GAUSSIAN PROCESSES

An example of

a Bayesian non parametric method

An example of application: Estimation of migrating bird population

Goal:

Estimate evolution of annual population of a given species by counting birds in migration corridors.

Hard regression problem:

- Modeled very roughly (annual trend)
- Random variations on different time scales (weather conditions, environmental factors)
- Observations are very few (volunteering)
- Observations are noisy (approximation)

More professional example: oil exploration

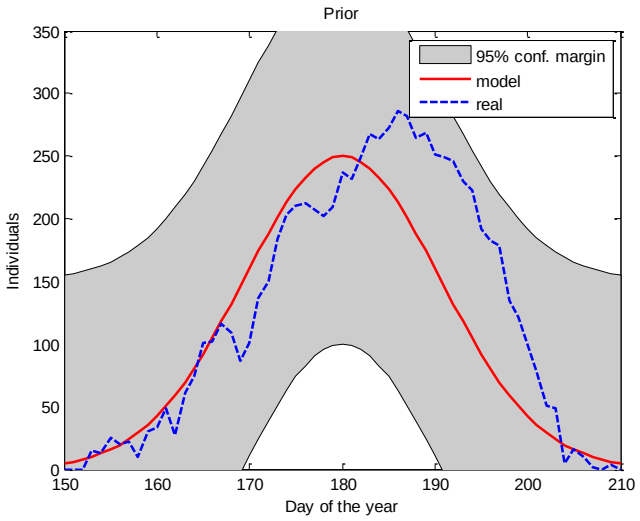
Common Crane

An example of application: Estimation of migrating bird population

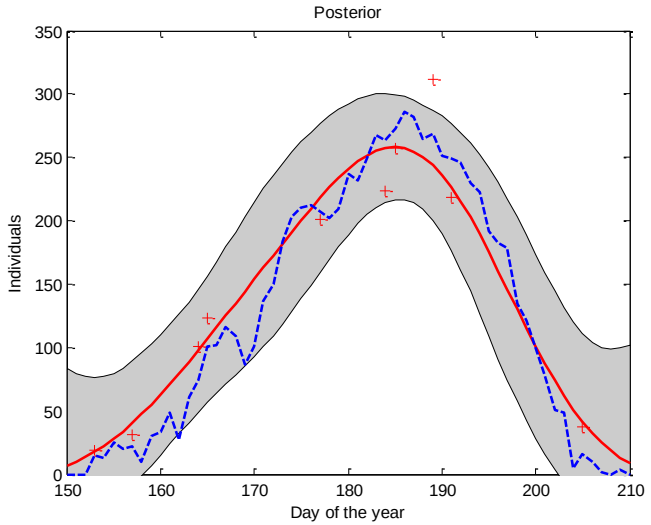
Requirements on regression method:

- Take into account an initial model
- Take into account temporal correlation
- Output degree of confidence on model predictions (confidence interval)
- Take into account noisy observations
- Fit Bayesian inference: reduce model uncertainty with observations

Gaussian process (or Kriging) is the right answer



Demo 1



Multivariate normal distribution: definition

Generalization to $\mathbb{R}^m$:

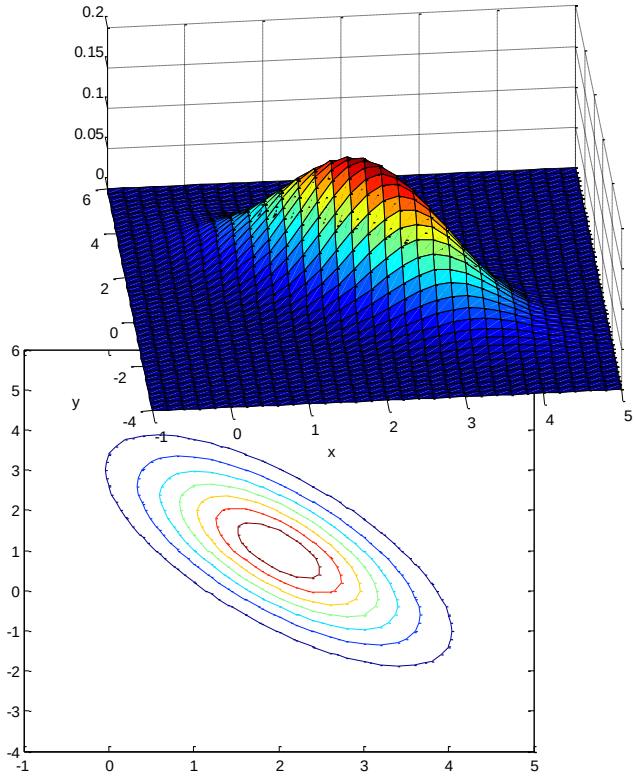
$$X_m \sim \mathcal{N}(\mu_m, \Sigma_{mm}) \Leftrightarrow f_{X_m}(X) = \frac{1}{\sqrt{(2\pi)^m |\Sigma_{mm}|}} e^{-\frac{1}{2}(X-\mu_m)^T \Sigma_{mm}^{-1}(X-\mu_m)}$$

Basic properties:

- $E(X_m) = \mu_m$
- $cov(X_m) = E\left((X-\mu_m)^T(X-\mu_m)\right) = \Sigma_{mm}$

Example:

- $\mu_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \Sigma_{22} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$



Multivariate normal distribution: fundamental properties

Closed under linear transformation:

$$\mathrm{X_m} \sim \mathcal{N}(\mu, \Sigma), \mu \in \mathbb{R}^{\mathrm{m}}, \Sigma \in M_{mm}, \mathrm{A} \in \mathrm{M_{nm}}, \mathrm{B} \in \mathrm{M_{n1}} \Longrightarrow$$

$$\boxed{AX + B \sim \mathcal{N}(A\mu + B, A\Sigma A^T)}$$

Particular cases:

Given $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}\right)$

- Addition: $\boxed{X_1 + X_2 \sim \mathcal{N}\left(\mu_1 + \mu_2, \Sigma_{11} + \Sigma_{22} + \Sigma_{12} + \Sigma_{12}^T\right)}$
- Marginalization: $\boxed{X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})}$

Multivariate normal distribution: fundamental properties

Closed under conditioning:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix} \right) \Rightarrow$$

$$\boxed{P(X_1 | X_2 = \vec{x}) \sim \mathcal{N}\left(\mu_1 + \Sigma_{12}{\Sigma_{22}}^{-1}(\vec{x} - \mu_2), \Sigma_{11} - \Sigma_{12}{\Sigma_{22}}^{-1}\Sigma_{12}^T\right)}$$

Closed under (pdf) multiplication:

$$\boxed{\mathcal{N}(\mu_A, \Sigma_A) \times \mathcal{N}(\mu_B, \Sigma_B) \equiv \mathcal{N}\left(\left(\Sigma_{\mathrm{A}}^{-1} + \Sigma_{\mathrm{B}}^{-1}\right)^{-1}\left(\Sigma_{\mathrm{A}}^{-1}\mu_A + \Sigma_{\mathrm{B}}^{-1}\mu_B\right), \left(\Sigma_{\mathrm{A}}^{-1} + \Sigma_{\mathrm{B}}^{-1}\right)^{-1}\right)}$$

Particular case: conjugate prior $\mu \sim \mathcal{N}(\mu_0, \Sigma_0)$ **of** $X \sim \mathcal{N}(\mu, \Sigma)$ **:**

$$\boxed{\mu | X = \vec{x} \sim \mathcal{N}\left((\Sigma_0^{-1} + \Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\vec{x}), (\Sigma_0^{-1} + \Sigma^{-1})^{-1}\right)}$$

Gaussian Process: Definition

A **Gaussian process** f is a stochastic process from $\mathbb{R}^m$ to$\mathbb{R}$ s.t every finite set of inputs is normally distributed:

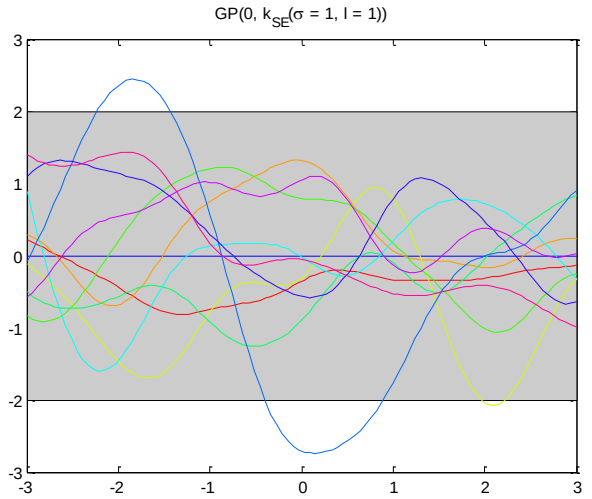
$$\forall n, \forall (X_1, \dots, X_n) \in (\mathbb{R}^m)^n, \big(f(X_1), \dots, f(X_n)\big) \text{ is normally distributed}$$

$$\text{i.e. } \big(f(X_1), \dots, f(X_n)\big) \sim \mathcal{N}\left(\begin{bmatrix} \mu(X_1) \\ \vdots \\ \mu(X_n) \end{bmatrix}, \begin{bmatrix} k(X_1, X_1) & \cdots & k(X_1, X_n) \\ \vdots & \ddots & \vdots \\ k(X_n, X_1) & \cdots & k(X_n, X_n) \end{bmatrix} \right)$$

$$\text{where } \begin{cases} \mu: & X \in \mathbb{R}^m & \longmapsto & E\big(f(X)\big) \\ k: & (X, X') \in (\mathbb{R}^m)^2 & \longmapsto & E\left(\big(f(X) - \mu(X)\big)\big(f(X') - \mu(X')\big)^T\right) \end{cases}$$

It is fully defined by μ and k and thus denoted by $\mathbf{f} \sim \boldsymbol{\mathcal{GP}}(\boldsymbol{\mu}, \boldsymbol{k})$

Gaussian Process: Interpretation



95 % confidence interval

- GP generalizes multivariate normal distribution to infinite dimension
- GP are **stochastic functions**: a random sample is a function $\mathbb{R}^m \to \mathbb{R}$

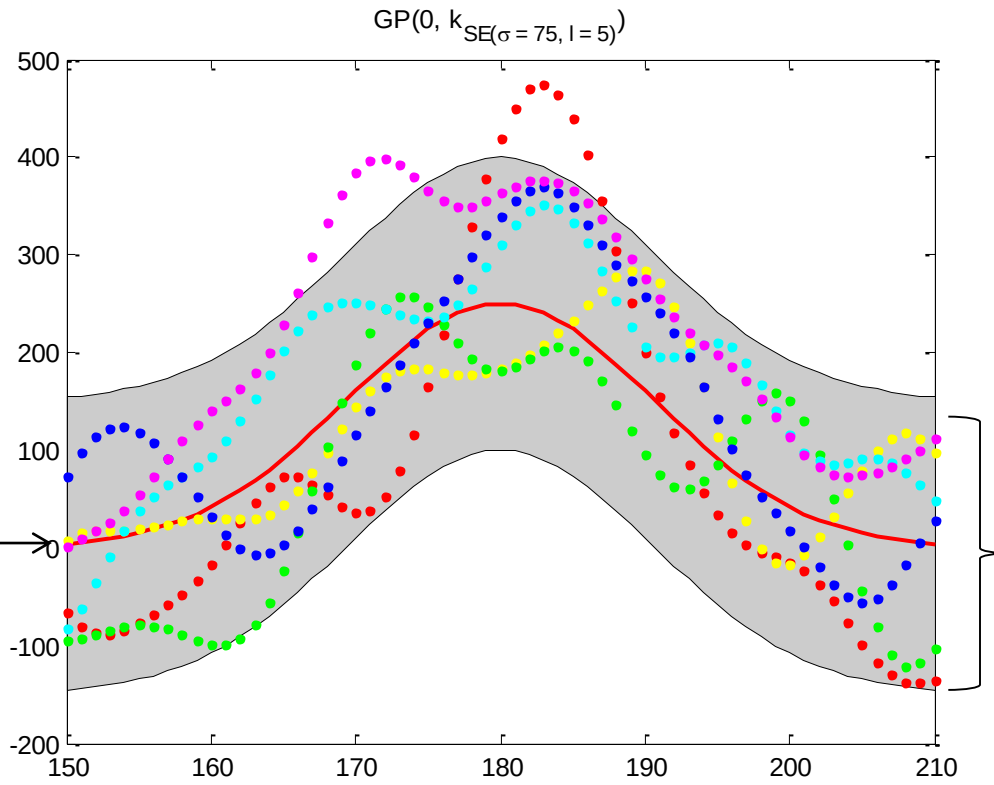
  $\rightarrow$ Chance for a sample to be far from $x \mapsto \mu(x)$ is small (see conf. interval)

- Random samples are smooth as $f(X) - f(X') \xrightarrow{X \to X'} \mathcal{N}(0,0)$

$$E\big(f(X) - f(X')\big) = \mu(X) - \mu(X') \xrightarrow{X \to X'} 0$$

$$E\left(\big(f(X) - f(X')\big)^2\right) = k(X,X) + k(X',X') - 2 \times k(X,X') \xrightarrow{X \to X'} 0$$

Gaussian Process:
How to represent $f \sim \mathcal{GP}(\mu, k)$?



Mean $\mu(x)$

95% confidence interval:
$[\mu(x) \pm 2\sigma(x)]$
with $\sigma(x) = \sqrt{k(x,x)}$

Method to draw samples:

Draw samples $f\left(X_i = (x_{max} - x_{\min})\frac{i}{n} + x_{min}\right)$ for $i = 0 \ldots n$

Gaussian Process: Bayesian Inference

Given $\mathrm{f} \sim \mathcal{GP}(\mu, k)$, suppose one observes $\mathcal{O} = \{(X_1, y_1), \ldots, (X_k, y_k)\}$

What is $\mathrm{f}|\mathcal{O}$?

It is still a Gaussian process!!

Proof?

$\forall (X'_1, \ldots, X'_n) \in (\mathbb{R}^m)^n, \big(f(X'_1), \ldots, f(X'_n), f(X_1), \ldots, f(X_k)\big)$ is normal

thus

$(f(X'_1), \ldots, f(X'_n) | \, f(X_1) = y_1, \ldots, f(X_k) = y_k)$ is also normal

Gaussian Process:
Posterior Gaussian Process

- Let $X_o = [X_1, \dots, X_k]^{\mathrm{T}}$ and $Y_o = [y_1, \dots, y_k]^{\mathrm{T}}$ be the observations $\mathcal{O}$
- Let $X_p = [X'_1, \dots, X'_n]^{\mathrm{T}}$ be the points for which $\mathrm{f}|\mathcal{O}$ has to be estimated
- Let $\mu_o = \begin{bmatrix} \mu(X_1) \\ \vdots \\ \mu(X_k) \end{bmatrix}, \Sigma_{oo} = \begin{bmatrix} k(X_1, X_1) & \cdots & k(X_1, X_k) \\ \vdots & \ddots & \vdots \\ k(X_k, X_1) & \cdots & k(X_k, X_k) \end{bmatrix}$, idem for μ_p, Σ_{pp} and Σ_{po}

Then $\mathrm{f}\left(\begin{bmatrix} X_p \\ X_o \end{bmatrix}\right) \sim \mathcal{N}\left(\begin{bmatrix} \mu_p \\ \mu_o \end{bmatrix}, \begin{bmatrix} \Sigma_{pp} & \Sigma_{po} \\ \Sigma_{po}^T & \Sigma_{oo} \end{bmatrix}\right)$ thus

$$f(X_p)|f(X_o) = Y_o \sim \mathcal{N}\left(\mu_p + \Sigma_{po}{\Sigma_{oo}}^{-1}(Y_o - \mu_o), \Sigma_{pp} - \Sigma_{po}{\Sigma_{oo}}^{-1}\Sigma_{po}^T\right)$$

Gaussian Process: How to choose prior $f \sim \mathcal{GP}(\mu, k)$?

1. Function $X \mapsto \mu(X)$ must be continuous
2. Function $k(X, X')$ must be a **positive definite kernel** :
   - For all finite point sampling, $\begin{bmatrix} k(X_1, X_1) & \cdots & k(X_1, X_n) \\ \vdots & \ddots & \vdots \\ k(X_n, X_1) & \cdots & k(X_n, X_n) \end{bmatrix}$ is positive semi-definite.
   - Usually decreasing function of $\|X - X'\|$

Examples of kernels:

– Exponential kernel: $k_{E(\gamma,\sigma,l)}(X, X') = \sigma^2 e^{-\left\|\frac{X-X'}{l}\right\|^{\gamma}}$

– Squared exponential kernel: $k_{SE(\sigma,l)}(X, X') = k_{E(2,\sigma,l)}(X, X')$

– Addition: $k(X, X') = k_1(X, X') + k_2(X, X')$

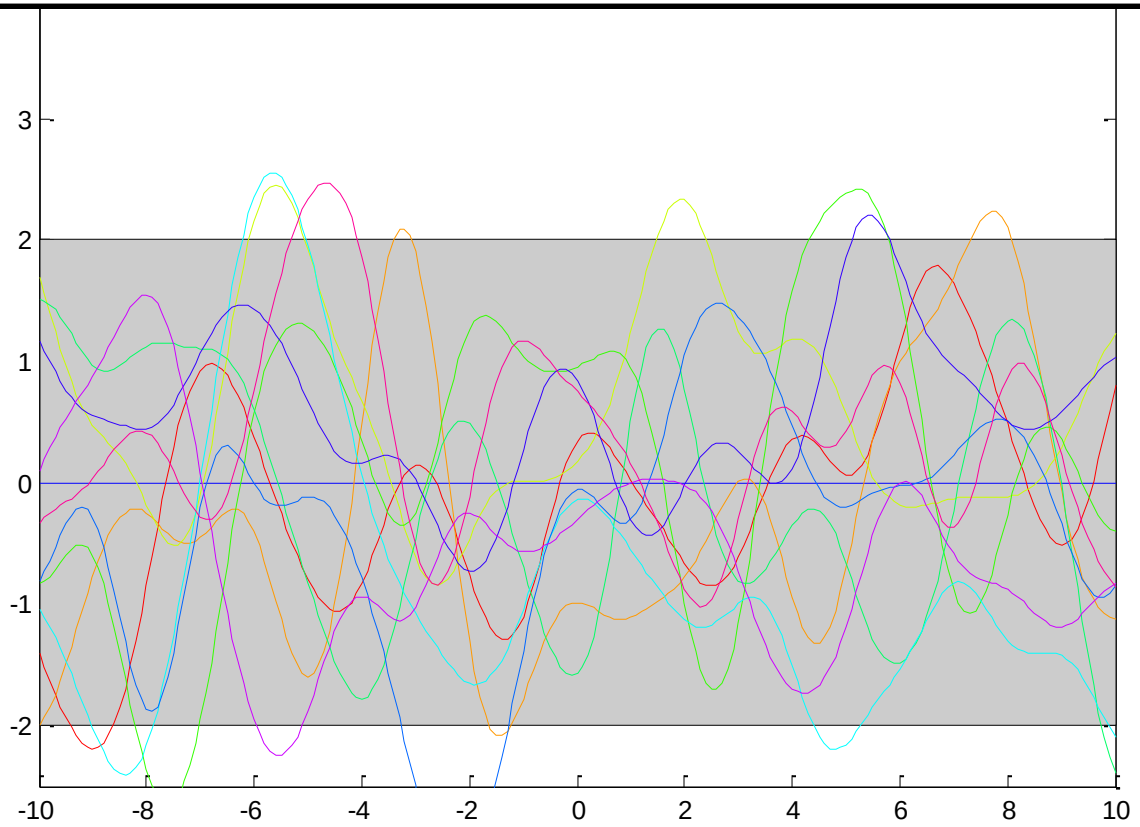
– Multiplication: $k(X, X') = k_1(X, X') \times k_2(X, X')$

– Change of referential: $k(X, X') = k_1\big(f(X), f(X')\big)$

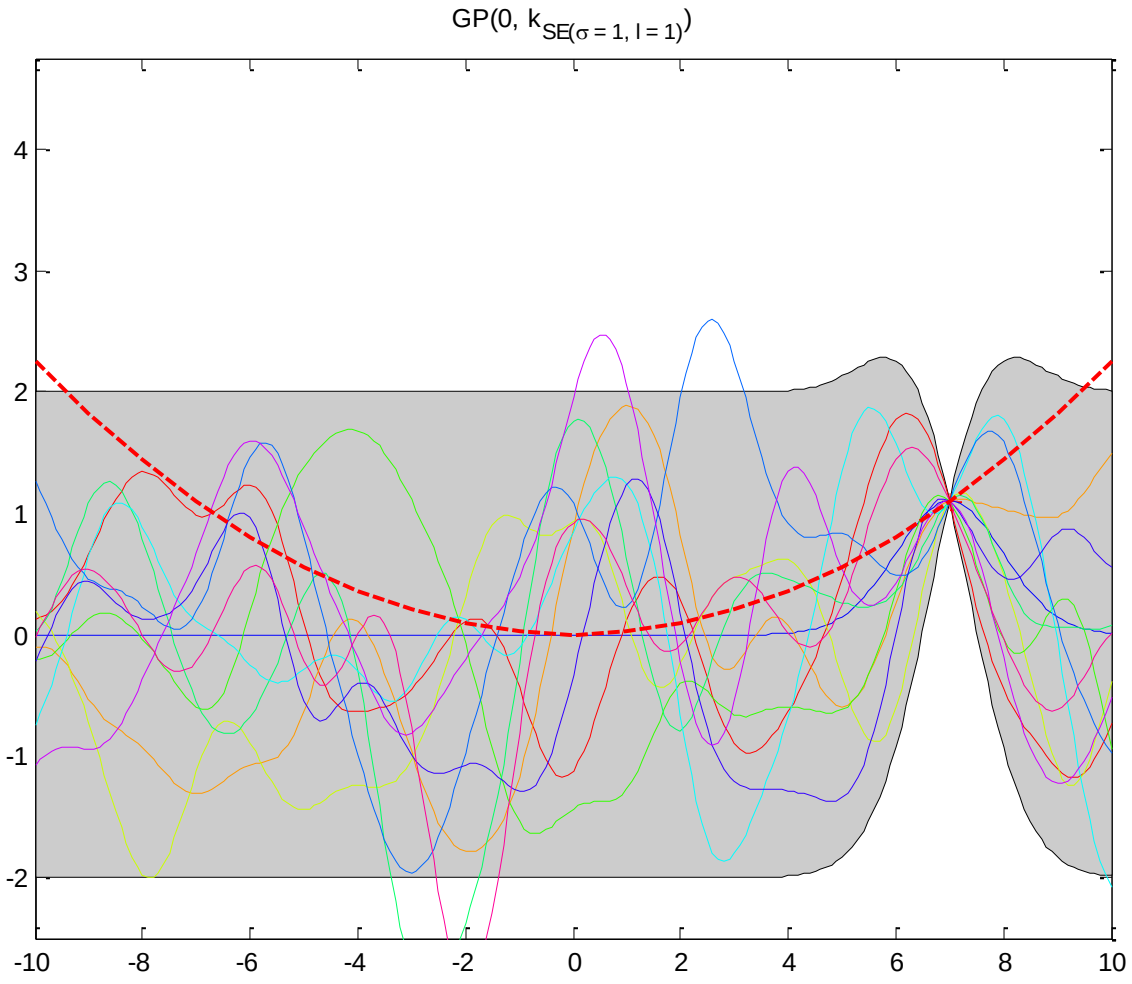
Gaussian Process:
Example of bayesian inference

Prior: $f \sim \mathcal{GP}\left(0^{\mathbb{R}}, k\right)$ with $k_{SE(\sigma,l)}(X, X') = \sigma^2 e^{-\frac{(X-X')^2}{2l^2}}, \sigma = 1, l = 1$

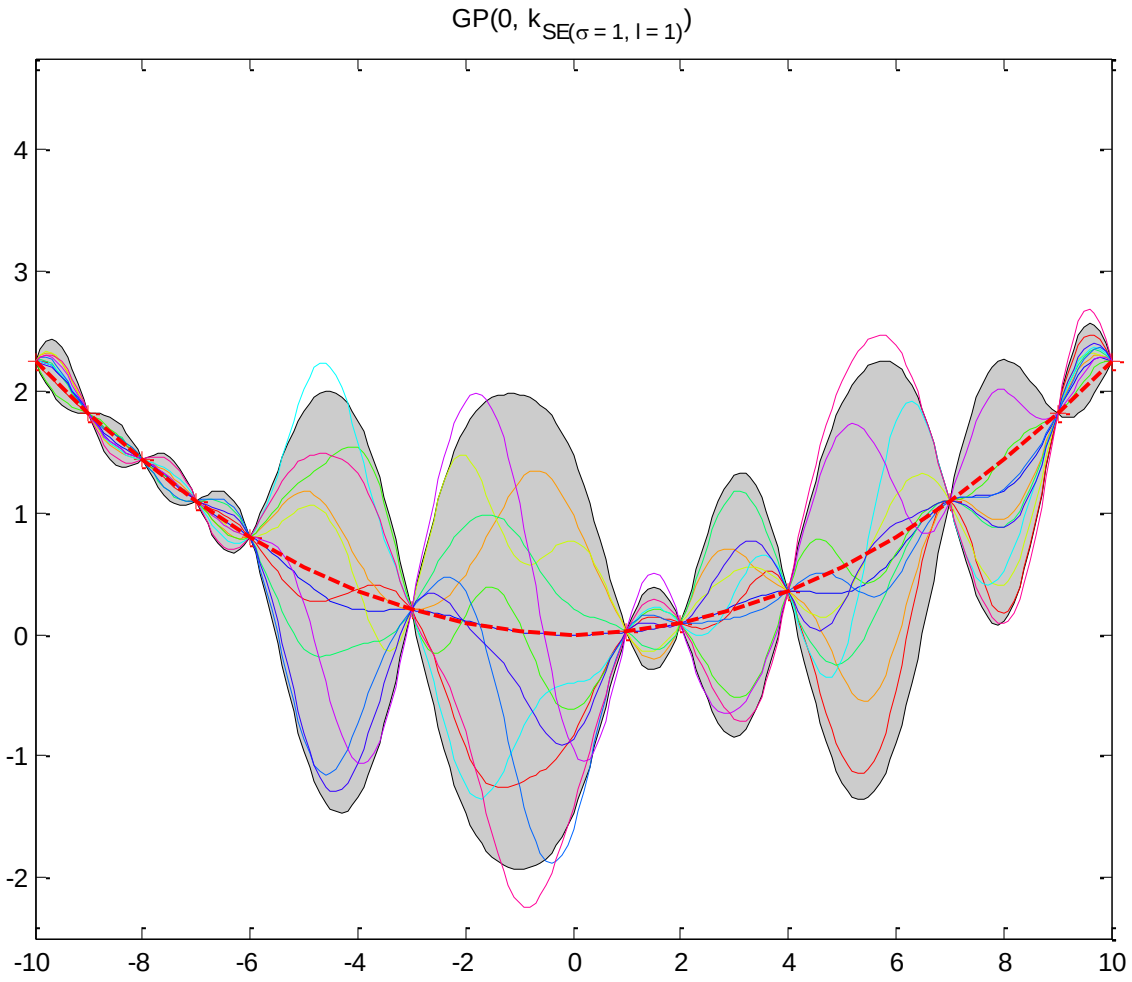
Demo 2



Gaussian Process:
Example of bayesian inference



Gaussian Process: Example of bayesian inference



Gaussian Process: Regularization through noisy observations

Recall:

$$\boxed{f(X_p)|f(X_o) = Y_o \sim \mathcal{N}\left(\mu_p + \Sigma_{po}{\Sigma_{oo}}^{-1}(Y_o - \mu_o), \Sigma_{pp} - \Sigma_{po}{\Sigma_{oo}}^{-1}\Sigma_{po}^T\right)}$$

Problem: Σ_{oo} might be ill-conditioned when observation points are too close.

Solution: introduce a measurement noise:

$$Y = f(X) + \varepsilon \text{ with white noise } \varepsilon \sim \mathcal{N}(0, \sigma_n^2)$$

- Theoretically regularizes Σ_{oo} by adding a ridge on main diagonal

$$\boxed{f(X_p)|f(X_o) = Y_o \sim \mathcal{N}\left(\mu_p + \Sigma_{po}(\Sigma_{oo} + \sigma_n^2 I_k)^{-1}(Y_o - \mu_o), \Sigma_{pp} - \Sigma_{po}(\Sigma_{oo} + \sigma_n^2 I_k)^{-1}\Sigma_{po}^T\right)}$$

- Practically, allows close points to disagree

Gaussian Process:
An example of non-parametric model

- **Parametric model:** model parameters Θ have to be specified

$$P(\mathrm{Y}|\mathrm{X},\mathcal{O}) = \int_{\Theta} P(Y|X,\Theta)\cdot P(\Theta|\mathcal{O})d\Theta = \int_{\Theta} f_{pred}(X,Y,\Theta)\cdot f_{learn}(\Theta,\mathcal{O})d\Theta$$

  E.g. linear models

- **Non-parametric model:** no model parameters (or non measurable)

$$P(\mathrm{Y}|\mathrm{X},\mathcal{O}) = f_{pred}(X,Y,\mathcal{O})$$

  E.g. Gaussian processes

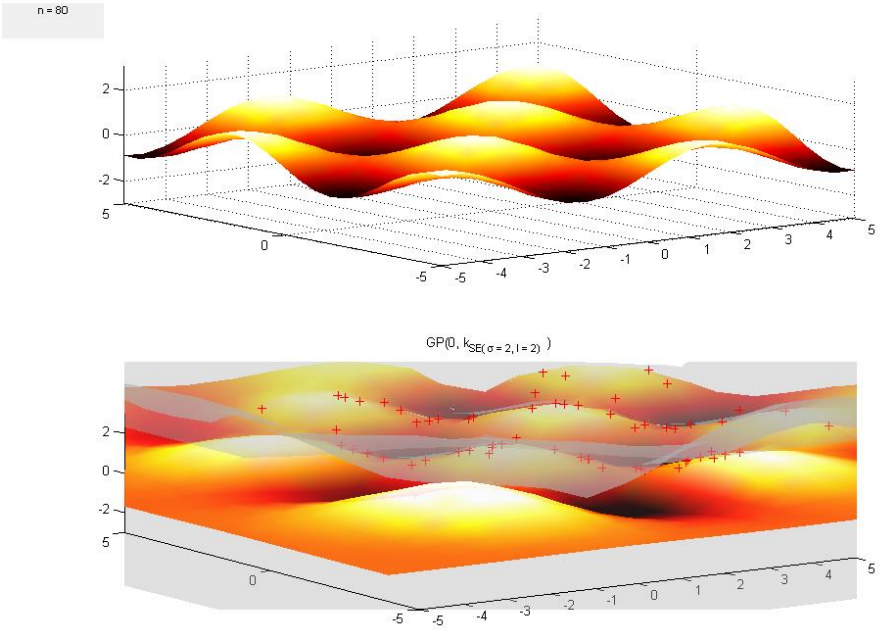
  - Merge of learning & prediction steps
  - Parameters are the observations: infinite number of parameters

☺ Light assumptions, fit input distribution
☹ Prediction gets intractable when n grows (not for real-time systems)

Gaussian Process: Generalization to higher dimensions

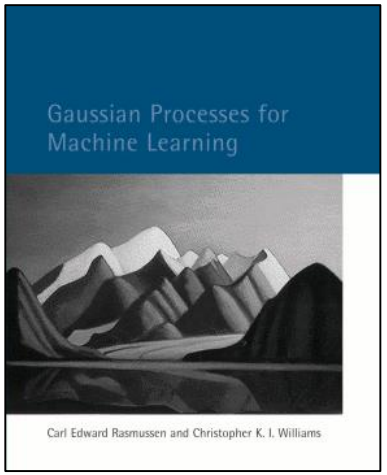
- Good scalability with number of dimensions $O(n^3 + n^2 m)$
- Dual properties of linear regression
- Very useful in geostatistics

Demo 3



Gaussian Process: A summary

- Generalizes multivariate normal distribution to processes
- Is a fully Bayesian regression tool:
  - Output is a distribution
  - Take into account prior distribution
- Is a non parametric model
- Very powerful regression tool
  - Based on modeling spatial / temporal correlations
  - Wide choice for prior hyper parameters
  - Take into account noisy observations
  - Scale nicely with number of dimensions
- Scale poorly with number of examples



Rasmussen 2006