Gaussian process models and kernel design

Applications to circadian rythm studies

Nicolas Durrande (Mines St-Étienne – durrande@emse.fr)

INRA Toulouse – MIAT Seminar, the 05th of June 2015

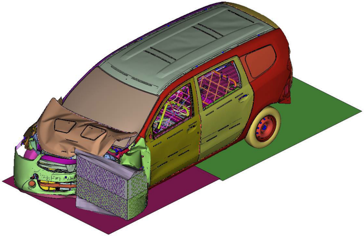
Introduction

Context

We consider a function f whose evaluation of f is costly :

examples

- physical experiment
- output of a large computer code
- ...



Using a limited number of observations we want to answer questions such as

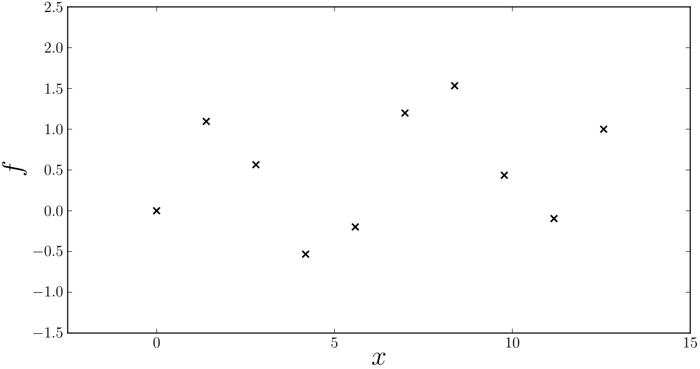
- what is the minimum of f ?
- what is the mean value ?
- what is the probablility to be above a given threshold ?
- Are there some non influent variables ?
- ...

Outline :

1. Introduction to GP models
2. A few words on RKHS
3. What is a kernel ?
4. Designing kernels
5. Application : eriodicity detection

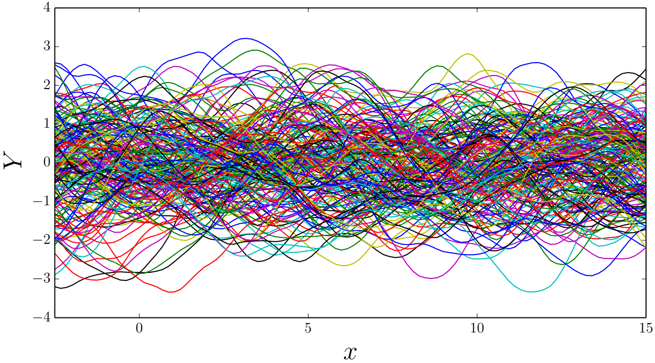
Gaussian process regression

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



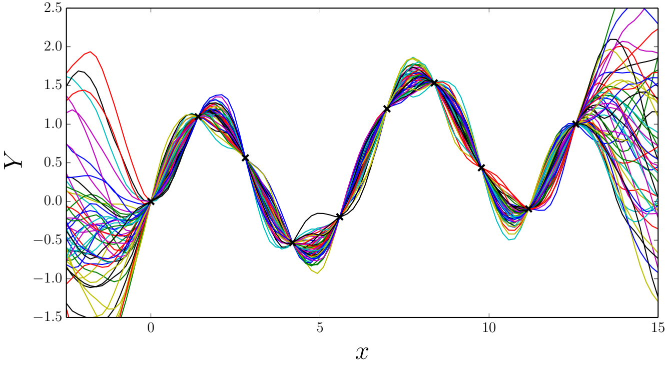
The observations are denoted by $f_i = f(x_i)$ (or $F = f(X)$).

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



Y is characterised by its mean and covariance function.

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

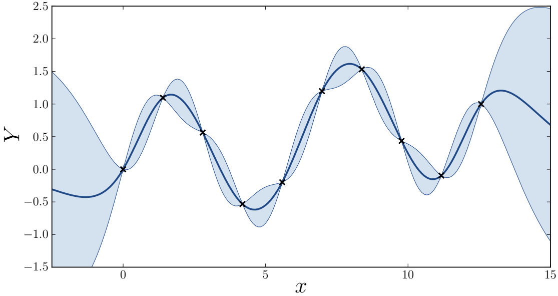


The conditional distribution is still Gaussian. It has mean and variance

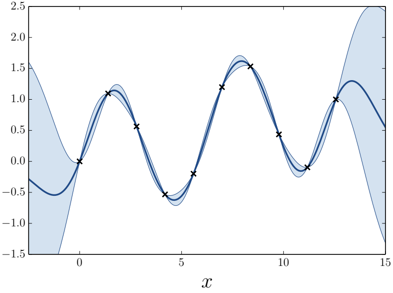
$$m(x) = \mathrm{E}\left(Y(x)|Y(X) = F\right) = k(x, X)k(X, X)^{-1}F$$
$$v(x) = \mathrm{var}\left(Y(x)|Y(X) = F\right) = k(x, x) - k(x, X)^t k(X, X)^{-1}k(x, X)$$

where k is the kernel : $k(x, y) = cov(Y(x), Y(y))$.

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

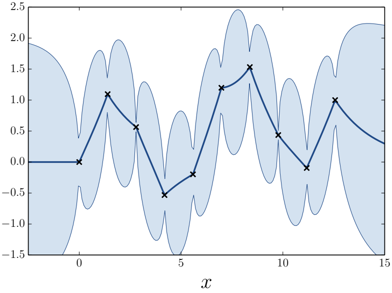


Changing the kernel has a huge impact on the model :



gaussian kernel

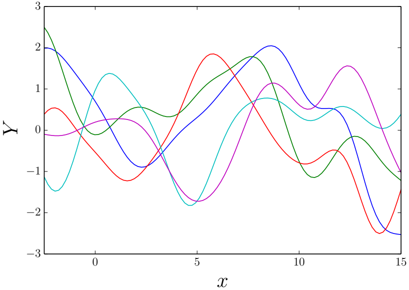
$$k(x,y) = \sigma^2 \exp\left(\frac{-(x-y)^2}{2\theta^2}\right)$$

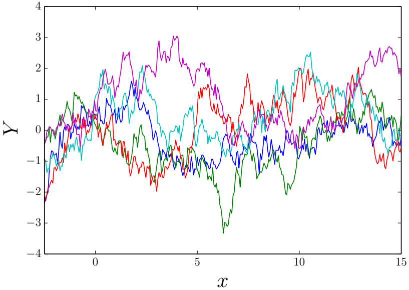


exponential kernel

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

This is because it means changing the prior on f :





gaussian kernel

exponential kernel

Given the observations, the model is entirely defined by the kernel. We will now focus on this object.

Theorem (Loeve)

$$\text{\textit{k corresponds to the covariance of a GP}} \\ \Updownarrow \\ \text{\textit{k is a symmetric positive semi-definite function}}$$

Definition

A function k is positive semi-definite if it satisfies

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

for all $n \in \mathbb{N}$, **for all** $x_i \in D$, **for all** $a_i \in \mathbb{R}$.

A symmetric positive semi-definite function is also the reproducing kernel of a RKHS :

Definition

$\mathcal{H}$ is a RKHS with kernel k if it is a Hilbert space such that :

- for all x, $k(x,.) \in \mathcal{H}$
- for all $f \in \mathcal{H}$, $\langle f(.), k(x,.) \rangle_{\mathcal{H}} = f(x)$

Given a kernel k, the associated RKHS 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 :

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

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

$$\begin{tikzcd}
 & Z \arrow[dr] & \\
k \arrow[ur, leftrightarrow] \arrow[dr, leftrightarrow] & & m \\
 & \mathcal{H} \arrow[ur] &
\end{tikzcd}$$

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

white noise : $k(x,y) = \delta_{x,y}$

bias : $k(x,y) = 1$

linear : $k(x,y) = xy$

exponential : $k(x,y) = \exp\left(-|x-y|\right)$

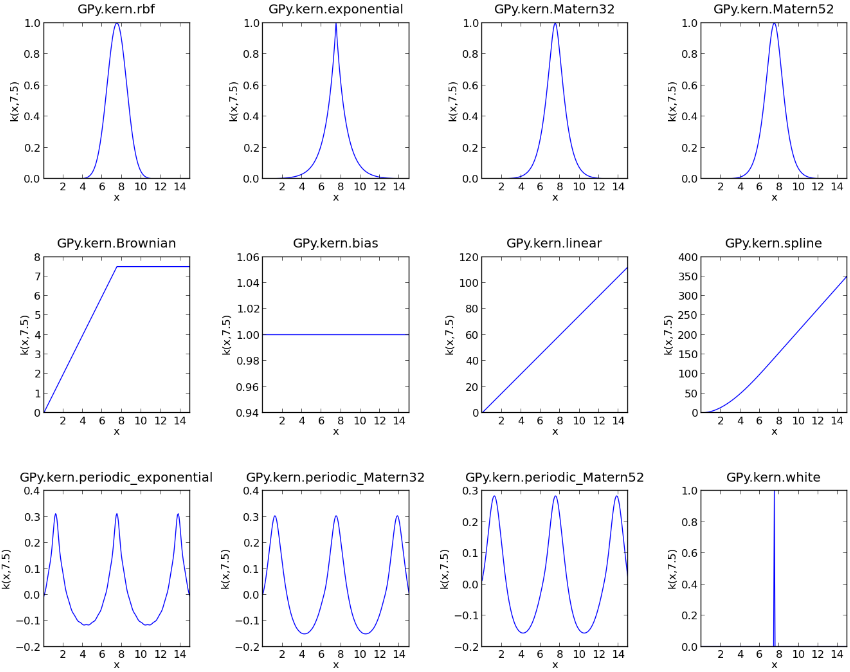
Brownian : $k(x,y) = \min(x,y)$

Gaussian : $k(x,y) = \exp\left(-(x-y)^2\right)$

Matérn 3/2 : $k(x,y) = (1+|x-y|) \times \exp\left(-|x-y|\right)$

sinc : $k(x,y) = \frac{\sin(|x-y|)}{|x-y|}$

$\vdots$


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
0.0 0.2 0.4 0.6 0.8 1.0
0 1 2 3 4 5 6 7 8
0.94 0.96 0.98 1.00 1.02 1.04 1.06
0 20 40 60 80 100 120
0 50 100 150 200 250 300 350 400
−0.2 −0.1 0.0 0.1 0.2 0.3 0.4
−0.2 −0.1 0.0 0.1 0.2 0.3
2 4 6 8 10 12 14

The kernel has to be chosen accordingly to the prior believe on the function to approximate :

- What is the regularity of the phenomenon ?
- Is it stationary ?
- ...

If we have some knowledge about the behaviour of f, can we design a kernel accordingly ?

We will discuss 3 options :

- Making new from old
- Linear operator
- extracting RKHS subspaces

Designing kernels

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

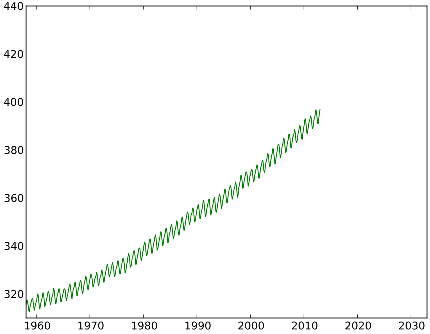
All these operations will preserve the positive definiteness.

How can this be useful ?

Sum of kernels over the same space

Example (The Mauna Loa observatory dataset)

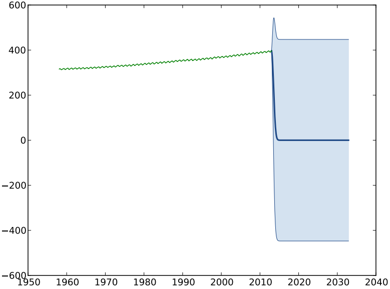
This famous dataset compiles the monthly CO_2 concentration in Hawaii since 1958.

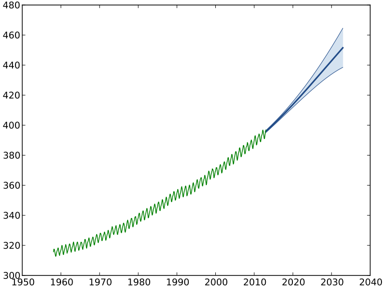


Let's try to predict the concentration for the next 20 years.

Sum of kernels over the same space

We first consider a squared-exponential kernel :





The results are terrible !

Sum of kernels over the same space

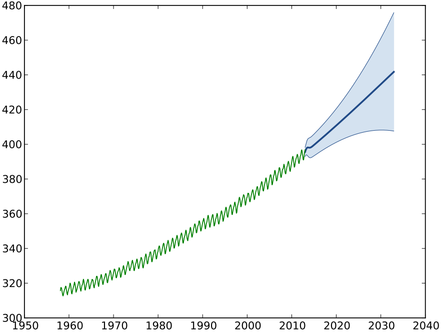
What happen if we sum both kernels ?

$$k(x,y) = \sigma_1^2 k_{rbf1}(x,y) + \sigma_2^2 k_{rbf2}(x,y)$$

Sum of kernels over the same space

What happen if we sum both kernels ?

$$k(x, y) = \sigma_1^2 k_{rbf1}(x, y) + \sigma_2^2 k_{rbf2}(x, y)$$



The model is drastically improved !

Sum of kernels over the same space

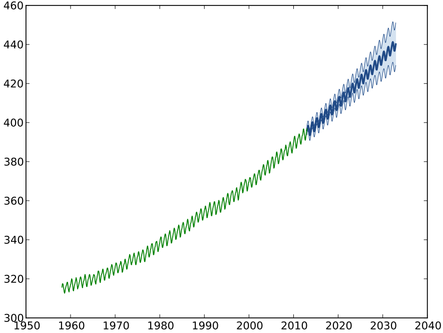
We can try the following kernel :

$$k(x,y) = \sigma_0^2 x^2 y^2 + \sigma_1^2 k_{rbf1}(x,y) + \sigma_2^2 k_{rbf2}(x,y) + \sigma_3^2 k_{per}(x,y)$$

Sum of kernels over the same space

We can try the following kernel :

$$k(x,y) = \sigma_0^2 x^2 y^2 + \sigma_1^2 k_{rbf1}(x,y) + \sigma_2^2 k_{rbf2}(x,y) + \sigma_3^2 k_{per}(x,y)$$



Once again, the model is significantly improved.

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

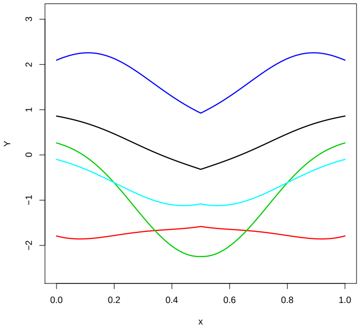
$$L_1 : f(x) \rightarrow \begin{cases} f(x) & x < 0.5 \\ f(1-x) & x \geq 0.5 \end{cases}$$

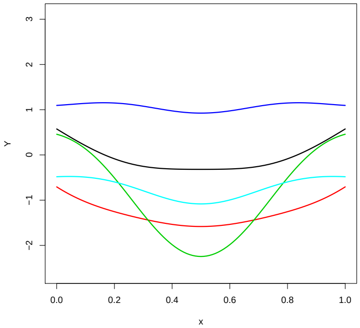
$$L_2 : f(x) \rightarrow \frac{f(x) + f(1-x)}{2}.$$

Effect of a linear operator : example (Ginsbourger, AFST 2013)

Examples of associated sample paths are

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

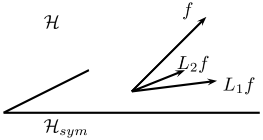




The differentiability is not always respected !

Effect of a linear operator

Ideally, we want to extract the subspace of symmetric functions in $\mathcal{H}$



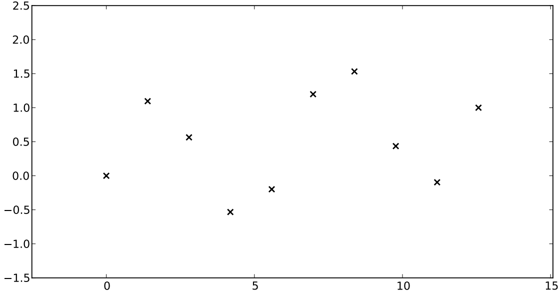
and to define L as the orthogonal projection onto $\mathcal{H}_{sym}$

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

Kernels for periodicity detection

General problem

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



As previously we will build an orthogonal decomposition of the RKHS :

$$\mathcal{H} = \mathcal{H}_p + \mathcal{H}_a$$

where $\mathcal{H}_p$ is the subspace of $\mathcal{H}$ spanned by the Fourier basis $B(t) = (\sin(t), \cos(t), \dots, \sin(nt), cos(nt))^t$.

Property

The reproducing kernel of $\mathcal{H}_p$ is

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

where G is the Gram matrix G associated to B.

We can deduce the following decomposition of the kernel :

$$k(x,y) = k_p(x,y) + \underbrace{k(x,y) - k_p(x,y)}_{k_a(x,y)}$$

Property : Decomposition of the model

The decomposition of the kernel gives directly

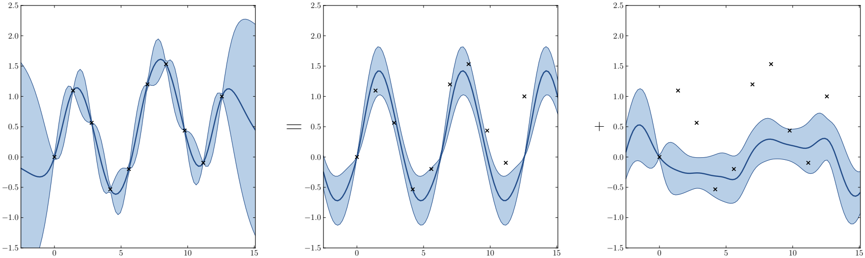
$$\begin{aligned} m(t) &= (k_p(t) + k_a(t))^t (K_p + K_a)^{-1} F \\ &= \underbrace{k_p(t)^t (K_p + K_a)^{-1} F}_{\text{periodic sub-model } m_p} + \underbrace{k_a(t)^t (K_p + K_a)^{-1} F}_{\text{aperiodic sub-model } m_a} \end{aligned}$$

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

$$\begin{aligned} v_p(t) &= k_p(t,t) - k_p(t)^t (K_p + K_a)^{-1} k_p(t) \\ v_a(t) &= k_a(t,t) - k_a(t)^t (K_p + K_a)^{-1} k_a(t) \end{aligned}$$

Example

For the observations shown previously we obtain :



Can we can do better ?

Previously, the kernels were parameterized 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.

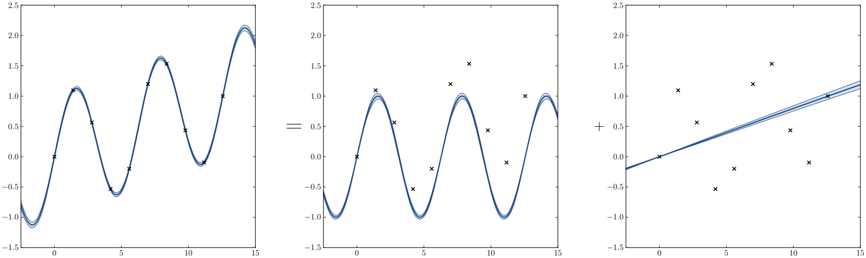
Let k^* be defined as

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

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

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

If we optimize the 5 parameters of k^* with maximum likelihood estimation we obtain :



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

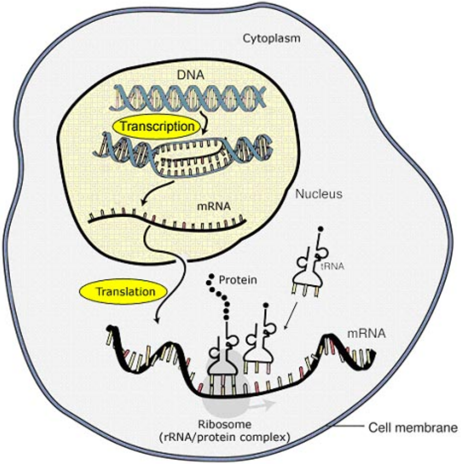
Examples

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

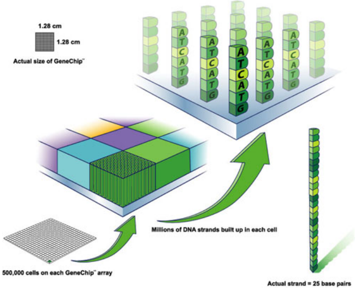
This phenomenon is called the circadian rhythm and the mechanism driving this cycle is the circadian clock.

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 various genes :



The mRNA concentration is measured with microarray experiments


1.28 cm
1.28 cm
Actual size of GeneChip™
Millions of DNA strands built up in each cell
500,000 cells on each GeneChip™ array
Actual strand = 25 base pairs

The chip is then scanned to determine the occupation of each cell and reveal the concentration of mRNA.

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.

⇒ identifying periodically expressed genes gives an insight on the overall mechanism.

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

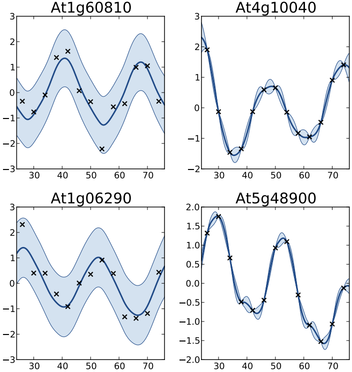
The dimension of the data is :

- 22810 genes
- 13 time points

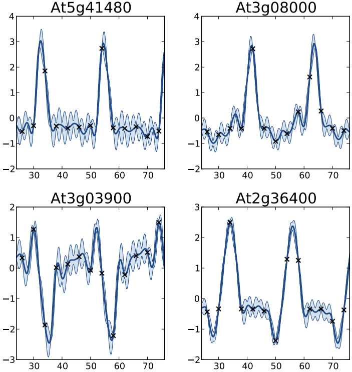
Edward 2006 gives a list of the 3504 most periodically expressed genes. The comparison with our approach gives :

- 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

Conclusion

Conclusion

We have seen that

- Gaussian process regression is a great tool for modeling
- Kernels can (and should) be tailored to the problem at hand

What cannot be done with GPs...

It is rather difficult to :

- impose non linear constrains
- deal with a (very) large number of observations
- ...

Conclusion

Bibliography

N. Aronszajn.
Theory of reproducing kernels.
Transaction of the AMS, 68,1950

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

Sensitivity analysis

The analysis of the influence of the various variables of a d-dimensional function f is often based on the HDMR :

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{i<j} f_{i,j}(x_i, x_j) + \dots + f_{1,\dots,d}(\mathbf{x})$$

where $\int f(x_I)\mathrm{d}x_i = 0$ if $i \in I$.

Can we obtain a similar decomposition for the model ?

A first idea is to consider ANOVA kernels [Stitson 97] :

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

A decomposition of the best predictor is naturally associated to those kernels.

Example : we have in 2D $K = 1 + K_1 + K_2 + K_1 K_2$ so the best predictor can be written as

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

This decomposition looks like the ANOVA representation of m but

the m_I **do not satisfy**

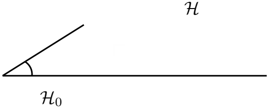
$$\int_{D_i} m_I(\mathbf{x}_I) \mathrm{d}x_i = 0$$

We need to build a kernel k_0 such that $\int k_0(x, y) \mathrm{d}x = 0$ for all y.

The RKHS framework will be very useful here

Can we extract the subspace of zero mean function in $\mathcal{H}$?

$$h \in \mathcal{H}_0 \Leftrightarrow \int h(x)\mathrm{d}x = 0$$



The integral operator is linear, and it is bounded if $\int k(x,x)\mathrm{d}x < \infty$.
$\Rightarrow$ We apply Riesz theorem. Let R be the representer.

$$h \in \mathcal{H}_0 \Leftrightarrow \int h(x)\mathrm{d}x = 0 \Leftrightarrow \langle h, R \rangle_{\mathcal{H}} = 0$$

Calculations give directly

$$R(x) = \langle R, k(x,.) \rangle_{\mathcal{H}} = \int_D k(x,s) \mathrm{d}s$$

$$L(h) = h - \frac{\langle R, k(x,.) \rangle_{\mathcal{H}}}{||R||^2_{\mathcal{H}}} R$$

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

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

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

The steps for approximating f with GPR are :

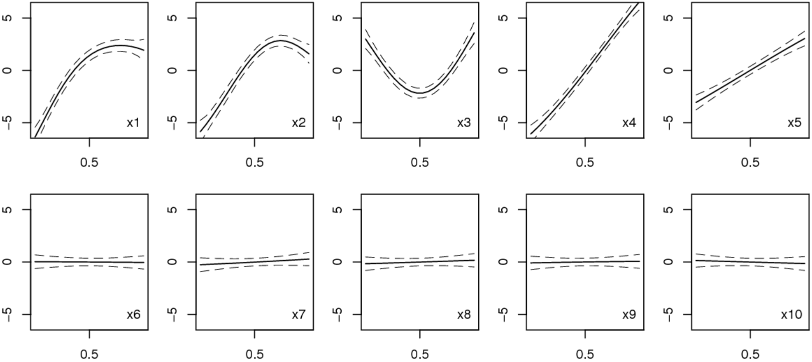
1 Learn f on a DoE (here LHS maximin with 180 points)

2 get the optimal values for the kernel parameters using MLE,

3 build the kriging predictor m based on $\prod(1 + k_0)$

As $\hat{f}$ 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 submodels.

The univariate sub-models are :



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

The sensitivity indices can be obtained analytically :

$$S_I = \frac{\mathrm{var}\left(m_I(X_I)\right)}{\mathrm{var}\left(m(X)\right)}$$
$$= \frac{F^T \mathrm{K}^{-1} \left(\bigodot_{i \in I} \Gamma_i\right) \mathrm{K}^{-1} F}{F^T \mathrm{K}^{-1} \left(\bigodot_{i=1}^{d} \left(1_{n \times n} + \Gamma_i\right) - 1_{n \times n}\right) \mathrm{K}^{-1} F}$$

where Γ_i is the matrix $\Gamma_i = \int_{D_i} k_i^0(s_i) k_i^0(s_i)^T \mathrm{d}s_i$, $1_{n \times n}$ is the $n \times n$ matrix of ones and where $\odot$ is an entrywise product.