

Gaussian Process Summer School

Kernel Design

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

Sheffield, September 2018



Introduction



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

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

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

Given some data, the conditional distribution is still Gaussian:

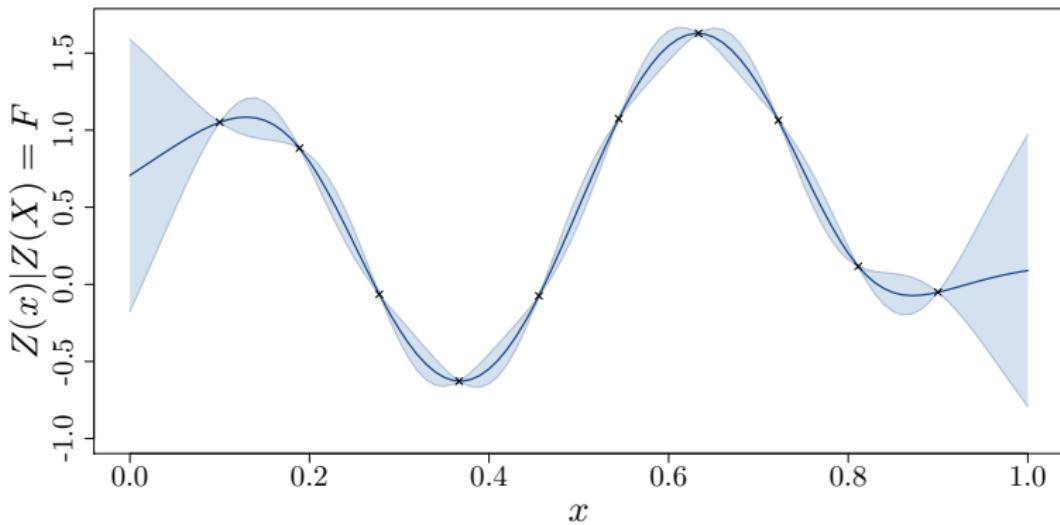
$$m(x) = E(Z(x)|Z(X) + \varepsilon=F)$$

$$= k(x, X)(k(X, X) + \tau^2 I)^{-1}F$$

$$c(x, x') = \text{cov}(Z(x), Z(x')|Z(X) + \varepsilon=F)$$

$$= k(x, x') - k(x, X)(k(X, X) + \tau^2 I)^{-1}k(X, x')$$

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



What is a kernel?



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

Example

$$k(x, x) = \text{cov}(Z(x), Z(x)) = \text{var}(Z(x)) \geq 0$$

$\Rightarrow k(x, x)$ is **positive**.

$$k(x, y) = \text{cov}(Z(x), Z(y)) = \text{cov}(Z(y), Z(x)) = k(y, x)$$

$\Rightarrow k(x, y)$ is **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:

$$\begin{aligned}\text{var}(T) &= \text{cov} \left(\sum_i a_i Z(x_i), \sum_j a_j Z(x_j) \right) = \sum_i \sum_j a_i a_j \text{cov}(Z(x_i), Z(x_j)) \\ &= \sum_i \sum_j a_i a_j k(x_i, x_j) \geq 0\end{aligned}$$

Since a variance is positive, we have for any arbitrary n , a_i and x_i .

Definition

The functions k satisfying

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

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 is a covariance \Rightarrow k is a symmetric positive semi-definite function

The reverse is also true:

Theorem (Loeve)

k corresponds to the covariance of a GP



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:

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

Matern 5/2 $k(x, y) = \sigma^2 \left(1 + \frac{\sqrt{5}|x - y|}{\theta} + \frac{5|x - y|^2}{3\theta^2}\right) \exp\left(-\frac{\sqrt{5}|x - y|}{\theta}\right)$

Matern 3/2 $k(x, y) = \sigma^2 \left(1 + \frac{\sqrt{3}|x - y|}{\theta}\right) \exp\left(-\frac{\sqrt{3}|x - y|}{\theta}\right)$

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

Brownian $k(x, y) = \sigma^2 \min(x, y)$

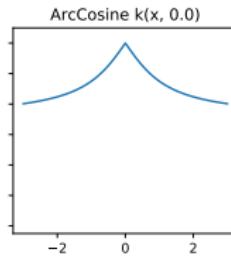
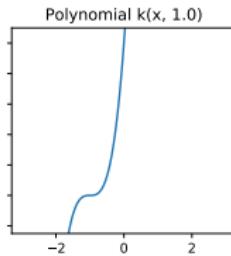
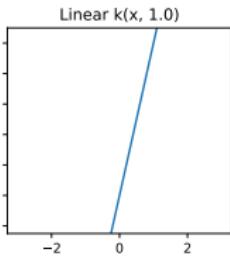
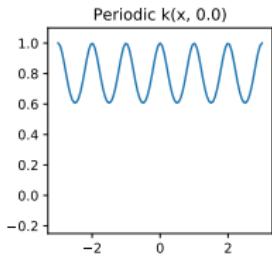
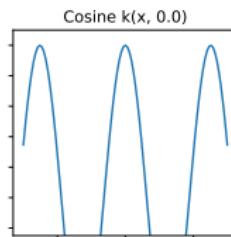
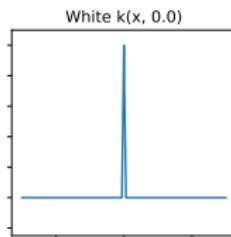
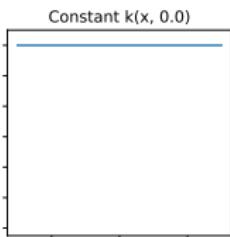
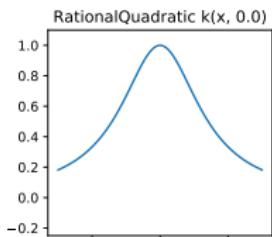
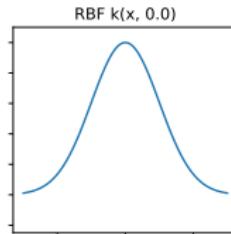
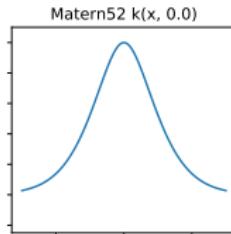
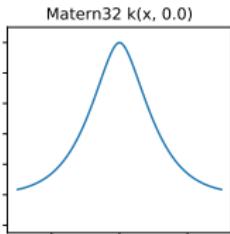
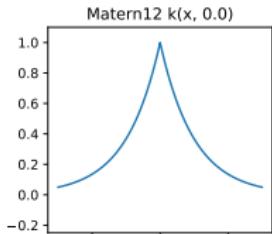
white noise $k(x, y) = \sigma^2 \delta_{x,y}$

constant $k(x, y) = \sigma^2$

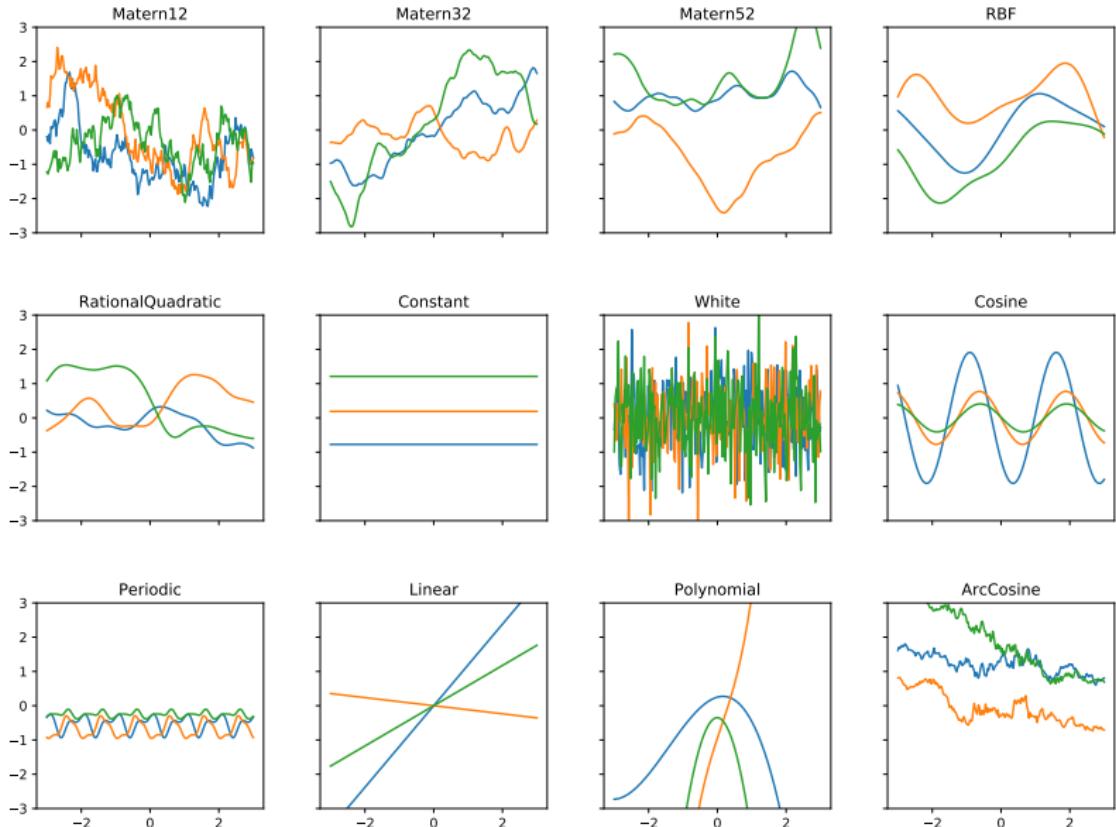
linear $k(x, y) = \sigma^2 xy$

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

Examples of kernels in gpflow:



Associated samples



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:

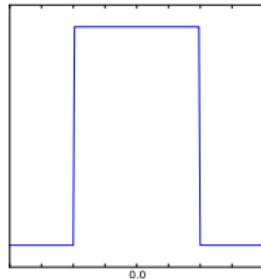
Theorem (Bochner)

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:

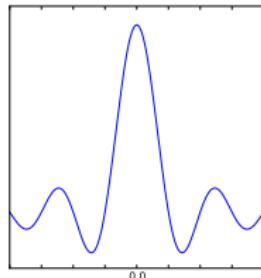
$$\tilde{k}(t) = \int_{\mathbb{R}} e^{-i\omega t} d\mu(\omega)$$

Example

We consider the following measure:



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



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

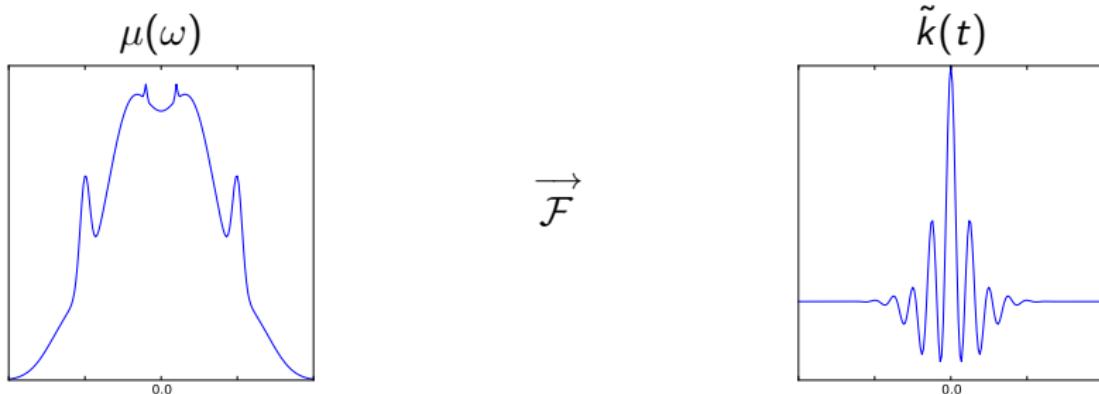
Usual kernels

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 $\frac{1}{(1+\omega^2)^\rho}$
 \Rightarrow they are psd.

Unusual kernels

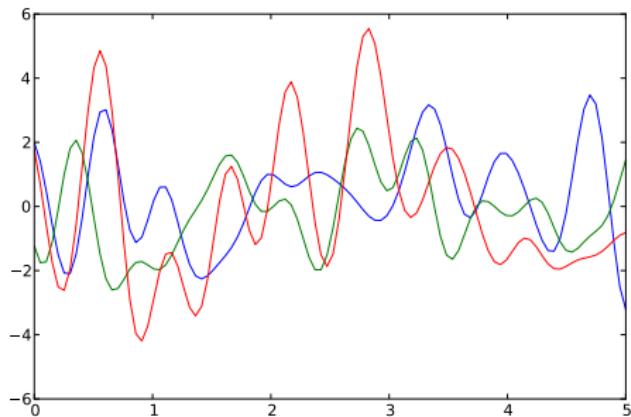
Inverse Fourier transform of a (symmetrised) sum of Gaussian gives
(A. Wilson, ICML 2013):



The obtained kernel is parametrised by its spectrum.

Unusual kernels

The sample paths have the following shape:



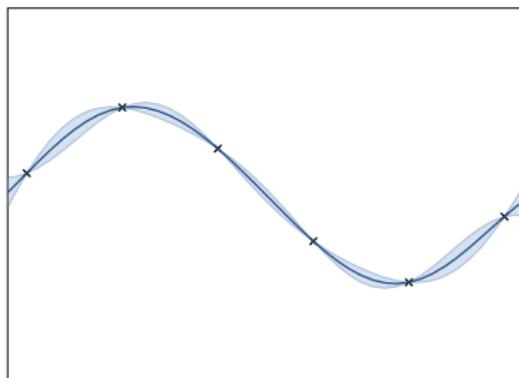


Choosing the appropriate kernel

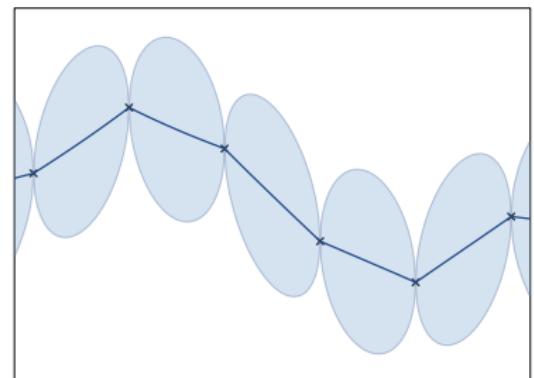


Changing the kernel has a huge impact on the model:

Gaussian kernel:

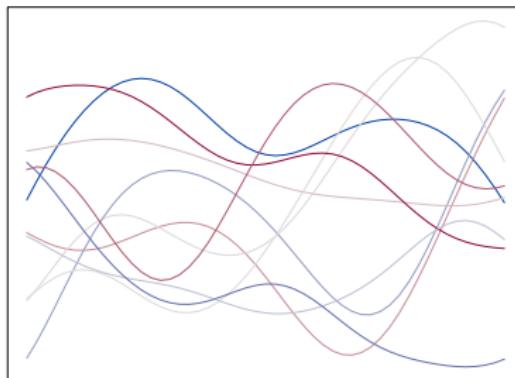


Exponential kernel:

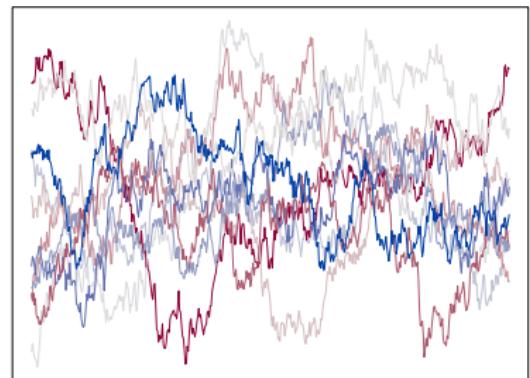


This is because changing the kernel implies changing the prior

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 assess 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 $x \rightarrow \log(x)$, $x \rightarrow \exp(x)$, ...



Making new from old



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 $k(x, y) = k_1(x_1, y_1) + k_2(x_2, y_2)$

- Multiplied together

- ▶ On the same space $k(x, y) = k_1(x, y) \times k_2(x, y)$
 - ▶ On the tensor space $k(x, y) = k_1(x_1, y_1) \times k_2(x_2, y_2)$

- 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 input space

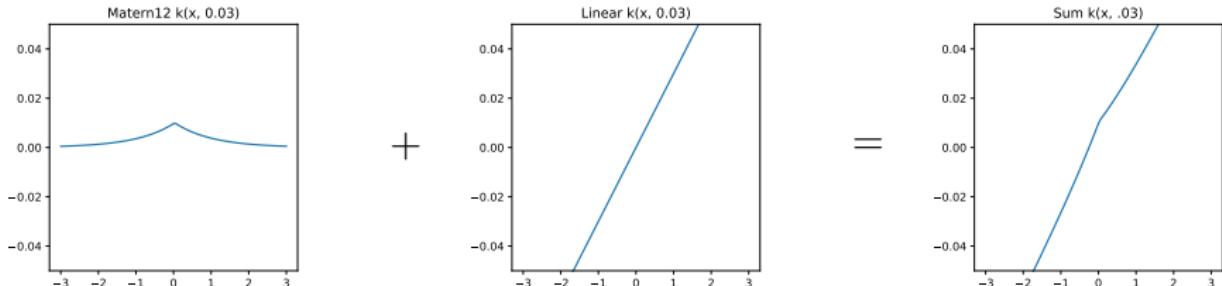
Property

$$k(x, y) = k_1(x, y) + k_2(x, y)$$

is a valid covariance structure.

This can be proved directly from the p.s.d. definition.

Example

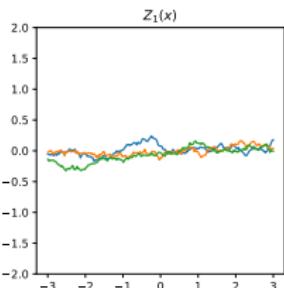


Sum of kernels over the same input space

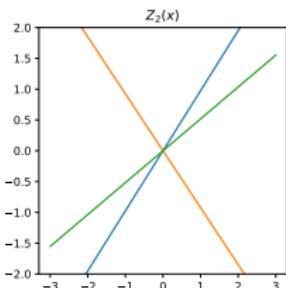
$Z \sim \mathcal{N}(0, k_1 + k_2)$ can be seen as $Z = Z_1 + Z_2$ where Z_1, Z_2 are independent and $Z_1 \sim \mathcal{N}(0, k_1)$, $Z_2 \sim \mathcal{N}(0, k_2)$

$$k(x, y) = k_1(x, y) + k_2(x, y)$$

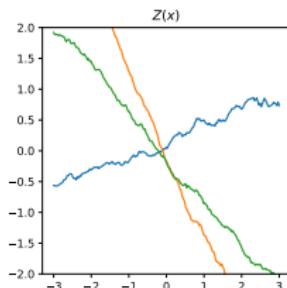
Example



+



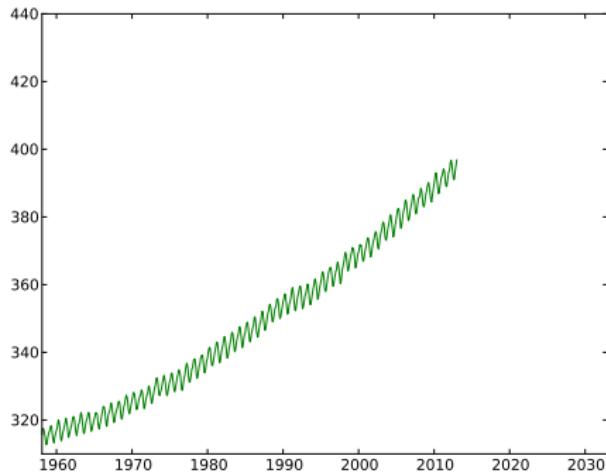
=



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.

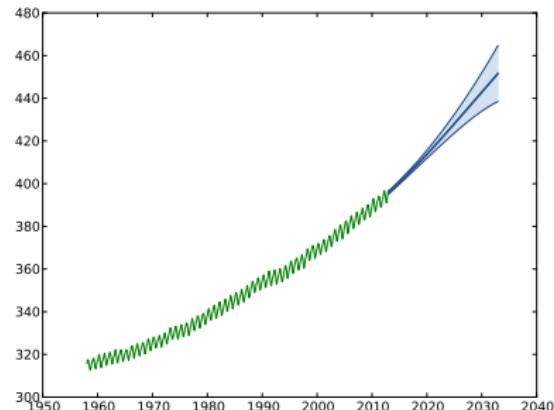
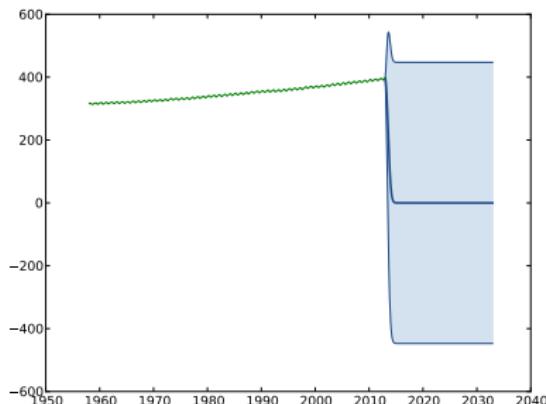


Let's try to predict the concentration for the next 20 years.

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

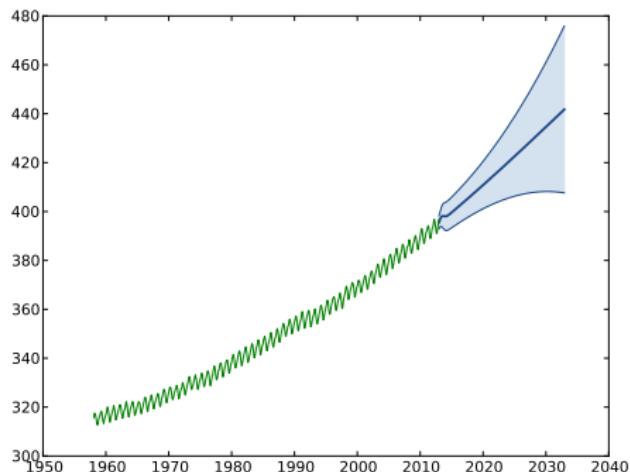
What happens if we sum both kernels?

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

Sum of kernels over the same space

What happens 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

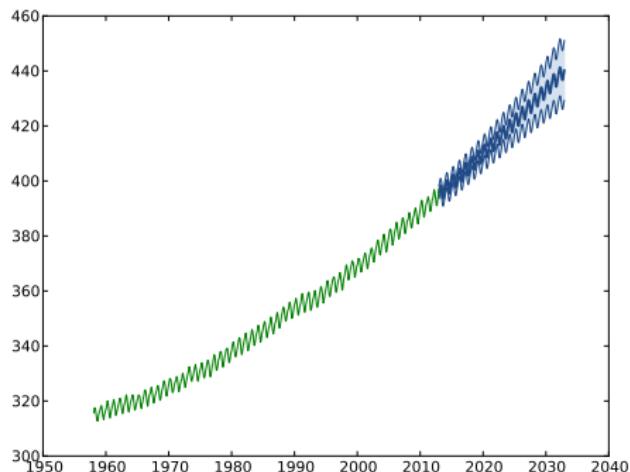
We can try the following kernel:

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

Sum of kernels over the same space

We can try the following kernel:

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



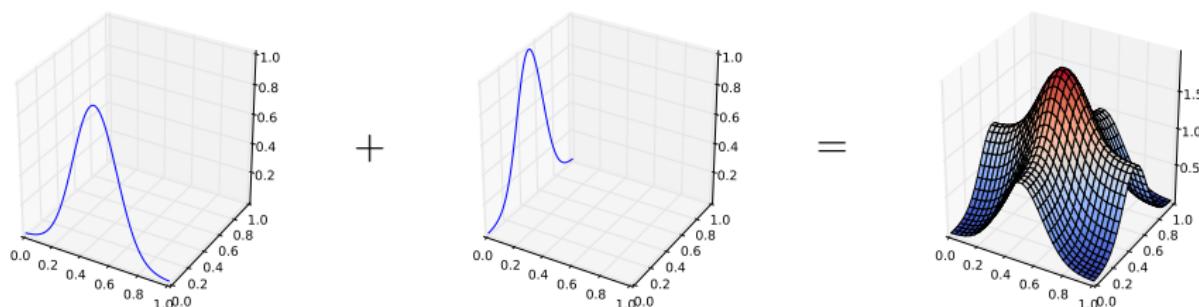
Once again, the model is significantly improved.

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.

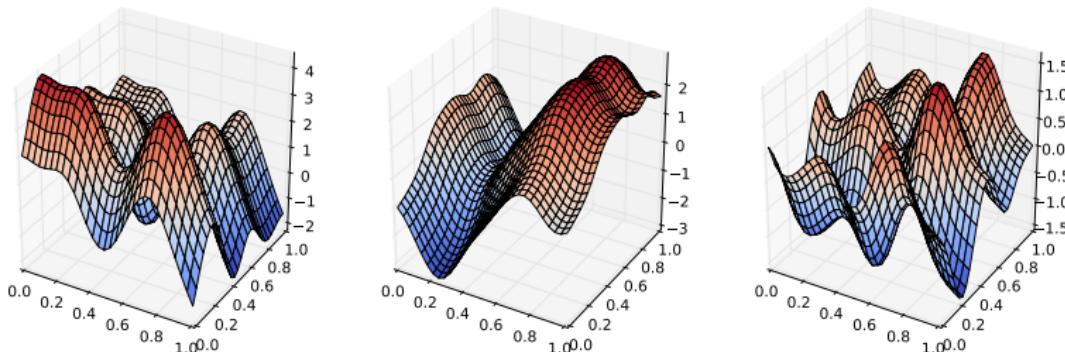


Remark:

- 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

Remarks

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

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

⇒ The model shares the prior behaviour.

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

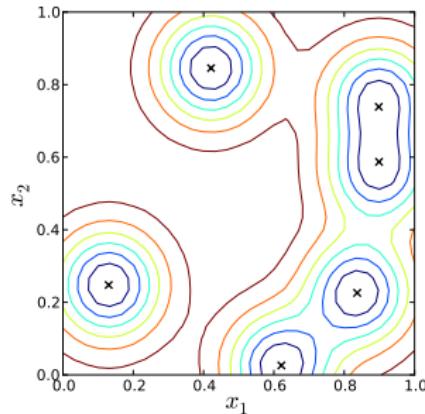
$$m_1(x_1) = E(Z_1(x_1) | Z_1(X_1) + Z_2(X_2) = \mathbf{F})$$

Sum of kernels over tensor space

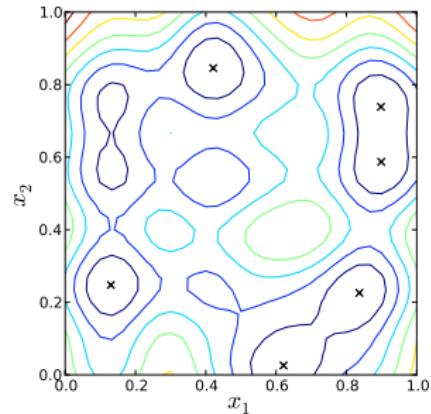
Remark

- The prediction variance has interesting features

pred. var. with kernel product

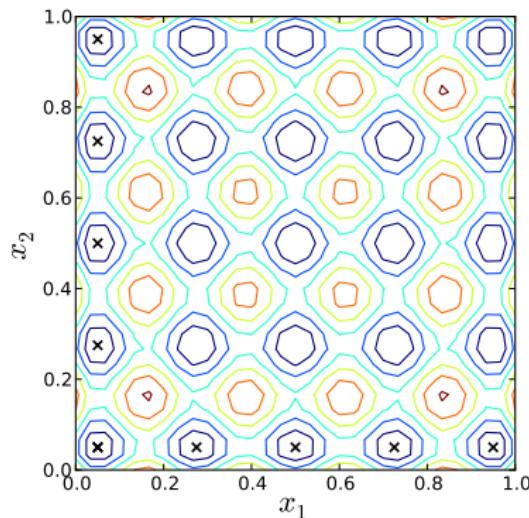


pred. var. with kernel sum



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

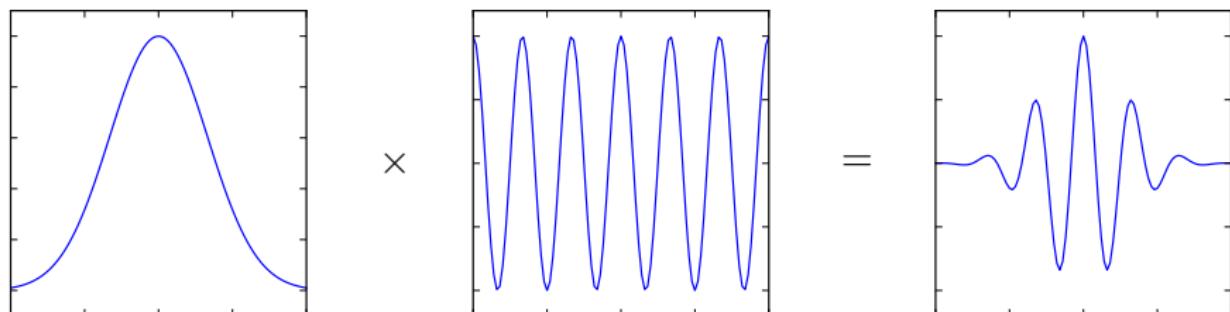
Property

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

is valid covariance structure.

Example

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



Product over the tensor space

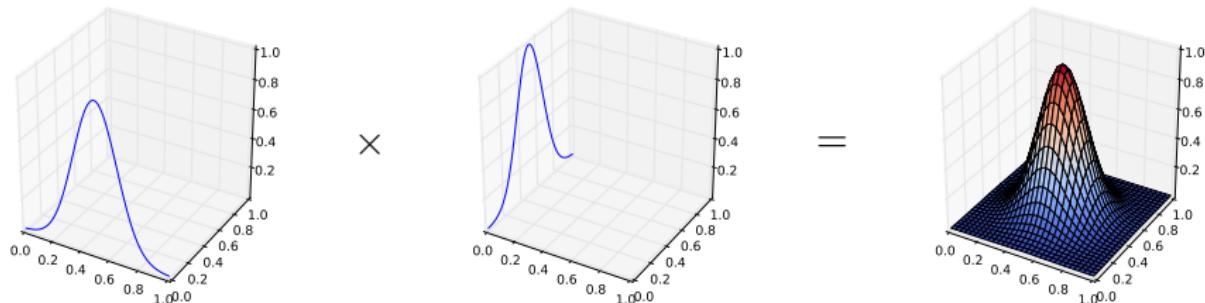
Property

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

is valid covariance structure.

Example

We multiply two squared exponential kernels



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 $D \rightarrow D_1$, then

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

is a kernel over $D \times D$.

proof

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

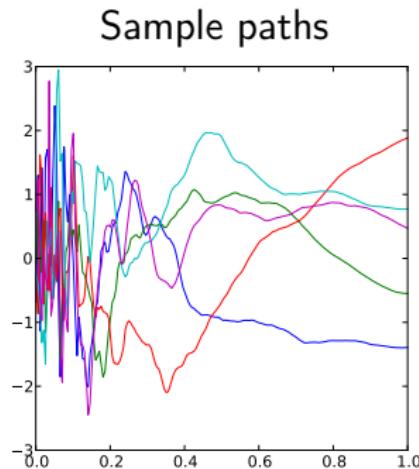
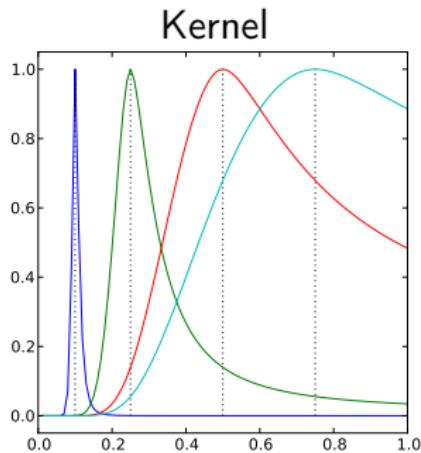
Remarks:

- 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:

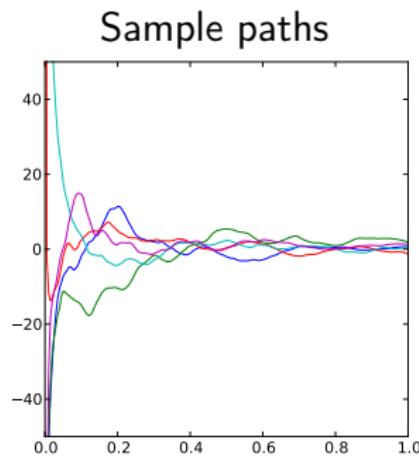
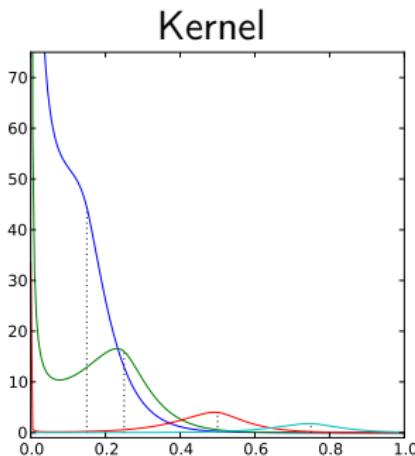


All these transformations can be combined!

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 linear operators



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

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

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

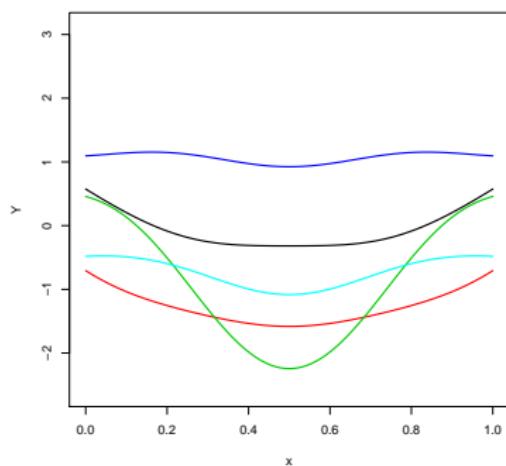
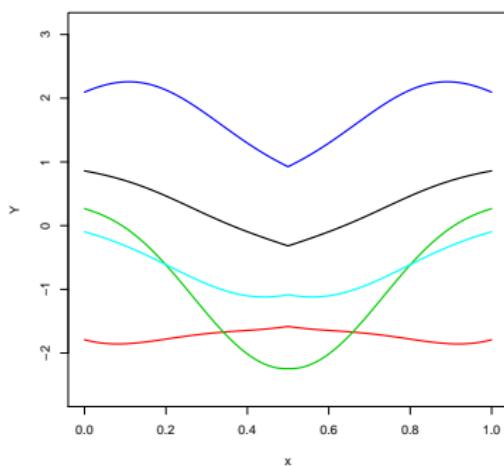
Effect of a linear operator

Example

Associated sample paths are

$$k_1 = L_1(L_1(k))$$

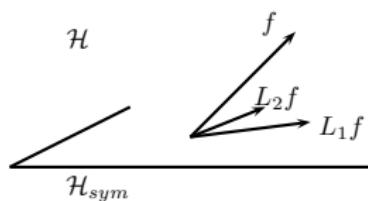
$$k_2 = L_2(L_2(k))$$



The differentiability is not always respected!

Effect of a linear operator

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



What about the optimal projection?

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



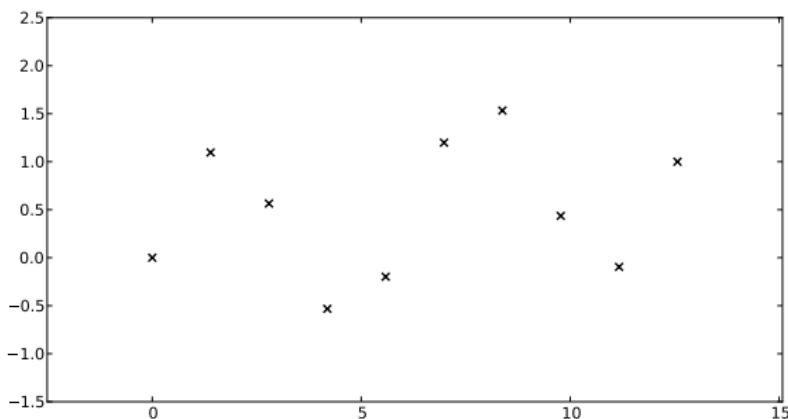
Application: Periodicity detection



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

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

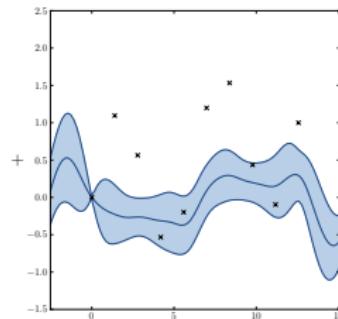
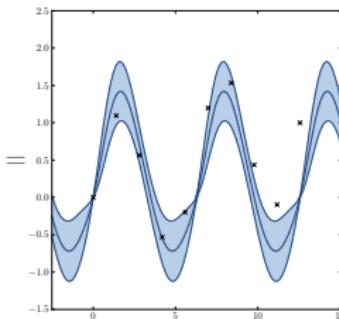
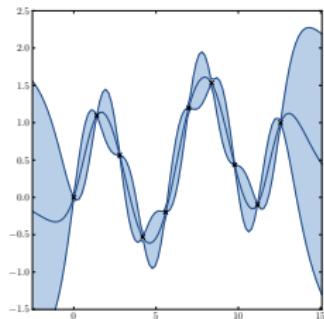
and we can associate a prediction variance to the sub-models:

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

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

Example

For the observations shown previously we obtain:



Can we 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.

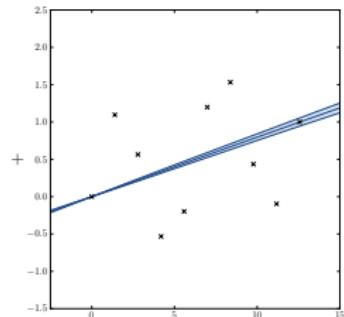
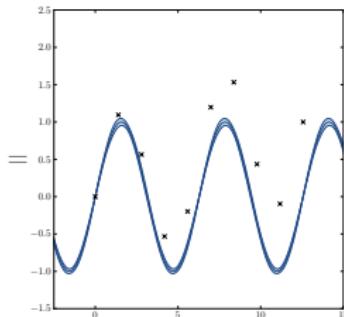
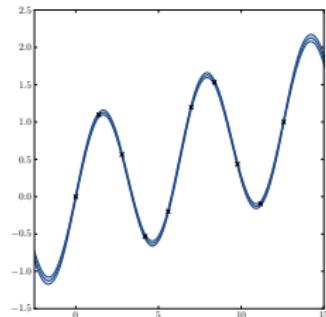
Let k^* be defined as

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

Furthermore, we include a 5th 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.

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



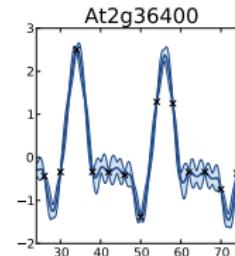
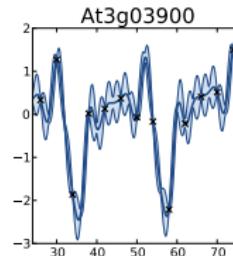
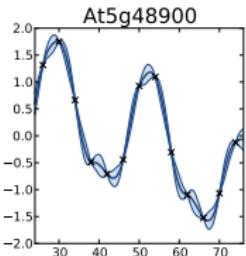
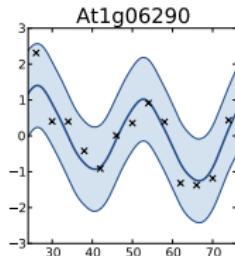
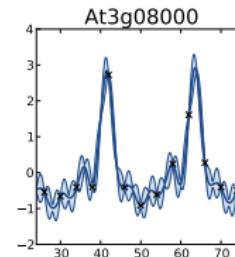
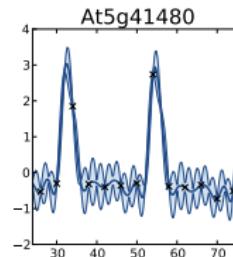
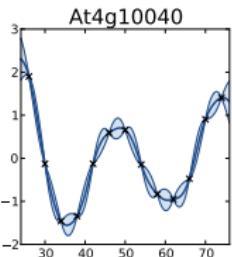
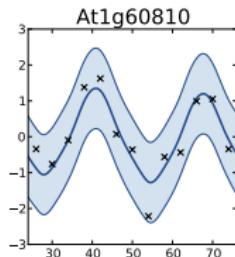
The dimension of the data is:

- 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

Conclusion

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, 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.