Idea

Stochastic regression based on Gaussian functions

Gaussian Processes

Most geophysical processes are **not deterministic**

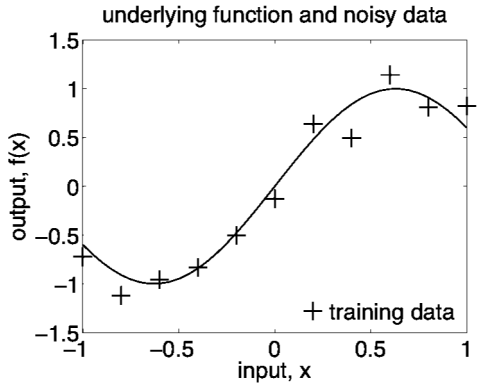
Need some representation of the **noise**

Uncertainty quantifications

→ **Gaussian Processes** (random/stochastic processes)

Gaussian Processes

Supervised learning: Regression



- Assume an underlying process which generates “clean” data.
- Goal: recover underlying process from noisy observed data.

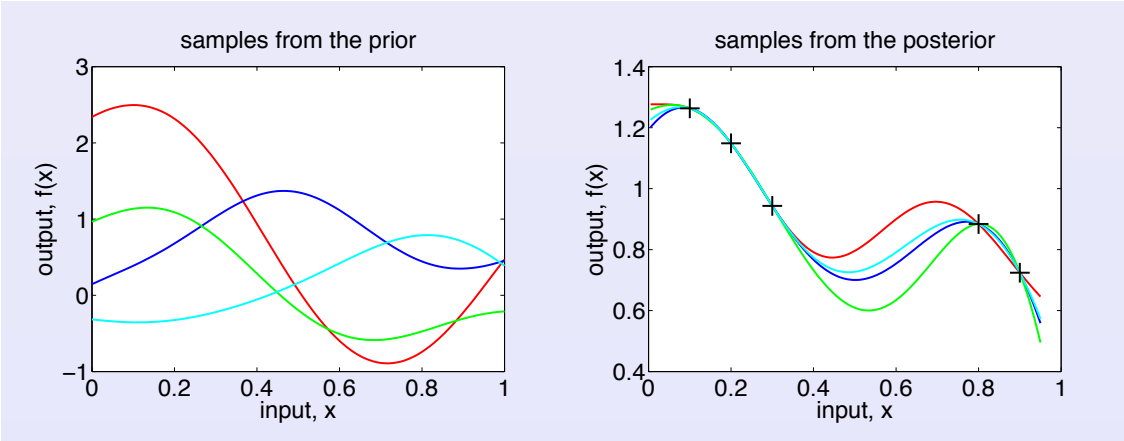
Gaussian Processes

Supervised learning: Regression

- Training data are $D = \{\mathbf{x}(i), y(i) \mid i = 1,\ldots,n\}$
- Each input is a vector $\mathbf{x}$ of dimension d.
- Each target is a real-valued scalar $y = f(x) + \text{noise}$.
- Collect inputs in $d \times n$ matrix, X , and targets in vector, $\mathbf{y}$: $D = \{X , \mathbf{y}\}$.
- Wish to infer $f^\star$ for unseen input $x^\star$, using $P(f^\star | x^\star, D)$.

Gaussian Processes

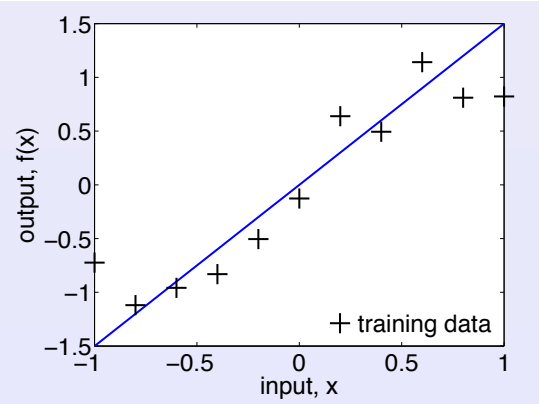
Gaussian Process Models: Inference in Function Space



- A Gaussian process defines a distribution over functions.
- Inference takes place directly in function space.

Linear regressions

Bayesian Linear Regression



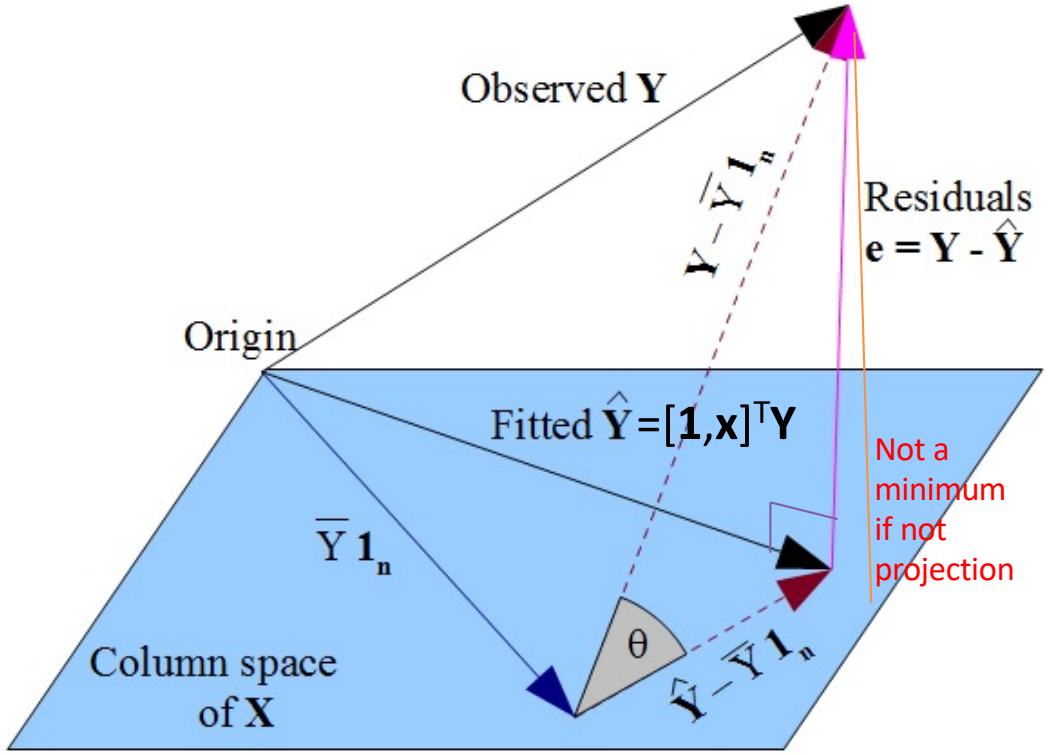
- Assuming noise $\varepsilon \sim N(0, \sigma^2)$, the linear regression model is:

$$f(\mathbf{x}|\mathbf{w}) = \mathbf{x}^T\mathbf{w} = \sum x_i w_i \quad (\mathbf{w}\text{: weights})$$

$$y = f + \varepsilon.$$

Linear regressions

Solution as projection into Span(**1**,**x**) subspace



Linear regressions

Bayes' theorem - posterior relates to prior

$$P(A|B) = P(B|A).P(A) / P(B)$$

As

$$P(A|B).P(B) = P(A \text{ and } B) = P(B|A).P(A)$$

Linear regressions

Bayesian Linear Regression

- Likelihood of parameters is:

$$P(\mathbf{y}|X,\mathbf{w}) = N(X^T\mathbf{w},\sigma^2 I).$$

- Assume a Gaussian prior over parameters:

$$P(w) = N(0,\Sigma_p).$$

- Apply Bayes' theorem to obtain posterior:

$$P(\mathbf{w}|\mathbf{y},X) \propto P(\mathbf{y}|X,\mathbf{w})P(\mathbf{w}).$$

Prior

Linear regressions

Bayesian Linear Regression

Posterior distribution over **w** is:

$$P(\mathbf{w}|\mathbf{y}, X) = \mathcal{N}(\frac{1}{\sigma^2}A^{-1}X\mathbf{y}, A^{-1}) \text{ where } A = \Sigma_p^{-1} + \frac{1}{\sigma^2}XX^\top.$$

Predictive distribution is:

$$\begin{aligned} P(f^\star|\mathbf{x}^\star, X, \mathbf{y}) &= \int f(\mathbf{x}^\star|\mathbf{w})P(\mathbf{w}|X, \mathbf{y})d\mathbf{w} \\ &= \mathcal{N}(\frac{1}{\sigma^2}\mathbf{x}^{\star\top}A^{-1}X\mathbf{y}, \mathbf{x}^{\star\top}A^{-1}\mathbf{x}^\star). \end{aligned}$$

Beyond linear regressions

Projection

• Use a set of basis functions $\Phi(x)$ to project a d dimensional input x into m dimensional feature space:
e.g. polynomial functions $\Phi(x) = (1, x, x^2, ...)$

• $P(f\star|\mathbf{x}\star, X, \mathbf{y})$ can be expressed in terms of inner products in feature space:
One trick: we can use kernels to approximate it (see later)

• Question: How many basis functions should we use?

Gaussian Process: definition

Definition

• A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

• Consistency/marginal distribution:
If the GP specifies $y^{(1)}, y^{(2)} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then it must also specify $y^{(1)} \sim N(\mu_1, \Sigma_{11})$:

• A GP is completely specified by a mean function $\boldsymbol{\mu}$ and a positive definite covariance function $\boldsymbol{\Sigma}$.

Gaussian Process: definition

Distribution over functions

- e.g. Choose mean function zero, and covariance function:

$$K_{p,q} = \mathrm{Cov}(f(x^{(p)}), f(x^{(q)})) = K(x^{(p)}, x^{(q)})$$

- For any set of inputs $x^{(1)}, \ldots, x^{(n)}$ we may compute K which defines a joint distribution over function values:

$$f(x^{(1)}), \ldots, f(x^{(n)}) \sim N(0, K).$$

- Therefore a GP specifies a distribution over functions.

Gaussian Processes

Simple example

•Can obtain a GP from the Bayesian linear regression model:

$$f(x) = x^\top \mathbf{w} \qquad \text{with } \mathbf{w} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p).$$

•Mean function is given by:

$$E[f(\mathbf{x})] = \mathbf{x}^\top E[\mathbf{w}] = 0.$$

•Covariance function is given by:

$$E[f(\mathbf{x}) f(\mathbf{x}')] = \mathbf{x}^\top E[\mathbf{w}\mathbf{w}^\top]\mathbf{x}' = \mathbf{x}^\top \boldsymbol{\Sigma}_p \mathbf{x}'.$$

Gaussian Processes

Weight space and function space correspondence

•For any set of m basis functions, Φ(x), the corresponding covariance function is:

$$K(\mathbf{x}^{(p)}, \mathbf{x}^{(q)}) = \Phi(\mathbf{x}^{(p)})^{\top} \Sigma_p \Phi(\mathbf{x}^{(q)}).$$

Conversely, for every covariance function k, there is a possibly infinite expansion in terms of basis functions:

$$K(\mathbf{x}^{(p)}, \mathbf{x}^{(q)}) = \sum_{i=1}^{\infty} \lambda_i \Phi_i(\mathbf{x}^{(p)}) \Phi_i(\mathbf{x}^{(q)}).$$

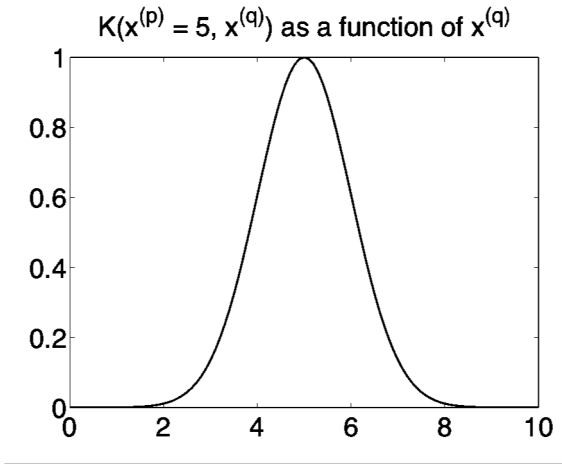
Gaussian Processes

Covariance function

•Specifies the covariance between pairs of random variables.

e.g. **Squared exponential covariance function**:

$$\mathrm{Cov}(f(\mathbf{x}^{(p)}), f(\mathbf{x}^{(q)})) = K(\mathbf{x}^{(p)}, \mathbf{x}^{(q)}) = \exp\left(-\frac{1}{2}|\mathbf{x}^{(p)} - \mathbf{x}^{(q)}|^2\right).$$



Translation invariant!

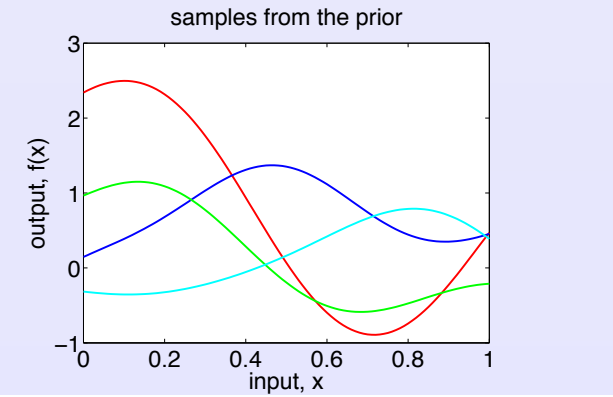
Gaussian Processes

Gaussian process Prior

•Given a set of inputs $x^{(1)}, \ldots, x^{(n)}$ we may draw samples $f(x^{(1)}), \ldots, f(x^{(n)})$ from the GP prior

$$f(\mathbf{x}^{(1)}), \ldots, f(\mathbf{x}^{(n)}) \sim \mathcal{N}(\mathbf{0}, K).$$

•Four samples:



Gaussian Processes

Posterior: Noise-free Observations

•Condition {X⋆,f⋆} on D = {X,f} obtain the posterior.

•Restrict prior to contain only functions which agree with D.

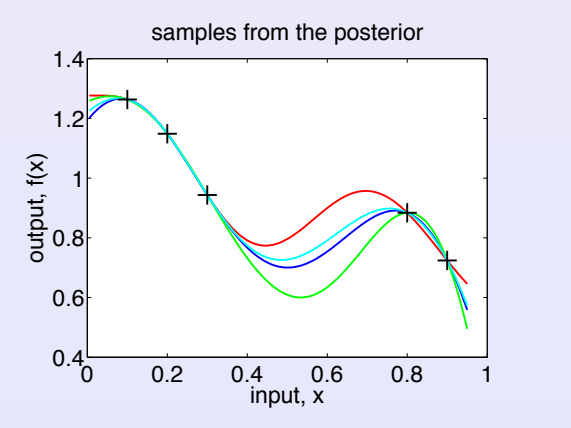
•The posterior, P(f⋆|X⋆,X,f), is Gaussian with:

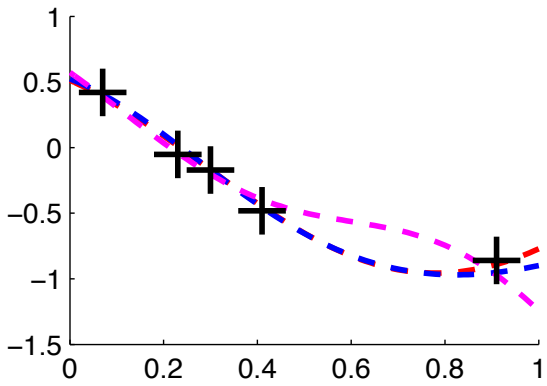
$$\boldsymbol{\mu} = K(X, X^*)K(X, X)^{-1}\mathbf{f}, \text{ and}$$
$$\Sigma = K(X^*, X^*) - K(X, X^*)K(X, X)^{-1}K(X^*, X).$$

Gaussian Processes

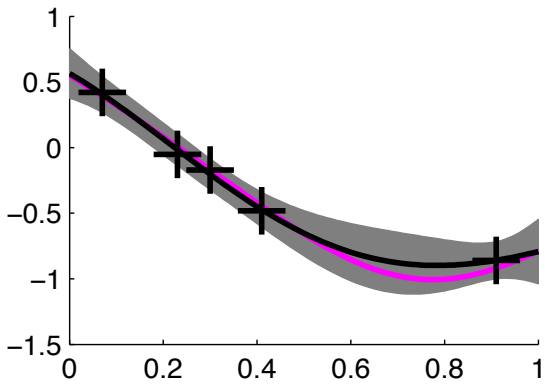
Posterior: Noise-free Observations

- Samples all agree with the observations D = {X,f}.
- Greatest variance is in regions with few training points.





Draws $\sim p(\mathbf{f}|\text{data})$



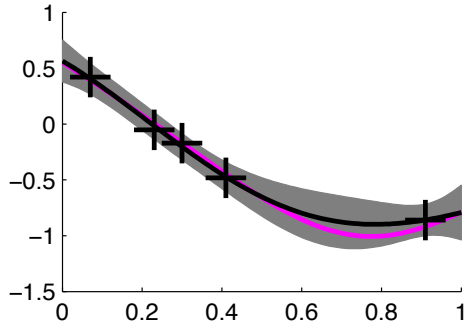
Mean and error bars

Gaussian Processes

Prediction: Noisy Observations

•Typically we have noisy observations:

D={X,**y**}, where **y**=**f**+**ε**



Mean and error bars

•Assume additive iid noise **ε** ~ N (**0**, $\sigma^2 I$).

Conditioning on D = {X , **y**} gives a Gaussian with:

$$\boldsymbol{\mu} = K(X, X^\star)[K(X, X) + \sigma^2 I]^{-1}\mathbf{y}, \text{ and}$$

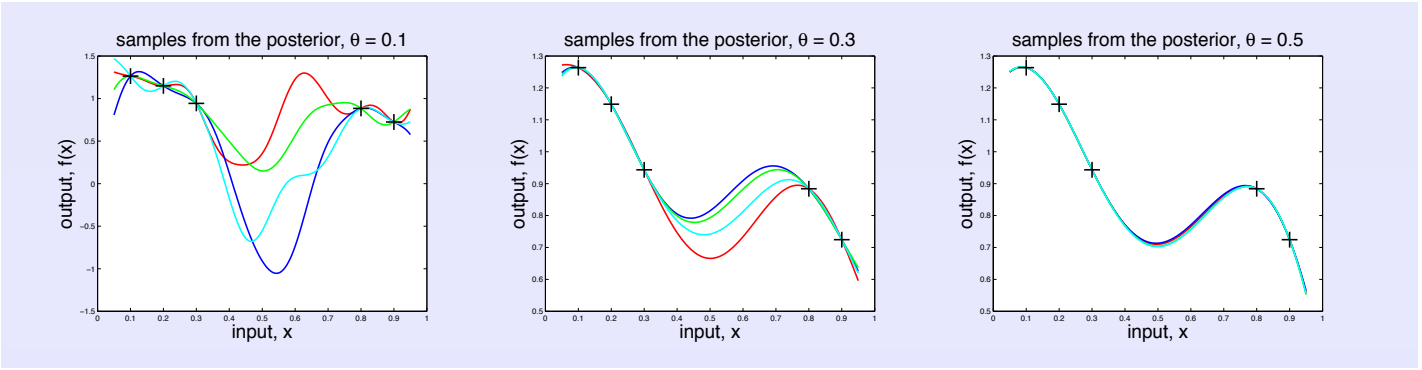
$$\Sigma = K(X^\star, X^\star) - K(X, X^\star)[K(X, X) + \sigma^2 I]^{-1}K(X^\star, X).$$

Gaussian Processes

Model selection: hyperparameters (same as NNs)

•E.g. the ARD covariance function

$$k(x^{(p)}, x^{(q)}) = \exp\left(-\frac{1}{2\theta^2}(x^{(p)} - x^{(q)})^2\right).$$



•How best to choose θ?

Gaussian Processes

Model selection: hyperparameters

•In absence of a strong prior P(θ), the posterior for hyperparameter θ is proportional to the marginal likelihood:

$$P(\theta|X,\mathbf{y}) \propto P(\mathbf{y}|X,\theta)$$

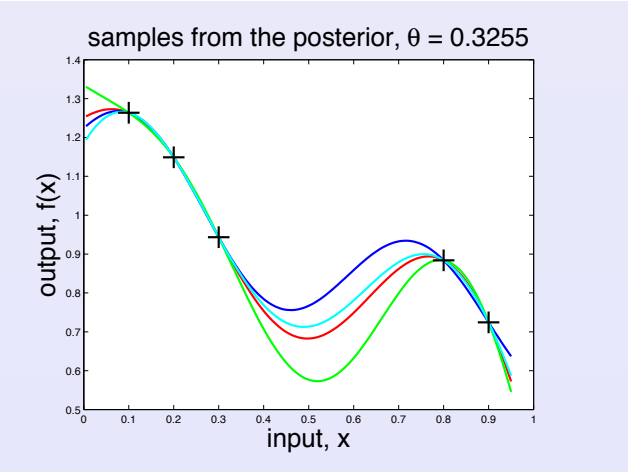
•Choose θ to optimize the marginal log-likelihood:

$$\log P(\mathbf{y}|X,\theta) = -\frac{1}{2}\log|K(X,X)+\sigma^2 I| - \frac{1}{2}\mathbf{y}^\top (K(X,X)+\sigma^2 I)^{-1}\mathbf{y} - \frac{n}{2}\log 2\pi.$$

Gaussian Processes

Model selection: hyperparameters

$\theta^{ML} = 0.3255$:

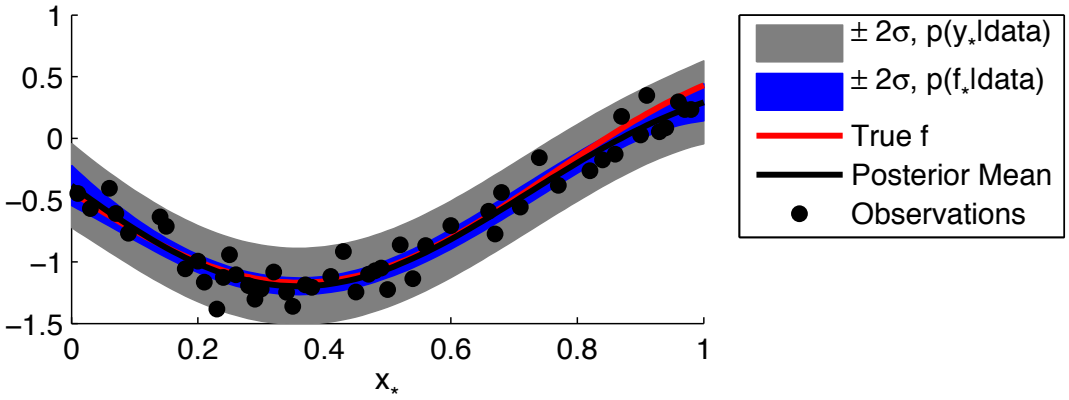


Using θ^{ML} is an approximation to the true Bayesian method of integrating over all θ values weighted by their posterior.

Gaussian Processes

Discovery or prediction

What should error-bars show?



$P(f_*|\text{data}) = \mathcal{N}(m, s^2)$ says what we know about the noiseless function.

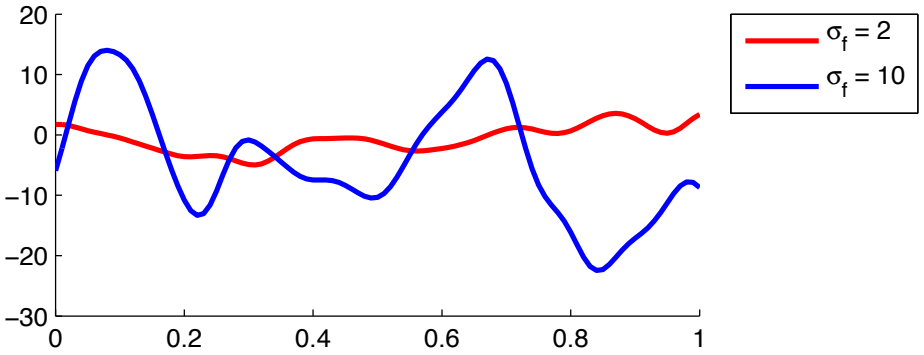
$P(y_*|\text{data}) = \mathcal{N}(m, s^2+\sigma_n^2)$ predicts what we'll see next.

Kernels

Many kernels have similar types of parameters:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{d=1}^{D} (x_{d,i} - x_{d,j})^2 / \ell_d^2 \right),$$

Consider $\mathbf{x}_i = \mathbf{x}_j$, $\Rightarrow$ marginal function variance is σ_f^2



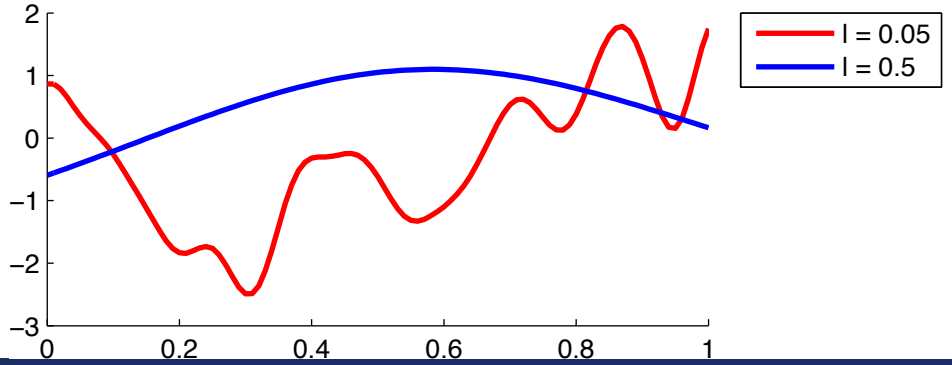
Gaussian Processes

Kernels

The ℓ_d parameters give the overall lengthscale in dimension-d

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{d=1}^{D} (x_{d,i} - x_{d,j})^2 / \ell_d^2 \right),$$

Typical distance between peaks $\approx \ell$



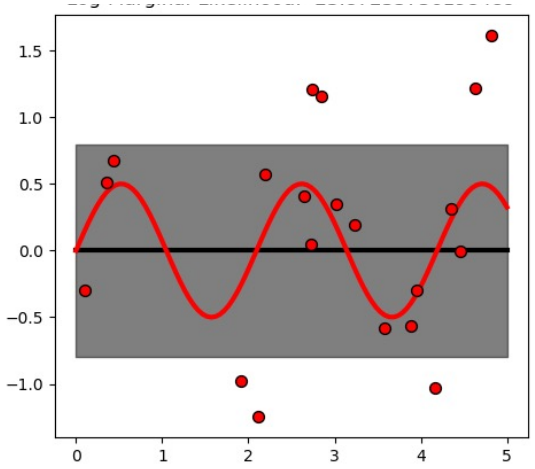
Gaussian Processes

Example of Kernels

Constant kernel

$$k(x_i, x_j) = \text{constant}$$

- The main use-case of the constant kernel is that it explains the noise-component of the signal. Tuning its parameter corresponds to estimating the noise-level



Example of Kernels

Squared exponential kernel

An ∞ number of radial-basis functions can give

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{d=1}^{D} (x_{d,i} - x_{d,j})^2 / \ell_d^2 \right),$$

the most commonly-used kernel in machine learning.

It looks like an (unnormalized) Gaussian, so is commonly called the Gaussian kernel. *Please* remember that this has nothing to do with it being a *Gaussian* process.

A Gaussian process need not use the "Gaussian" kernel. In fact, other choices will often be better.

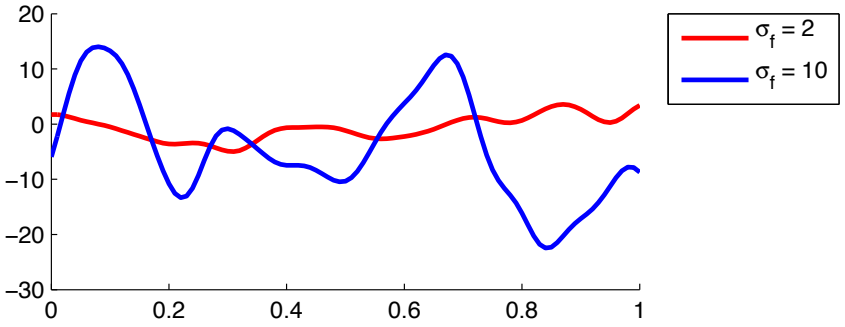
Example of Kernels

Squared exponential kernel

Many kernels have similar types of parameters:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{d=1}^{D} (x_{d,i} - x_{d,j})^2 / \ell_d^2 \right),$$

Consider $\mathbf{x}_i = \mathbf{x}_j$, $\Rightarrow$ marginal function variance is σ_f^2



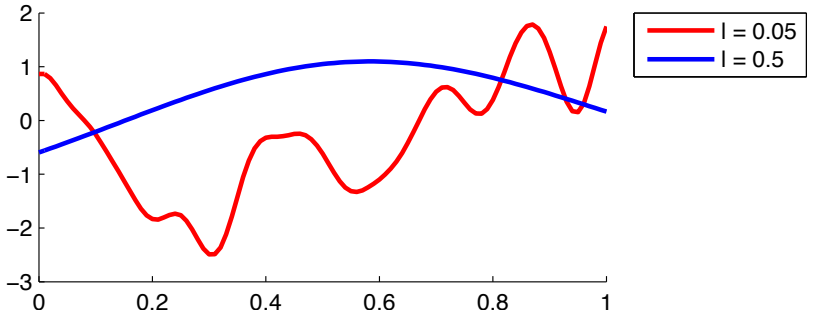
Example of Kernels

Squared exponential kernel

The ℓ_d parameters give the overall lengthscale in dimension-d

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{d=1}^{D} (x_{d,i} - x_{d,j})^2 / \ell_d^2 \right),$$

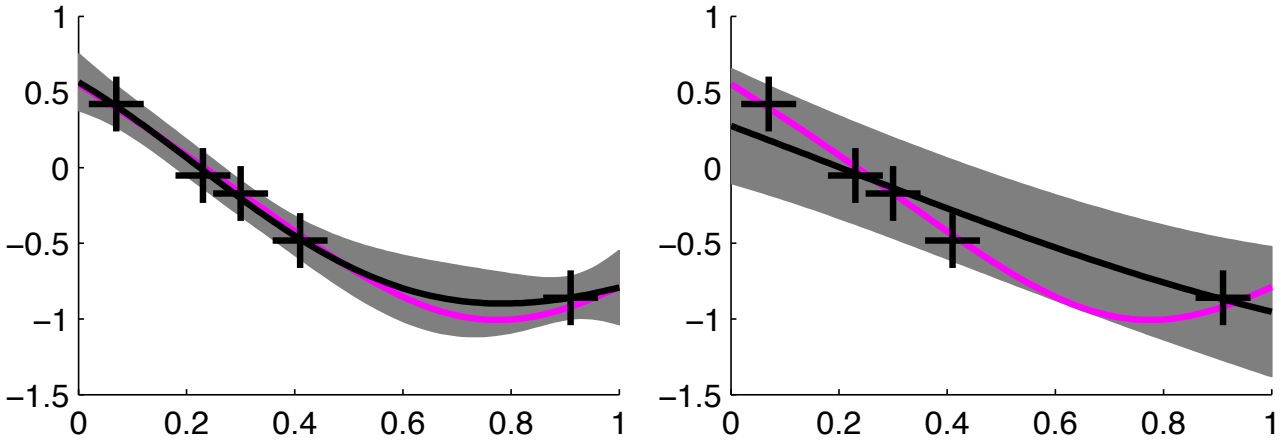
Typical distance between peaks $\approx \ell$



Example of Kernels

Interpretation

Different (SE) kernel parameters give different explanations of the data:



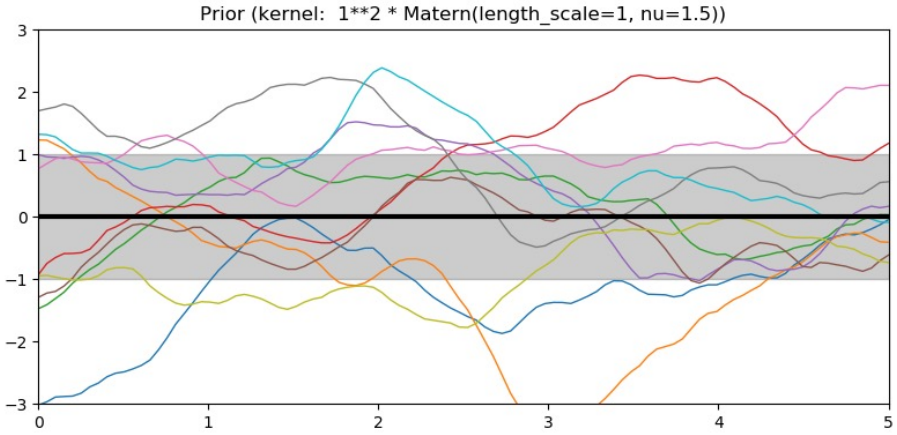
$\ell = 0.5, \quad \sigma_n = 0.05$ $\qquad$ $\ell = 1.5, \quad \sigma_n = 0.15$

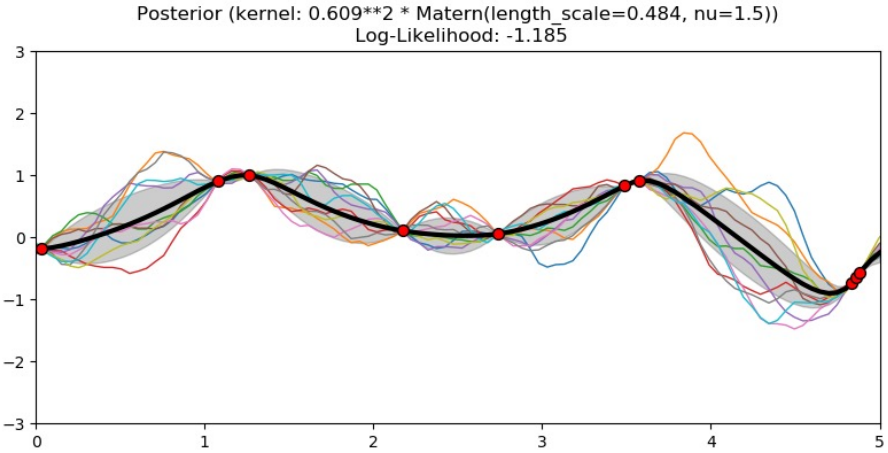
Gaussian Processes

Example of Kernels

Matérn kernel – generalize RBF kernel with smoothing parameter ν

$$k(x_i, x_j) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\gamma\sqrt{2\nu} d(x_i/l, x_j/l) \right)^{\nu} K_\nu \left(\gamma\sqrt{2\nu} d(x_i/l, x_j/l) \right),$$



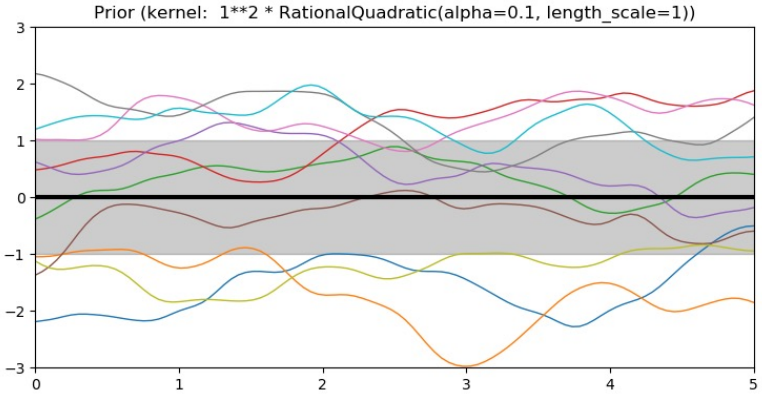


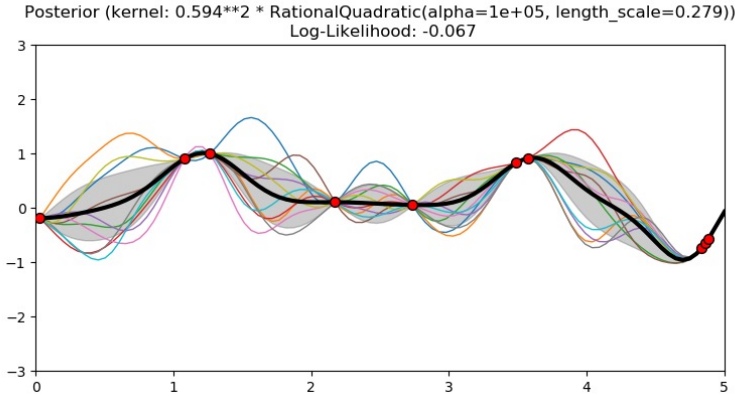
Gaussian Processes

Example of Kernels

Rational quadratic kernel – scale mixture (an infinite sum) ofRBF kernels with different characteristic length-scales and mixture parameter α

$$k(x_i, x_j) = \left(1 + \frac{d(x_i, x_j)^2}{2\alpha l^2}\right)^{-\alpha}$$



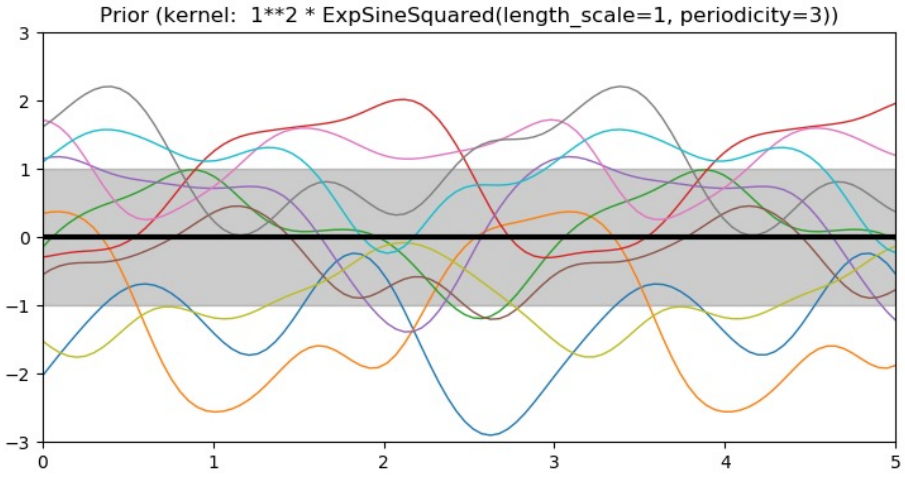


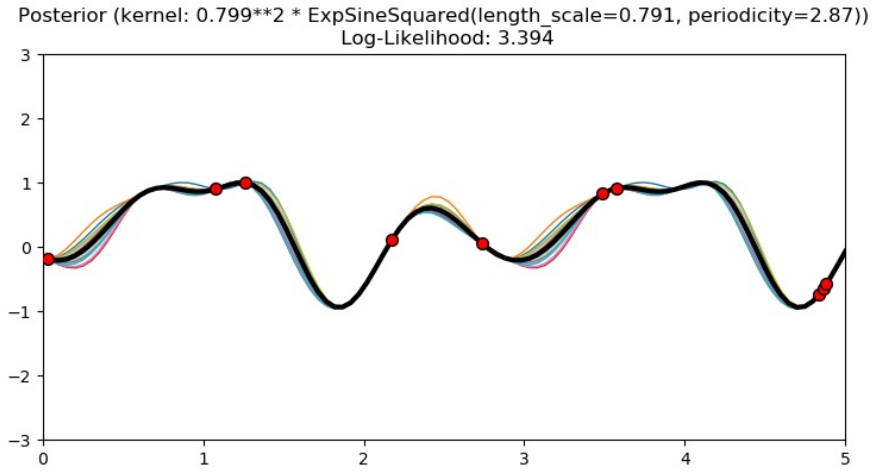
Gaussian Processes

Example of Kernels

Exp-Sine-Squared kernel – allows modeling periodic functions.

$$k(x_i, x_j) = \mathbf{exp}\left(-2\left(\mathbf{sin}(\pi/p * d(x_i, x_j))/l\right)^2\right)$$





Gaussian Processes

Example of use: GPR on Mauna Loa CO2 data

Monthly average atmospheric CO2 concentrations (in parts per million by volume – ppm) collected at the Mauna Loa Observatory in Hawaii, between 1958 and 1997

Kernel made of

- a **long term**, smooth trend by an RBF kernel. The RBF kernel with a large length-scale enforces this component to be smooth (not necessarily rising)
- a **seasonal component**, with Exp Sine Squared kernel with a fixed periodicity of 1 year. The length-scale of this periodic component, controlling its smoothness, is a free parameter.
- smaller, **medium term irregularities** are to be explained by a Rational Quadratic kernel component
- **a "noise" term**, consisting of an RBF kernel contribution, which shall explain the correlated noise components such as local weather phenomena, and a Constant contribution for the white noise. The relative amplitudes and the RBF's length scale are further free parameters.