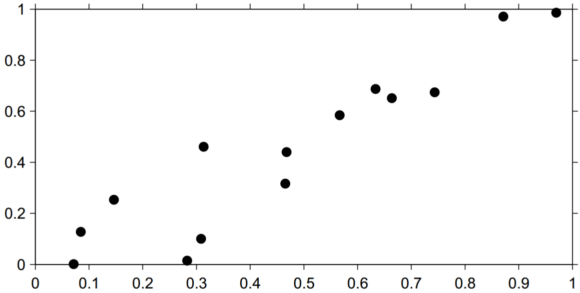
Gaussian processes introduction

Victor Kocheganov

February 19, 2016

Problem statement



Training data: $\left\{(x_i, y_i) \colon i = \overline{1, N}\right\}$, $x_i, y_i \in R$
Task: given new x^* predict y^*

Three ways to go

Linear regression

Bayesian regression

Gaussian Process

Linear regression

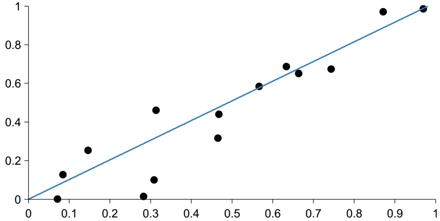
Utilize **model**

$$y = mx,$$

One should solve

$$m^* = arg \min_{m \in R} \sum_{i=1}^{N} (y_i - mx_i)^2$$

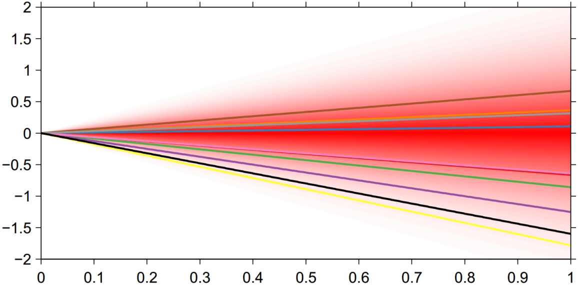
and get $m^* = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$



Bayesian regression

Utilize **model**

$$y_i = mx_i + \varepsilon_i, \quad \varepsilon_i \sim_{\text{iid}} N(0, \sigma_\varepsilon^2), \quad m \sim N(0, 1).$$



The goal is to find $p(m|y, x)$

Bayesian formula

$$p(m|y,x) = \frac{p(y|m,x)p(m)}{\int p(y|m,x)p(m)dm}$$
$$\propto p(y|m,x)p(m)$$

Bayesian formula

$$
\begin{aligned}
p(m|y,x) &= \frac{p(y|m,x)p(m)}{\int p(y|m,x)p(m)dm} \\
&\propto p(y|m,x)p(m) \\
&\propto \left(\prod_i \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} e^{-(y_i-mx_i)^2/(2\sigma_\varepsilon^2)}\right) \frac{1}{\sqrt{2\pi}e^{-m^2/2}}
\end{aligned}
$$

Bayesian formula

$$
\begin{aligned}
p(m|y,x) &= \frac{p(y|m,x)p(m)}{\int p(y|m,x)p(m)dm} \\
&\propto p(y|m,x)p(m) \\
&\propto \left(\prod_i \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} e^{-(y_i-mx_i)^2/(2\sigma_\varepsilon^2)}\right)\frac{1}{\sqrt{2\pi}e^{-m^2/2}} \\
&\propto e^{-\sum_i (y_i-mx_i)^2/(2\sigma_\varepsilon^2)-m^2/2}
\end{aligned}
$$

Bayesian formula

$$
\begin{aligned}
p(m|y,x) &= \frac{p(y|m,x)p(m)}{\int p(y|m,x)p(m)dm} \\
&\propto p(y|m,x)p(m) \\
&\propto \left(\prod_i \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} e^{-(y_i - m x_i)^2/(2\sigma_\varepsilon^2)}\right) \frac{1}{\sqrt{2\pi} e^{-m^2/2}} \\
&\propto e^{-\sum_i (y_i - m x_i)^2/(2\sigma_\varepsilon^2) - m^2/2} \\
&\propto e^{-\frac{\sum_i x_i^2 + \sigma_\varepsilon^2}{2\sigma_\varepsilon^2}\left(m - \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}\right)^2}
\end{aligned}
$$

Bayesian formula

$$
\begin{aligned}
p(m|y,x) &= \frac{p(y|m,x)p(m)}{\int p(y|m,x)p(m)dm} \\
&\propto p(y|m,x)p(m) \\
&\propto \left(\prod_i \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} e^{-(y_i-mx_i)^2/(2\sigma_\varepsilon^2)}\right) \frac{1}{\sqrt{2\pi}e^{-m^2/2}} \\
&\propto e^{-\sum_i (y_i-mx_i)^2/(2\sigma_\varepsilon^2)-m^2/2} \\
&\propto e^{-\frac{\sum_i x_i^2+\sigma_\varepsilon^2}{2\sigma_\varepsilon^2}\left(m-\frac{\sum_i x_i y_i}{\sum_i x_i^2+\sigma_\varepsilon^2}\right)^2}
\end{aligned}
$$

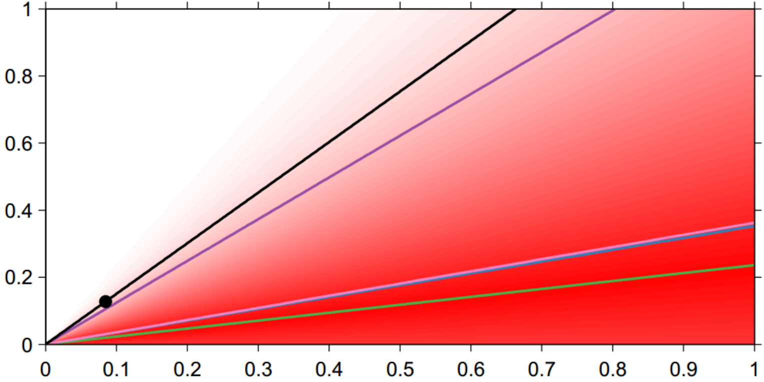
That is: $m|y,x \sim \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2+\sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2+\sigma_\varepsilon^2}\right)$

Bayesian formula

$$
\begin{aligned}
p(m|y,x) &= \frac{p(y|m,x)p(m)}{\int p(y|m,x)p(m)dm} \\
&\propto p(y|m,x)p(m) \\
&\propto \left(\prod_i \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} e^{-(y_i-mx_i)^2/(2\sigma_\varepsilon^2)}\right) \frac{1}{\sqrt{2\pi} e^{-m^2/2}} \\
&\propto e^{-\sum_i (y_i-mx_i)^2/(2\sigma_\varepsilon^2)-m^2/2} \\
&\propto e^{-\frac{\sum_i x_i^2+\sigma_\varepsilon^2}{2\sigma_\varepsilon^2}\left(m-\frac{\sum_i x_i y_i}{\sum_i x_i^2+\sigma_\varepsilon^2}\right)^2}
\end{aligned}
$$

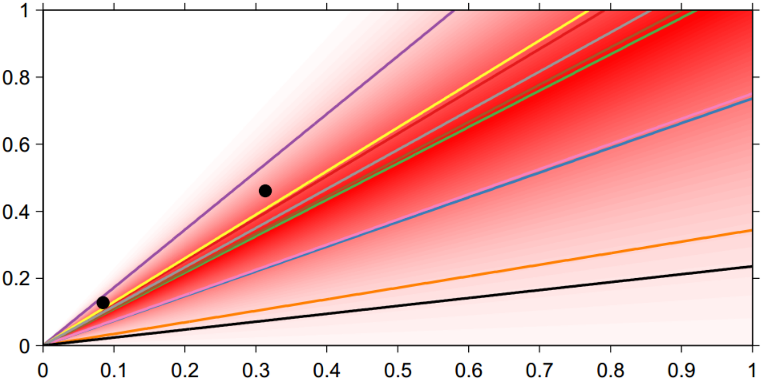
That is: $m|y,x \sim \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2+\sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2+\sigma_\varepsilon^2}\right)$

Parameters posterior



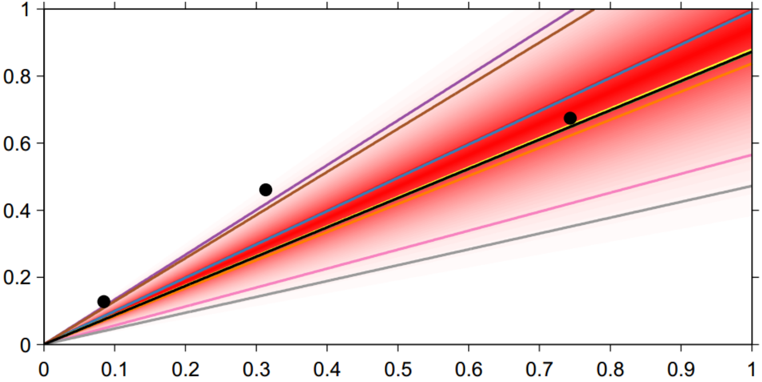
$$m|y,x \sim \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2}\right)$$

Parameters posterior



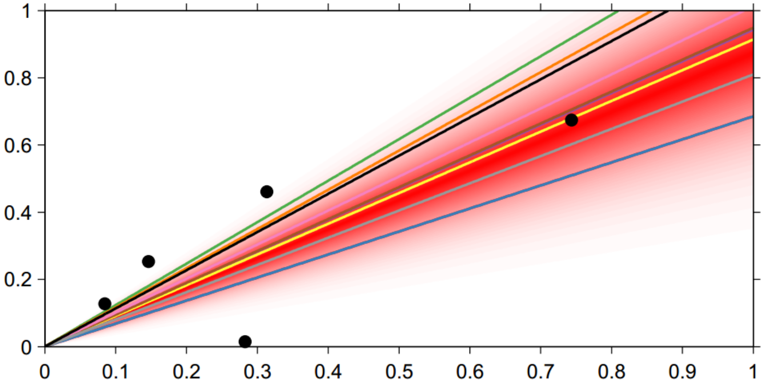
$$m|y, x \sim \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2}\right)$$

Parameters posterior



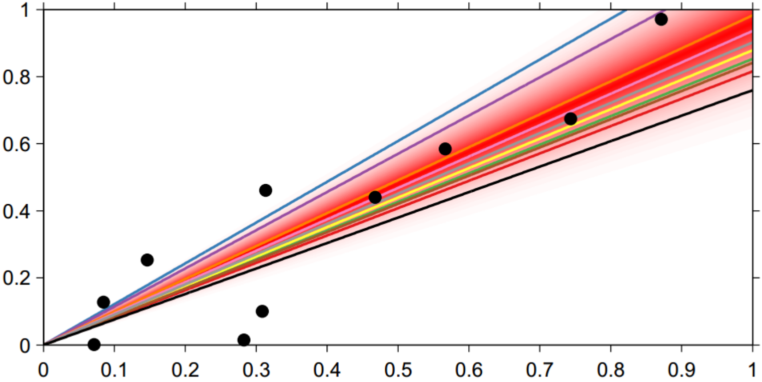
$$m|y,x \sim \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2}\right)$$

Parameters posterior



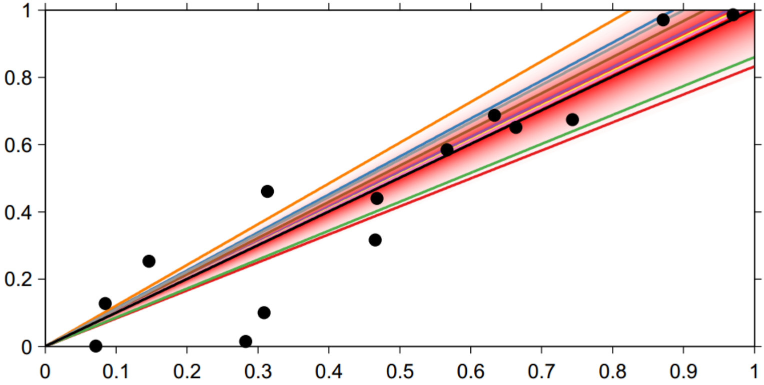
$$m|y,x \sim \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2}\right)$$

Parameters posterior



$$m|y, x \sim \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2}\right)$$

Parameters posterior



$$m|y,x \sim \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2}\right)$$

Output posterior

$$p(y^*|x^*, x, y) = \int p(y^*|x^*, m)p(m|x, y)dm$$
$$= \mathcal{N}\left(x^* \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, (x^*)^2 \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2}\right)$$

In more general case $m \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$ and input space is multidimensional:

$$y^*|x^*, x, y \sim \mathcal{N}\left(\frac{1}{\sigma_\varepsilon^2}\mathbf{x}^{*\intercal}A^{-1}X\mathbf{y}, \mathbf{x}^{*\intercal}A^{-1}\mathbf{x}^*)\right),$$

where $A = \sigma_\varepsilon^{-2}XX^\intercal + \Sigma_p^{-1}$

Output posterior

$$
\begin{aligned}
p(y^*|x^*, x, y) &= \int p(y^*|x^*, m)p(m|x, y)dm \\
&= \mathcal{N}\left(x^* \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sigma_\varepsilon^2}, (x^*)^2 \frac{\sigma_\varepsilon^2}{\sum_i x_i^2 + \sigma_\varepsilon^2}\right)
\end{aligned}
$$

In more general case $m \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$ and input space is multidimensional:

$$
y^*|x^*, x, y \sim \mathcal{N}\left(\frac{1}{\sigma_\varepsilon^2}\mathbf{x}^{*\intercal} A^{-1} X\mathbf{y}, \mathbf{x}^{*\intercal} A^{-1}\mathbf{x}^*)\right),
$$

where $A = \sigma_\varepsilon^{-2} XX^\intercal + \Sigma_p^{-1}$

Input space to feature space

Project input space with transformation $\phi(\mathsf{x})\colon \mathrm{R}^n \to \mathrm{R}^s$.

$$\Phi = \Phi(X) = (\phi(\mathsf{x}_1), \phi(\mathsf{x}_2), \ldots, \phi(\mathsf{x}_N))$$

Model takes the form:

$$\mathbf{y} = \phi(\mathsf{x})^\intercal \mathbf{m}$$

Input space to feature space

Project input space with transformation $\phi(\mathbf{x})\colon \mathrm{R}^n \to \mathrm{R}^s$.

$$\Phi = \Phi(X) = (\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_N))$$

Model takes the form:

$$\mathbf{y} = \phi(\mathbf{x})^\intercal \mathbf{m}$$

$$y^*|x^*, x, y \sim \mathcal{N}\left(\phi^{*\intercal}\Sigma_p\Phi^{-1}(K + \sigma_\varepsilon^2 I)^{-1}\mathbf{y},\right.$$
$$\left.\phi^{*\intercal}\Sigma_p\phi^* - \phi^{*\intercal}\Sigma_p\Phi(K + \sigma_\varepsilon^2 I)^{-1}\Phi^\intercal\Sigma_p\phi^*\right),$$

Input space to feature space

Project input space with transformation $\phi(\mathbf{x})\colon \mathrm{R}^n \to \mathrm{R}^s$.

$$\Phi = \Phi(X) = (\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_N))$$

Model takes the form:

$$\mathbf{y} = \phi(\mathbf{x})^\intercal \mathbf{m}$$

$$y^*|x^*, x, y \sim \mathcal{N}\left(\phi^{*\intercal}\Sigma_p\Phi^{-1}(K + \sigma_\varepsilon^2 I)^{-1}\mathbf{y},\right.$$
$$\left.\phi^{*\intercal}\Sigma_p\phi^* - \phi^{*\intercal}\Sigma_p\Phi(K + \sigma_\varepsilon^2 I)^{-1}\Phi^\intercal\Sigma_p\phi^*\right),$$

where $\phi^* = \phi(\mathbf{x}^*)$ and $K = \Phi^\intercal\Sigma_p\Phi$

Input space to feature space

Project input space with transformation $\phi(\mathbf{x})\colon \mathrm{R}^n \to \mathrm{R}^s$.

$$\Phi = \Phi(X) = (\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \dots, \phi(\mathbf{x}_N))$$

Model takes the form:

$$\mathbf{y} = \phi(\mathbf{x})^\intercal \mathbf{m}$$

$$y^*|x^*, x, y \sim \mathcal{N}\left(\phi^{*\intercal}\Sigma_p \Phi^{-1}(K + \sigma_\varepsilon^2 I)^{-1}\mathbf{y},\right.$$
$$\left.\phi^{*\intercal}\Sigma_p\phi^* - \phi^{*\intercal}\Sigma_p\Phi(K + \sigma_\varepsilon^2 I)^{-1}\Phi^\intercal\Sigma_p\phi^*\right),$$

where $\phi^* = \phi(\mathbf{x}^*)$ and $K = \Phi^\intercal\Sigma_p\Phi$

From Bayesian regression to GP

- We defined a Bayesian linear regression model by specifying priors on m and ε_i
  - $m \sim \mathcal{N}$, $\varepsilon_i \sim_{\text{iid}} \mathcal{N}(0, \sigma_\varepsilon^2)$

From Bayesian regression to GP

- We defined a Bayesian linear regression model by specifying priors on m and ε_i
  - $m \sim \mathcal{N}$, $\varepsilon_i \sim_{\text{iid}} \mathcal{N}(0, \sigma_\varepsilon^2)$
  - $y_i | m, \varepsilon_i = m x_i + \varepsilon_i$

From Bayesian regression to GP

- We defined a Bayesian linear regression model by specifying priors on m and ε_i
  - $m \sim \mathcal{N}$, $\varepsilon_i \sim_{\text{iid}} \mathcal{N}(0, \sigma_\varepsilon^2)$
  - $y_i | m, \varepsilon_i = mx_i + \varepsilon_i$
- This implicitly defined a joint prior on $\{y_i : i = \overline{1, n}\}$

From Bayesian regression to GP

- We defined a Bayesian linear regression model by specifying priors on m and ε_i
  - $m \sim \mathcal{N}$, $\varepsilon_i \sim_{\text{iid}} \mathcal{N}(0, \sigma_\varepsilon^2)$
  - $y_i | m, \varepsilon_i = m x_i + \varepsilon_i$
- This implicitly defined a joint prior on $\{y_i \colon i = \overline{1, n}\}$
  - $y_i \sim \mathcal{N}(0, x_i^2 + \sigma_\varepsilon^2)$

From Bayesian regression to GP

- We defined a Bayesian linear regression model by specifying priors on m and ε_i
    - $m \sim \mathcal{N}$, $\varepsilon_i \sim_{\text{iid}} \mathcal{N}(0, \sigma_\varepsilon^2)$
    - $y_i | m, \varepsilon_i = m x_i + \varepsilon_i$
- This implicitly defined a joint prior on $\{y_i \colon i = \overline{1, n}\}$
    - $y_i \sim \mathcal{N}(0, x_i^2 + \sigma_\varepsilon^2)$
    - $cov(y_i, y_j) = x_i x_j, \forall i \neq j$

From Bayesian regression to GP

- We defined a Bayesian linear regression model by specifying priors on m and ε_i
  - $m \sim \mathcal{N}$, $\varepsilon_i \sim_{\text{iid}} \mathcal{N}(0, \sigma_\varepsilon^2)$
  - $y_i | m, \varepsilon_i = m x_i + \varepsilon_i$
- This implicitly defined a joint prior on $\{y_i \colon i = \overline{1, n}\}$
  - $y_i \sim \mathcal{N}(0, x_i^2 + \sigma_\varepsilon^2)$
  - $cov(y_i, y_j) = x_i x_j$, $\forall i \neq j$

That is $(y_1, y_2, \ldots, y_n)$ has multivariate normal distribution:

$$y \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \quad k_{ij} = x_i x_j + \delta_{ij}\sigma_\varepsilon^2$$

From Bayesian regression to GP

- We defined a Bayesian linear regression model by specifying priors on m and ε_i
  - $m \sim \mathcal{N}$, $\varepsilon_i \sim_{\text{iid}} \mathcal{N}(0, \sigma_\varepsilon^2)$
  - $y_i | m, \varepsilon_i = m x_i + \varepsilon_i$
- This implicitly defined a joint prior on $\{y_i \colon i = \overline{1, n}\}$
  - $y_i \sim \mathcal{N}(0, x_i^2 + \sigma_\varepsilon^2)$
  - $cov(y_i, y_j) = x_i x_j$, $\forall i \neq j$

That is $(y_1, y_2, \ldots, y_n)$ has multivariate normal distribution:

$$y \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \quad k_{ij} = x_i x_j + \delta_{ij} \sigma_\varepsilon^2$$

Gaussian processes

Definition

Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution

We can write this collection of random variables as

$$\{f(x) \colon x \in \mathcal{X}\}$$

i.e. a function f evaluated at inputs x.

Gaussian processes

Definition

Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution

We can write this collection of random variables as

$$\{f(x) : x \in \mathcal{X}\}$$

i.e. a function f evaluated at inputs x.

GP(μ,k) is completely specified by

Gaussian processes

Definition

Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution

We can write this collection of random variables as

$$\{f(x) \colon x \in \mathcal{X}\}$$

i.e. a function f evaluated at inputs x.

GP(μ,k) is completely specified by

- Mean function $\mu(x) = \mathrm{E}(f(x))$

Gaussian processes

Definition

Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution

We can write this collection of random variables as

$$\{f(x) \colon x \in \mathcal{X}\}$$

i.e. a function f evaluated at inputs x.

GP(μ,k) is completely specified by

- Mean function $\mu(x) = \mathrm{E}(f(x))$
- Covariance/kernel function $k(x, x') = Cov(f(x), f(x'))$

Gaussian processes

Definition

Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution

We can write this collection of random variables as

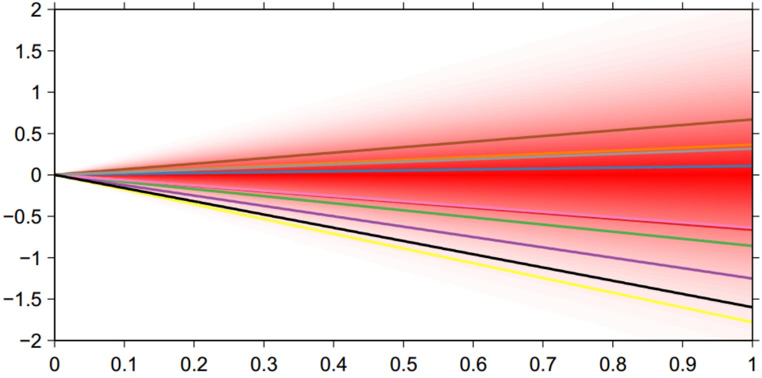
$$\{f(x) \colon x \in \mathcal{X}\}$$

i.e. a function f evaluated at inputs x.

GP(μ,k) is completely specified by

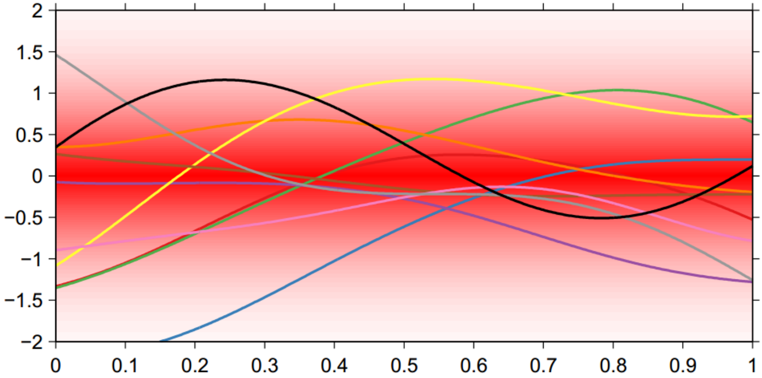
- Mean function $\mu(x) = \mathrm{E}(f(x))$
- Covariance/kernel function $k(x, x') = Cov(f(x), f(x'))$

Linear kernel



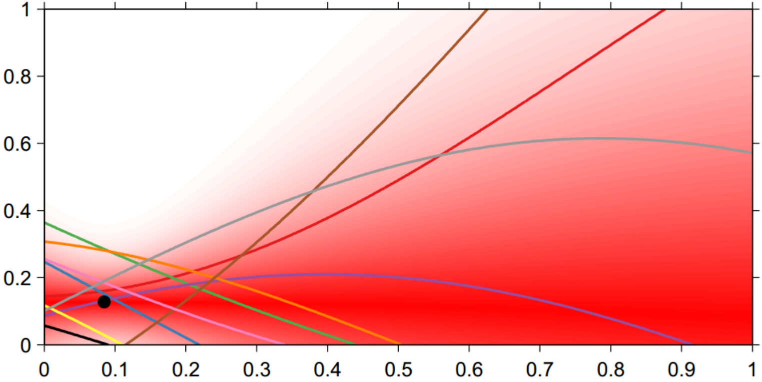
$$k(x, x') = xx'$$

Exponential kernel

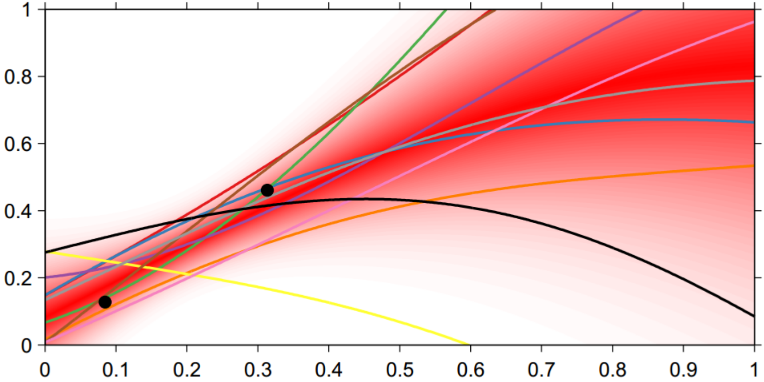


$$k(x, x') = e^{-(x-x')^2}$$

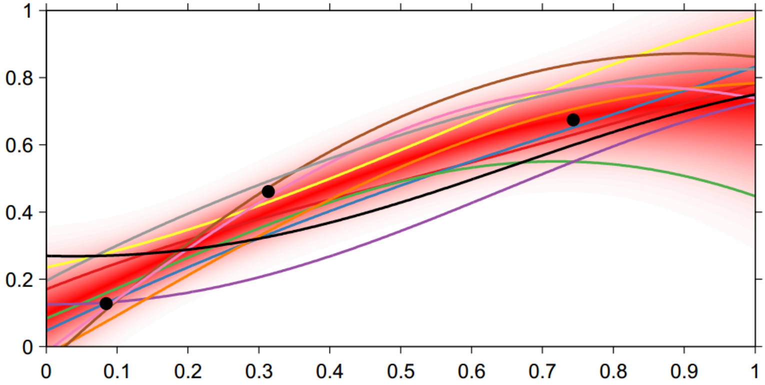
Non-linear kernel



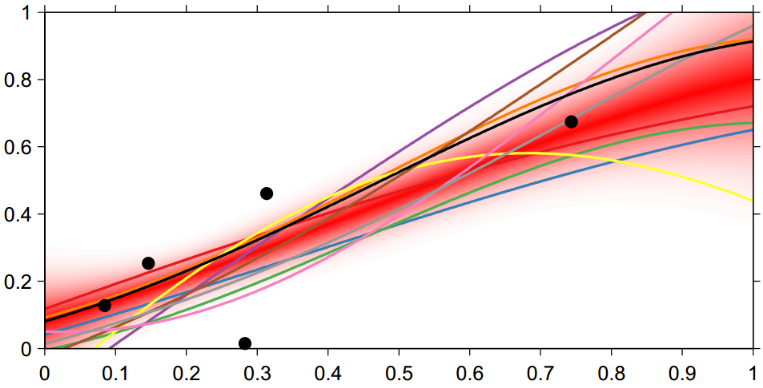
Non-linear kernel



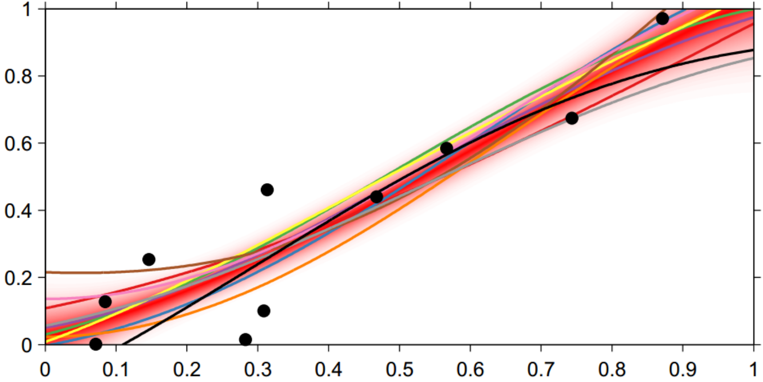
Non-linear kernel



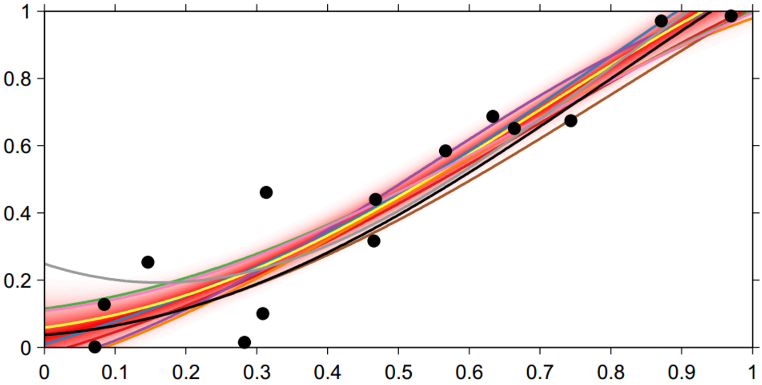
Non-linear kernel



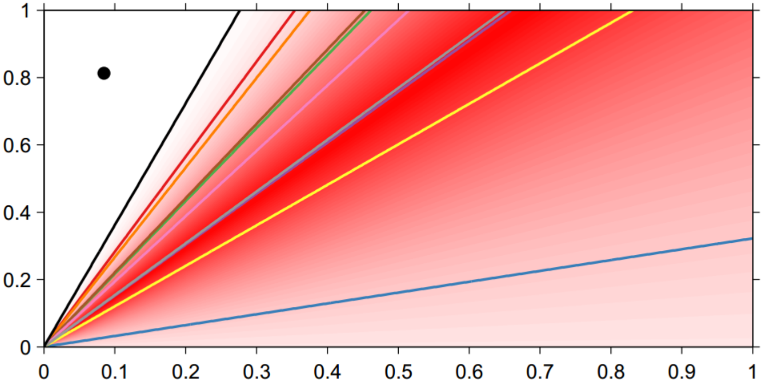
Non-linear kernel



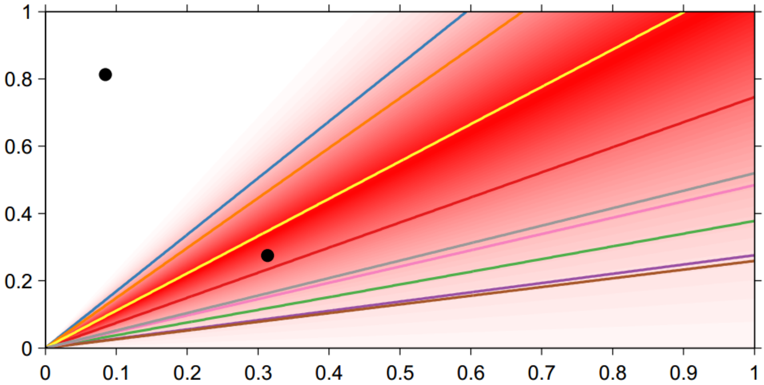
Non-linear kernel



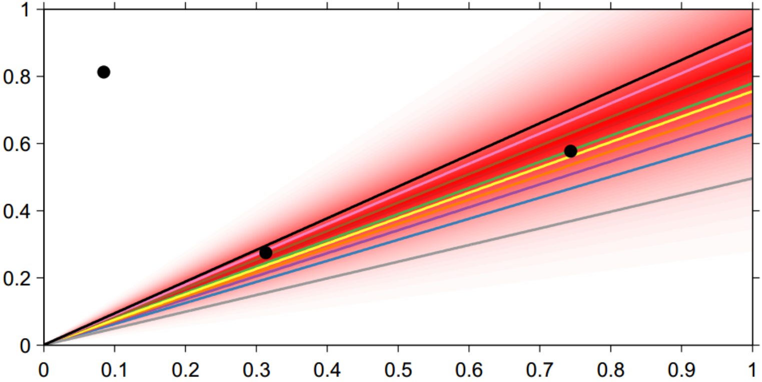
Linear model gone wrong



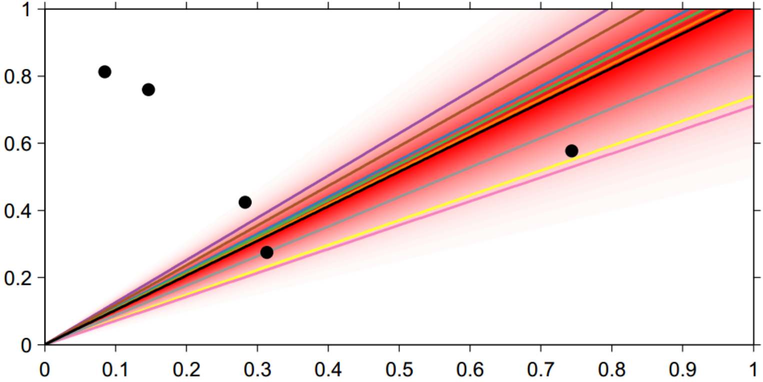
Linear model gone wrong



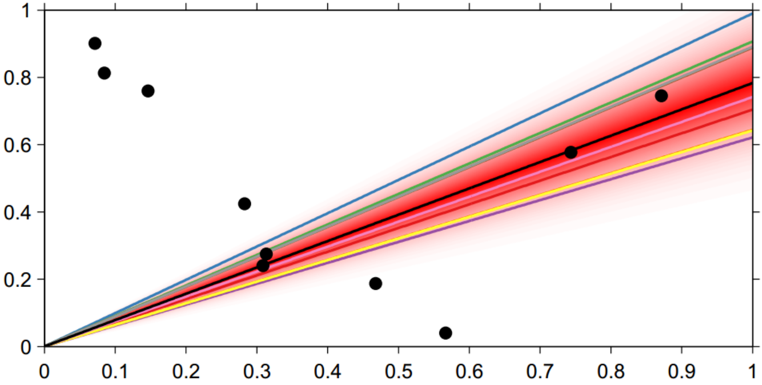
Linear model gone wrong



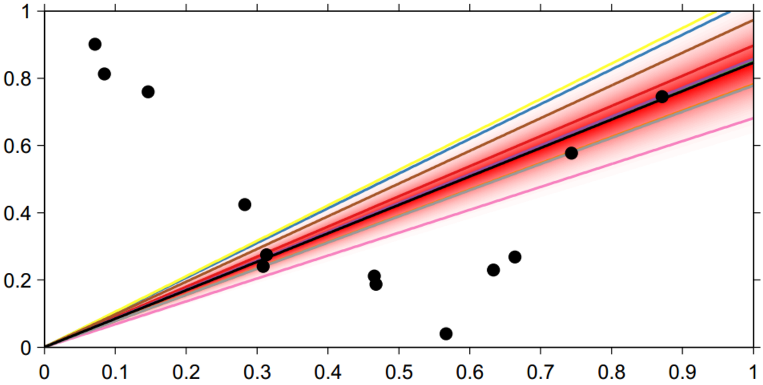
Linear model gone wrong



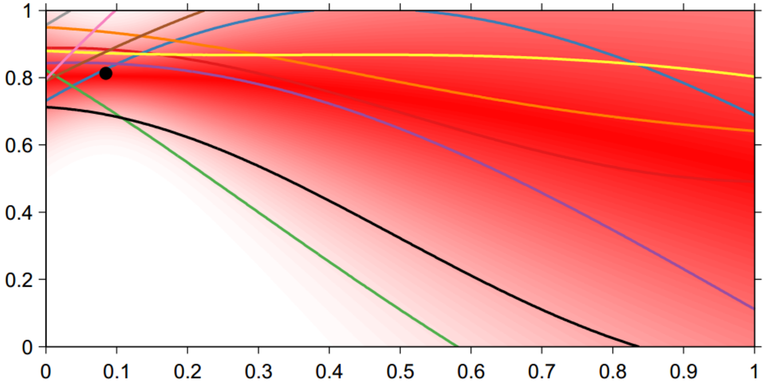
Linear model gone wrong



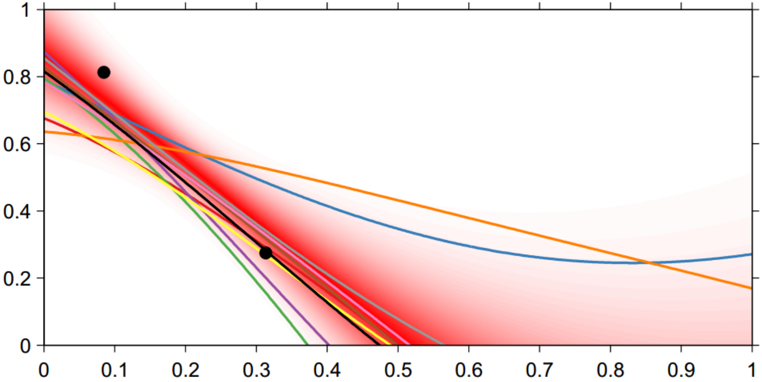
Linear model gone wrong



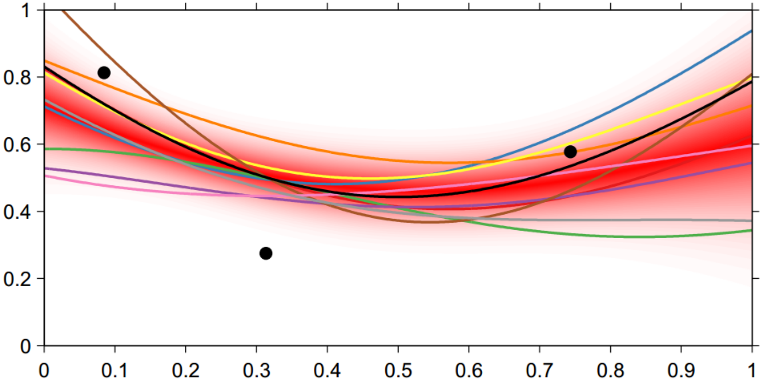
Non-linearity to the rescue



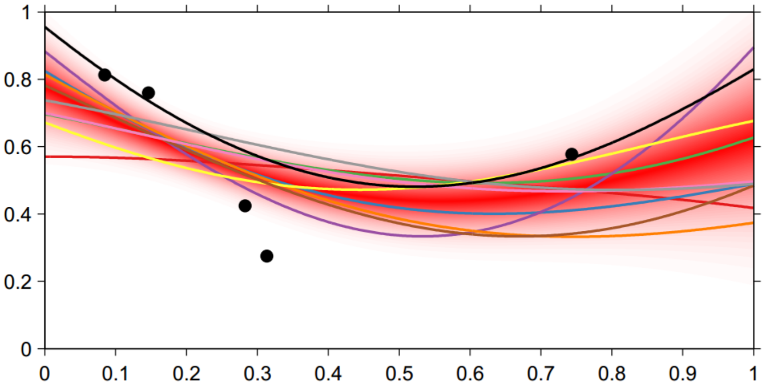
Non-linearity to the rescue



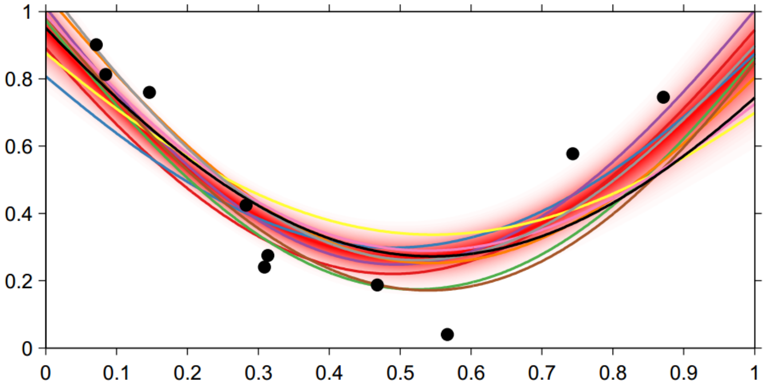
Non-linearity to the rescue



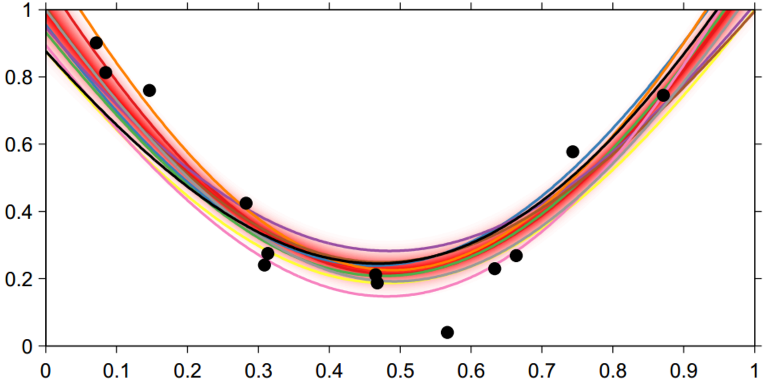
Non-linearity to the rescue



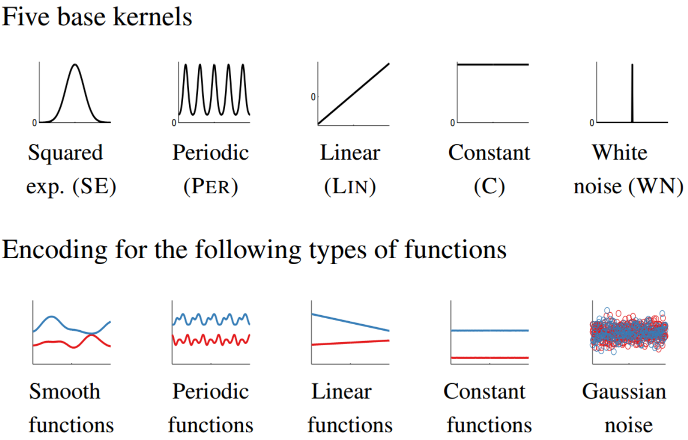
Non-linearity to the rescue



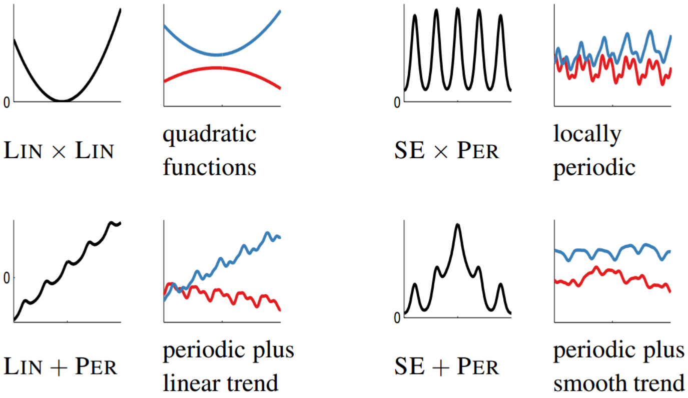
Non-linearity to the rescue



Kernel types



Kernels compositions



Classification problem

Training data: $\left\{(\mathbf{x}_i, y_i)\colon i = \overline{1, N}\right\}$, $\mathbf{x}_i \in R^s$, $y_i \in C = \{-1, +1\}$.
Task: given new x^* predict distribution y^* on C.
That is we want to know

$$\pi(x) = p(y = +1|\mathbf{x}).$$

Classification problem solving

Main idea:

$$\pi(x) = \sigma(f(\mathbf{x})),$$

$f(\cdot)$ — latent variable.

Step 1

$$p(f^*|X, y, \mathbf{x}^*) = \int p(f^*|X, \mathbf{x}^*, \mathbf{f})p(\mathbf{f}|X, y)df$$

Step 2

$$p(y^* = +1|X, y, \mathbf{x}^*) = \int \sigma(f^*)p(f^*|X, y, \mathbf{x}^*)df^*$$