Friedrich-Alexander-Universität
Technische Fakultät

Machine Learning
Data Analytics

Machine Learning for Time Series

(MLTS or MLTS-Deluxe Lectures)

Dr. Dario Zanca
Machine Learning and Data Analytics (MaD) Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
25.10.2022

Topics overview

- Time series fundamentals and definitions (2 lectures)
- Bayesian Inference (1 lecture)
- Gussian processes (2 lectures) ←
- State space models (2 lectures)
- Autoregressive models (1 lecture)
- Data mining on time series (1 lecture)
- Deep learning on time series (4 lectures)
- Domain adaptation (1 lecture)

In this lecture…

1. Prior on functions
2. Gaussian processes
3. Gaussian process regression

Gaussian Process Regression

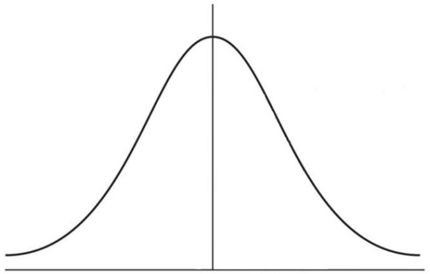
Prior on functions

Concepts review: Gaussian Distribution

Univariate vs. Multivariate

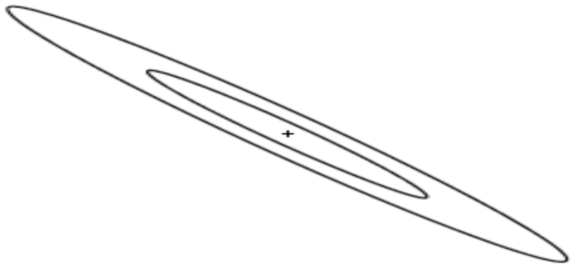
The **univariate Gaussian distribution** is given by

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



The **multivariate Gaussian distribution** is given by

$$\mathcal{N}(x|\mu,\Sigma) = (2\pi)^{-D/2}|\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$



Concept review: Conditional and Marginal of a Gaussian

If x and y are jointly Gaussian

$$p(x,y) = p\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right)$$

we get the marginal distribution of x by

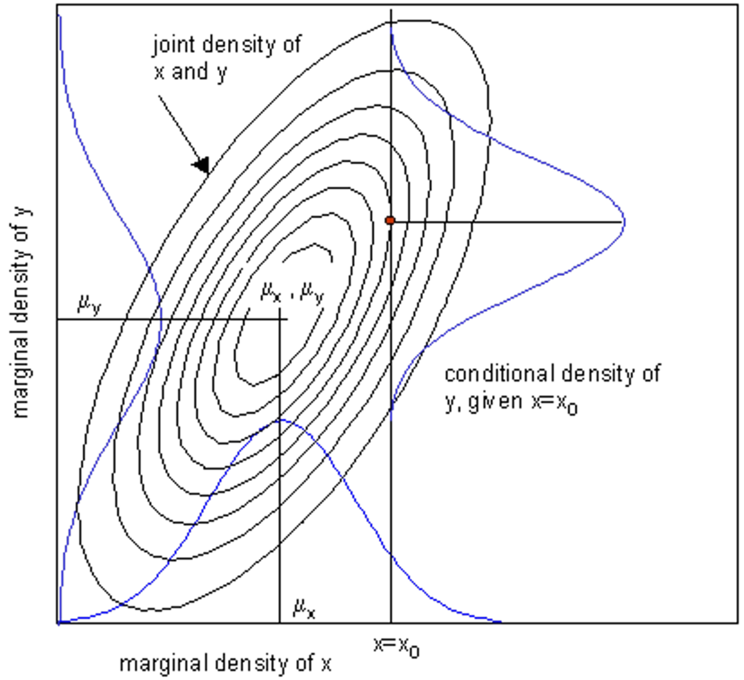
$$p(x,y) = \mathcal{N}\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right) \Rightarrow p(x) = \mathcal{N}(\mathrm{a}, \mathrm{A})$$

and the conditional distribution of x given y by

$$p(x,y) = \mathcal{N}\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A & B \\ B^\tau & C \end{bmatrix}\right) \Rightarrow p(x|y) = \mathcal{N}(\mathrm{a} + \mathrm{B}C^{-1}(y-b), A - BC^{-1}B^\tau)$$

where x and y can be scalars or vectors.

Both the conditional $p(x|y)$ and the marginal $p(x)$ of a joint Gaussian $p(x,y)$ are Gaussian.



Supervised learning

Bayesian view

In **supervised learning**, we:

- observe some inputs x_i and some outputs y_i
- Assume that $y_i = f(x_i)$
  - for some unknown function f
  - Possibly subject to noise

The optimal approach is to:

- infer a distribution over functions given the data, $p(f\,|X, y)$
- Then use it for prediction

$$p(y_*|x_*, X, y) = \int p(y_*|f, x_*)p(f|X, y)\, df$$

Prior on parameters

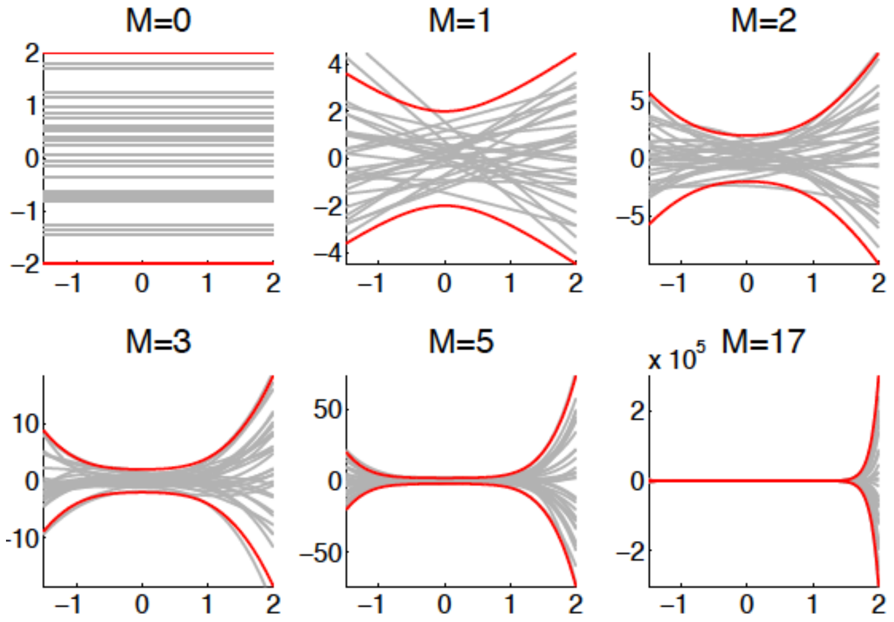
A model M is the result of the choice of:

- A **model structure**
- The **model's parameters**

In the example:

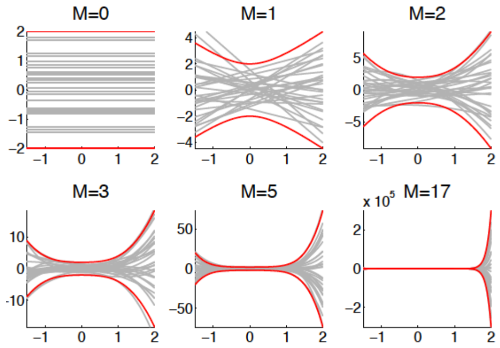
$$f_w(x) = \sum_{m=0}^{M} \omega_m \Phi_m(x), \quad \text{with } \Phi_m(x) = x^m$$

We have defined a prior distribution over functions but **in an indirect way**



Priors on functions

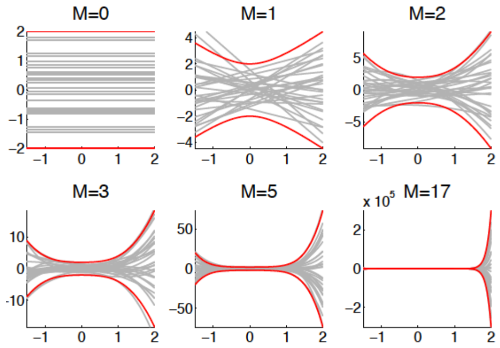
Models with priors on the weights *indirectly* specify priors over functions.

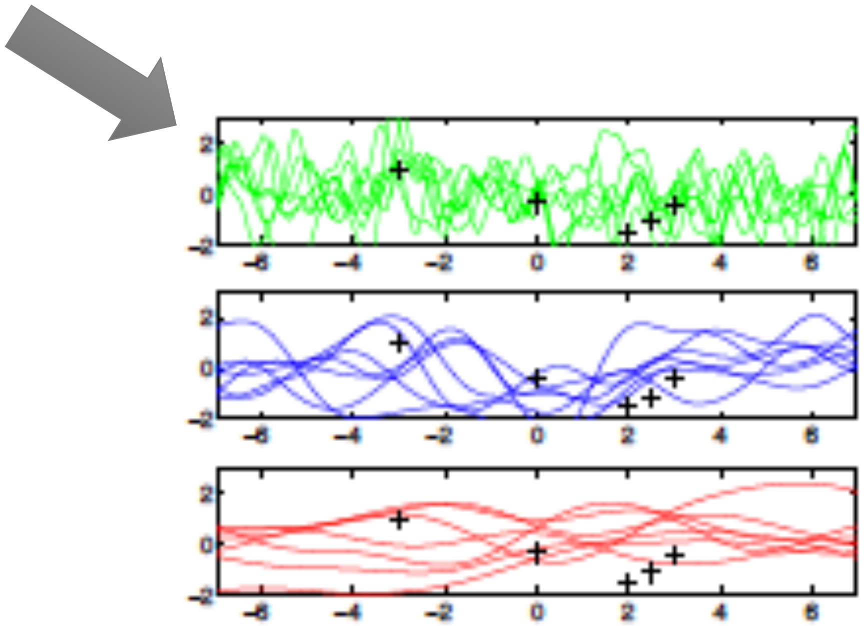


Priors on functions

Models with priors on the weights *indirectly* specify priors over functions.

- **What about specifying priors on functions directly?**
- **What does a probability density over functions even look like?**

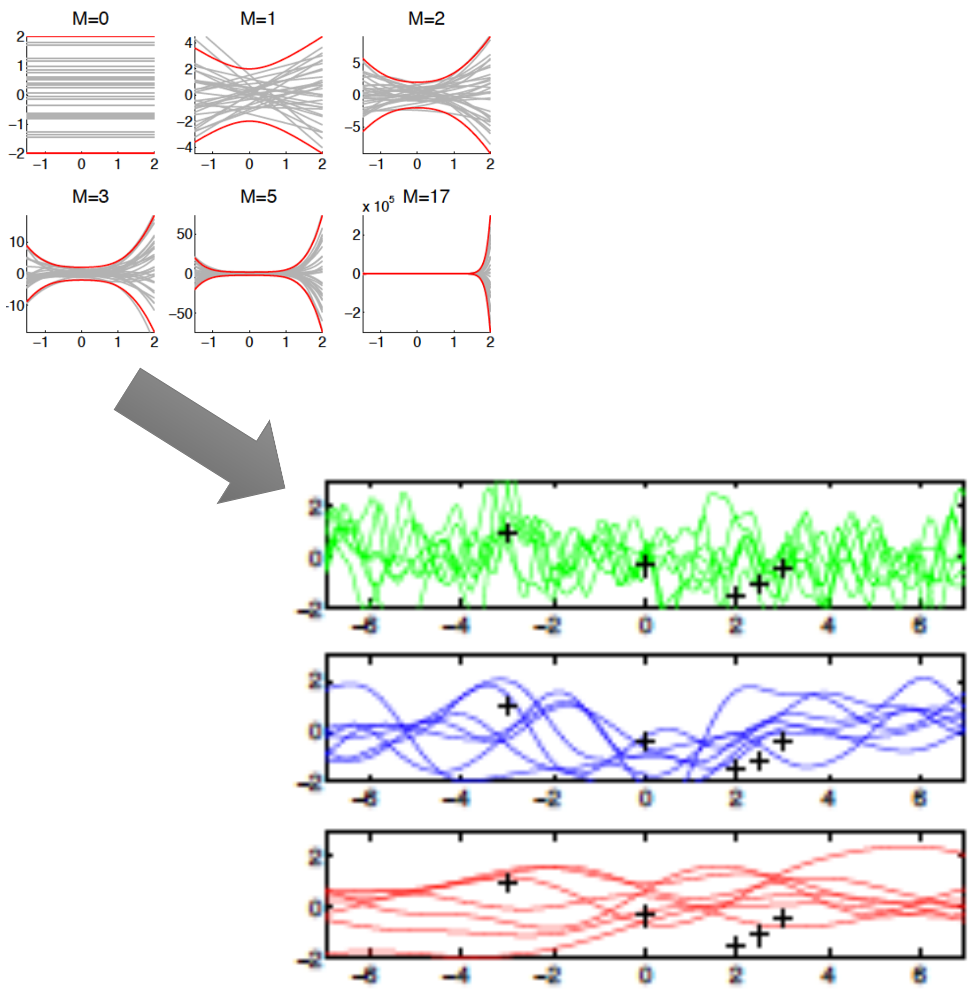




Priors on functions

Models with priors on the weights *indirectly* specify priors over functions.

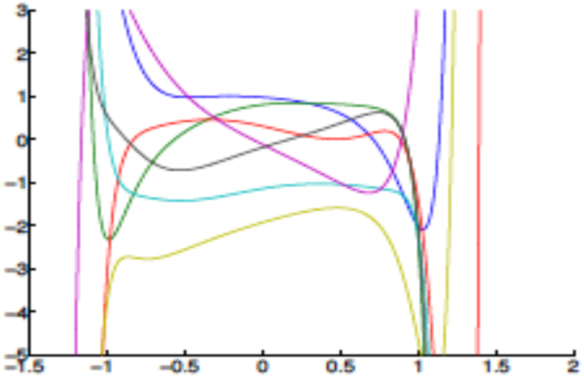
- **What about specifying priors on functions directly?**
- **What does a probability density over functions even look like?**



Prior on parameters

The Bayes rule can be written as:

$$p(f|y) = \frac{p(y|f)p(f)}{p(y)}$$



Samples from the prior

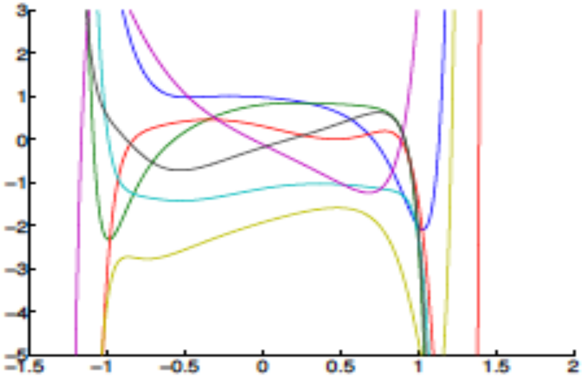
Prior on parameters

The Bayes rule can be written as:

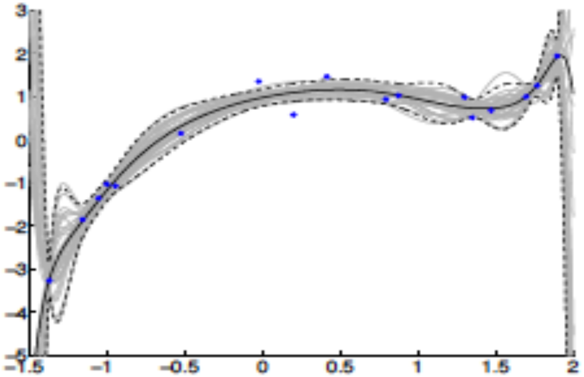
$$p(f|y) = \frac{p(y|f)p(f)}{p(y)}$$

We keep the functions which are “closer” to the data

→ Notion of **closeness** is given by the likelihood $p(y|f)$



Samples from the prior

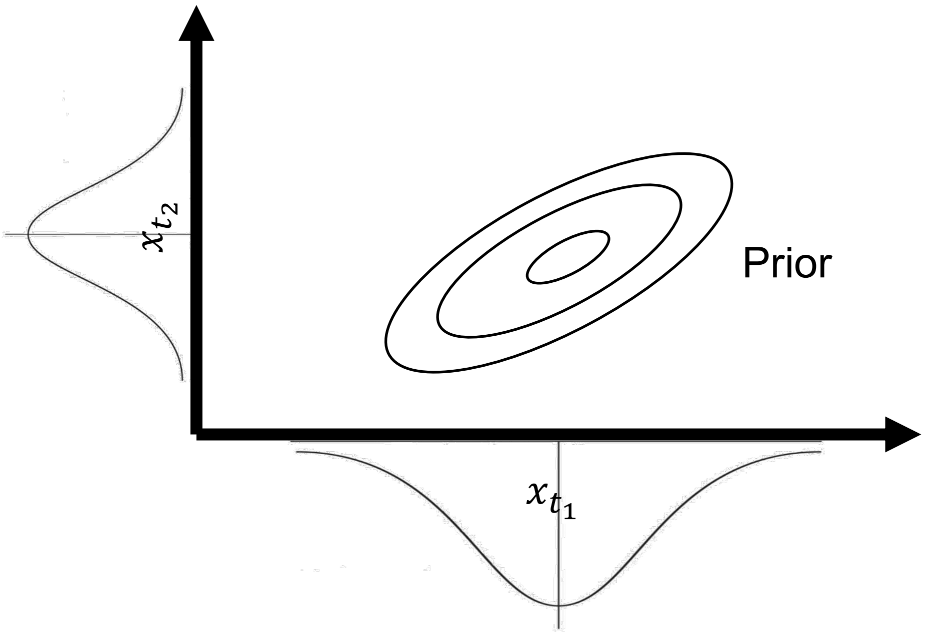


Samples from the posterior

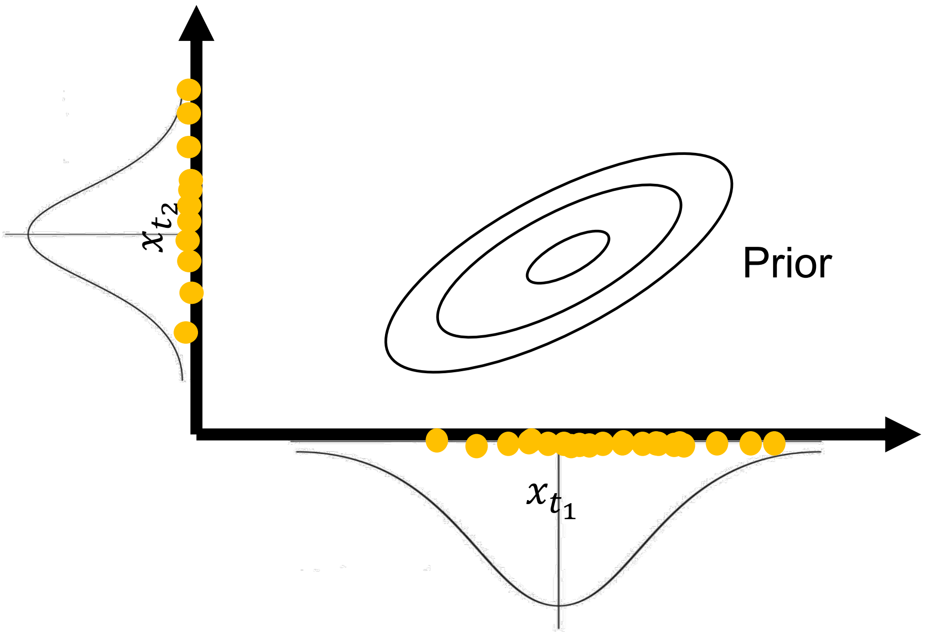
Gaussian Process Regression

Gaussian Processes (GP)

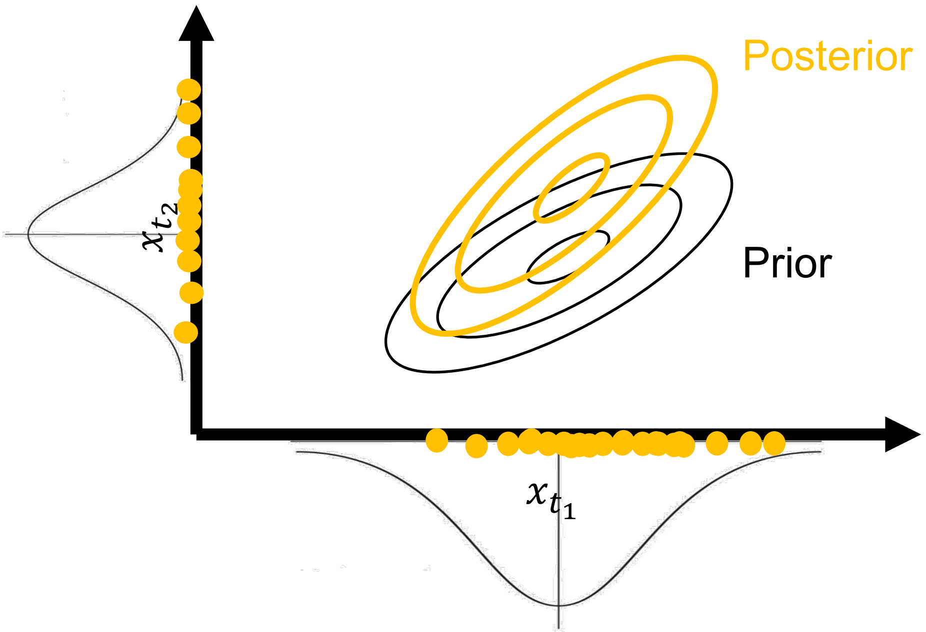
Towards Gaussian Processes



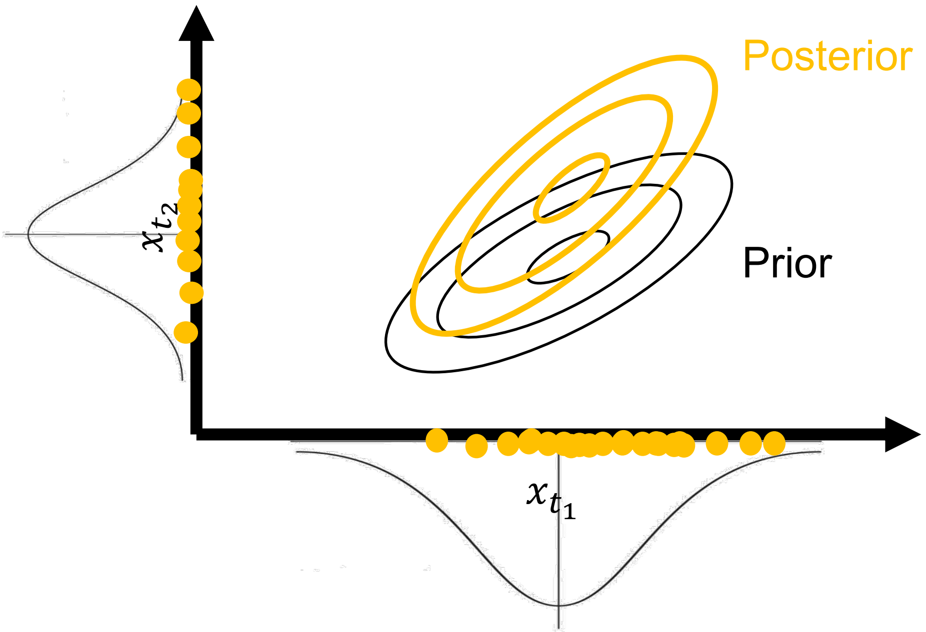
Towards Gaussian Processes



Towards Gaussian Processes

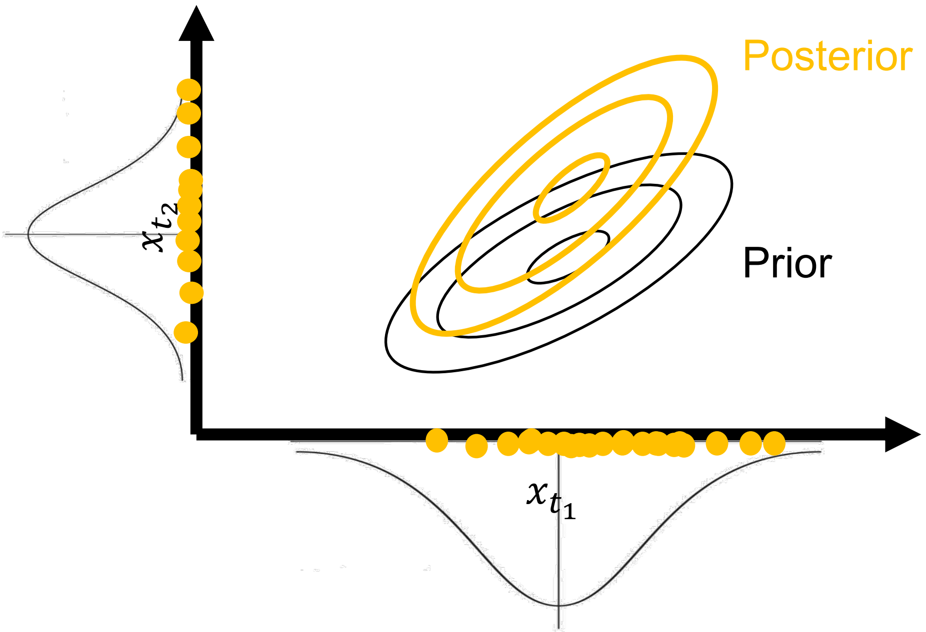


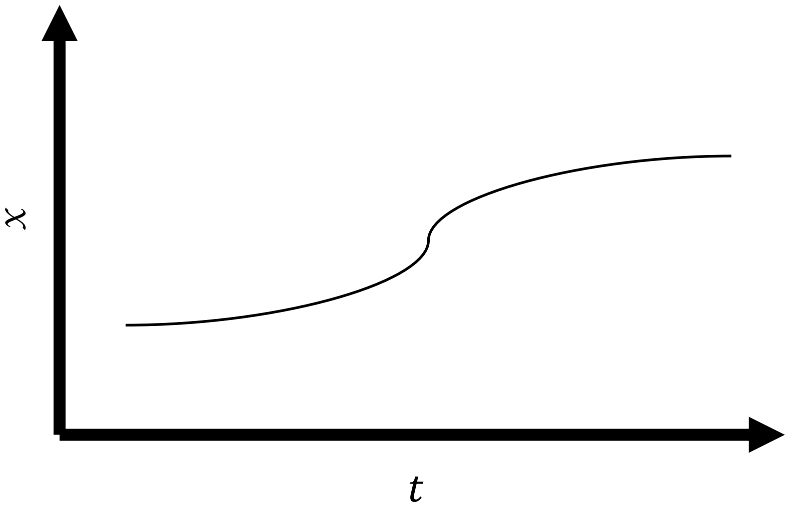
Towards Gaussian Processes



For **multivariate Gaussian distributions** we look at groups of real-valued variables.

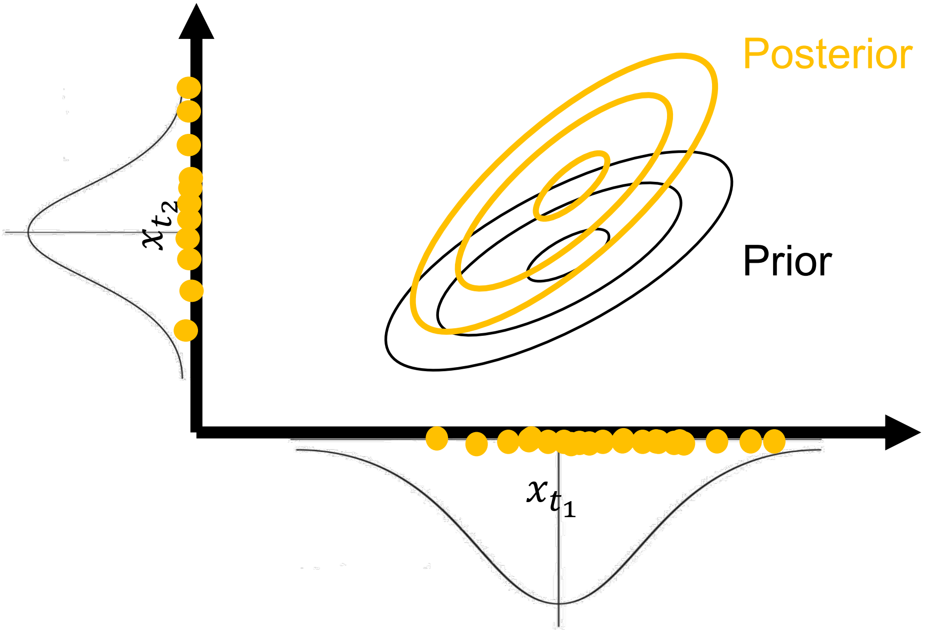
Towards Gaussian Processes

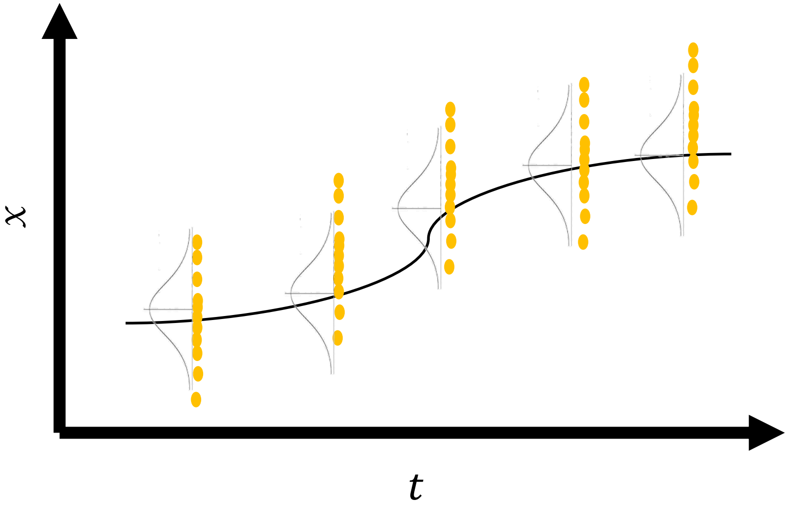




For **multivariate Gaussian distributions** we look at groups of real-valued variables.

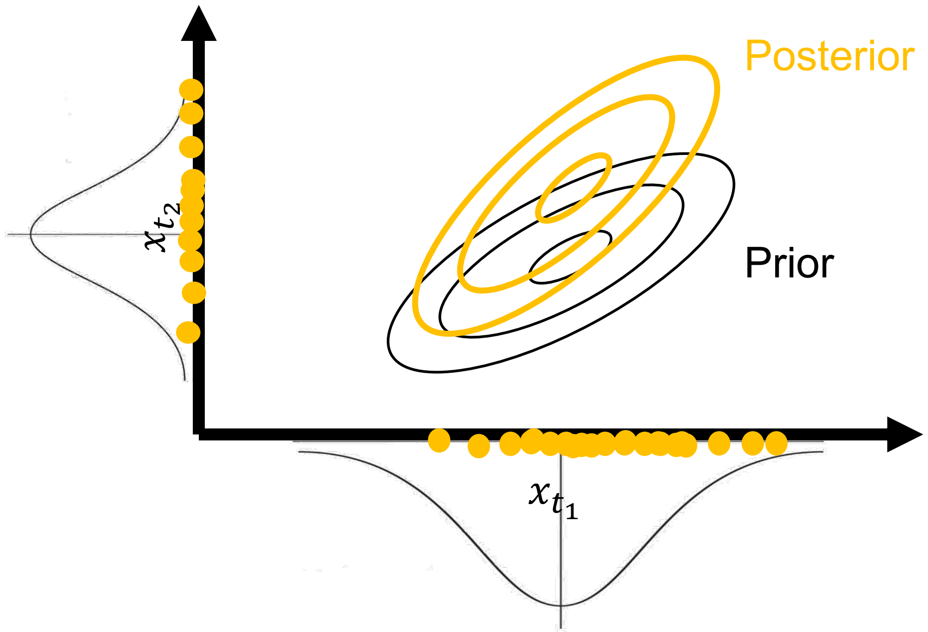
Towards Gaussian Processes

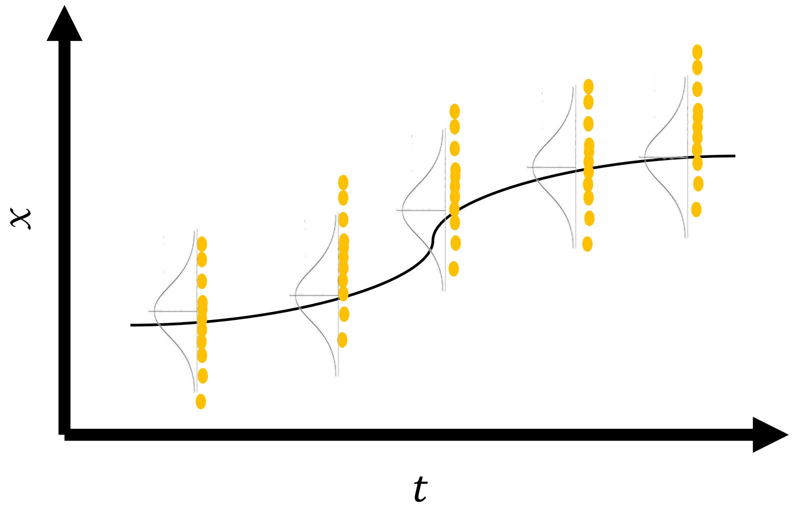




For **multivariate Gaussian distributions** we look at groups of real-valued variables.

Towards Gaussian Processes





For **multivariate Gaussian distributions** we look at groups of real-valued variables.

For **Gaussian processes** we look at very many random variables with Gaussian distribution.

→ GP are functions of (potentially infinite) number of real-valued variables.

Gaussian Process (GP)

<u>Definition:</u>

A function $\boldsymbol{f}$ is a Gaussian process if $\boldsymbol{f(t) = [f(t_1), \dots, f(t_N)]^T}$ has multivariate distribution for each $\boldsymbol{t = [t_1, \dots, t_N]^T}$.

For any subset of t: $f(t) \sim N(\mu(t), \Sigma(t, t'))$

Gaussian Process (GP)

Notice: here we use t for time, but in general we can have a $x \in \mathbb{R}^d$.

The **mean function** is defined as

$$\mu: \mathbb{R} \rightarrow \mathbb{R} \quad (\text{or, } \mathbb{R}^d \rightarrow \mathbb{R})$$

- Often, we subtract the mean from the data to have $\mu(t) = 0, \quad \forall\, t$

The **covariance function** is defined as $\Sigma:\ \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$; positive semidefinite matrix.

- Often for GP we refer to a **kernel function** $k(t, t')$.

Gaussian Process (GP)

Notice: here we use t for time, but in general we can have a $x \in \mathbb{R}^d$.

The **mean function** is defined as

$$\mu: \mathbb{R} \rightarrow \mathbb{R} \quad (\text{or, } \mathbb{R}^d \rightarrow \mathbb{R})$$

- Often, we subtract the mean from the data to have $\mu(t) = 0, \quad \forall\, t$

The **covariance function** is defined as Σ: $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$; positive semidefinite matrix.

- Often for GP we refer to a **kernel function** $k(t, t')$.

We can, then, rewrite the Gaussian process as:

- $\boldsymbol{f(t) \sim \mathcal{N}(\mu(t), k(t, t')}$ or $\boldsymbol{f(t) \sim \mathcal{N}(0, k(t, t'))}$

Gaussian Process (GP)

Notice: here we use t for time, but in general we can have a $x \in \mathbb{R}^d$.

The **mean function** is defined as

$$\boxed{\mu: \mathbb{R} \to \mathbb{R}} \text{ (or, } \mathbb{R}^d \to \mathbb{R})$$

- Often, we subtract the mean from the data to have $\mu(t) = 0, \quad \forall\, t$

The **covariance function** is defined as Σ: $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$; positive semidefinite matrix.

- Often for GP we refer to a **kernel function** $\boxed{k(t, t')}$

We can, then, rewrite the Gaussian process as:

- $\boldsymbol{f(t) \sim \mathcal{N}(\mu(t), k(t, t')}$ or $\boldsymbol{f(t) \sim \mathcal{N}(0, k(t, t'))}$

A GP is defined by its mean and kernel function, so we can write:

$$f \sim GP(\mu, k)$$

Gaussian process

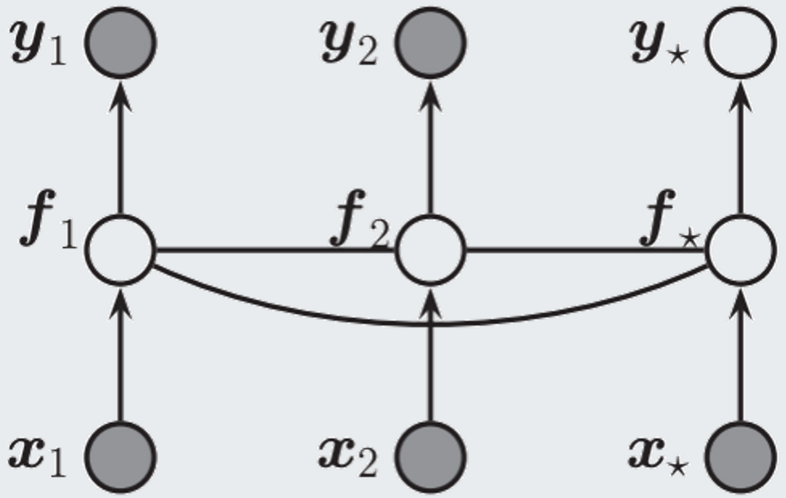
Graphical illustration

Then, a Gaussian Process assumes that the distribution over the function's on a finite (and arbitrary) set of points,

$$p\big(f(x_1), \ldots, f(x_N)\big),$$

is jointly Gaussian, with mean $\mu(x)$ and covariance $\sum(x)$, where the covariance is given by $\sum_{ij} = \kappa(x_i, x_j)$ and κ being a positive definite kernel function.

Key idea: if x_i and x_j are similar w.r.t. the kernel, their output through f will also be similar.



Gaussian process (graphical illustration)

$$p(y, f\,|x) = \mathcal{N}(0, \kappa(x)) \prod_i p(x_i, f_i)$$

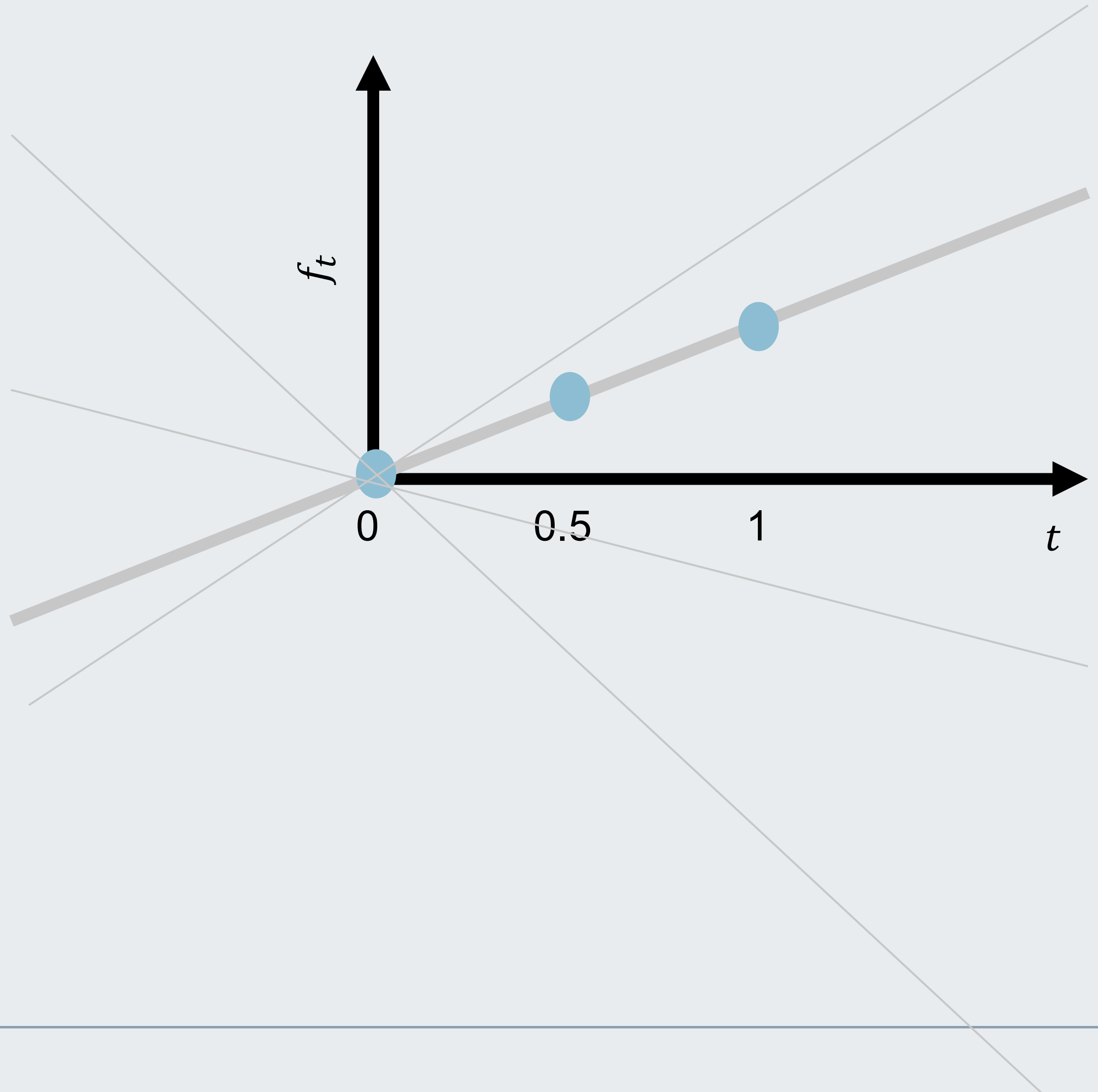
Example of Gaussian process

Random lines

Let's $t = \mathbb{R}$, be the entire real line. We define:

$$f_t = t \cdot w$$

with $w \sim \mathcal{N}(0,1)$, $w \in \mathbb{R}$.



We verify that f is a GP:

$$[f_{t_1}, \dots, f_{t_N}]^T =$$

$$= [wt_1, \dots, wt_N]^T =$$

$$= w[t_1, \dots, t_N]^T$$

→ Since $w \sim \mathcal{N}(0,1)$ is (multivariate) Gaussian, the result is also (multivariate) Gaussian.

Gaussian Process Regression

Gaussian Processes (GP) regression

Suppose:

$$\begin{bmatrix} f \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_f \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{ff} & \Sigma_{fy} \\ \Sigma_{fy}^T & \Sigma_{yy} \end{bmatrix}\right)$$

Then,

$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

- $\mu_{f|y} = \mu_f + \Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{f|y} = \Sigma_{ff} - \Sigma_{fy}\,\Sigma_{yy}^{-1}\Sigma_{fy}^T$

Gaussian processes

Conditioning and Inference

Suppose:

$$\begin{bmatrix} f \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_f \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{ff} & \Sigma_{fy} \\ \Sigma_{fy}^T & \Sigma_{yy} \end{bmatrix}\right)$$

Then,

$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

- $\mu_{f|y} = \mu_f + \boxed{\Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)}$
- $\Sigma_{f|y} = \Sigma_{ff} - \Sigma_{fy}\,\Sigma_{yy}^{-1}\Sigma_{fy}^T$

The mean "update" is linear function of the observation y

Gaussian processes

Conditioning and Inference

Suppose:

$$\begin{bmatrix} f \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_f \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{ff} & \Sigma_{fy} \\ \Sigma_{fy}^T & \Sigma_{yy} \end{bmatrix} \right)$$

Then,

$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

- $\mu_{f|y} = \mu_f + \boxed{\Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)}$
- $\Sigma_{f|y} = \Sigma_{ff} - \boxed{\Sigma_{fy}\, \Sigma_{yy}^{-1}\Sigma_{fy}^T}$

The mean “update” is linear function of the observation y

How much does the data explain?

Small → it can approach 0 → Uncertain ~ Σ_{ff}

Large → it can approach Σ_{ff} → zero covariance!

Gaussian processes

Prediction

Suppose:

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \mu_{y^*} \end{bmatrix}, \begin{bmatrix} \Sigma_{yy} & \Sigma_{y^*y} \\ \Sigma_{y^*y}^T & \Sigma_{y^*y^*} \end{bmatrix}\right)$$

Then,

$$p(y^*|y) = \mathcal{N}(\mu_{y^*|y}, \Sigma_{y^*|y})$$

where:

- $\mu_{y^*|y} = \mu_{y^*} + \Sigma_{y^*y}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{y^*|y} = \Sigma_{y^*y^*} - \Sigma_{y^*y}\,\Sigma_{yy}^{-1}\Sigma_{y^*y}^T$

Suppose:

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \mu_{y^*} \end{bmatrix}, \begin{bmatrix} \Sigma_{yy} & \Sigma_{y^*y} \\ \Sigma_{y^*y}^T & \Sigma_{y^*y^*} \end{bmatrix} \right)$$

Then,

$$p(y^*|y) = \mathcal{N}(\mu_{y^*|y}, \Sigma_{y^*|y})$$

where:

- $\mu_{y^*|y} = \boxed{\mu_{y^*} + \Sigma_{y^*y}\Sigma_{yy}^{-1}(y - \mu_y)}$
- $\Sigma_{y^*|y} = \boxed{\Sigma_{y^*y^*} - \Sigma_{y^*y}\,\Sigma_{yy}^{-1}\Sigma_{y^*y}^T}$

Predictive mean

Predictive covariance

Example for GP regression

Function to be estimated: $y(t) \; = \; t\sin(t)$

Sampling interval: $t \in [0, 10]$

Remember the conditioning:

$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

- $\mu_{f|y} = \mu_f + \Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{f|y} = \Sigma_{ff} - \Sigma_{fy}\,\Sigma_{yy}^{-1}\Sigma_{fy}^T$

Example for GP regression

Function to be estimated: $y(t) \; = \; t \sin(t)$

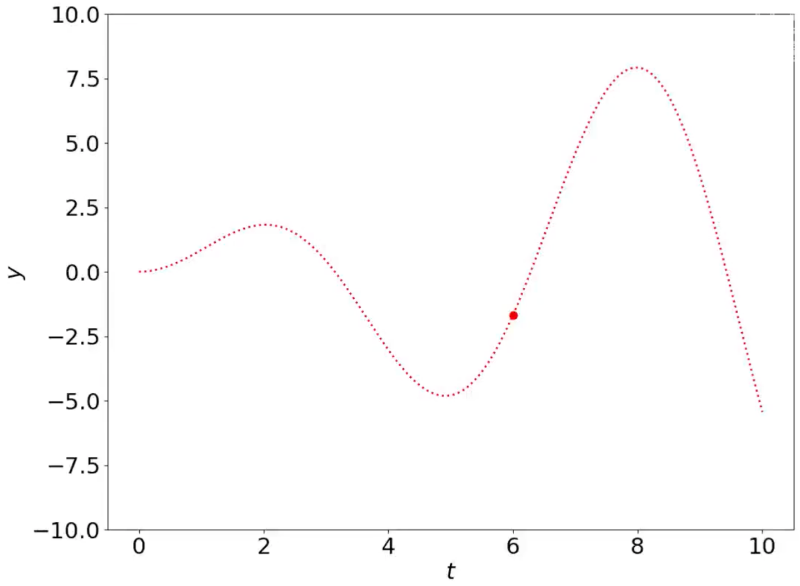
Sampling interval: $t \in [0, 10]$

Remember the conditioning:

$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

- $\mu_{f|y} = \mu_f + \Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{f|y} = \Sigma_{ff} - \Sigma_{fy}\,\Sigma_{yy}^{-1}\Sigma_{fy}^T$



Now, we can plug in particular values for y, e.g.,

$$p(f|\mathbf{y(t=6)}) = \mathcal{N}(\mu_{f|\mathbf{y(t=6)}}, \Sigma_{f|\mathbf{y(t=6)}})$$

Example for GP regression

Function to be estimated: $y(t) \; = \; t\sin(t)$

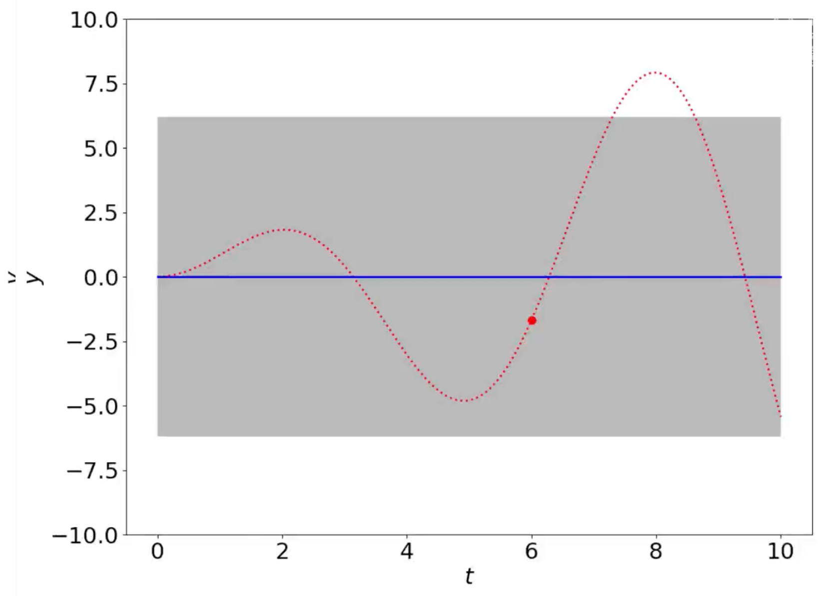
Sampling interval: $t \in [0, 10]$

Remember the conditioning:

$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

- $\mu_{f|y} = \mu_f + \Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{f|y} = \Sigma_{ff} - \Sigma_{fy}\,\Sigma_{yy}^{-1}\Sigma_{fy}^T$



Now, we can plug in particular values for y, e.g.,

$$p(f|\mathbf{y(t=6)}) = \mathcal{N}(\mu_{f|\mathbf{y(t=6)}}, \Sigma_{f|\mathbf{y(t=6)}})$$

Example for GP regression

Function to be estimated: $y(t) \; = \; t \sin(t)$

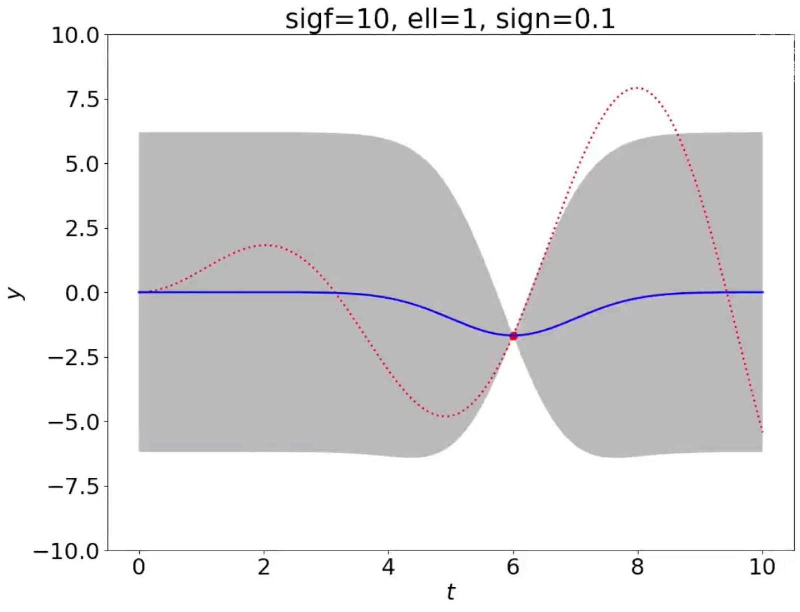
Sampling interval: $t \in [0, 10]$

Remember the conditioning:

$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

- $\mu_{f|y} = \mu_f + \Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{f|y} = \Sigma_{ff} - \Sigma_{fy}\,\Sigma_{yy}^{-1}\Sigma_{fy}^T$



Now, we can plug in particular values for y, e.g.,

$$p(f|\boldsymbol{y(t=6)}) = \mathcal{N}(\mu_{f|\boldsymbol{y(t=6)}}, \Sigma_{f|\boldsymbol{y(t=6)}})$$

Example for GP regression

Machine Learning Data Analytics

Function to be estimated: $y(t) \;=\; t\sin(t)$

Sampling interval: $t \in [0, 10]$

Remember the conditioning:

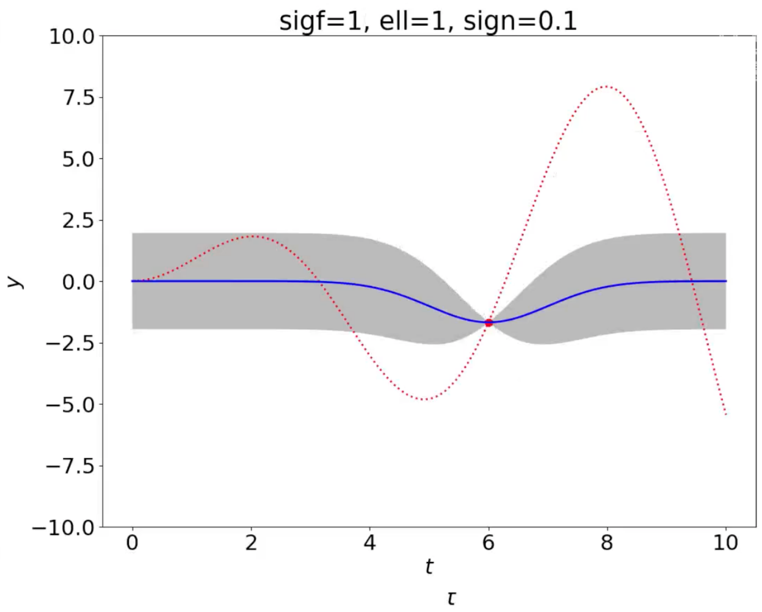
$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

- $\mu_{f|y} = \mu_f + \Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{f|y} = \Sigma_{ff} - \Sigma_{fy}\,\Sigma_{yy}^{-1}\Sigma_{fy}^{T}$

Now, we can plug in particular values for y, e.g.,

$$p(f|\boldsymbol{y(t=6)}) = \mathcal{N}(\mu_{f|\boldsymbol{y(t=6)}}, \Sigma_{f|\boldsymbol{y(t=6)}})$$



Example for GP regression

Function to be estimated: $y(t) \;=\; t\,\sin(t)$

Sampling interval: $t \in [0, 10]$

Remember the conditioning:

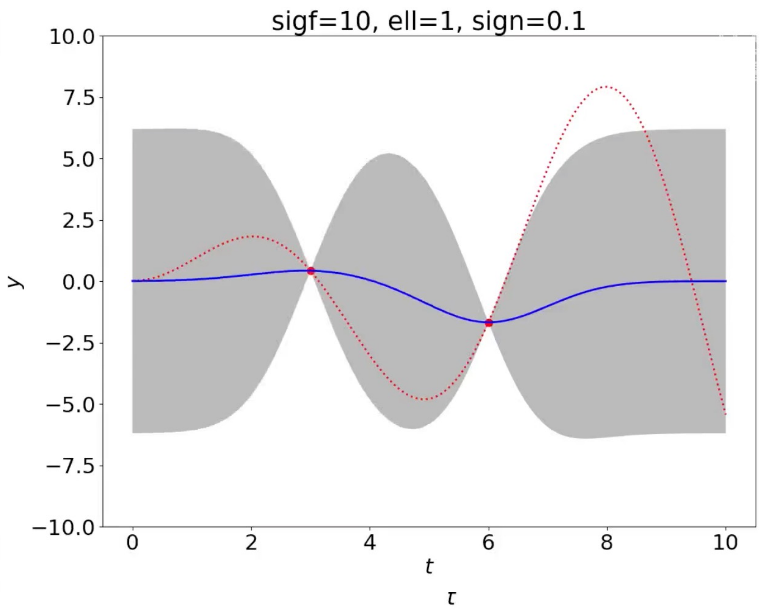
$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

- $\mu_{f|y} = \mu_f + \Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{f|y} = \Sigma_{ff} - \Sigma_{fy}\,\Sigma_{yy}^{-1}\Sigma_{fy}^{T}$

Now, we can plug in particular values for y, e.g.,

$$p(f|\boldsymbol{y(t=6)}) = \mathcal{N}(\mu_{f|\boldsymbol{y(t=6)}}, \Sigma_{f|\boldsymbol{y(t=6)}})$$



Example for GP regression

Function to be estimated: $y(t) \; = \; t\,\sin(t)$

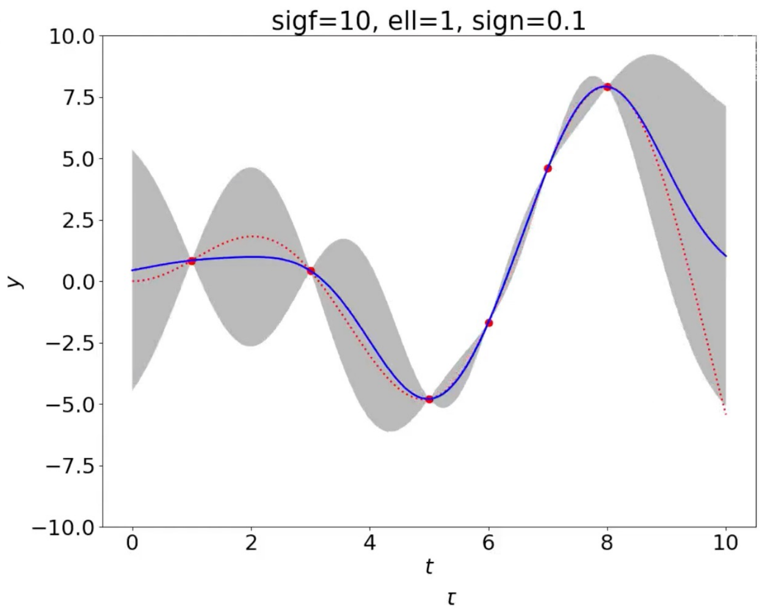
Sampling interval: $t \in [0, 10]$

Remember the conditioning:

$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

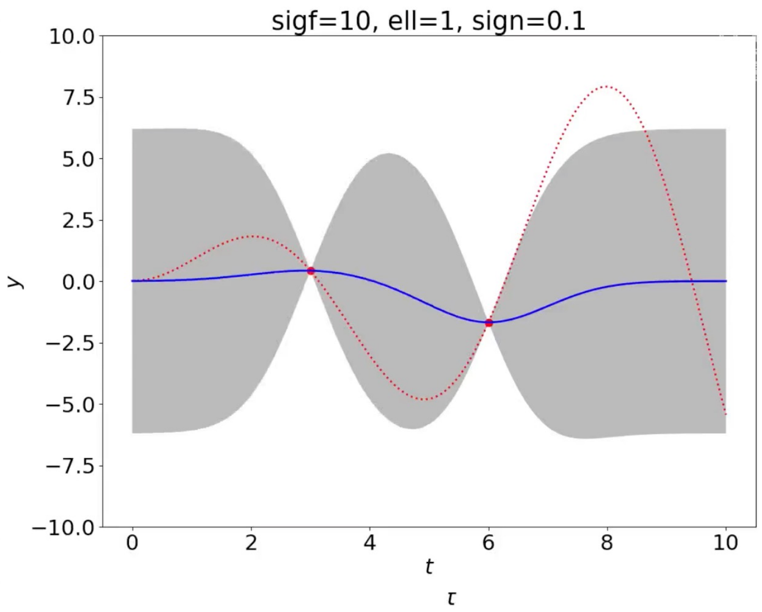
- $\mu_{f|y} = \mu_f + \Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{f|y} = \Sigma_{ff} - \Sigma_{fy}\,\Sigma_{yy}^{-1}\Sigma_{fy}^T$



Now, we can plug in particular values for y, e.g.,

$$p(f|\boldsymbol{y(t=6)}) = \mathcal{N}(\mu_{f|\boldsymbol{y(t=6)}}, \Sigma_{f|\boldsymbol{y(t=6)}})$$

Example for GP regression



Example for GP regression

Remember the conditioning:

$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

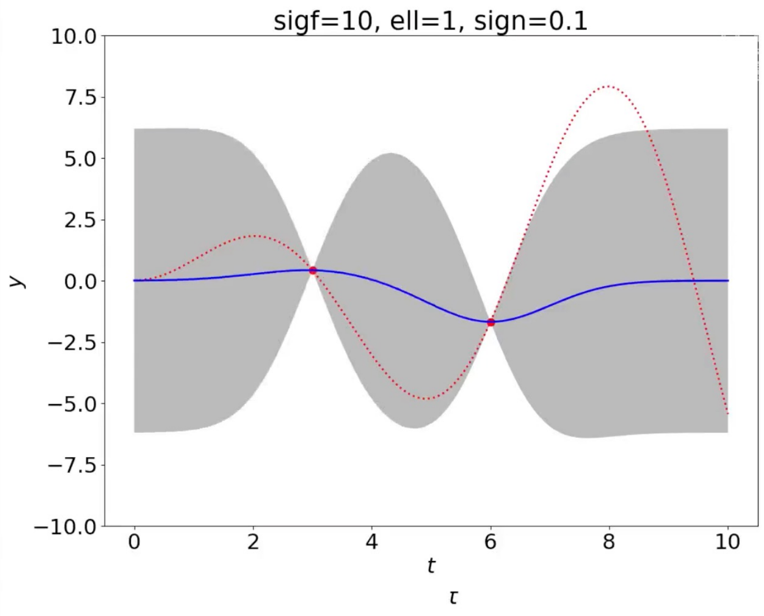
- $\mu_{f|y} = \mu_f + \Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{f|y} = \Sigma_{ff} - \Sigma_{fy}\,\Sigma_{yy}^{-1}\Sigma_{fy}^T$

Remember the prediction:

$$p(y^*|y) = \mathcal{N}(\mu_{y^*|y}, \Sigma_{y^*|y})$$

where:

- $\mu_{y^*|y} = \mu_{y^*} + \Sigma_{y^*y}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{y^*|y} = \Sigma_{y^*y^*} - \Sigma_{y^*y}\,\Sigma_{yy}^{-1}\Sigma_{y^*y}^T$



Example for GP regression

Remember the conditioning:

$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

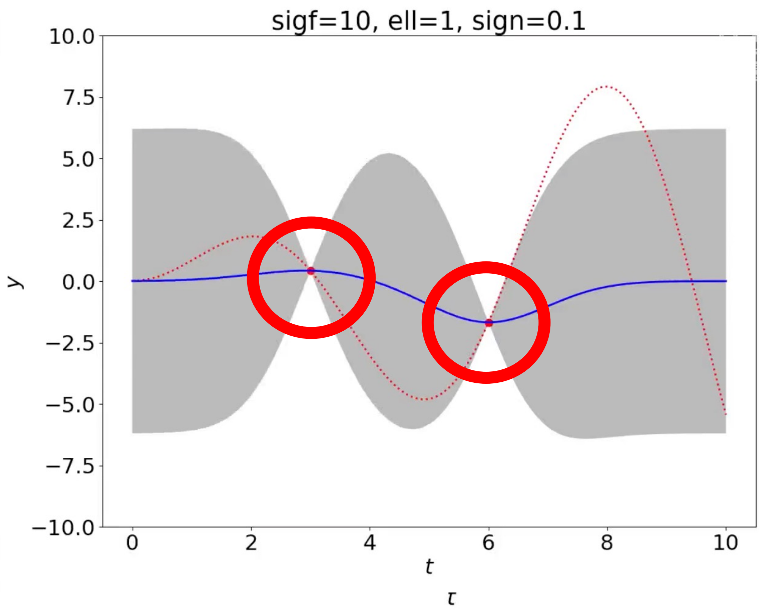
- $\mu_{f|y} = \mu_f + \Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{f|y} = \Sigma_{ff} - \Sigma_{fy}\,\Sigma_{yy}^{-1}\Sigma_{fy}^T$

Remember the prediction:

$$p(y^*|y) = \mathcal{N}(\mu_{y^*|y}, \Sigma_{y^*|y})$$

where:

- $\mu_{y^*|y} = \mu_{y^*} + \Sigma_{y^*y}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{y^*|y} = \Sigma_{y^*y^*} - \Sigma_{y^*y}\,\Sigma_{yy}^{-1}\Sigma_{y^*y}^T$



Test the prediction on $y = \begin{bmatrix} t = 3 \\ t = 6 \end{bmatrix}$

Example for GP regression

Remember the conditioning:

$$p(f|y) = \mathcal{N}(\mu_{f|y}, \Sigma_{f|y})$$

where:

- $\mu_{f|y} = \mu_f + \Sigma_{fy}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{f|y} = \Sigma_{ff} - \Sigma_{fy}\Sigma_{yy}^{-1}\Sigma_{fy}^T$

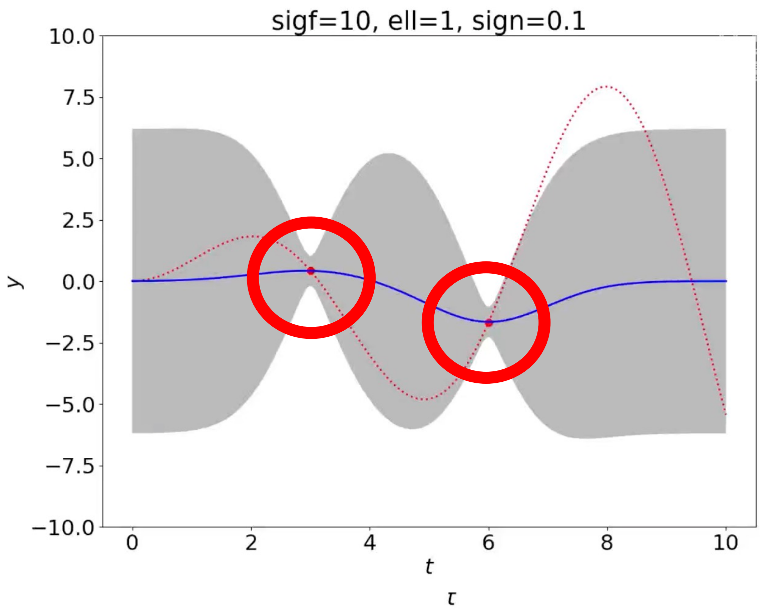
Remember the prediction:

$$p(y^*|y) = \mathcal{N}(\mu_{y^*|y}, \Sigma_{y^*|y})$$

where:

- $\mu_{y^*|y} = \mu_{y^*} + \Sigma_{y^*y}\Sigma_{yy}^{-1}(y - \mu_y)$
- $\Sigma_{y^*|y} = \Sigma_{y^*y^*} - \Sigma_{y^*y}\Sigma_{yy}^{-1}\Sigma_{y^*y}^T$

Test the prediction on $y = \begin{bmatrix} t = 3 \\ t = 6 \end{bmatrix}$



Lecture title

Recap

Recap

- Prior on functions
  - Recap of Gaussian distributions
  - Prior on parameters (indirect prior on functions)
- Gaussian processes
  - Multivariate Gaussian vs. GP
  - Definition
  - Example
- Gaussian process regression
  - Conditioning and inference
  - Prediction
  - Example