4. Model Selection : Maximizing the Log Marginal Likelihood

Bookmark this page

Exercises due May 21, 2021 19:59 EDT

Maximizing the Log Marginal Likelihood

With this approach for model selection, one directly considers the data as a whole. For each possible parameter, one obtains a Gaussian model, and one tests how likely it is that the data comes from such a Gaussian model. Thus, one computes the score for the model according to

$$\log p\left(\mathbf{X}_2|\mathbf{Y}_2,\theta\right) = -\frac{1}{2}\mathbf{X}_2^\intercal\Sigma_{22}^{-1}\mathbf{X}_2 - \frac{1}{2}\log|\Sigma_{22}| - \frac{N-d}{2}\log\left(2\pi\right)$$

The relation above is essentially computing how well the covariance of the data matches the covariance we are building with our kernel function, as parameterized by θ.

We can recognize two terms for which some intuition can be gained:

$$\log p\left(\mathbf{X}_2|\mathbf{Y}_2,\theta\right) = \underbrace{-\frac{1}{2}\mathbf{X}_2^\intercal\Sigma_{22}^{-1}\mathbf{X}_2}_{\text{Data fitting}} - \underbrace{\frac{1}{2}\log|\Sigma_{22}|}_{\text{Complexity Penalty}} - \frac{N-d}{2}\log\left(2\pi\right)$$

The first term, $\frac{1}{2}\mathbf{X}_2^\intercal\Sigma_{22}^{-1}\mathbf{X}_2$, can be roughly interpreted as part of the cost that represents how closely the data fits the prediction. Meanwhile, the second term, $\frac{1}{2}\log|\Sigma_{22}|$, serves as a measure of the complexity of the model.

Fitting term

1 point possible (graded)
Consider the fitting term

$$-\frac{1}{2}\mathbf{X}_2^\intercal\Sigma_{22}^{-1}\mathbf{X}_2.$$

For this problem, let $\Sigma_{22} = sI$. As s increases, what happens to the fitting term as defined above? (Remember to account for the negative sign.)

- Fitting term increases
- Fitting term decreases

Submit You have used 0 of 1 attempt

Complexity term

1 point possible (graded)

Consider the complexity term

$$-\frac{1}{2}\log|\Sigma_{22}|$$

For this problem, let $\Sigma_{22} = sI$. As s increases, what happens to the complexity term as defined above?

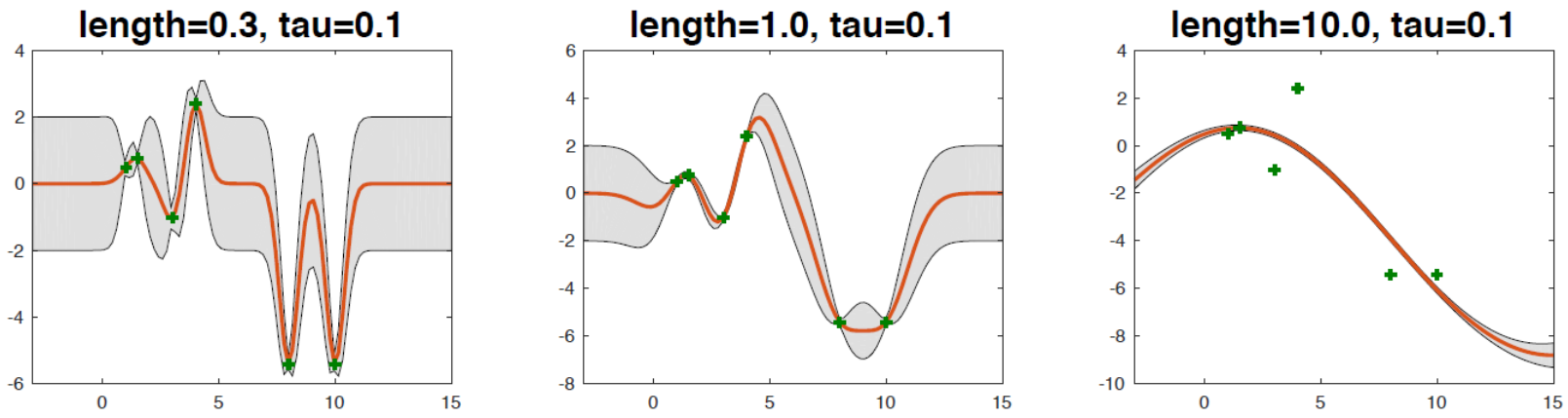
- Complexity term increases
- Complexity term decreases

Submit

You have used 0 of 1 attempt

To visualize such behavior in the two terms described above, let us recall the impact of the parameter ℓ in the example from the last lecture. The below figure shows the predicted distributions with different values for the parameter ℓ in the kernel function.

$$k(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)$$



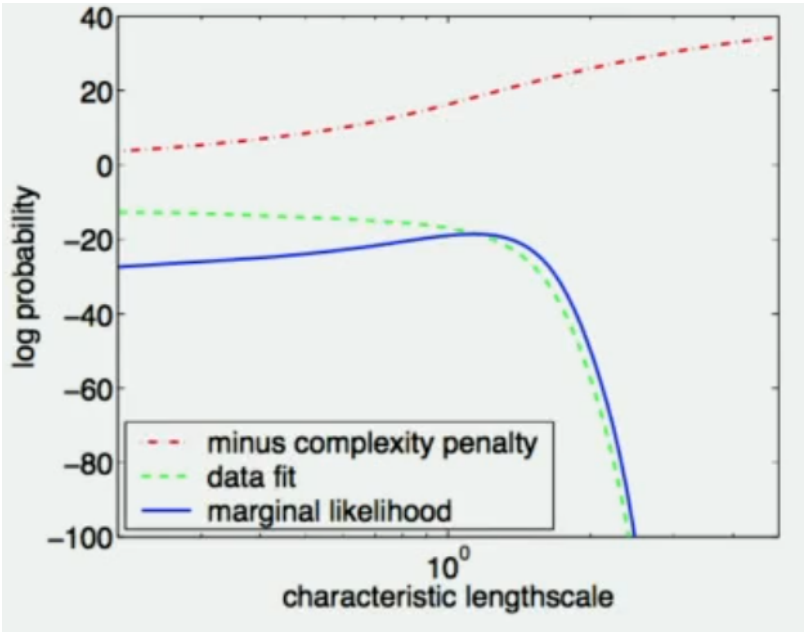
kernel function determines shape of interpolation

40: Effects of the parametrization of a kernel function

Initially, we can observe that as the parameter ℓ is smaller, the red line – corresponding to the predicted mean – fits the observed points precisely, but the red curve interpolates poorly as it deviates back to zero frequently creating a complex shape. On the other hand, for large values of ℓ, the model becomes “simpler" as the predicted behavior is rather smooth with slow changes. However, as a consequence, the predicted mean does not match the observed values.

Consider, also, that one should expect that if one removes an observed point, then the resulting model should not be “too different" under some appropriate smoothness assumptions. Note that the model with a small ℓ could produce a dramatically different model by removing one of the points. While for the model generated with a large ℓ this might not be the case. Thus, we can create some intuition on the tension between data fitting and model complexity.

The below figure shows the log marginal likelihood for the model described in the above figure, with the two separate components of the likelihood shown independently. The first, data fitting term is shown in green. We can see that as the value of the parameter ℓ is increased, the data fitting term reduces, while at the same time, the second, negative complexity penalty term, as shown in red, increases as the complexity is reduced.



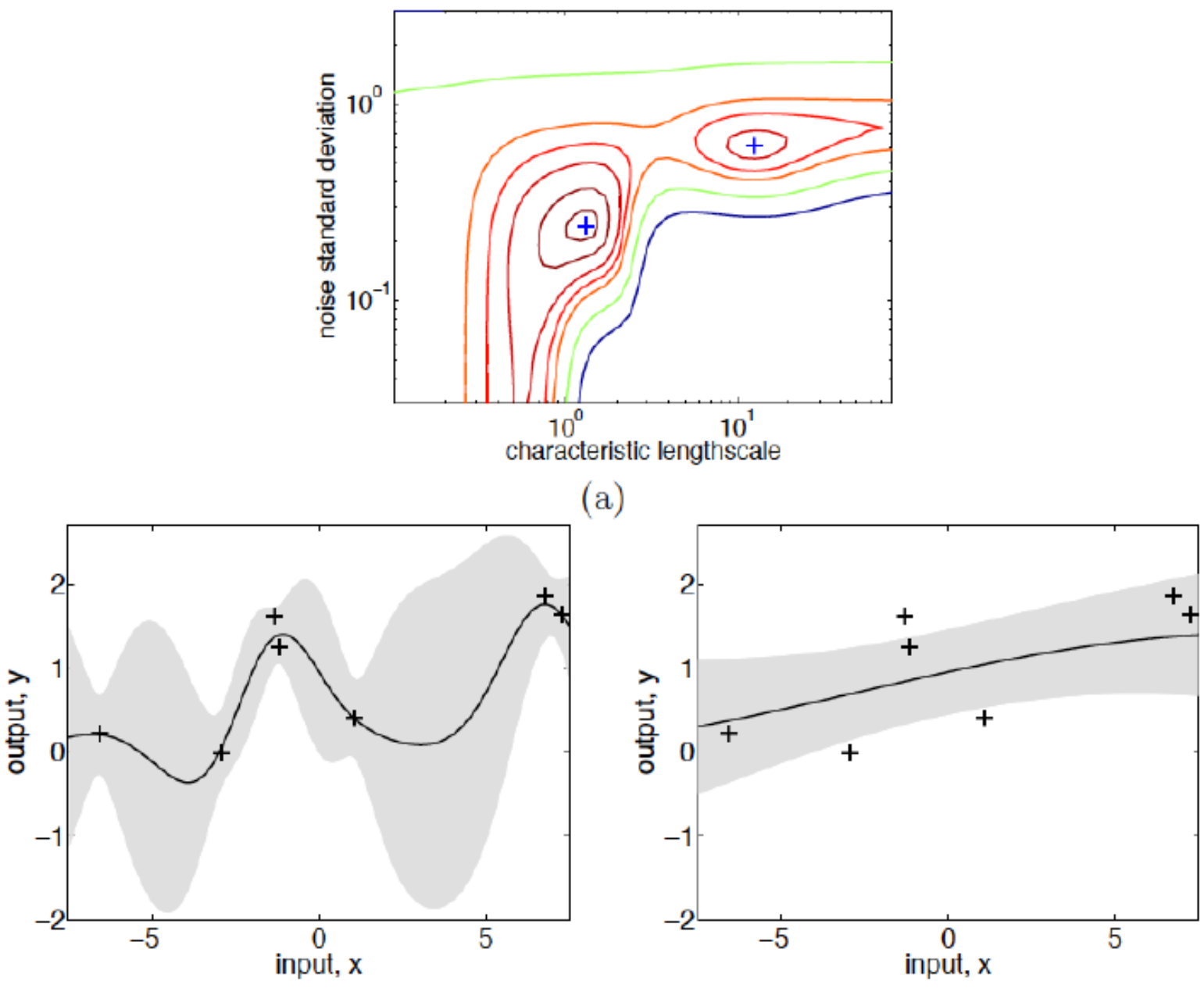
41: Effects of the parameter on data fitting and complexity of the obtained model.

Ideally, one would solve the optimization problem

$$\max_{\theta} \log p\left(\mathbf{X}_2 | \mathbf{Y}_2, \theta\right).$$

That is, find the parameter θ that maximizes the log-likelihood of the observed random variables. However, such an optimization problem is not simple, and generally non-convex. Thus, advanced optimization procedures may need to be used.

In the below figure, the top image shows the marginalized log-likelihood for different parameter values of ℓ and noise τ.



42: Finding the best parameter.

One can observe that two points correspond to the local maximum. The predictions generated by the parameters corresponding to those maximum points are shown in the bottom images of the figure to the left and right respectively. From this, we can see that these maxima produce substantially different models.

Discussion

Hide Discussion

Topic: Module 5: Environmental Data and Gaussian Processes:Sensing and Analyzing global patterns of dependence / 4. Model Selection : Maximizing the Log Marginal

Likelihood

Add a Post

Show all posts by recent activity

There are no posts in this topic yet.

✖

‹ Previous Next ›