ADVANCED MACHINE LEARNING
GAUSSIAN PROCESSES
LECTURE 1

Mattias Villani

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

LiU

TOPIC OVERVIEW

- Lecture 1
  - Recall: **The multivariate normal distribution**
  - Recall: Bayesian inference for **Gaussian linear**/**nonlinear regression**
  - Introduction to **Gaussian Process Regression**

- Lecture 2
  - **More on kernel functions**
  - Estimating the **GP hyperparameters**
  - **Large scale GPs**

- Lecture 3
  - **Gaussian Process Classification**
  - Some examples of **GP applications**

THE MULTIVARIATE NORMAL DISTRIBUTION

▶ The **density function** of a p-variate normal vector $\mathbf{x} \sim N(\mu, \Sigma)$ is

$$f(\mathbf{x}) = \left(\frac{1}{2\pi} \right)^{p/2} \frac{1}{\sqrt{\det \Sigma}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu) \right\}$$
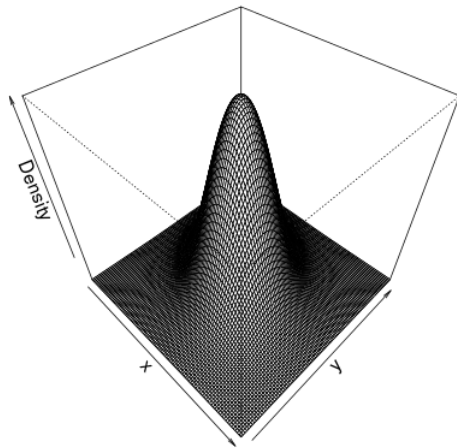
▶ Example: **Bivariate normal** $(p = 2)$

$$\Sigma = \left(\begin{array}{cc} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{array} \right)$$

▶ Mean and variance
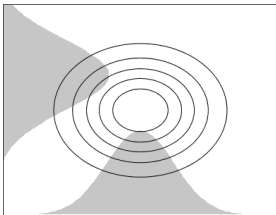
$$E(\mathbf{x}) = \mu \quad Var(\mathbf{x}) = \Sigma$$

Multivariate normal

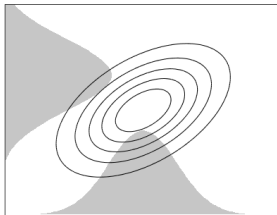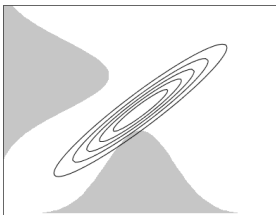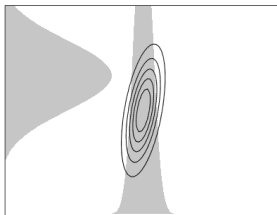# MULTIVARIATE NORMAL

$\rho = 0, \sigma_1 = 1, \sigma_2 = 1$

$\rho = 0.5, \sigma_1 = 1, \sigma_2 = 1$

$\rho = 0.95, \sigma_1 = 1, \sigma_2 = 1$

$\rho = 0.5, \sigma_1 = 1/4, \sigma_2 = 1$

THE MULTIVARIATE NORMAL DISTRIBUTION, CONT.

- Let $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ where x_1 is $p_1 \times 1$ and x_2 is $p_2 \times 1$ $(p_1 + p_2 = p)$.

- Partition μ and Σ accordingly as

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

- **Marginals are normal**. Let $x \sim N(\mu, \Sigma)$, then

$$x_1 \sim N(\mu_1, \Sigma_{11})$$

- **Conditionals are normal**. Let $x \sim N(\mu, \Sigma)$, then

$$x_1 | x_2 = x_2^* \sim N\left[\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2^* - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right]$$

NONLINEAR REGRESSION

▶ **Linear regression**[1]

$$y = f(\mathbf{x}) + \epsilon$$

$$f(\mathbf{x}) = \mathbf{w}^T \cdot \mathbf{x}$$

and $\epsilon \sim N(0, \sigma_n^2)$ and iid over observations.

▶ The weights $\mathbf{w}$ are called regression coefficients (β) in statistics.

▶ **Polynomial regression**: $\phi(\mathbf{x}) = (1, x, x^2, x^3, ..., x^k)$:

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) \cdot$$

▶ More generally: **splines** with **basis functions**.

▶ Polynomial and spline models are linear in $\mathbf{w}$. Least squares!

---

[1] I follow the notation in RW rather than PRML. In PRML: y is the noise-free response. $t = y + \epsilon$ is the response with noise. β^{-1} is the noise variance (σ_n^2).

BAYESIAN LINEAR REGRESSION - INFERENCE

▶ Linear regression for all n observations

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times p}{\mathbf{X}}\,\underset{p\times 1}{\mathbf{w}} + \underset{n\times 1}{\varepsilon}$$

▶ $\mathbf{w}$ is unknown. σ_n is assumed known.

▶ **Prior**

$$\mathbf{w} \sim N\left(0, \Sigma_p\right)$$

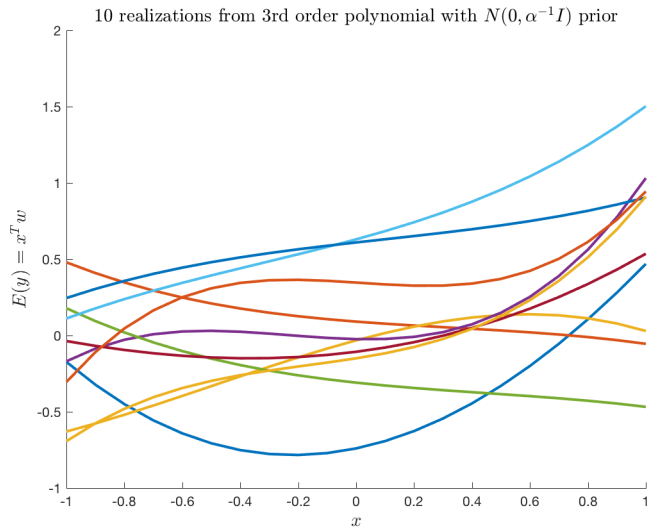▶ Common choice (Ridge regression): $\Sigma_p = \alpha^{-1}\mathbf{I}$.

▶ **Posterior**

$$\mathbf{w}|\mathbf{X},\mathbf{y} \sim N\left(\bar{\mathbf{w}}, \mathbf{A}^{-1}\right)$$
$$\mathbf{A} = \sigma_n^{-2}\mathbf{X}^T\mathbf{X} + \Sigma_p^{-1}$$
$$\bar{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X} + \sigma_n^2\Sigma_p^{-1}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

▶ Recall: **Posterior precision = Data Precision + Prior Precision**.

A PRIOR ON **w** IS A PRIOR ON FUNCTIONS

10 realizations from 3rd order polynomial with $N(0, \alpha^{-1}I)$ prior

NON-PARAMETRIC REGRESSION

- **Non-parametric regression**: avoiding a parametric form for $f(\cdot)$. Treat $f(\mathbf{x})$ as an unknown parameter for every $\mathbf{x}$.

- **Weight space** view
  - Restrict attention to a grid of (ordered) x-values: $x_1, x_2, .., x_k$.
  - Put a joint prior on the k function values: $f(x_1), f(x_2), ..., f(x_k)$.

- **Function space** view
  - Treat f as an **unknown function**.
  - Put a **prior over a set of functions**.

NONPARAMETRIC = ONE PARAMETER FOR EVERY X!

SE kernel - $\ell = 1/2$
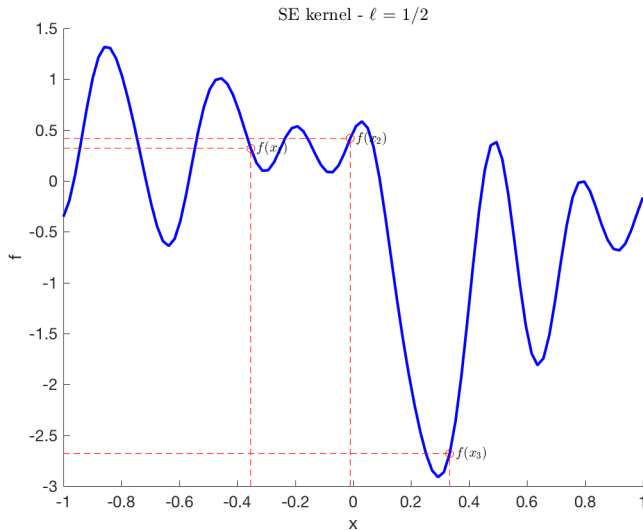
GAUSSIAN PROCESS REGRESSION

- Weight-space view. GP assumes

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

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

$$Cov(f(x_p), f(x_q))$$

- **Squared exponential** covariance function

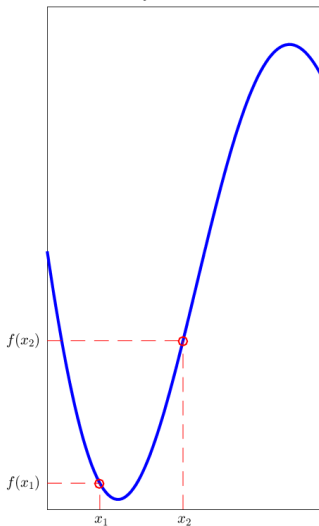$$Cov(f(x_p), f(x_q)) = k(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\frac{x_p - x_q}{\ell}\right)^2\right)$$

- Nearby x's have highly correlated function ordinates $f(x)$.
- We can compute $Cov(f(x_p), f(x_q))$ for *any* x_p and x_q.
- Extension to multiple covariates: $(x_p - x_q)^2$ replaced by $(\mathbf{x}_p - \mathbf{x}_q)^T(\mathbf{x}_p - \mathbf{x}_q)$.
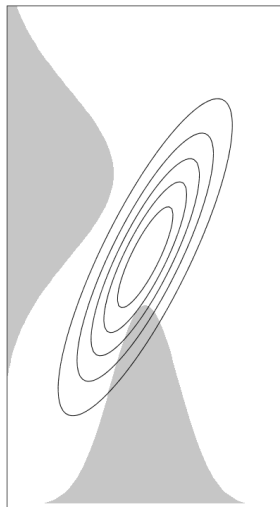
SMOOTH FUNCTION - POINTS NEARBY

$\sigma_f = 1, \ell = 1$

Correlation coefficient = 0.99

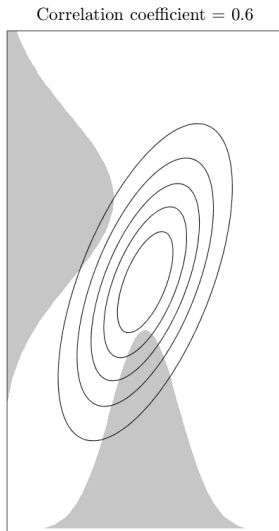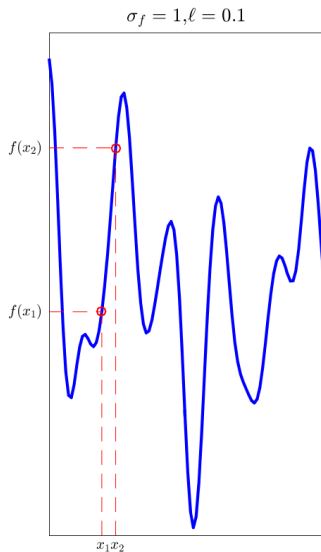$f(x_1)$

$f(x_2)$

$x_1 x_2$
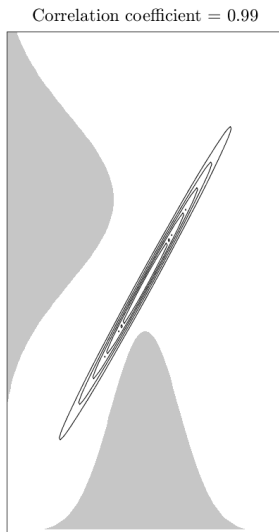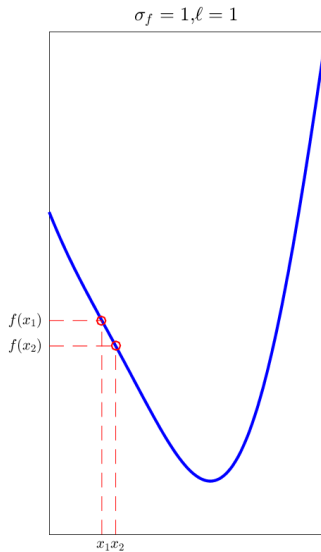
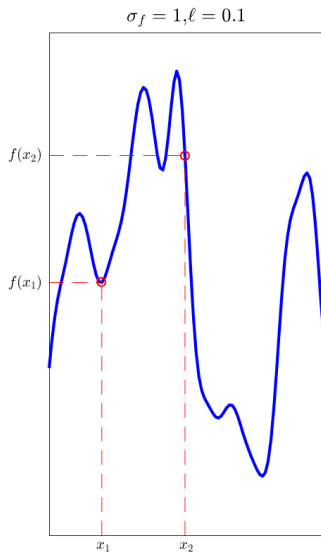# SMOOTH FUNCTION - POINTS FAR APART
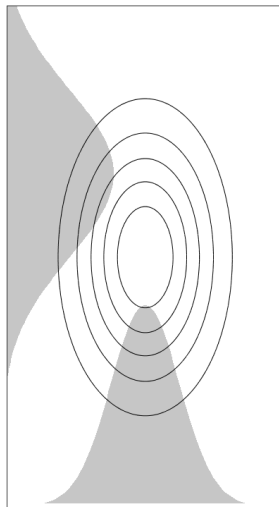

$\sigma_f = 1, \ell = 1$

Correlation coefficient = 0.83

JAGGED FUNCTION - POINTS NEARBY

$\sigma_f = 1, \ell = 0.1$

Correlation coefficient $= 0.6$

JAGGED FUNCTION - POINTS FAR APART

$\sigma_f = 1, \ell = 0.1$

Correlation coefficient = 1.1e-08

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 really a **probability distribution over functions** (curves).
- 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** is denoted by

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

- **Bayesian**: $f(x) \sim GP$ encodes **prior beliefs** about the unknown $f(\cdot)$.

SIMULATING A GP

- Example:

$$m(x) = \sin(10x)$$

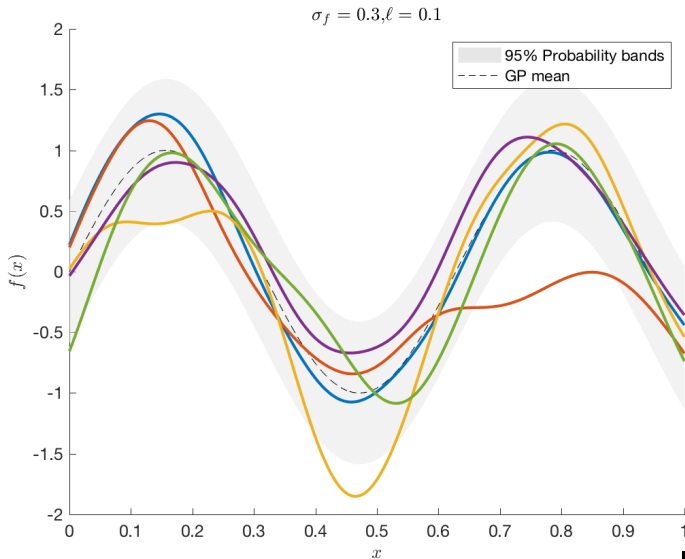$$k(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\frac{x - x'}{\ell}\right)^2\right)$$

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

- Larger ℓ gives more smoothness in $f(x)$.
- Simulate draw from $f(x) \sim GP\left(m(x), k(x, x')\right)$ over a grid $\mathbf{x}_* = (x_1, ..., x_n)$ by using that

$$f(\mathbf{x}_*) \sim N\left(m(\mathbf{x}_*), K(\mathbf{x}_*, \mathbf{x}_*)\right)$$

- Note that the **kernel** $k(x, x')$ produces a **covariance matrix** $K(\mathbf{x}_*, \mathbf{x}_*)$ when evaluated at the vector $\mathbf{x}_*$.

SIMULATING A GP

$\sigma_f = 0.3, \ell = 0.1$

THREE COMMONLY USED COVARIANCE KERNELS

- Let $r = \|x - x'\|$.
- **Squared exponential (SE)** $(\ell > 0, \sigma_f > 0)$

$$K_{SE}(r) = \sigma_f^2 \exp\left(-\frac{r^2}{2\ell^2}\right)$$

  - Infinitely mean square differentiable. Very smooth.
- **Rational Quadratic (RQ)** $(\ell > 0, \sigma_f > 0, \alpha > 0)$

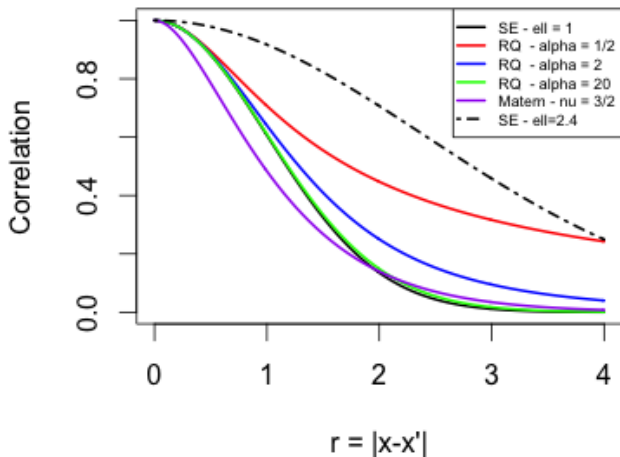$$K_{RQ}(r) = \sigma_f^2 \left(1 + \frac{r^2}{2\alpha\ell^2}\right)^{-\alpha}$$

  - RQ is sum of SE with different ℓ. When $\alpha \to \infty$, $K_{RQ}(r) \to K_{SE}(r)$.
- **Matérn** $(\ell > 0, \sigma_f > 0, \nu > 0)$

$$K_{Matern}(r) = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{\ell}\right)^\nu K_\nu \left(\frac{\sqrt{2\nu}r}{\ell}\right)$$
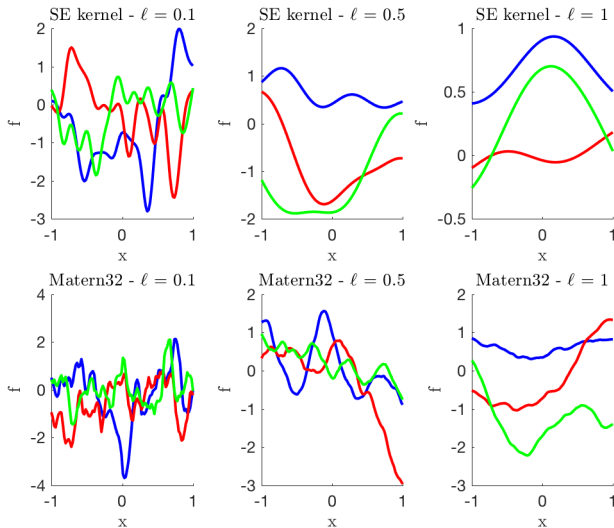
  - $\nu = 3/2$ and $\nu = 5/2$ common. As $\nu \to \infty$, $K_{Matern}(r) \to K_{SE}(r)$.
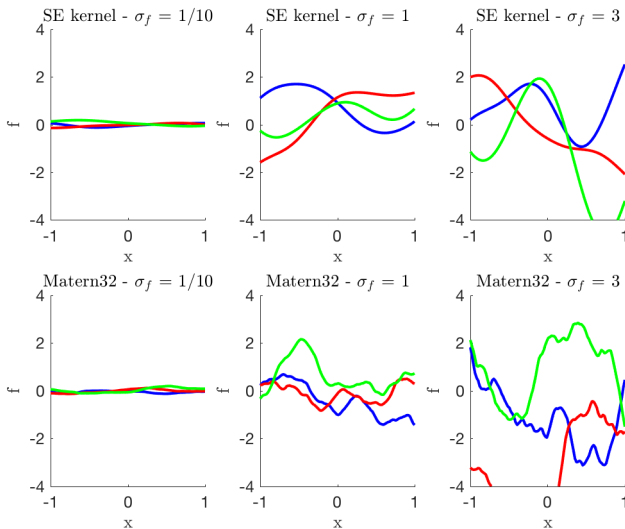
CORRELATION AS A FUNCTION OF DISTANCE

Correlation functions

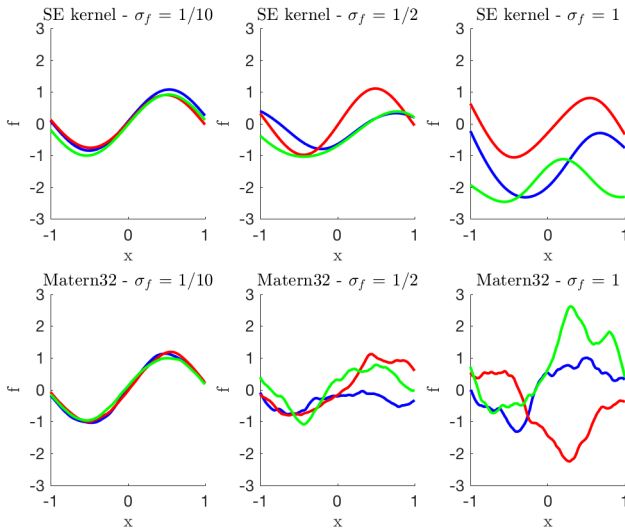# THE LENGTH SCALE ℓ DETERMINES THE SMOOTHNESS

THE SCALE FACTOR σ_f DETERMINES THE VARIANCE

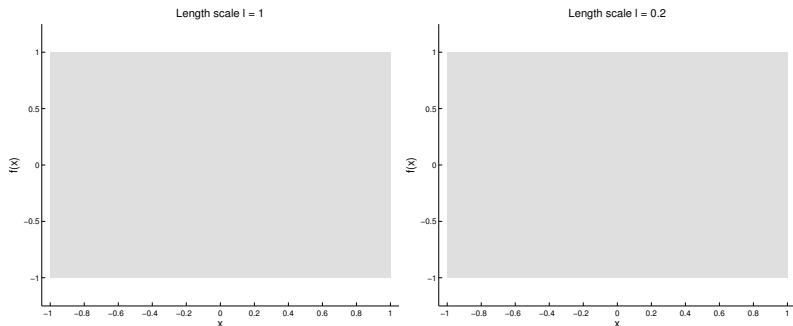# THE MEAN CAN BE *sin*(3*x*). OR WHATEVER.

SIMULATING A GP

▶ The joint way: Choose a grid $x_1, ..., x_k$. Simulate the k-vector

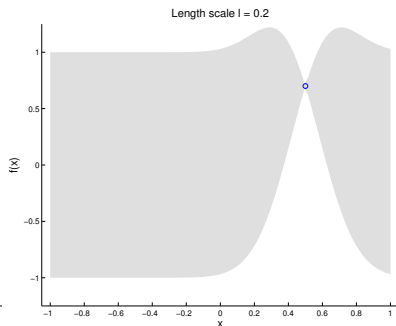$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{pmatrix} \sim N(\mathbf{m}, \mathbf{K})$$
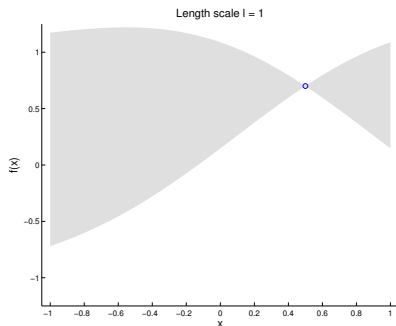
▶ More intuition from the conditional decomposition

$$p(f(x_1), f(x_2), ..., f(x_k)) = p(f(x_1)) \, p(f(x_2)|f(x_1)) \cdots$$
$$\times \, p(f(x_k)|f(x_1), ..., f(x_{k-1}))$$

SIMULATION FROM $\ell=1$ VS $\ell=0.2$. BEFORE FIRST DRAW.

Length scale l = 1
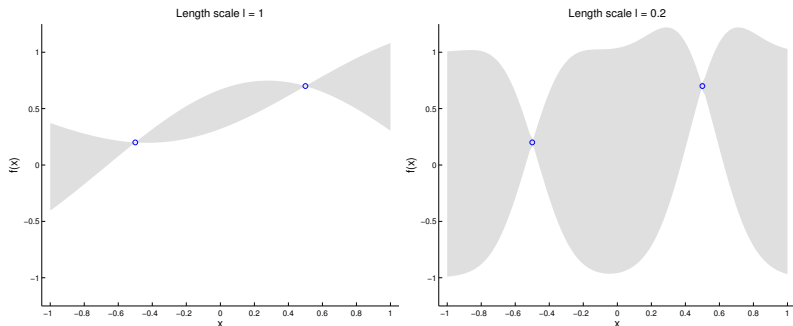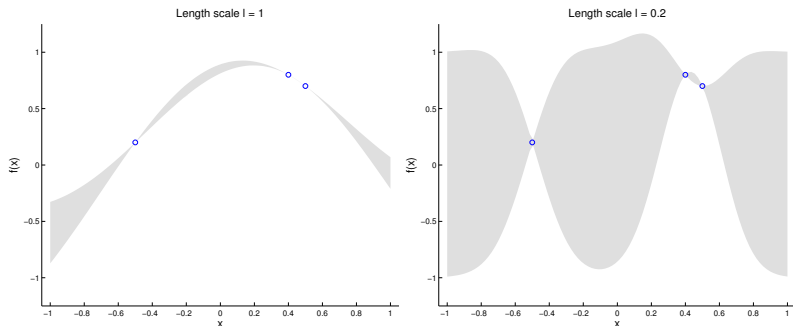
Length scale l = 0.2

SIMULATION FROM $\ell=1$ VS $\ell=0.2$. BEFORE SECOND DRAW.

SIMULATION FROM ℓ=1 VS ℓ=0.2. BEFORE THIRD DRAW.

SIMULATION FROM ℓ=1 VS ℓ=0.2. BEFORE FOURTH DRAW.

THE POSTERIOR FOR A GAUSSIAN PROCESS REGRESSION

- ▶ **Model**

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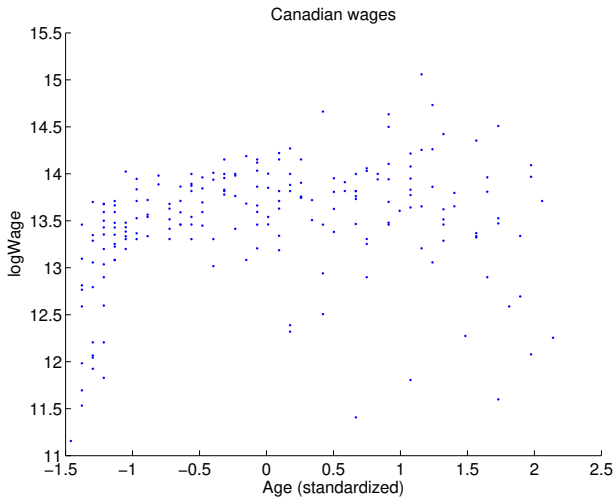
- ▶ **Prior**

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

- ▶ You have observed the data: $\mathbf{x} = (x_1, ..., x_n)^T$ and $\mathbf{y} = (y_1, ..., y_n)^T$.
- ▶ Goal: the posterior of $f(\cdot)$ over a grid of x-values: $\mathbf{f}_* = \mathbf{f}(\mathbf{x}_*)$.
- ▶ The **posterior** (use formula for conditional Gaussian above)

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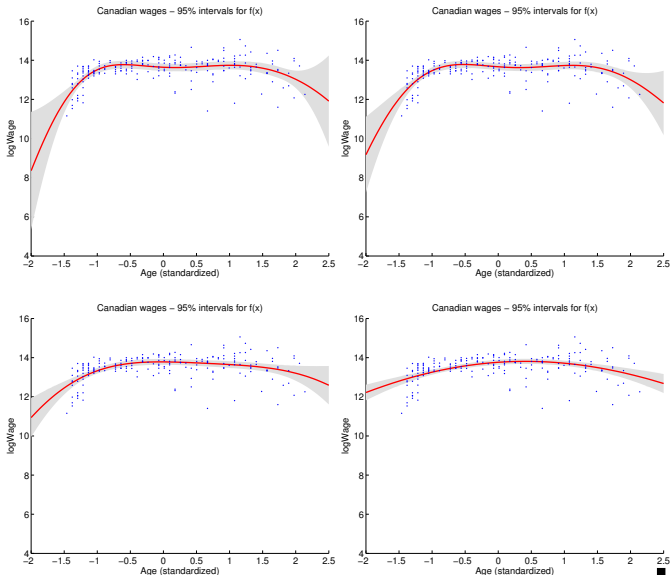
$$\bar{\mathbf{f}}_* = K(\mathbf{x}_*, \mathbf{x})\left[K(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right]^{-1} \mathbf{y}$$

$$\text{cov}(\mathbf{f}_*) = K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, \mathbf{x})\left[K(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right]^{-1} K(\mathbf{x}, \mathbf{x}_*)$$
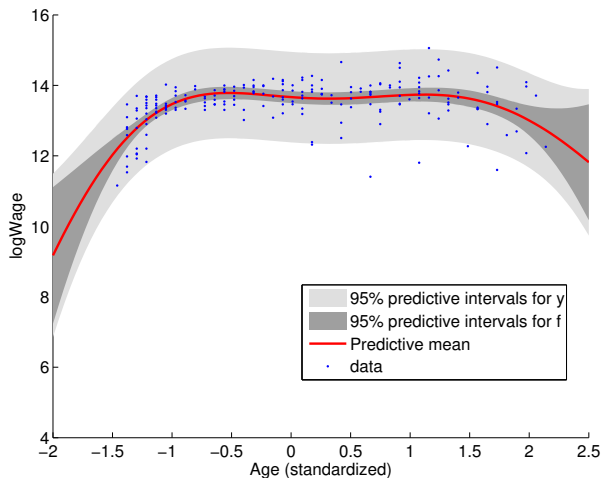
EXAMPLE - CANADIAN WAGES

Canadian wages

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

CANADIAN WAGES - PREDICTION WITH $\ell = 0.5$

SOFTWARE

- Python: GPy
- Matlab: Statistics and Machine Learning Toolbox, GPML, GPstuff.
- R: Kernlab,