4F13 Probabilistic Machine Learning - Gaussian Processes

Lawrence Tray
St John's College

November 3, 2020

Abstract

1 Introduction

2 Questions

2.a Squared Exponential Covariance Function

We start with a simple squared exponential (SE) covariance function. As we start by working in one dimension this is necessarily isotropic. The covariance function is given by:

$$k(x, x') = \nu^2 \exp\left\{-\frac{(x - x')^2}{2l^2}\right\} \tag{1}$$

The hyperparameters are ν and l which control the baseline variance level and length scale of variation respectively. We load in the training data from *'cw1a.mat'* and train a GP model, with zero mean and covariance function given by equation 1. We train the model by minimising the negative log marginal likelihood (denoted $\mathcal{L}$). Table 1 scenario A shows the how the parameters are optimised.

Parameter	**A Initial**	**A Final**	**B Initial**	**B Final**
$\log l$	-1	-2.054	-0.45	2.085
l	0.368	0.128	0.638	8.045
$\log \nu$	0	-0.109	0	-0.363
ν	1	0.897	1	0.696
$\mathcal{L}$	0	-2.139	0	-0.411

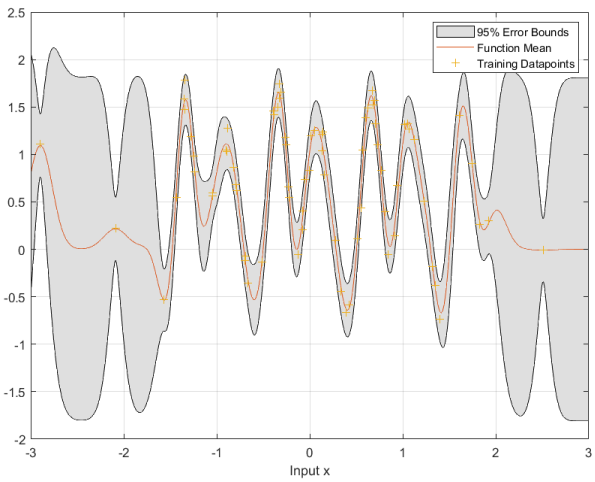
Table 1: Parameter optimisation (log values included for reference as reported by gpml toolbox)

This yields a predictor as in figure 1a. The 95% error bound is computed by $[\mu(x) - 2\sigma(x), \mu(x) + 2\sigma(x)]$. In other words, as $y \sim \mathcal{N}(\mu(x), \sigma(x)^2)$, there is a 95% chance of y falling within two standard deviations of the mean (all evaluated at a specific x).

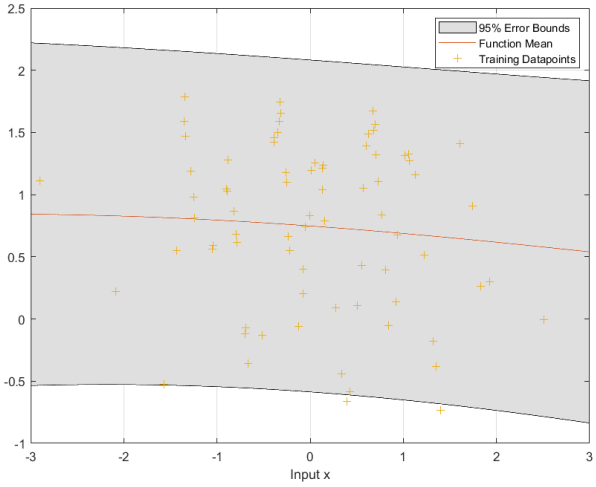
We see that the error bars are always centred on the mean and that they have small standard deviations for regions in which we have many datapoints observed. This makes intuitive sense as we cannot make confident predictions in areas where the training data is sparse (such as for $|x| \geq 3$). The hyperparameters do not change enormously form the optimisation. We have that the length scale of variation l shrinks slightly to 0.128 - which agrees with the length scale of variation in the dataset. The scale factor ν also shrinks slightly to 0.897 as the model is able to match the data quite accurately.

2.b Hyperparameter Initialisation

However, the optimisation only finds a local minimum of the negative log-likelihood $\mathcal{L}$. Therefore, a different intilisation of the hyperparameters can yield different results. This is illustrated in scenario B of table 1. It was found that $\log l = -0.45$ was a critical point. Any setting of $\log l$ above this would converge to the case-B optimum; anything below converges to the original case-A optimum. Varying ν only seemed to change the position of this critical point but would not converge to an altogether different solution.



(a) Basic hyperparameter initialisation



(b) Alternative hyperparameter initialisation

Figure 1: Squared Exponential covariance Gaussian Process trained on data

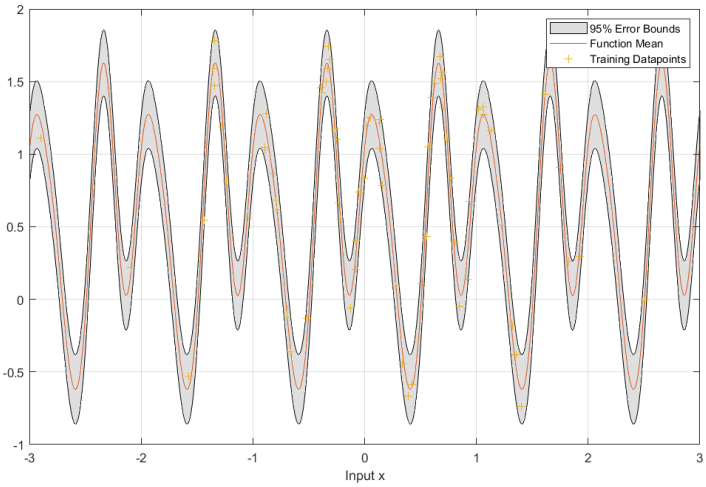


Figure 2: Periodic covariance GP on same training data

2.c Periodic Covariance Function

2.d Cholesky Decomposition

2.e Model Comparison

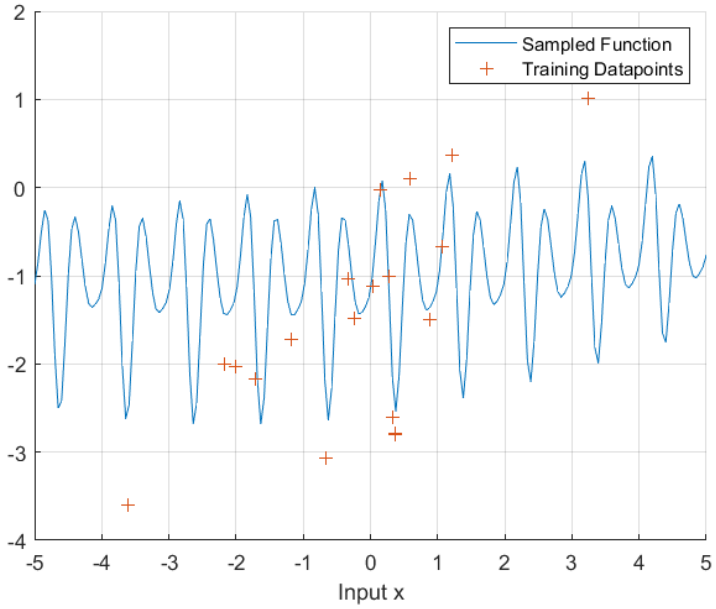
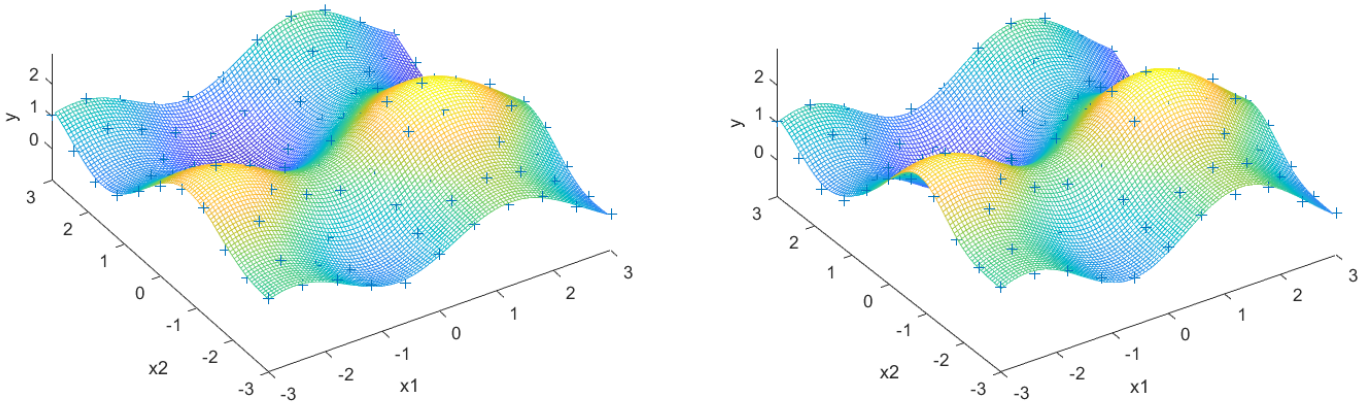


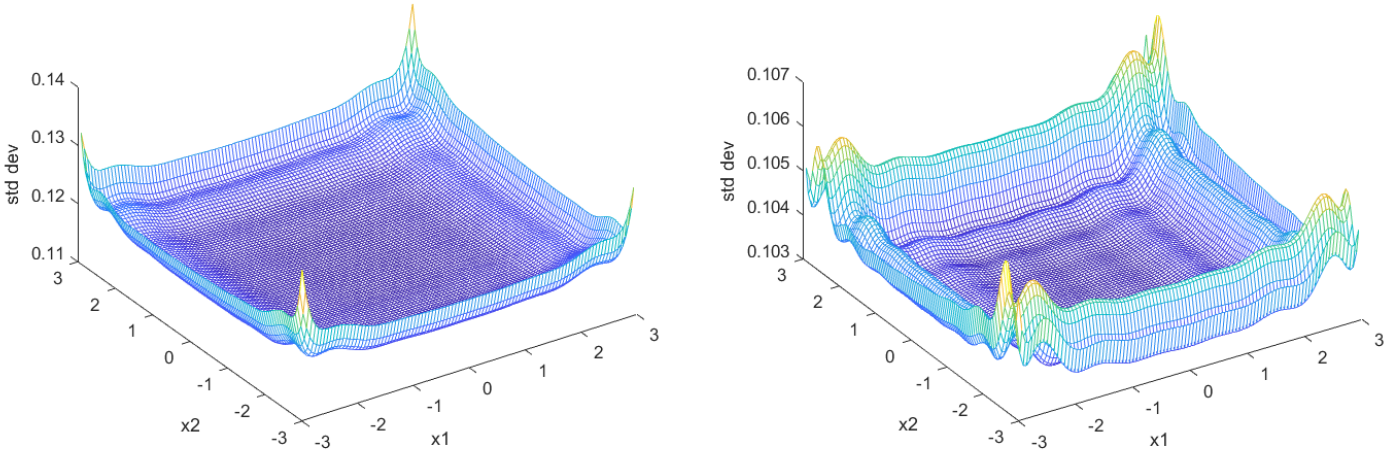
Figure 3: Trained GP with initial hyperparameter settings



(a) Basic covariance function: single covSEard model

(b) Additive covariance function: two covSEard model

Figure 4: Comparison of two covariance function fits on training data



(a) Basic covariance function: single covSEard model

(b) Additive covariance function: two covSEard model

Figure 5: Comparison of standard deviation for