What is a kernel?

Theorem (Leove):

k corresponds to the covariance of a GP.

k is a symmetric positive semi-definite function.

when k is a function of $x - y$, the kernel is called stationary, σ is called the variance and θ is called the lengthscale. And it is quite import to look at the length scale after the optimization step, if the value is quite small that means your model will gain no information from the surrounding, and thus no good prediction for the testing points.

Choosing appropriate kernel

In order to choose a kernel, one should gather all possible information about the function to approximate:

- Is it stationary ?
- Is it differentiable, what's its regularity ?
- Do we expect particular trends ?
- Do we expect particular patterns (periodicity, cycles, additivity) ?

Kernels often include rescaling parameters : θ for the x axis (length-scale) and σ for the y (σ^2 often corresponds to the GP variance). They can be tuned by

- maximizing the likelihood
- minimizing the prediction error

It is common to try various kernels and to asses the model accuracy. The idea is to compare some model predictions against actual values :

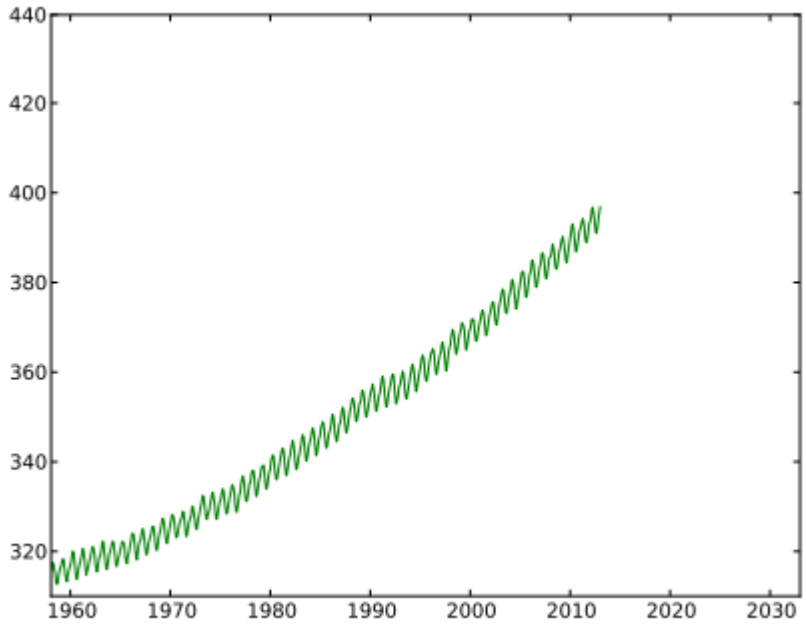
- On a test set
- Using leave-one-out

Furthermore, it is often interesting to try some input remapping such as $x \to \log(x), x \to \exp(x)$ to make our data set stationary, and then choose to use the stationary kernel.

Making new from old

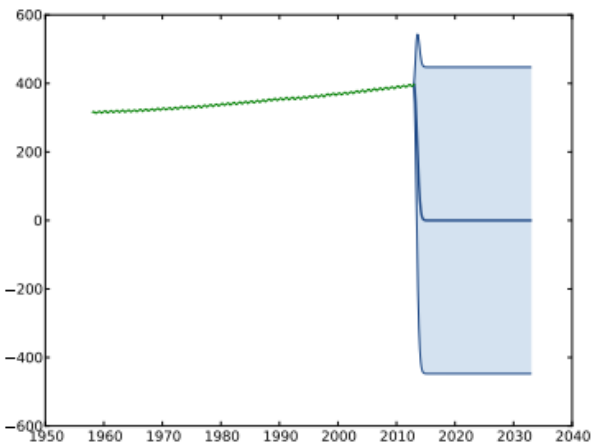
- Summed together
  - On the same space k(x, y) = k1(x, y) + k2(x, y)
  - On the tensor space k(x, y) = k1(x1, y1) + k2(x2, y2)
- Multiplied together
  - On the same space k(x, y) = k1(x, y) × k2(x, y)
  - On the tensor space k(x, y) = k1(x1, y1) × k2(x2, y2)
- Composed with a function
  - k(x, y) = k1(f (x), f (y))

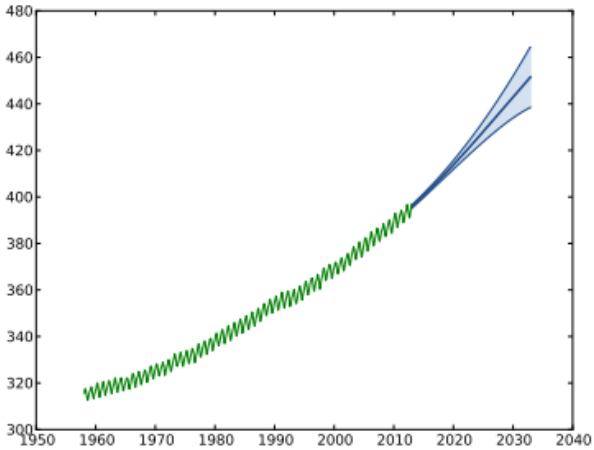
Example: CO_2



- First, we consider a squared-exponential kernel:

$$k(x, y) = \sigma^2 \exp\left(-\frac{(x-y)^2}{\theta^2}\right) \tag{10}$$



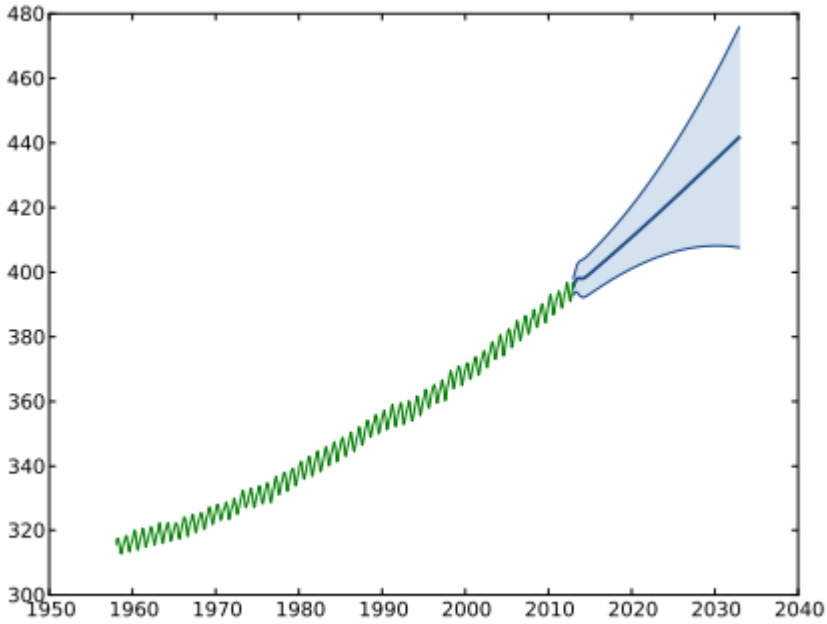


First, we would like to say that we observe the high frequency in the data, so we would like to choose a very small length scale value, the result is shown as the left picture. The reason is that, you choose the small scale value, when you are in 2020, the model will get no information from the data set and not influenced by the past values.

Second choice is to be focus on the trend, with low frequency which would lead to a very large length scale value, as shown in picture right. But the confidence interval is over confident.

- Second, we sum both kernels together.

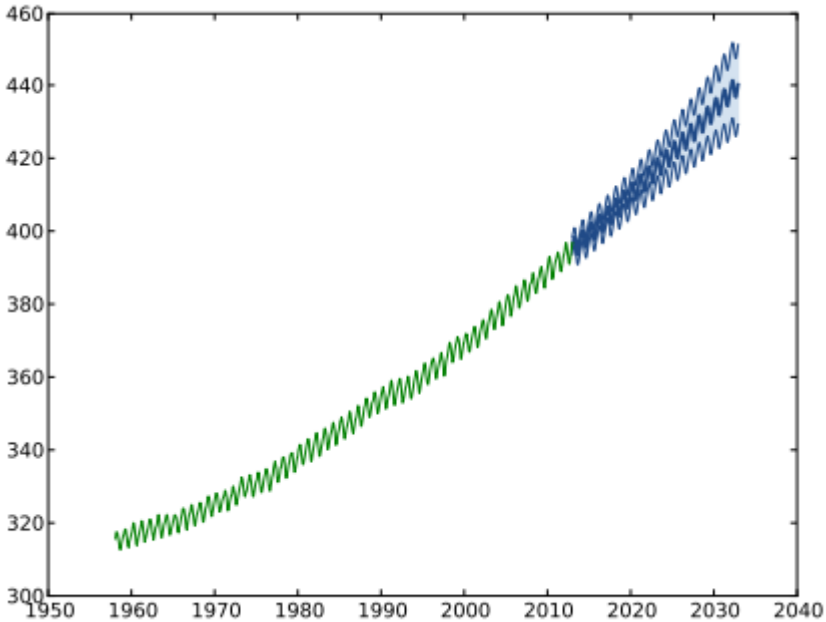
$$k(x, y) = k_{rbf1}(x, y) + k_{rbf2}(x, y) \tag{1}$$



One thing to notice that, even in the second choice (combination choice), it seems that we would have more parameters, but indeed the optimization process will get easier than the first choice. Because for the first two model, the likelihood will get very very small, since both assumptions make sense for data we have.

- Then, adding the periodic term into the kernel.

$$k(x,y) = \sigma_0^2 x^2 y^2 + k_{rbf1}(x,y) + k_{rbf2}(x,y) + k_{per}(x,y) \tag{2}$$

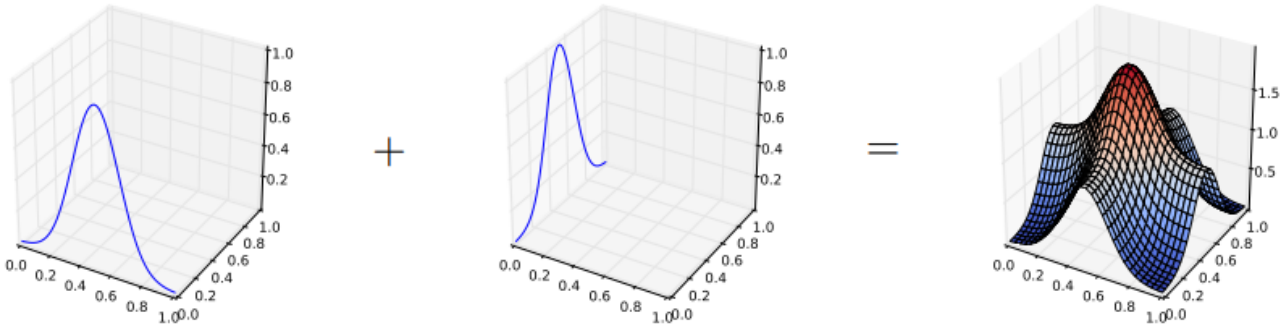


Sum of kernels over tensor space

Property:

$$k(\mathbf{x},\mathbf{y}) = k_1(x_1,y_1) + k_2(x_2,y_2) \tag{3}$$

is a valid covariance structure.



Tensor Additive kernels are very useful for:

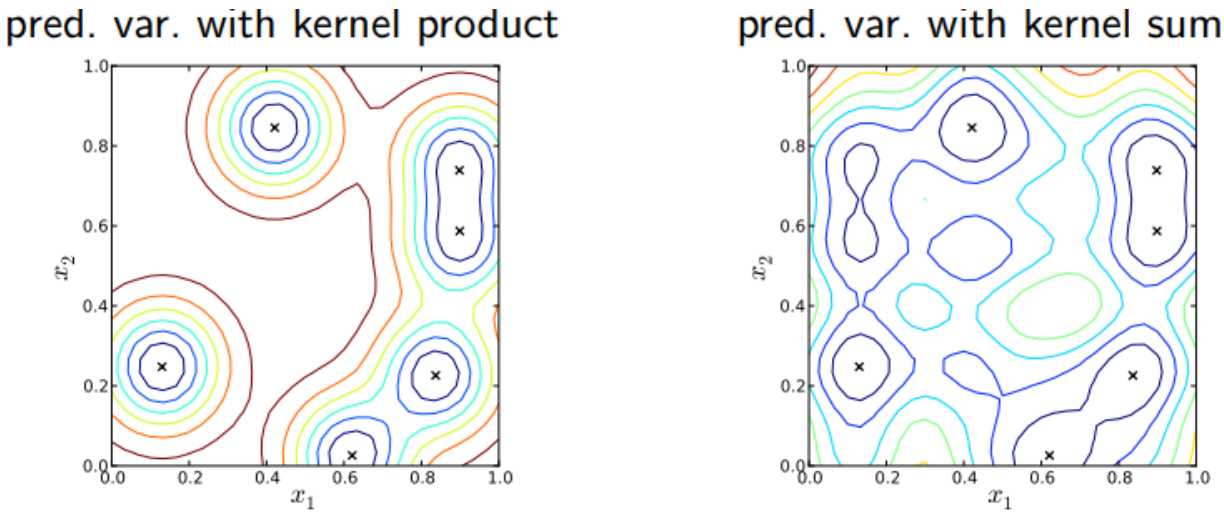
- Approximating additive functions
- Building models over high dimensional input space

Remark:

1. From a GP point of view, k is the kernel of $Z(x) = Z(x_1) + Z(x_2)$.
2. It is straightforward to show that the mean predictor is additive.

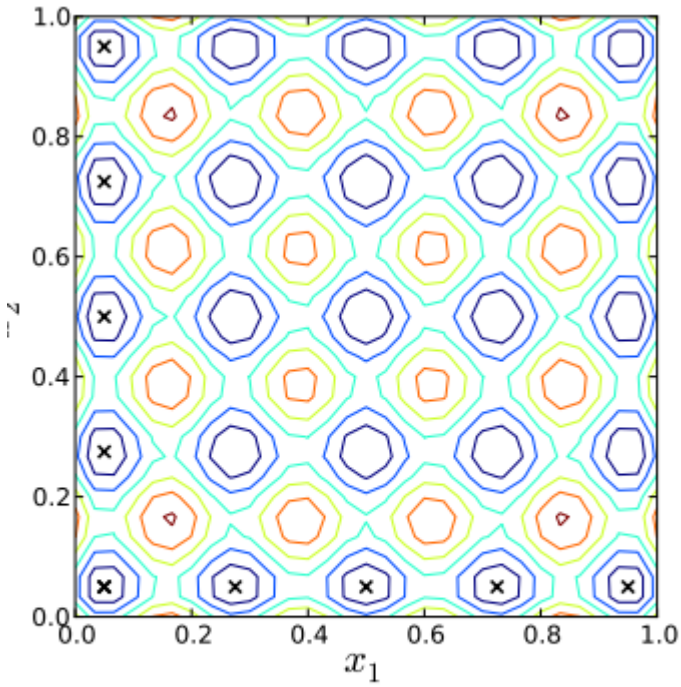
$$
\begin{aligned}
m(\mathbf{x}) &= \left(k_1(x, X) + k_2(x, X)\right)\left(k(X, X)\right)^{-1} F \\
&= \underbrace{k_1\left(x_1, X_1\right)\left(k(X, X)\right)^{-1} F}_{m_1(x_1)} + \underbrace{k_2\left(x_2, X_2\right)\left(k(X, X)\right)^{-1} F}_{m_2(x_2)}
\end{aligned}
$$

3. The prediction variance has interesting features.



The right one comes from a additive kernel, as we can see even in the area which is away from the observation points, the variance is not too high. The reason for that our prior, e.g. kernel is additive, we already three observations which would form a rectangle, and our prediction would be the fourth vertex, thus the variance would be small. **All the prior would retrieve it in the posterior**.

This property can be used to construct a design of experiment that covers the space especially for the high-D input space, with only $cst \times d$ points.



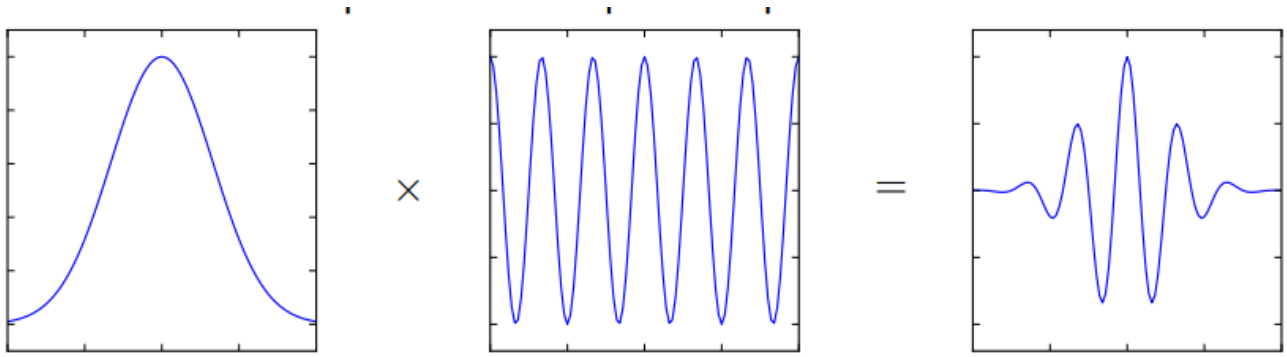
Prediction variance

Product over the same space

Property:

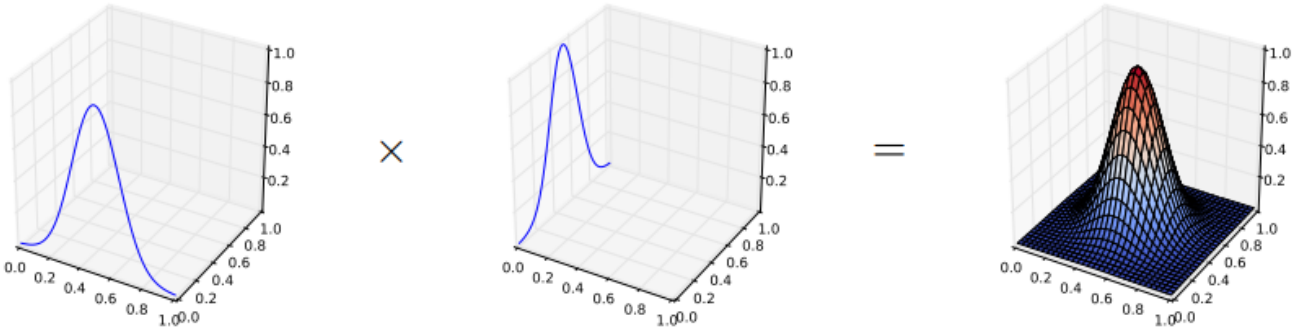
$$k(x, y) = k_1(x, y) \times k_2(x, y) \tag{4}$$

is valid covariance structure.



Product over the tensor space

$$k(\mathbf{x}, \mathbf{y}) = k_1\left(x_1, y_1\right) \times k_2\left(x_2, y_2\right) \tag{5}$$



Composition with a function

$$k(x,y) = k_1(f(x), f(y)) \tag{6}$$
$$Proof:$$
$$\sum\sum a_i a_j k(x_i, x_j) = \sum\sum a_i a_j k_1\left(\underbrace{f(x_i)}_{y_i}, \underbrace{f(x_j)}_{y_j}\right) \geq 0$$

This can be seen as a nonlinear rescaling of the input space.

Periodicity detection

Given a few observations can we extract the periodic part of a signal ?

As previously we will build a decomposition of the process in two independent GPs :

$$Z = Z_p + Z_a \tag{7}$$

where Z_p is a GP in the span of the Fourier basis

$$B(t) = (\sin(t), \cos(t), \ldots, \sin(nt), \cos(nt))^t \tag{8}$$

Note that the aperiodic means the projection of the $cos - sin$ space will end up with zero.

And it can be proved that

$$\begin{aligned} k_p(x,y) &= B(x)^t G^{-1} B(y) \\ k_a(x,y) &= k(x,y) - k_p(x,y) \end{aligned} \tag{9}$$

where G is the Gram Matrix associated to B in the RKHS.

As previously, a decomposition of the model comes with a decomposition of the kernel:

$$\begin{aligned} m(t) =& (k_p(x,X) + k_a(x,X))\, k(X,X)^{-1} F \\ =& k_p(x,X)k(X,X)^{-1}F + \underbrace{k_a(x,X)k(X,X)^{-1}F}_{\text{aperiodic sub-model } m_a} \end{aligned}$$

and we can associate a prediction variance to the sub-models:

$$v_p(t) = k_p(x,x) - k_p(x,X)^t k(X,X)^{-1} k_p(t)$$
$$v_a(t) = k_a(x,x) - k_a(x,X)^t k(X,X)^{-1} k_a(t)$$