UNIVERSITY OF
CAMBRIDGE

Gaussian Processes

Mihai Varsandan

mv436
Girton College

10th November 2019

1 Theory

Gaussian Process represents a different method of performing regression by allocating a prior probability to functions($\mathbf{f}$) that we think could fit the data($\mathbf{y}$) very well. The covariance is used to store these priors as *kernels*.However, there an infinite number of functions.This is tackled by the the marginal likelihood which is computed as the integral of the likelihood times the prior over the all the possible functions.[RW06a]

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)d\mathbf{f} \tag{1}$$

The Gaussian process implies that the prior is Gaussian, $\mathbf{f}|X \sim \mathcal{N}(\mathbf{0}, K)$ and the likelihood is a factorised Gaussian $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma_n^2 I)$[RW06b]. The result of the integral can be expressed as a log likelihood of the form:

$$\log p(\mathbf{y}|X) = \underbrace{-\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 I)^{-1}\mathbf{y}}_{\text{data fit}} \underbrace{-\frac{1}{2}\log|K + \sigma_n^2 I|}_{\text{model complexity}} \underbrace{-\frac{n}{2}\log 2\pi}_{\text{normalisation const.}} \tag{2}$$

It is observed the that marginal likelihood in Gaussian process represents a trade off between *data fit* and *model complexity*(*Occam's razor*). To predict the value of new outputs($\mathbf{y}_*$) provided that new inputs ($\mathbf{x}_*$) are given, the following equation is used:

$$\mathbf{y}_* = \sum_{i=1}^{n}(K + \sigma_n^2 I)^{-1}\mathbf{y}k(\mathbf{x}_i, \mathbf{x}_*) \tag{3}$$

2 Task A

Data contained in **cw1a**.mat is loaded into *Matlab* and a GP is trained using the **GPML** toolbox. A covariance function of the form of a *Squared Exponential(SE)* is used:

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp -\frac{1}{2}(\mathbf{x}_p - \mathbf{x}_q)^T M(\mathbf{x}_p - \mathbf{x}_q)) + \sigma_n^2 \delta_{pq} \tag{4}$$

$$M = l^{-2} I \tag{5}$$

Initially the hyperparameters l,σ_f,σ_n are initialised at 0.37,1,1 and the negative log marginal likelihood (nlml) is optimised. The result can be seen in the *figure 1* below:

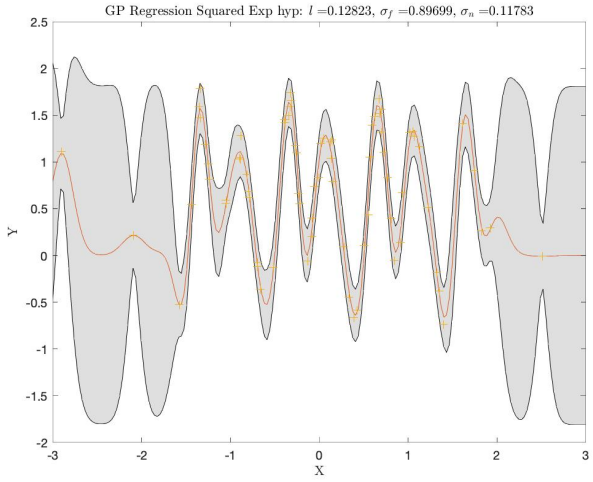
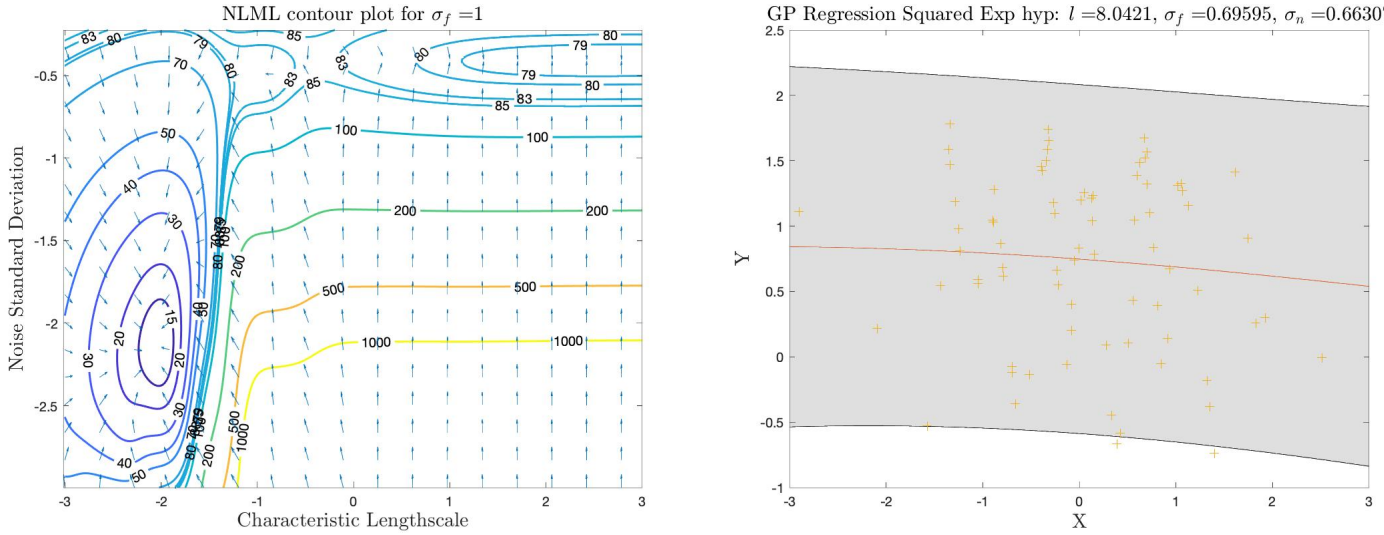


Figure 1: GP regression with SE covariance at low length-scale and low noise with $nlml = 11.9$

As observed in the figure the optimised hyperparameters are found to be 0.13,0.9,0.118. Therefore the length-scale is relatively small resulting in growing errors when the points are not very close to each other. The noise level is very low which shows that the model does not expect a lot of noise in the data.

3 Task B

By initialising the hyperparameters at some different values the optimisation converges at different local minima. In the *figure 2a* below the contour plot of the NLML is shown at $\sigma_f = 1$. As observed there are two local minima where the optimisation could converge to. This is shown by the arrows in *figure 2a* where the gradient of the function points toward the minimum point. From the contour plot, it is observed that at high noise values, the marginal likelihood becomes almost independent of the length-scale. Similarly, at low length-scales and low noise, the marginal likelihood becomes almost independent of the noise([RW06c]).



(a) Contour Plot of the Negative Log Marginal Likelihood of GP function [LV]

(b) GP regression for the local minima centered at large lengthscale and large noise with $nlml = 78.2$

Figure 2

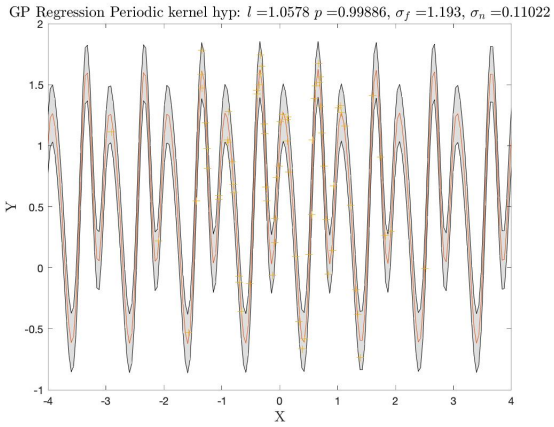
In the *figure 2b* the GP regression for the other local minima is shown. As observed, because both the length-scale and the noise are very big the data is explained by a slowly varying function. Comparing the two types of regression, it is observed that the data-fit decreases with the length-scale as the model becomes less and less flexible to variations in the inputs. On the other hand, the complexity penalty increases with the length-scale, because the model gets less and less complex as observed in *figure 2b*([RW06c]). The difference between the *nlml* in both of GP is significant in favour of the *figure 1*. Also from *2b* it is observed that generally, this GP fails to describe the data very well due to high noise. Therefore a based on this evidence the model with a low length-scale and low noise would be in favour to become the representative model of the data.

4 Task C

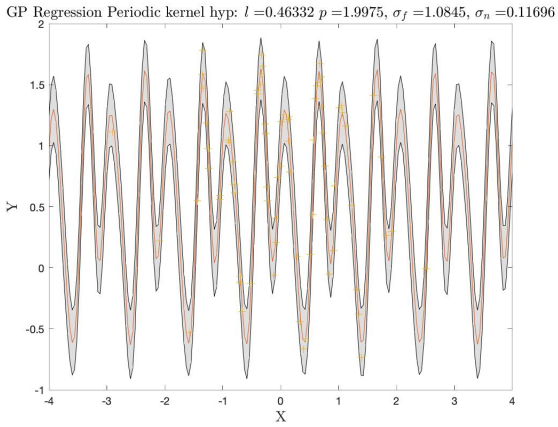
Having explored the *SE kernel*, a periodic covariance function is now explored. The function is showed below:

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp(-2\frac{\sin(\frac{\pi(\mathbf{x}_p - \mathbf{x}_q)}{p})^2}{l^2}) + \sigma_n^2 \delta_{pq} \tag{6}$$

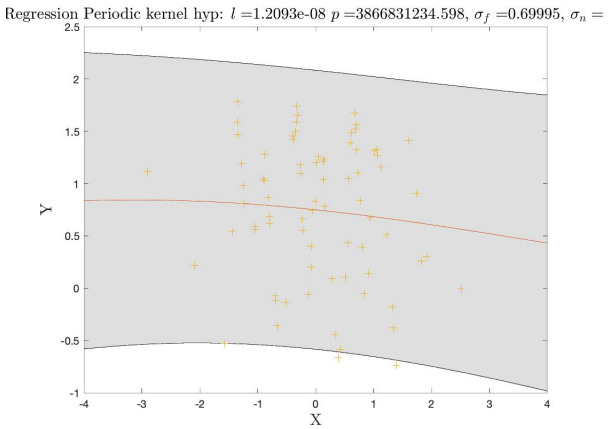
By applying this kernel to the same data as in the previous task and initialising the hyperparameters at different points, multiple regression models are available, each corresponding to a local minima. In the figures below some of the models are displayed. In the *Appendix 1* a contour plot of the length-scale against the periodicity(p) is showed.From the contour plot is observed that at low noise and relatively low length-scale the *nlml* only depends on the periodicity. Making the periodicity very big has the same effect as increasing the length-scale as observed in *figures 3c and 3d*.

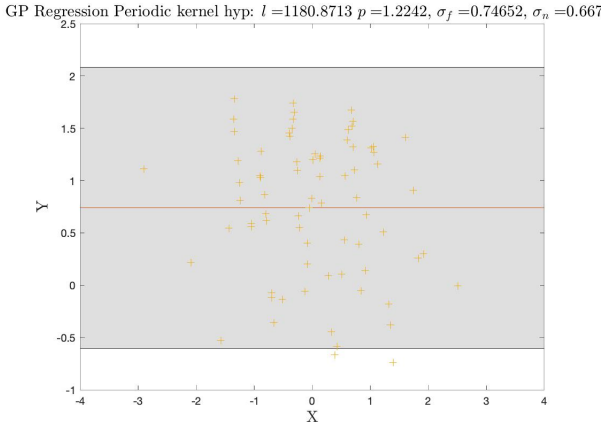


(a) GP regression with $nlml = -35.26$



(b) GP regression with $nlml = -16.7$





6.jpg

(c) GP regression at low length-scale and high periodicity with $nlml = 78.2$

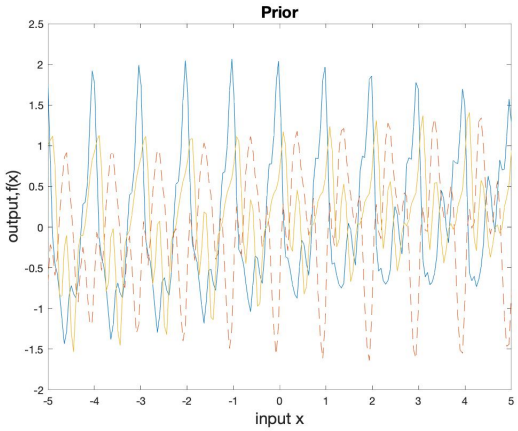
(d) GP regression at high length-scale and low periodicity with $nlml = 78.34$

Figure 3

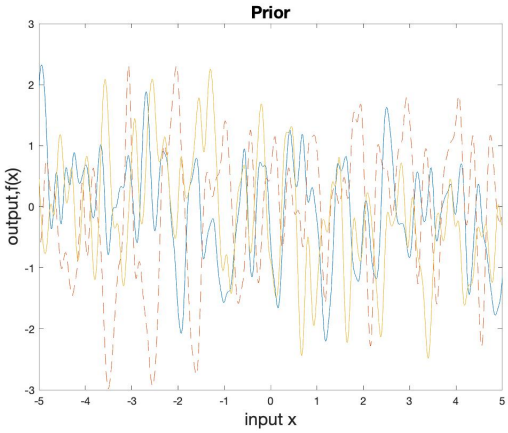
As observed from figures 3a and 3b the value of *nlml* is very low compared to the one in Task A. According the *nlml* value using a periodic function as kernel is a better representation of the data-set.

5 Task D

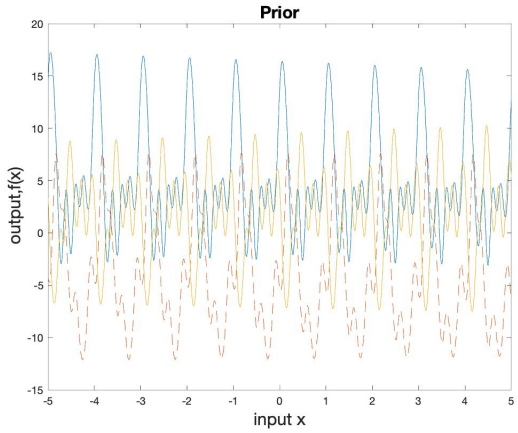
So far, both the *SE kernel* and *Periodic kernel* have been studied independently. Now, a covariance function which has the product of the two kernels as its value is studied. To visualise this function a GP prior is built containing this covariance function. The value of the prior is found by drawing from Gaussian distribution with the covariance function that it is wanted. The form of the covariance is a Cholesky decomposition of $K_x + \sigma_n^2 I$. A diagonal matrix is added to form a positive definite matrix out of which a Cholesky decomposition can be made. Three random functions for different hyperparemeters are displayed below:



(a) Prior functions with a low periodicity and a high length-scale of SE kernel



(b) Prior functions with a low periodicity and a low length-scale of SE kernel



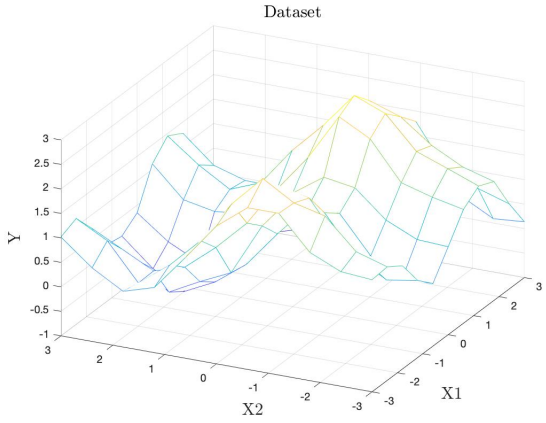
(c) Prior functions with a low periodicity and a high length-scale and a high signal variance of SE kernel

Figure 4

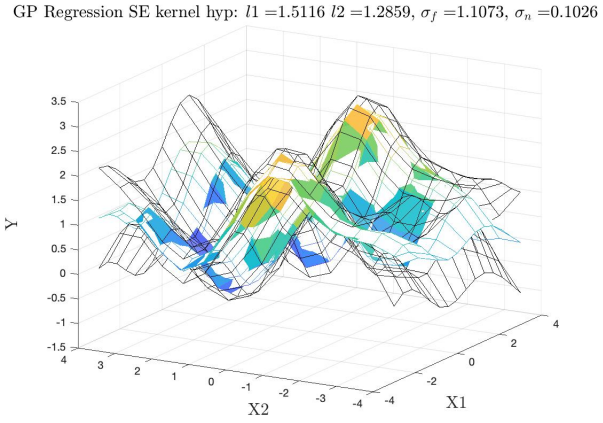
As observed from the figure above there two characteristics which correspond to all functions drawn from this covariance. The periodicity coming from the Periodic covariance and the smoothness coming from the SE covariance. By having the length-scale of the SE kernel high the functions become more periodic and they will vary less, which is observed in *figure 4a*. Decreasing the length-scale of the SE covariance(*figure 4b* shows that the periodic varies more with the change in the input x. If instead the signal variance is modified the amplitude of the functions is being modified(*figure 4c*).

6 Task E

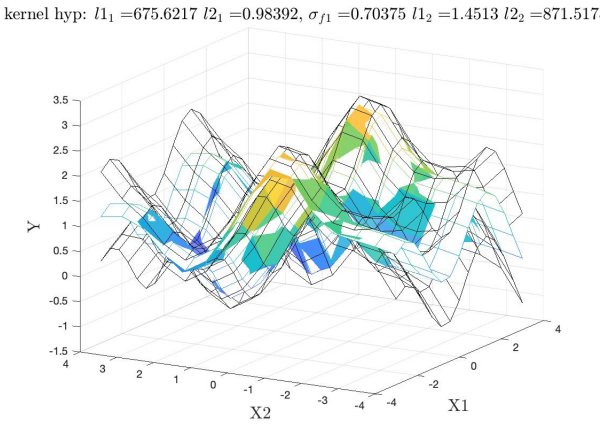
A 2D input data has now been loaded from **cw1e.mat**(it can be seen in *figure 5a*.This data is modelled using by two GP functions, each with a different covariance function. The first one can be seen in *figure 5b*. It observed that the length-scale matches the variations in data from *figure 5a*.



(a) Dataset



(b) GP regression with SE kernel with $nlml = -19.22$



(c) GP regression with SE+SE kernel with $nlml = -66.4$

Figure 5

Observing *figure 5c*, it can be seen from the hyperparameters that in one direction there is a trade-off between a high length-scale SE with a low light-scale SE. This would explain why the *nlml* is lower in this case. Because the model is now becoming more smoother while still adjusting for the data. The main improvement in the prediction can be seen at the extremes of the $X1$ and $X2$ were with only SE kernel the prediction rapidly diverges while for SE + SE kernel it tries to keep its shape. However, this comes at the expense of the computing time which for this covariance it is $\approx 90\%$ longer than the SE only covariance.

WORDCOUNT = 997

References

[RW06a] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. the MIT Press, 2006. Chap. Chapter 1.

[RW06b] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. the MIT Press, 2006. Chap. Chapter 2.

[RW06c] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. the MIT Press, 2006. Chap. Chapter 5.

[LV] Samuel Liebana(scl63) and Mihai Varsandan. *Graph*. Matlab. [Computed Together on 01/11/2019].

7 Appendix

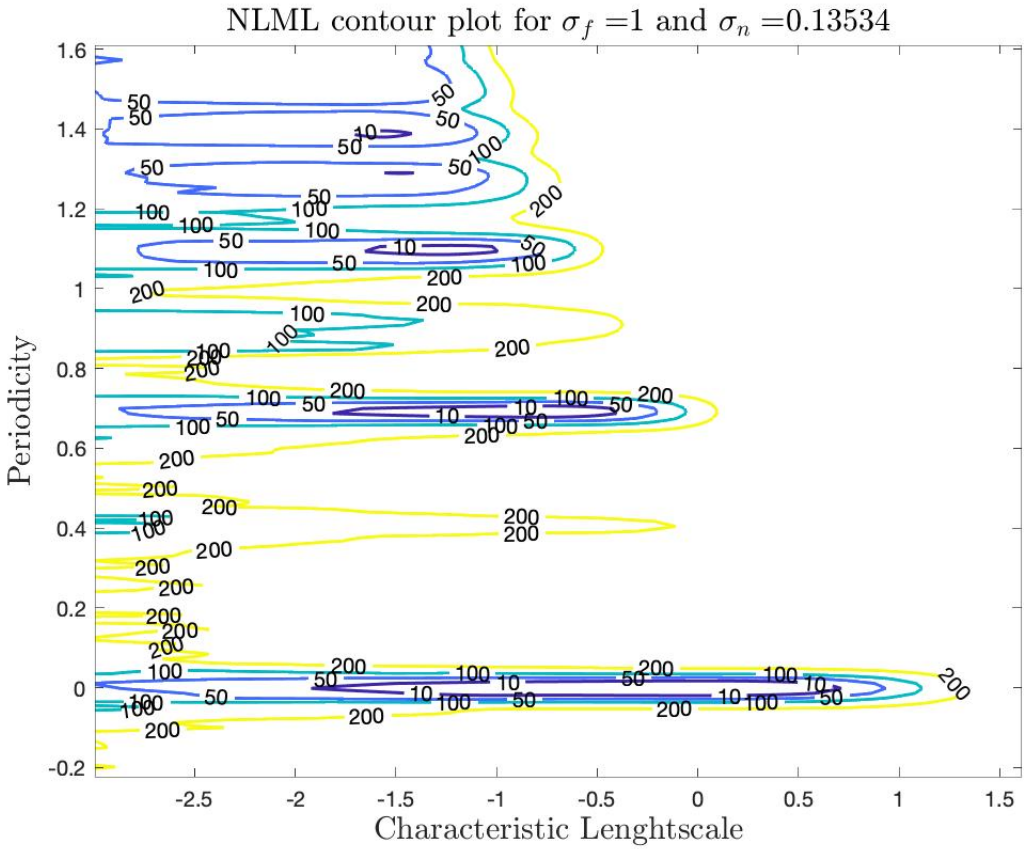


Figure 6: Contour plot for minimising nlml with a periodic kernel