

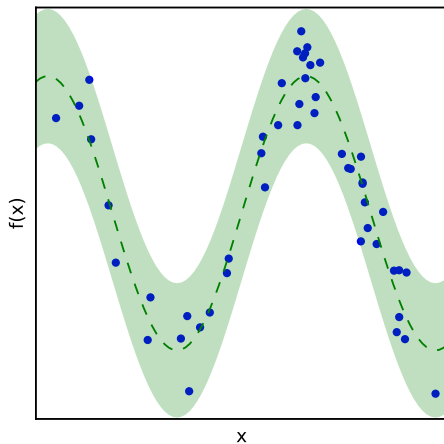
# Supervised Learning

## Gaussian processes for regression

Gilles Louppe (@glouppe)

October 17, 2016

# Problem statement



Regression, with error bars

# Problem statement

- Training data  $\mathcal{L} = \{\mathbf{x}_i, f_i\}_{i=1}^N$
- $\mathbf{x}_i \in \mathbb{R}^d$
- $y_i = f(\mathbf{x}_i)$  for some unknown  $f$
- We wish to recover the underlying process  $f$  from the observed data, i.e infer  $f$  for some unseen  $\mathbf{x}$ , using  $p(f|\mathcal{L})$ .

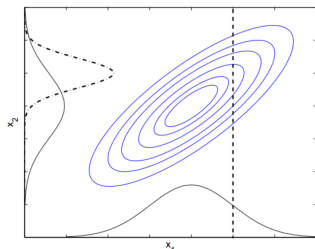
# The multivariate gaussian distribution

Let assume a random variable  $\mathbf{f}$  following a multivariate Gaussian distribution, and let partition its dimensions into two sets  $A$  and  $B$ :

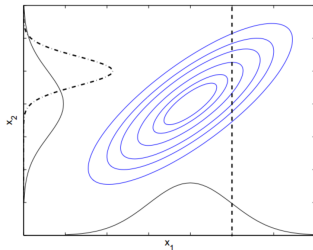
$$\underbrace{f_1, \dots, f_i}_{\mathbf{f}_A}, \underbrace{f_{i+1}, \dots, f_N}_{\mathbf{f}_B} \sim \mathcal{N}(\boldsymbol{\mu}, K)$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix} \in \mathbb{R}^N$$

$$K = \begin{bmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{bmatrix} \in \mathbb{R}^{N \times N}$$



# Marginal and conditional distributions



- The *marginal* distribution of a multivariate gaussian is a multivariate gaussian:

$$\mathbf{f}_A \sim \mathcal{N}(\mu_A, K_{AA})$$

- The *conditional* distribution of a multivariate gaussian is a multivariate gaussian:

$$\mathbf{f}_A | \mathbf{f}_B \sim \mathcal{N}(\mu_A + K_{AB} K_{BB}^{-1} (\mathbf{f}_B - \mu_B), K_{AA} - K_{AB} K_{BB}^{-1} K_{BA})$$

# Gaussian Processes

**Definition.** A *Gaussian process* is a (potentially infinite) collection of random variables such that the joint distribution of any finite number them is multivariate Gaussian.

# Gaussian Processes: the simpler explanation

*A Gaussian process is a*

**HUGE<sup>1</sup>**

multivariate gaussian distribution.

---

<sup>1</sup>The dimension is the number of data points.

# Gaussian distributions vs. Gaussian processes

## Gaussian distribution

$$x \sim \mathcal{N}(\mu, K)$$

- Distribution over vectors.
- Fully specified by a mean and covariance.
- The position of the random variable in the vector plays the role of the index.

## Gaussian process

$$f \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot))$$

- Distribution over functions.
- Fully specified by a mean *function*  $m$  and a covariance *function*  $k$ .
- The argument of the random function plays the role of the index.



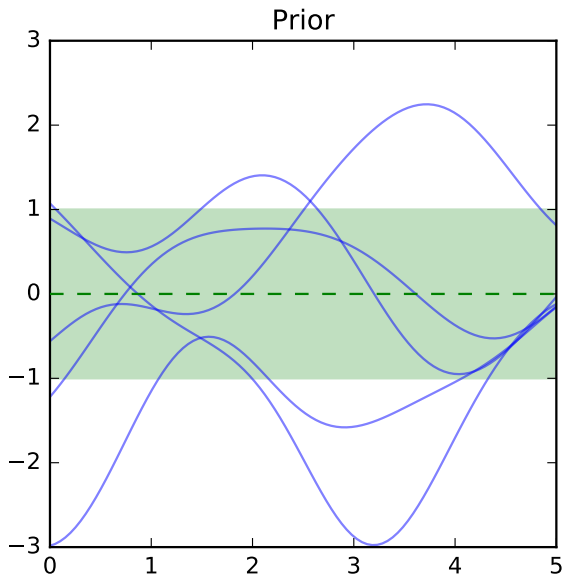
## Gaussian process prior

For  $m(\cdot) = 0$  and for any set  $A = \mathbf{x}_1, \dots, \mathbf{x}_M$  of test points, we may compute the covariance matrix  $K_{AA}$ , which defines a joint distribution  $p(\mathbf{f}_A)$  over function values at those points:

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_M) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, K_{AA})$$

That is, we are marginalizing over the random variables not including in the test points.

## Gaussian process prior



## Gaussian process posterior

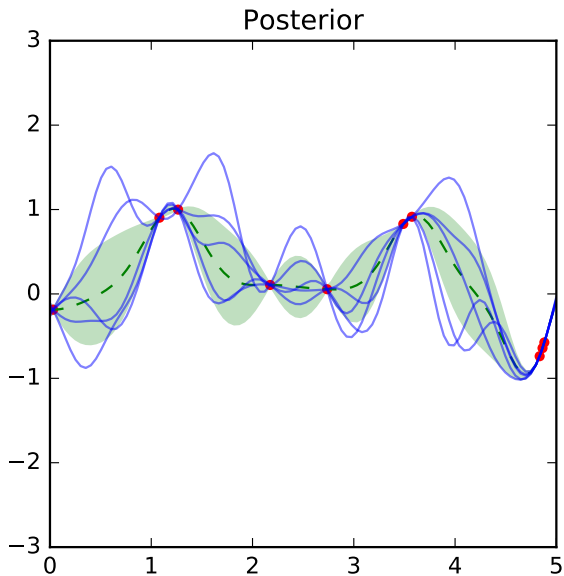
Given training data  $\mathcal{L} = (B = \mathbf{x}_1, \dots, \mathbf{x}_N; \mathbf{f}_B = f(\mathbf{x}_1), \dots, f(\mathbf{x}_N))$  and test points  $A = \mathbf{x}_1, \dots, \mathbf{x}_M$ , we may similarly derive the joint distribution

$$\begin{bmatrix} \mathbf{f}_A \\ \mathbf{f}_B \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \begin{bmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{bmatrix})$$

on which we can condition  $\mathbf{f}_B$  on the known values from  $\mathcal{L}$ , resulting in the posterior distribution  $p(\mathbf{f}_A|\mathcal{L})$ :

$$\mathbf{f}_A \sim \mathcal{N}(K_{AB}K_{BB}^{-1}\mathbf{f}_B, K_{AA} - K_{AB}K_{BB}^{-1}K_{BA})$$

## Gaussian process posterior



## Covariance functions (or kernels)

The covariance function  $k(\cdot, \cdot)$  encodes the covariance between pairs of random variables  $\mathbf{x}_i, \mathbf{x}_j$ . It must be positive semi-definite and symmetric.

Popular examples include:

- The squared exponential function (RBF)
- The Matern kernel
- The linear kernel
- The polynomial kernel
- The white noise kernel

Kernels can be composed together to describe complex interactions.

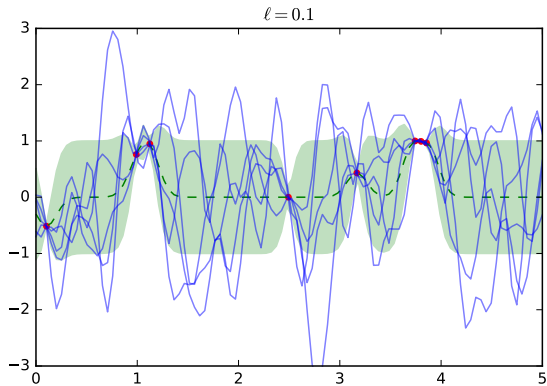
# Squared exponential function

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\ell^2}\right)$$

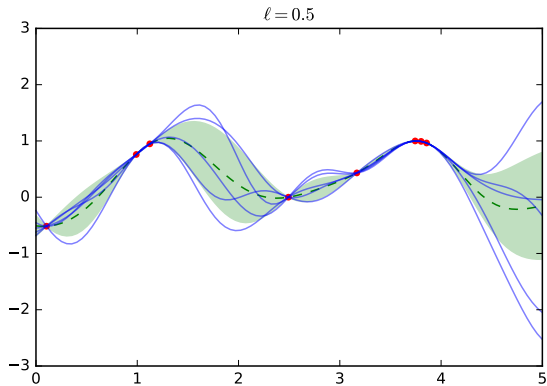
Hyper-parameters:

- The length scale  $\ell$  describes the smoothness of the function.
- The output variance  $\sigma^2$  determines the average distance of the function away from its mean.

## Squared exponential function ( $\ell = 0.1$ )

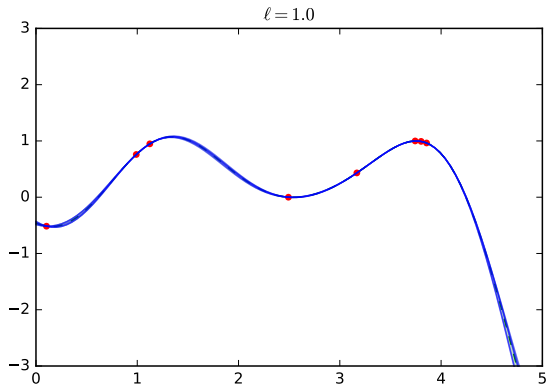


## Squared exponential function ( $\ell = 0.5$ )





## Squared exponential function ( $\ell = 1.0$ )



# Hyper-parameters

- So far we have assumed that the gaussian process prior  $p(\mathbf{f})$  was specified a priori.
- However, this distribution has itself parameters. E.g.,  $\ell$  and  $\sigma$  when using the squared exponential function.
- Let  $\theta$  denote the vector of hyper-parameters. How do we learn  $\theta$ ?

## Marginal likelihood

Given training data  $\mathcal{L} = (B = \mathbf{x}_1, \dots, \mathbf{x}_N; \mathbf{f}_B = f(\mathbf{x}_1), \dots, f(\mathbf{x}_N))$ , we can derive the prior

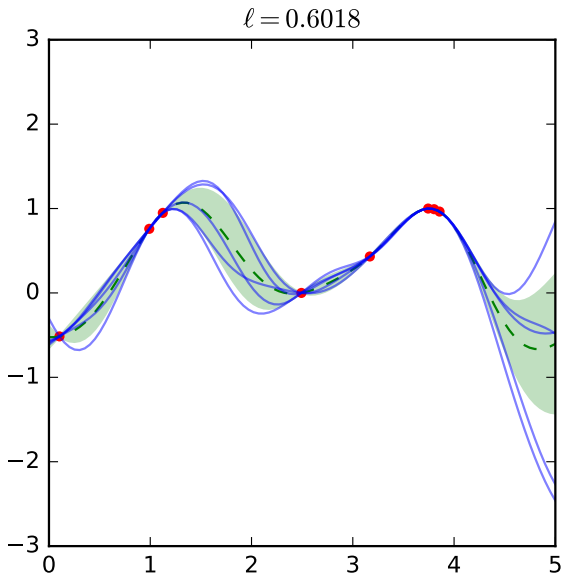
$$\begin{bmatrix} f(\mathbf{x}_1) \\ \dots \\ f(\mathbf{x}_N) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, K_{BB;\theta})$$

at the training points.

Let select  $\theta$  to maximize the likelihood  $p(\mathbf{f}_B; \theta)$  of the training observations under that prior:

$$\begin{aligned} \theta^* &= \arg \min_{\theta} -\log p(\mathbf{f}_B; \theta) \\ &= \arg \min_{\theta} -\frac{1}{2} \log \det K_{BB;\theta} - \frac{1}{2} \mathbf{f}_B^T K_{BB;\theta}^{-1} \mathbf{f}_B + c \end{aligned}$$

## Squared exponential function ( $\ell^*$ )



# Summary

- Gaussians processes = multivariate gaussian in infinite dimension.
- Provide a principled approach to derive a posterior distribution  $p(\mathbf{f}|\mathcal{L})$ .
- They are non-parametric, but often require a careful design of the covariance function.
- Gaussian processes extend to classification (not covered)
- They do not scale well to many observations and/or many features.