Analytical calculations of Sobol indices for the Gaussian process metamodel

→ Comparison of 2 approaches : use of predictor only and global Gp model

Amandine Marrel★

INSA supervisor : B. Laurent

CEA supervisors: B. Iooss◊, M. Jullien★

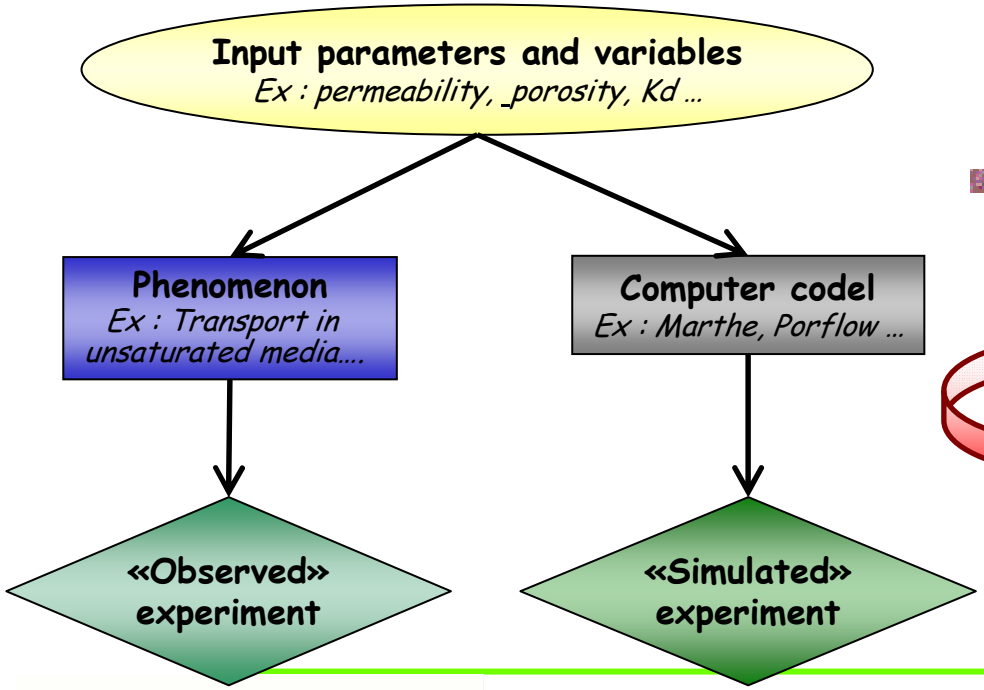
◊ CEA/DEN/CAD/DER/SESI/LCFR

★ CEA/DTN/CAD/DTN/SMTM/LMTE

Framework and purposes (1)

Modeling process :

- Real phenomena represented by deterministic equations
  Ex : transport equations, Darcy equations…
- Input parameters and variables $X = [X_1, \ldots, X_d]$
  Ex : model parameters (estimation or literature), physical variables (in situ or laboratory)…
- Implementation : computer code $Y_{code}(X)$
  Ex : migration of pollutant in saturated porous media…



Purposes :

- Identify the most influent variables,
  → Improvement of the knowledge of these variables in priority

SENSITIVITY ANALYSIS

Framework and purposes (2)

Problem :

- Computer code complex & time expensive
- High number of inputs
- Large number of simulations required for sensitivity analysis

→ **Direct use of computer code = Very Difficult**

Solution : Replace computer code by a statistic model called **metamodel**

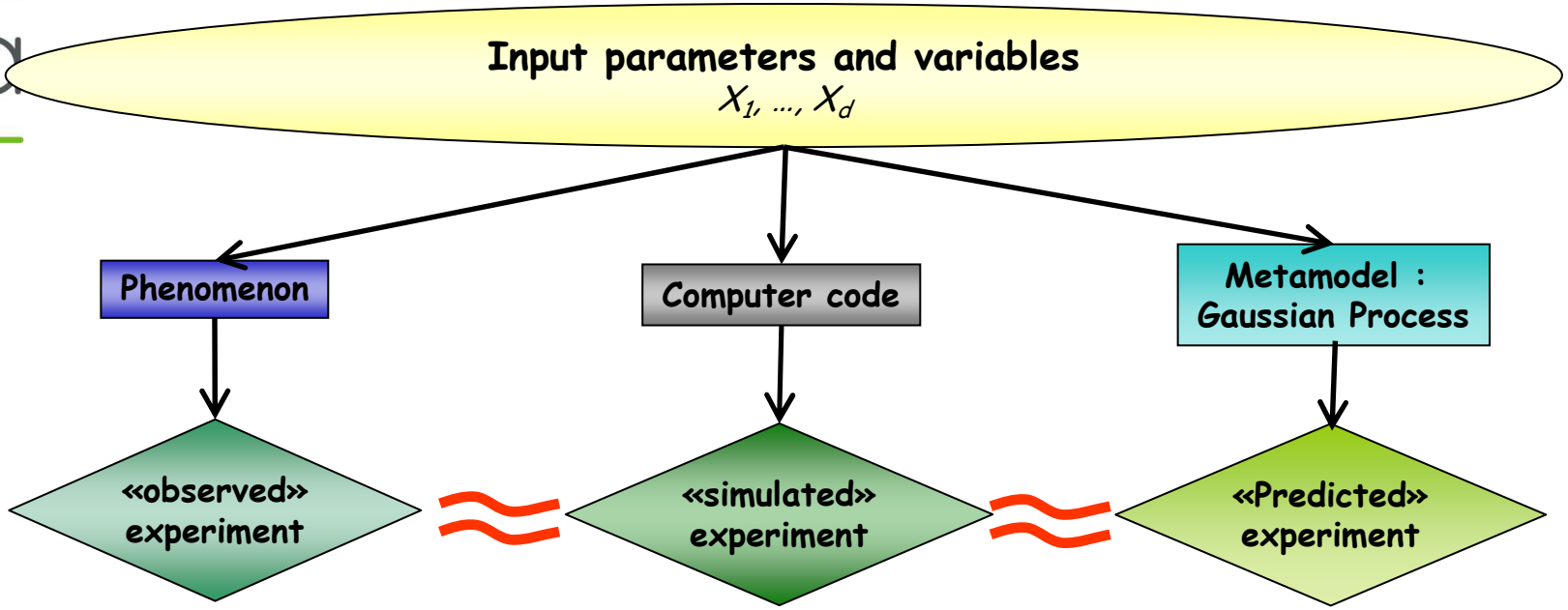
- Ex : Polynomials, splines, neural networks, regression trees…
- Choice : conditional Gaussian Process (GP)

Application of metamodel

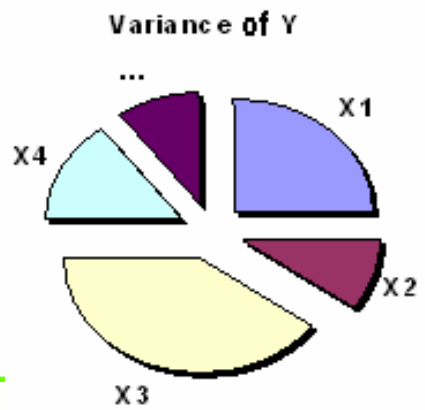
- Sensitivity analysis

→ **Computation of Sobol indices**

Framework and purposes (3)



SENSITIVITY ANALYSIS



Computation of **Sobol indices**

Content

1. **Gaussian process Metamodel**
   - Gp model
   - Joint and conditional distributions

2. **Sensitivity analysis**
   - Sobol indices
   - Sensitivity analysis with GP model : 2 approaches

3. **Applications**
   - Ishigami function
   - Irregular function
   - g-Sobol function

Conclusion and prospects

Gaussian process model

- **Definition :**

  Gaussian process defined on $R^d \times \Omega$

  $Y(x,\omega) = F(x) + Z(x,\omega)$

  Regression **stochastic part**

  Stochastic process Z with :

  $E_\Omega[Z(x)] = 0$

  $Cov_\Omega(Z(x), Z(u)) = \sigma^2 R(x, u)$

  where σ^2 is the variance

  and R the correlation function

  $Z \sim \mathcal{N}(0, \sigma^2 R)$

- **Parametric choices :**
  - F : polynomial of degree 1 $\mathbf{F(x)} = \beta_0 + \sum_{i=1}^{d} \beta_i x_i$
  - R : stationary process with generalized exponential covariance

$$\mathbf{R(x, u)} = \mathbf{R(x - u)} = \exp\left(- \sum_{i=1}^{d} \theta_i \left| x_i - u_i \right|^{p_i} \right)$$

Joint and conditional distributions

- Joint distribution :
  - GP model : $Y(x,\omega) = F(x) + Z(x,\omega)$
  - Learning sample (LS) of n simulations : (X_{LS}, Y_{LS})

$$X_{LS} = \left[x^{(1)},\ldots,x^{(n)}\right],\ F_{LS} = F(X_{LS}),\ R_{LS} = \left(R\left(x^{(i)},x^{(k)}\right)\right)_{i,k}$$

  - Joint distribution of LS : $Y_{LS} \sim \mathcal{N}(\beta F_{LS}, \sigma^2 R_{LS})$
  - Conditional Gp metamodel :

$$\Rightarrow \quad Y(x,\omega)_{|X_{LS},Y_{LS}} \sim PG$$

$$\begin{cases} E_\Omega\left[Y(x,\omega)_{|X_{LS},Y_{LS}}\right] = \beta F(x) + r(x) R_{LS}^{-1}\left[Y_{LS} - \beta F_{LS}\right] \\ \qquad \text{with } r(x) = \left[R\left(x^{(1)},x\right),\ldots,R\left(x^{(n)},x\right)\right] \\ Cov_\Omega\left(Y(u,\omega)_{|X_{LS},Y_{LS}}, Y(v,\omega)_{|X_{LS},Y_{LS}}\right) = \sigma^2\left(R(u,v) + {}^t r(u) R_{LS}^{-1} r(v)\right) \\ \text{Predictor notation : } \hat{Y}(x) = E_\Omega\left[Y(x,\omega)_{|X_{LS},Y_{LS}}\right] \end{cases}$$

Content

1. Gaussian process Metamodel
   - Gp model
   - Joint and conditional distributions

2. **Sensitivity analysis**
   - Sobol indices
   - Sensitivity analysis with GP model : 2 approaches

3. Applications
   - Ishigami function
   - Irregular function
   - g-Sobol function

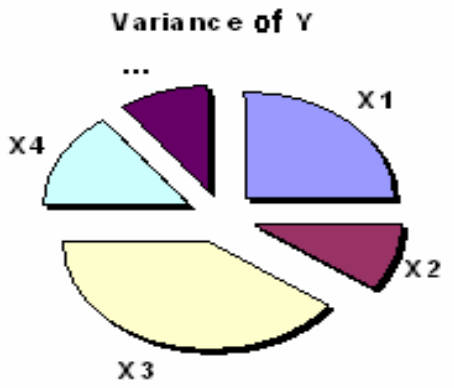
Conclusion and prospects

Sobol indices

- Global sensitivity analysis

  → Based on variance decomposition :

  Part of the variance of each input in the output variance



  → Input/output relation neither linear, nor monotonous :

  Computation of Sobol indices

- Definitions for a deterministic function $g(X_1, \ldots, X_d)$

  - Main Effects :

$$a(X_i) = \oint g(x_1, \ldots, x_{i-1}, X_i, x_{i+1}, \ldots, x_d) \prod_{j=1, j\neq i}^{d} dx_j = E[g(X_1, \ldots, X_d) / X_i]$$

  - Sobol indices :

$$S_i = \frac{Var_{X_i}[E(g(X_1, \ldots, X_d) / X_i)]}{Var(g)} = \frac{Var_{X_i}[a(X_i)]}{Var(g)}$$

  - Notation : $X_{-i} = [X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d]$

Sensitivity analysis with Gp model : 2 approaches (1)

- Gp model conditionally to LS points :

$$\Rightarrow Y(x,\omega)_{|X_{LS},Y_{LS}} \sim PG \begin{cases} \hat{Y}(x) = E_\Omega\left[Y(x,\omega)_{|X_{LS},Y_{LS}}\right] \\ Cov_\Omega\left(Y(u,\omega)_{|X_{LS},Y_{LS}}, Y(v,\omega)_{|X_{LS},Y_{LS}}\right) \end{cases}$$

- Computation of Sobol indices :

From predictor formula

$$\hat{Y}(X) = E_\Omega\left[Y(X,\omega)_{|X_{LS},Y_{LS}}\right]$$

$$\begin{cases} a(X_i) = E_{X_{-i}}\left[\hat{Y}(X)/X_i\right] \\ S_i = \dfrac{Var_{X_i}[E_{X_{-i}}(\hat{Y}(X)/X_i)]}{Var_X(\hat{Y})} \end{cases}$$

$a(X_i)$: deterministic function of X_i

S_i : deterministic indices

From global Gp model $Y(X,\omega)_{|X_{LS},Y_{LS}}$

$$\begin{cases} a(X_i,\omega) = E_{X_{-i}}\left[Y(X,\omega)_{|X_{LS},Y_{LS}}/X_i\right] \\ V_i(\omega) = Var_{X_i}\left[E_{X_{-i}}\left(Y(X,\omega)_{|X_{LS},Y_{LS}}/X_i\right)\right] \end{cases}$$

$a(X_i,\omega)$: stochastic process of X_i

V_i : random variables

$$\tilde{S}_i = \frac{E_\Omega(V_i)}{E_\Omega Var_X Y(X,\omega)_{|X_{LS},Y_{LS}}}$$

Sensitivity analysis with Gp model : 2 approaches (2)

- Sobol indices:

From predictor formula

$$\hat{Y}(X) = E_{\Omega}\left[Y(X,\omega)_{|X_{LS},Y_{LS}}\right]$$

$$S_i = \frac{Var_{X_i}\left[E_{X_{-i}}\left(E_{\Omega}\left(Y_{|X_{LS},Y_{LS}}(X,\omega)\right)/X_i\right)\right]}{Var_X\left(E_{\Omega}Y_{|X_{LS},Y_{LS}}\right)}$$

From global Gp model

$$Y(X,\omega)_{|X_{LS},Y_{LS}}$$

$$\tilde{S}_i = \frac{E_{\Omega}\left(Var_{X_i}\left[E_{X_{-i}}\left(Y_{|X_{LS},Y_{LS}}(X,\omega)/X_i\right)\right]\right)}{E_{\Omega}Var_X Y(X,\omega)_{|X_{LS},Y_{LS}}}$$

- Computation:
  - Analytical calculations
  - Numerical integrals
  - Independent inputs + product of one-dim covariances :
    - ⇨ S_i : simple integrals
    - ⇨ $\tilde{S}_i$: simple and double integrals

Content

1. Gaussian process Metamodel
   - Gp model
   - Joint and conditional distributions

2. Sensitivity analysis
   - Sobol indices
   - Sensitivity analysis with GP model : 2 approaches

3. **Applications**
   - Ishigami function
   - Irregular function
   - g-Sobol function

Conclusion and prospects

Application on an analytical function fx_{test}

- Purpose:

→ Comparison of the 2 approaches to compute Sobol indices

- predictor only
- global model

- Method

→ Study of the convergence of Sobol indices in function of :

- the number of simulations of the learning sample **N**
- the predictivity $\mathbf{Q_2}$ of metamodel on a test sample (K points)

$$Q_2(Y, \hat{Y}) = 1 - \frac{\sum_{i=1}^{K} \left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{K} \left(\overline{Y} - Y_i\right)^2}$$

- Protocol

→ Simulations of N points for the LS with fx_{test}

→ Estimation of Gp metamodel

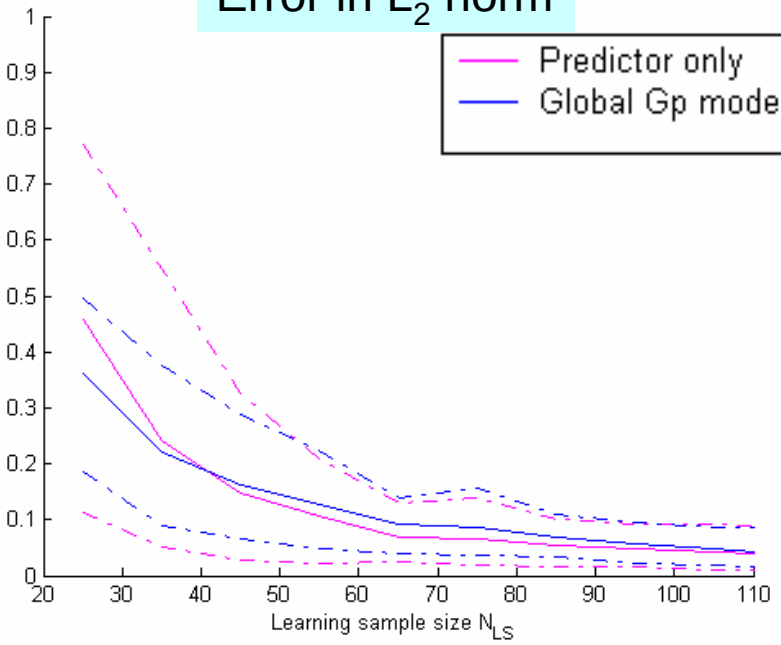
→
  - Computation of Q_2 on a test sample (10000 points)
  - Computation of Sobol indices following the 2 approaches
    + Error with theoretical indices (in L_2 norm)

Application 1 : Homma-Saltelli Function

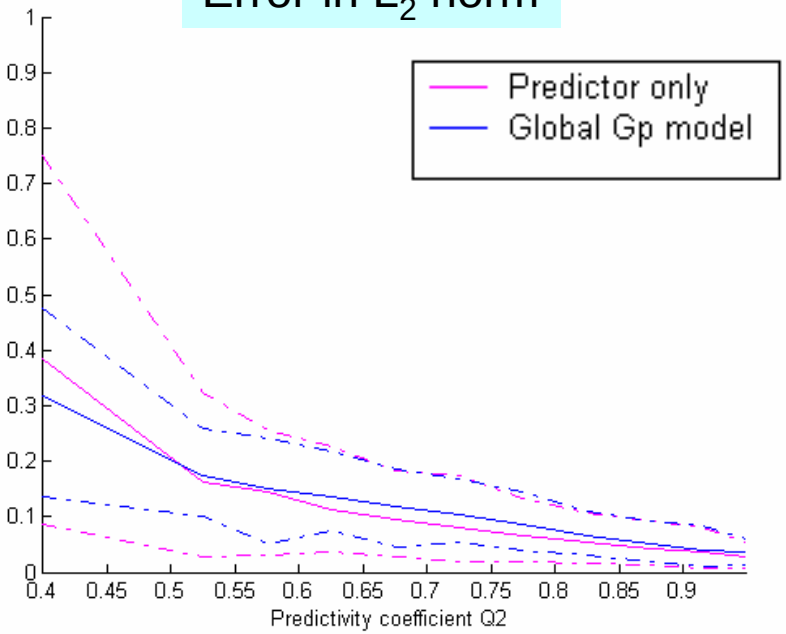
- Function of 3 inputs with 2 coefficients :

$$\begin{cases} g(X_1, X_2, X_3) = \sin X_1 + 7(\sin X_2)^2 + 0.1 X_3^4 \sin X_1 \\ X_i \sim U_{[-\pi;\pi]} \quad \text{for} \quad i = 1, \ldots, 3 \end{cases}$$

Error in L_2 norm



Error in L_2 norm



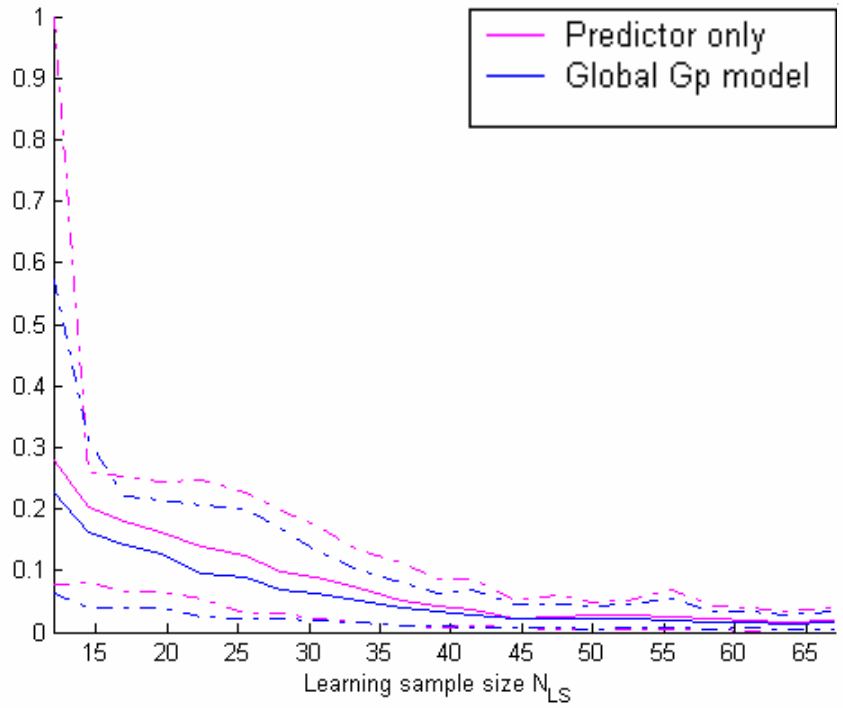
For each method :
- Empirical mean
- Empirical 0.95%-quantile
- Empirical 0.05%-quantile

Application 2 : Irregular Function

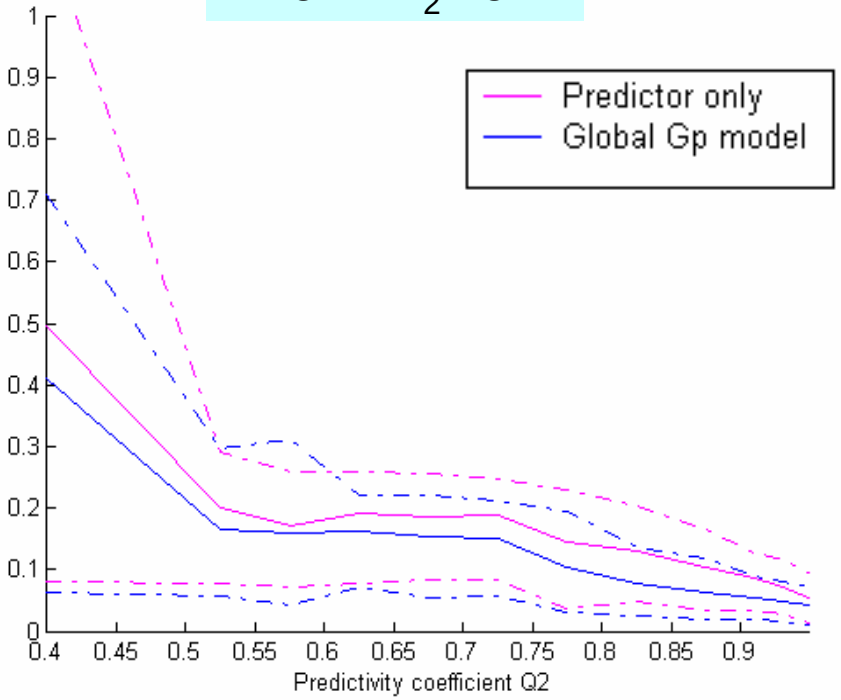
- Function of 2 inputs :

$$\begin{cases} g(X_1, X_2) = 0.2e^{X_1} - 0.2X_2 + \frac{1}{3}X_2^6 + 4X_2^4 - 4X_2^2 + \frac{7}{10}X_1^2 + X_1^4 + \frac{3}{4X_1^2 + 4X_2^2 + 1} \\ X_i \sim U_{[-1;1]} \quad \text{for} \quad i = 1,2 \end{cases}$$

Error in L_2 norm

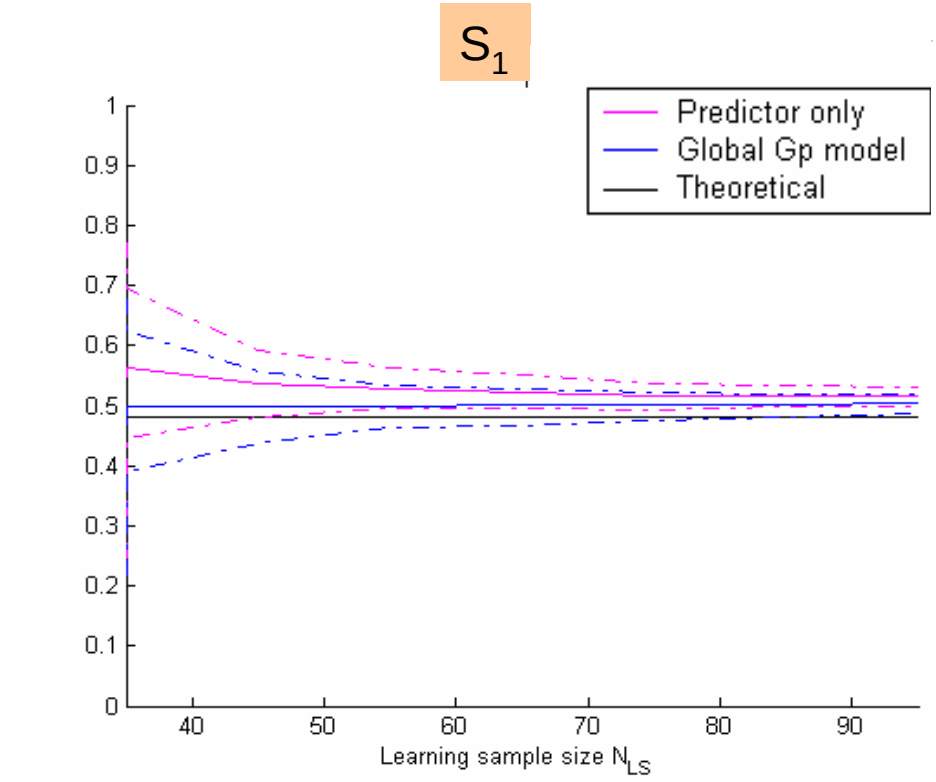


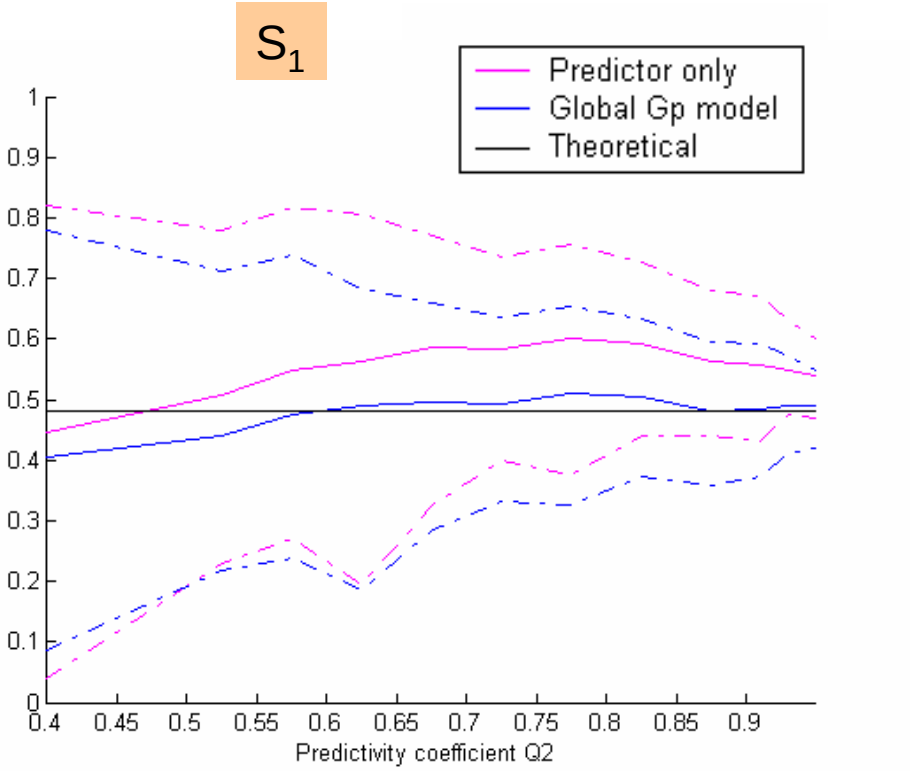
Error in L_2 norm



Application 3 : g-Sobol Function (1)

- **Function of 5 inputs :** $g(X_1,..., X_5) = \prod_{i=1}^{5} \frac{|4X_i - 2| + i}{1+i}$ with $X_i \sim U_{[0;1]}$ for $i = 1...5$

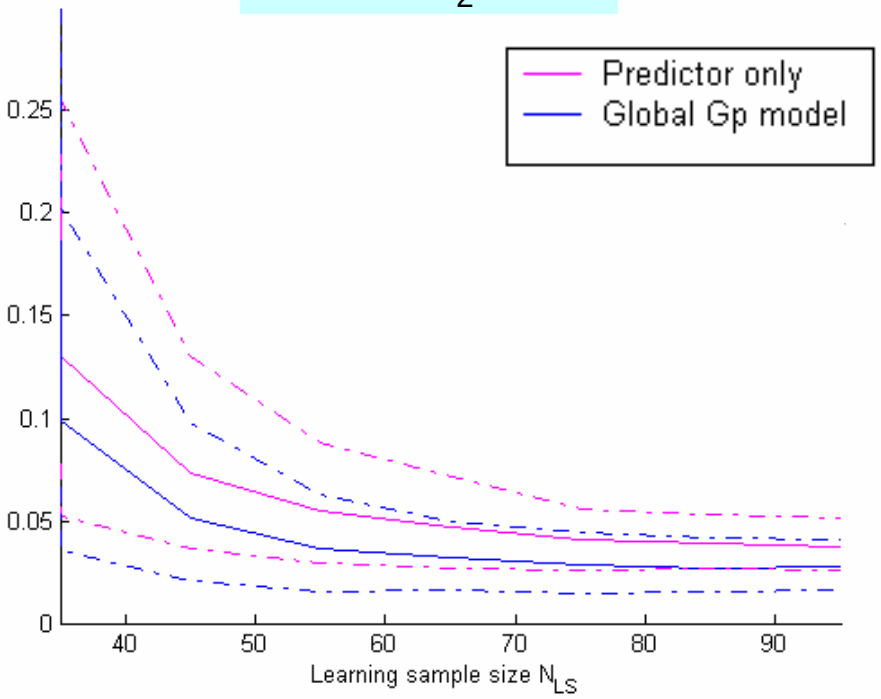




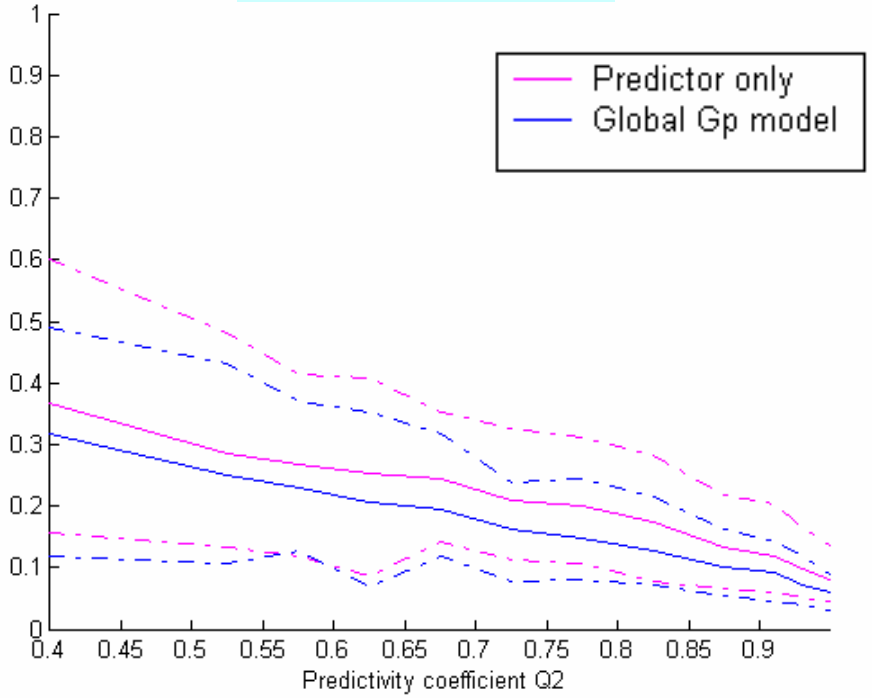
Application 3 : g-Sobol Function (2)

- **Function of 5 inputs :** $g(X_1,\ldots,X_5)=\prod_{i=1}^{5}\frac{|4X_i-2|+i}{1+i}$ with $X_i \sim U_{[0;1]}$ for $i=1\ldots5$

Error in L_2 norm



Error in L_2 norm



Content

1. Gaussian process Metamodel
   - Gp model
   - Joint and conditional distributions

2. Sensitivity analysis
   - Sobol indices
   - Sensitivity analysis with GP model : 2 approaches

3. Applications
   - Ishigami function
   - Irregular function

Conclusion and prospects

Conclusion and prospects

- $\tilde{S}_i$ (global Gp model) better in mean than S_i (predictor only)
  - vs. learning sample size N_{LS}
  - vs. prediction accuracy (Q_2)

- Lower sampling deviation and variability for $\tilde{S}_i$ particularly for low N_{LS} or Q_2

- More computer time required to compute $\tilde{S}_i$ (double integrals)

→ Significant interest when few data available and inaccurate Gp metamodel

- Perspectives :
  - Use of all the distribution of $V_i(\omega)$ not only mean of V_i
  ⇨ computation of variance and confidence interval on $\tilde{S}_i$

$$V_i(\omega) = Var_{X_i}\left[E_{X_{-i}}\left(Y(X,\omega)_{|X_{LS},Y_{LS}} / X_i\right)\right] \Rightarrow \begin{cases} E_\Omega(V_i) \Rightarrow \tilde{S}_i \\ Var_\Omega(V_i) \Rightarrow \text{CI for } \tilde{S}_i \text{ ?} \end{cases}$$

ANNEXE

Framework and purposes (2)

Problem :

- Computer code complex & time expensive
- High number of inputs

→ **Direct use of computer code = Very Difficult**

Solution : Replace computer code by an approximate statistic model called **metamodel**

- Ex : Polynomials, splines, neural netwoks, regression trees…
- Choice : conditional Gaussian Process (GP)

Why Gaussian Process ?

- Extension of kriging to computer codes (Sacks & al., Vasquez)
- Exact interpolator
- Statistic framework (MSE available, gaussian framework,…)
- Analytic formula and fast computation of the predictor
- Efficiency and flexibility of the model (Master2 training period)

Sensitivity analysis with GP model : 2 approaches (1)

- GP model conditionally to LS points :

$$\Rightarrow Y(x,\omega)_{|X_{LS},Y_{LS}} \sim PG \begin{cases} \hat{Y}(x) = E_{\Omega}\left[Y(x,\omega)_{|X_{LS},Y_{LS}}\right] \\ Cov_{\Omega}\left(Y(u,\omega)_{|X_{LS},Y_{LS}}, Y(v,\omega)_{|X_{LS},Y_{LS}}\right) \end{cases}$$

- Computation of Sobol indices :

From predictor formula

$$\hat{Y}(X) = E_{\Omega}\left[Y(X,\omega)_{|X_{LS},Y_{LS}}\right]$$

$$a(X_i) = \oint \hat{Y}(x_1,\ldots,x_{i-1},X_i,x_{i+1},\ldots,x_d)\prod_{j=1,j\neq i}^{d} dx_j$$

$$S_i = \frac{Var_{X_i}[E(\hat{Y}(X_1,\ldots,X_d)/X_i)]}{Var(\hat{Y})}$$

$a(X_i)$: deterministic function of X_i

S_i : deterministic indices

From global model $Y(X,\omega)_{|X_{LS},Y_{LS}}$

$$a(X_i,\omega) = \oint Y(X,\omega)_{|X_{LS},Y_{LS}}(x_1,\ldots,x_{i-1},X_i,x_{i+1},\ldots,x_d)\prod_{j=1,j\neq i}^{d} dx_j$$

$$V_i = Var_{X_i}\left[E_{X_1,..,X_{i-1},X_{i+1},..,X_d}\left(Y(X,\omega)_{|X_{LS},Y_{LS}}(X_1,..,X_d)/X_i\right)\right]$$

$a(X_i,\omega)$: stochastic process of X_i

V_i : random variables

$$\tilde{S}_i = \frac{E_{\Omega}(V_i)}{E_{\Omega}Var_X Y(X,\omega)_{|X_{LS},Y_{LS}}}$$

Sensitivity analysis with GP model : 2 approaches (2)

- Sobol indices:

From predictor formula

$$\hat{Y}(X) = E_\Omega\left[Y(X,\omega)_{|X_{LS},Y_{LS}}\right]$$

From global model

$$Y(X,\omega)_{|X_{LS},Y_{LS}}$$

$$S_i = \frac{Var_{X_i}\left[E_{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_d}\left(E_\Omega\left(Y_{|X_{LS},Y_{LS}}(X_1,\ldots,X_d,\omega)\right)/X_i\right)\right]}{Var(E_\Omega Y_{|X_{LS},Y_{LS}})}$$

$$\tilde{S}_i = \frac{E_\Omega\left(Var_{X_i}\left[E_{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_d}\left(Y_{|X_{LS},Y_{LS}}(X_1,\ldots,X_d)/X_i\right)\right]\right)}{E_\Omega Var_X Y(X,\omega)_{|X_{LS},Y_{LS}}}$$

- Computation:
  - Analytic computation
  - Numerical integrals
    (independent inputs + one-dim covariance ⇨ simple integrals)
  - Monte-Carlo (global variance)

Application 1 : Ishigami Function (1)

- Function of 3 inputs with 2 coefficients :

$$\begin{cases} g(X_1, X_2, X_3) = \sin X_1 + a(\sin X_2)^2 + bX_3^4 \sin X_1 \\ X_i \sim U_{[-\pi;\pi]} \quad \text{for} \quad i = 1,\dots,3 \end{cases}$$

- Theoretical Sobol indices

$$\begin{cases} S_1 = \left(b\frac{\pi^4}{5} + b^2\frac{\pi^8}{50} + 0.5 \right) \Big/ D \\ S_2 = \frac{a^2}{8D} \\ S_3 = 0 \\ D = \frac{a^2}{8} + b\frac{\pi^4}{5} + b^2\frac{\pi^8}{18} + 0.5 \end{cases}$$

- Numerical Application : a = 7 and b = 0.1

$$\begin{cases} S_1 = 0.3139 \\ S_2 = 0.4424 \\ S_3 = 0 \end{cases}$$