Gaussian Process Summer School

Kernel Design

Nicolas Durrande – PROWLER.io (nicolas@prowler.io)

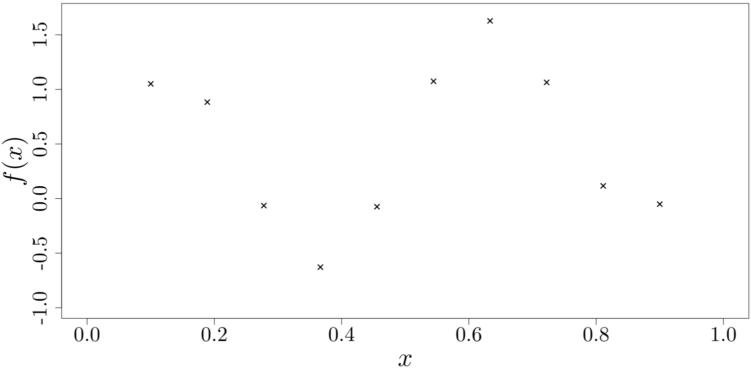
Sheffield, September 2017

We have seen during the introduction lectures that the distribution of a GP Z depends on two functions :

- the mean $m(x) = \mathrm{E}\left(Z(x)\right)$
- the covariance $k(x, x') = \mathrm{cov}\left(Z(x), Z(x')\right)$

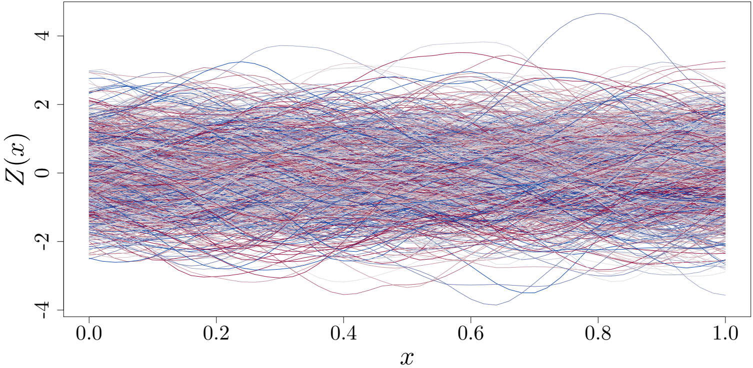
In this talk, we will focus on the **covariance function**, which is often call the **kernel**.

We assume we have observed a function f for a limited number of time points $x_1, \dots, x_n$:

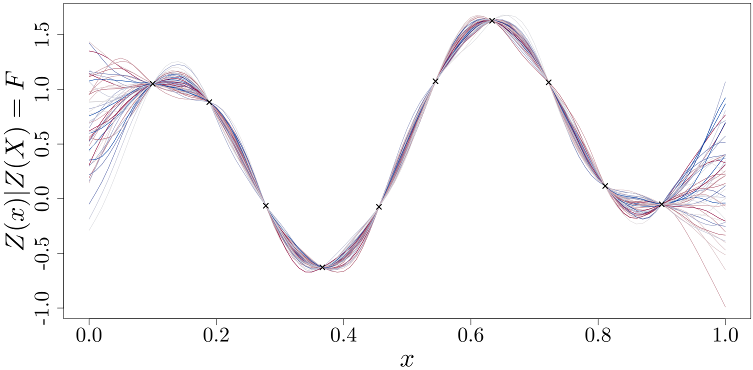


The observations are denoted by $f_i = f(x_i)$ (or $F = f(X)$).

Since f in unknown, we make the general assumption that it is a sample path of a Gaussian process Z :



Combining these two informations means keeping the samples interpolating the data points :

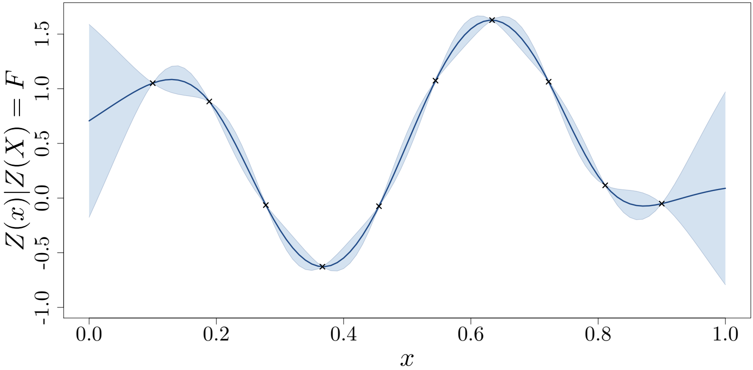


The conditional distribution is still Gaussian with moments :

$$m(x) = \mathrm{E}\left(Z(x)|Z(X)\text{=}F\right) = k(x, X)k(X, X)^{-1}F$$

$$c(x, x') = \mathrm{cov}\left(Z(x), Z(x')|Z(X)\text{=}F\right) = k(x, x') - k(x, X)k(X, X)^{-1}k(X, x')$$

It can be represented as a mean function with confidence intervals.



Let Z be a random process with kernel k. Some properties of kernels can be obtained directly from their definition.

Example

$$k(x,x) = \mathrm{cov}\left(Z(x), Z(x)\right) = \mathrm{var}\left(Z(x)\right) \geq 0$$
$$\Rightarrow k(x,x) \text{ is } \textbf{positive}.$$

$$k(x,y) = \mathrm{cov}\left(Z(x), Z(y)\right) = \mathrm{cov}\left(Z(y), Z(x)\right) = k(y,x)$$
$$\Rightarrow k(x,y) \text{ is } \textbf{symmetric}.$$

We can obtain a thinner result...

We introduce the random variable $T = \sum_{i=1}^{n} a_i Z(x_i)$ where n, a_i and x_i are arbitrary. Computing the variance of T gives :

$$\begin{aligned} \mathrm{var}(T) &= \mathrm{cov}\left(\sum_i a_i Z(x_i), \sum_j a_j Z(x_j)\right) = \sum_i \sum_j a_i a_j \mathrm{cov}\left(Z(x_i), Z(x_j)\right) \\ &= \sum \sum a_i a_j k(x_i, x_j) \end{aligned}$$

Since a variance is positive, we have

$$\sum_i \sum_j a_i a_j k(x_i, x_j) \geq 0$$

for any arbitrary n, a_i and x_i.

Definition

The functions satisfying the above inequality **for all** $n \in \mathbb{N}$, **for all** $x_i \in D$, **for all** $a_i \in \mathbb{R}$ are called positive semi-definite functions.

We have just seen :

$$k \text{ is a covariance} \Rightarrow k \text{ is a positive semi-definite function}$$

The reverse is also true :

Theorem (Loeve)

k corresponds to the covariance of a GP

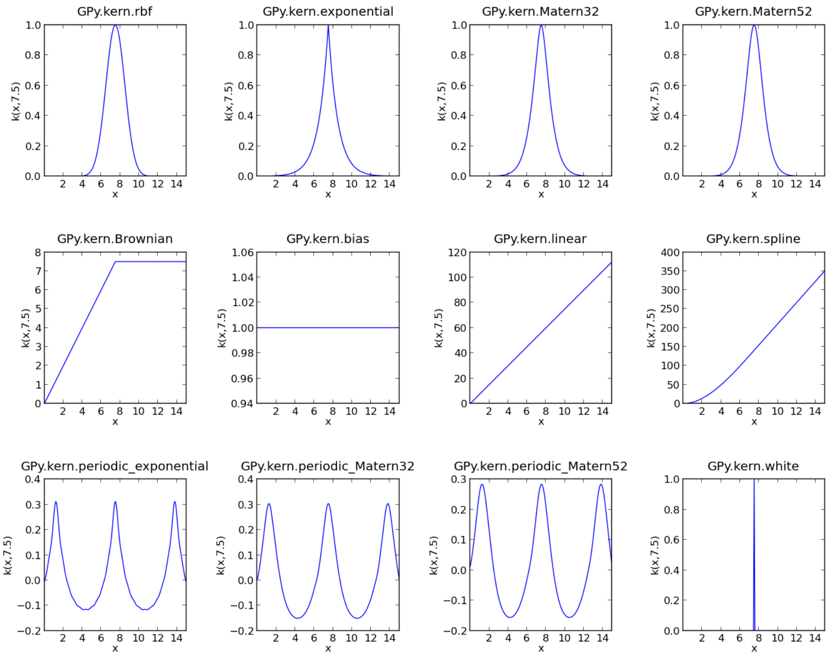
$$\Updownarrow$$

k is a symmetric positive semi-definite function

Proving that a function is psd is often difficult. However there are a lot of functions that have already been proven to be psd :

$$
\begin{aligned}
\text{squared exp.} \quad & k(x, y) = \sigma^2 \exp\left(-\frac{(x-y)^2}{2\theta^2}\right) \\
\text{Matern 5/2} \quad & k(x, y) = \sigma^2 \left(1 + \frac{\sqrt{5}|x-y|}{\theta} + \frac{5|x-y|^2}{3\theta^2}\right) \exp\left(-\frac{\sqrt{5}|x-y|}{\theta}\right) \\
\text{Matern 3/2} \quad & k(x, y) = \sigma^2 \left(1 + \frac{\sqrt{3}|x-y|}{\theta}\right) \exp\left(-\frac{\sqrt{3}|x-y|}{\theta}\right) \\
\text{exponential} \quad & k(x, y) = \sigma^2 \exp\left(-\frac{|x-y|}{\theta}\right) \\
\text{Brownian} \quad & k(x, y) = \sigma^2 \min(x, y) \\
\text{white noise} \quad & k(x, y) = \sigma^2 \delta_{x,y} \\
\text{constant} \quad & k(x, y) = \sigma^2 \\
\text{linear} \quad & k(x, y) = \sigma^2 xy
\end{aligned}
$$

When k is a function of $x - y$, the kernel is called **stationary**.
σ^2 is called the **variance** and θ the **lengthscale**.


GPy.kern.rbf
GPy.kern.exponential
GPy.kern.Matern32
GPy.kern.Matern52
GPy.kern.Brownian
GPy.kern.bias
GPy.kern.linear
GPy.kern.spline
GPy.kern.periodic_exponential
GPy.kern.periodic_Matern32
GPy.kern.periodic_Matern52
GPy.kern.white
k(x,7.5)
x

For a few kernels, it is possible to prove they are psd directly from the definition.

- $k(x, y) = \delta_{x,y}$
- $k(x, y) = 1$

For most of them a direct proof from the definition is not possible. The following theorem is helpful for stationary kernels :

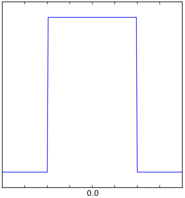
Theorem (Bochner)

A continuous stationary function $k(x, y) = \tilde{k}(|x - y|)$ is positive definite if and only if $\tilde{k}$ is the Fourier transform of a finite positive measure :

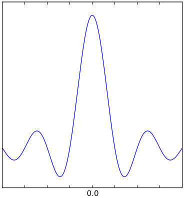
$$\tilde{k}(t) = \int_{\mathbb{R}} e^{-i\omega t} \mathrm{d}\mu(\omega)$$

Example

We consider the following measure :



Its Fourier transform gives $\tilde{k}(t) = \dfrac{\sin(t)}{t}$:



As a consequence, $k(x, y) = \dfrac{\sin(x - y)}{x - y}$ is a valid covariance function.

Usual kernels

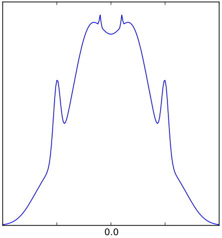
Bochner theorem can be used to prove the positive definiteness of many usual stationary kernels

- The Gaussian is the Fourier transform of itself
  $\Rightarrow$ it is psd.
- Matérn kernels are the Fourier transforms of $\frac{1}{(1+\omega^2)^p}$
  $\Rightarrow$ they are psd.

Unusual kernels

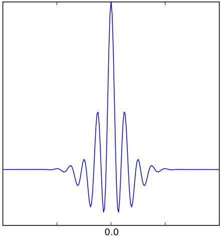
Inverse Fourier transform of a (symmetrised) sum of Gaussian gives (A. Wilson, ICML 2013) :

$\mu(\omega)$



$$\overrightarrow{\mathcal{F}}$$

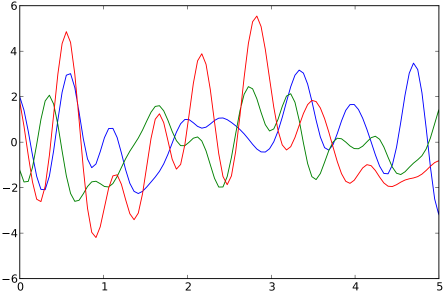
$\tilde{k}(t)$



The obtained kernel is parametrised by its spectrum.

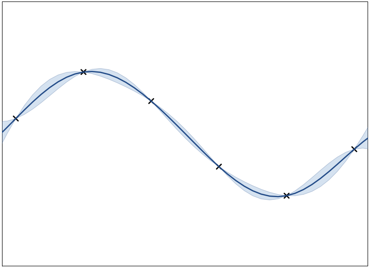
Unusual kernels

The sample paths have the following shape :

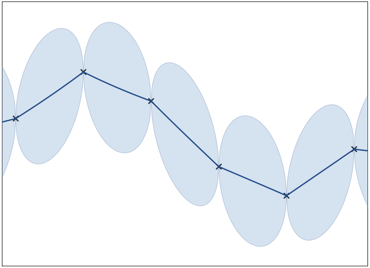


Changing the kernel has a huge impact on the model :

Gaussian kernel:

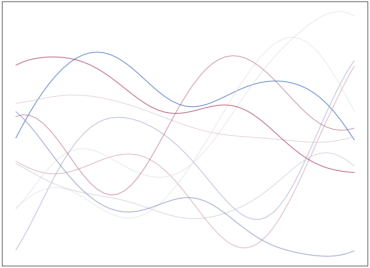


Exponential kernel:

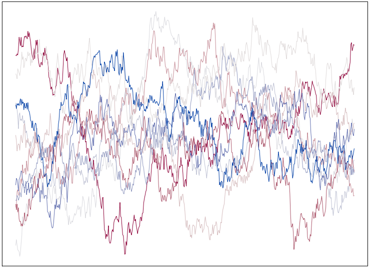


This is because changing the kernel implies changing the prior

Gaussian kernel:



Exponential kernel:



In order to choose a kernel, one should gather all possible informations 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

Two (ideally three) things should be checked :

- Is the mean accurate (MSE, Q^2) ?
- Do the confidence intervals make sense ?
- Are the predicted covariances right ?

Furthermore, it is often interesting to try some input remapping such as $x \to \log(x)$, $x \to \exp(x)$, ...

Making new from old :

Kernels can be :

- Summed together
  - On the same space $k(x, y) = k_1(x, y) + k_2(x, y)$
  - On the tensor space $k(\mathbf{x}, \mathbf{y}) = k_1(x_1, y_1) + k_2(x_2, y_2)$
- Multiplied together
  - On the same space $k(x, y) = k_1(x, y) \times k_2(x, y)$
  - On the tensor space $k(\mathbf{x}, \mathbf{y}) = k_1(x_1, y_1) \times k_2(x_2, y_2)$
- Composed with a function
  - $k(x, y) = k_1(f(x), f(y))$

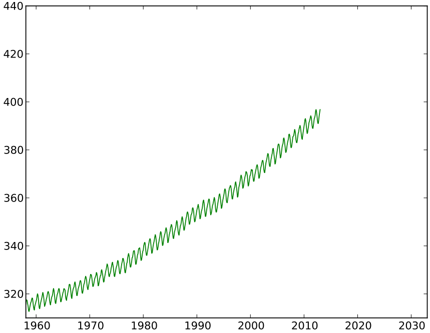
All these operations will preserve the positive definiteness.

How can this be useful ?

Sum of kernels over the same space

Example (The Mauna Loa observatory dataset)

This famous dataset compiles the monthly CO_2 concentration in Hawaii since 1958.

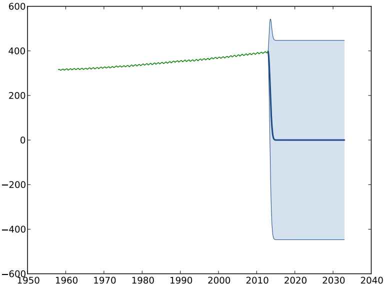


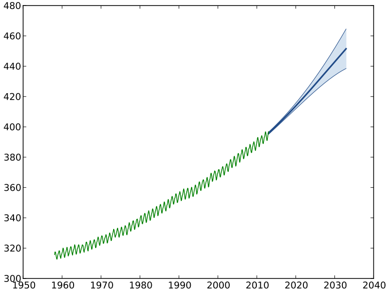
Let's try to predict the concentration for the next 20 years.

Sum of kernels over the same space

We first consider a squared-exponential kernel :

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





The results are terrible !

Sum of kernels over the same space

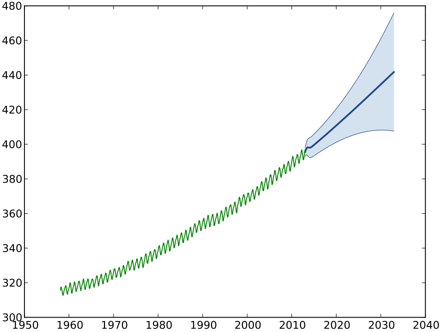
What happen if we sum both kernels ?

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

Sum of kernels over the same space

What happen if we sum both kernels ?

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



The model is drastically improved !

Sum of kernels over the same space

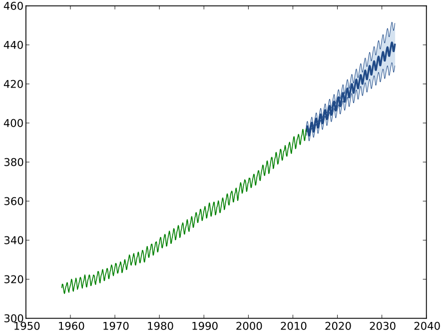
We can try the following kernel :

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

Sum of kernels over the same space

We can try the following kernel :

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



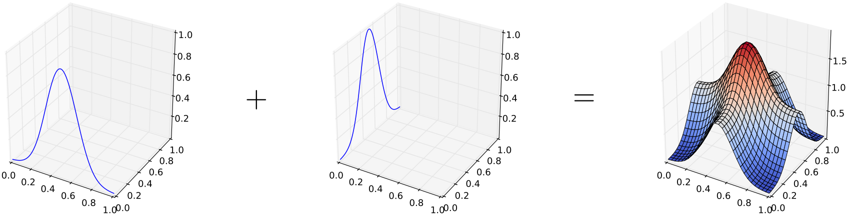
Once again, the model is significantly improved.

Sum of kernels over tensor space

Property

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

is a valid covariance structure.

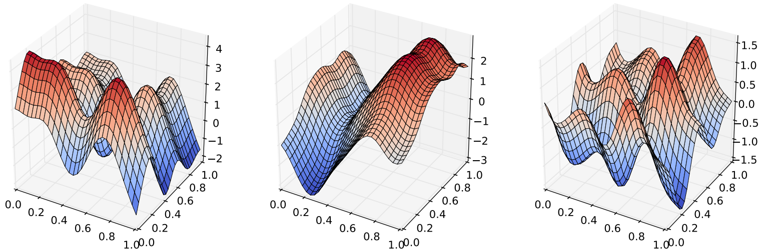


Remark :

- From a GP point of view, k is the kernel of $Z(\mathbf{x}) = Z_1(x_1) + Z_2(x_2)$

Sum of kernels over tensor space

We can have a look at a few sample paths from Z :



⇒ They are additive (up to a modification)

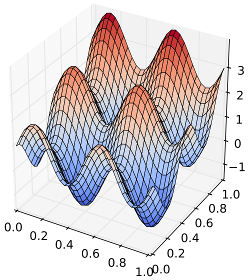
Tensor Additive kernels are very useful for

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

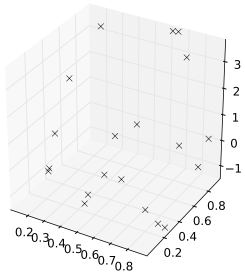
Sum of kernels over tensor space

We consider the test function $f(x) = \sin(4\pi x_1) + \cos(4\pi x_2) + 2x_2$ and a set of 20 observation in $[0,1]^2$

Test function



Observations

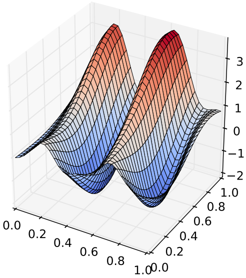


Sum of kernels over tensor space

We obtain the following models :

Gaussian kernel

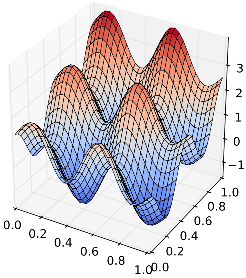
Mean predictor



RMSE is 1.06

Additive Gaussian kernel

Mean predictor



RMSE is 0.12

Sum of kernels over tensor space

Remarks

- It is straightforward to show that the mean predictor is additive

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

$\Rightarrow$ The model shares the prior behaviour.

- The sub-models can be interpreted as GP regression models with observation noise :

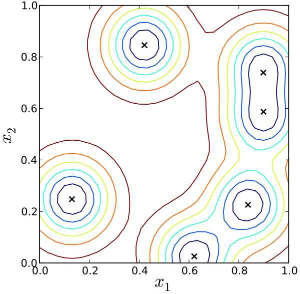
$$m_1(x_1) = \mathrm{E}\left(\ Z_1(x_1) \mid Z_1(X_1) + Z_2(X_2)\text{=}F\ \right)$$

Sum of kernels over tensor space

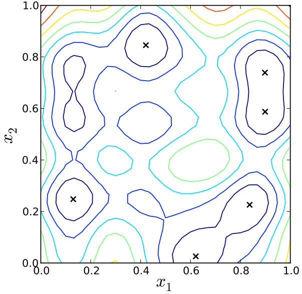
Remark

- The prediction variance has interesting features

pred. var. with kernel product

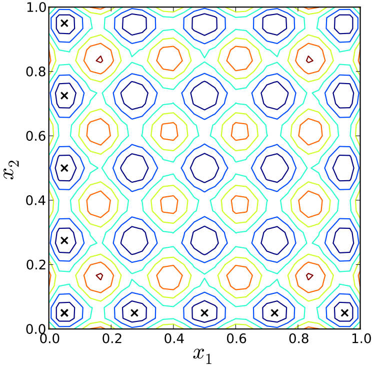


pred. var. with kernel sum



Sum of kernels over tensor space

This property can be used to construct a design of experiment that covers the space with only $cst \times d$ points.



Prediction variance

Product over the same space

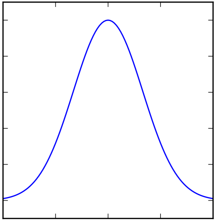
Property

$$k(x, y) = k_1(x, y) \times k_2(x, y)$$

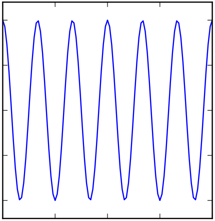
is valid covariance structure.

Example

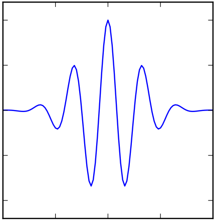
We consider the product of a squared exponential with a cosine :



$\times$



$=$



Product over the tensor space

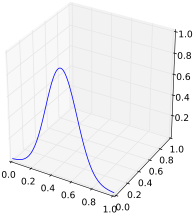
Property

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

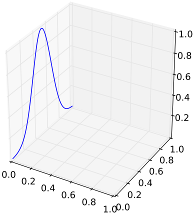
is valid covariance structure.

Example

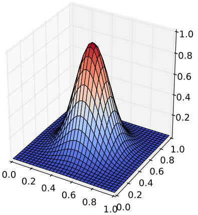
We multiply two squared exponential kernels



$\times$



$=$



Calculation shows we obtain the usual 2D squared exponential kernels.

Composition with a function

Property

Let k_1 be a kernel over $D_1 \times D_1$ and f be an arbitrary function $D \to D_1$, then

$$k(x,y) = k_1(f(x), f(y))$$

is a kernel over $D \times D$.
proof

$$\sum \sum a_i a_j k(x_i, x_j) = \sum \sum a_i a_j k_1(\underbrace{f(x_i)}_{y_i}, \underbrace{f(x_j)}_{y_j}) \geq 0$$

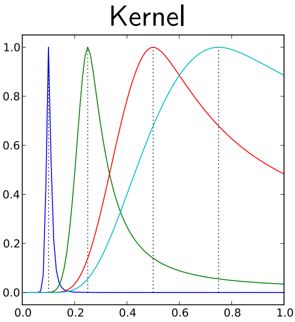
Remarks :

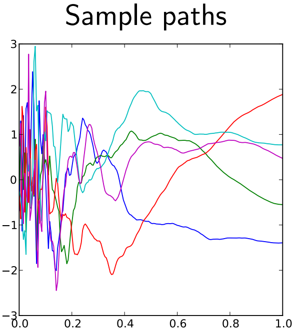
- k corresponds to the covariance of $Z(x) = Z_1(f(x))$
- This can be seen as a (nonlinear) rescaling of the input space

Example

We consider $f(x) = \frac{1}{x}$ and a Matérn 3/2 kernel $k_1(x, y) = (1 + |x - y|)e^{-|x-y|}$.

We obtain :



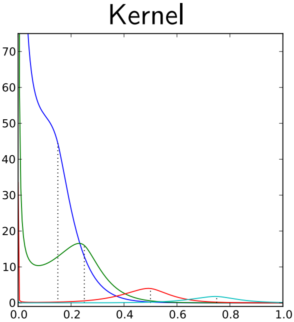


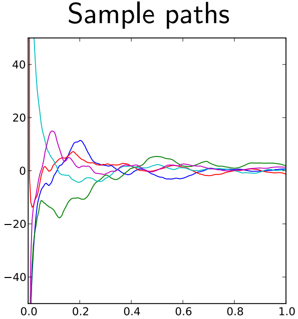
All these transformations can be combined !

Example

$k(x, y) = f(x)f(y)k_1(x, y)$ is a valid kernel.

This can be illustrated with $f(x) = \frac{1}{x}$ and $k_1(x, y) = (1 + |x - y|)e^{-|x-y|}$:





Effect of a linear operator

Property (Ginsbourger 2013)

Let L be a linear operator that commutes with the covariance, then $k(x, y) = L_x(L_y(k_1(x, y)))$ is a kernel.

Example

We want to approximate a function $[0, 1] \to \mathbb{R}$ that is symmetric with respect to 0.5. We will consider 2 linear operators :

$$L_1 : f(x) \to \begin{cases} f(x) & x < 0.5 \\ f(1-x) & x \geq 0.5 \end{cases}$$

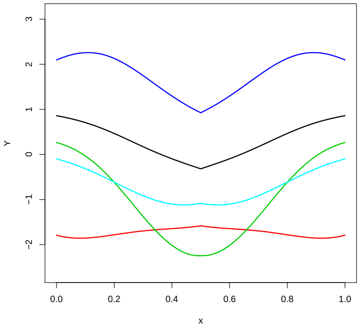
$$L_2 : f(x) \to \frac{f(x) + f(1-x)}{2}.$$

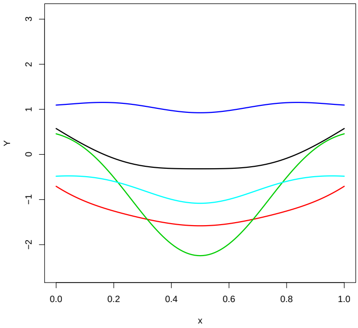
Effect of a linear operator

Example

Associated sample paths are

$$k_1 = L_1(L_1(k)) \qquad\qquad k_2 = L_2(L_2(k))$$

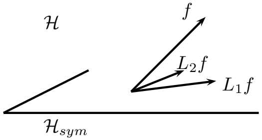




The differentiability is not always respected !

Effect of a linear operator

These linear operator are projections onto a space of symmetric functions :



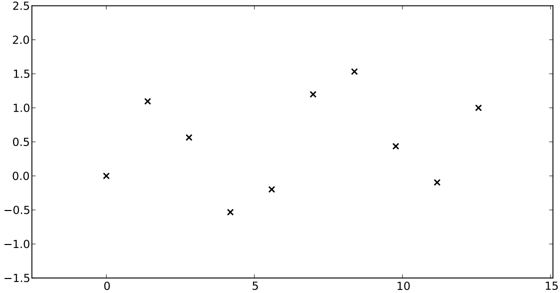
What about the optimal projection ?

⇒ This can be difficult... but it raises interesting questions !

Periodicity detection

We will now discuss the detection of periodicity

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 :

$$Z = Z_p + Z_a$$

where Z_p is a GP in the span of the Fourier basis $B(t) = (\sin(t), \cos(t), \dots, \sin(nt), cos(nt))^t$.

Property

It can be proved that the kernel of Z_p and Z_a are

$$k_p(x, y) = B(x)^t G^{-1} B(y)$$
$$k_a(x, y) = k(x, y) - k_p(x, y)$$

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

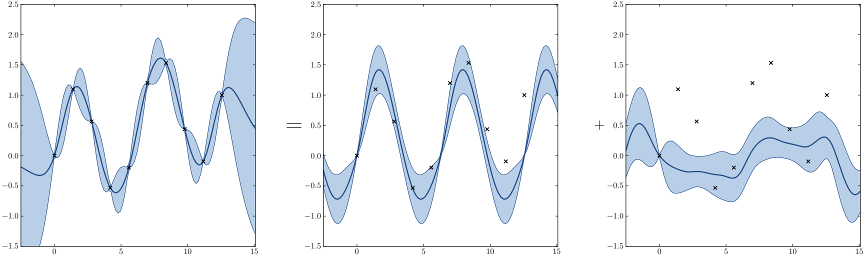
$$\begin{aligned} m(t) &= (k_p(x,X) + k_a(x,X))k(X,X)^{-1}F \\ &= \underbrace{k_p(x,X)k(X,X)^{-1}F}_{\text{periodic sub-model } m_p} + \underbrace{k_a(x,X)k(X,X)^{-1}F}_{\text{aperiodic sub-model } m_a} \end{aligned}$$

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

$$\begin{aligned} v_p(t) &= k_p(x,x) - k_p(x,X)^t k(X,X)^{-1} k_p(t) \\ v_a(t) &= k_a(x,x) - k_a(x,X)^t k(X,X)^{-1} k_a(t) \end{aligned}$$

Example

For the observations shown previously we obtain :



Can we can do any better ?

Initially, the kernels are parametrised by 2 variables :

$$k(x, y, \sigma^2, \theta)$$

but writing k as a sum allows to tune independently the parameters of the sub-kernels.

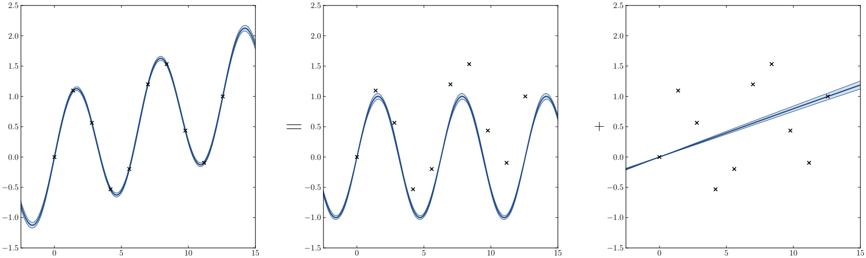
Let k^* be defined as

$$k^*(x, y, \sigma_p^2, \sigma_a^2, \theta_p, \theta_a) = k_p(x, y, \sigma_p^2, \theta_p) + k_a(x, y, \sigma_a^2, \theta_a)$$

Furthermore, we include a 5^{th} parameter in k^* accounting for the period by changing the Fourier basis :

$$B_\omega(t) = (\sin(\omega t), \cos(\omega t), \ldots, \sin(n\omega t), cos(n\omega t))^t$$

MLE of the 5 parameters of k^* gives :



We will now illustrate the use of these kernels for gene expression analysis.

We can apply this method to study the circadian rythm in organisms. We used *arabidopsis* data from Edward 2006.

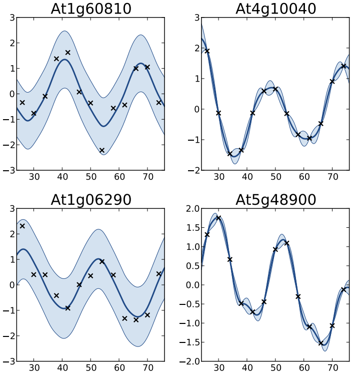
The dimension of the data is :

- 22810 genes
- 13 time points

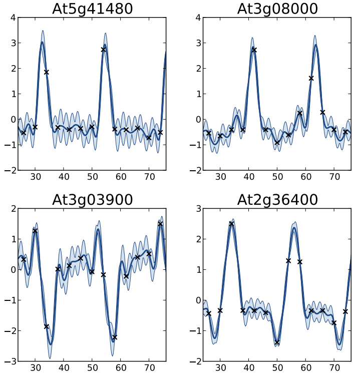
Edward 2006 gives a list of the 3504 most periodically expressed genes. The comparison with our approach gives :

- 21767 genes with the same label (2461 per. and 19306 non-per.)
- 1043 genes with different labels

Let's look at genes with different labels :



periodic for Edward



periodic for our approach

Small recap

We have seen that

- Kernels have a huge impact on the model
- They have to reflect the prior belief on the function to approximate.
- Kernels can (and should) be tailored to the problem at hand.

Although a direct proof of the positive definiteness of a function is often intractable, Bochner theorem allows to build kernels from their power spectrum.

Various operations can be applied to kernels while keeping p.s.d.ness :

Making new from old

- sum
- product
- composition with a function
- these can be combined

Linear operator

If we have a linear application that transforms any function into a function satisfying the desired property, it is possible to build a GP fulfilling the requirements.

C. E. Rasmussen and C. Williams
Gaussian Processes for Machine Learning, The MIT Press, 2006.

A. Berlinet and C. Thomas-Agnan
RKHS in probability and statistics, Kluwer academic, 2004.

N. Durrande, D. Ginsbourger, O. Roustant
Additive covariance kernels for high-dimensional Gaussian process modeling, AFST 2012.

N. Durrande, D. Ginsbourger, O. Roustant, L. Carraro
ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis, JMA 2013.

N. Durrande, J. Hensman, M. Rattray, N. D. Lawrence
Detecting periodicities with Gaussian processes. PeerJ Computer Science 2016.

D. Ginsbourger, X. Bay, L. Carraro and O. Roustant
Argumentwise invariant kernels for the approximation of invariant functions, AFST 2012.