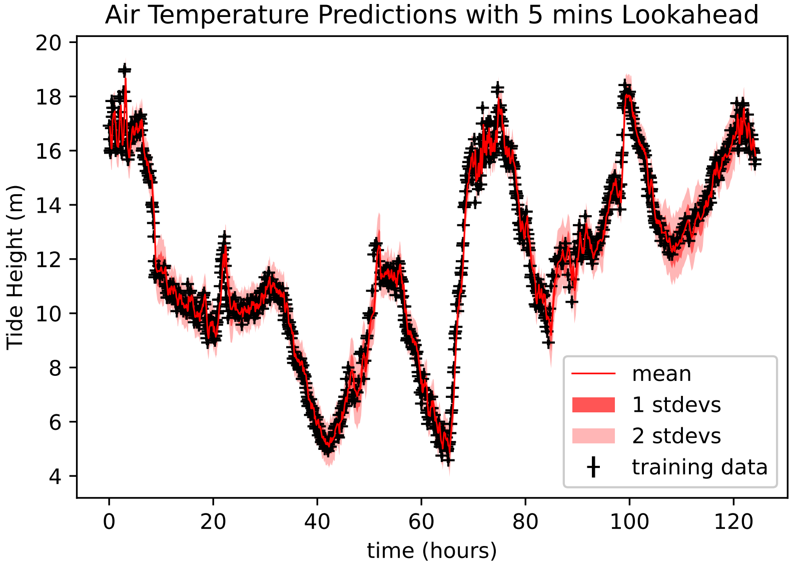

Figure 3: This plot shows the posterior mean function and the standard deviation that the posterior distribution assigns to each point when predicting air temperature measurements with a lookahead of five minutes. The mean function shows a very good fit to the ground truth but one can already see larger variance in regions with relatively rapid changes in the ground truths.

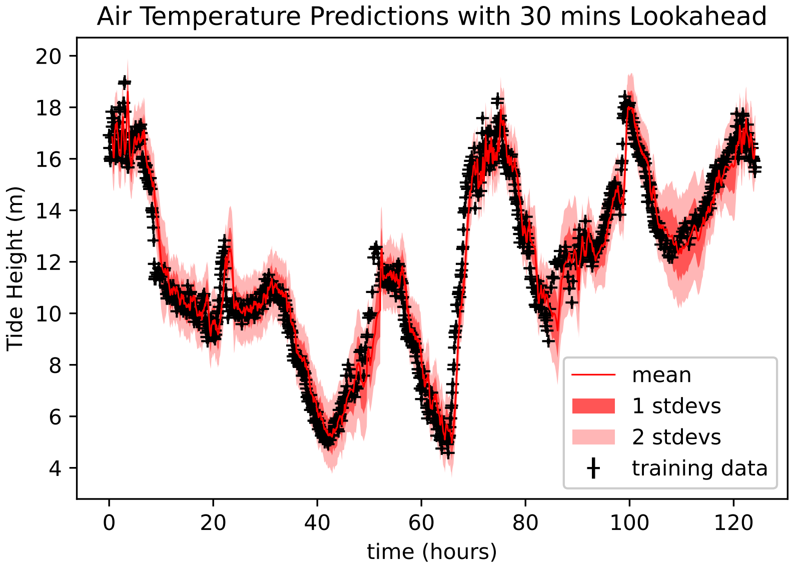

Figure 4: This plot shows the posterior mean function and the standard deviation that the posterior distribution assigns to each point when predicting air temperature measurements with a lookahead of 30 minutes. The fit of the mean function to the ground truth data is still reasonably good, and the posterior variance has grown to account for the errors that do develop.

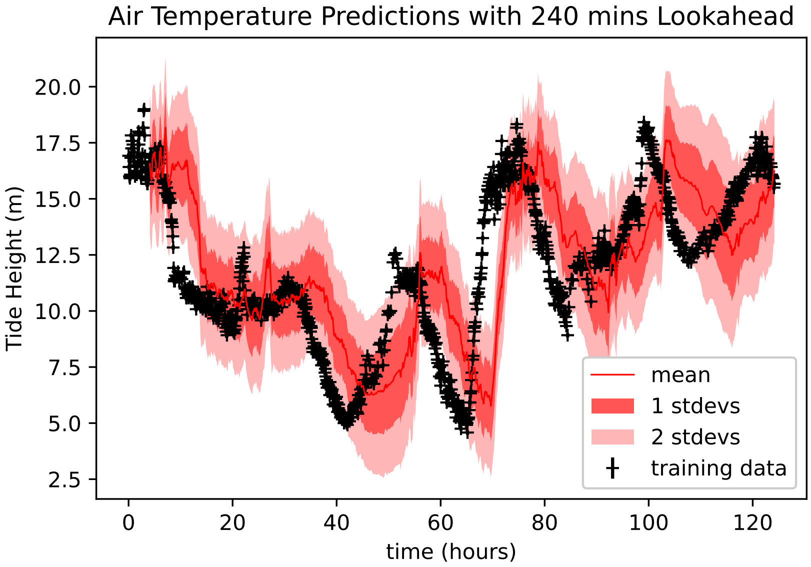

Figure 5: This plot shows the posterior mean function and the standard deviation that the posterior distribution assigns to each point when predicting air temperature measurements with a lookahead of two hours. One can now clearly see that the mean function is translated to the right with respect to the ground truth datapoints.

3.2 Sequential Prediction of Tide Height

As can be seen in fig. 6, fig. 7, and fig. 8, even for a large lookahead the posterior mean function fits the ground truth data quite well. This is likely a consequence of the very regular and periodic tide height being easy to predict given a well-calibrated covariance function.

It is noticeable that as the lookahead grows, the uncertainty grows significantly for the first period in particular (compare fig. 6 and fig. 7), likely because the GP does not have many training datapoints for the first few datapoints. This is especially pronounced for a large lookahead (fig. 8). The GP also has trouble anticipating slightly lower or larger amplitudes when using a large lookahead because it is difficult to anticipate the tide height levelling off earlier than before when the training data barely includes any points from the current period.

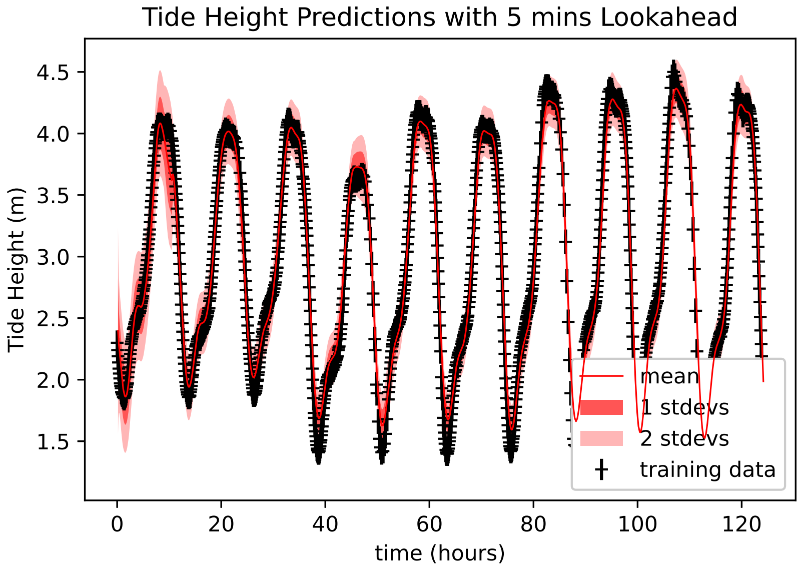

Figure 6: This plot shows the posterior mean function and the standard deviation that the posterior distribution assigns to each point when predicting tide height measurements with a lookahead of five minutes. The mean function shows a very good fit to the ground truth, with generally small variances.

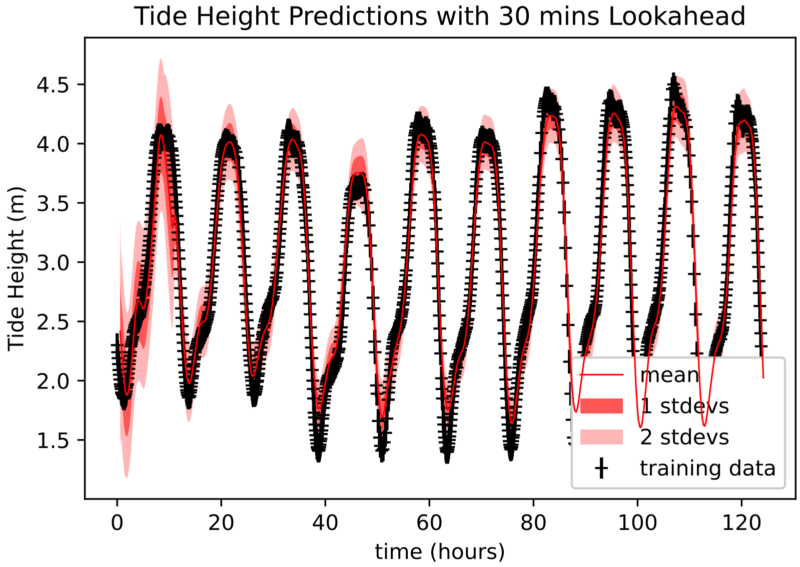

Figure 7: This plot shows the posterior mean function and the standard deviation that the posterior distribution assigns to each point when predicting tide height measurements with a lookahead of 30 minutes. The fit of the mean function to the ground truth data is still reasonably good but we can see the variances to grow, especially for the first period.

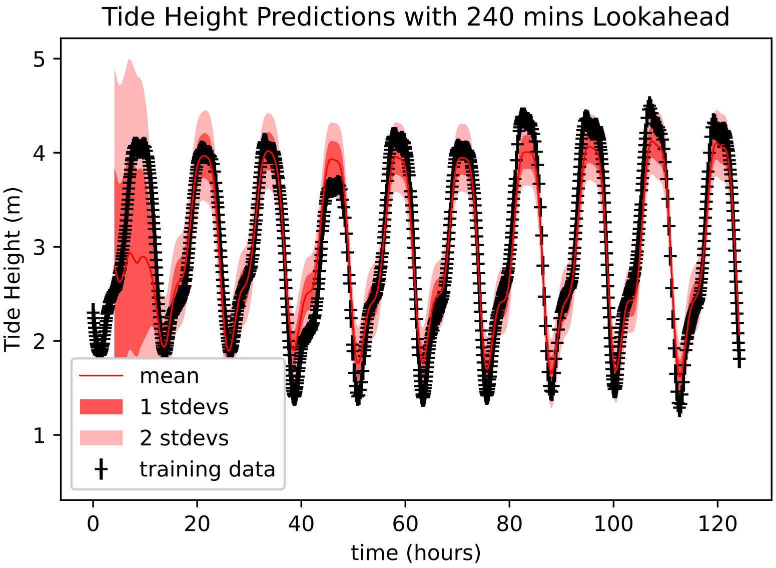

Figure 8: This plot shows the posterior mean function and the standard deviation that the posterior distribution assigns to each point when predicting tide height measurements with a lookahead of two hours. The mean function still roughly follows the ground truth data, but fails to predict the tide height accurately at the peaks of the plot and especially for the first period where the Gaussian Process only has very few training datapoints to predict the amplitude of the function. This is reflected in much larger posterior variances.