Kernel Design

GP Summer School, Sheffield, September 2015

Nicolas Durrande, École des Mines de St-Étienne

We will first recall some definitions

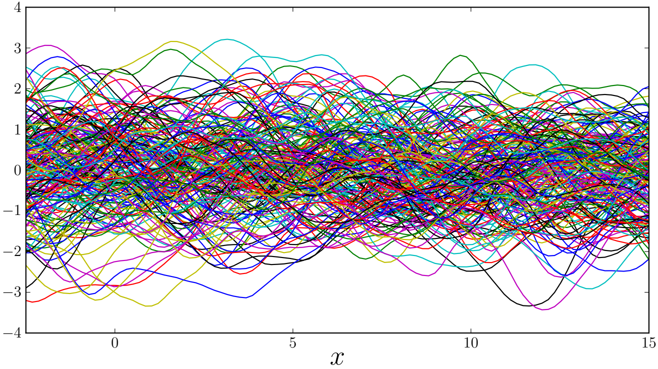
Gaussian process

A random process Z indexed by D is said to be Gaussian iif $(Z(x_1), \dots, Z(x_n))$ is multivariate normal $\forall x_i \in D$, $\forall n \in \mathbb{N}$.

Multivariate normal

A d-dimensional random vector Y is multivariate normal of $a^t Y$ is Gaussian distributed $\forall a \in \mathbb{R}^d$

Examples of a Gaussian process sample paths :

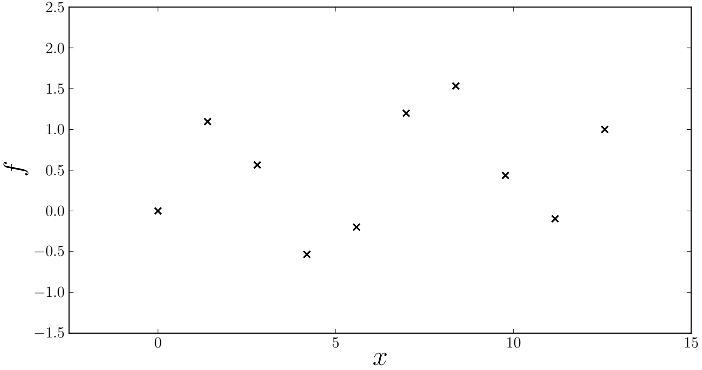


We have seen during the introduction lectures that the distribution of a GP Z depends on two functions :

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

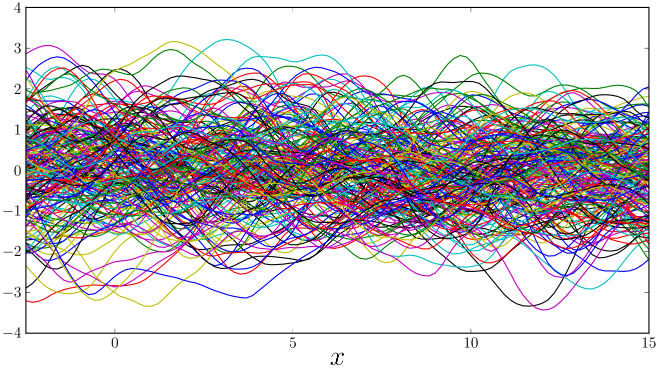
In this talk, we will focus on the **covariance function**, which is often call the **kernel**.

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

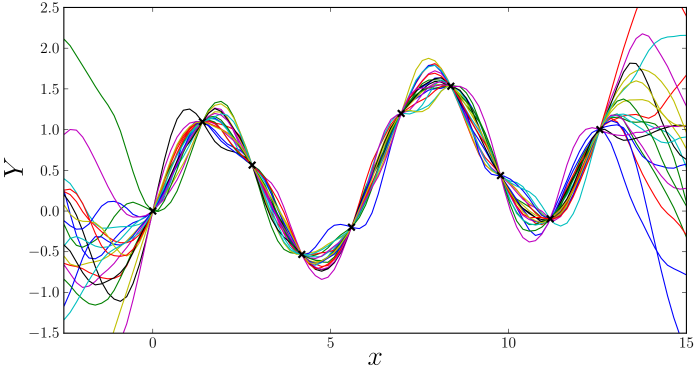


The observations are denoted by $f_i = f(x_i)$ (or $F = f(X)$).

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



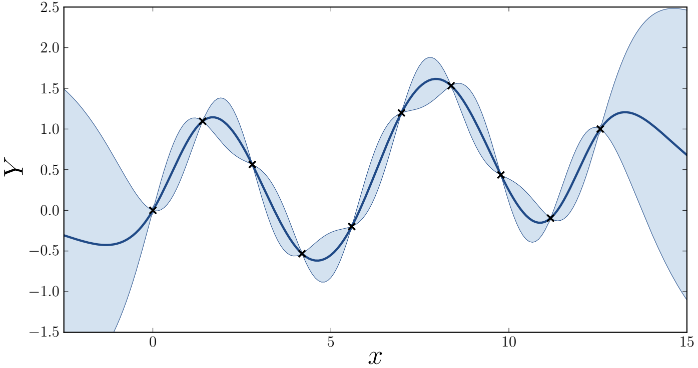
We can look at the sample paths of Z that interpolate the data points :



The conditional distribution is still Gaussian with moments :

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

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 :

$$\mathrm{var}\,(T) = \sum\sum a_i a_j \mathrm{cov}\,(Z(x_i), Z(x_j)) = \sum\sum a_i a_j k(x_i, x_j)$$

We thus have :

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

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 not assumed here that Z is Gaussian !

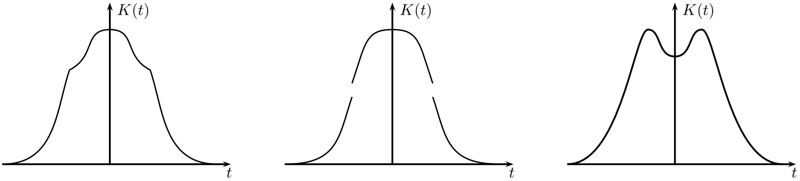
If k is stationary (ie $k(x, y) = \tilde{k}(|x - y|)$) psd implies further results :

Properties

- If $\tilde{k}$ is n times differentiable in 0, then it is n times differentiable everywhere.
- The maximum value of $\tilde{k}(t)$ is reached in $t = 0$.

Example

The following functions are not valid covariance structures



We have seen :

k is a covariance $\Rightarrow$ k 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

Positive semi definiteness is also a key concept in functional analysis leading to the theory of Reproducing Kernel Hilbert Spaces (RKHS).

Definition

The RKHS associated to a kernel k is the completion of

$$\left\{ \sum_{i=1}^{n} a_i k(x_i, .); n \in \mathbb{N}, a_i \in \mathbb{R}, x_i \in D \right\}$$

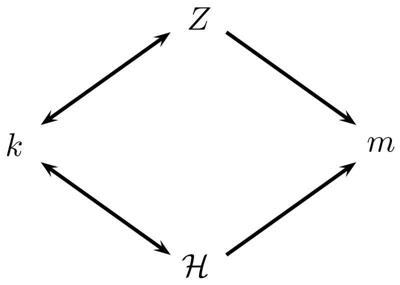
for the inner product

$$\left\langle \sum_{i=1}^{n} a_i k(x_i, .), \sum_{i=1}^{m} b_i k(x_i, .) \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j k(x_i, x_j)$$

Given some observations, the best predictor is defined as the interpolator with minimal norm :

$$m = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \{ ||h||_{\mathcal{H}}, h(x_i){=}f(x_i) \} = \cdots = k(x, X) k(X, X)^{-1} F$$

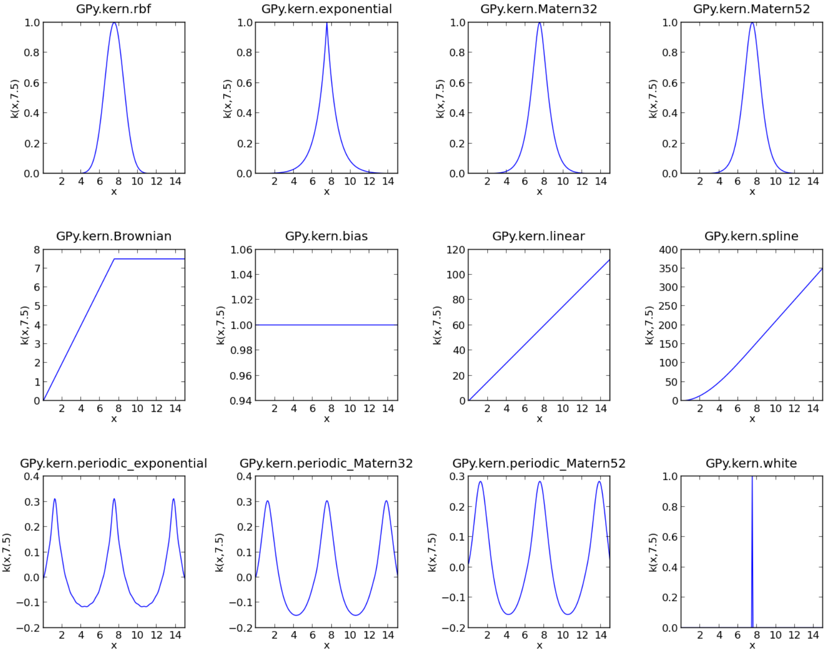
The expression is the same as the conditional expectation of the GP !



In order to build m we can consider any off the shelf kernel :

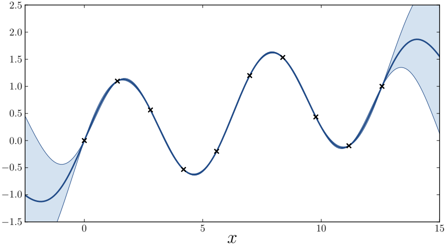
$$
\begin{aligned}
\text{white noise :} \quad & k(x,y) = \delta_{x,y} \\
\text{bias :} \quad & k(x,y) = 1 \\
\text{linear :} \quad & k(x,y) = xy \\
\text{exponential :} \quad & k(x,y) = \exp\left(-|x-y|\right) \\
\text{Brownian :} \quad & k(x,y) = \min(x,y) \\
\text{Gaussian :} \quad & k(x,y) = \exp\left(-(x-y)^2\right) \\
\text{Matérn 3/2 :} \quad & k(x,y) = (1+|x-y|) \times \exp\left(-|x-y|\right) \\
\text{sinc :} \quad & k(x,y) = \frac{\sin(|x-y|)}{|x-y|} \\
& \vdots
\end{aligned}
$$

Most of the above kernels are **stationary**.

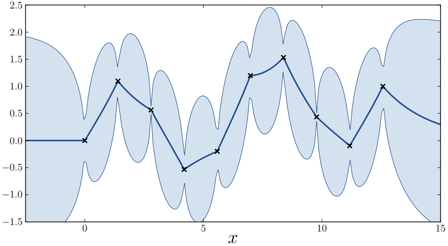

GPy.kern.rbf
GPy.kern.exponential
GPy.kern.Matern32
GPy.kern.Matern52
GPy.kern.Brownian
GPy.kern.bias
GPy.kern.linear
GPy.kern.spline
GPy.kern.periodic_exponential
GPy.kern.periodic_Matern32
GPy.kern.periodic_Matern52
GPy.kern.white
k(x,7.5)
x

Changing the kernel has a huge impact on the model :

exponentiated quadratic kernel

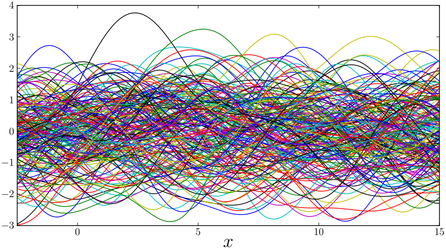


exponential kernel

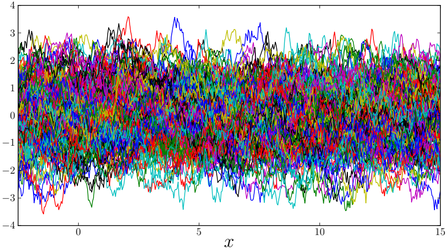


This is because changing the kernel implies changing the prior

exponentiated quadratic kernel



exponential kernel



Theorem (Bochner)

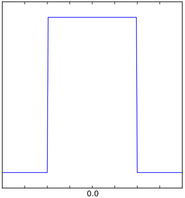
A continuous stationary function $k(x, y) = \tilde{k}(|x - y|)$ *is positive definite if and only if* $\tilde{k}$ *is the Fourier transform of a finite positive measure :*

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

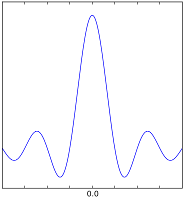
This result is very useful to prove the positive definiteness of stationary functions.

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.

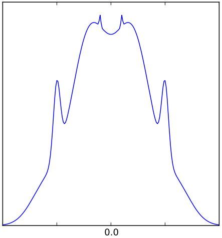
It can also be generalised to distributions :

- $\delta_{x,y}$ is the Fourier transform of the constant function
  $\Rightarrow$ it is psd.
- the constant function is the Fourier transform of $\delta_{x,y}$
  $\Rightarrow$ it is psd.

Unusual kernels

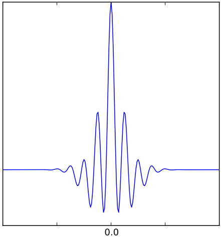
Inverse Fourier transform of a (symmetrised) sum of Gaussian gives

$\mu(\omega)$



$\overrightarrow{\mathcal{F}}$

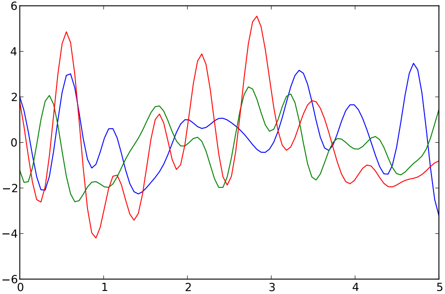
$\tilde{k}(t)$



The obtained kernel is parametrised by its spectrum [A. Wilson, ICML 2013].

Unusual kernels

The sample paths have the following shape :



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

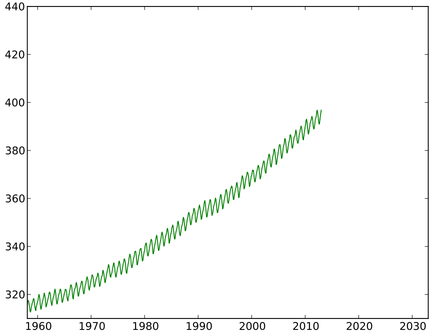
All these operations will preserve the positive definiteness.

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.

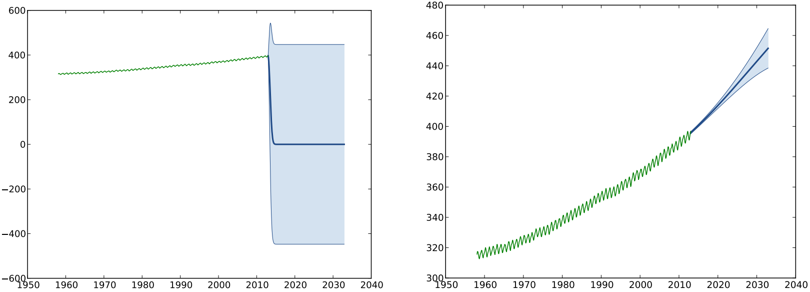


Let's try to predict the concentration for the next 20 years.

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

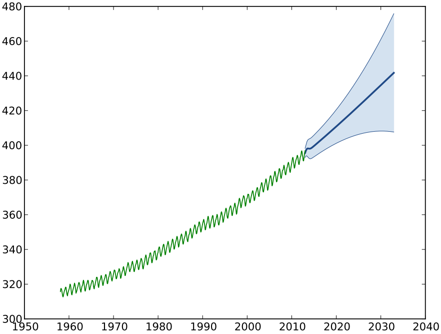
What happen if we sum both kernels ?

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

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

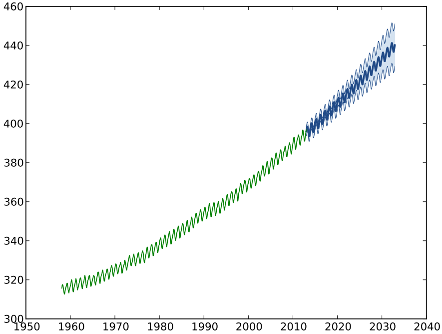
We can try the following kernel :

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

Sum of kernels over the same space

We can try the following kernel :

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



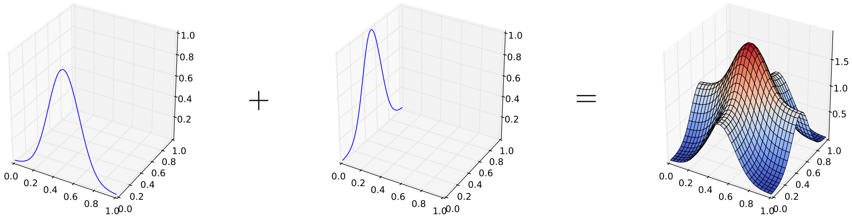
Once again, the model is significantly improved.

Sum of kernels over tensor space [Durrande 2012]

Property

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

is a valid covariance structure.

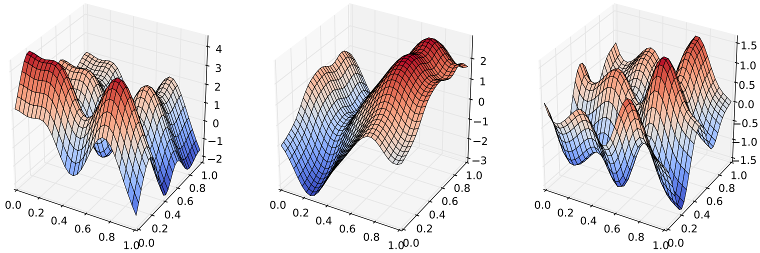


Remark :

- 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 :



$\Rightarrow$ They are additive (up to a modification)

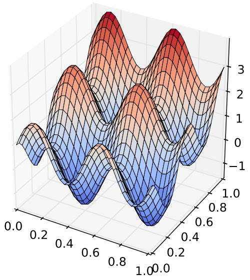
Tensor Additive kernels are very useful for

- Approximating additive functions
- Building models over high dimensional inputs spaces

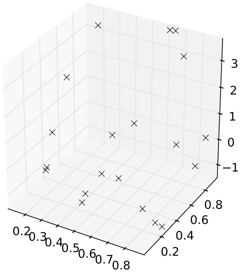
Sum of kernels over tensor space

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

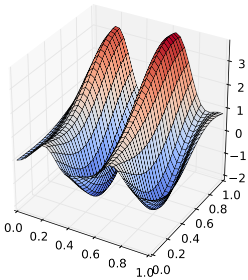


Sum of kernels over tensor space

We obtain the following models :

Gaussian kernel

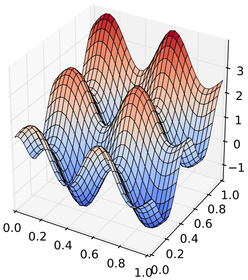
Mean predictor



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 :

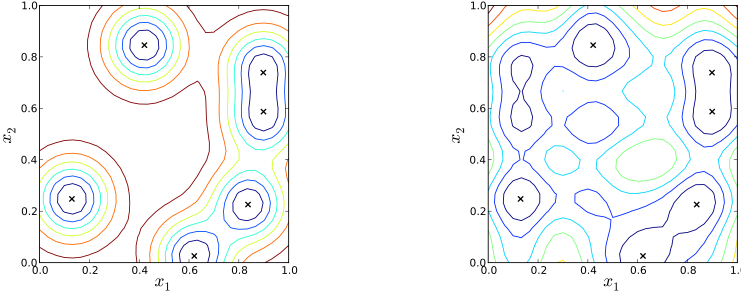
$$m_1(x_1) = \mathrm{E}(\ Z_1(x_1) \mid Z_1(X_1) + Z_2(X_2){=}F\ )$$

Sum of kernels over tensor space

Remark

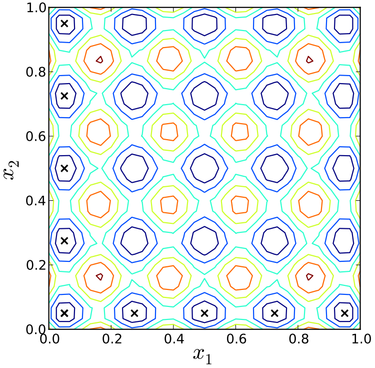
- The prediction variance has interesting features

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

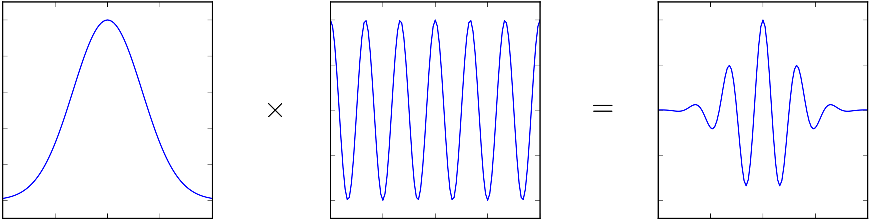
Property

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

is valid covariance structure.

Example

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



Product over the tensor space

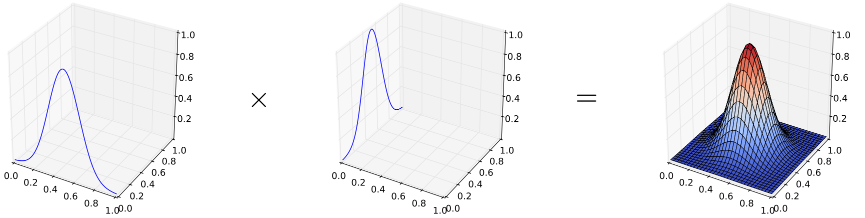
Property

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

is valid covariance structure.

Example

We multiply 2 squared exponential kernel



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

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

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

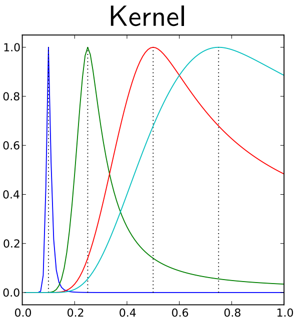
Remarks :

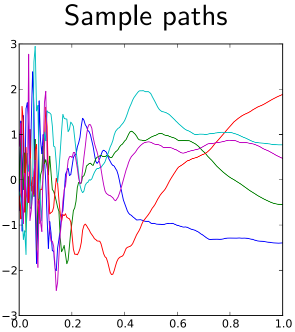
- k corresponds to the covariance of $Z(x) = Z_1(f(x))$
- This can be seen as a (nonlinear) rescaling of the input space

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 :



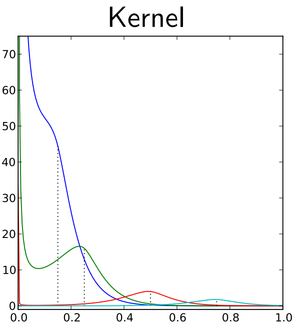


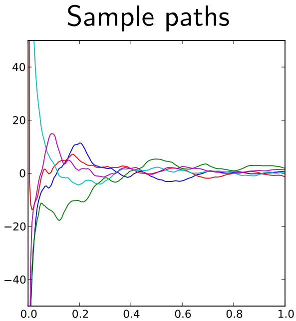
All these transformations can be combined !

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

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 :

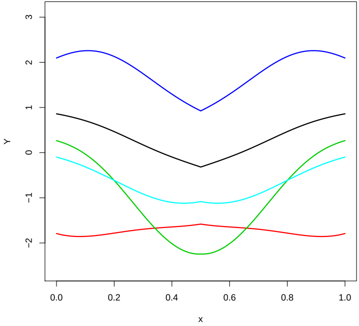
$$L_1 : f(x) \to \begin{cases} f(x) & x < 0.5 \\ f(1-x) & x \geq 0.5 \end{cases}$$

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

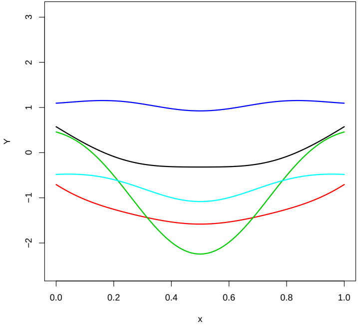
Effect of a linear operator : example [Ginsbourger 2013]

Examples of associated sample paths are

$k_1 = L_1(L_1(k))$



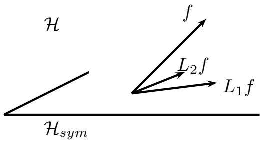
$k_2 = L_2(L_2(k))$



The differentiability is not always respected !

Effect of a linear operator

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



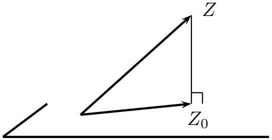
What about the optimal projection ?

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

samples with zero integrals

We are interested in building a GP such that the integral of the samples are exactly zero.

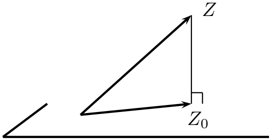
idea : project a GP onto a space of functions with zero integrals :



samples with zero integrals

We are interested in building a GP such that the integral of the samples are exactly zero.

idea : project a GP onto a space of functions with zero integrals :



It can be proved [Durrande 2013] that the orthogonal projection is given by

$$Z_0(x) = Z(x) - \frac{\int k(x,s)\mathrm{d}s \int Z(s)\mathrm{d}s}{\iint k(s,t)\mathrm{d}s\mathrm{d}t}$$

The associated kernel is :

$$k_0(x, y) = k(x, y) - \frac{\int k(x, s)\mathrm{d}s \int k(y, s)\mathrm{d}s}{\iint k(s, t)\mathrm{d}s\mathrm{d}t}$$

Such 1-dimensional kernels are of great importance for sensitivity analysis when combined with ANOVA kernels :

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= \prod_{i=1}^{d}(1 + k_0(x_i, y_i)) \\ &= 1 + \underbrace{\sum_{i=1}^{d} k(x_i, y_i)}_{\textit{additive part}} + \underbrace{\sum_{i<j} k(x_i, y_i)k(x_j, y_j)}_{2^{nd} \textit{ order interactions}} + \dots + \underbrace{\prod_{i=1}^{d} k(x_i, y_i)}_{\textit{full interaction}} \end{aligned}$$

2d example

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= \prod_{i=1}^{2} (1 + k_0(x_i, y_i)) \\ &= 1 + k_0(x_1, y_1) + k_0(x_2, y_2) + k_0(x_1, y_1) k_0(x_2, y_2) \end{aligned}$$

The best predictor can be written as

$$\begin{aligned} m(\mathbf{x}) &= (1 + k_0(x_1, X_1) + k_0(x_2, X_2) + k_0(x_1, X_1) k_0(x_2, X_2))^t k(X, X)^{-1} F \\ &= m_0 + m_1(x_1) + m_2(x_2) + m_{12}(\mathbf{x}) \end{aligned}$$

These terms correspond to the FANOVA representation of m.

10d example

Let us consider the random test function $f : [0,\ 1]^{10} \rightarrow \mathbb{R}$:

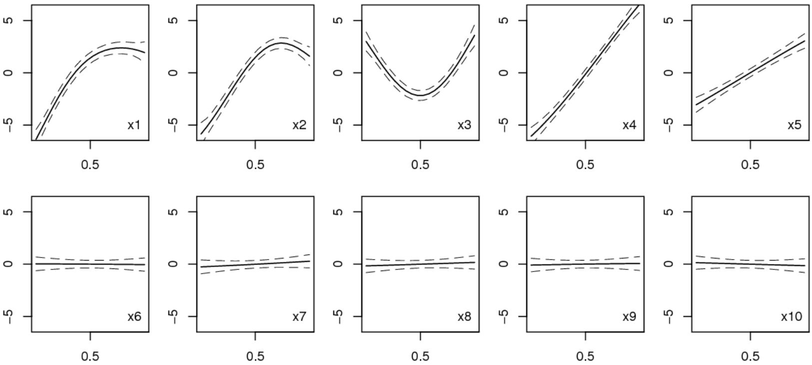
$$x \mapsto 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \mathcal{N}(0,1)$$

The steps for approximating f with GPR are :

1. Learn f on a DoE (here LHS maximin with 180 points)
2. get the optimal values for the kernel parameters using MLE,
3. build a model based on kernel $\prod(1 + k_0)$

As m is a function of 10 variables, the model can not easily be represented : it is usually considered as a "blackbox". However, the structure of the kernel allows to split m in sub-models.

The univariate sub-models are :

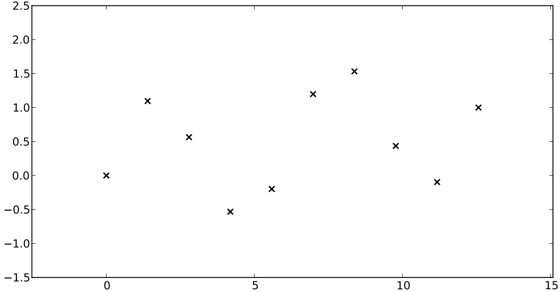


$$\left(\text{we had } f(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \mathcal{N}(0,1) \right)$$

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 [Durrande 2013].

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

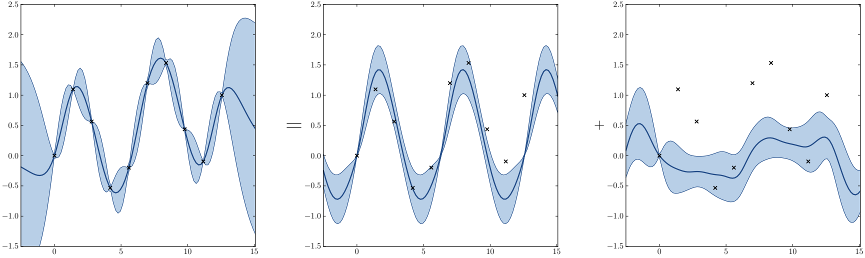
$$\begin{aligned} m(t) &= (k_p(x,X) + k_a(x,X))k(X,X)^{-1}F \\ &= \underbrace{k_p(x,X)k(X,X)^{-1}F}_{\text{periodic sub-model } m_p} + \underbrace{k_a(x,X)k(X,X)^{-1}F}_{\text{aperiodic sub-model } m_a} \end{aligned}$$

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.

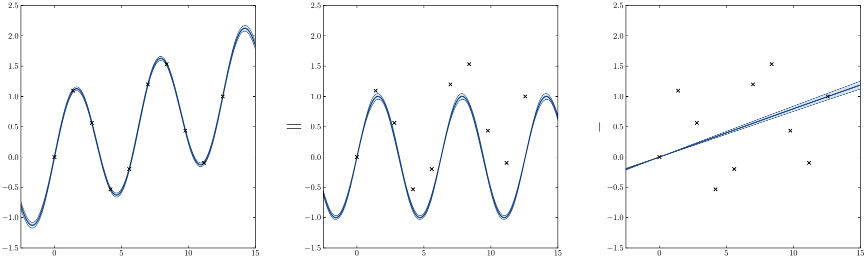
Let k^* be defined as

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

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

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

MLE of the 5 parameters of k^* gives :



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

The 24 hour cycle of days can be observed in the oscillations of many physiological processes of living beings.

Examples

Body temperature, jet lag, sleep, ... but also observed for plants, micro-organisms, etc.

This phenomenon is called the circadian rhythm and the mechanism driving this cycle is the circadian clock.

To understand how the circadian clock operates at the gene level, biologist look at the temporal evolution of gene expression.

The aim of gene expression is to measure the activity of genes :

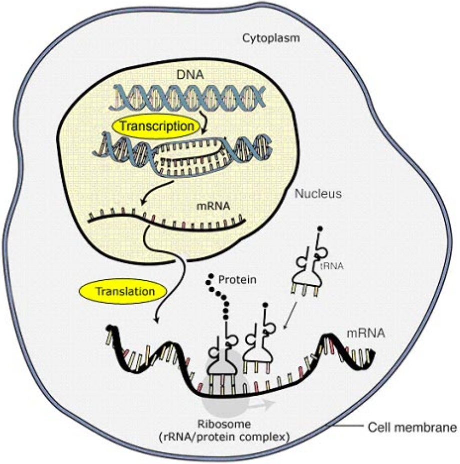
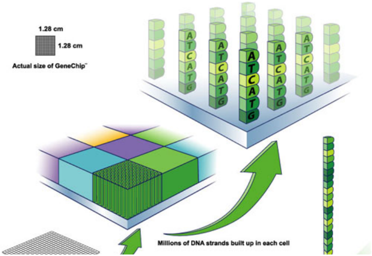


Image adapted from: National Human Genome Research Institute

The mRNA concentration is measured with microarray experiments



Experiments to study the circadian clock are typically :

1. Expose the organism to a 12h light / 12h dark cycle
2. at t=0, transfer to constant light
3. perform a microarray experiment every 4 hours to measure gene expression

Regulators of the circadian clock are often rhythmically regulated.
$\Rightarrow$ identifying periodically expressed genes gives an insight on the overall mechanism.

We used data from Edward 2006, based on *arabidopsis*.

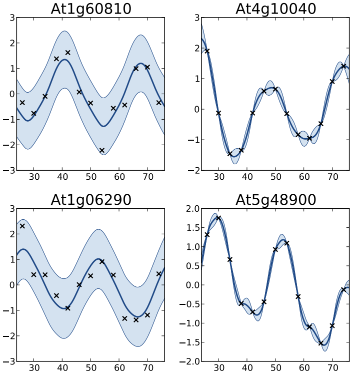
The dimension of the data is :

- 22810 genes
- 13 time points

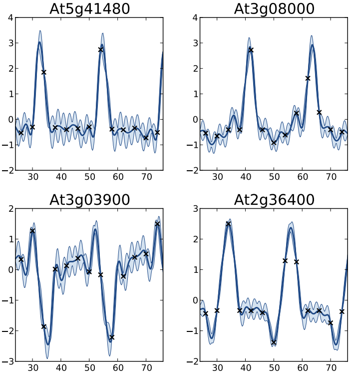
Edward 2006 gives a list of the 3504 most periodically expressed genes. The comparison with our approach gives :

- 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 the psd :

Making new from old

- sum
- product
- composition with a function

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
Gaussian process models for periodicity detection, Hal preprint.

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