Gaussian Process

Dripta Mj
Department of Mathematics
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, India

Machine Learning
Sem 3, 2018-19

Sem 3, 2018-19

Introduction

- Uncertainty is ubiquitous in most real-world problems.
- Different forms of uncertainty:
  - Measurement noise
  - Parameter uncertainty
  - Structural uncertainty
- Ignoring uncertainty risk poor prediction, decision making.
- Bayesian approach provides a principled framework for handling uncertainty.
- The application of probability theory to learning from data is called Bayesian learning[1].

[1] Z. Ghahramani, Nature, 2015.

Bayes Rule

Product Rule

$$p(a,b) = p(b|a)p(a)$$

Sum Rule

$$p(a) = \sum_b p(a,b)$$

Bayes Rule

$$p(a|b) = \frac{p(b|a)p(a)}{p(b)} = \frac{p(b|a)p(a)}{\sum_a p(b|a)p(a)}$$

$$p(\text{☀}) = 0.13$$
$$p(\text{🌧}) = 0.85$$
$$p(\text{⛈}) = 0.02$$

$$p(\textbf{IN}|\text{☀}) = 0.05$$
$$p(\textbf{IN}|\text{🌧}) = 0.80$$
$$p(\textbf{IN}|\text{⛈}) = 0.99$$

$$\begin{aligned} p(\textbf{IN}) &= p(\textbf{IN},\text{☀}) + p(\textbf{IN},\text{🌧}) + p(\textbf{IN},\text{⛈}) \\ &= p(\textbf{IN}|\text{☀})p(\text{☀}) + p(\textbf{IN}|\text{🌧})p(\text{🌧}) + p(\textbf{IN}|\text{⛈})p(\text{⛈}) \\ &= 0.71 \end{aligned}$$

$$\begin{aligned} p(\text{☀}|\textbf{IN}) &= \frac{p(\textbf{IN}|\text{☀})\, p(\text{☀})}{p(\textbf{IN})} \\ &= \frac{p(\textbf{IN}|\text{☀})\, p(\text{☀})}{p(\textbf{IN}|\text{☀})p(\text{☀}) + p(\textbf{IN}|\text{🌧})p(\text{🌧}) + p(\textbf{IN}|\text{⛈})p(\text{⛈})} \\ &= 0.009 \end{aligned}$$

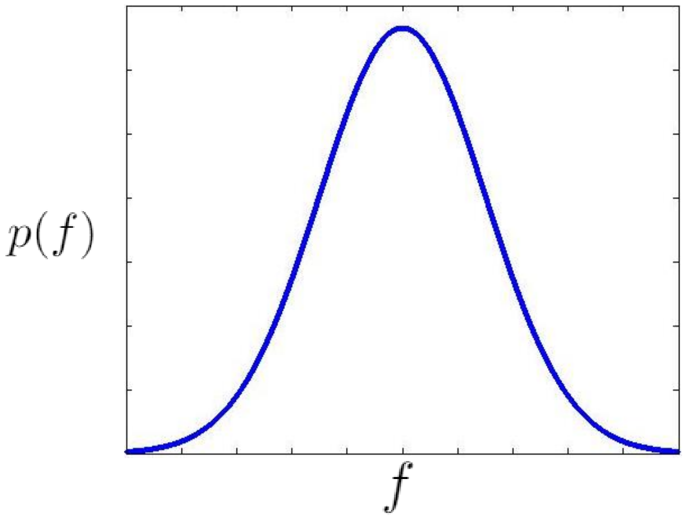
Bayesian ML

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

- The posterior distribution expresses the model knowledge after incorporating the data and the prior assumption.
- The likelihood is the probability density of the observations given the parameters.
- The prior distribution expresses our prior beliefs of the model before observing the data.
- The marginal likelihood (or evidence) is the integral of the likelihood times the prior.

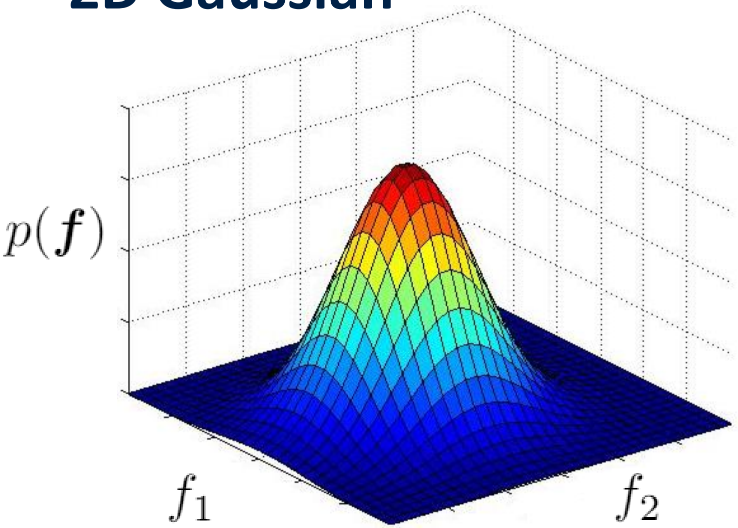
Gaussian Distribution

1D Gaussian



$$p(f) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(f-\mu)^2 \right]$$

2D Gaussian

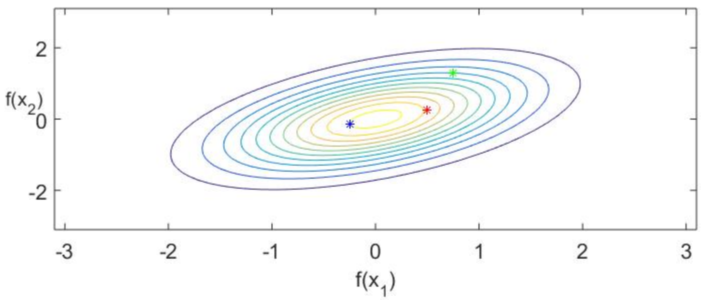


$$p(\boldsymbol{f}) = \frac{1}{2\pi|\boldsymbol{K}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{\mu})^T \boldsymbol{K}^{-1}(\boldsymbol{f}-\boldsymbol{\mu}) \right]$$

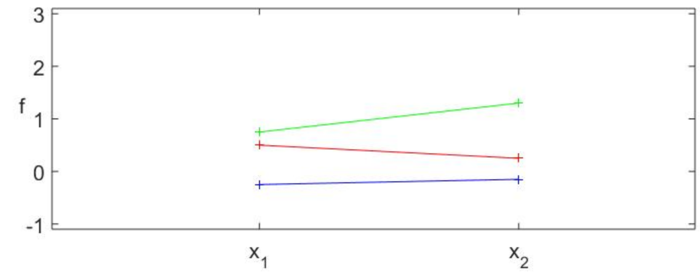
$$\boldsymbol{f} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \boldsymbol{K} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

Multivariate Gaussian Distribution

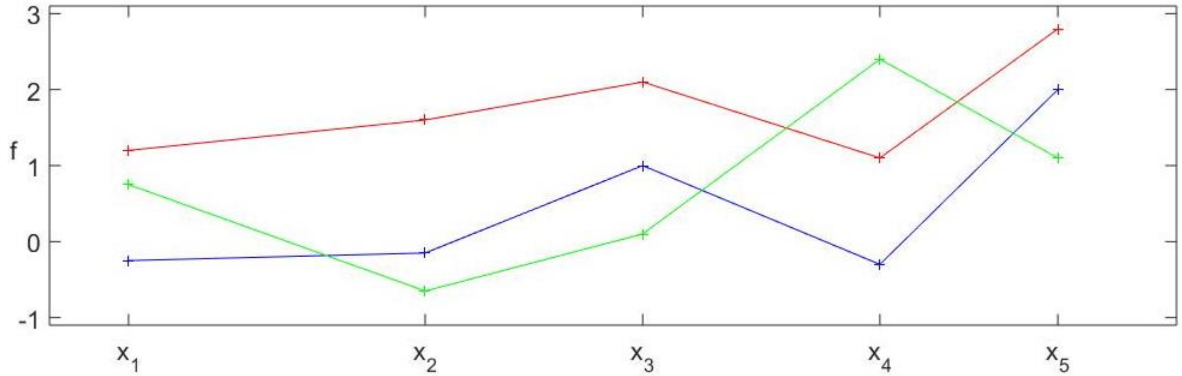
2D Gaussian



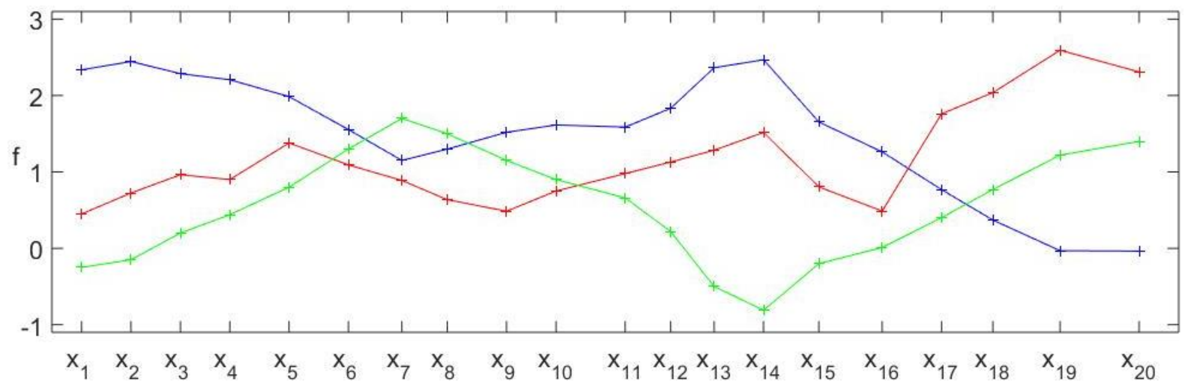
Draws



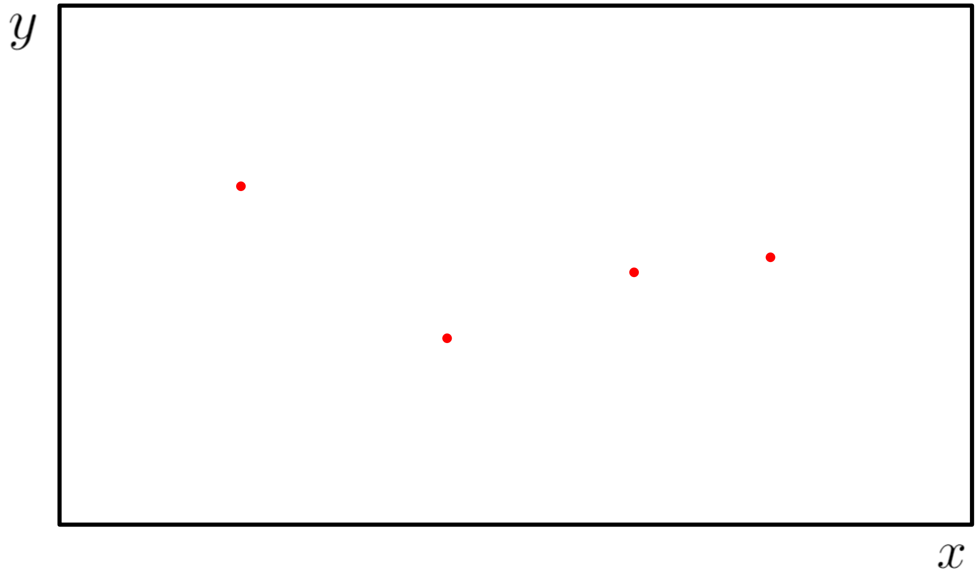
Draws from 5D Gaussian



Draws from 20D Gaussian



Regression with Bayesian ML



Training points:

(given)

Inputs $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}\}$

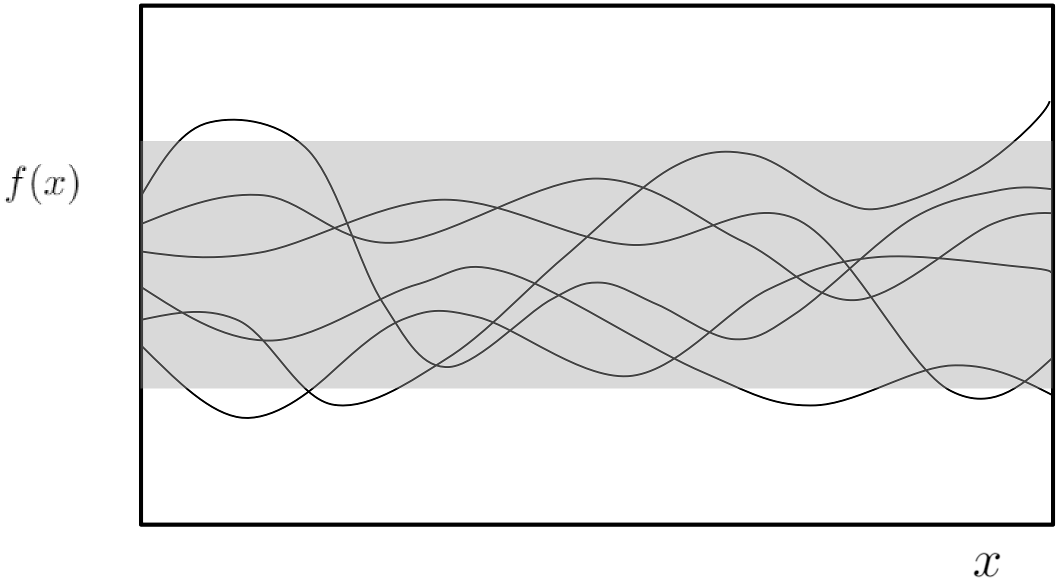
Outputs $\mathbf{y} = \{y^{(1)}, y^{(2)}, ..., y^{(N)}\}$

- Want predictions $\boldsymbol{y}_*$ at unobserved locations $\boldsymbol{x}_*$

Model: $y_i = f(x_i) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_n^2)$

- Evaluate $p(\boldsymbol{y}_* | \boldsymbol{y})$

Prior distribution

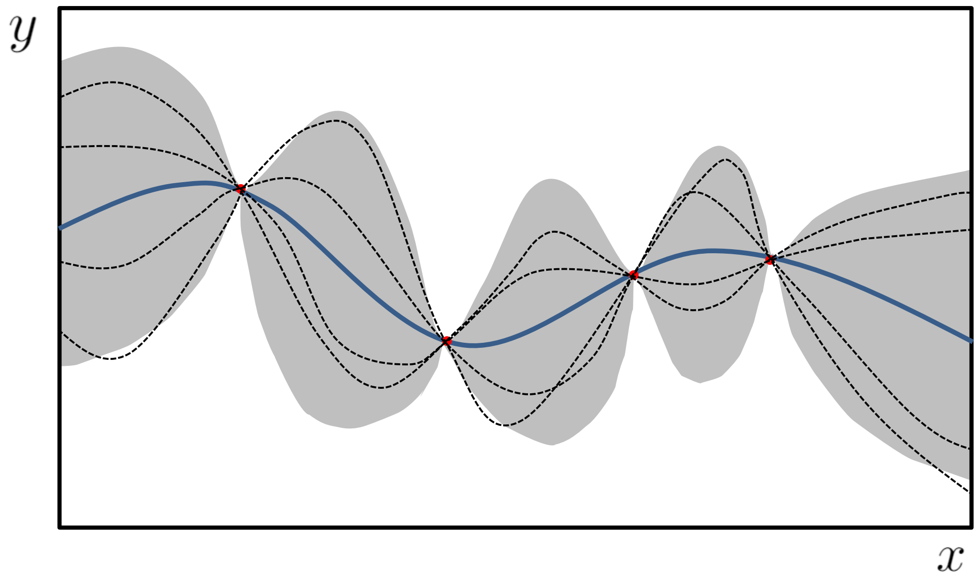


Model: $y^{(i)} = f(\mathbf{x}^{(i)}) + \epsilon^{(i)}, \qquad \epsilon^{(i)} \sim \mathcal{N}(0, \sigma_n^2)$

Prior: $f \sim \mathcal{GP}(m, k)$

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{m}, \mathbf{K})$$

Regression with Bayesian ML



Test points: $$\mathbf{X}_* = \left\{\mathbf{x}_*^{(1)}, \mathbf{x}_*^{(2)}, ..., \mathbf{x}_*^{(M)}\right\}$$

- Want predictions $\mathbf{y}_* = \left\{y_*^{(1)}, y_*^{(2)}, ..., y_*^{(M)}\right\}$ at $\mathbf{X}_*$.
- Evaluate $p(\mathbf{y}_*|\mathbf{y})$.

Kernel matrix

- Also known as Gram matrix.

- Formed by applying the kernel function k to all pairs of data points in $\mathbf{X}$.

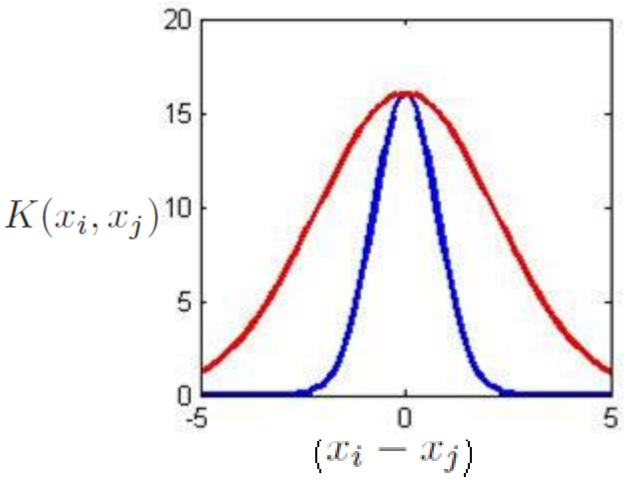
$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}) & k(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) & \cdot & \cdot & k(\mathbf{x}^{(1)}, \mathbf{x}^{(N)}) \\ k(\mathbf{x}^{(2)}, \mathbf{x}^{(1)}) & k(\mathbf{x}^{(2)}, \mathbf{x}^{(2)}) & \cdot & \cdot & k(\mathbf{x}^{(2)}, \mathbf{x}^{(N)}) \\ k(\mathbf{x}^{(3)}, \mathbf{x}^{(1)}) & k(\mathbf{x}^{(3)}, \mathbf{x}^{(2)}) & \cdot & \cdot & k(\mathbf{x}^{(3)}, \mathbf{x}^{(N)}) \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ k(\mathbf{x}^{(N)}, \mathbf{x}^{(1)}) & k(\mathbf{x}^{(N)}, \mathbf{x}^{(2)}) & \cdot & \cdot & k(\mathbf{x}^{(N)}, \mathbf{x}^{(N)}) \end{bmatrix}$$

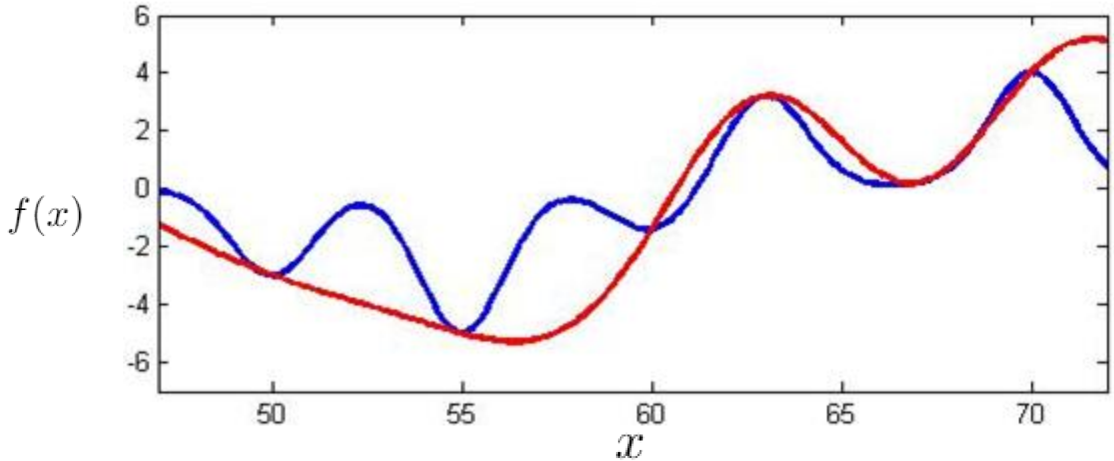
- Square matrix of size $N \times N$.

- Symmetric.

Kernel functions

Exponentiated Quadratic: $K(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{|x_i - x_j|^2}{2l^2}\right)$

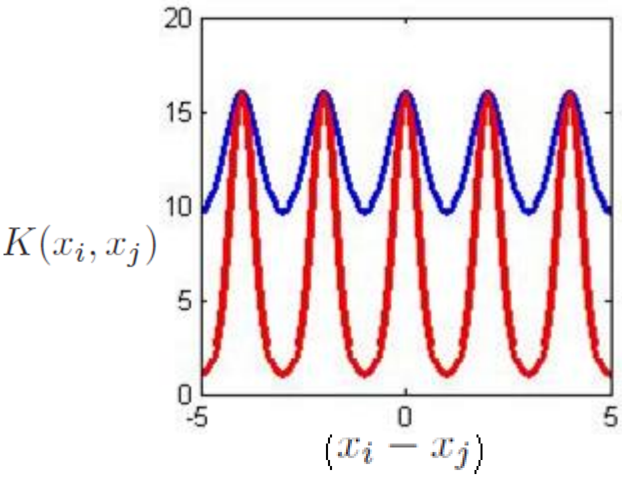


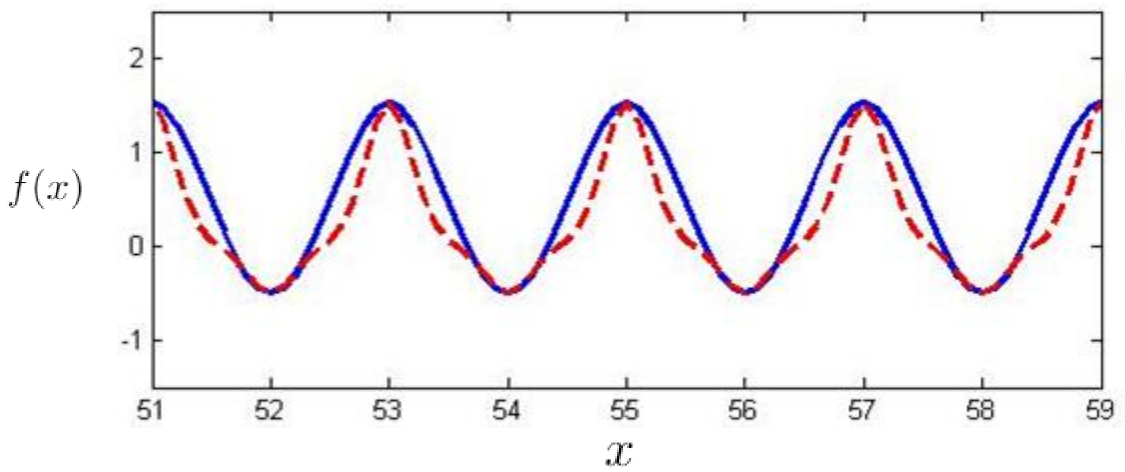


Figures from: Prediction of tidal currents using Bayesian machine learning, Ocean Engineering, 2018.

Kernel functions

Periodic: $K(x_i, x_j) = \sigma_f^2 \exp\left(\frac{-2}{l^2}\sin^2\left(\frac{\pi|x_i - x_j|}{p}\right)\right)$

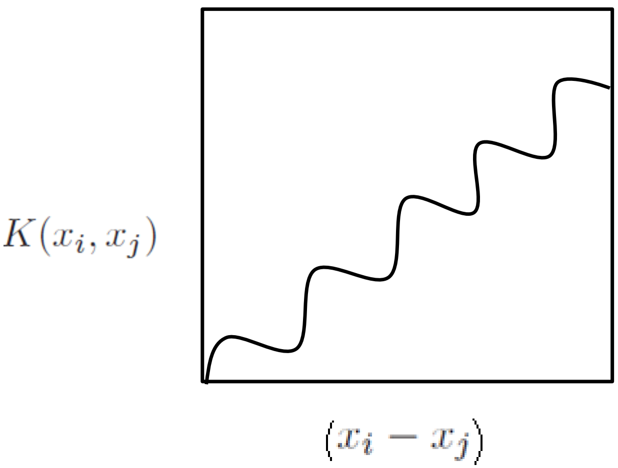


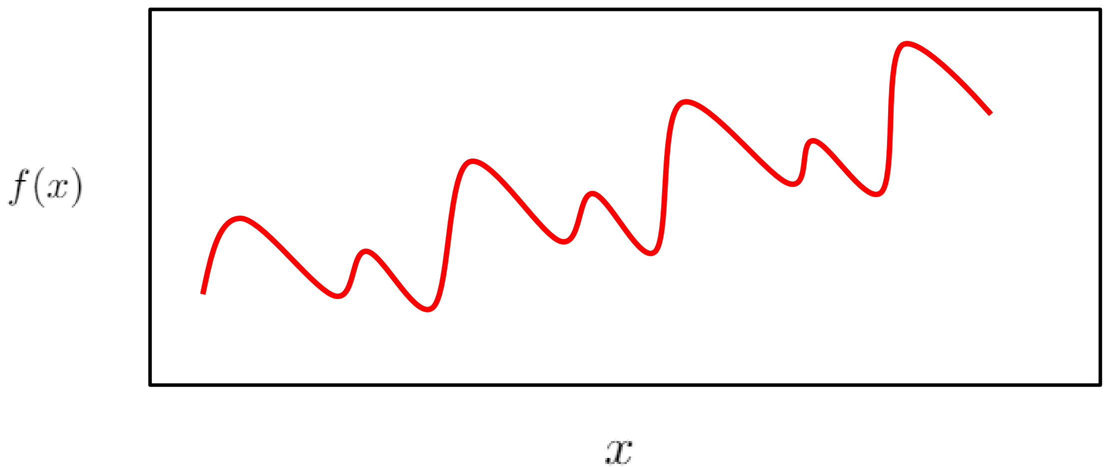


Figures from: Prediction of tidal currents using Bayesian machine learning, Ocean Engineering, 2018.

Combining kernels

Linear + Periodic: $K(x_i, x_j) = K_{\text{linear}}(x_i, x_j) + K_{\text{periodic}}(x_i, x_j)$





Procedure

Dataset

$$\mathbf{X} = \left[\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}\right] \qquad \mathbf{y} = \left[y^{(1)}, y^{(2)}, ..., y^{(N)}\right]$$

Model

$$y^{(i)} = f\left(\mathbf{x}^{(i)}\right) + \epsilon^{(i)} \qquad \epsilon^{(i)} \sim \mathcal{N}(0, \sigma_N^2)$$

Prior

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{m}, \mathbf{K})$$

Likelihood

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma_N^2 \mathbf{I})$$

f posterior

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\mathrm{d}\mathbf{f}}$$

Procedure

- Want to make prediction at M points:

$$\mathbf{X}_* = \left[\mathbf{x}_*^{(1)}, \mathbf{x}_*^{(2)}, ..., \mathbf{x}_*^{(M)}\right]$$

- Let $\mathbf{f}_*$ be the vector of latent function values at $\mathbf{X}_*$.

- Joint distribution of $\mathbf{f}$ and $\mathbf{f}_*$:

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

- Conditional distribution of $\mathbf{f}_*$ given $\mathbf{f}$:

$$p(\mathbf{f}_*|\mathbf{f}) = \mathcal{N}\big(\mathbf{m}_* + \mathbf{K}(\mathbf{X}_*,\mathbf{X})\mathbf{K}(\mathbf{X},\mathbf{X})^{-1}(\mathbf{f}-\mathbf{m}),$$
$$\mathbf{K}(\mathbf{X}_*,\mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*,\mathbf{X})\mathbf{K}(\mathbf{X},\mathbf{X})^{-1}\mathbf{K}(\mathbf{X},\mathbf{X}_*)\big)$$

- Compute $p(\mathbf{f}_*|\mathbf{y})$ as

$$p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}_*|\mathbf{f})p(\mathbf{f}|\mathbf{y})\mathrm{d}\mathbf{f}$$

Posterior distribution

- Compute $p(\mathbf{f}_*|\mathbf{y})$ as

$$p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}_*|\mathbf{f})p(\mathbf{f}|\mathbf{y})\mathrm{d}\mathbf{f}$$

- Posterior distribution:

$$\begin{aligned} p(\mathbf{y}^*|\mathbf{y}) &= \int p(\mathbf{y}^*|\mathbf{f}^*)p(\mathbf{f}^*|\mathbf{y})\mathrm{d}\mathbf{f}^* \\ &= \mathcal{N}(\boldsymbol{\mu}, \sigma^2) \end{aligned}$$

where

$$\boldsymbol{\mu} = \mathbf{m}_* + \mathbf{K}(\mathbf{X}_*, \mathbf{X})\big(\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_N^2\mathbf{I}\big)^{-1}(\mathbf{y} - \mathbf{m})$$
$$\sigma^{\mathbf{2}} = \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*, \mathbf{X})\big(\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_N^2\mathbf{I}\big)^{-1}\mathbf{K}(\mathbf{X}, \mathbf{X}_*)$$

Learning with Gaussian process

- Any kernel function has a number of parameters (hyperparameters) which are unknown. For example, in the exponentiated quadratic kernel function

$$K(x_i, x_j) = \sigma_f^2 \exp\left(- \frac{|x_i - x_j|^2}{2l^2} \right)$$

  the hyperparameters are the variance σ_f^2 and lengthscale l.

- Also the parameters of the likelihood function are unknown.
- Jointly representing these hyperparamters as $\boldsymbol{\theta}$.
- Learning with Gaussian process is equivalent to learning these hyperparameters.
- Inference can be made once the hyperparameters are learnt.

Learning with Gaussian process

- The derived posterior distribution is actually function of **unknown** hyperparameters – $p(\mathbf{y}_*|\mathbf{y}, \boldsymbol{\theta})$, and they to be tackled.
- Marginalization of the hyperparameters:

$$p(\mathbf{y}_*|\mathbf{y}) = \int p(\mathbf{y}_*|\mathbf{y}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta}$$

where $p(\boldsymbol{\theta}|\mathbf{y})$ is the posterior distribution of the hyperparameters.

- Employing Bayes theorem on $p(\boldsymbol{\theta}|\mathbf{y})$ we get

$$p(\mathbf{y}_*|\mathbf{y}) = \frac{\int p(\mathbf{y}_*|\mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}{\int p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}$$

- Finding a solution to the intractable integrals is one of the major challenges in GP.
- Two well known approaches of determining an approximate solution:
  - Maximum likelihood estimation (MLE)
  - Maximum a-posteriori (MAP) approach

Maximum Likelihood Estimation

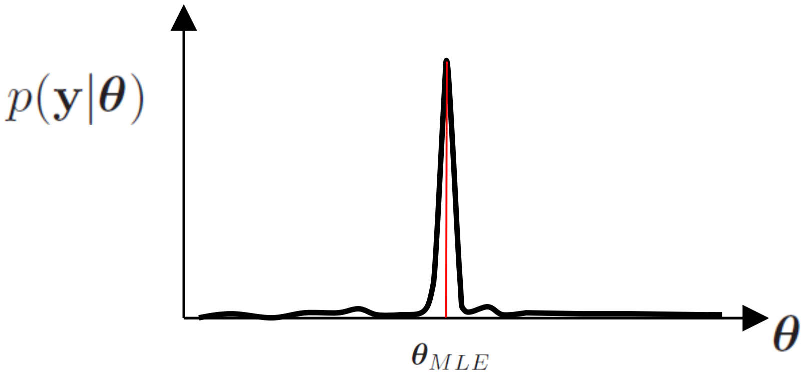
- Evaluate:

$$p(\mathbf{y}_*|\mathbf{y}) = \frac{\int p(\mathbf{y}_*|\mathbf{y},\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}$$

- Approximations:

  - $p(\mathbf{y}|\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{MLE})$ where

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta})$$



Maximum Likelihood Estimation

- Evaluate:

$$p(\mathbf{y}_*|\mathbf{y}) = \frac{\int p(\mathbf{y}_*|\mathbf{y},\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}$$

- Approximations:
  - $p(\mathbf{y}|\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{MLE})$ where

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta})$$

  - $p(\boldsymbol{\theta}) = c$

- On substitution of the approximations in we get

$$p(\mathbf{y}_*|\mathbf{y}) = \frac{\int p(\mathbf{y}_*|\mathbf{y},\boldsymbol{\theta})\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{MLE})c\mathrm{d}\boldsymbol{\theta}}{\int \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{MLE})c\mathrm{d}\boldsymbol{\theta}}$$

$$= p(\mathbf{y}_*|\mathbf{y},\boldsymbol{\theta}_{MLE})$$

Maximizing log-likelihood

- Convention: Maximize $\log p(\mathbf{y}|\boldsymbol{\theta})$ instead of $p(\mathbf{y}|\boldsymbol{\theta})$.
  - $\log()$ is a monotonically increasing function, so the maximum of log-likelihood is the maximum of likelihood.
- Eventually we have

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = \underbrace{-\frac{1}{2}(\mathbf{y}-\mathbf{m})^T(\mathbf{K}+\sigma_N^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{m})}_{\boxed{1}} \underbrace{-\frac{1}{2}\log|\mathbf{K}+\sigma_N^2\mathbf{I}|}_{\boxed{2}} - \frac{N}{2}\log 2\pi$$

- $\boxed{1}$ penalizes the mismatch between data and prediction.
- $\boxed{2}$ penalizes the model complexity.

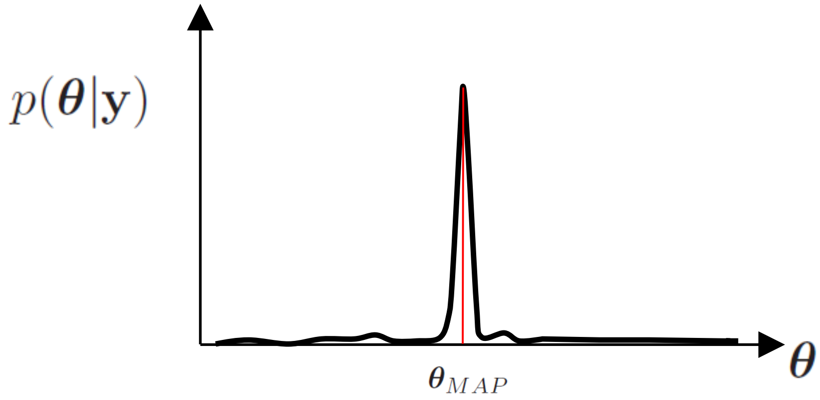
Maximum a-posteriori

- Evaluate:

$$p(\mathbf{y}_*|\mathbf{y}) = \int p(\mathbf{y}_*|\mathbf{y}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}$$

- Approximation:
  - $p(\boldsymbol{\theta}|\mathbf{y}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})$ where

$$\boxed{\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y})}$$



Maximum a-posteriori

- Evaluate:

$$p(\mathbf{y}_*|\mathbf{y}) = \int p(\mathbf{y}_*|\mathbf{y}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta} \quad \text{-------------------} \quad \color{red}{\blacksquare}$$

- Approximation:
  - $p(\boldsymbol{\theta}|\mathbf{y}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})$ where

$$\boxed{\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y})}$$

- On substitution of the approximations in $\color{red}{\blacksquare}$ we get

$$p(\mathbf{y}_*|\mathbf{y}) = \int p(\mathbf{y}_*|\mathbf{y}, \boldsymbol{\theta}) \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP}) \mathrm{d}\boldsymbol{\theta}$$

$$= p(\mathbf{y}_*|\mathbf{y}, \boldsymbol{\theta}_{MAP})$$

Maximizing hyperparameter posterior

- Convention: Maximize the log of the hyperparameter posterior:

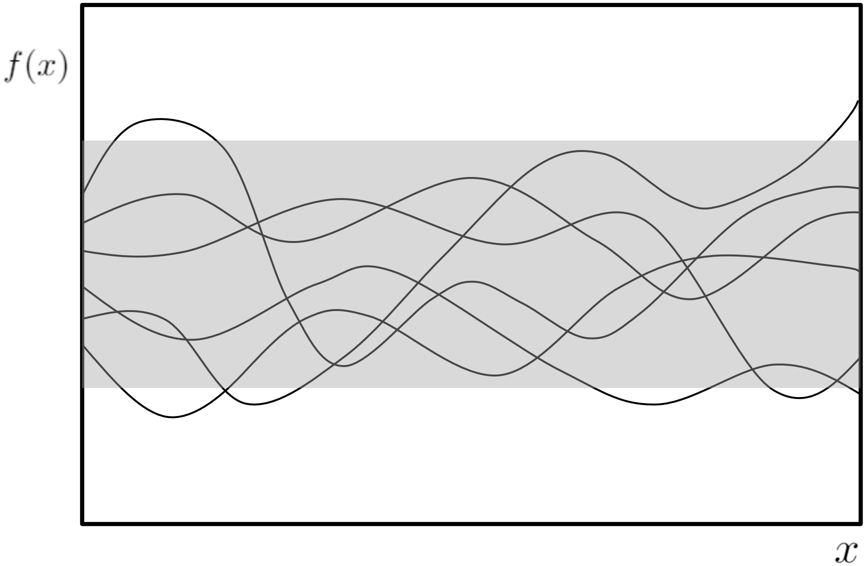
$$\log p(\boldsymbol{\theta}|\mathbf{y}) = -\frac{1}{2}(\mathbf{y}-\mathbf{m})^T(\mathbf{K}+\sigma_N^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{m}) - \frac{1}{2}\log|\mathbf{K}+\sigma_N^2\mathbf{I}| - \frac{N}{2}\log 2\pi + \log p(\boldsymbol{\theta})$$

- The main difference between maximizing $\log p(\boldsymbol{\theta}|\mathbf{y})$ and $\log p(\mathbf{y}|\boldsymbol{\theta})$ is the prior term $\log p(\boldsymbol{\theta})$.

- $p(\boldsymbol{\theta})$ can be used to represent our prior belief/knowledge of hyperparameter values.

Summary

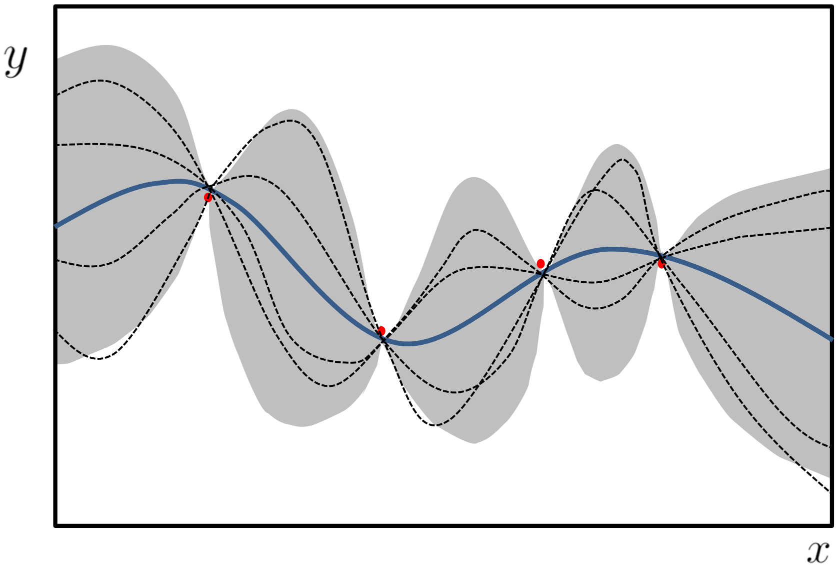
Prior



$$f \sim \mathcal{GP}(m, k)$$

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{m}, \mathbf{K})$$

Posterior



$$\boldsymbol{\mu} = \mathbf{m}_* + \mathbf{K}(\mathbf{X}_*, \mathbf{X})\big(\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_N^2 \mathbf{I}\big)^{-1}(\mathbf{y} - \mathbf{m})$$

$$\boldsymbol{\sigma}^2 = \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*, \mathbf{X})\big(\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_N^2 \mathbf{I}\big)^{-1}\mathbf{K}(\mathbf{X}, \mathbf{X}_*)$$

Approximate inference techniques

- Computations become intractable when using non-Gaussian likelihood function $p(\mathbf{y}|\mathbf{f})$.
  - In such a case, closed form expressions of the $\mathbf{f}$-posterior $p(\mathbf{f}|\mathbf{y})$ and marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ are not available.
- Exact inference is not possible and approximate inference techniques need to be used.
- Approaches:
  - Approximate deterministic inference:
    * Laplace approximation
    * Expectation propagation
    * Variational methods
  - Approximate sampling inference:
    * Markov chain Monte Carlo (MCMC) sampling

Example

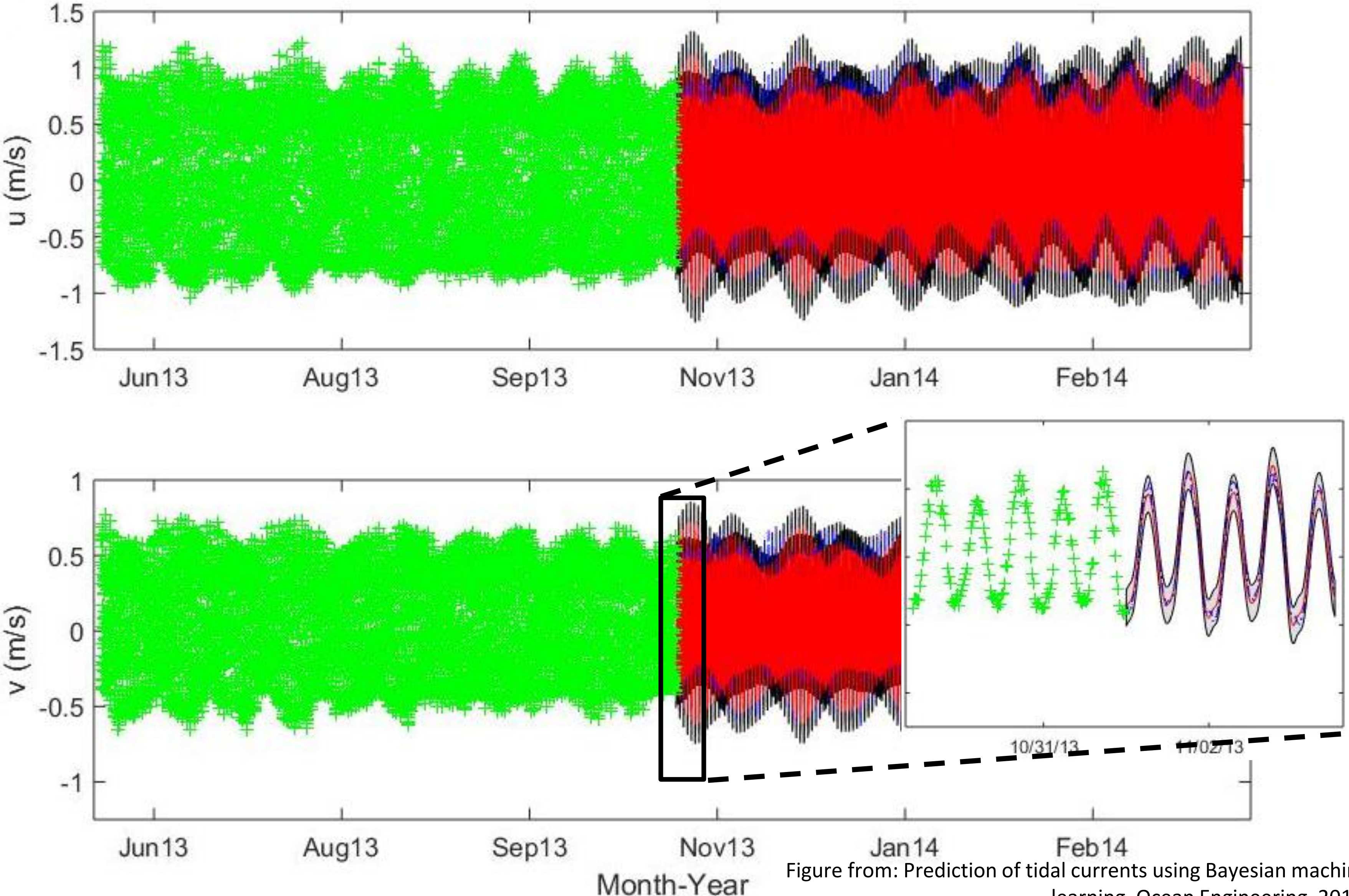


Figure from: Prediction of tidal currents using Bayesian machine learning, Ocean Engineering, 2018.