Surrogate models and Gaussian Process regression – lecture 5/5

Advanced GP models

Mines St-Étienne – Majeure Data Science – 2016/2017

Nicolas Durrande (durrande@emse.fr)

Multioutputs GPR / GPR with categorical inputs

Example

We observe the temperature in two cities A and B for a few time points X_A and X_B. We assume a Gaussian process prior for these $T_A(t)$ and $T_B(t)$. What would be your prediction for the temperature in A at a new time point t?

Example

We observe the temperature in two cities A and B for a few time points X_A and X_B. We assume a Gaussian process prior for these $T_A(t)$ and $T_B(t)$. What would be your prediction for the temperature in A at a new time point t?

Ideally, we are interested in $T_A(t)|T_A(X_A), T_B(X_B)$. If $(T_A(t),\ T_A(X_A),\ T_B(X_B))$ is a Gaussian vector, we know how to compute the conditional distribution. However, it requires the cross covariance $k_{AB}(t, t') = \text{cov}[T_A(t), T_B(t')]$.

Exercise

Compute the distribution of $T_A(t)|\, T_A(X_A), T_B(X_B)$.

Exercise

Compute the distribution of $T_A(t)|\, T_A(X_A),\, T_B(X_B)$.

Solution

The conditional mean is:

$$m_A(t) = \mathrm{E}[T_A(t)|\, T_A(X_A){=}F_A,\, T_B(X_B){=}F_B]$$

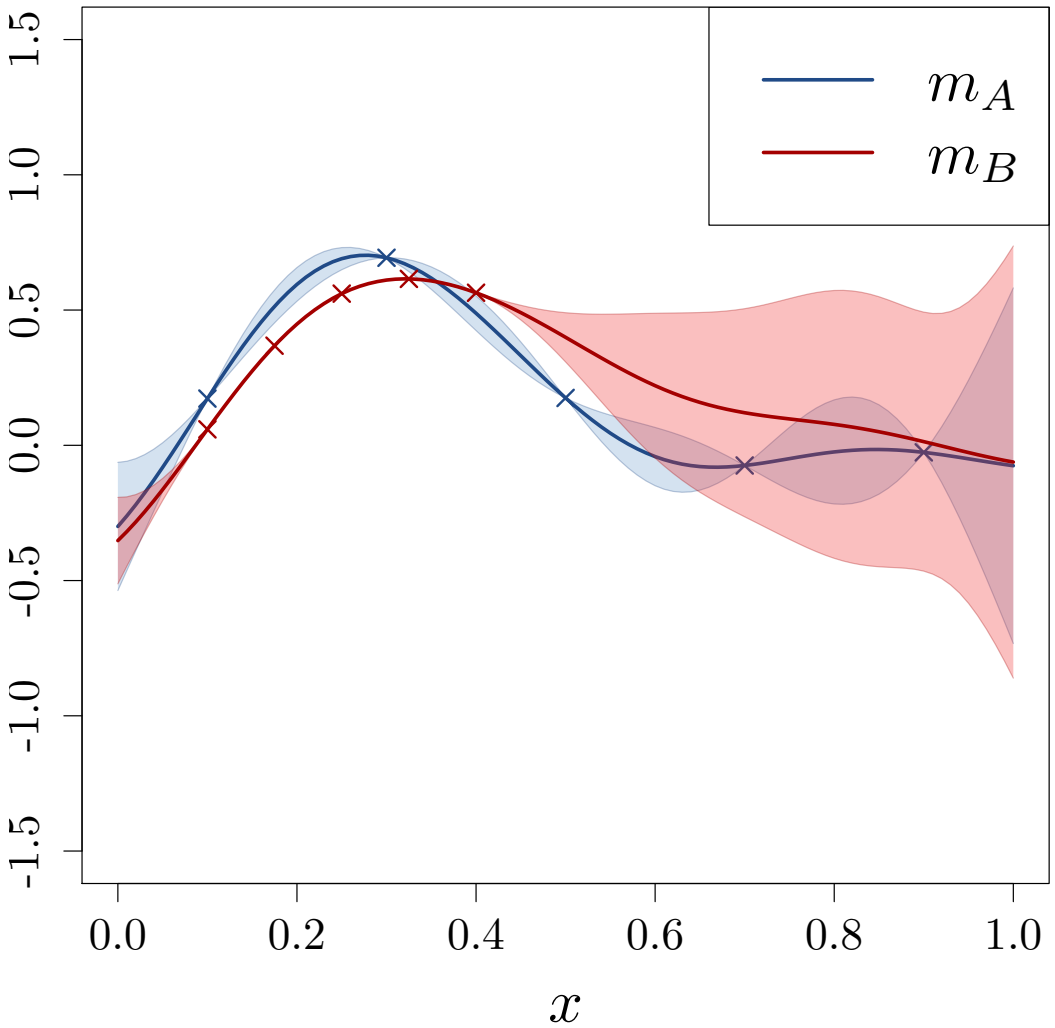
$$\begin{pmatrix} k_A(t, X_A) & k_{AB}(t, X_B) \end{pmatrix} \begin{pmatrix} k_A(X_A, X_A) & k_{AB}(X_A, X_B) \\ k_{AB}(X_A, X_B)^t & k_B(X_B, X_B) \end{pmatrix}^{-1} \begin{pmatrix} F_A \\ F_B \end{pmatrix}$$

The conditional covariance is:

$$\begin{aligned} c_A(t, t') &= \mathrm{cov}[T_A(t),\, T_A(t')|\, T_A(X_A){=}F_A,\, T_B(X_B){=}F_B] \\ &= k_A(t, t') - \begin{pmatrix} k_A(t, X_A) & k_{AB}(t, X_B) \end{pmatrix} \\ &\qquad \times \begin{pmatrix} k_A(X_A, X_A) & k_{AB}(X_A, X_B) \\ k_{AB}(X_A, X_B)^t & k_B(X_B, X_B) \end{pmatrix}^{-1} \begin{pmatrix} k_A(t', X_A)^t \\ k_{AB}(t', X_B)^t \end{pmatrix} \end{aligned}$$

Example

If we do the same thing fot T_B we obtain:



Instead of considering the GP to be multioutput, it is possible to see the GP as having one input but one extra categorical variable:

$$Z(t, c) = \begin{cases} Z_A(t) & \text{if } c = A \\ Z_B(t) & \text{if } c = B. \end{cases}$$

Exercise:

Compute the kernel of Z.

Instead of considering the GP to be multioutput, it is possible to see the GP as having one input but one extra categorical variable:

$$Z(t,c) = \begin{cases} Z_A(t) & \text{if } c = A \\ Z_B(t) & \text{if } c = B. \end{cases}$$

Exercise:
Compute the kernel of Z.

With this settings, the conditional mean

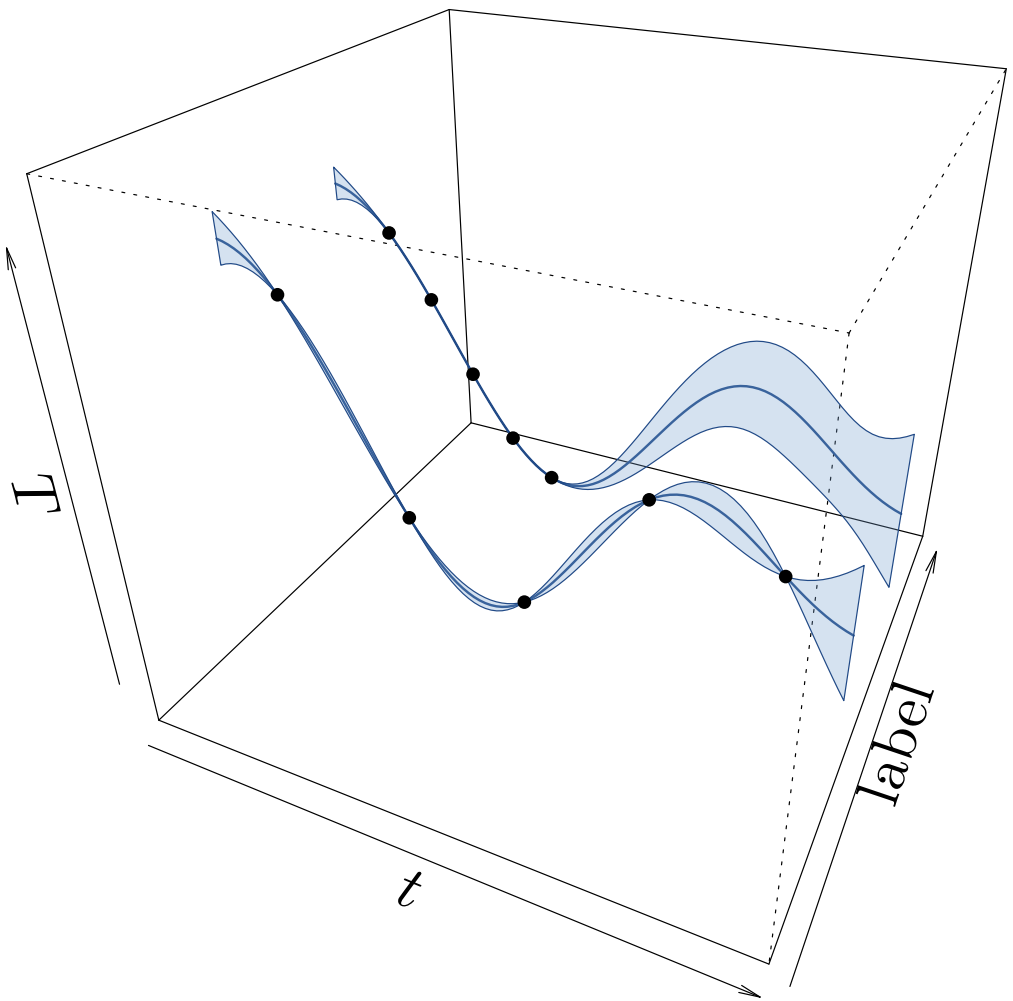
$$m_A(t) = \begin{pmatrix} k_A(t, X_A) & k_{AB}(t, X_B) \end{pmatrix} \begin{pmatrix} k_A(X_A, X_A) & k_{AB}(X_A, X_B) \\ k_{AB}(X_A, X_B)^t & k_B(X_B, X_B) \end{pmatrix}^{-1} \begin{pmatrix} F_A \\ F_B \end{pmatrix}$$

writes as an usual conditional mean

$$m_A(t) = m(t, A) = \mathrm{E}[T(t, A) | T(X) = F] = k(\begin{pmatrix} t \\ A \end{pmatrix}, X) k(X, X)^{-1} F$$

Example

We obtain this representation for the model



In the end, multioutputs GPs can be seen as GPs with one extra categorical variable indicating the output label.

All the math stay the same, we just need to specify a covariance function that takes into account this extra variable. A common approach is to consider a product covariance structure

$$k\left(\begin{pmatrix} t \\ c \end{pmatrix}, \begin{pmatrix} t' \\ c' \end{pmatrix}\right) = k_{cont}(t, t')k_{disc}(c, c')$$

where $k_{disc}(c, c')$ can be described by a covariance matrix. In practice, this covariance matrix has to be estimated.

If there are **2 outputs** (or 2 levels for the categorical variable), it is a 2×2 covariance matrix. It can be parameterised by

$$k_{disc} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$$

with $\sigma_1,\ \sigma_2 \geq 0$ and $\rho \in [-1, 1]$. The latter can be estimated by ML.

In higher dimension (say k), it is possible to consider the following parameterization for k_{disc}:

$$k_{disc} = WW^T$$

where W is a $k \times l$ matrix. The choice of l allows to tune the complexity of the estimation.

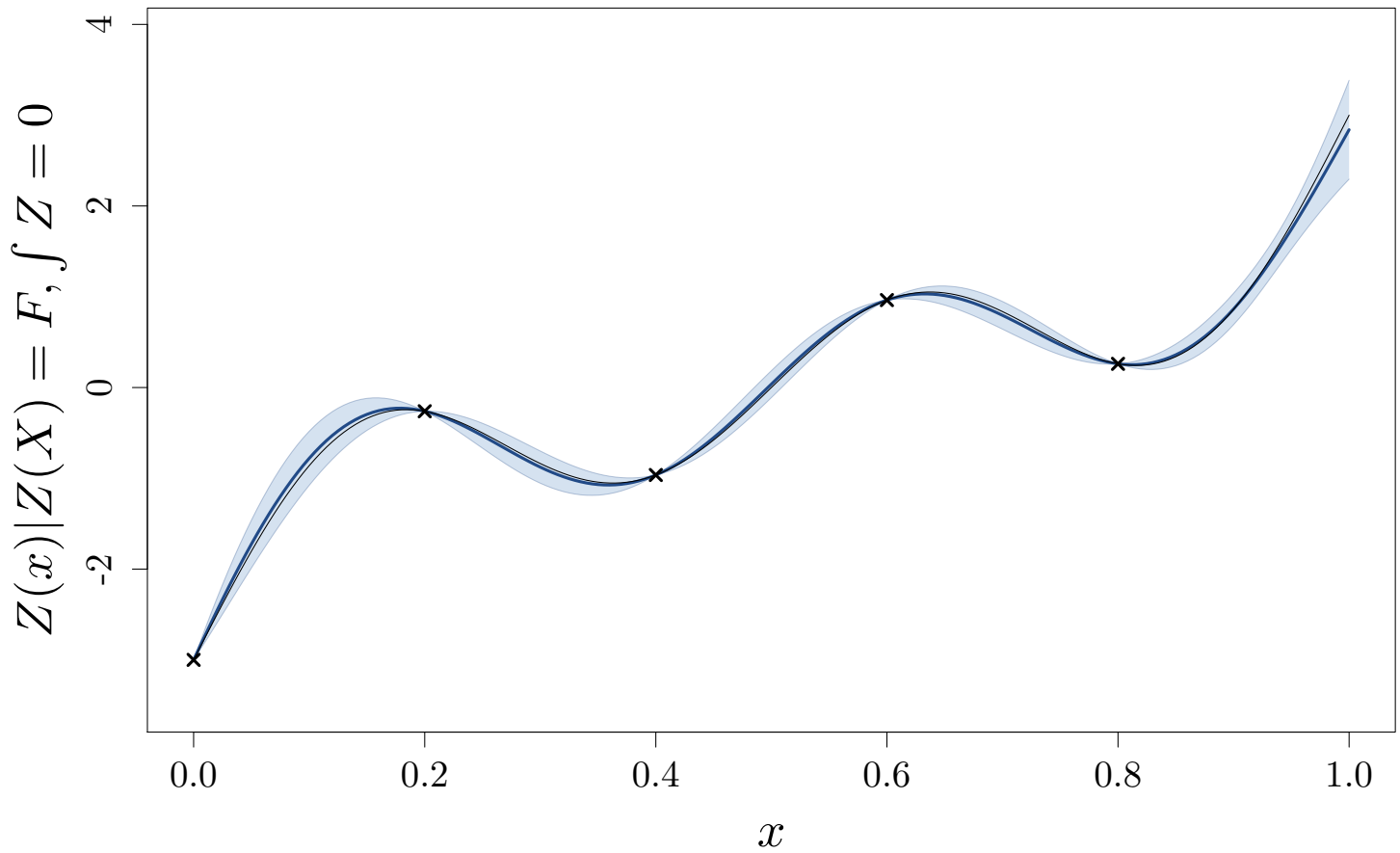
It is also possible to include in models observations more sophisticated than $Z(X) = F$...

For instance, if we know the integral of the function to approximate and it's derivative in a few points, we want to consider

$$Z \;\Big|\; Z(X) = F, \int Z = a, \frac{\mathrm{d}Z}{\mathrm{d}x}(X') = F'$$

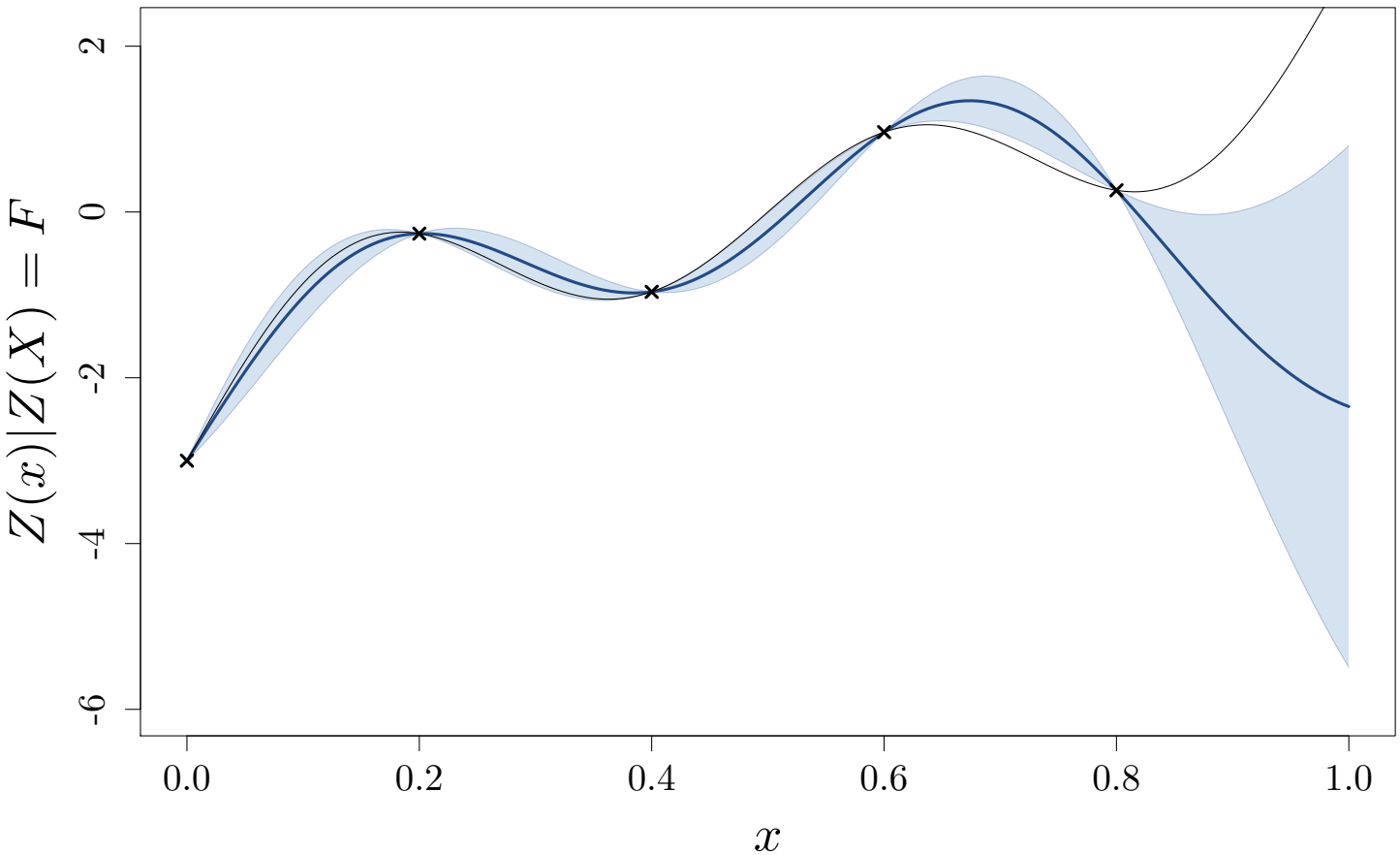
Example

If we take into account that the function is centred, we obtain:

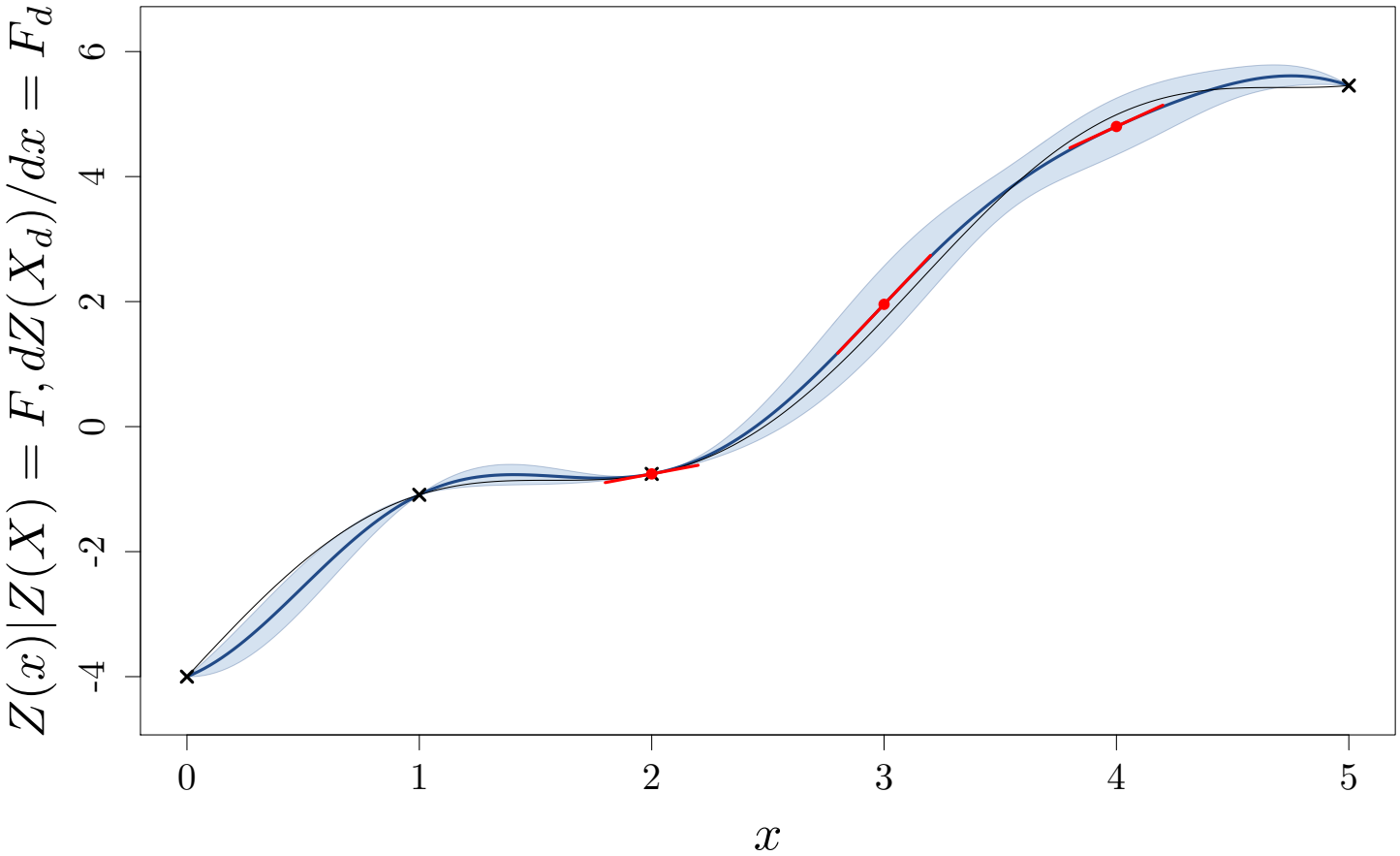


Example

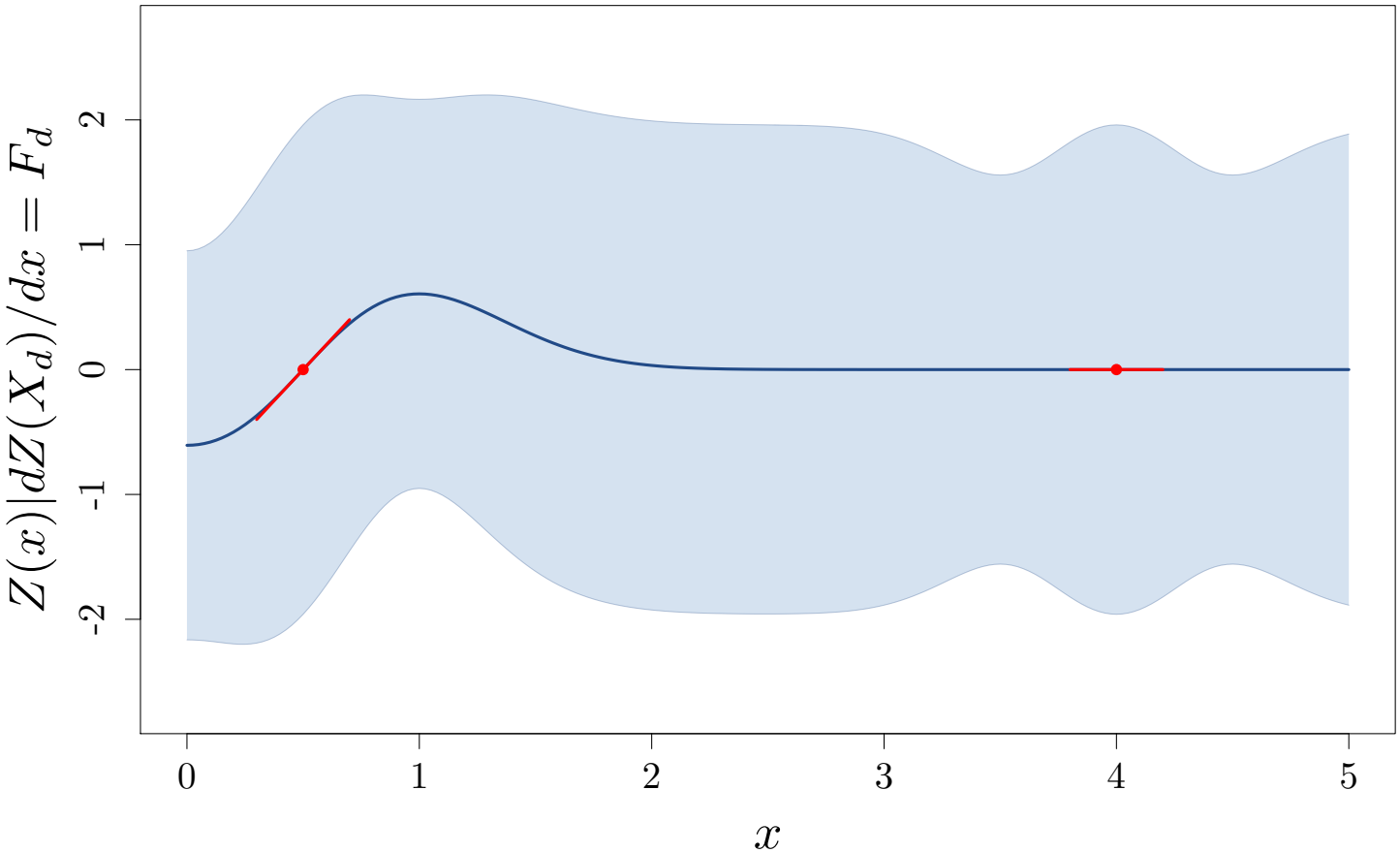
Whereas if we ignore it:



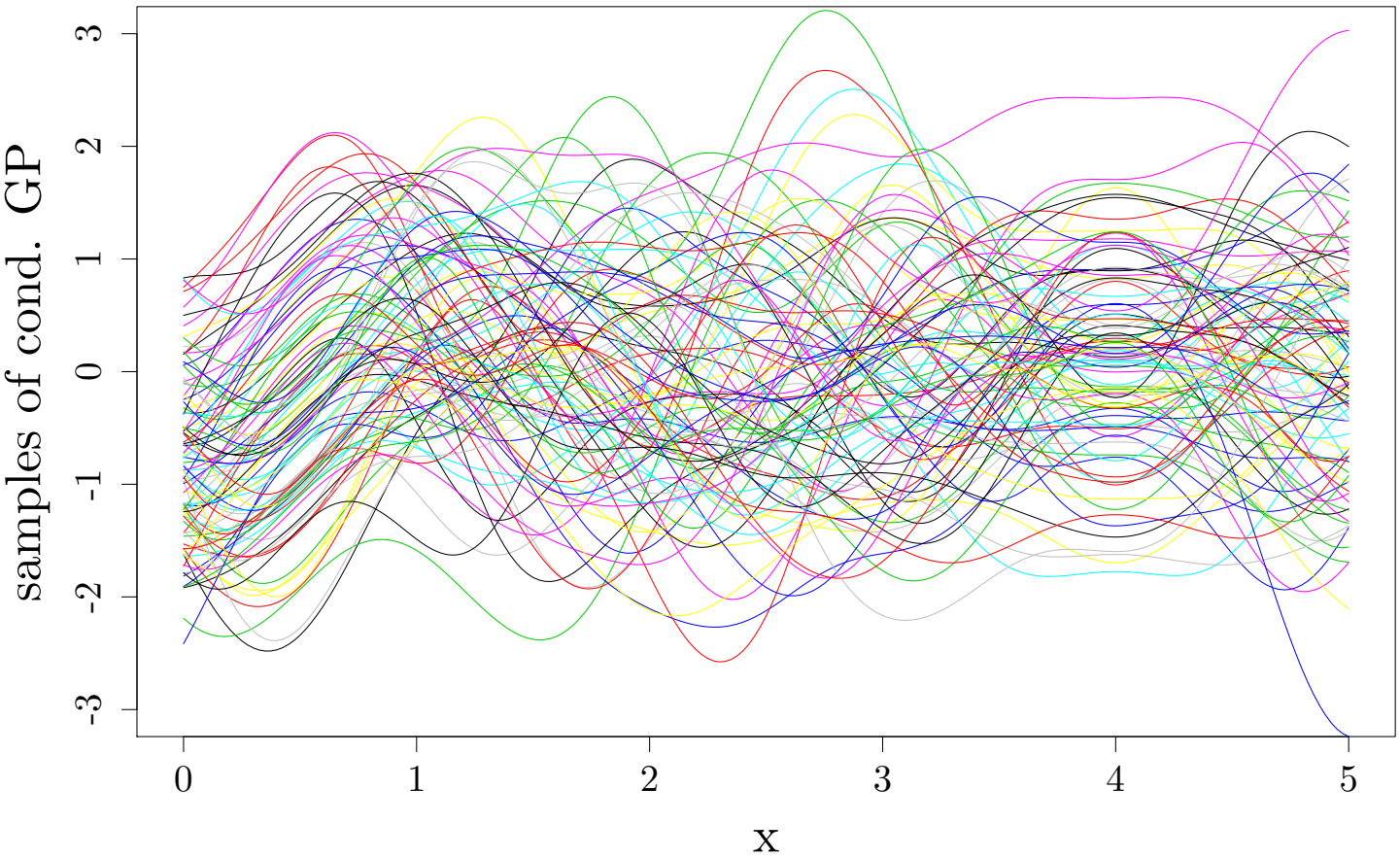
Similarly, we can include in a model some derivative observations:



We can see interesting behaviour if we look at a model with only derivatives.



As always, we can simulate conditional paths:



Example of GPR application: Detecting periodicity in gene expression

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:

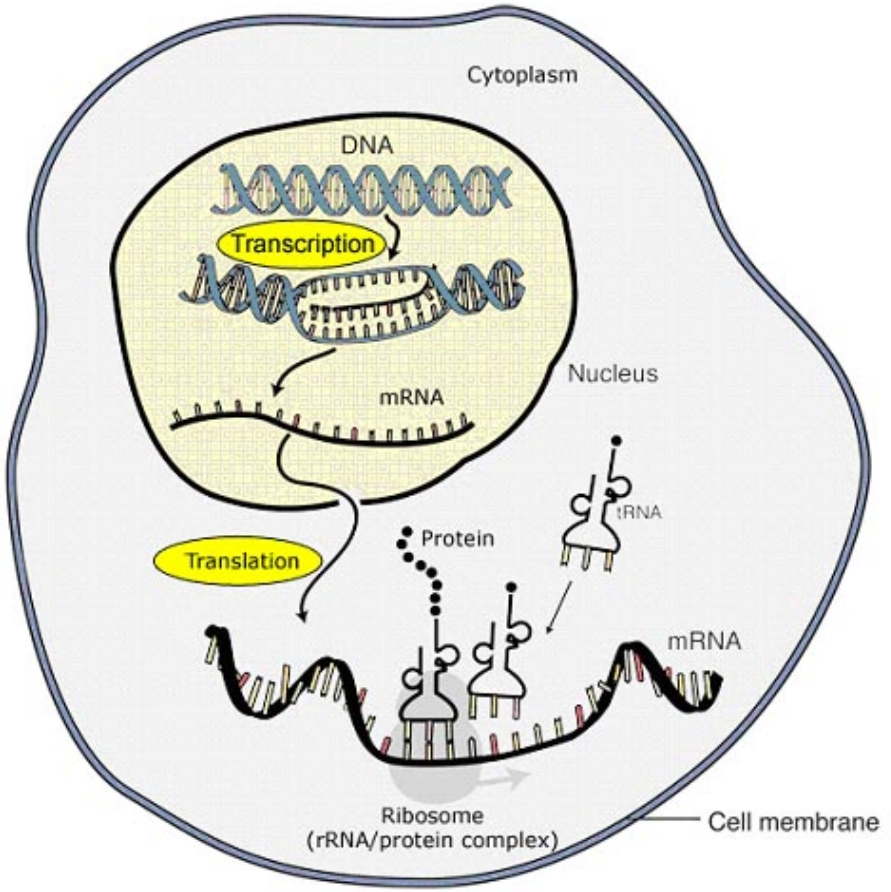
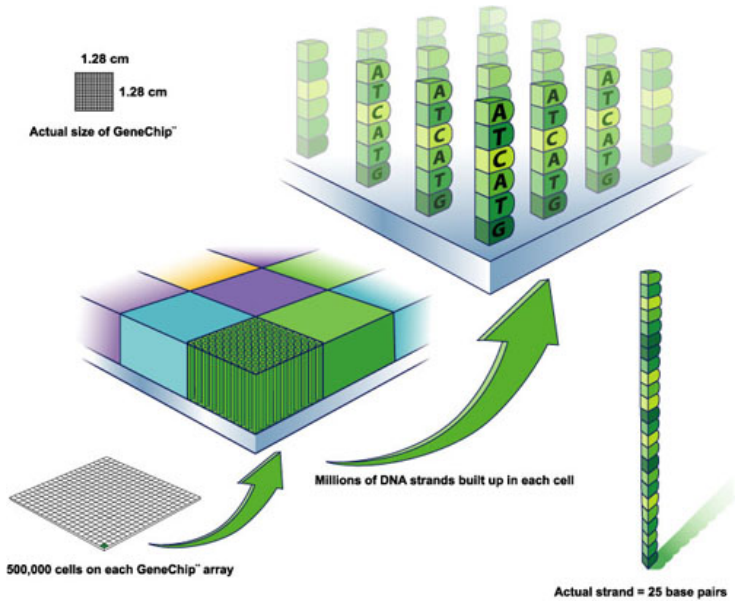


Image adapted from: National Human Genome Research Institute.

The mRNA concentration is measured with microarray experiments



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

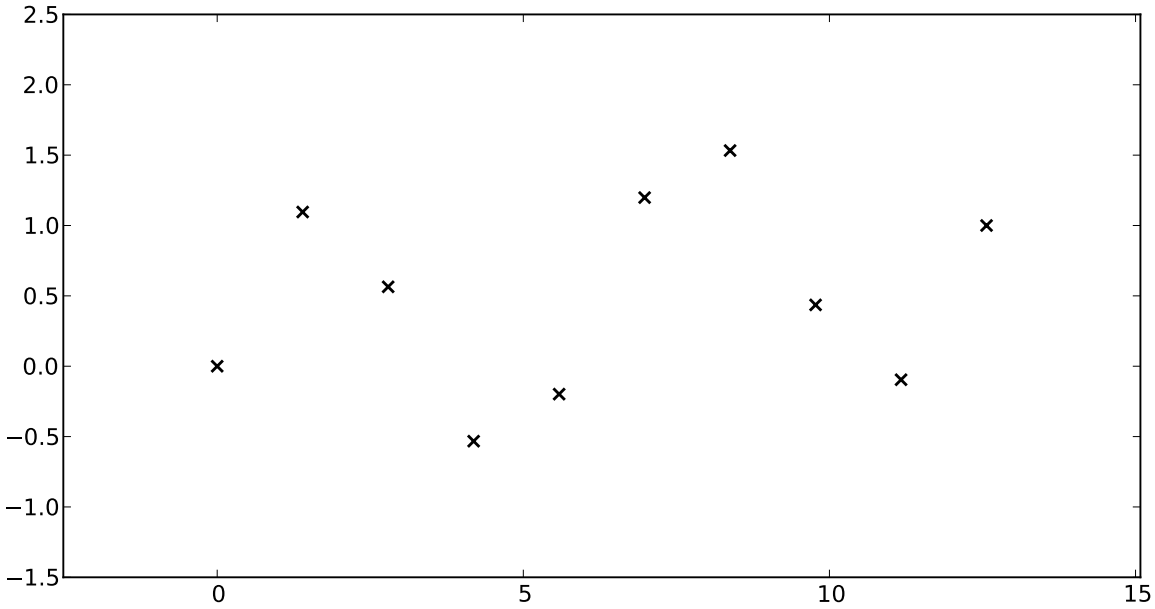
Experiments to study the circadian clock are typically:

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.

$\Rightarrow$ identifying periodically expressed genes gives an insight on the overall mechanism.

In practice, we have for each gene:



Can we extract the periodic part of a signal ?

Let Z be a GP and $B(t) = (\sin(t), \cos(t), \ldots, \sin(nt), cos(nt))^t$ be the fourier basis. We consider the projection of Z onto the basis:

$$Z_p(t) = \frac{\langle Z, \sin \rangle}{\langle \sin, \sin \rangle} \sin(t) + \frac{\langle Z, \cos \rangle}{\langle \cos, \cos \rangle} \cos(t) + \cdots + \frac{\langle Z, \cos(n.) \rangle}{\langle \cos(n.), \cos(n.) \rangle} \cos(nt)$$

This give a decomposition of the GP:

$$Z = Z_p + \underbrace{Z - Z_p}_{Z_a}.$$

By considering the appropriate inner product, we can ensure that Z_p and Z_a are independant.

Property

The reproducing kernel of Z_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

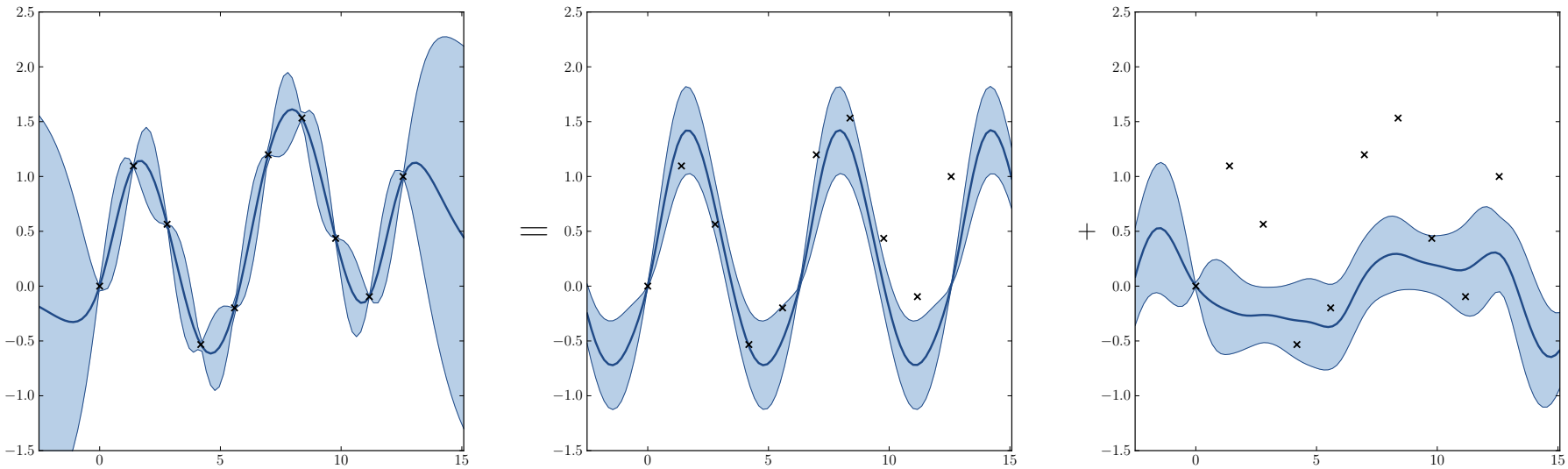
$$\begin{aligned} m(t) &= (k_p(t) + k_a(t))^t (K_p + K_a)^{-1} F \\ &= \underbrace{k_p(t)^t (K_p + K_a)^{-1} F}_{\text{periodic sub-model } m_p} + \underbrace{k_a(t)^t (K_p + K_a)^{-1} F}_{\text{aperiodic sub-model } m_a} \end{aligned}$$

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.

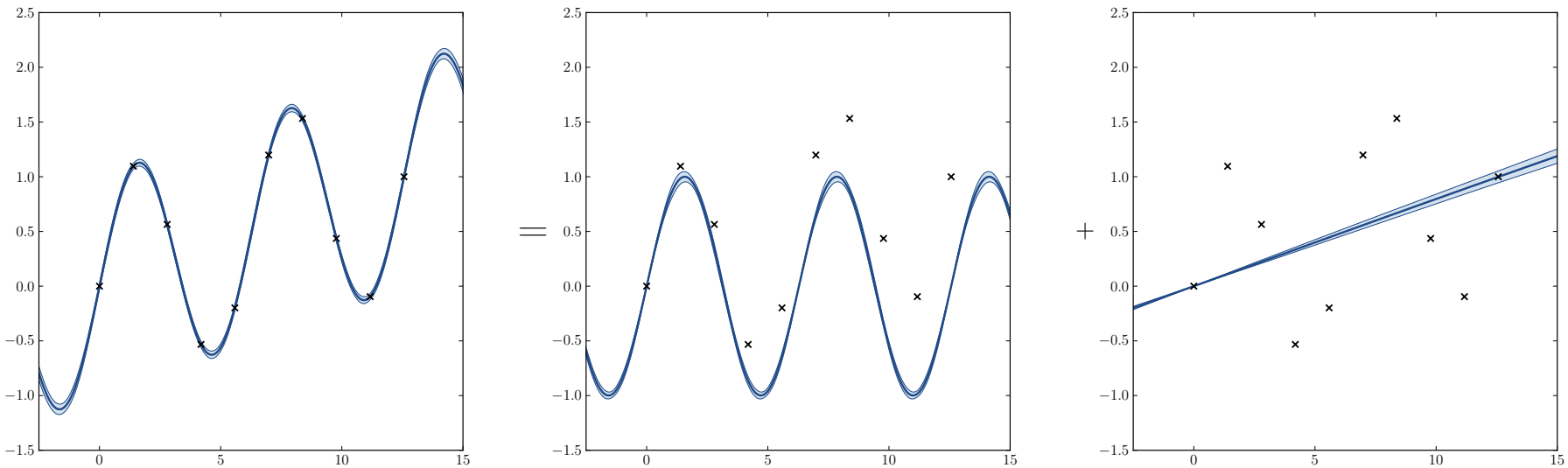
Let k^* be defined as

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

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

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

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



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

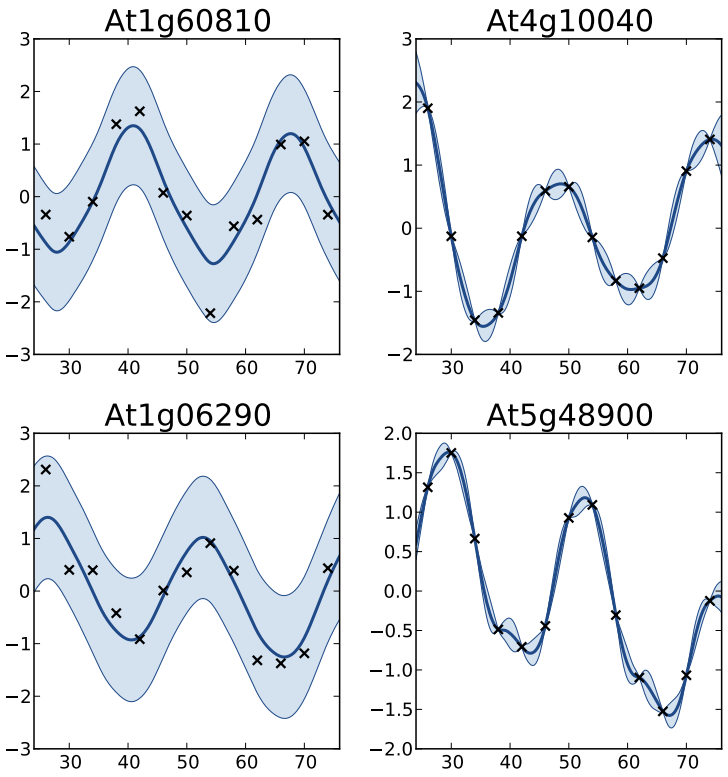
The dimension of the data is:

- 22810 genes
- 13 time points

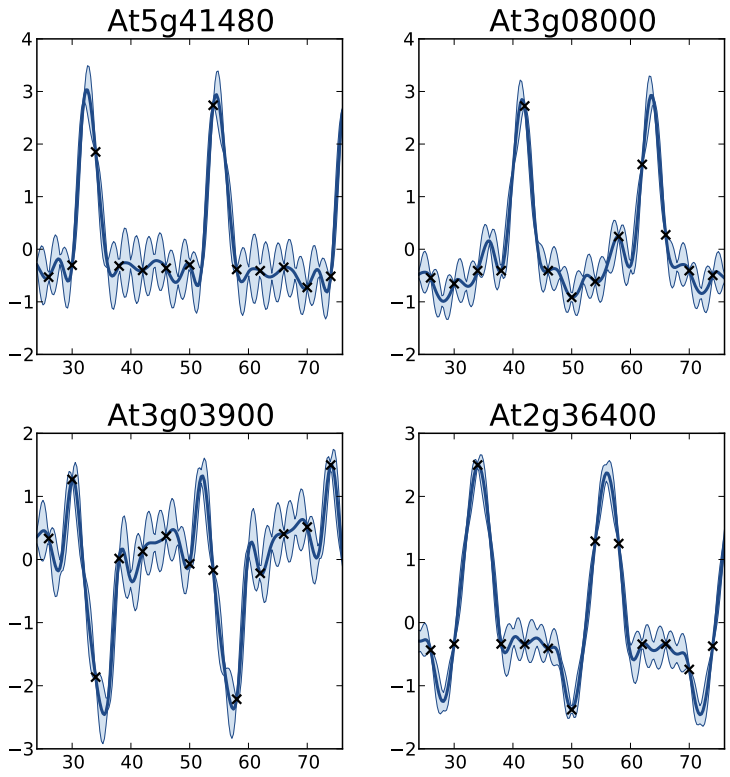
Edward 2006 gives a list of the 3504 most periodically expressed genes. The comparison with our approach gives:

- 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

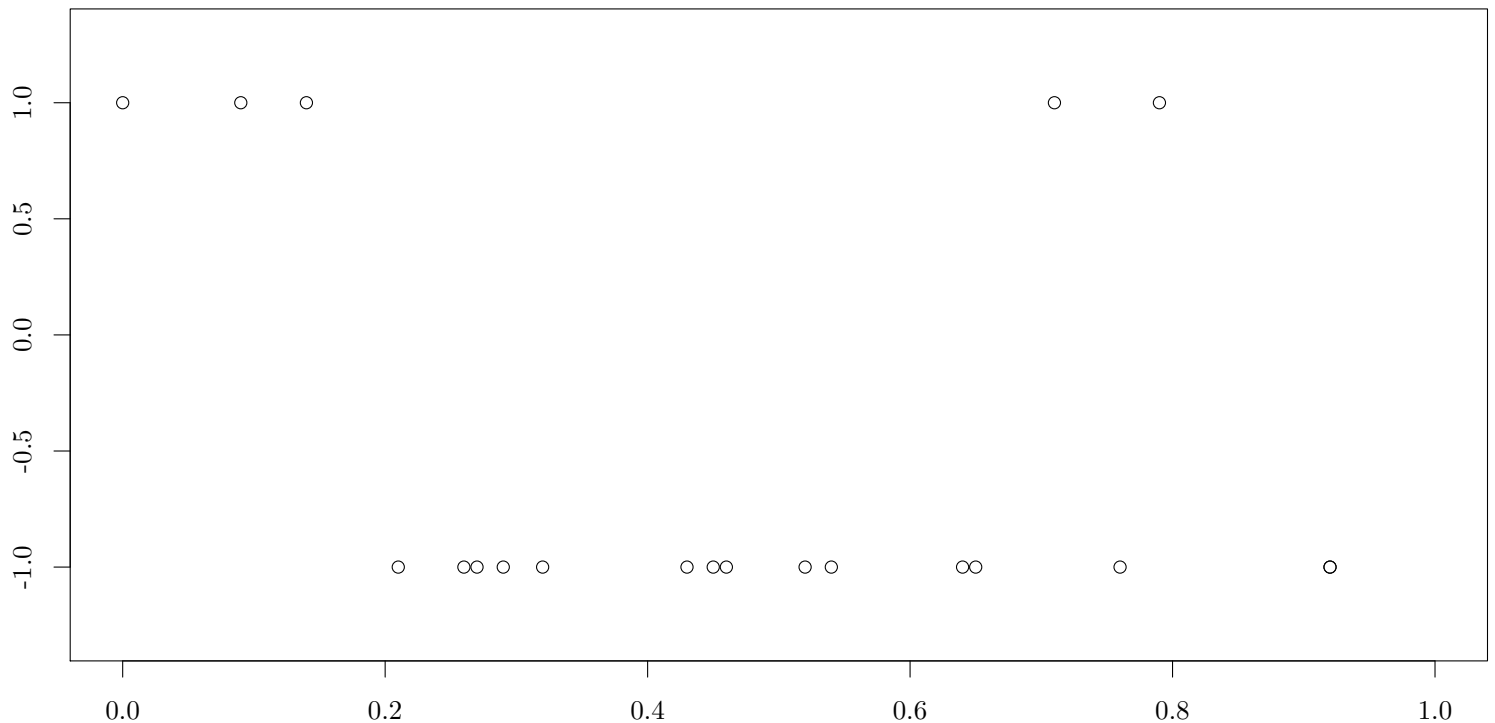


periodic for our approach

GP models for classification

Until now, we have focused on GP **Regression**: we were using GPs to predict a continuous output given some input values.

Gaussian process models are also useful for **classification** problems:



We consider the following probabilistic model for the data

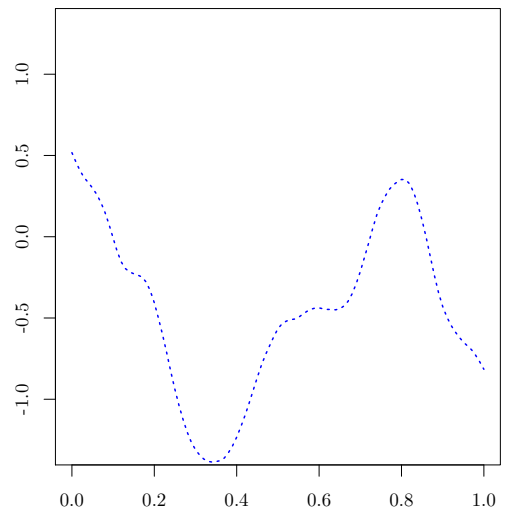
1. Let Y be a Gaussian process over $\mathbb{R}$.
2. Let Φ be a sigmoid transformation such as

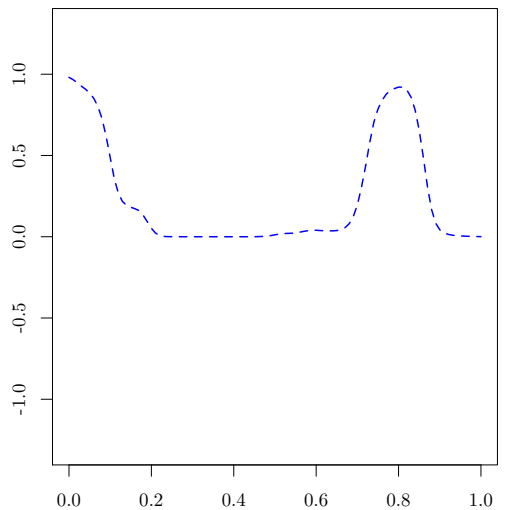
$$\Phi(y) = \frac{1}{1+e^{-y}} \qquad \textbf{or} \qquad \Phi \text{ is the Gaussian cdf.}$$

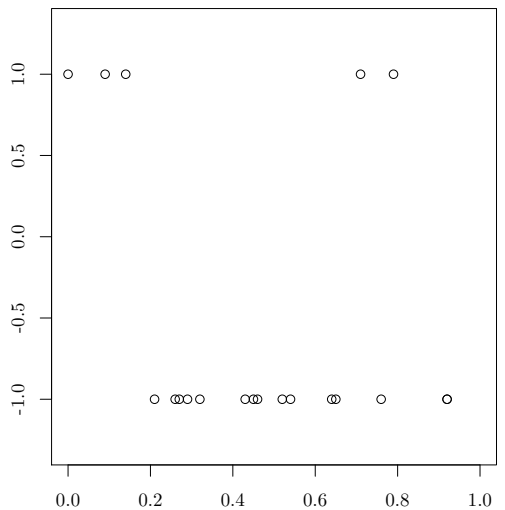
3. We denote by Z the image of Y by Φ: $Z(x) = \Phi(Y(x))$.
4. The observation $F_i \in \{-1,\ 1\}$ at input X_i is given by a Bernouilli sample with parameter $Z(X_i)$.

Same thing with images

GP $Y(x)$ $Z(x) = \Phi(Y(x))$ $F_i \sim \mathcal{B}(Z(X_i))$ iid





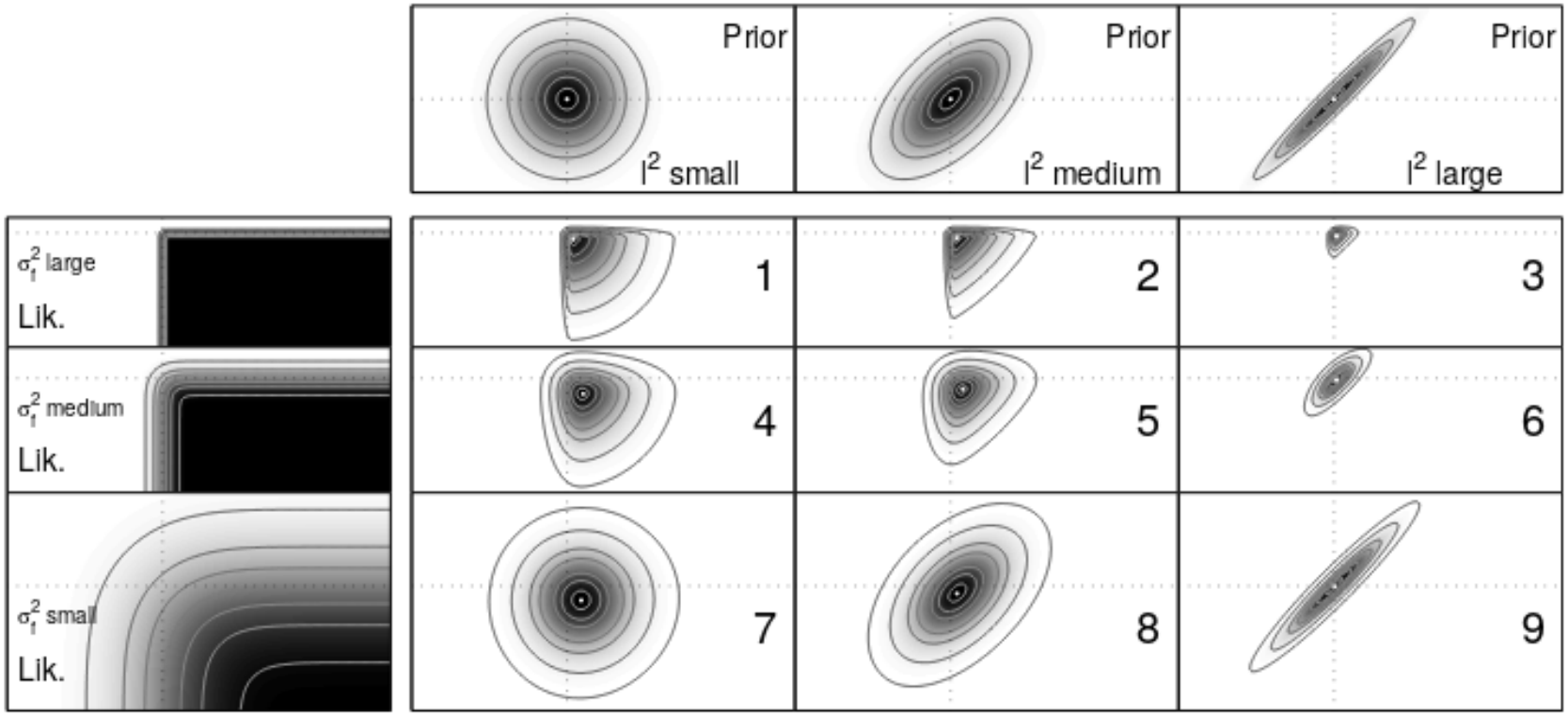


In order to make predictions, we need to compute the conditional distribution of Y (or Z) given the observed data: Y and Z are called latent variables because we never have observations of them.

Exercise

1. Is the vector (Y, F) a Gaussian vector?
2. What can you say about $P(F_i|Y(X_i))$? Deduce the probability of $P((F_1, F_2)|Y(X_1), Y(X_2))$.
3. Use Bayes rule to re-write the conditional pdf of $(Y(X_1), Y(X_2))$ given the observations (F_1, F_2).
4. Give a graphical representation of this conditional distribution.

Examples of posteriors



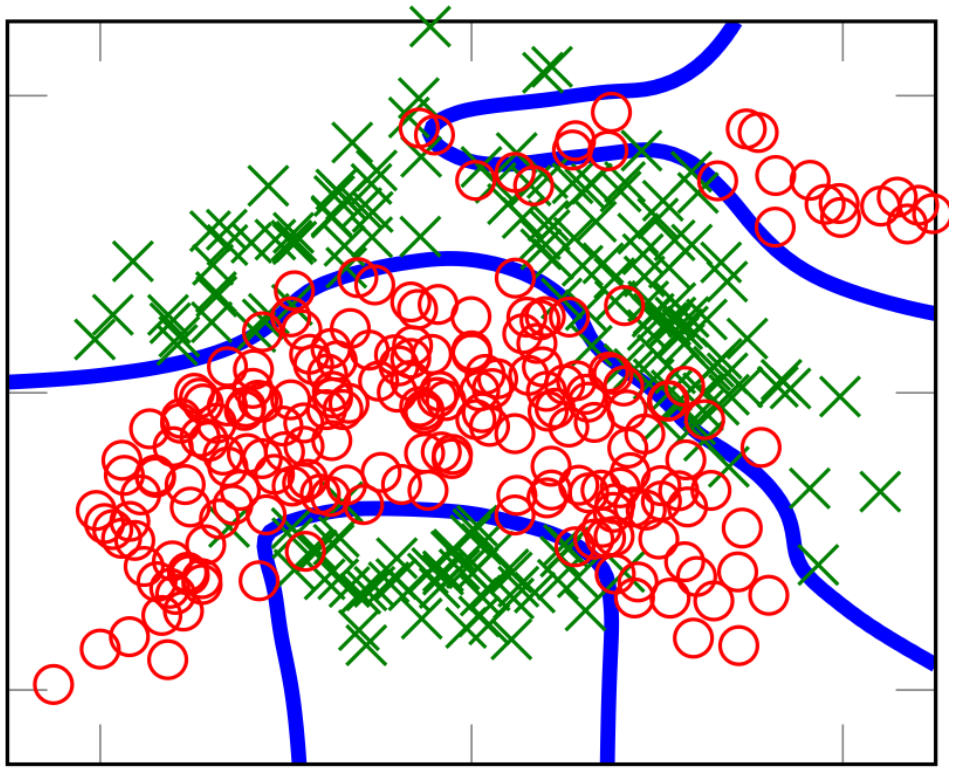
source: Nickisch and Rasmussen, JMLR 2008.

The conditional distribution of $Y|F$ is not Gaussian... In practice it is possible to:

- Use the mode of the distribution.
- Sample from the distribution using MCMC.
- Make a Gaussian approximation of this non Gaussian distribution:
  - Laplace method
  - Variational inference
  - ...

Once the distribution of $Y(X)|F$ has been approximated, we can deduce the distribution of $Y(x)|F$. It is then possible to obtain predictions for the labels of future observations at x.

As for GPR, GP classification models are straightforward to generalize in higher dimension



source: Hensman, Durrande and Solin, arXiv 2016.

Conclusion

We have seen that

- Gaussian processes are a great tool for modeling
  - Regression and classification
- Kernels can (and should) be tailored to the problem at hand
- It is possible to include in models more that function values

What we have not seen...

- Unsupervised models: non-linear generalization of PCA
- How to deal with a (very) large number of observations
- links with the RKHS theory
- ...

Gaussian process models are of **particular interest in industry**, and they are also an **active research topic**...