Gaussian Processes: Master Normality

PhD. Juan Ungredda, ESTECO SpA

Content

- Multivariate Gaussian Distribution
- Gaussian process
- Gaussian process regression
- Hyperparameter optimization
- Benefits & Difficulties

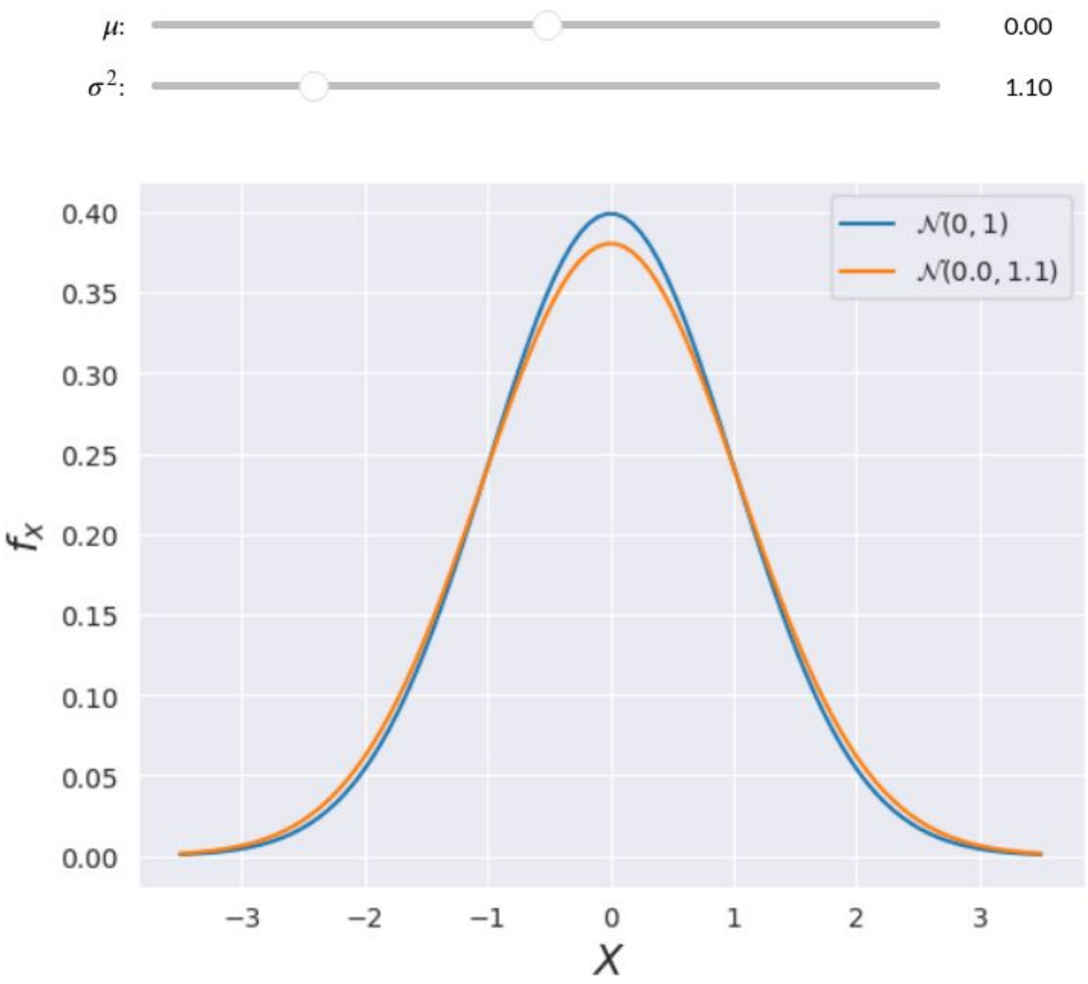
Univariate Gaussian Density

A random variable x with density $x \sim \mathcal{N}(\mu, \sigma^2)$

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Where:

- x is the random variable.
- $\mu = \mathbb{E}[X]$ is the mean, representing the central tendency of the distribution.
- $\sigma^2 = \mathrm{Var}(X)$ is the variance, determining the spread or dispersion of the distribution.


μ:
0.00
σ²:
1.10
𝒩(0, 1)
𝒩(0.0, 1.1)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
f_X
−3
−2
−1
0
1
2
3
X

Multivariate Gaussian Distribution

A random variable $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ presents a density function

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Where:

- $\mathbf{x}$: d-dimensional vector representing the random variables.
- $\boldsymbol{\mu}$: mean vector, representing the expected value of each random variable.
- $\boldsymbol{\Sigma}$: covariance matrix, represents the relationships between random variables.

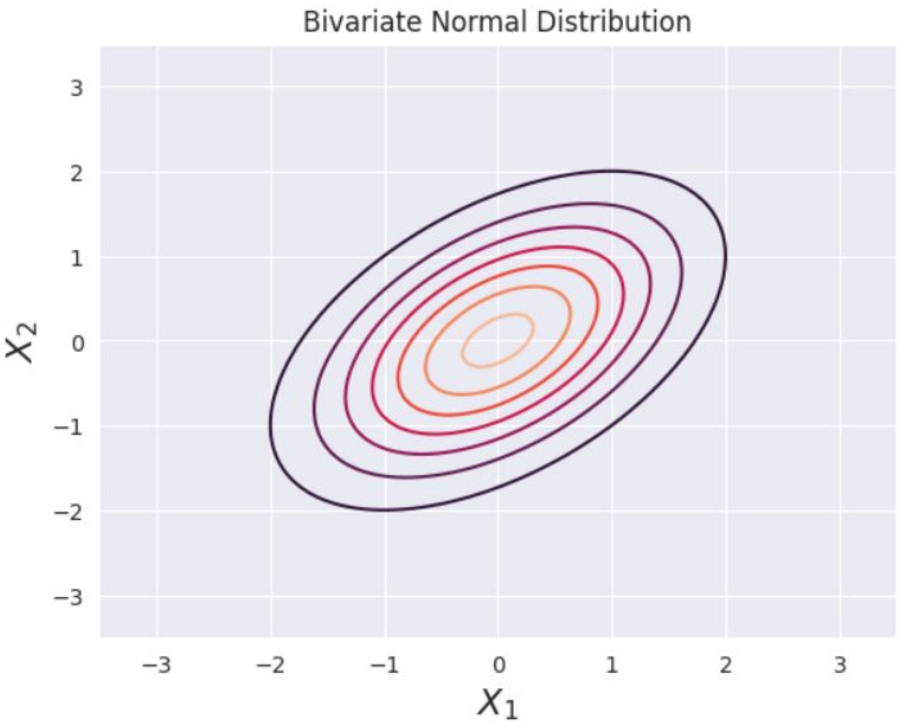
μ_1: 0.00

μ_2: 0.00

$Var(x_1)$: 1.00

$Var(x_2)$: 1.00

$Cov(x_1, x_2)$: 0.50



Covariance Matrix

For n random variables $X_1, X_2, \ldots, X_n$, the multivariate covariance matrix Σ is:

$$\Sigma = \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_2) & \ldots & \mathrm{Cov}(x_1, x_n) \\ \mathrm{Cov}(x_2, x_1) & \mathrm{Var}(x_2) & \ldots & \mathrm{Cov}(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_n, x_1) & \mathrm{Cov}(x_n, x_2) & \ldots & \mathrm{Var}(x_n) \end{bmatrix}$$

- **Symmetry**: $\mathrm{cov}(x, x') = \mathrm{cov}(x', x)$ for all x and x'.
- **Positive Semi-definite**: $x^T \Sigma x \geq 0$ for any vector $x \neq 0$.

Conditioning

Given the two random vectors $\mathbf{x}_A$ and $\mathbf{x}_B$, the conditional probability of $\mathbf{x}_A$ is defined as,

$$p(\mathbf{x}_A|\mathbf{x}_B) = \frac{p(\mathbf{x}_A, \mathbf{x}_B)}{p(\mathbf{x}_B)}$$

defined for $p(\mathbf{x}_B) > 0$

Exercise: Gaussian Conditioning

Assume an n-dimensional random vector has a normal distribution,

$$N\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}, \begin{bmatrix}\mu_X\\ \mu_Y\end{bmatrix}, \begin{bmatrix}A & C\\ C^T & B\end{bmatrix}\right)$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are two subvectors of respective dimensions p and q with $p+q=n$. Then, conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is also normal with mean vector

$$\mu_{\mathbf{y}|\mathbf{x}} = \mu_Y + C^T A^{-1}(\mathbf{x} - \mu_X)$$

and covariance matrix

$$\Sigma_{\mathbf{y}|\mathbf{x}} = B - C^T A^{-1} C$$

Proof:

The joint density of $\mathbf{x}$ is:

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}Q(\tilde{\mathbf{x}})\right]$$

where Q is defined as

$$Q(\tilde{\mathbf{x}}) = (\tilde{\mathbf{x}} - \tilde{\mu})^T \Sigma^{-1} (\tilde{\mathbf{x}} - \tilde{\mu}) = [(\mathbf{x} - \mu_X)^T, (\mathbf{y} - \mu_Y)^T] \begin{bmatrix} A & C \\ C^T & B \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mu_X \\ \mathbf{y} - \mu_Y \end{bmatrix}$$

Here we have assumed

$$\Sigma^{-1} = \begin{bmatrix} \tilde{A} & \tilde{C} \\ \tilde{C^T} & \tilde{B} \end{bmatrix}$$

where

$$\tilde{A} = (A - CB^{-1}C^T)^{-1}C^T A^{-1}$$

$$\tilde{B} = (B - C^T A^{-1} C)^{-1} C B^{-1}$$

$$\tilde{C} = -A^{-1}C(B - C^T A^{-1} C)^{-1} = \tilde{C^T}$$

Substituting into $Q(\tilde{\mathbf{x}})$ to get:

$$Q(\tilde{\mathbf{x}}) = (\mathbf{x} - \mu_X)^T A^{-1} (\mathbf{x} - \mu_X) + [(\mathbf{y} - \mu_Y) - C^T A^{-1} (\mathbf{x} - \mu_X)]^T (B - C^T A^{-1} C)^{-1} [(\mathbf{y} - \mu_Y) - C^T A^{-1} (\mathbf{x} - \mu_X)]$$

Now the joint distribution can be written as:

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi)^{n/2} |A|^{1/2}} \exp\left[-\frac{1}{2} Q(\tilde{\mathbf{x}}) \right] = N(\mathbf{x} | \mu_X, A) \cdot N(\mathbf{y} | b, M)$$

Now the joint distribution can be written as:

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi)^{n/2}|A|^{1/2}} \exp\left[-\frac{1}{2}Q(\tilde{\mathbf{x}})\right] = N(\mathbf{x}|\mu_X, A) \cdot N(\mathbf{y}|b, M)$$

The conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})} \\ &= \frac{1}{(2\pi)^{q/2}|M|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - b)^T M^{-1}(\mathbf{y} - b)\right] \end{aligned}$$

with

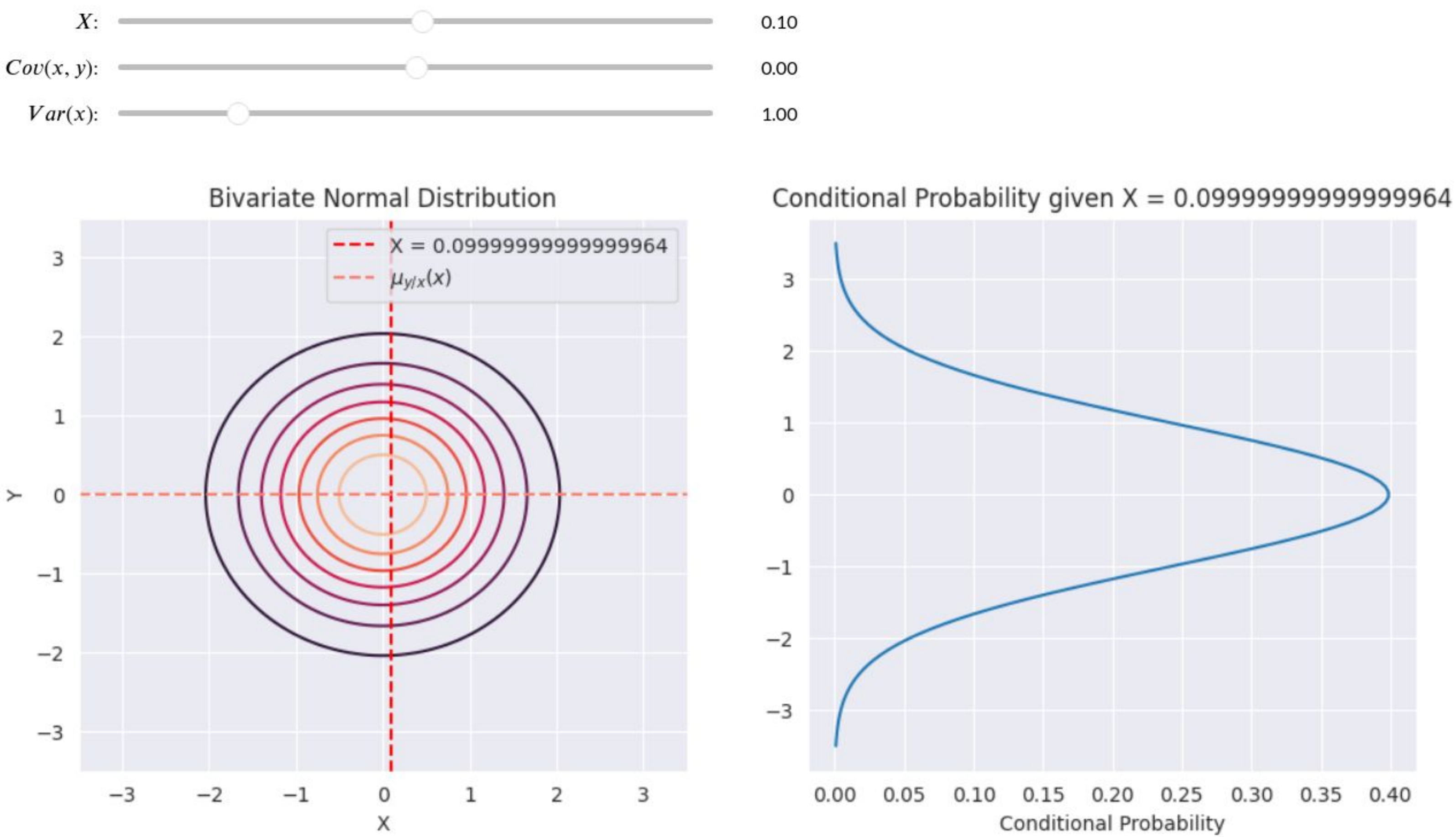
$$b = \mu_Y + C^T A^{-1}(\mathbf{x} - \mu_X)$$

$$M = B - C^T A^{-1} C$$

Consider $n = 2$, then,

$$b = \mu_Y + \frac{Cov(x, y)}{Var(x)}(x - \mu_X)$$

$$M = Var(y) - \frac{Cov(x, y)^2}{Var(x)}$$



X: 0.10
Cov(x, y): 0.00
Var(x): 1.00
Bivariate Normal Distribution
X = 0.09999999999999964
μy/x(x)
X
Y
Conditional Probability given X = 0.09999999999999964
Conditional Probability

Marginalisation

Given the two random vectors $\mathbf{x}_A$ and $\mathbf{x}_B$, the marginal probability of $\mathbf{x}_A$ is given by,

$$p(\mathbf{x}_A) = \int p(\mathbf{x}_A, \mathbf{x}_B) d\mathbf{x}_B$$

Exercise: Gaussian Marginalisation

Let $\mathbf{x}$ and $\mathbf{y}$ be jointly Gaussian random vector with dimension m and n, respectively.

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}\right)$$

show that $x \sim \mathcal{N}(\mu_X, A)$

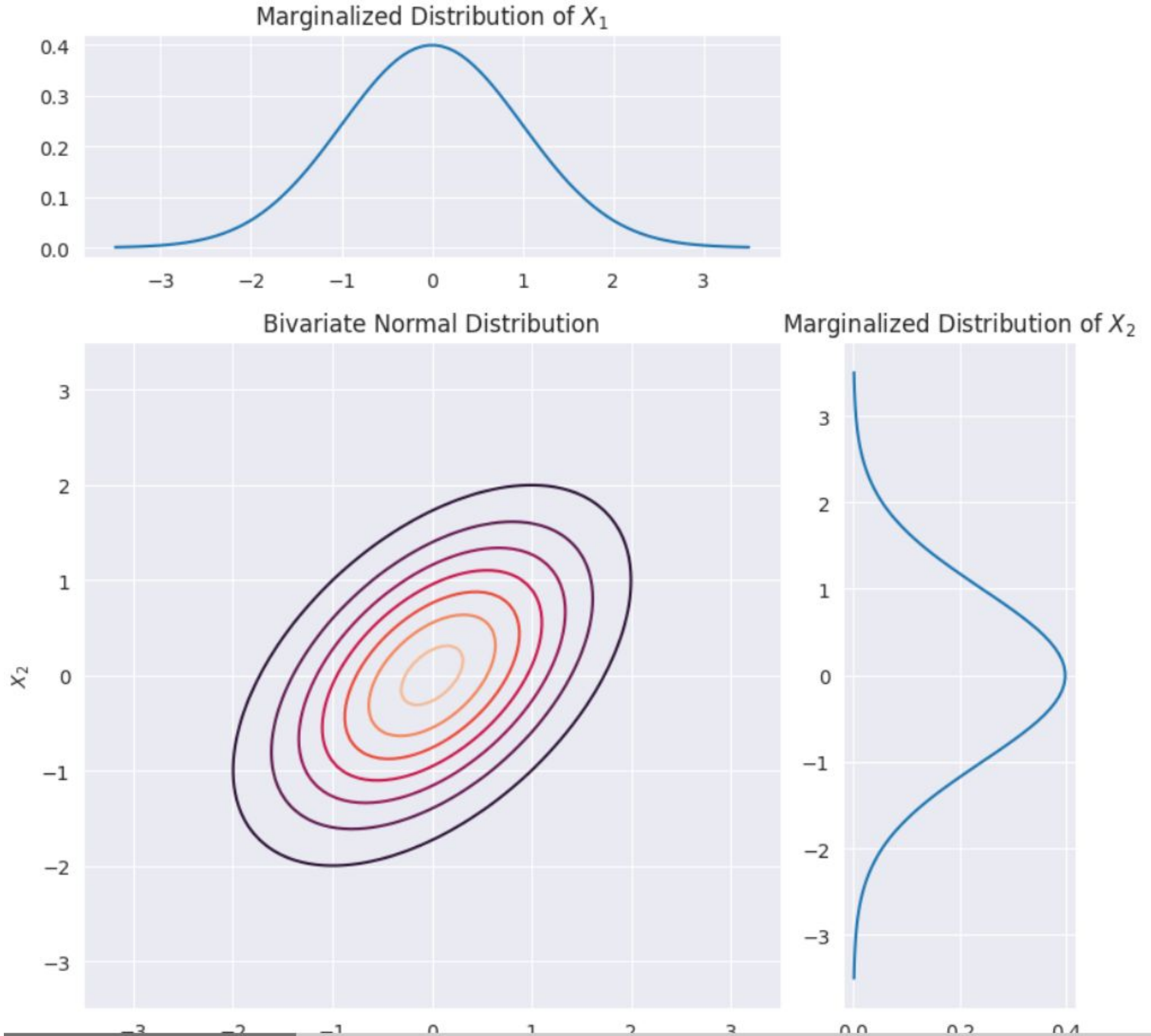
$\mathbf{Solution}$:

$$\mathbf{x} = A\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} = A\left(\begin{bmatrix}\mu_X\\ \mu_Y\end{bmatrix} + MZ\right) = A\begin{bmatrix}\mu_X\\ \mu_Y\end{bmatrix} + AMZ$$

where,

$$A = [I_{m,m}, \mathbf{0}_{m,n}]$$

Therefore, $\mathbf{x}$ is normally distributed with $\mathbb{E}[\mathbf{x}] = A\begin{bmatrix}\mu_X\\ \mu_Y\end{bmatrix} = \mu_X$ and $Cov(\mathbf{x}) = A\begin{bmatrix}A & C\\ C^T & B\end{bmatrix}A^T = A$.



Marginalized Distribution of X_1
Bivariate Normal Distribution
Marginalized Distribution of X_2
X_2

Gaussian Normal Samples

Given a Cholesky decomposition of the covariance matrix to obtain the lower triangular matrix L,

$$\Sigma = LL^T$$

Then, you can generate samples from a standard normal distribution as:

$$\mathbf{x} = \mu + L\mathbf{z}$$

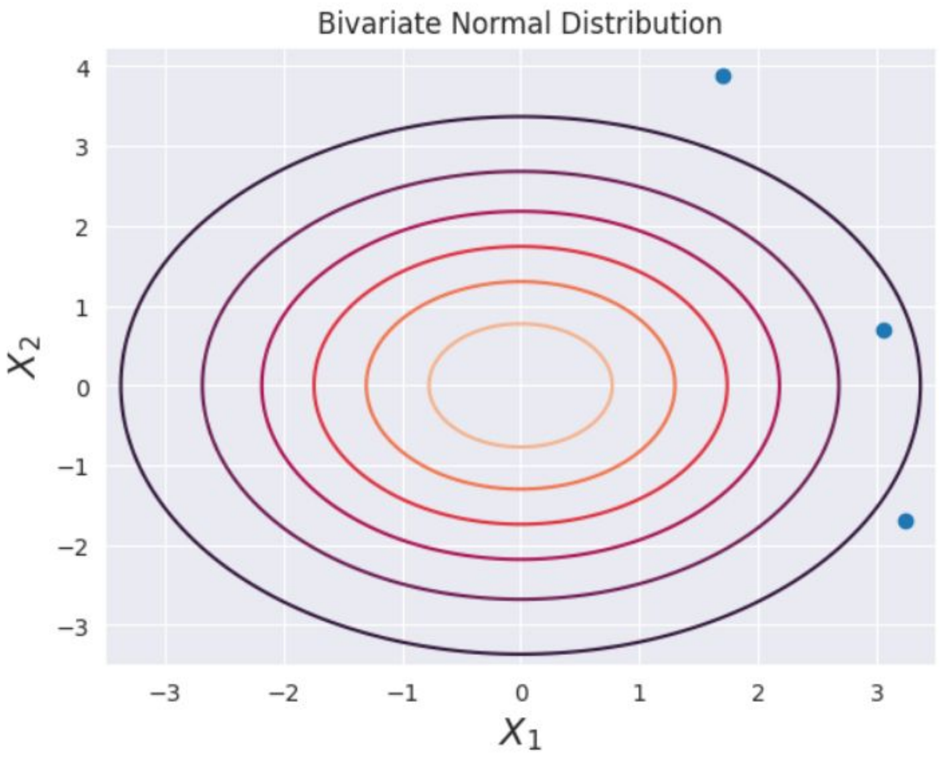
where $\mathbf{z} \sim N(\mathbf{0}, I)$. For a single dimension we have, $x = \mu_x + \sigma_x z$

n 3

$Var(x_1)$ 3.00

$Var(x_2)$ 3.00

$Cov(x_1, x_2)$ 0.00



Summary

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \Sigma = \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}\right)$$

Marginalisation & Conditioning:

$$\mathbf{x} \sim \mathcal{N}(\mu_x, A)$$

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}\left(\mu_x + CB^{-1}(\mathbf{y} - \mu_y), A - CB^{-1}C^T\right)$$

Sampling:

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \mu + L\mathbf{z}, \text{ where } \Sigma = LL^T$$

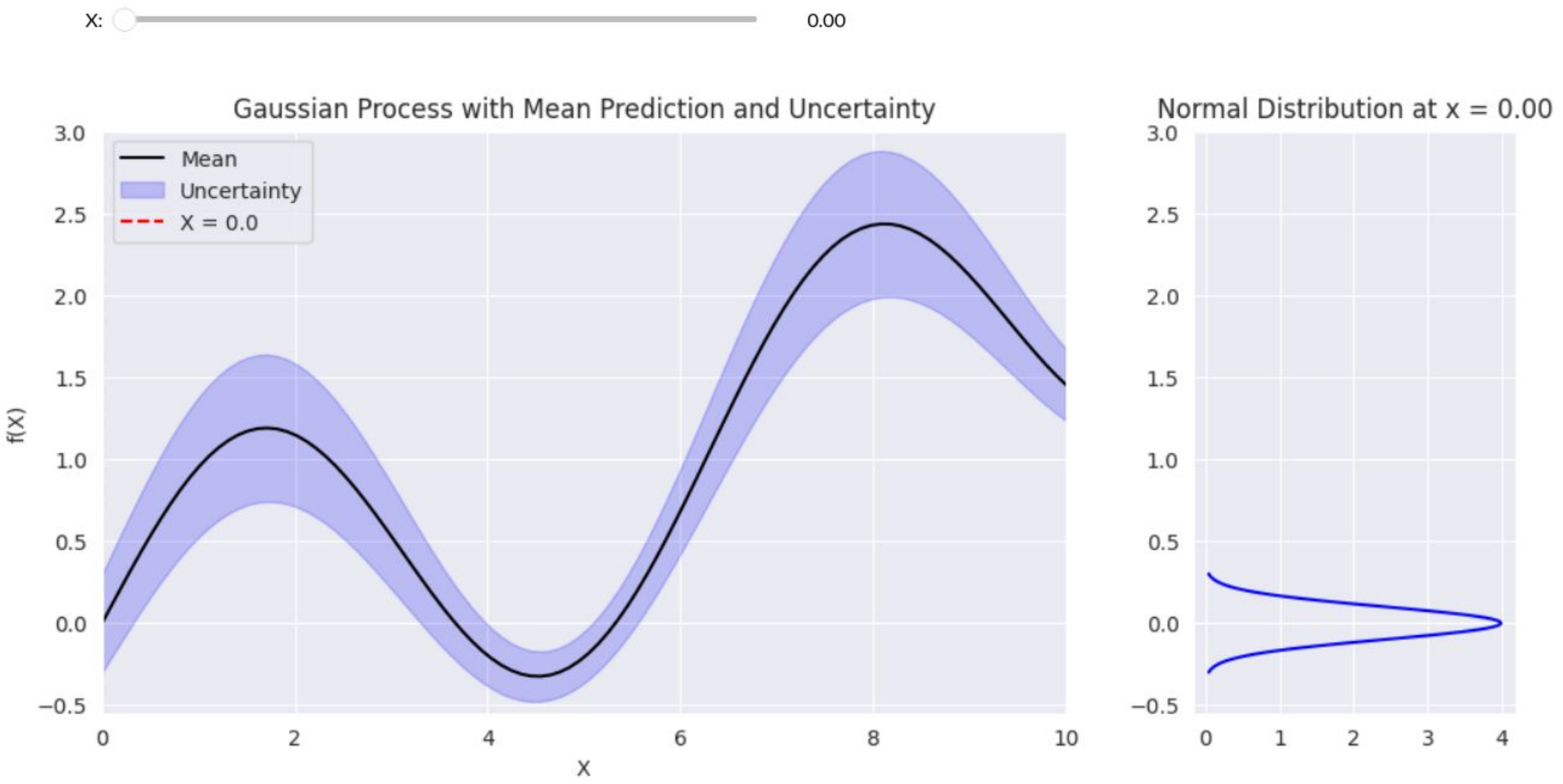
Gaussian Process

It is a collection of random variables, where any finite number of variables have a joint Gaussian distribution. A Gaussian process (GP) is defined by its mean function $m(x)$ and covariance function $k(x, x')$ as,

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

-**Mean Function**: $m(x) = \mathbb{E}[f(x)]$

-**Covariance Function**: $k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x) - m(x'))]$



X:
0.00
Gaussian Process with Mean Prediction and Uncertainty
Mean
Uncertainty
X = 0.0
f(X)
X
Normal Distribution at x = 0.00

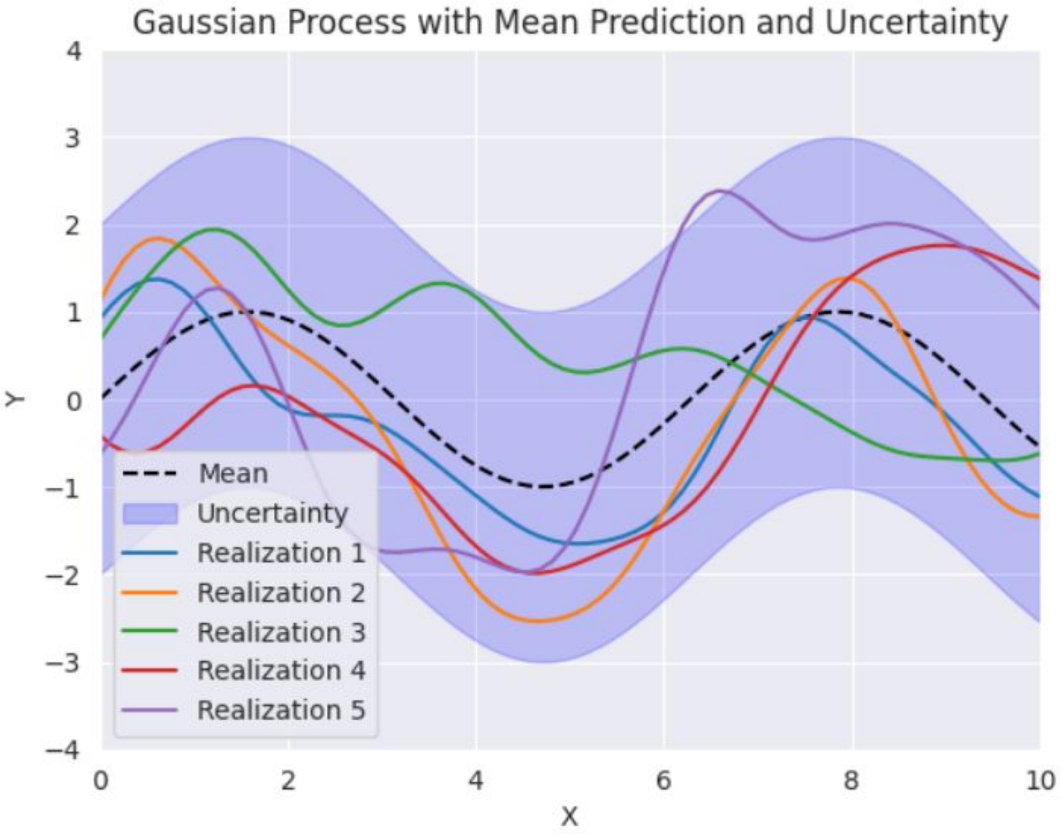
Mean Function

The mean function represents the expected value of the process at any given point.

- **Zero Mean Function**: The simplest assumption is to assume that it is zero everywhere, i.e., $m(x) = 0$ for all x.
- **Non-Zero Mean Function**: Prior knowledge, basis functions, etc.

Sin Mean Function

Linear Mean Function



Covariance Function

Define:

- Similarity/correlation between data points
- Smoothness & Periodicity

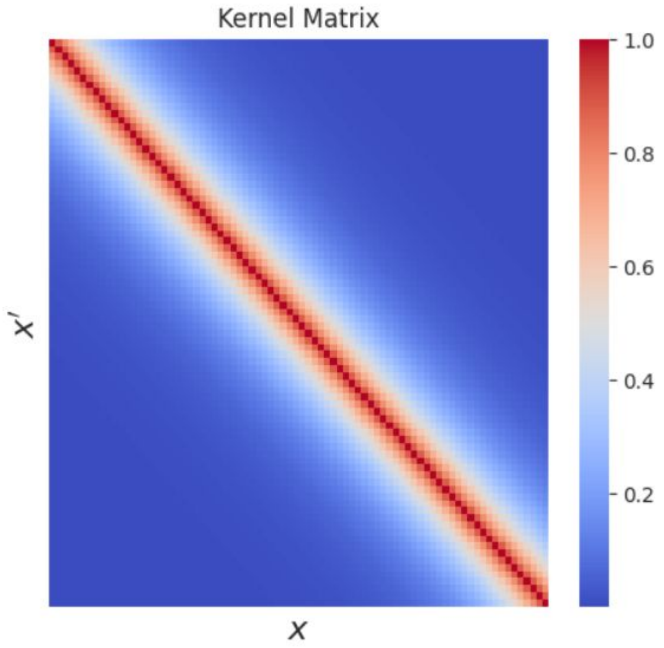
Properties:

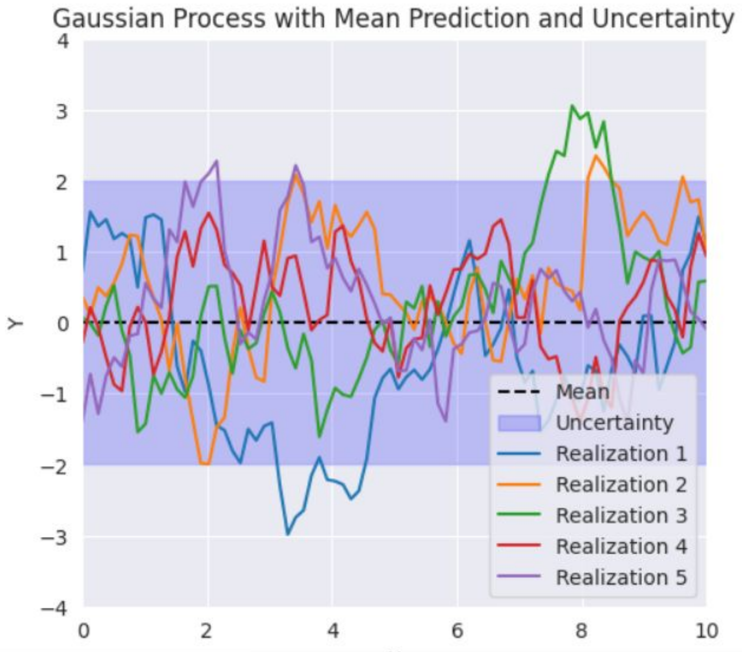
- Symmetric & Positive Semi-definite

Gaussian Kernel	l	1	σ^2	1
Linear Kernel	σ_v^2	0.	σ_b^2	1
White Noise Kernel	σ^2	1		

Matérn Kernel:

$$k_{\text{Matern}}(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{l} |x - x'| \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{l} |x - x'| \right)$$





Combining Kernels

- $\mathbf{Summing\ Kernels}$: The resulting covariance allows to capture various patterns simultaneously.

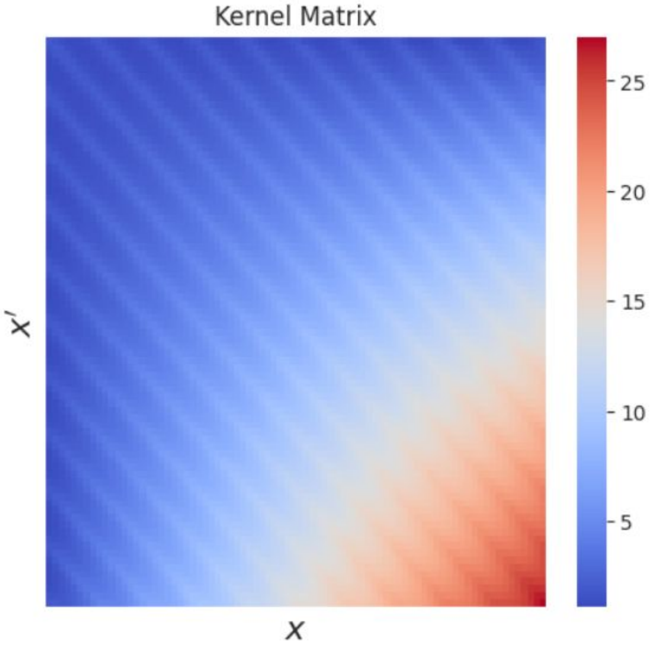
$$k_{\text{sum}}(x, x') = k_1(x, x') + k_2(x, x') + \cdots + k_n(x, x')$$

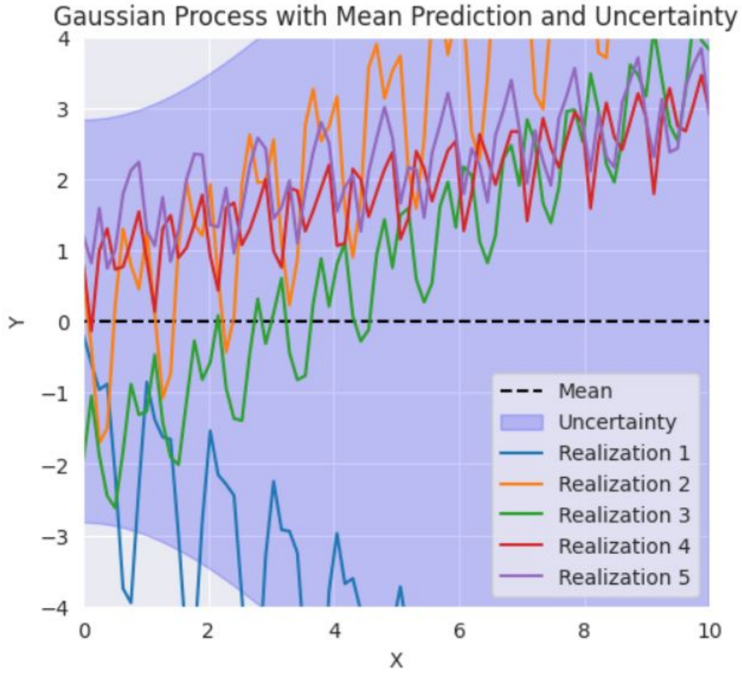
- $\mathbf{Multiplying\ Kernels}$: This approach is useful for modeling interactions between different patterns present in the data.

$$k_{\text{mult}}(x, x') = k_1(x, x') \times k_2(x, x') \times \cdots \times k_n(x, x')$$

Operation addition

$$Operation(k_{Linear}, k_{Periodic})$$





Predictive Distribution

Given the noise-free observations $\mathbf{f}$, the joint distribution of observed locations and test points X and X^* is,

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

The predictive distribution is obtained by conditioning on the observed data:

$$\mathbf{f}^*|\mathbf{f} \sim \mathcal{N}(\mu^*, \Sigma^*)$$

$$\mu^* = \mathbf{m}^* + K(X^*,X)^T K(X,X)^{-1}(\mathbf{f} - \mathbf{m})$$

$$\Sigma^* = K(X^*,X^*) - K(X^*,X)^T K(X,X)^{-1} K(X,X^*)$$

Predictive Distribution using Noisy Observation

Consider, $y(x) = f(x) + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma_\nu^2)$. Therefore,

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m} \\ \mathbf{m}^* \end{bmatrix}, \begin{bmatrix} K(X,X) + \sigma_\nu^2 & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix} \right)$$

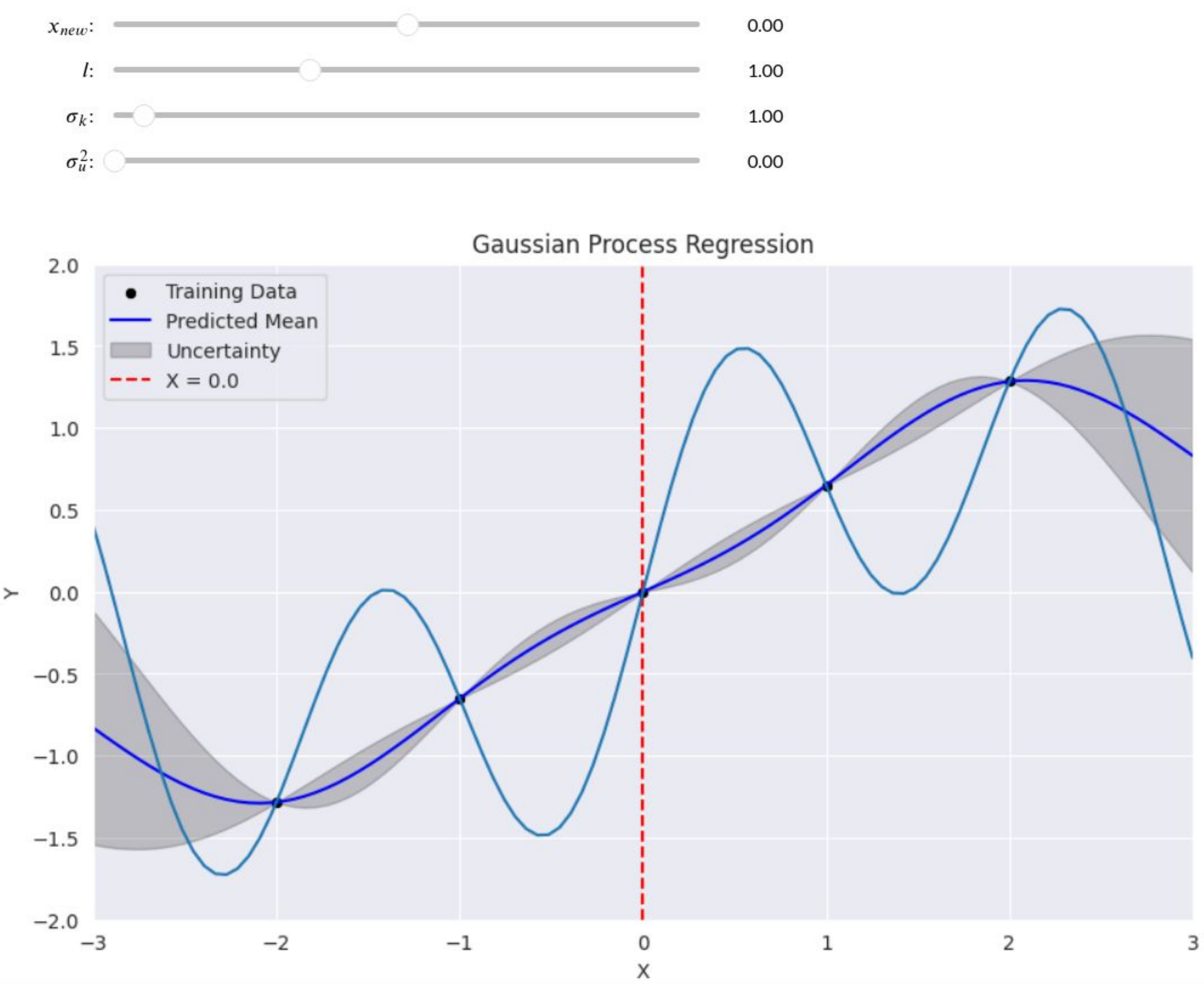
and,

$$\mathbf{f}^* | \mathbf{y} \sim \mathcal{N}(\mu^*, \Sigma^*)$$

$$\mu^* = \mathbf{m}^* + K(X^*,X)^T [K(X,X) + \sigma_\nu^2 I]^{-1} (\mathbf{y} - \mathbf{m})$$

$$\Sigma^* = K(X^*,X^*) - K(X^*,X)^T [K(X,X) + \sigma_\nu^2 I]^{-1} K(X,X^*)$$

ESTECO



x_{new}: 0.00
l: 1.00
σ_k: 1.00
σ_u^2: 0.00
Gaussian Process Regression
Training Data
Predicted Mean
Uncertainty
X = 0.0
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Y
−3
−2
−1
0
1
2
3
X

Learning Hyperparameters

Given the marginal likelihood of the observed data.

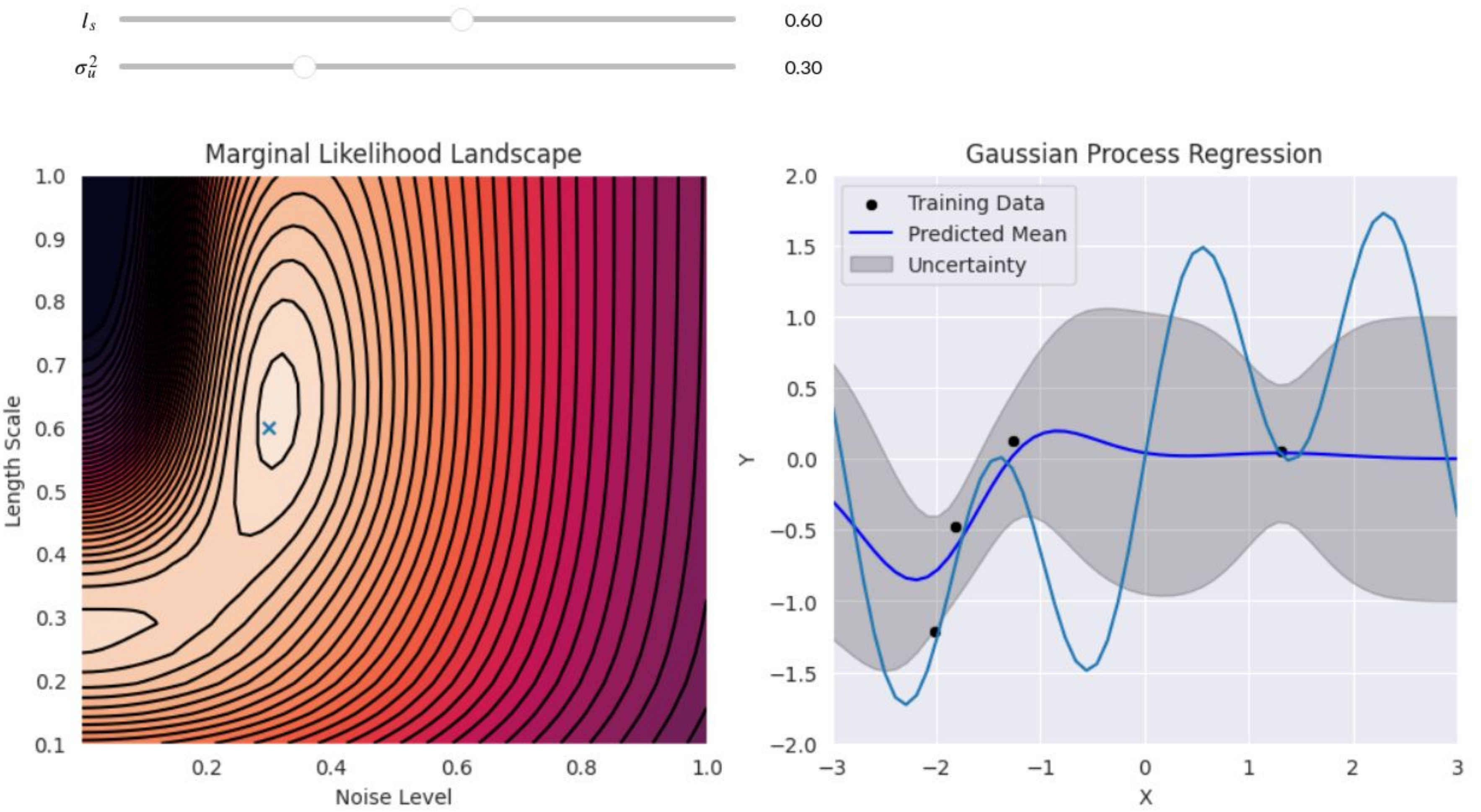
$$p(\mathbf{f}^*|X, \mathbf{y}, X^*, \theta) = \frac{p(\mathbf{y}, \mathbf{f}^*|X^*, X, \theta)}{\underbrace{p(\mathbf{y}|\mathbf{X}, \theta)}_{\text{marginal likelihood}}} = \frac{p(\mathbf{y}, \mathbf{f}^*|X^*, X, \theta)}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X, \theta)d\mathbf{f}}$$

where,

$$\log p(y|X, \theta) = -\frac{1}{2}\left(\mathbf{y}^\top (K_\theta(X, X) + \sigma_n^2 I)^{-1}\mathbf{y} + \log|K_\theta(X, X) + \sigma_n^2 I| + n\log(2\pi)\right)$$

Aim:

$$\theta^* = \operatorname{argmax}_\theta \left(-\log p(\mathbf{y}|X, \theta)\right)$$


l_s
0.60
σ_u^2
0.30
Marginal Likelihood Landscape
Length Scale
Noise Level
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.2
0.4
0.6
0.8
1.0
Gaussian Process Regression
Training Data
Predicted Mean
Uncertainty
Y
X
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−3
−2
−1
0
1
2
3

Benefits

- **Flexibility**: GPs can model complex relationships between inputs and outputs without imposing a specific functional form.
- **Uncertainty Estimation**: GPs provide not only point predictions but also estimate uncertainty in predictions.

Difficulties

- Presents a $\mathcal{O}(n^3)$ computational cost and $\mathcal{O}(n^2)$ in memory.
  - Sparse approximation.
- Challenging in high number of dimensions.
  - Structural assumptions
- The marginal likelihood is often multi-modal, i.e, local optima.
  - Random start points, using prior distributions, marginalise over hyperparameters.