Gaussian Process regression (Krigin)

Georgios Karagiannis

Department of Mathematics, Purdue

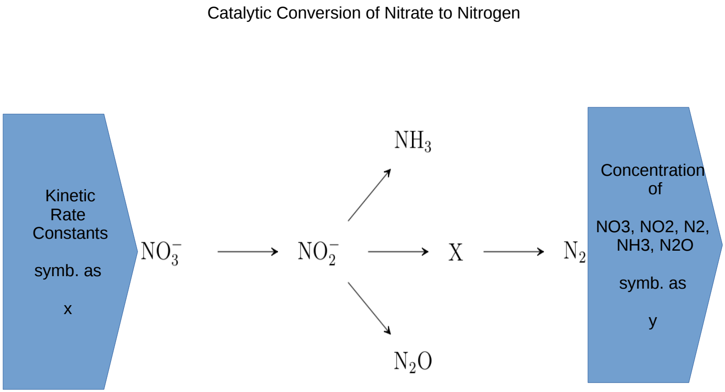
July 8, 2016

SURF 2016

Why is it useful in UQ ?

E.g.:consider the framework of Computer Experiments:

- There is an expensive Simulator that describes a Physical procedure



Why is it useful in UQ ?

It can be used as an emulator (probabilistic surrogate) in:

- Prediction of the output of expensive simulators (***)
- Optimization of expensive objective functions
- Sensitivity Analysis in expensive simulators
- Uncertainty Propagation in expensive simulators

Preliminaries

Multivariate Normal distribution (I)

Notation

$$f \sim N_n(\mu, \Sigma), \quad f := (f_1, ..., f_n)^\intercal$$

Mean (vector)

$$\mu := (\mathrm{E}(f_1), ..., \mathrm{E}(f_n))^\intercal, \qquad \mu_i = \mathrm{E}(f_i)$$

Covariance (matrix)

$$\Sigma = \begin{bmatrix} \mathrm{Cov}(f_1, f_1) & \cdots & \mathrm{Cov}(f_1, f_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(f_n, f_1) & \cdots & \mathrm{Cov}(f_n, f_n) \end{bmatrix}, \qquad \Sigma_{i,i'} = \mathrm{Cov}(f_i, f_{i'})$$

Multivariate Normal distribution (II)

Notation

$$f \sim \mathrm{N}_n(\mu, \Sigma), \quad f := (f_1, ..., f_n)^\intercal$$

Density function

$$\mathrm{N}_n(f|\mu, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp(-\frac{1}{2}(f-\mu)^\intercal \Sigma^{-1}(f-\mu))$$

Cumulative function

$$\Pr(f \leqslant U|\mu, \Sigma) = \int_{-\infty}^{U} \mathrm{N}_n(f|\mu, \Sigma)\mathrm{d}f$$

Multivariate Normal distribution (III)

If

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \sim \mathrm{N}_{n_1+n_2}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12}^{\top} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix} \right)$$

then marginalizing implies

$$f_1 \sim \mathrm{N}_{n_1}(\mu_1, \Sigma_{11})$$

and

$$f_2 \sim \mathrm{N}_{n_2}(\mu_2, \Sigma_{22})$$

and ...

Multivariate Normal distribution (IV)

If

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \sim \mathrm{N}_{n_1+n_2}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12}^{\intercal} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}\right)$$

then conditioning implies

$$f_1|(f_2 = t) \sim \mathrm{N}_{n_1}(\mu_{1|2}, \Sigma_{1|2})$$

where

$$\mu_{1|2} = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(t - \mu_2)$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{\intercal}$$

Gaussian process (GP)

Definition GP is a collection of random variables $\{f(x); x \in \mathcal{X}\}$, indexed by label x, where any finite collection of those variables has a multivariate normal distribution

Namely We denote the GP as

$$f(\cdot) \sim \mathrm{GP}(\mu(\cdot), c(\cdot, \cdot))$$

with mean

$$\mu(x) := \mathrm{E}(f(x))$$

and covariance function

$$c(x, x') := \mathrm{Cov}(f(x), f(x'))$$

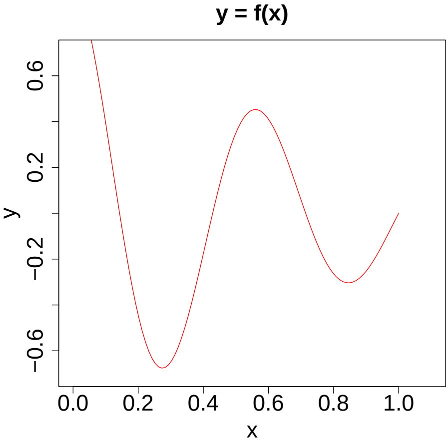
Notes Essentially, GP is a distribution defined over functions

GP is specified by its mean and covariance functions.

Gaussian process regression (Krigin)

Running example (A boring 1D function)

Assume the unknown function, we wish to recover, is:



The prior GP model

Prior information about $f(\cdot)$ is represented as a GP, as:

$$f(\cdot)|\beta, r, \tau^2, \sigma^2 \sim \mathrm{GP}(\mu_0(\cdot), c_0(\cdot, \cdot)),$$

with

- mean function

$$\mu_0(x) = \sum_{k=0}^{p} \beta_k h_k(x) = h^{\top}(x)\beta$$

  E.g., $\mu_0(x) = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + ... + \beta_p x^p$

- covariance function

$$c_0(x, x') := \tau^2 \prod_{j=1}^{d} R_j(x_j, x_j'|r_j) + \delta_{x,x'} g$$

Some $R(x,x'|\psi)$ leading to valid covariance functions

- Gaussian covariance functions:

$$R(x,x'|r)=\exp(-\frac{1}{2}\frac{|x-x'|^2}{r^2})$$

- Exponential covariance function

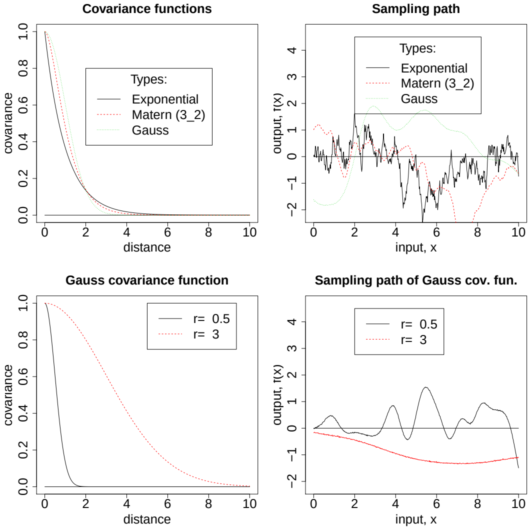
$$R(x,x'|r)=\exp(-\frac{|x-x'|}{r})$$

- Matern covariance function

$$R(x,x'|r)=\frac{2^{1-v}}{\Gamma(v)}(\frac{\sqrt{2v}|x-x'|}{r})^{v}K_{v}(\frac{\sqrt{2v}|x-x'|}{r})$$

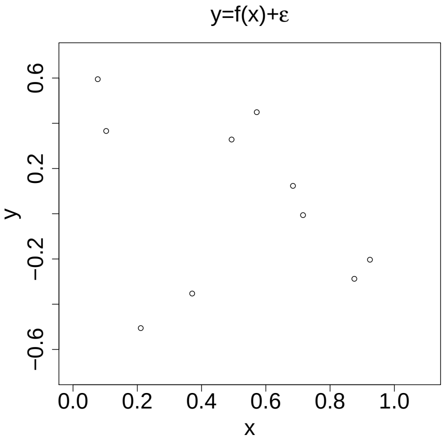
K_v modified Sobol function, for $v=3/2,5/2,...$

Running example



The statistical model (likelihood function)

- Suppose available training data-set $D = \{(x_i, y_i); i = 1, ..., n\}$



The statistical model (likelihood function)

- Suppose available training data-set $D = \{(x_i, y_i); i = 1, ..., n\}$

$$y_i = \underbrace{f(x_i)}_{f(\cdot)\sim \mathrm{GP}(\mu_0(\cdot), c_0(\cdot,\cdot))} + \underbrace{\epsilon_i}_{\epsilon_i \sim \mathrm{N}(0,\sigma^2)}$$

- The joint distribution of $\boldsymbol{y}$ (a.k.a. likelihood function) is

$$\mathcal{L}(\boldsymbol{y}|\beta, r, \tau^2, \sigma^2) = \mathrm{N}_n(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{C} + \mathbb{I}_n\sigma^2)$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top;$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_0(x_1) & \cdots & \mu_0(x_n) \end{bmatrix}^\top;$$

$$\boldsymbol{C} = \begin{bmatrix} c_0(x_1, x_1) & \cdots & c_0(x_1, x_n) \\ \vdots & \ddots & \vdots \\ c_0(x_1, x_n) & \cdots & c_0(x_n, x_n) \end{bmatrix}$$

Towards a probabilistic surrogate model

- Let $f(x)$ & $f(x')$ be function values at 'unseen' inputs $x \,\&\, x' \in \mathcal{X}$.

- Then joint distribution of $(f(x), f(x'), y)^\top$ is

$$\begin{bmatrix} f(x) \\ f(x') \\ \boldsymbol{y} \end{bmatrix} \sim \mathrm{N}_{n+2}\left(\begin{bmatrix} \mu_0(x) \\ \mu_0(x') \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} c_0(x,x) & c_0(x,x') & \boldsymbol{c}(x)^\top \\ c_0(x',x) & c_0(x',x') & \boldsymbol{c}(x')^\top \\ \boldsymbol{c}(x) & \boldsymbol{c}(x') & \boldsymbol{C} + \mathbb{I}\sigma^2 \end{bmatrix} \right)$$

where $\boldsymbol{c}(x) = (c_0(x, x_i); i = 1, ..., n)^\top$

The Posterior GP model (probabilistic surrogate model)

By conditioning on $\boldsymbol{y}$, it can be shown that $f(\cdot)|D, \beta, L, \tau^2, \sigma^2$ is a GP

$$f(\cdot)|D, \beta, r, \tau^2, \sigma^2 \sim \text{GP}(\mu_n(\cdot), c_n(\cdot, \cdot))$$

with ...

The Posterior GP model (probabilistic surrogate model)

By conditioning on $\boldsymbol{y}$, it can be shown that $f(\cdot)|D, \beta, L, \tau^2, \sigma^2$ is a GP

$$f(\cdot)|D, \beta, r, \tau^2, \sigma^2 \sim \text{GP}(\mu_n(\cdot), c_n(\cdot, \cdot))$$

with ...

- mean function

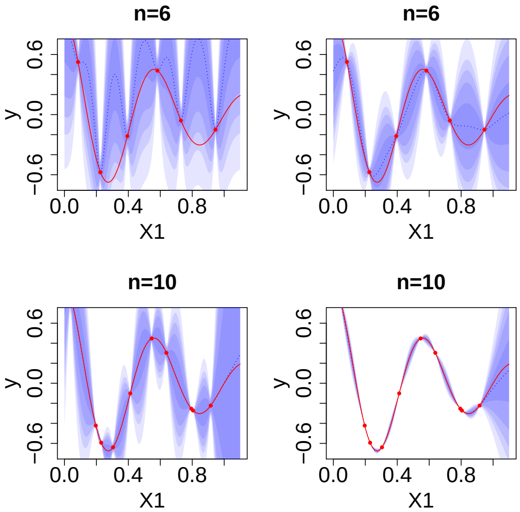
$$\mu_n(x) = \mu_0(x) + \boldsymbol{c}(x)(\boldsymbol{C} + \mathbb{I}\sigma^2)^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$$

- covariance function

$$c_n(x, x') = c_0(x, x') + \boldsymbol{c}(x)(\boldsymbol{C} + \mathbb{I}\sigma^2)^{-1}\boldsymbol{c}(x')^\top$$

Recall:
$$\boldsymbol{\mu} = (\mu_0(x_1), ..., \mu_0(x_n))^\top,$$
$$\boldsymbol{C} = (c_0(x_i, x_j); i = 1, ..., n, j = 1, ..., n)$$
$$\boldsymbol{c}(x) = (c_0(x, x_1), ..., c_0(x, x_n))^\top$$

Running example



How to train the GP regression ?

How to learn $\beta, r, \tau^2, \sigma^2$?

- Classical statistical inference:

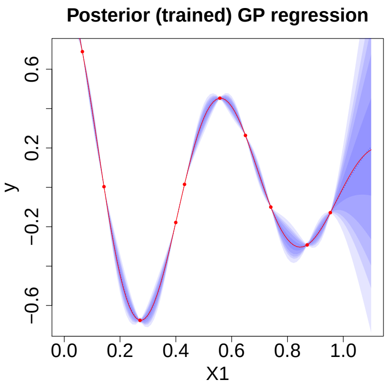
  E.g by Maximum Likelihood Estimation (MLE):

$$(\hat{\beta}, \hat{r}, \hat{\tau}^2, \hat{\sigma}^2) = \arg \min_{\forall(\beta,\ell,\tau^2,\sigma^2)} (-2\log(\mathcal{L}(\boldsymbol{y}|\beta, r, \tau^2, \sigma^2)))$$

- Bayesian statistical inference:
  - By Maximum A posteriori Estimation (MAP)
  - By evaluating the posterior distributions

Here, we focus on MLE ... easier to digest

Running example



Mean $\mu_n(\cdot)$:
Estimate (Intercept) 0.3460

Covariance $c_n(\cdot,\cdot)$:
Type : Matern $v = 5/2$

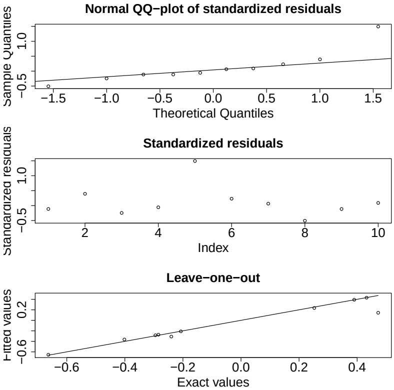
Estimate $\hat{r}$: 0.2846
Variance $\hat{\sigma}^2$: 0.6674608

Nugget effect $\hat{g}$: 1e-07

How to assess the GPR model ?

Check for:

- Normality assumption
- Goodness of fit
- Predictive ability



Compare different models (Eg. Gauss vs Matern cov. funct.)

Leave-one-out Cross Validation (LOO-CV)

For $i = 1, ..., n$:

1. Train the GP regression model against data-set $D^{(-i)} = \{(x_j, y_j); \forall i \neq j\}$
2. Predict $\hat{y}_i$ at input x_i, based on the GP regression model

Compute a performance criterion ($\text{CV} = \text{CV}(y, \hat{y})$) measuring how close your predictions ($\hat{y}_i$) to the real values (y_i) are.

- E.q.: R^2, RMSE, MAE

Performance criteria for LOO-CV

- Coefficient of determination :

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  - The percentage of the total variation explained by the predictions

- Root mean squared error :

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2}$$

  - penalizes larger differences

- Mean absolute error:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

  - more robust to outlines

Running example

		LOO-CV criterion		
		R2	RMSE	MAE
GPR Model	Gaussian	0.99	0.03	0.01
	Exponential	0.45	0.29	0.21
	Matern 5/2	0.96	0.07	0.04

Table: Model comparison

And now,

let's practice ...