Gaussian Process

Jee-Hyong Lee

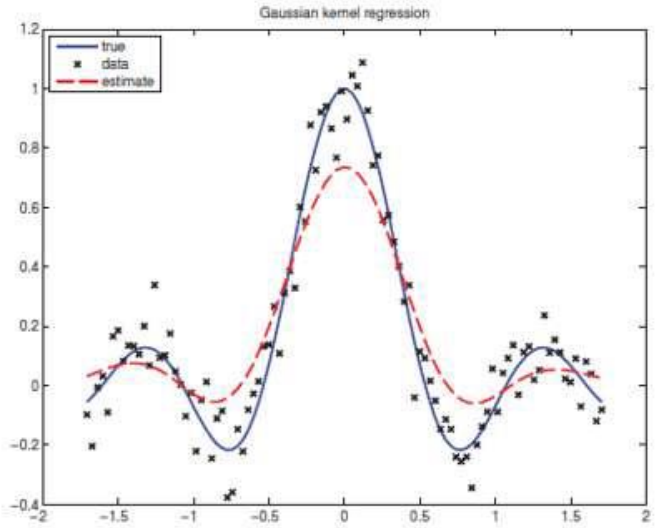
Sungkyunkwan Univ.

Introduction

- **Parametric Models vs. Non-parametric Models**
  - Non-parametric model은 parametric model 과 다르게 model 의 구조를 가정하지 않고,데이터로부터 모든 것을 알아낸다.
- **Parametric models:**
  - Linear Regression
  - GMM
- **Non-parametric models:**
  - KNN
  - Kernel Regression
  - Gaussian Process

Introduction

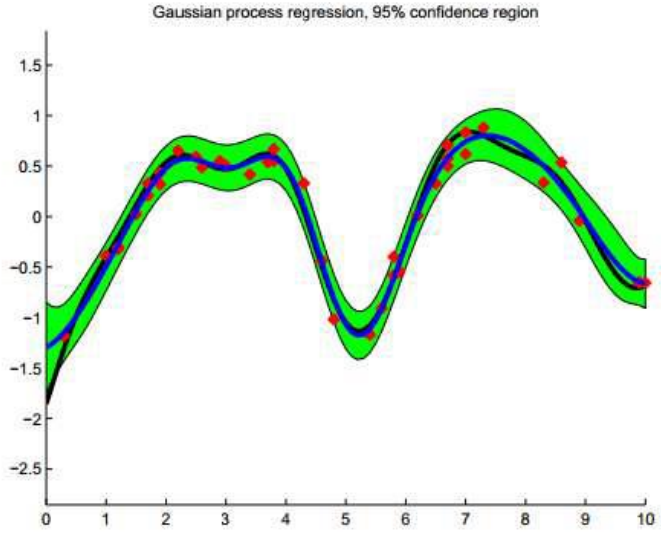
Kernel Regression
(Non-Parametric, Non-Bayes)



$$f(\mathbf{x}) = \sum_{i=1}^{N} w_i(\mathbf{x}) y_i$$

$$w_i(\mathbf{x}) \triangleq \frac{\kappa_h(\mathbf{x} - \mathbf{x}_i)}{\sum_{i'=1}^{N} \kappa_h(\mathbf{x} - \mathbf{x}_{i'})}$$

GP Regression
(Non-Parametric, Bayes)

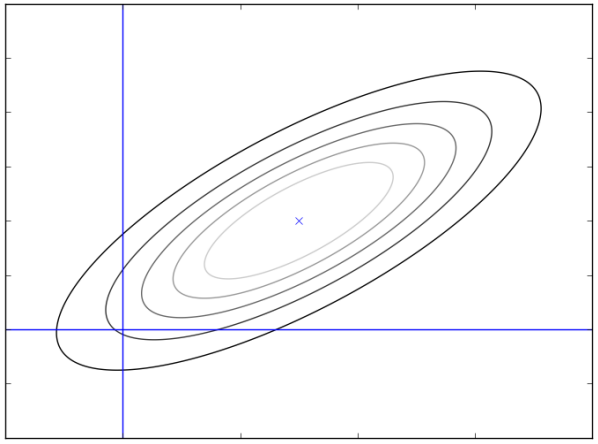


함수에 대한 분포를 알기 때문에
예측의 **confidence**가 나옴

Gaussian Distribution

- **Gaussian Distribution (Normal Distribution)**

$$X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$



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

Gaussian Distribution

- **Covariance**

$$\mathrm{cov}(X,Y) = \sigma_{X,Y}^2 = E(XY) - \mu_X \mu_Y$$

$$= E((X-\mu_X)(Y-\mu_Y)) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_X)(y_i-\mu_Y)$$

- **Covariance Matrix**
  - *k* variables are observed together

$$(x_1^1, x_2^1, \cdots, x_k^1), (x_1^2, x_2^2, \cdots, x_k^2), \cdots, (x_1^n, x_2^n, \cdots, x_k^n)$$

$$\Sigma = \begin{pmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 & \cdots & \sigma_{1,k}^2 \\ \sigma_{2,1}^2 & \sigma_{1,1}^2 & \cdots & \sigma_{2,k}^2 \\ \cdots & \cdots & \ddots & \cdots \\ \sigma_{k,1}^2 & \cdots & \cdots & \sigma_{k,k}^2 \end{pmatrix} \quad \text{where } \sigma_{i,j} = \sigma_{X_i, X_j}$$

$$\text{Symetric because } \sigma_{i,j} = \sigma_{j,i}$$

Gaussian Distribution

- **Covariance Matrix**

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_{1,1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{1,1}^2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \sigma_{k,k}^2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 & \cdots & \sigma_{1,k}^2 \\ \sigma_{2,1}^2 & \sigma_{1,1}^2 & \cdots & \sigma_{2,k}^2 \\ \cdots & \cdots & \ddots & \cdots \\ \sigma_{k,1}^2 & \cdots & \cdots & \sigma_{k,k}^2 \end{pmatrix}$$

Gaussian Distribution

- **Posterior Gaussian Distribution**

$$\begin{pmatrix} Y_A \\ Y_B \end{pmatrix} = N\left(\begin{pmatrix} \boldsymbol{\mu_A} \\ \boldsymbol{\mu_B} \end{pmatrix}, \begin{pmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{pmatrix}\right)$$

$$\boldsymbol{p(Y_B | Y_A) = N(\mu, \Sigma)}$$

$$\boldsymbol{\mu = \mu_B + K_{BA} K_{AA}^{-1} Y (Y_A - \mu_A)}$$

$$\boldsymbol{\Sigma = K_{BB} - K_{BA} K_{AA}^{-1} K_{AB}}$$

황당한 f(x) 값 추정

- **Q: 우리가 모르는 함수 f(x) 가 있습니다.**
  f(1)의 값이 얼마일까요?
- **A: 장난치냐? ㅡㅡ;;**
- **Q: 아뇨.. 장난은 아니구요.. ^^;;;**
- **Q: 그럼 이렇게 합시다. 어짜피 모르는 것이니까,**
  모든 x에 대해서 f(x)를 확률변수라고 가정하고요.
  f(x)는 다음 확률 분포를 따른다고 가정하지요.
  - f(x) ~ N(0, 1)
  - 가장 만만한 분포가 Gaussian이니까 이것에 따른다고 하고요
  - f(x)가 무슨 값일지 모르니 그 평균은 0일테고요. 분산은 다른 값으로 해도 되지만 우선 1이라고 할께요.
- **A: OK.. 인정할 수 있어**

황당한 f(x) 값 추정

- **Q: 그러면, f(1)=1라고 하면, f(2)의 값은 얼마일까요?**
- **A: 뭐.. 지금 f(1) 하나만 알고 있으니, f(2)도 평균적으로 1이라고 해야하지 않을까?**
- **A: 그래도, f(2)는 확률변수이니까 분포를 알아야 할 텐데..**
- **Q: 음.. 계속 맞는 말씀하시네요..**
- **Q: 그러면 이렇게 합시다. 확률변수 f(1)과 f(2) 사이에 correlation이 좀 있다고 하지요.**
  예를 들면 f(1)이 평균에서 1정도 떨어진 값이 관찰되었다면 f(2)는 평균에서 0.7 정도 떨어진 값이 관찰된다고 합시다.
  - 즉, f(1)과 f(2)의 covariance는 0.7

$$cov(y_1, y_2) = E\left((y_1 - \mu_{y_1})(y_2 - \mu_{y_2})\right)$$

황당한 f(x) 값 추정

- **Q: 그리고, (f(1), f(2))는 bivariate Gaussian distribution 을 따른다고 하지요.**
  - bivariate Gaussian distribution는 아래와 같은 수식이지요

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = N\left(\begin{pmatrix} \mu_{y_1} \\ \mu_{y_2} \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{pmatrix}\right)$$

- **A: 가정이니까 알아서 해요.**
  그런데, 그렇게 가정하면 뭐가 좀 나아지나요?
- **Q: 에이.. 잘 모르시네.. 복잡한 수학 안하고 결론만 말 할게요**

$$p(y_2|y_1) = N(\mu, \sigma^2) \qquad \mu = \mu_{y_2} - \sigma_{21}^2 \cdot (\sigma_{12}^2)^{-1} \cdot (y_1 - \mu_{y_1})$$
$$\sigma^2 = \sigma_{22}^2 - \sigma_{21}^2 \cdot (\sigma_{11}^2)^{-1} \cdot \sigma_{12}^2$$

- **A: 뭔 소린지??**

황당한 f(x) 값 추정

- **Q: 아래 수식에서 y_1 = f(1), y_2 = f(2)라고 하고 각 공분산을 대입하면,**

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = N\left(\begin{pmatrix} \mu_{y_1} \\ \mu_{y_2} \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{pmatrix}\right) \longrightarrow \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}\right)$$

이고, 우리는 y_1=1로 알고 있으니까

$$p(y_2|y_1 = 1) = N(\mu, \sigma^2) \longrightarrow p(y_2|y_1 = 1) = N(1, 0.51)$$

$$\mu = 0 + \sigma_{21}^2 \cdot (\sigma_{12}^2)^{-1} \cdot (y_1 - 0)$$
$$\sigma^2 = \sigma_{22}^2 - \sigma_{21}^2 \cdot (\sigma_{11}^2)^{-1} \cdot \sigma_{12}^2$$

- **A: 오호.. f(2)의 분포를 알아 낼 수가 있네 @@**

황당한 f(x) 값 추정

- **A: 그런데, $f(x_1)$과 $f(x_2)$의 공분산을 어떻게 알지요?**
- **Q: 사실 알 수가 없어요. 그러니까 가정을 하나 주가 해 봅시다.**
  - x_1과 x_2가 similar 할수록 $f(x_1)$과 $f(x_2)$는 유사한 값을 가진다
    - x_1과 x_2가 가까워질수록 $f(x_1)$과 $f(x_2)$의 covariance는 커진다
  - x_1과 x_2가 덜 similar 할수록 $f(x_1)$과 $f(x_2)$는 다른 값을 가진다
    - x_1과 x_2가 멀어질수록, $f(x_1)$과 $f(x_2)$의 covariance는 작아진다
  - 일반적으로 위의 가정에 부합하는 함수를 하나 정의해서 공분산으로 사용해요. 예를 들면,

$$f(x_1)\text{과 } f(x_2) \text{ 공분산} = \exp\left(-\frac{\|x_1 - x_2\|_2^2}{l^2}\right)$$

황당한 f(x) 값 추정

- **A: 어.. 그러니까..**
  - 예를 들어, x_1=x_2 가 되면 $\sigma_{12}^2 = \sigma_{21}^2$ 가 되고, 그러면

$$p(y_2|y_1) = N(1,0)$$

  - x_2가 x_1에서 멀어질수록 $p(y_2|y_1)$의 분산은 커집니다. σ_{21}^2와 σ_{12}^2값은 가정에 의해서 작아지니까요.

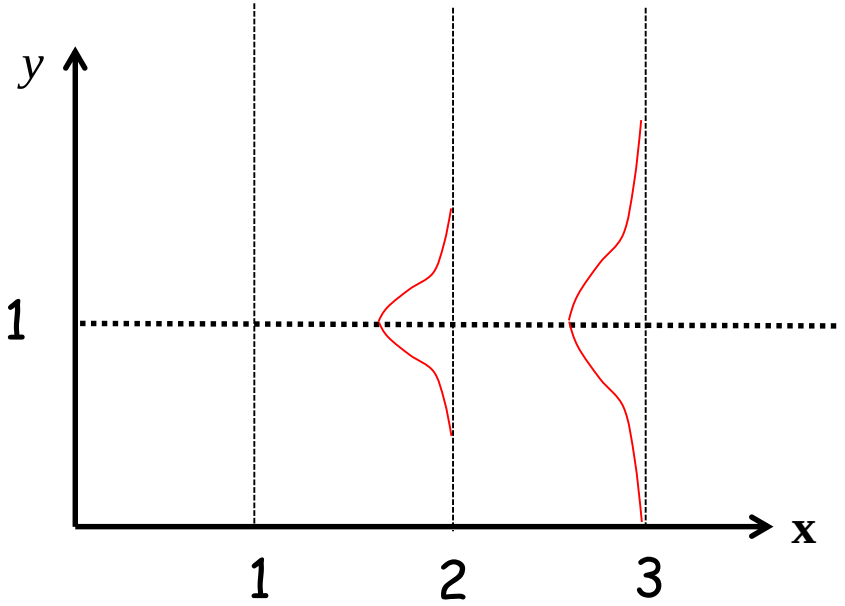
$$\sigma_{12}^2 = \sigma_{21}^2 = \exp\left(-\frac{\|x_1 - x_2\|_2^2}{l^2}\right)$$

$$\sigma^2 = \sigma_{22}^2 - \sigma_{21}^2 \cdot (\sigma_{11}^2)^{-1} \cdot \sigma_{12}^2$$

- **A: 이것은 우리 가정과도 일치하는 것이네요**

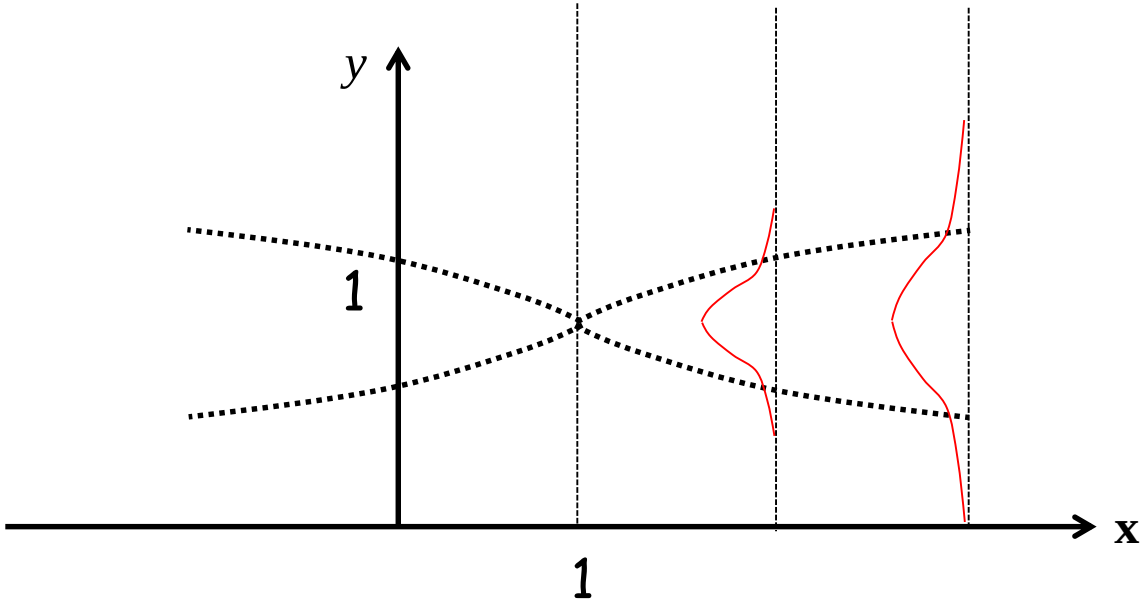
황당한 f(x) 값 추정

- **Q: 그렇지요, x가 1에 가까우면 당연히 f(x)는 1에 가까워질 것이고, 멀어지면 더 불분명해진다고 할 수 있으니까요**



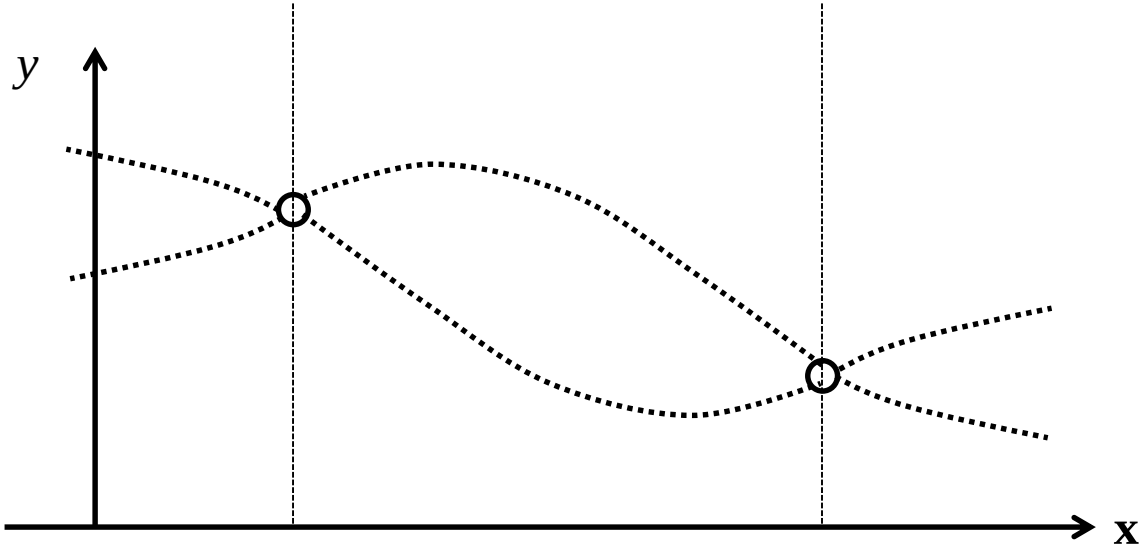
황당한 f(x) 값 추정

- **Q: 이때 f(x)의 uncertainty는 아래처럼 표시할 수도 있겠지요**



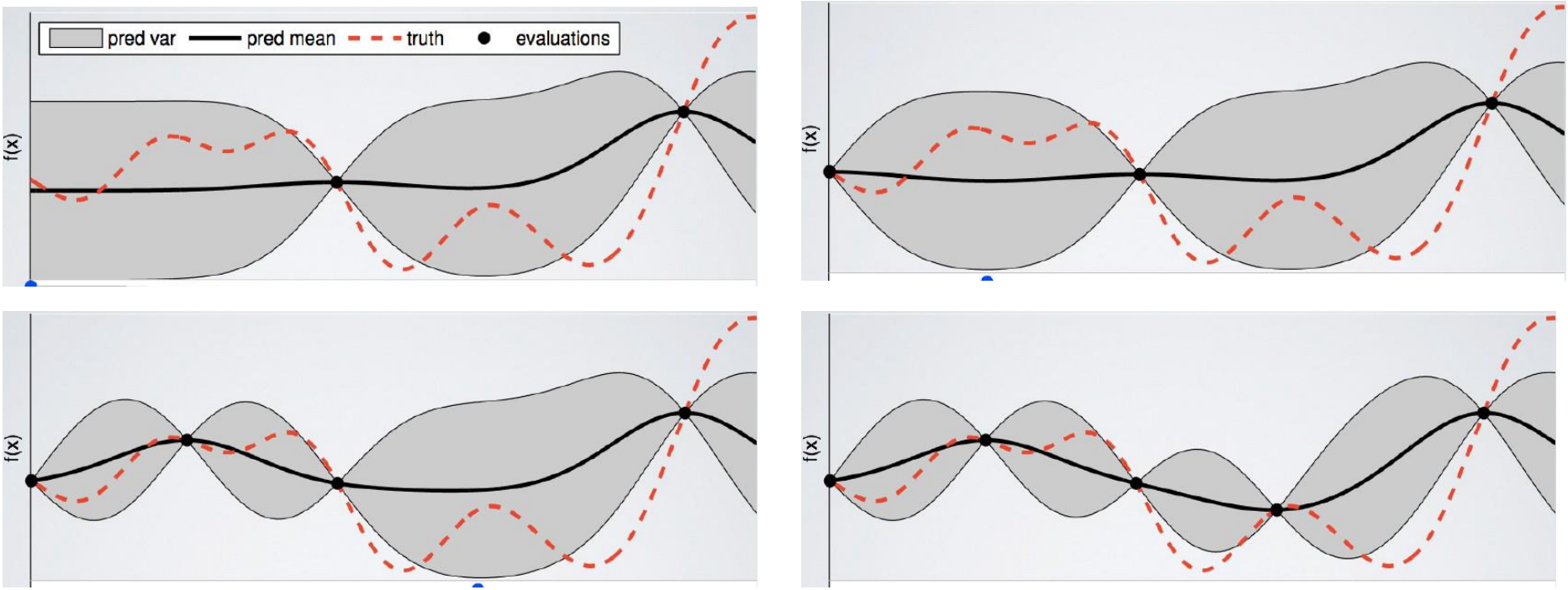
황당한 f(x) 값 추정

- **Q: 만약 우리가 두 지점의 값을 알고 있다면, 다른 지점의 값은 아래와 같이 분포한다고 할 수 있어요.**



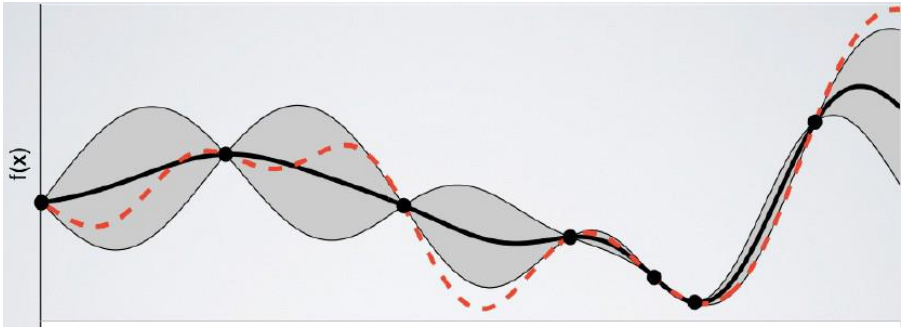
황당한 f(x) 값 추정

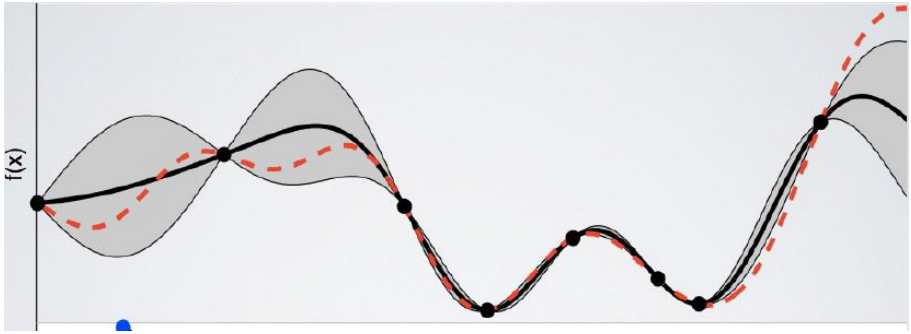
- **Q: 그리고 더 많은 지점의 값을 알수록 f(x)를 더 잘 추정할 수 있어요.**

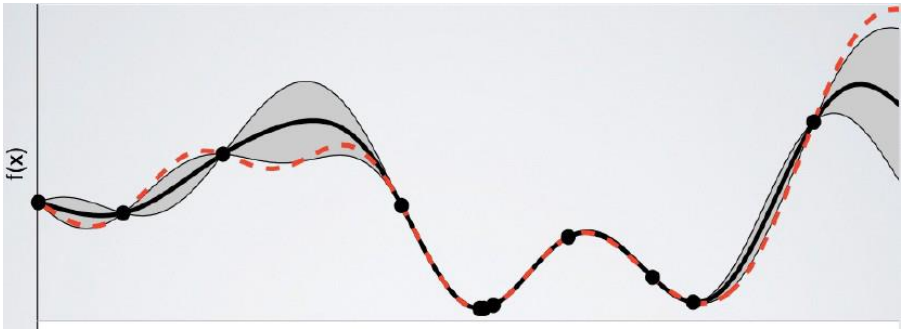


황당한 f(x) 값 추정

- **Q: 관찰한 값이 n개일 때의 경우로 확장하면, Gaussian Process입니다.**



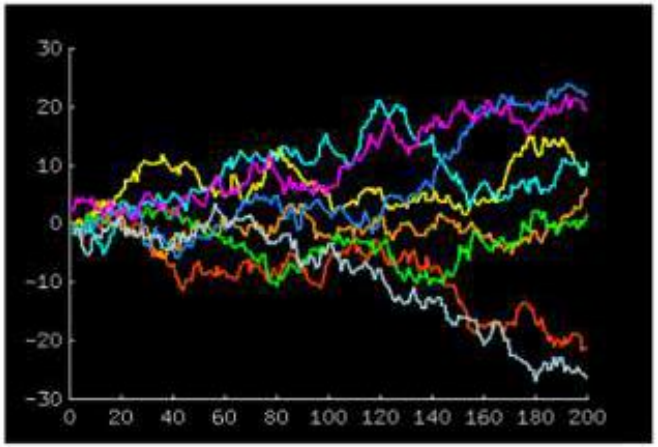




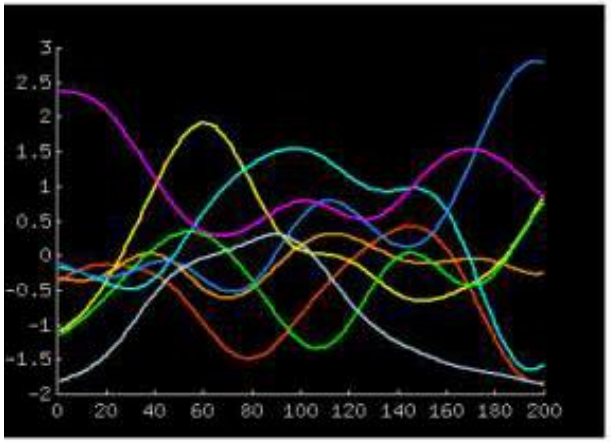
황당한 f(x) 값 추정

- **A: 네.. 쉽네요. 그런데, Covariance 함수가 바뀌면 추정되는 함수의 모양이 바뀌겠네요.**
- **Q: 맞습니다. Covariance 함수는 사용자가 잘 선택해야 합니다.**

Brownian



Squared exponential



Gaussian Process

- **Gaussian process is a collection of random variables**
  - Any finite number of which have joint Gaussian distributions.

$$(y_1, y_2, \ldots, y_n) \sim N(\mu, \Sigma)$$

where

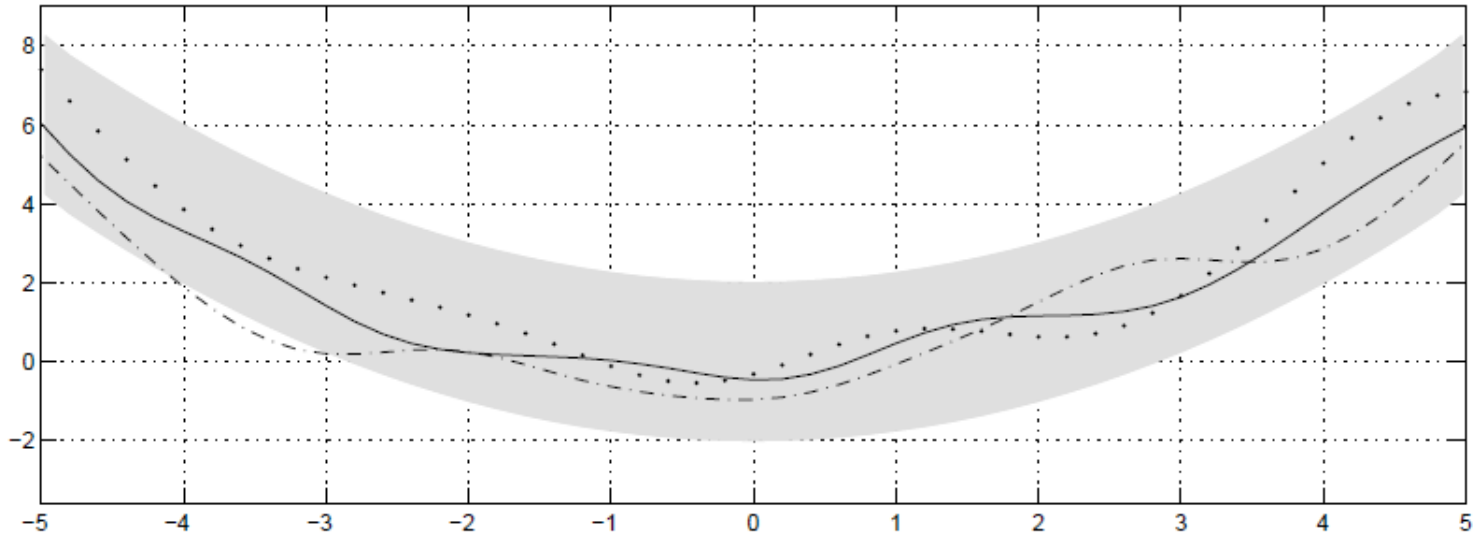
$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is given

$$\mu_i = m(x_i), \Sigma_{ij} = \kappa(x_i, x_j)$$

- m and κ are given by user
- Usually, we set $m(\cdot) = 0$

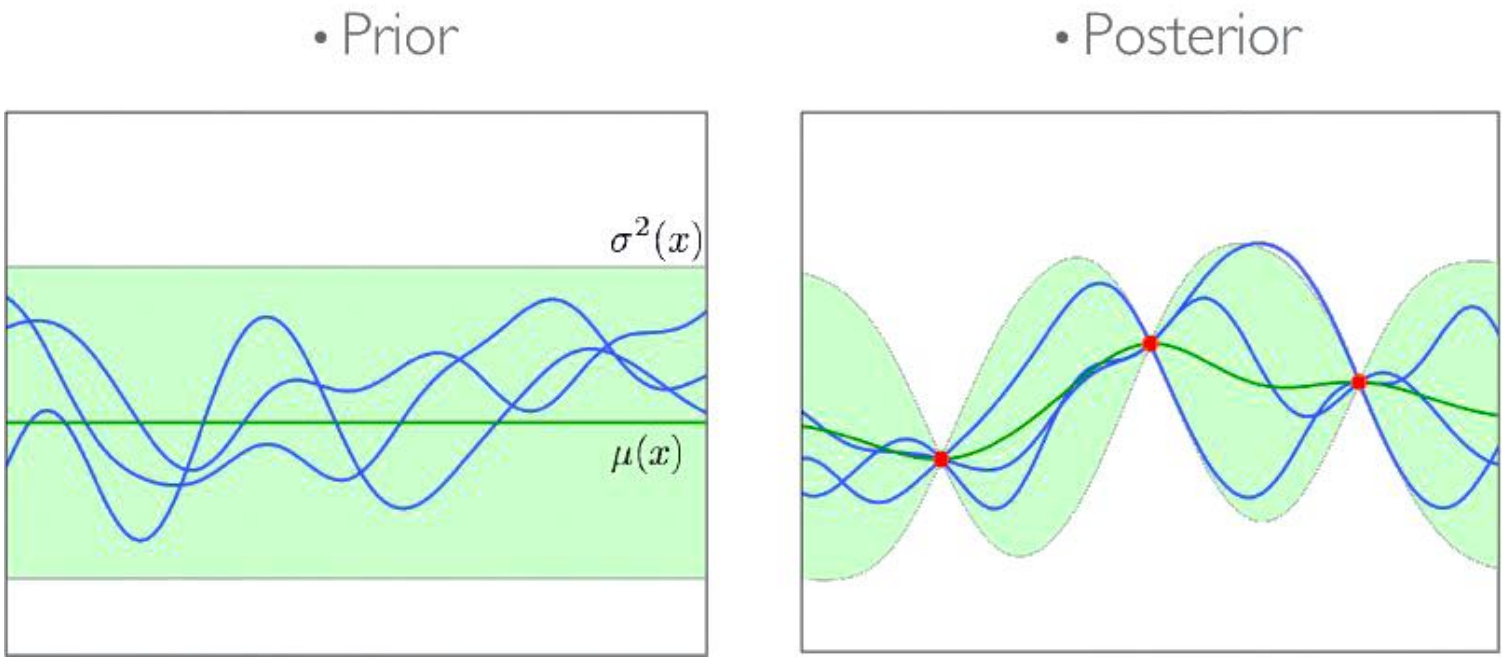
Gaussian Process

- **A distribution over functions**
  - a collection of random variables == a function



Gaussian Process

- **Prior Gaussian Process vs Posterior Gaussian Process**



Gaussian Process

- **Posterior Gaussian Process**
  - Given Training Data: $\boldsymbol{D} = \{(\boldsymbol{x_i}, \boldsymbol{y_i}), \boldsymbol{i} = \boldsymbol{1}: \boldsymbol{N}\}$

$$X = \{x_1, x_2, \dots, x_N\}, Y = \{y_1, y_2, \dots, y_N\}$$

  - Test Data: $\boldsymbol{D}_* = \{(\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}), \boldsymbol{i} = \boldsymbol{1}: \boldsymbol{N}_*\}$

$$X_* = \{\alpha_1, \alpha_2, \dots, \alpha_{N_*}\}, Y_* = \{\beta_1, \beta_2, \dots, \beta_{N_*}\}$$

  - Assumption

$$\begin{pmatrix} Y \\ Y_* \end{pmatrix} = N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} K & K_* \\ K_*^T & K_{**} \end{pmatrix}\right)$$

$$K = \kappa(X, X) \quad K_* = \kappa(X, X_*) \quad K_{**} = \kappa(X_*, X_*)$$

  - Then

$$\boldsymbol{p(Y_*|X_*, X, Y) = N(\mu_*, \Sigma_*)}$$

$$\boldsymbol{\mu_* = K_*^T K^{-1} Y} \qquad \boldsymbol{\Sigma_* = K_{**} - K^T K^{-1} K_*}$$

Gaussian Process

- **If there is one test data**
  - Given Training Data: $\boldsymbol{D} = \{(\boldsymbol{x_i}, \boldsymbol{y_i}), \boldsymbol{i} = \boldsymbol{1}\!:\!\boldsymbol{N}\}$

$$X = \{x_1, x_2, \dots, x_N\}, Y = \{y_1, y_2, \dots, y_N\}$$

  - Test Data: $(\boldsymbol{x}_*, y_*)$

$$\boldsymbol{\mu}_* = \boldsymbol{K}_*^{\boldsymbol{T}} \boldsymbol{K}^{-\boldsymbol{1}} \boldsymbol{Y}$$

similarity

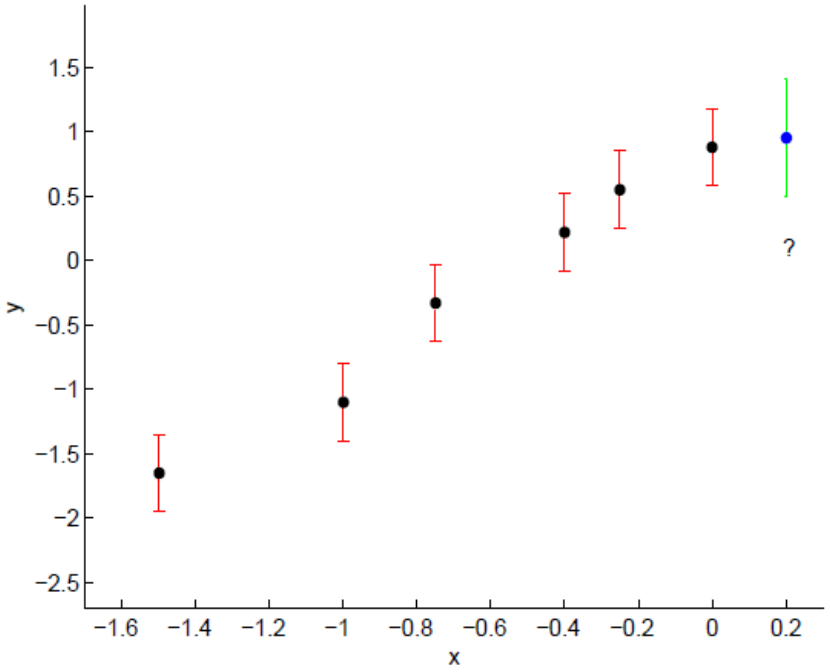
Some value

$$\boldsymbol{\mu}_* = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{\kappa}(\boldsymbol{x_i}, \boldsymbol{x_j}) \cdot \boldsymbol{\alpha_i}$$

$$\boldsymbol{\alpha_i} = \boldsymbol{K}^{-\boldsymbol{1}} \boldsymbol{Y}$$

Gaussian Process

- **Example**
  - Calculate K, K_*, K_{**}



$$\mathbf{x} = \begin{bmatrix} -1.50 & -1.00 & -0.75 & -0.40 & -0.25 & 0.00 \end{bmatrix}$$

$$k(x, x') = \sigma_y^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right) + \sigma_n^2 \delta_{ii'}$$

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix}$$

$$K_* = \begin{bmatrix} k(x_*, x_1) & k(x_*, x_2) & \cdots & k(x_*, x_n) \end{bmatrix} \qquad K_{**} = k(x_*, x_*)$$

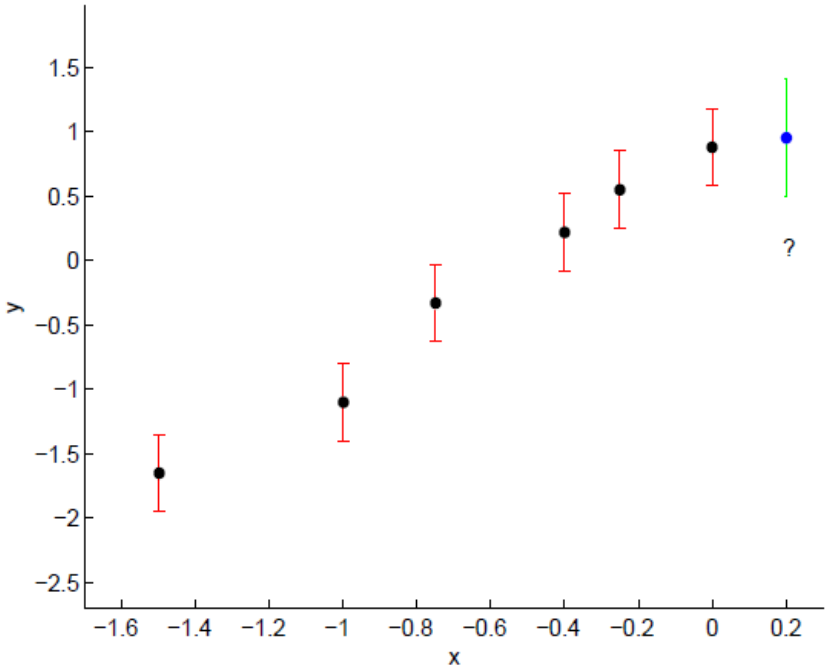
$$K = \begin{bmatrix} 1.70 & 1.42 & 1.21 & 0.87 & 0.72 & 0.51 \\ 1.42 & 1.70 & 1.56 & 1.34 & 1.21 & 0.97 \\ 1.21 & 1.56 & 1.70 & 1.51 & 1.42 & 1.21 \\ 0.87 & 1.34 & 1.51 & 1.70 & 1.59 & 1.48 \\ 0.72 & 1.21 & 1.42 & 1.59 & 1.70 & 1.56 \\ 0.51 & 0.97 & 1.21 & 1.48 & 1.56 & 1.70 \end{bmatrix}$$

$$K_* = \begin{bmatrix} 0.38 & 0.79 & 1.03 & 1.35 & 1.46 & 1.58 \end{bmatrix}$$

$$K_{**} = 1.70$$

Gaussian Process

- **Example**
  - Calculate y_*, var(y_*)



$$\mathbf{x} = \begin{bmatrix} -1.50 & -1.00 & -0.75 & -0.40 & -0.25 & 0.00 \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K & K_*^{\mathrm{T}} \\ K_* & K_{**} \end{bmatrix}\right)$$

$$y_* | \mathbf{y} \sim \mathcal{N}(K_* K^{-1}\mathbf{y},\ K_{**} - K_* K^{-1} K_*^{\mathrm{T}})$$

$$\overline{y}_* = K_* K^{-1}\mathbf{y}$$

$$\mathrm{var}(y_*) = K_{**} - K_* K^{-1} K_*^{\mathrm{T}}$$

$$\overline{y}_* = 0.95 \text{ and } \mathrm{var}(y_*) = 0.21$$

왜 Gaussian Process가 좋은가?

- **장점**
  - Bayesian method 이다
  - 예측의 uncertainty 를 수치화 할 수 있다
  - 여러 model selection 과 hyperparameter selection 과 같은 Bayesian method를 그대로 사용할 수 있다
  - input point 에 대한 임의의 함수를 모델링한다 (No model assumption)