Bayesian Learning - Lecture 12

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Overview

- **Multivariate normal**
- **Gaussian process regression**

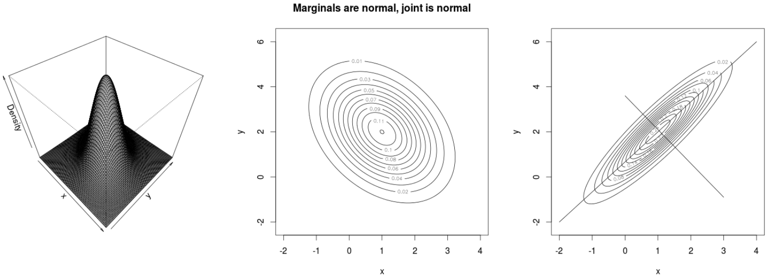
Multivariate normal

- **Multivariate normal**

$$\mathbf{x} \sim N_p(\mu, \Sigma)$$

where $\mu = (\mu_1, ..., \mu_p)'$ and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1p}\sigma_1\sigma_p \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & & \rho_{2p}\sigma_2\sigma_2 \\ \vdots & & \ddots & \\ \rho_{p1}\sigma_p\sigma_1 & \rho_{p2}\sigma_p\sigma_2 & \cdots & \sigma_p^2 \end{pmatrix}$$



Multivariate normal - some properties

- Let $\mathbf{x} \sim N_p(\mu, \Sigma)$.
- Let $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$ where $\mathbf{x}_1$ is $p_1 \times 1$ and $\mathbf{x}_2$ is $p_2 \times 1$ $(p_1 + p_2 = p)$.
- Let $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

- **Marginal distributions** are normal

$$\mathbf{x}_1 \sim N_{p_1}(\mu_1, \Sigma_1)$$

- **Conditional distributions** are normal

$$\mathbf{x}_1 | \mathbf{x}_2 \sim N_{p_1}\left[\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right]$$

Non-parametric regression

- **Linear regression**

$$y = \beta \cdot x + \varepsilon$$

  where $\varepsilon \sim N(0, \sigma^2)$ and iid over observations.

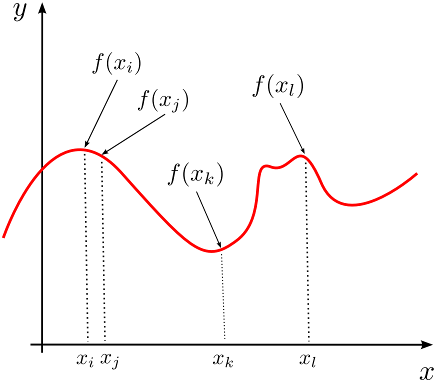
- **Nonlinear regression**

$$y = f(x) + \varepsilon$$

  where $f(\cdot)$ is some nonlinear function (ex $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$).

- **Non-parametric regression**: avoiding a parametric form for $f(\cdot)$.
- How do we put a **prior over a set of functions**?
- Restrict attention to a grid of (ordered) x-values: $x_1, x_2, .., x_k$.
- We can now put a joint prior on the k function values: $f(x_1), f(x_2), ..., f(x_k)$.

NONPARAMETRIC = ONE PARAMETER FOR EVERY X!



Gaussian process regression

- We clearly need to impose **smoothness**.
- Multivariate normal (Gaussian) prior:

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{pmatrix} \sim N(\mathbf{m}, \mathbf{K})$$

- But how do we specify the $k \times k$ **covariance matrix** $\mathbf{K}$?

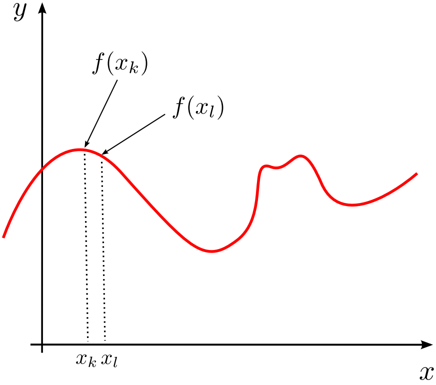
$$Cov\left(f(x_p), f(x_q)\right)$$

- **Squared exponential covariance function**

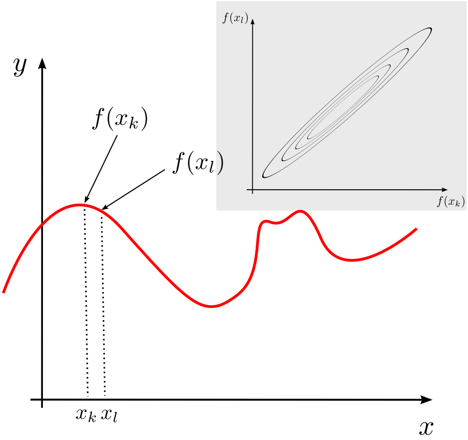
$$Cov\left(f(x_p), f(x_q)\right) = K(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\frac{x_p - x_q}{\ell}\right)^2\right)$$

- The covariance between $f(x_p)$ and $f(x_q)$ is a function of x_p and x_q.
- Nearby x's have highly correlated function ordinates $f(x)$.

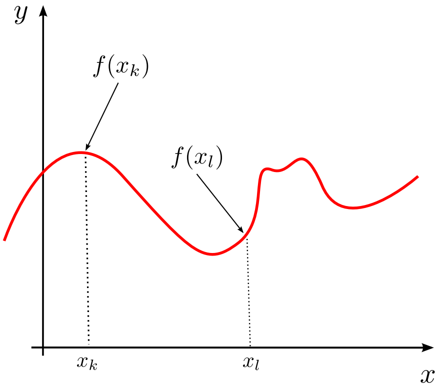
Smooth function - points nearby



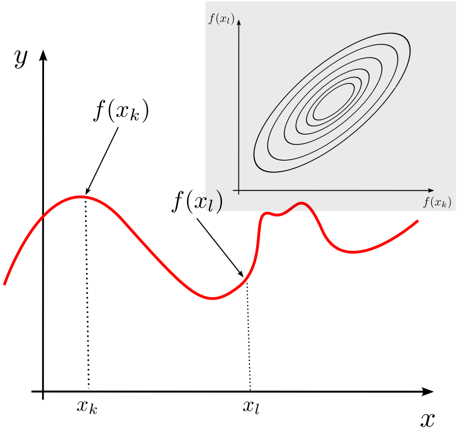
Smooth function - points nearby



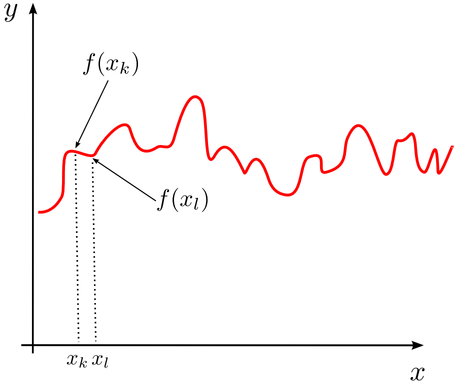
Smooth function - points far apart



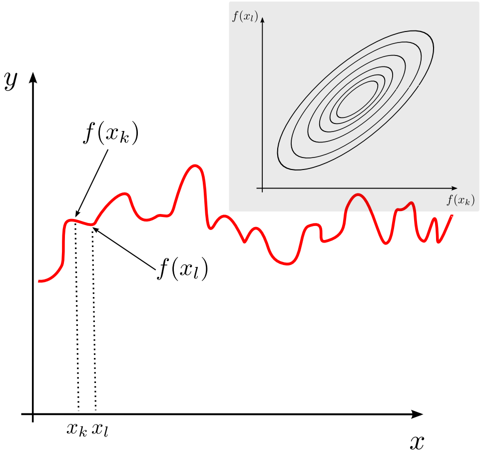
Smooth function - points far apart



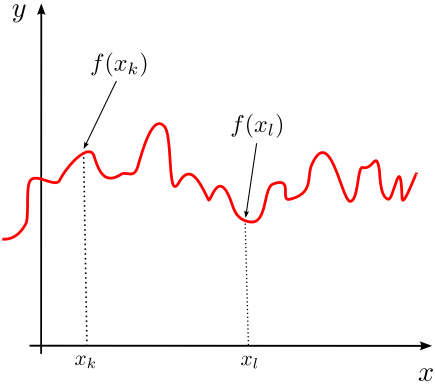
Jagged function - points nearby



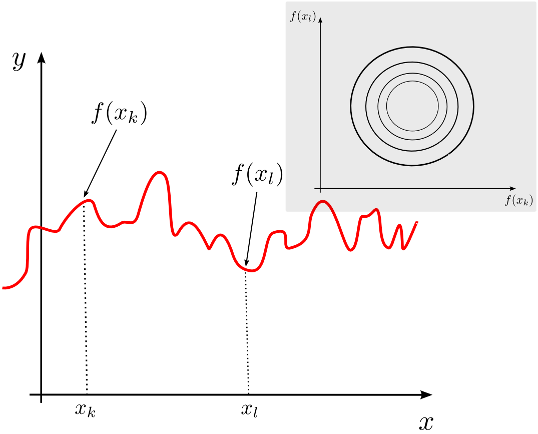
Jagged function - points nearby



Jagged function - points far apart



Jagged function - points far apart



Gaussian process regression, cont.

Definition

A **Gaussian process** (**GP**) is a collection of random variables, any finite number of which have a multivariate Gaussian distribution.

- A Gaussian process is a **probability distribution over functions**.
- A GP is completely specified by a **mean** and a **covariance function**

$$m(x) = \mathrm{E}\left[f(x)\right]$$

$$K(x, x') = E\left[\left(f(x) - m(x)\right)\left(f(x') - m(x')\right)\right]$$

  for any two inputs x and x' (note: this is *not* the transpose here).

- A **Gaussian process** (prior) is denoted by

$$f(x) \sim GP\left(m(x), K(x, x')\right)$$

Gaussian process regression, cont.

- Example:

$$m(x) = \sin(x)$$
$$K(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\frac{x_p - x_q}{\ell}\right)^2\right)$$

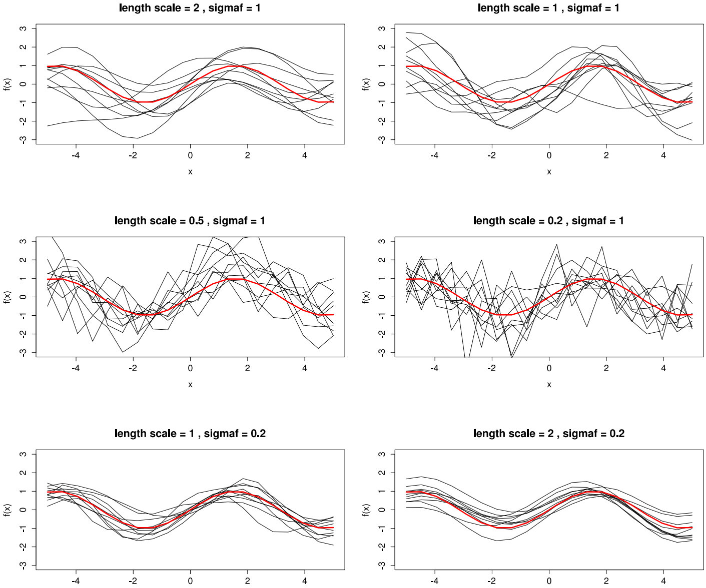
  where $\ell > 0$ is the length scale.

- Larger ℓ gives more smoothness in $f(x)$.

- **Simulate** a draw from $f(x) \sim GP(m(x), K(x, x'))$ over any grid $x_* = (x_1, ..., x_n)$ by using that

$$f(x_*) \sim N(m(x_*), K(x_*, x_*))$$

SIMULATING A GP - SINE MEAN AND SE KERNEL



Gaussian process regression, cont.

- **Model**

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon \overset{iid}{\sim} N(0, \sigma^2)$$

- **Prior**

$$f(x) \sim GP\left(0, K(x, x')\right)$$

- You have observed the data: $\mathbf{x} = (x_1, ..., x_n)'$ and $\mathbf{y} = (y_1, ..., y_n)'$.
- Goal: the posterior of $f(\cdot)$ over a grid of x-values:

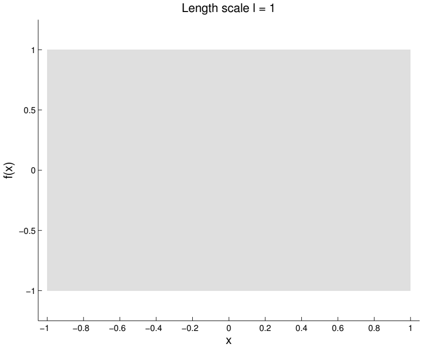
$$\mathbf{f}_* = \mathbf{f}(\mathbf{x}_*) = (f(x_{1*}), f(x_{2*}), ..., f(x_{m*}))$$

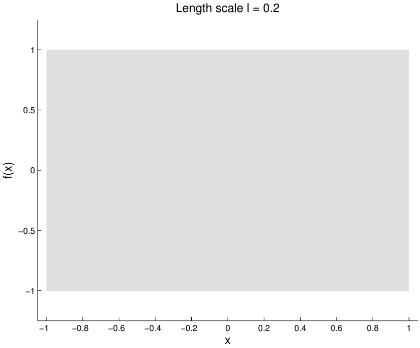
- The **posterior**

$$\mathbf{f}_* | \mathbf{x}, \mathbf{y}, \mathbf{x}_* \sim N\left(\bar{\mathbf{f}}_*, \mathrm{cov}(\mathbf{f}_*)\right)$$

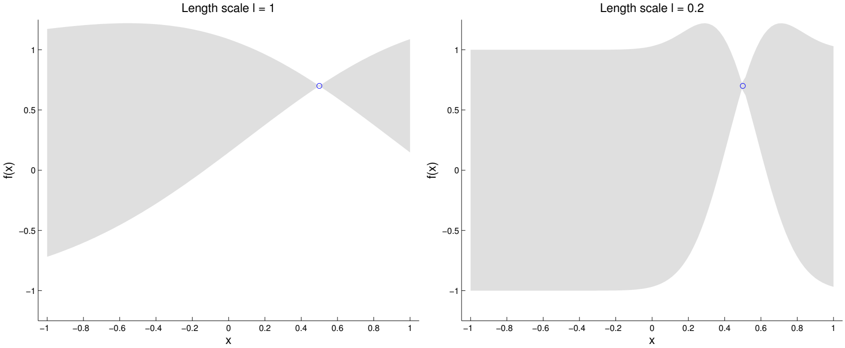
$$\bar{\mathbf{f}}_* = K(\mathbf{x}_*, \mathbf{x})\left[K(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right]^{-1} \mathbf{y}$$
$$\mathrm{cov}(\mathbf{f}_*) = K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, \mathbf{x})\left[K(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right]^{-1} K(\mathbf{x}, \mathbf{x}_*)$$

Learning a noise-free Gaussian process

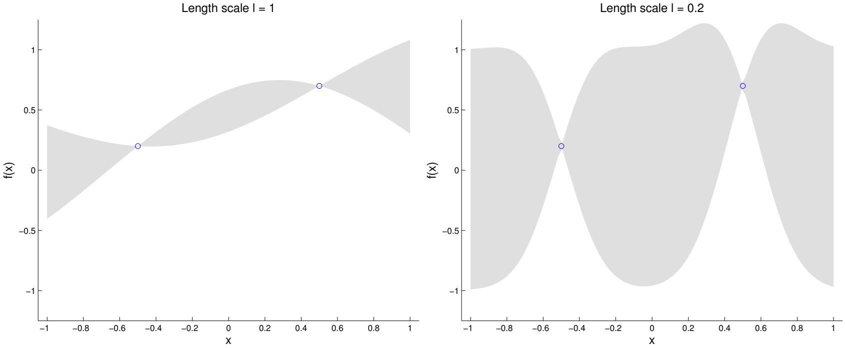




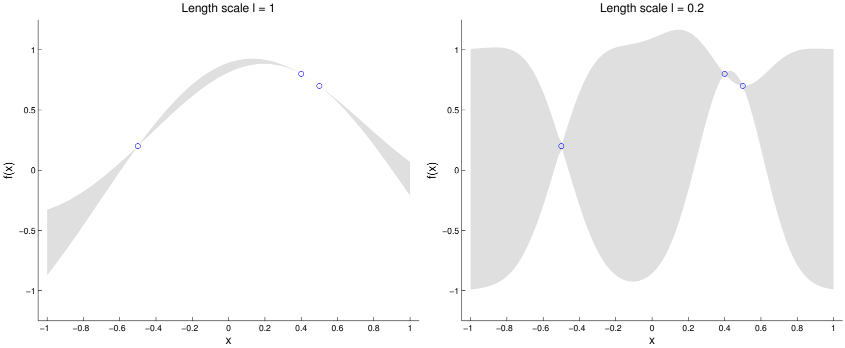
Learning a noise-free Gaussian process



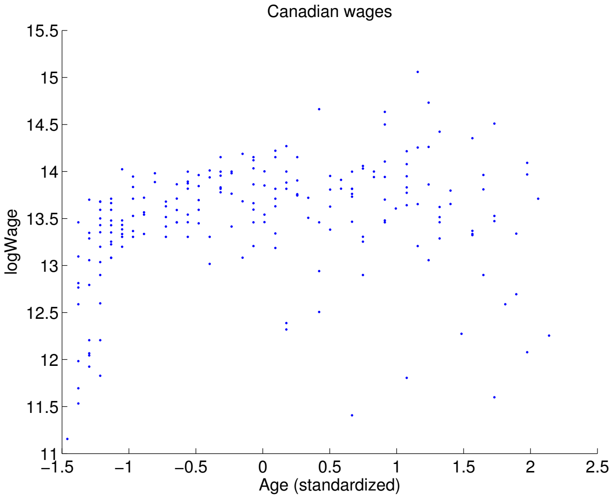
Learning a noise-free Gaussian process



Learning a noise-free Gaussian process



Example - Canadian wages



POSTERIOR OF F - $\ell = 0.2, 0.5, 1, 2$