Bayesian Learning - Lecture 6

Mattias Villani

Division of Statistics
Department of Computer and Information Science
Linköping University

Lecture overview

- Flexible nonlinear regression and splines
- Smoothness/shrinkage priors
- Gaussian process regression

Polynomial regression

- Recall the linear regression model with a single covariate

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

- Polynomial regression:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_k x_i^k + \varepsilon_i.$$

  This can be written as a linear regression

$$y = X_P \beta + \varepsilon,$$

  where

$$X_P = (1, x, x^2, ..., x^k).$$

- The posterior of β is obtained like any linear regression.

Splines

- A more local basis: **truncated polynomials**

$$b_{ij} = \begin{cases} (x_i - k_j)^p & \text{if } x_i > k_j \\ 0 \text{ otherwise} \end{cases}$$

- $k_1, k_2, ..., k_m$ are the **knots**.
- Splines are nonlinear in x, but linear in basis space

$$y = X_b \beta + \varepsilon,$$

where X_b is the basis regression matrix

$$X_b = (b_1, ..., b_m).$$

- Common extension

$$X_b = (1, x, b_1, ..., b_m).$$

- Still just linear in X_b. Linear regression fitting.

Smoothness prior for splines

- Problem: too many knots leads to **over-fitting**.
- Solution: **smoothness/shrinkage/regularization prior**

$$\beta_i|\sigma^2 \overset{iid}{\sim} N\left(0, \frac{\sigma^2}{\lambda}\right)$$

- Larger λ gives smoother fit. Note: here we have $\Omega_0 = \lambda I$.
- Equivalent to a penalized likelihood:

$$-2 \cdot LogPost \propto RSS(\beta) + \lambda\beta'\beta$$

- Posterior mean gives **ridge regression** estimator

$$\tilde{\beta} = \left(X'X + \lambda I\right)^{-1} X'y$$

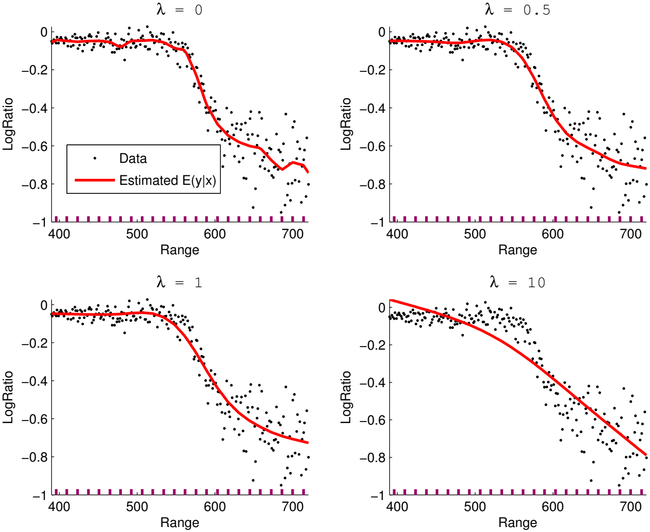
- **Shrinkage** toward zero

$$\text{As } \lambda \to \infty,\ \tilde{\beta} \to 0$$

- When $X'X = I$

$$\tilde{\beta} = \frac{1}{1+\lambda}\hat{\beta}_{OLS}$$

Bayesian spline with smoothness prior



Smoothness prior for splines, cont.

- The famous **Lasso** variable selection method is equivalent to using the posterior mode estimate under the prior:

$$\beta_i|\sigma^2 \overset{iid}{\sim} \text{Laplace}\left(0, \frac{\sigma^2}{\lambda}\right)$$

where the general Laplace density is

$$p(\beta_i) = \frac{1}{2b}\exp\left(-\frac{|\beta_i - \mu|}{b}\right)$$

- The Bayesian shrinkage prior is **interpretable**, **not ad hoc**.
- Laplace distribution have heavy tails.
- Laplace prior: many β_i close to zero, but some β_i may be very large.
- Normal distribution have light tails.
- Normal prior: most β_i are fairly equal in size, and no single β_i can be very much larger than the other ones.

Estimating the shrinkage

- How do we determine the degree of smoothness, λ? Cross-validation.
- Bayesian: I cannot specify $\lambda \Rightarrow \lambda$ is unknown $\Rightarrow$ use a prior for λ.
- One possibility: $\lambda \sim Inv-\chi^2(\eta_0, \lambda_0)$. The user specifies η_0 and λ_0.
- Alternative approach: specify the prior on the *degrees of freedom*.
- Hierarchical setup:

$$
\begin{aligned}
y|\beta, x &\sim N(x'\beta, \sigma^2) \\
\beta|\sigma^2 &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} \delta_0^{-1} I_q & 0 \\ 0 & \lambda^{-1} I_m \end{pmatrix}\right) \\
\sigma^2 &\sim Inv-\chi^2(\nu_0, \sigma_0^2) \\
\lambda &\sim Inv-\chi^2(\eta_0, \lambda_0)
\end{aligned}
$$

Note: different shrinkage on the original q covariates (δ_0) and the covariates that comes from the knots (λ).

Estimating the shrinkage, cont.

- Joint posterior

$$p(\beta, \sigma^2, \lambda | y, x) = p(\beta, \sigma^2 | \lambda, y, x) p(\lambda | y, x)$$

where

$$p(\lambda | y, x) = \int \int p(\beta, \sigma, \lambda | y, x) d\beta d\sigma^2$$

is the marginal posterior of λ.

- The conditional posterior $p(\beta, \sigma^2 | \lambda, y, x)$ is a special case linear regression with conjugate prior $\mu_0 = (0, 0)'$ and

$$\Omega_0 = \begin{pmatrix} \delta_0 I_q & 0 \\ 0 & \lambda I_m \end{pmatrix}$$

Summary of the posterior with normal shrinkage prior

- The joint posterior of β, σ^2 and λ is

$$\beta|\sigma^2,\lambda,y \sim N\left(\mu_n,\Omega_n^{-1}\right)$$
$$\sigma^2|\lambda,y \sim Inv-\chi^2\left(\nu_n,\sigma_n^2\right)$$
$$p(\lambda|y) \propto \sqrt{\frac{|\Omega_0|}{|X'X+\Omega_0|}}\left(\frac{\nu_n\sigma_n^2}{2}\right)^{-\nu_n/2}\cdot p(\lambda)$$

where $p(\lambda)$ is the prior for λ, and

$$\mu_n = \left(X'X+\Omega_0\right)^{-1}X'y$$
$$\Omega_n = X'X+\Omega_0$$
$$\nu_n = \nu_0+n$$
$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2+y'y-\mu_n'\Omega_n\mu_n$$

Bayesian variable selection and estimating knot locations

- Selecting among a set of fixed knots $k_1, ..., k_m$ is a variable selection problem. More on **Bayesian variable selection** in the last module.

- The **location of the knots** can be treated as **unknown**, and estimated from the data.

- The joint posterior of parameters and knot locations

$$p(\beta, \sigma^2, \lambda, \xi_1, ..., \xi_q | y, x)$$

  where ξ_i is the location of the ith knot.

- Posterior is complex but can be sampled from by Markov Chain Monte Carlo (MCMC). Li and Villani (2013, SJS).

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

- **Linear regression**

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

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

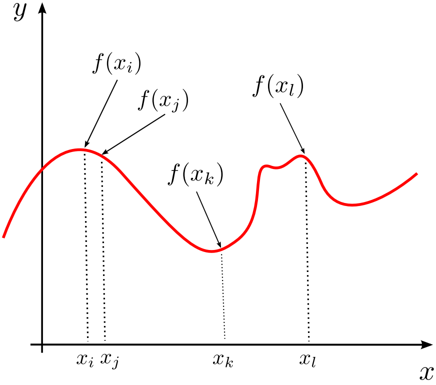
- **Nonlinear regression**

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

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!



Nonparametric regression - smooth interpolation

- Treat all n ordinates as unknown parameters:

$$f(x_i) = \gamma_i.$$

- Problem: too many parameters. Estimated curve wiggles way too much.
- Solution: use a (multivariate) prior on $\gamma = (\gamma_1, ..., \gamma_n)'$ that carries the info that the regression curve is smooth:

$$\text{if } x_i \text{ and } x_k \text{ are close then } \gamma_i \text{ is close to } \gamma_k$$

- Order the data with respect to covariates and assign the prior

$$p(\gamma_i|\gamma_{i-1}) \sim N\left(\gamma_{i-1}, \tau_0^2 \cdot |x_i - x_{i-1}|\right), \text{ for } i = 2, ..., n.$$

- The hyperparameter τ_0^2 controls the degree of prior smoothness.

Gaussian process regression

- Generalization of smooth interpolation. 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 K**?

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

Gaussian process regression

- Generalization of smooth interpolation. 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 K**?

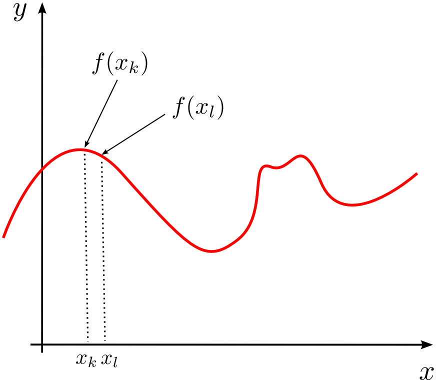
$$Cov\left(f(x_p), f(x_q)\right)$$

- **Squared exponential covariance function**

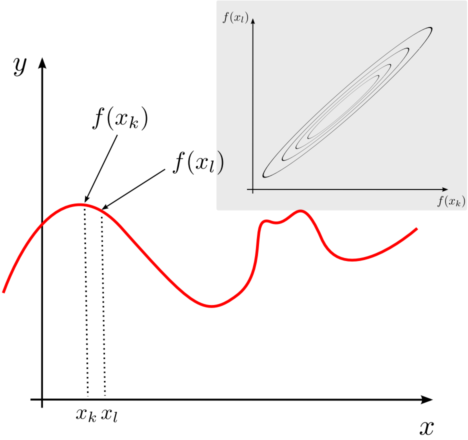
$$Cov\left(f(x_p), f(x_q)\right) = K(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\frac{x_p - x_q}{\ell}\right)^2\right)$$

- 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 $f(x)$.
- We can compute $Cov\left(f(x_p), f(x_q)\right)$ for *any* x_p and x_q.

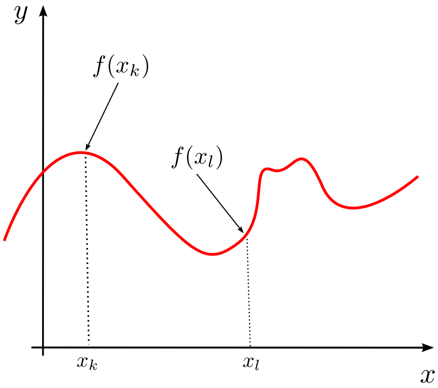
SMOOTH FUNCTION - POINTS NEARBY



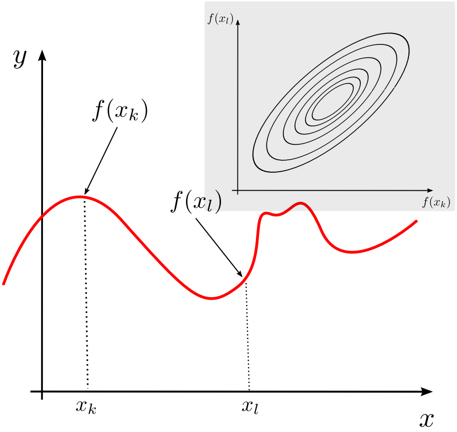
SMOOTH FUNCTION - POINTS NEARBY



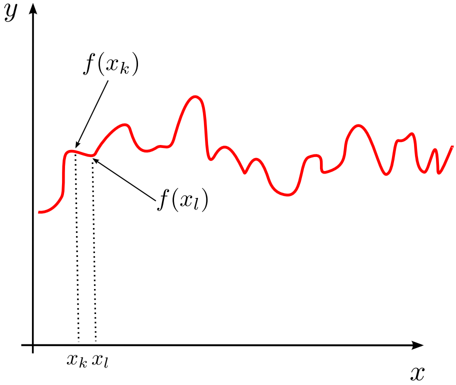
Smooth function - points far apart



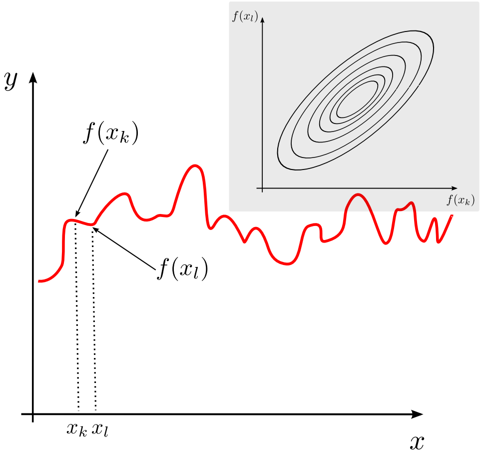
SMOOTH FUNCTION - POINTS FAR APART



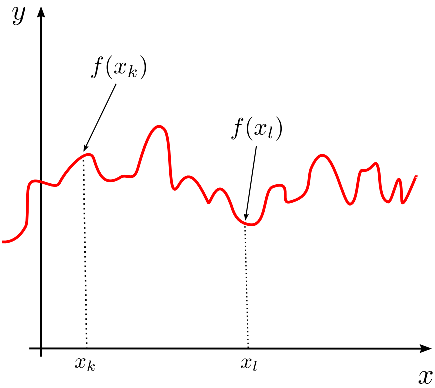
Jagged function - points nearby



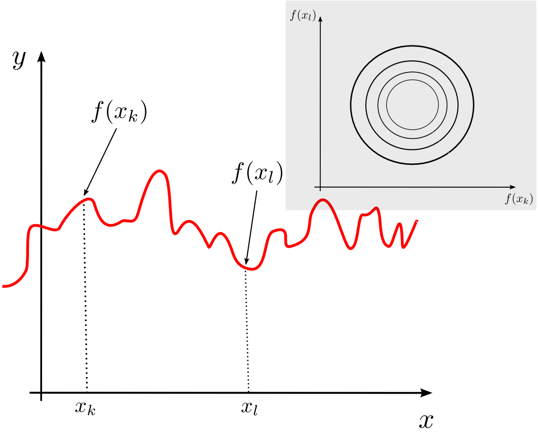
JAGGED FUNCTION - POINTS NEARBY



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 really a **probability distribution over functions** (curves). This is exactly what we want! No need for a grid!
- 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:

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

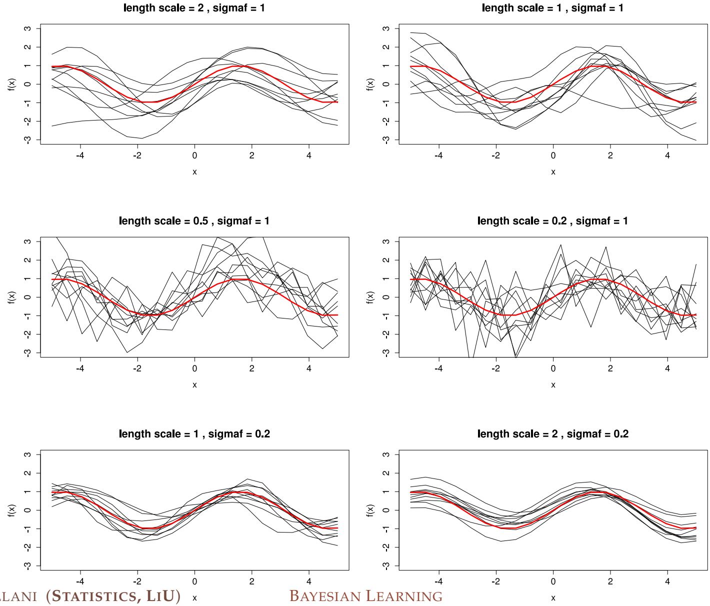
$$K(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\frac{x_p - x_q}{\ell}\right)^2\right)$$

  where $l > 0$ is the length scale.

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

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

- **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: $\mathbf{f}_* = \mathbf{f}(\mathbf{x}_*)$.

Gaussian process regression, cont.

- **Model**

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

- **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: $\mathbf{f}_* = \mathbf{f}(\mathbf{x}_*)$.
- Intermediate step: joint distribution of $\mathbf{y}$ and $\mathbf{f}_*$

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim N\left\{\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} K(\mathbf{x}, \mathbf{x}) + \sigma^2 I & K(\mathbf{x}, \mathbf{x}_*) \\ K(\mathbf{x}_*, \mathbf{x}) & K(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right\}$$

Gaussian process regression, cont.

- **Model**

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

- **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: $\mathbf{f}_* = \mathbf{f}(\mathbf{x}_*)$.
- Intermediate step: joint distribution of $\mathbf{y}$ and $\mathbf{f}_*$

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} K(\mathbf{x}, \mathbf{x}) + \sigma^2 I & K(\mathbf{x}, \mathbf{x}_*) \\ K(\mathbf{x}_*, \mathbf{x}) & K(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right\}$$

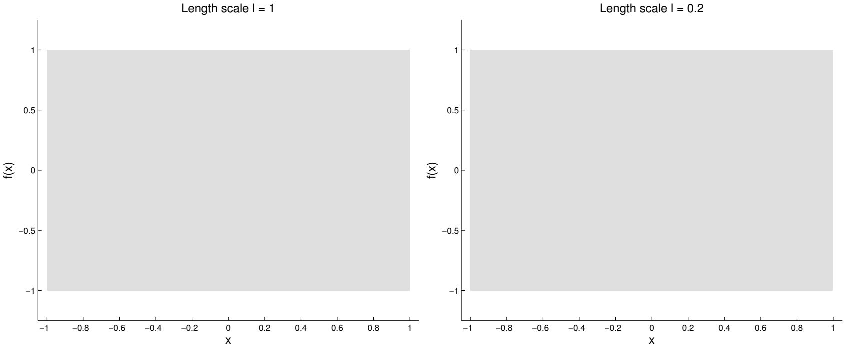
- The **posterior**

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

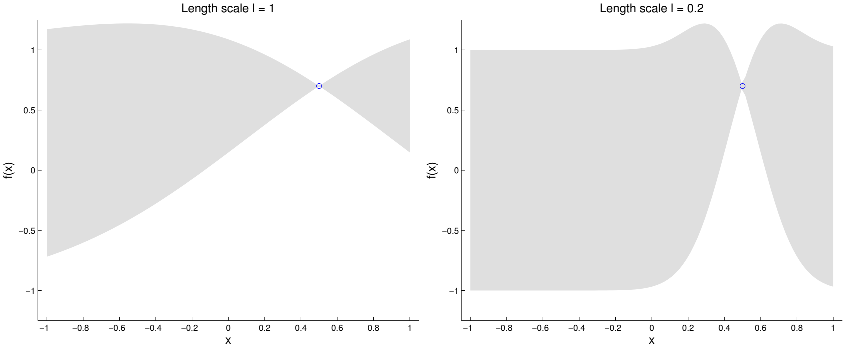
$$\bar{\mathbf{f}}_* = K(\mathbf{x}_*, \mathbf{x}) \left[K(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right]^{-1} \mathbf{y}$$

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

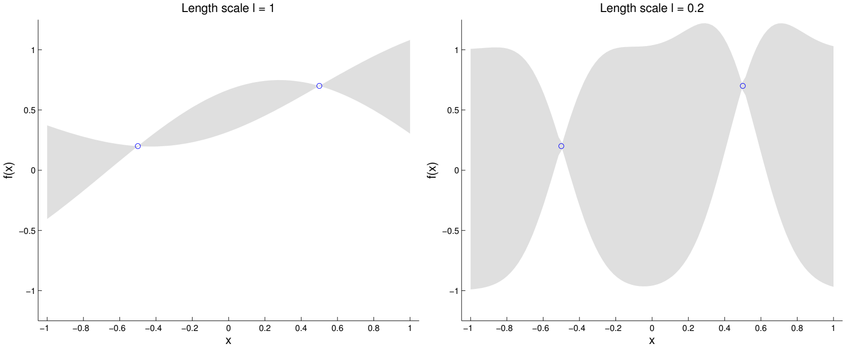
Learning a noise-free Gaussian process



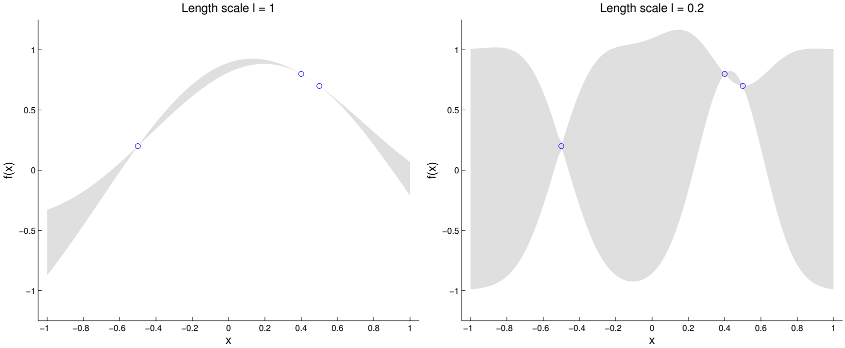
Learning a noise-free Gaussian process



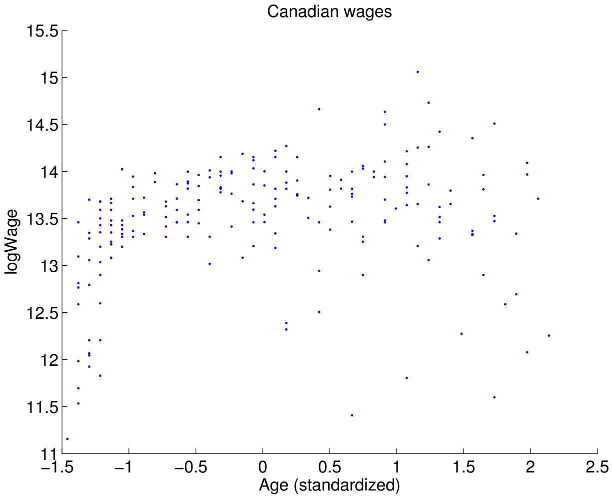
LEARNING A NOISE-FREE GAUSSIAN PROCESS



Learning a noise-free Gaussian process



Example - Canadian wages



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