Scalable ML

10605-10805

Gaussian Processes for Regression

Barnabás Póczos

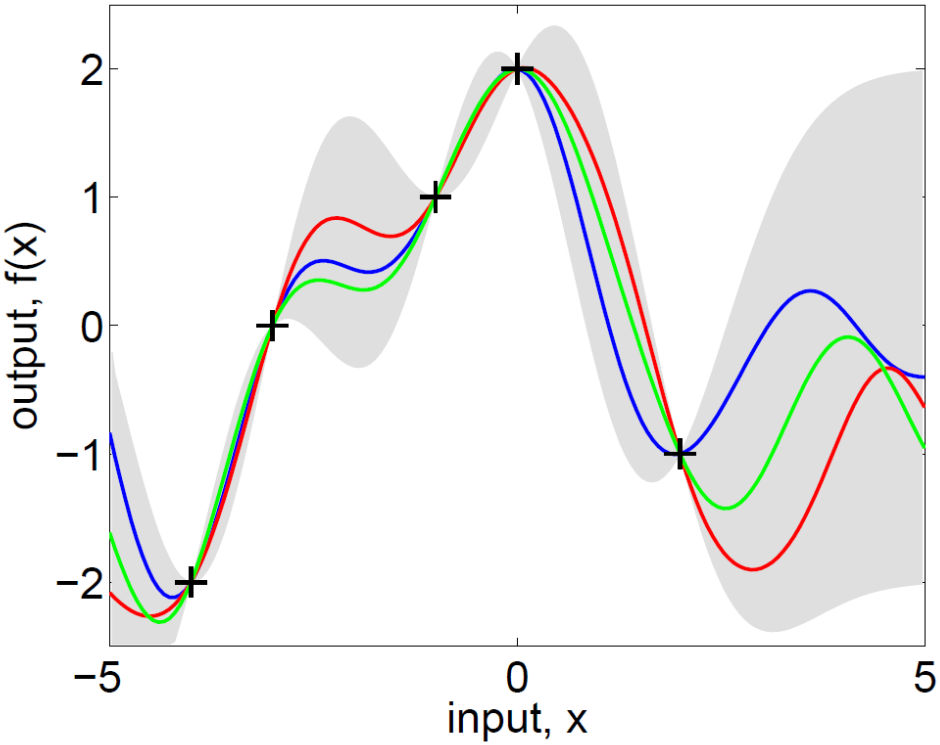
Why GPs for Regression?

Regression methods:

Linear regression, ridge regression, support vector regression, kNN regression, etc…

Motivation:

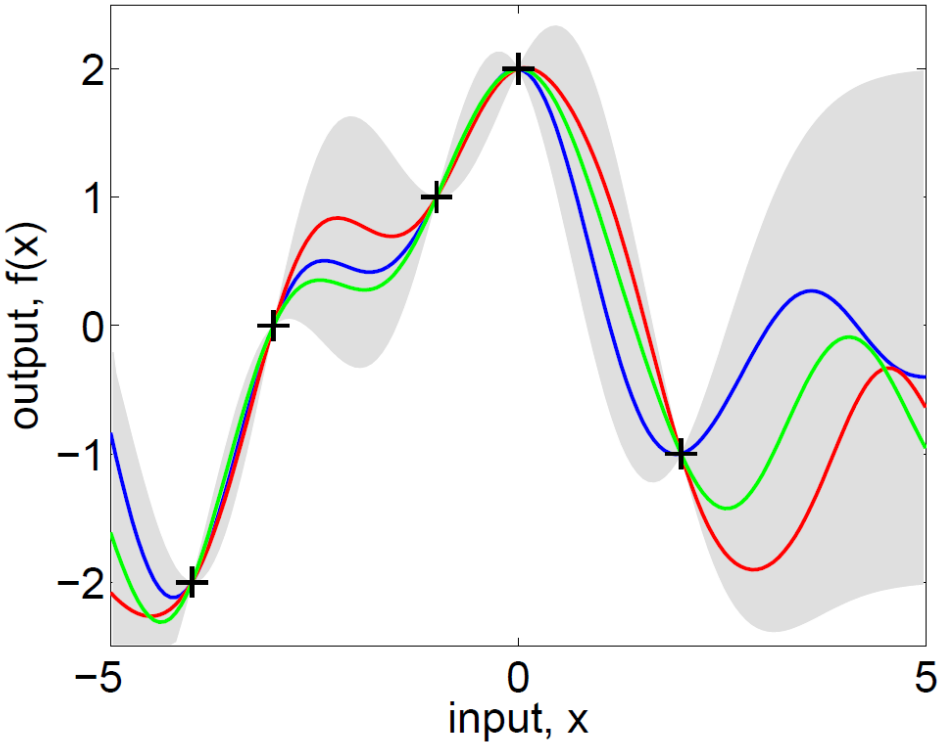
All the above regression method give point estimates. We would like a method that could also provide confidence during the estimation.



Why GPs for Regression?

GPs can answer the following questions

- Here's where the function will **most likely be.** (expected function)
- Here are some **examples** of what it might look like. (sampling from the posterior distribution)
- Here is a prediction of what you'll see if you evaluate your function at x, **with confidence**

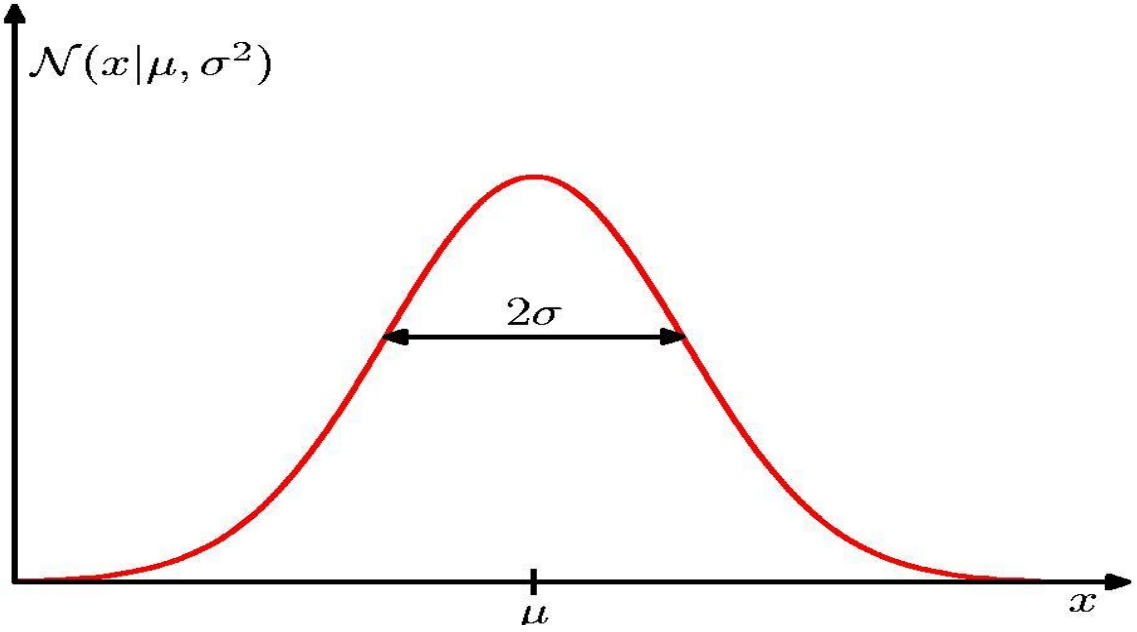


Properties of Multivariate Gaussian Distributions

1D Gaussian Distribution

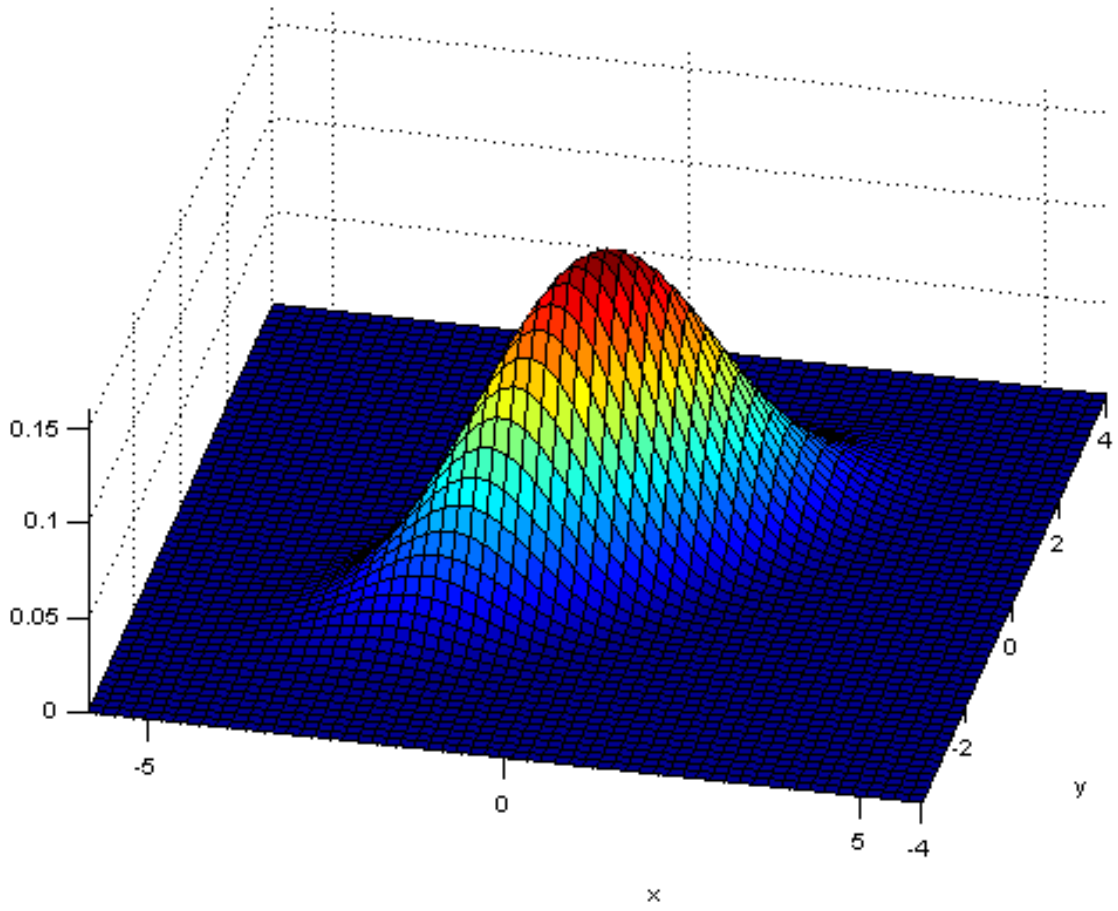
Parameters

- Mean, μ
- Variance, σ^2



$$P(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Multivariate Gaussian



$$p(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left\{\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Multivariate Gaussian

- A 2-dimensional Gaussian is defined by
  - a mean vector $\mu = [\ \mu_1, \mu_2\ ]$
  - a covariance matrix: $\Sigma = \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{2,1}^2 \\ \sigma_{1,2}^2 & \sigma_{2,2}^2 \end{bmatrix}$

  where $\sigma_{i,j}^2 = E[\ (x_i - \mu_i)\ (x_j - \mu_j)\ ]$ is (co)variance

- Note: Σ is symmetric,
  "positive semi-definite": $\forall x$: $x^T \Sigma x \geq 0$

Useful Properties of Gaussians

- Marginal distributions of Gaussians are Gaussian

- Given:

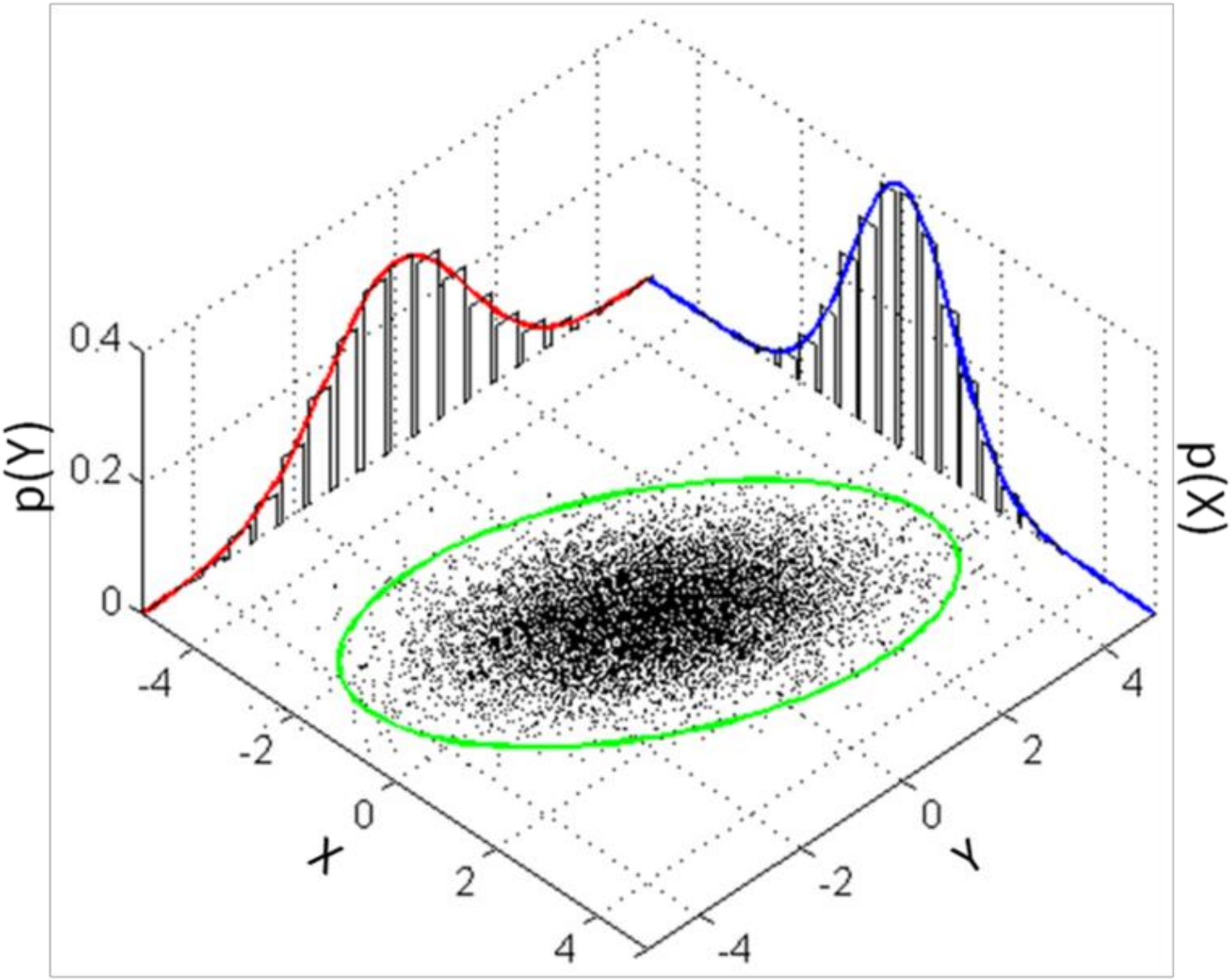
$$x = (x_a, x_b), \mu = (\mu_a, \mu_b)$$

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

- Marginal Distribution:

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

Marginal distributions of Gaussians are Gaussian



Block Matrix Inversion

Theorem

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} S_D^{-1} & -A^{-1}BS_A^{-1} \\ -D^{-1}CS_D^{-1} & S_A^{-1} \end{bmatrix}$$

Definition: Schur complements

Schur complements of A: $S_A = D - CA^{-1}B$

Schur complements of D: $S_D = A - BD^{-1}C$

Useful Properties of Gaussians

- Conditional distributions of Gaussians are Gaussian
- Notation:

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \quad \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

- Conditional Distribution:

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

$$\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b - \mu_b) = \mu_a - \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b - \mu_b)$$

$$\Lambda_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

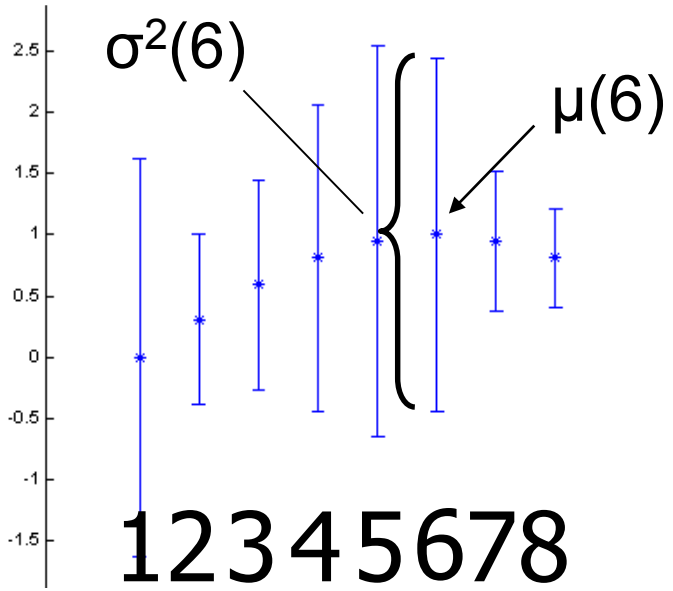
Schur complement of Σ_{bb} in Σ

Higher Dimensions

- **Visualizing > 3 dimensions is… difficult**

- Marginals are Gaussian, e.g., $f(6) \sim N(\mu(6), \sigma^2(6))$

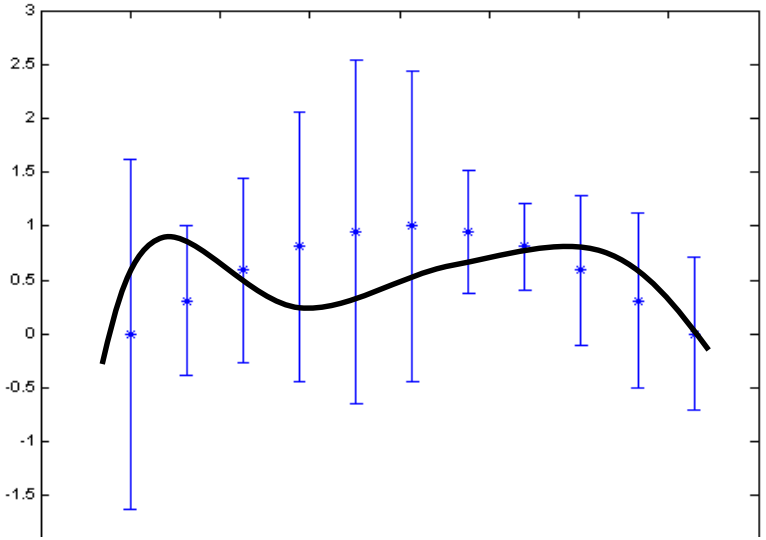
Visualizing an 8-dimensional Gaussian:



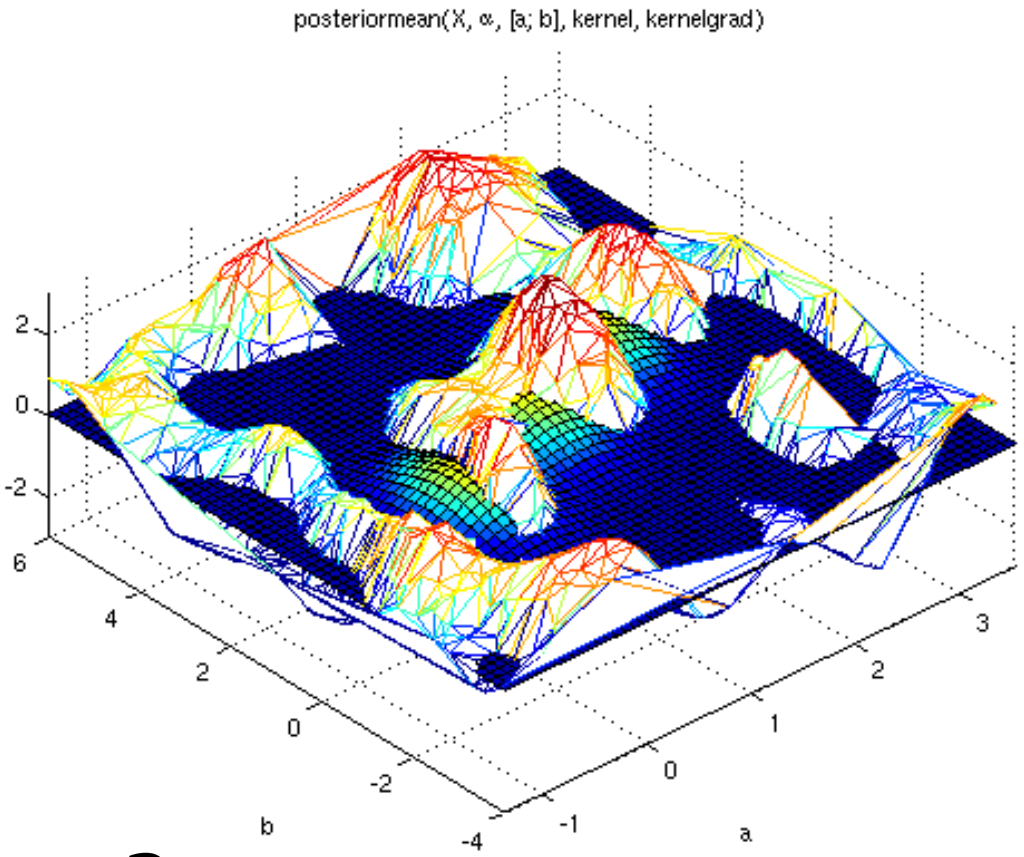
Yet Higher Dimensions

Why stop there?

- We indexed before with $\{1, 2, 3, 4, 5, 6, 7, 8\}$.
- Why noy indexing with $\mathbb{Z}$, or $\mathbb{R}$?
- Need functions $\mu(x), k(x, z)$, $\forall x, z \in \mathbb{R}$
- x and z are indexes over the random variables
- f is now an uncountably infinite dimensional vector
- **Smoothness**: If $z \to x$, then i) $\mathbb{E}[f(z)] \to \mathbb{E}[f(x)]$, and ii) $corr(f(z), f(x))$ becomes high.



Getting Ridiculous



Why stop there?

- We indexed before with $\mathbb{R}$, why not with $\mathbb{R}^D$?
- Need functions $\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{z}),\ \forall \mathbf{x}, \mathbf{z} \in \mathbb{R}^D$

Gaussian Process

Definition:

- Probability distribution *indexed by* an arbitrary set (integer, real, finite dimensional vector, etc)
- Each element gets a Gaussian distribution over the reals with mean $\mu(x)$
- These distributions are dependent/correlated as defined by $k(x,z)$
- Any finite subset of indices defines a multivariate Gaussian distribution

Gaussian Process

- Distribution over *functions….*

  If our regression model is a GP, then it won't be a point estimate anymore! It can provide regression estimates with confidence

- Domain (index set) of the functions can be pretty much whatever
  - Reals
  - Real vectors
  - Graphs
  - Strings
  - Sets
  - …

GP pseudo code

Inputs:

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \in \mathbb{R}^{n \times D}, \; n \text{ training inputs}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^{n}, \; n \text{ training targets}$$

$k(\cdot, \cdot) : \mathbb{R}^{D \times D} \rightarrow \mathbb{R}$ covariance function (kernel)

$\mathbf{x}_*$ test input

σ^2 noise level on the observations

$$[y(\mathbf{x}) = f(\mathbf{x}) + \epsilon, \; \epsilon \sim \mathcal{N}(0, \sigma^2)]$$

GP pseudo code (continued)

1., $K \in \mathbb{R}^{n \times n}$ Gram matrix. $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$

$$k(\mathbf{x}_*) = k_* = k(X, \mathbf{x}_*) = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_*) \\ \vdots \\ k(\mathbf{x}_n, \mathbf{x}_*) \end{bmatrix} \in \mathbb{R}^n$$

2., $\boldsymbol{\alpha} = (K + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y} \in \mathbb{R}^n$

3., $\bar{f}_* = k_*^T \boldsymbol{\alpha} \in \mathbb{R}$

4., $cov(f_*) = \underbrace{k(\mathbf{x}_*, \mathbf{x}_*)}_{\mathbb{R}} - \underbrace{k_*^T}_{\mathbb{R}^{1 \times n}} \underbrace{[K + \sigma^2 I_n]^{-1}}_{\mathbb{R}^{n \times n}} \underbrace{k_*}_{\mathbb{R}^n} \in \mathbb{R}$

Outputs: $\bar{f}_*,\ cov(f_*)$

Function Space View of GPs

Notations:

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \tilde{\mathbf{x}})) \in \mathbb{R},\ \mathbf{x} \in \mathbb{R}^D$$

$$m(\mathbf{x}) = \mathbb{E}[f(x)] \in \mathbb{R},\ \text{(mean function)}$$

$$k(\mathbf{x}, \tilde{\mathbf{x}}) = \mathbb{E}[(f(x) - m(\mathbf{x}))(f(\tilde{\mathbf{x}}) - m(\tilde{\mathbf{x}}))^T] \in \mathbb{R}$$

(covariance function)

GP is **completely specified** by its
mean function $m(\mathbf{x})$, and
covariance function $k(\mathbf{x}, \tilde{\mathbf{x}})$

Function Space View of GPs

Gaussian Processes:

For each $\mathbf{x} \in \mathbb{R}^D$ we associate a Gaussian variable $f(\mathbf{x})$ such that $\mathbb{R} \ni f(\mathbf{x}) \sim \mathcal{N}_{f(\mathbf{x})}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$, and its correalation with other $f(\tilde{\mathbf{x}})$ variables is $k(\mathbf{x}, \tilde{\mathbf{x}})$.

$$\mathbb{R} \ni f(\mathbf{x}) \sim \mathcal{N}_{f(\mathbf{x})}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$$

$$\begin{bmatrix} f(\mathbf{x}) \\ f(\tilde{\mathbf{x}}) \end{bmatrix} \sim \mathcal{N}_{\begin{bmatrix} f(\mathbf{x}) \\ f(\tilde{\mathbf{x}}) \end{bmatrix}} \left\{ \begin{bmatrix} m(\mathbf{x}) \\ m(\tilde{\mathbf{x}}) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}, \mathbf{x}) & k(\tilde{\mathbf{x}}, \mathbf{x}) \\ k(\mathbf{x}, \tilde{\mathbf{x}}) & k(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) \end{bmatrix} \right\}$$

Prediction with noise free observations

Training set: $D = \{(\mathbf{x}_i, f_i) | i = 1, \ldots, n\}$

noise free observations

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \in \mathbb{R}^{n \times D}, \; m \text{ training inputs}$$

$$X_* = \begin{bmatrix} \mathbf{x}_{*1}^T \\ \vdots \\ \mathbf{x}_{*m}^T \end{bmatrix} \in \mathbb{R}^{m \times D}, \; m \text{ test inputs}$$

$$f = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix} \in \mathbb{R}^n, \; n \text{ training targets}$$

$$f_* = \begin{bmatrix} f_{*1} \\ \vdots \\ f_{*m} \end{bmatrix} \in \mathbb{R}^m, \; m \text{ test targets}$$

Prediction with noise free observations

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}_{\begin{bmatrix} f \\ f_* \end{bmatrix}} \left\{ \begin{bmatrix} \mathbf{0}_n \\ \mathbf{0}_m \end{bmatrix}, \underbrace{\begin{bmatrix} k(X,X) & k(X,X_*) \\ k(X_*,X) & k(X_*,X_*) \end{bmatrix}}_{\in \mathbb{R}^{(m+n)\times(m+n)}} \right\}$$

Goal:

We want to calculate the posterior distribution $f_*|X_*, X, f$

Prediction with noise free observations

Lemma:

$$P(f_*|X_*, X, f) =$$

$$\mathcal{N}_{f_*}\left(k(X_*, X)k(X, X)^{-1}f, k(X_*, X_*) - k(X_*, X)k(X, X)^{-1}k(X, X_*)\right)$$

Proof: a bit of calculation using the joint (n+m) dim density

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

Prediction using noisy observations

$$y = f(\mathbf{x}) + \epsilon \in \mathbb{R} \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \in \mathbb{R}$$

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

$$\Rightarrow cov([y_1, \ldots, y_n]) = k(X,X) + \sigma^2 \mathbf{I}_n \in \mathbb{R}^{n \times n}$$

The joint distribution:

$$\Rightarrow \begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}_{\begin{bmatrix} y \\ f_* \end{bmatrix}} \left\{ \begin{bmatrix} \mathbf{0}_n \\ \mathbf{0}_m \end{bmatrix}, \begin{bmatrix} k(X,X) + \sigma^2 \mathbf{I}_n & k(X,X_*) \\ k(X_*,X) & k(X_*,X_*) \end{bmatrix} \right\}$$

Prediction using noisy observations

Short notations:

$$K = k(X, X) \in \mathbb{R}^{n \times n}$$

$$K_* = k(X, X_*) \in \mathbb{R}^{n \times m}$$

$$k(\mathbf{x}_*) = k_* = k(X, \mathbf{x}_*) = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_*) \\ \vdots \\ k(\mathbf{x}_n, \mathbf{x}_*) \end{bmatrix} \in \mathbb{R}^n$$

$\Rightarrow$ for a single test point $\mathbf{x}_*$:

$$\bar{f}_* = \underbrace{k_*^T}_{\mathbb{R}^{1 \times n}} \underbrace{[K + \sigma^2 I_n]^{-1}}_{\mathbb{R}^{n \times n}} \underbrace{\mathbf{y}}_{\mathbb{R}^n} \in \mathbb{R}$$

$$cov(f_*) = \underbrace{k(\mathbf{x}_*, \mathbf{x}_*)}_{\mathbb{R}} - \underbrace{k_*^T}_{\mathbb{R}^{1 \times n}} \underbrace{[K + \sigma^2 I_n]^{-1}}_{\mathbb{R}^{n \times n}} \underbrace{k_*}_{\mathbb{R}^n} \in \mathbb{R}$$

Thanks for the Attention! ☺