

732A96/TDDE15 Advanced Machine Learning

Gaussian Process Regression and Classification

Jose M. Peña
IDA, Linköping University, Sweden

Lectures 10: Gaussian Process Regression

Contents

- ▶ Linear Regression
- ▶ Bayesian Linear Regression
- ▶ Gaussian Process Regression
- ▶ Squared Exponential Covariance Function
- ▶ Gaussian Process Regression: Canadian Wages

Literature

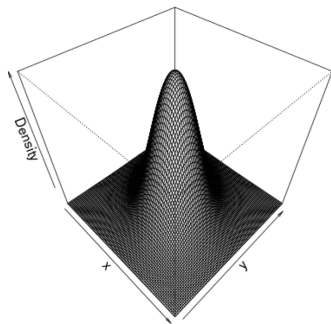
- ▶ Main source
 - ▶ Rasmussen, C. E. and Williams, K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006. Chapters 2.1-2.5.
- ▶ Additional source
 - ▶ Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006. Chapters 6.4.1-6.4.2.

Gaussian Distribution

- Density function of the Gaussian (a.k.a normal) distribution for a n -dimensional random variable \mathbf{x} :

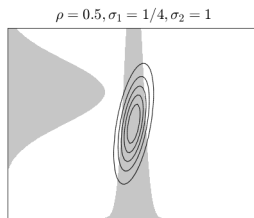
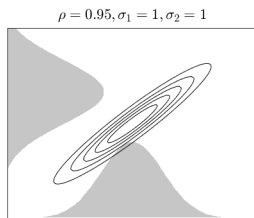
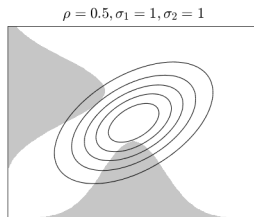
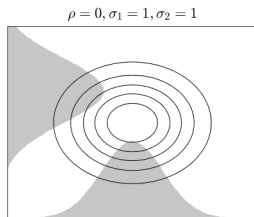
$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right\}$$

- Recall that $E[\mathbf{x}] = \mu$ and $\text{var}(\mathbf{x}) = \Sigma$.



Gaussian Distribution

- Example: $\mathcal{N}(x_1, x_2 | \mu, \Sigma)$ with $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$.



Gaussian Distribution

- Recall that if

$$p(x) = \mathcal{N}(x|\mu, \Lambda^{-1})$$

$$p(y|x) = \mathcal{N}(y|Ax + B, L^{-1})$$

then

$$p(x, y) = \mathcal{N}(x, y|A\mu + B, R^{-1})$$

where

$$R^{-1} = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{pmatrix}$$

- Recall also that if $p(x) = \mathcal{N}(x|\mu, \Sigma)$ and $\Lambda = \Sigma^{-1}$ and

$$\begin{aligned} x &= (x_a, x_b)^T & \mu &= (\mu_a, \mu_b)^T \\ \Sigma &= \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} & \Lambda &= \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \end{aligned}$$

then

$$p(x_a) = \mathcal{N}(x_a|\mu_a, \Sigma_{aa})$$

$$p(x_a|x_b) = \mathcal{N}(x_a|\mu_{a|b}, \Lambda_{aa}^{-1})$$

$$p(x_a|x_b) = \mathcal{N}(x_a|\mu_{a|b}, \Sigma_{a|b})$$

where $\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)$ or

where $\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$

and $\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$

Linear Regression

- ▶ Training data: $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \dots, n\} = (X, \mathbf{y})$.
- ▶ Deterministic function: $f(\mathbf{x}) = \mathbf{x}^T \mathbf{w}$.
- ▶ Additive noisy observations: $y = f(\mathbf{x}) + \epsilon$.
- ▶ Gaussian noise: $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$.
- ▶ Likelihood function: $p(\mathbf{y} | X, \mathbf{w}) = \mathcal{N}(X^T \mathbf{w}, \sigma_n^2 I) \propto \exp \left\{ \frac{1}{2\sigma_n^2} \|\mathbf{y} - X^T \mathbf{w}\|^2 \right\}$.
- ▶ To obtain \mathbf{w}^{ML} ,
 - ▶ take the derivative of the log lik function wrt \mathbf{w} , and
 - ▶ set it to zero, and
 - ▶ solve to obtain $\mathbf{w}^{ML} = (XX^T)^{-1} X\mathbf{y}$.
- ▶ Minimizing the least squared error (i.e., $\frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \mathbf{w})^2$) gives the same result. This justifies the use of LSE.

Bayesian Linear Regression

- ▶ Prior distribution: $\mathbf{w} \sim \mathcal{N}(0, \Sigma_p)$, e.g. ridge regression $\Sigma_p = \alpha^{-1}I$.
- ▶ Posterior distribution:

$$\log p(\mathbf{w}|X, \mathbf{y}) \propto \log p(\mathbf{y}|X, \mathbf{w}) + \log p(\mathbf{w}) \propto \frac{1}{2\sigma_n^2} \|\mathbf{y} - X^T \mathbf{w}\|^2 - \frac{1}{2} \mathbf{w}^T \Sigma_p^{-1} \mathbf{w}.$$

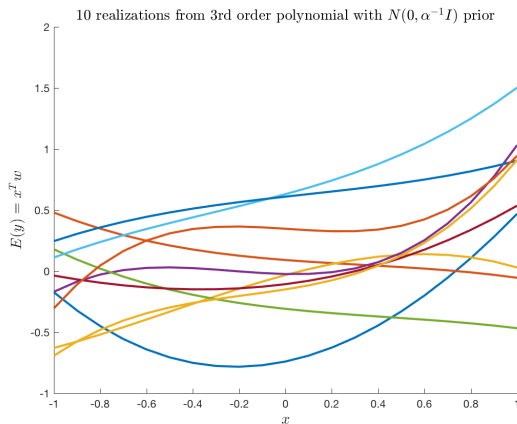
- ▶ So, \mathbf{w}^{MAP} can be seen as penalized/regularized ML estimates.
- ▶ Specifically, $p(\mathbf{w}|X, \mathbf{y}) = \mathcal{N}(\bar{\mathbf{w}} = \frac{1}{\sigma_n^2} A^{-1} X \mathbf{y}, A^{-1})$ where $A = \sigma_n^{-2} X X^T + \Sigma_p^{-1}$, and thus $\mathbf{w}^{MAP} = \bar{\mathbf{w}}$.
- ▶ A full Bayesian approach does not use \mathbf{w}^{MAP} but the predictive distribution:

$$p(f_*|\mathbf{x}_*, X, \mathbf{y}) = \int p(f_*|\mathbf{x}_*, \mathbf{w}) p(\mathbf{w}|X, \mathbf{y}) d\mathbf{w} = \mathcal{N}\left(\frac{1}{\sigma_n^2} \mathbf{x}_* A^{-1} X \mathbf{y}, \mathbf{x}_* A^{-1} \mathbf{x}_*\right).$$

- ▶ The above carries over to the feature space $\phi(\mathbf{x})$. The kernel trick applies. See p. 12 of Rasmussen and Williams.

Bayesian Linear Regression

- A prior on \mathbf{w} is a prior on f .



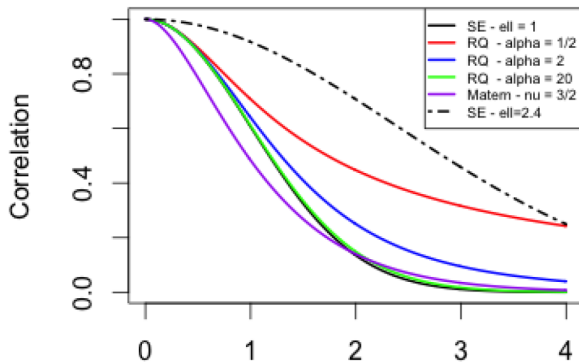
Gaussian Processes Regression

- ▶ A GP defines a prior distribution over functions directly, instead of indirectly through weights as before. Therefore, a GP operates on the space of functions rather than on the space of weights. Operating in either space is equivalent. A GP defines a prior over functions by defining a prior over a finite number of points.
- ▶ Formally, a GP is a collection of random variables, any finite number of which have a joint Gaussian distribution.¹ Hence, a GP is defined as
 - ▶ $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ where
 - ▶ $m(\mathbf{x}) = E[f(\mathbf{x})]$ is the mean function (assumed to be zero hereinafter), and
 - ▶ $k(\mathbf{x}, \mathbf{x}') = E[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$ is the covariance function.
- ▶ E.g., squared exponential covariance function:
$$k(\mathbf{x}, \mathbf{x}') = \text{cov}(f(\mathbf{x}), f(\mathbf{x}')) = \sigma_f^2 \exp \left\{ - \frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2} \right\},$$
 i.e. highly correlated function values for close points.
- ▶ Note that each random variable or dimension in a GP is a function value on a point. Hence, a GP specifies a probability distribution over functions at a finite number of points.
- ▶ We can sample the function space by sampling the GP at any number of chosen points X_* . To do so, we sample a multivariate Gaussian distribution with the corresponding covariance matrix, i.e. $\mathbf{f}_* | X_* \sim \mathcal{N}(0, K(X_*, X_*))$.
- ▶ Demo of GaussianProcesses.R.

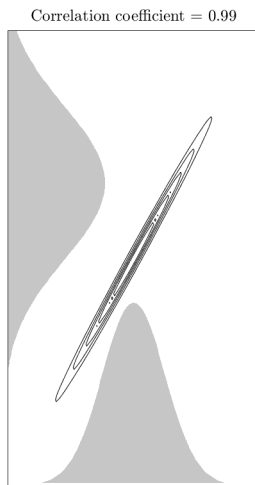
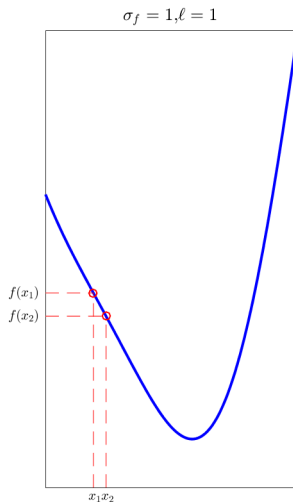
¹If the collection of random variables is finite, then a GP is a Gaussian distribution.

Squared Exponential Covariance Function

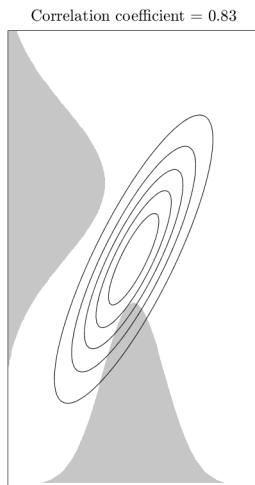
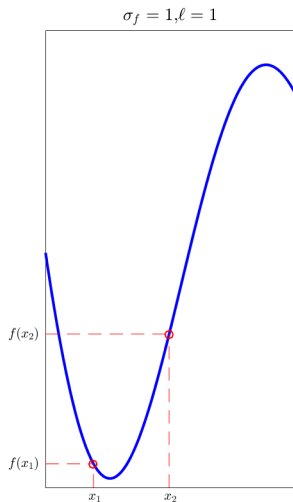
Correlation functions



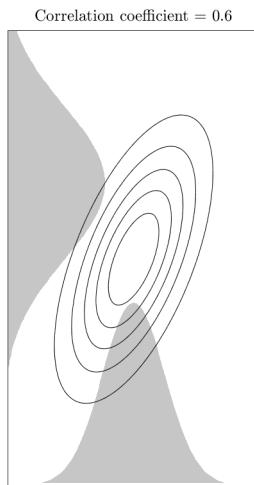
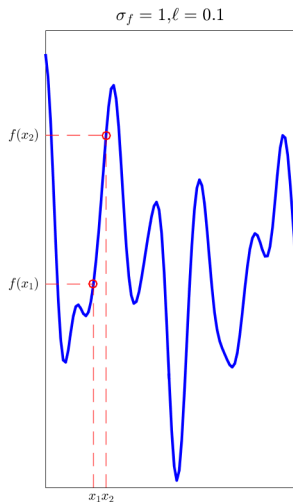
Squared Exponential Covariance: Smooth Function, Close Points



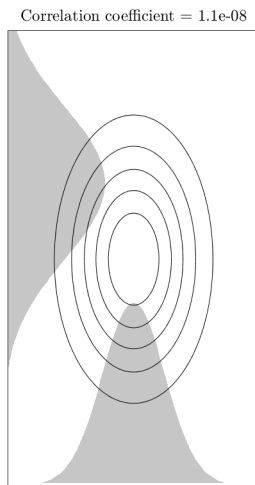
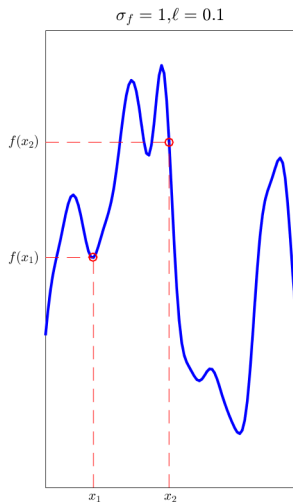
Squared Exponential Covariance: Smooth Function, Distant Points



Squared Exponential Covariance: Jagged Function, Close Points

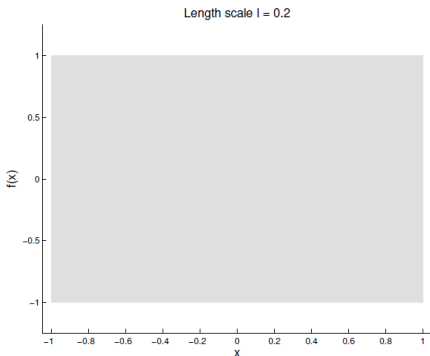
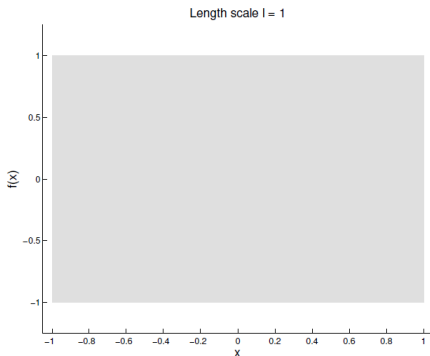


Squared Exponential Covariance: Jagged Function, Distant Points

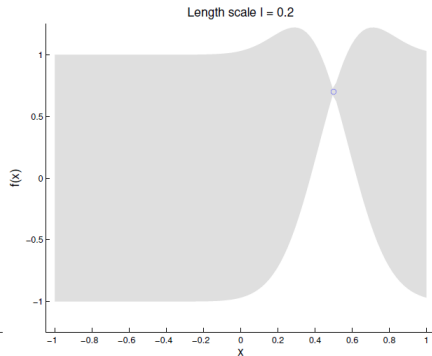
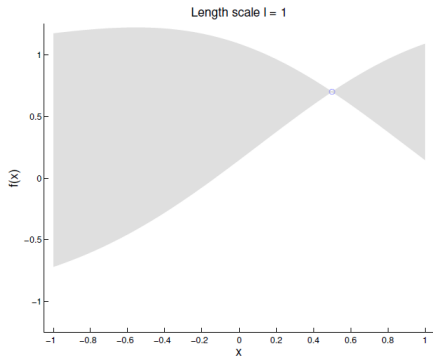


Gaussian Process Sampling: Before First Draw

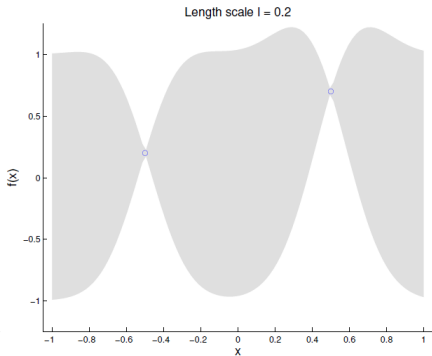
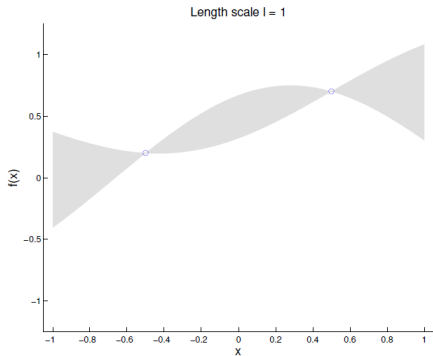
- ▶ To sample a GP at points $X_* = \{x_1, \dots, x_n\}$, we sample a multivariate Gaussian distribution $\mathcal{N}(0, K(X_*, X_*))$.
- ▶ Alternatively, we can sample univariate Gaussian distributions, since $p(f(x_1), \dots, f(x_n)) = \prod_{i=1}^n p(f(x_i) | f(x_1), \dots, f(x_{i-1}))$.



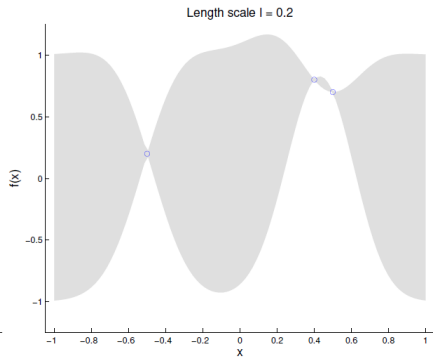
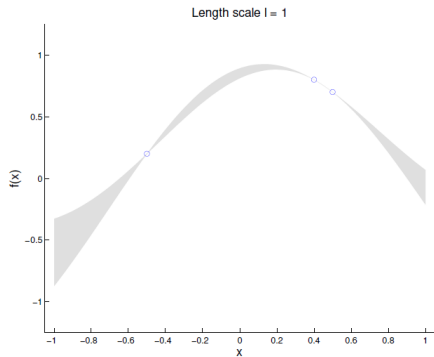
Gaussian Process Sampling: Before Second Draw



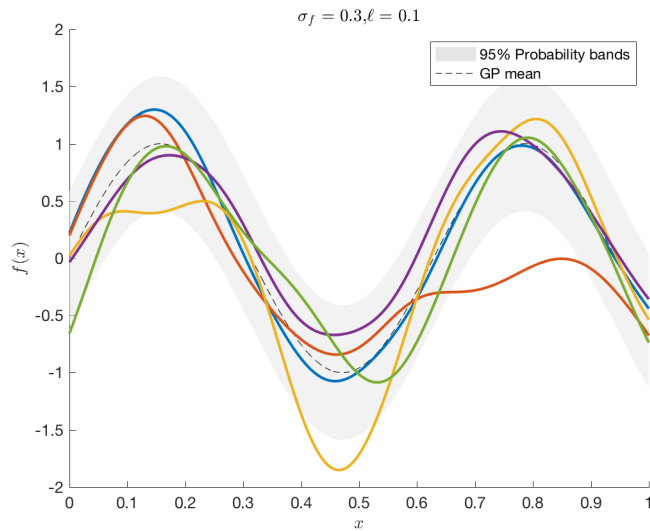
Gaussian Process Sampling: Before Third Draw



Gaussian Process Sampling: Before Fourth Draw



Gaussian Process Sampling



Gaussian Processes Regression

- ▶ With no data, sample from $\mathbf{f}_* | X_* \sim \mathcal{N}(0, K(X_*, X_*))$.
- ▶ With noise-free training data $\mathcal{D} = \{(\mathbf{x}_i, f_i) | i = 1, \dots, n\} = (X, \mathbf{y})$, build

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right)$$

and sample from $\mathbf{f}_* | X_*, X, \mathbf{f} \sim \mathcal{N}(K(X_*, X)K(X, X)^{-1}\mathbf{f}, K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*))$.

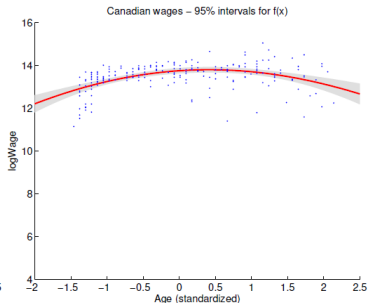
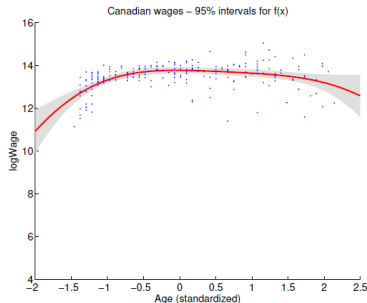
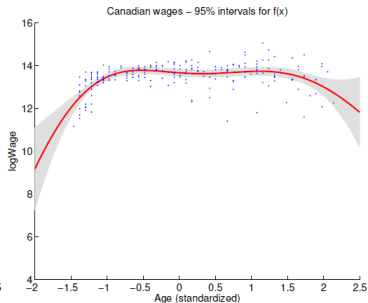
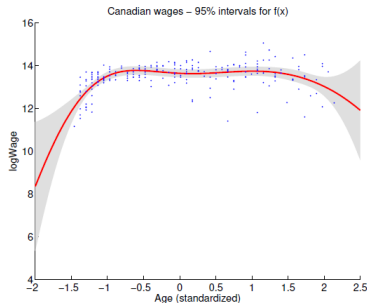
- ▶ With noisy training data $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \dots, n\} = (X, \mathbf{y})$, build

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X, X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right)$$

and sample from $\mathbf{f}_* | X_*, X, \mathbf{y} \sim \mathcal{N}(K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y}, K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, X_*))$.

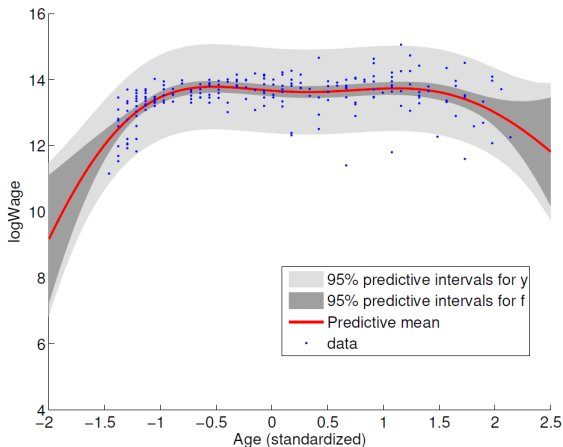
- ▶ See p. 17 of Rasmussen and Williams for the correspondence between the weight and function space views: Every covariance function can be mapped into a set of features, and vice versa.
- ▶ Demo of KernLabDemo.R.

Gaussian Process Regression: Canadian Wages ($\ell = 0.2, 0.5, 1, 2$)



Gaussian Process Regression: Canadian Wages ($\ell = 0.5$)

- ▶ Predictive interval for \mathbf{f}_* : $\text{mean}(\mathbf{f}_*) \pm 1.96 \text{ sd}(\mathbf{f}_*)$.
- ▶ Predictive interval for \mathbf{y}_* : $\text{mean}(\mathbf{f}_*) \pm 1.96 [\text{sd}(\mathbf{f}_*) + \sigma_n]$.



Contents

- ▶ Linear Regression
- ▶ Bayesian Linear Regression
- ▶ Gaussian Processes Regression
- ▶ Squared Exponential Covariance Function
- ▶ Gaussian Process Regression: Canadian Wages

Thank you