Statistical methods and the SIMLAB program

Éva Valkó

2021.10.07.

1. Introduction
2. Sample generation
3. Sensitivity analysis
4. Uncertainty analysis

1. Introduction

Why the sensitivity and uncertainty analysis are important?

1. Applicable for any type of model (not just chemical)
2. In every model the parameters have uncertainty→ calculation of the nominal values is not sufficient, the uncertainty belongs to these nominal values has to be defined→ the effect of this uncertainty to the model results has to be investigated
3. Sometimes the definition of the uncertainty range is more difficult than the definition of the nominal values

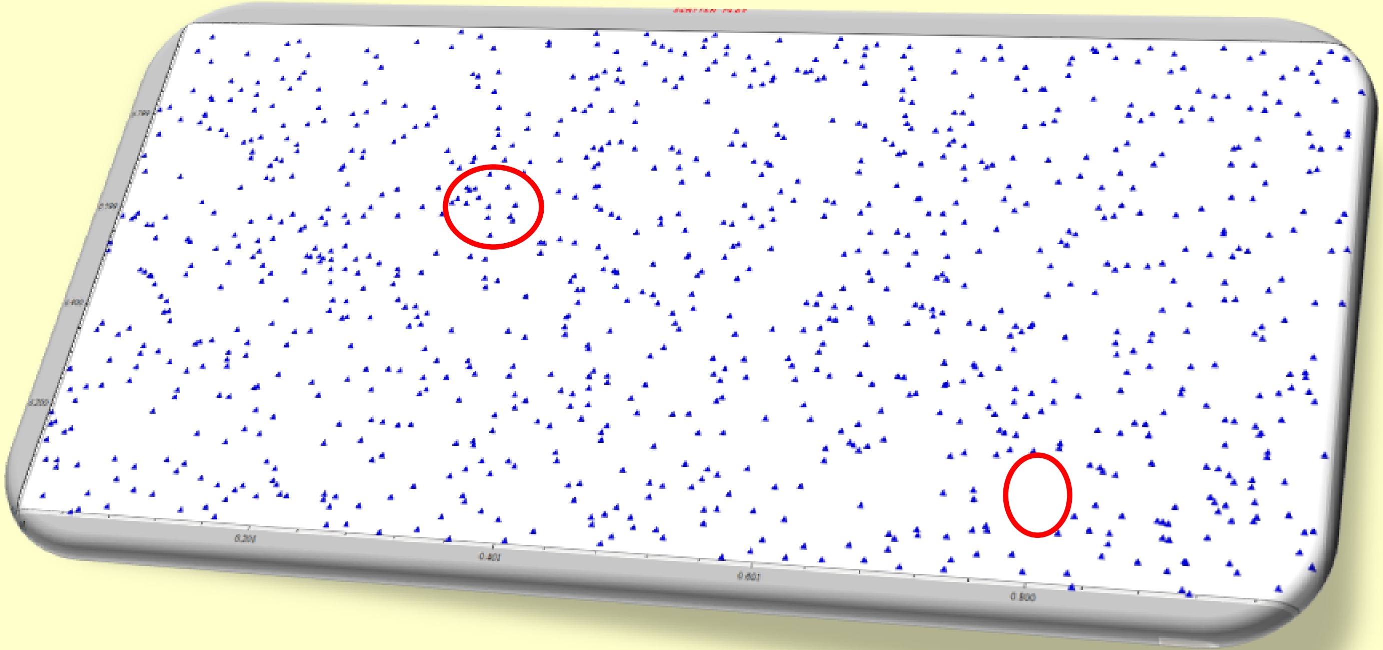
2. Sample generation

I. Random sampling

(pseudo-random)

unbiased estimation to the expected value and the standard deviation

Problem:
„white places and clustering”

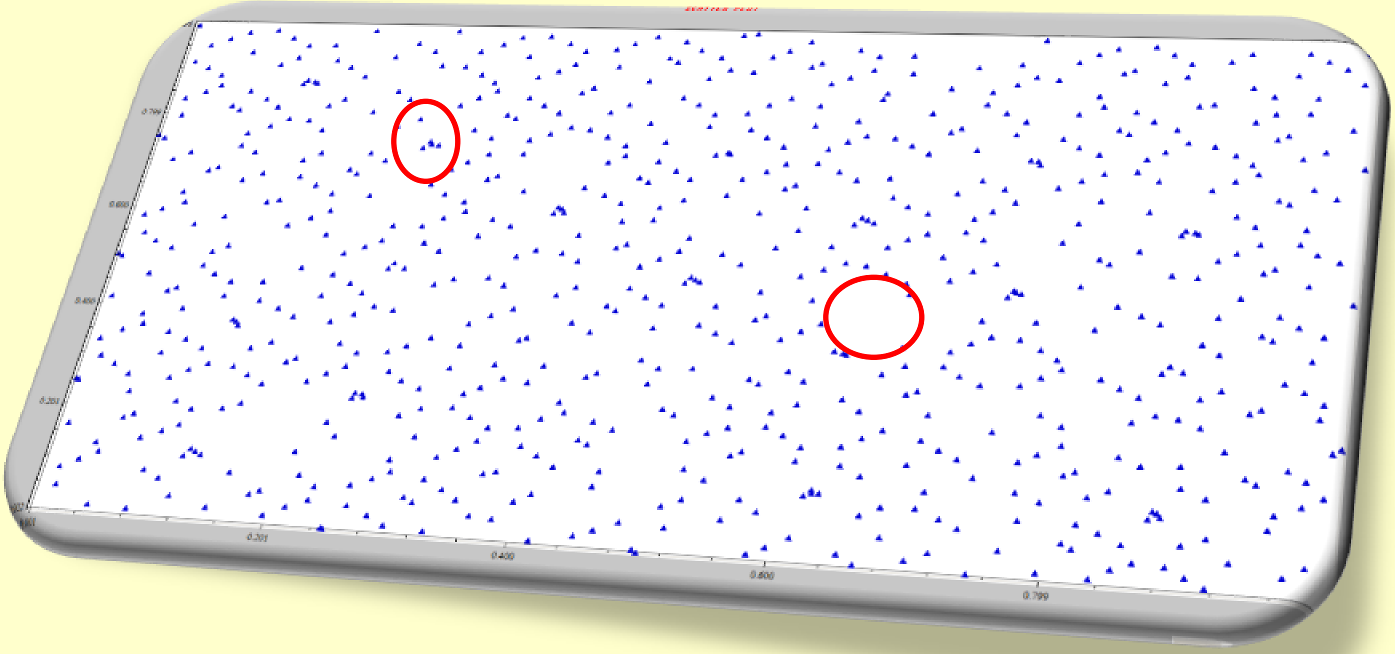


The random sampling method is applicable independently of the distributions of the parameters

2. Sample generation

II. Quasi-Random LpTau

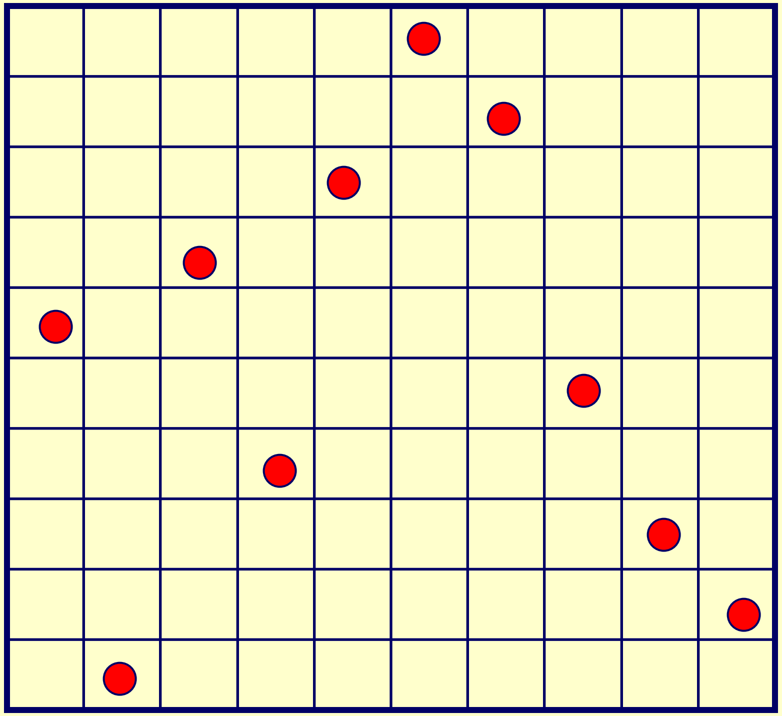
uniformly distributed sets in the [0;1]x[0;1]x[0;1]… unit cube →
each point generated is then used as an input to the transformation that calculates the inverse cumulative function of each element factor set

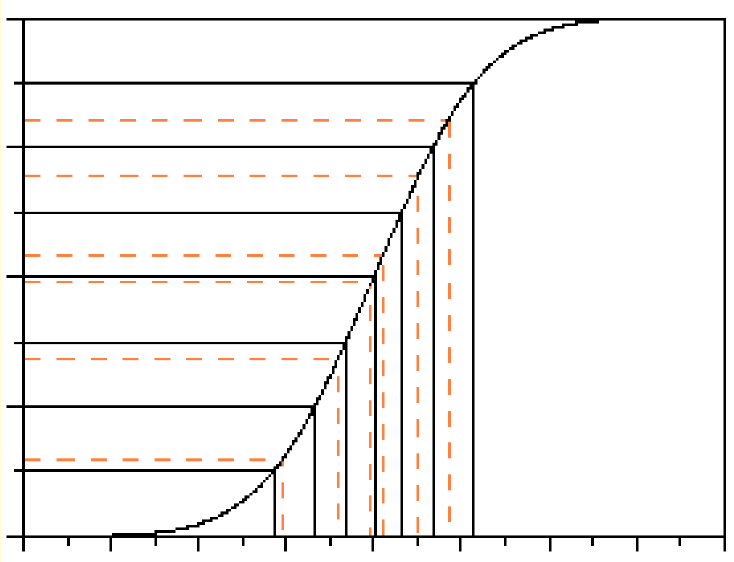


The Quasi-Random LpTau sample generation is applicable to uncorrelated parameters only. The number of parameters has to be ≤ 52

2. Sample generation

III. Latin Hypercube sampling



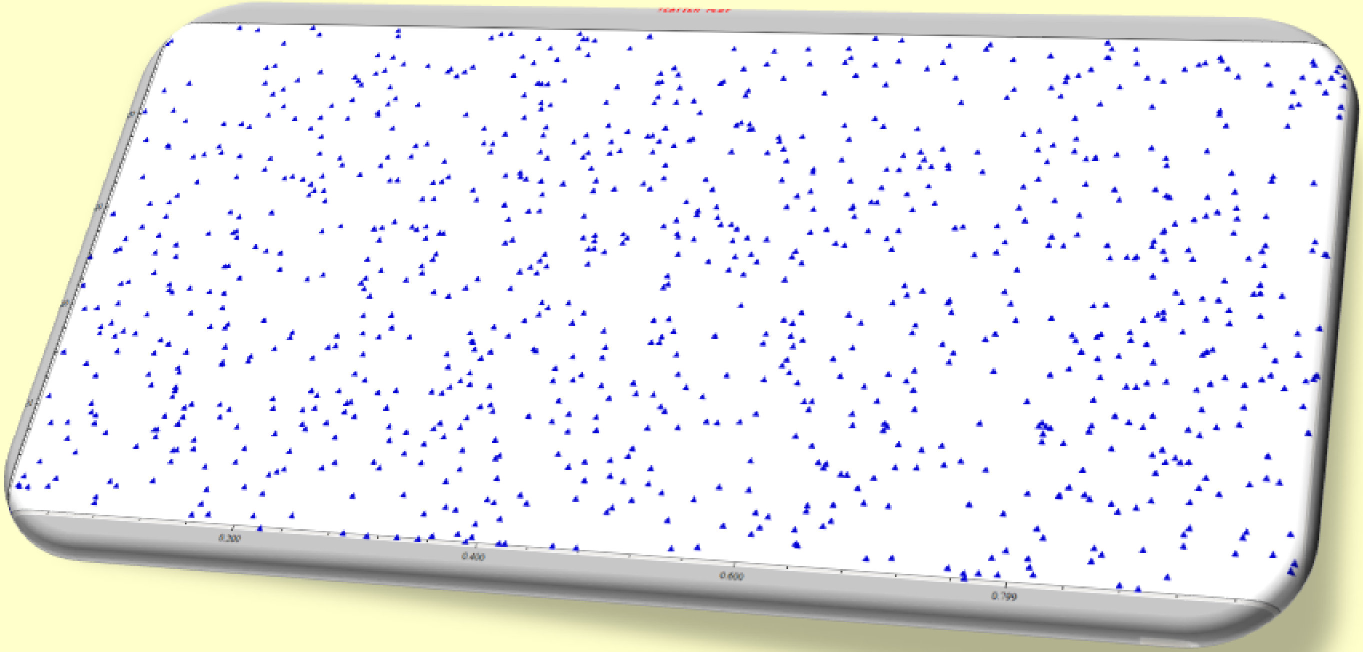


In the Latin Hypercube the range of each input factor, X_j, j=1,2…k, is divided into N intervals of equal marginal probability, 1/N, and one observation of each input factor is made in each interval using random sampling within that interval

2. Sample generation

III. Latin Hypercube sampling

More uniform distribution in the sample space



The Latin Hypercube sampling is applicable independently of the distribution of the parameters

3. SENSITIVITY ANALYSIS

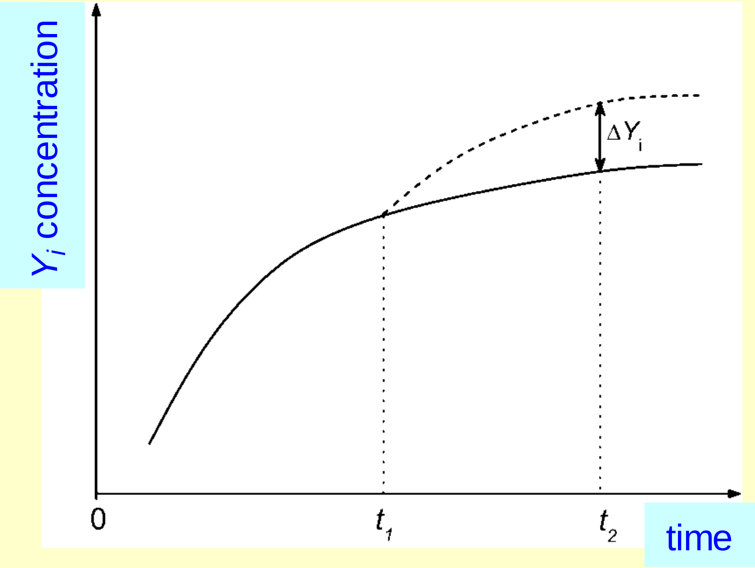
Local sensitivity analysis

Sensitivity analysis is a family of mathematical methods.
It investigates the dependence of the model results
on the values of the parameters

Local sensitivity analysis: investigates the effect of the small change of parameters

Local sensitivity coefficients can be investigated by a
finite difference approximation:

$$\frac{\partial Y_i}{\partial p_j}(t_1,t_2) \approx \frac{\Delta Y_i(t_2)}{\Delta p_j} = \frac{Y_i'(t_2) - Y_i(t_2)}{\Delta p_j}$$



parameter is changed at time t_1
the result is observed at time t_2

T. Turányi: Sensitivity analysis of complex kinetic systems: Tools and applications, *J. Math. Chem.*, **5**, 203-248(1990)

Local sensitivity analysis

Another approach: Taylor series expansion

$$Y_i(t,\mathbf{p}+\Delta\mathbf{p}) = Y_i(t,\mathbf{p}) + \sum_{j=1}^{m} \frac{\partial Y_i}{\partial p_j}\Delta p_j + \frac{1}{2}\sum_{k=1}^{m}\sum_{j=1}^{m} \frac{\partial^2 Y_i}{\partial p_k \partial p_j}\Delta p_k \Delta p_j + \ldots$$

Local sensitivity coefficient: $s_{ik} = \dfrac{\partial Y_i}{\partial p_k}$

Local sensitivity matrix: $\mathbf{S} = \left\{ \dfrac{\partial Y_i}{\partial p_k} \right\}$

The effect of parameter changes can be estimated using local sensitivities:

Changing a single parameter: $Y_i'(t_2) = Y_i(t_2) + \dfrac{\partial Y_i}{\partial p_j}\Delta p_j$

Changing several parameters: $\mathbf{Y}'(t_2) = \mathbf{Y}(t_2) + \mathbf{S}(t_1,t_2)\Delta\mathbf{p}(t_1)$

Local sensitivity analysis

$$\frac{\mathrm{d}\,\mathbf{Y}}{\mathrm{d}\,t} = \mathbf{f}\left(\mathbf{Y},\mathbf{p}\right) \qquad \mathbf{Y}(t_0) = \mathbf{Y}_0$$

Differentiation with respect p_j

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\frac{\partial\mathbf{Y}}{\partial p_j} = \mathbf{J}\frac{\partial\mathbf{Y}}{\partial p_j} + \frac{\partial\mathbf{f}}{\partial p_j} \qquad \frac{\partial\ \mathbf{Y}}{\partial\ p_j}\left(t_0\right) = 0 \qquad j = 1,\ 2,\ \ldots,\ m$$

The same equation with matrix-vector notation:

$$\dot{\mathbf{S}} = \underbrace{\mathbf{J}\,\mathbf{S}}_{\boxed{\text{indirect effect}}} + \underbrace{\mathbf{F}}_{\boxed{\text{direct effect}}}, \quad \mathbf{S}\left(0\right) = \mathbf{0}$$

where $\mathbf{J} = \left\{\frac{\partial f_i}{\partial Y_j}\right\}$ $\mathbf{F} = \left\{\frac{\partial f_j}{\partial p_k}\right\}$

Local sensitivity analysis

1 Brute force method (finite difference approximation)

$$\frac{\partial Y_i}{\partial p_j(t_1)}(t_2) \approx \frac{\Delta Y_i(t_2)}{\Delta p_j(t_1)} = \frac{Y_i'(t_2) - Y_i(t_2)}{\Delta p_j(t_1)}$$

Δp_j small: large error caused by the limited number of digits handled by the computer

Δp_j large: large error due to nonlinearity

2 Direct method

2a. Coupled direct method:

coupled solution of the kinetic and sensitivity differential equations:

$$\frac{\mathrm{d}\,\mathbf{Y}}{\mathrm{d}\,t} = \mathbf{f}(\mathbf{Y},\mathbf{p}) \qquad \mathbf{Y}(t_0) = \mathbf{Y}_0$$

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\frac{\partial \mathbf{Y}}{\partial p_j} = \mathbf{J}\frac{\partial \mathbf{Y}}{\partial p_j} + \frac{\partial \mathbf{f}}{\partial p_j} \qquad \frac{\partial\,\mathbf{Y}}{\partial\,p_j}(t_0) = 0$$

The coupled solution is repeated for each parameter: $j = 1, 2, \ldots, m$

Lots of unnecessary calculations.

Local sensitivity analysis

2b. Decoupled Direct Method (DDM):

joint solution of the kinetic and sensitivity diff. equations in each step:

$$\frac{\mathrm{d}\,\mathbf{Y}}{\mathrm{d}\,t} = \mathbf{f}\left(\mathbf{Y},\mathbf{p}\right) \qquad \mathbf{Y}(t_0) = \mathbf{Y}_0$$

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\frac{\partial \mathbf{Y}}{\partial p_j} = \mathbf{J}\frac{\partial \mathbf{Y}}{\partial p_j} + \frac{\partial \mathbf{f}}{\partial p_j} \qquad \frac{\partial\,\mathbf{Y}}{\partial\,p_j}\left(t_0\right) = 0 \qquad j = 1, 2, \ldots, m$$

The Jacobian of these equations are identical, therefore in each step

⇒ transformation of the Jacobian to a triangle matrix

⇒ selection of step size Δ*t* based on the Jacobian

⇒ solution of the stiff ODE: calculation of new **Y**

⇒ calculation of the new sensitivity vector for parameter *j* =1
using the same triangle matrix

⇒ ⇒ ⇒ ⇒ ⇒ repeating for all parameters $j = 1, 2, \ldots, m$

⇒ ⇒ repeating for new time steps from the transformation of **J**

features:

- very fast method; the computer time only slightly increases with the number of parameters *m* (because the transformation of J is the most time consuming)
- the accuracy of solution can be controlled

Local sensitivity analysis

$$s_{ik} = \frac{\partial Y_i}{\partial p_k}$$

(Original) local sensitivity coefficients:
the parameter is changed by one unit
inspected: the result is changed by how many units
[unit of result / unit of parameter]

Normalized local sensitivity coefficients:

$$\tilde{s}_{ik} = \frac{p_k}{Y_i}\frac{\partial Y_i}{\partial p_k} = \frac{\partial \ln Y_i}{\partial \ln p_k}$$

investigates relative changes
How much % change of the result
due to 1 % change of the parameter?
dimension free

So far: single parameter is changed
effect on a single model result is investigated

Further information can also be extracted from sensitivity matrix **S** using principal component analysis, like the case when
several parameters are changed simultaneously, and
the effect on multiple model results is investigated.

Local sensitivity analysis

- Linear approximation of the variance of the model result
- Does not take into account the nonlinear effects
- The result belongs to the nominal set of model parameters
- Realistic results, if the model behaves qualitatively similarly in the whole domain of parameters
- Non-realistic results, if the model is qualitatively different in the various parts of the parameter domain
- Provides separately the contribution of parameters
- Can be calculated fast

Global uncertainty analysis

Local uncertainty analysis:

Gives information at given nominal values

Well applicable if there is no significant behavior changing in the domain of the parameters

Exact results in case of linear models

Global uncertainty analysis:

The full domain of the parameters is investigated

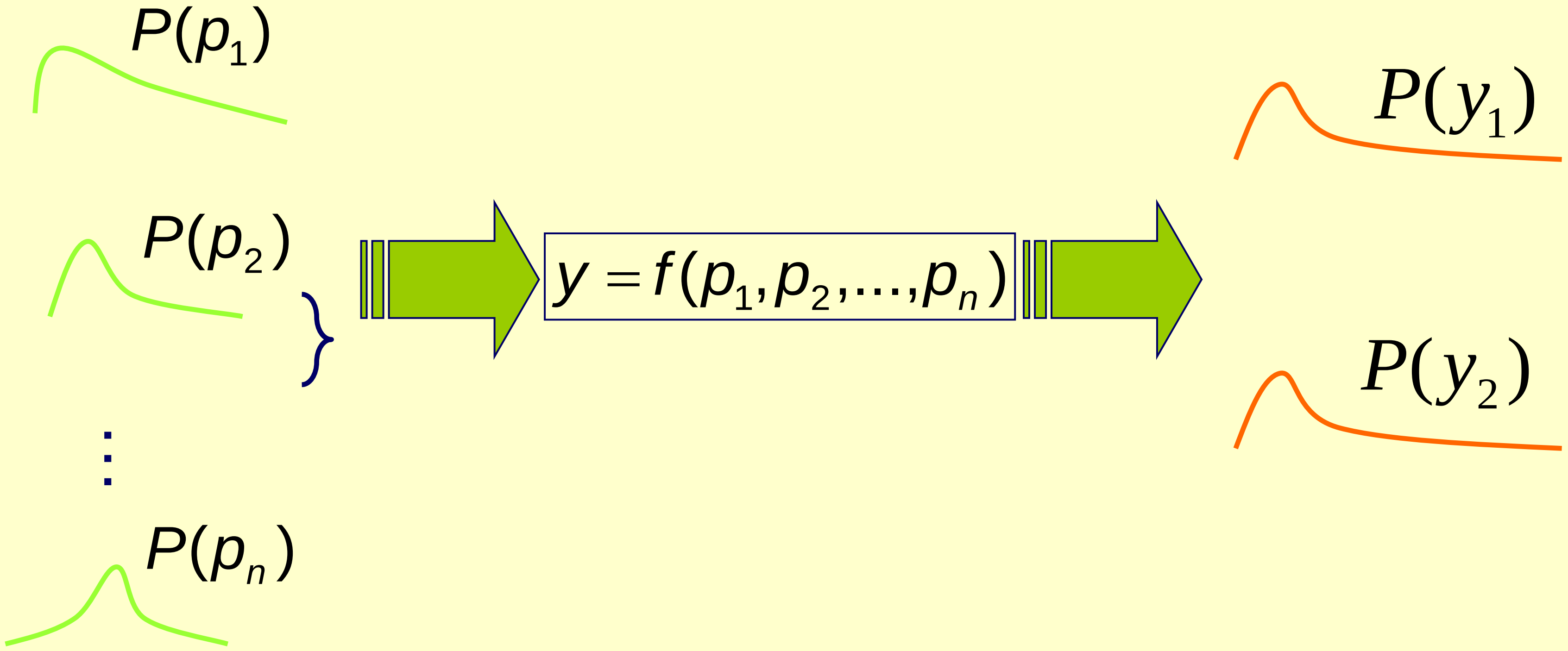
To use global uncertainty analysis more computational time needs than to use local uncertainty analysis

Global uncertainty analysis

The uncertainty of the parameters are written down by their probability density function (*pdf*)

Aims of global uncertainty analysis:

1. What is the *pdf* of the model result according to the parameter's *pdf*?
2. What fraction of the variance of the model result caused by the parameter's uncertainty?



$P(p_1)$

$P(p_2)$

$\vdots$

$P(p_n)$

$$y = f(p_1, p_2, \ldots, p_n)$$

$P(y_1)$

$P(y_2)$

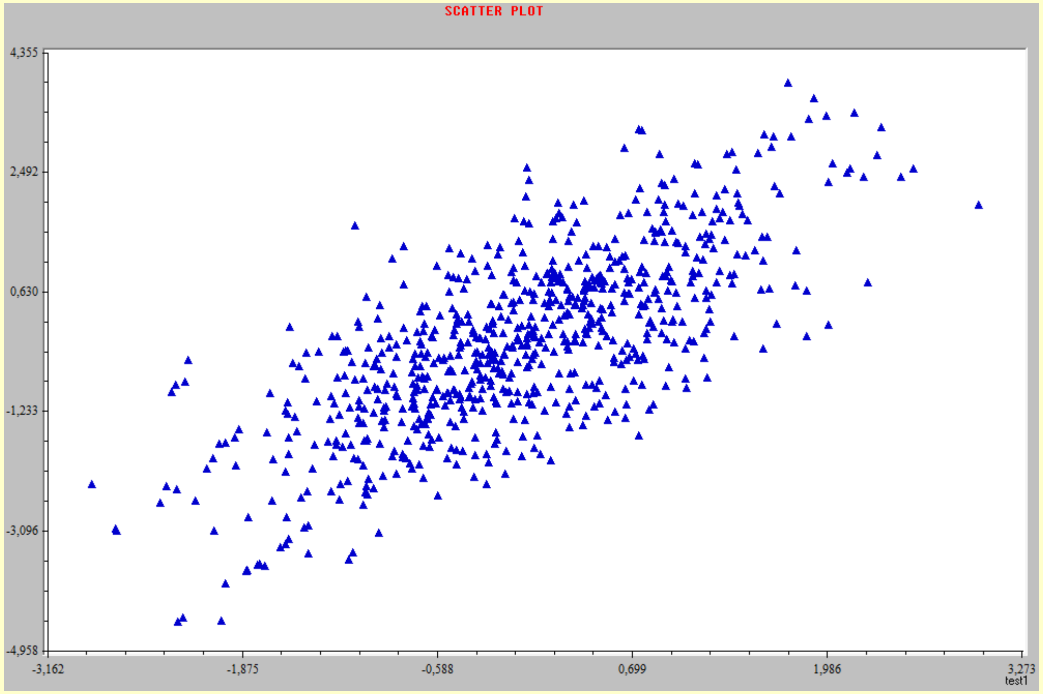
Global uncertainty analysis

I. Scatter plots

One of the easiest methods

Denote x_i (i=1..n) the parameters of the model and y denotes the result of the model

Create a figure belongs to (x_i,y)



Shows the relation between the parameters and the model result (linear, nonlinear, monotonic, non-monotonic), and the strength of relation

Helps to understand the behavior of the model

Disadvantages: Lots of figures have to be created and investigated

The importance of the parameters couldn’t be estimated

Global uncertainty analysis

II. Pearson product moment correlation coefficient (PEAR)

Simply method

Denote x_i (i=1..n) the parameters of the model and y denotes the result of the model

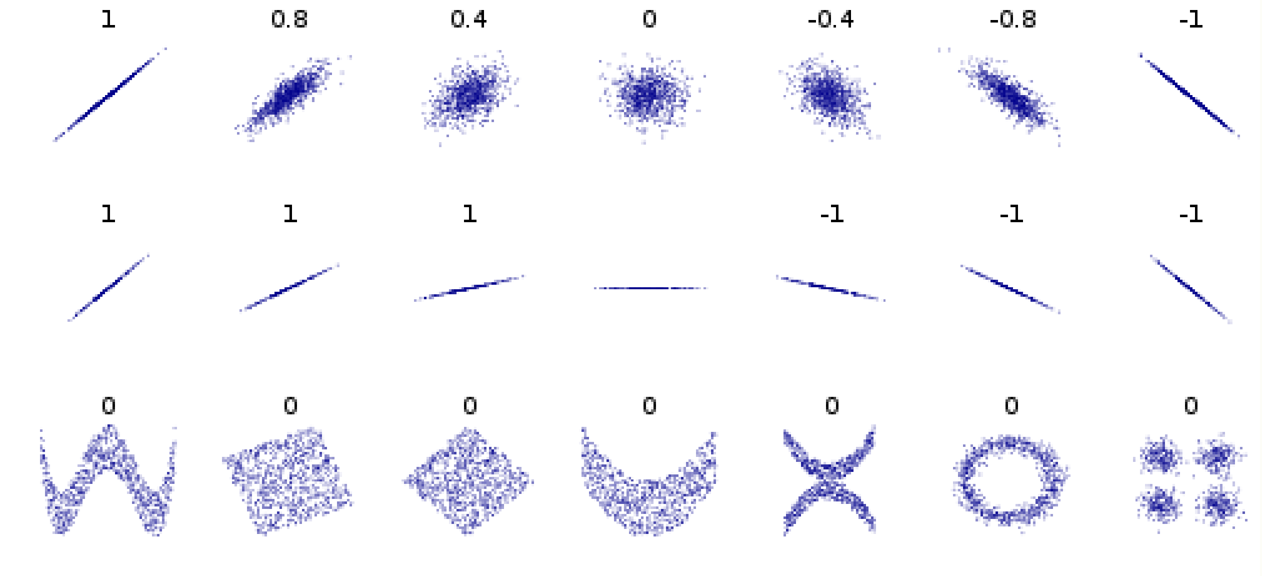
Calculate the correlation between x_i and y

$$PEAR(x_i\,,y) = \frac{\operatorname{cov}(x_i\,,y)}{\sigma_x\sigma_y} = \frac{E[(x_i-\mu_{x_i})(y-\mu_y)]}{\sigma_x\sigma_y}$$

The correlation coefficient is a measure of the linear dependence between two variables x_i and y, giving a value between +1 and −1 inclusive, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation

Global uncertainty analysis

II. Pearson product moment correlation coefficient (PEAR)



The correlation coefficient is a measure of the linear dependence between two variables x_i and y, giving a value between +1 and −1 inclusive, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation

Global uncertainty analysis

III. Spearman coefficient (SPEA)

Simply method

Denote x_i (i=1..n) the parameters of the model and y denotes the result of the model

Calculate the next correlation coefficient:

$$SPEA(x_i, y) = PEAR(R(x_i), R(y))$$

where $R(.)$ indicates the transformation which substitutes the variable value with its rank.

For non-linear models the Spearman coefficient is preferred as a measure of correlation

Basic assumptions:

Both the x_i and y are random samples from their respective populations

The measurement scale of both variables is at least ordinal

Global uncertainty analysis

IV. Standardised Regression Coefficient (SRC)

More quantitative measures of sensitivity are based on regression analysis.

If a linear regression model is being sought, it takes the form

$$y_i = b_0 + \sum_j b_j x_{ij} + \varepsilon_i$$

where y_i =1,…, are the output values of the model, b_j, j=1,…,k (k being the number of input variables) are coefficients that must be determined and ε_i is the error (residual) due to approximation. One common way of determining the coefficients b_j is using the least square method.

Global uncertainty analysis

IV. Standardised Regression Coefficient (SRC)

Assuming that **b** has been computed using least square method

$$\hat{s} = \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{m-1}} \qquad \hat{s}_j = \sqrt{\frac{\sum_i (x_{ij} - \bar{x}_j)^2}{m-1}}$$

$$\bar{x}_j = \frac{\sum_i x_{ij}}{m} \qquad \bar{y} = \frac{\sum_i y_i}{m} \qquad \text{Standard Regression Coefficients} : b_j \frac{\hat{s}}{\hat{s}_j}$$

$$R_y{}^2 = \sum_{i=1}^{m} (\hat{y}_i - \bar{y}) / \sum_{i=1}^{m} (y_i - \bar{y})^2$$

→ Provides a measure of how well the linear regression model based on SRC’s can reproduce the actual output *y*

Global uncertainty analysis

IV. Standardised Regression Coefficient (SRC)

$$R_y{}^2 = \sum_{i=1}^{m} (\hat{y}_i - \bar{y}) / \sum_{i=1}^{m} (y_i - \bar{y})^2 \longrightarrow$$

Provides a measure of how well the linear regression model based on SRC’s can reproduce the actual output *y*

$R_y{}^2 \longrightarrow$ Represents the fraction of the variance of the output explained by the regression. The closer R_y^2 is to unit, the better is the model performance.

Important: The variables of the model have to be independent!

Global uncertainty analysis

V. Partial Correlation Coefficient(PCC)

Denote x_{i} (i=1..n) the parameters of the model and y denotes the result of the model

1.Calculate the coefficients of the next regression functions:

$$\hat{y} = b_0 + \sum_{h \neq j} b_h x_h \qquad \hat{x}_j = c_0 + \sum_{h \neq j} c_h x_h$$

2. The partial correlation coefficient between variable x_j and model result y

$$PEAR(y - \hat{y}, x_j - \hat{x}_j) \text{ or } SPEAR(y - \hat{y}, x_j - \hat{x}_j)$$

PCC gives the strength of the correlation between y and a given input x_j cleaned of any effect due to any correlation between x_j and any of the x_i, $i \neq j$.

Global uncertainty analysis

VI. Standardised Rank Regression Coefficients(SRRC)

Regression analysis often performs poorly when the relationships between the input variables are non-linear.

The rank transform is a simple procedure which involves replacing the data with their corresponding ranks. The usual least square regression analysis is then performed entirely on these ranks

The new value of ${R_y}^2$ (on ranks) is computed.

$${R_y}^2 = \sum_{i=1}^{m}(\hat{y}_i - \bar{y}) / \sum_{i=1}^{m}(y_i - \bar{y})^2$$

If this new value is higher, then the new coefficients SRRC, can be used for sensitivity analysis instead of SRC's.

Global uncertainty analysis

VII. Partial Rank Correlation Coefficients (PRCC)

The Partial Correlation Coefficients can be computed on the ranks (Partial Rank Correlation Coefficients). The performance of the PRCC shows the same features of performance as the SRCC: good for monotonic models, and not fully satisfactory in the presence of non-monotonicity.

Global uncertainty analysis

VIII. Kolmogorov-Smirnov test (SMIRNOV)

Model simulations are classified as either behavioral (B) or non-behavioral ($\overline{B}$). A set of binary elements are defined distinguish between two sub-sets of each input factor $X_i : (X_i \,|\, B)$ of m elements and $(X_i \,|\, \overline{B})$ of n elements (n+m=N, the total number of Monte Carlo runs performed). Under the null hypothesis that the two distributions are identical $H_0 : f_m(X_i \,|\, B = f_n(X_i \,|\, \overline{B})$

$$d_{m,n}(X_i) = \sup \| F_m(X_i \,|\, b) - F_n(X_i \,|\, \overline{B}) \|$$

To perform the Smirnov test, we choose the significance level α, which is the probability of rejecting H_0 when it is not true. From α we derive D_α at which the computed value $d_{m,n}$ determines the rejection of H_0.

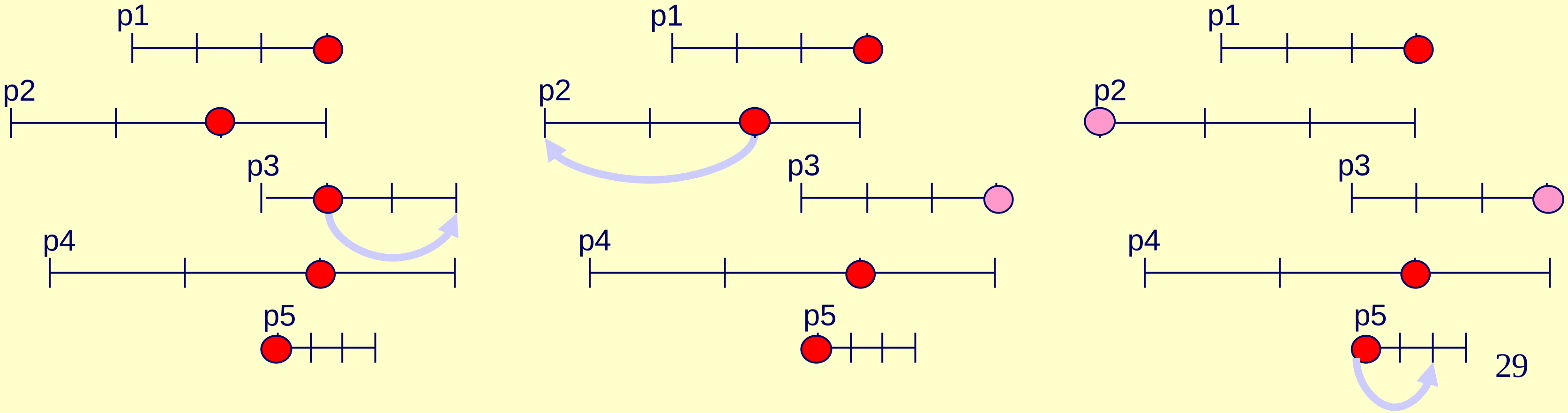
If $d_{m,n} > D_\alpha$ then H_0 is rejected at significance level α and the factor X_i is considered as important.

Simlab uses α=5%.

Global uncertainty analysis

IX. Morris Method

Morris estimates the main effect of a factor by computing a number d_{ij} of local measures, at different points in the input space, and taking their average. These d_{ij} values are selected such that each factor is varied over its interval of experimentation. Morris wishes to determine which factors have negligible effects, linear, additive effects, non-linear or interaction effects.



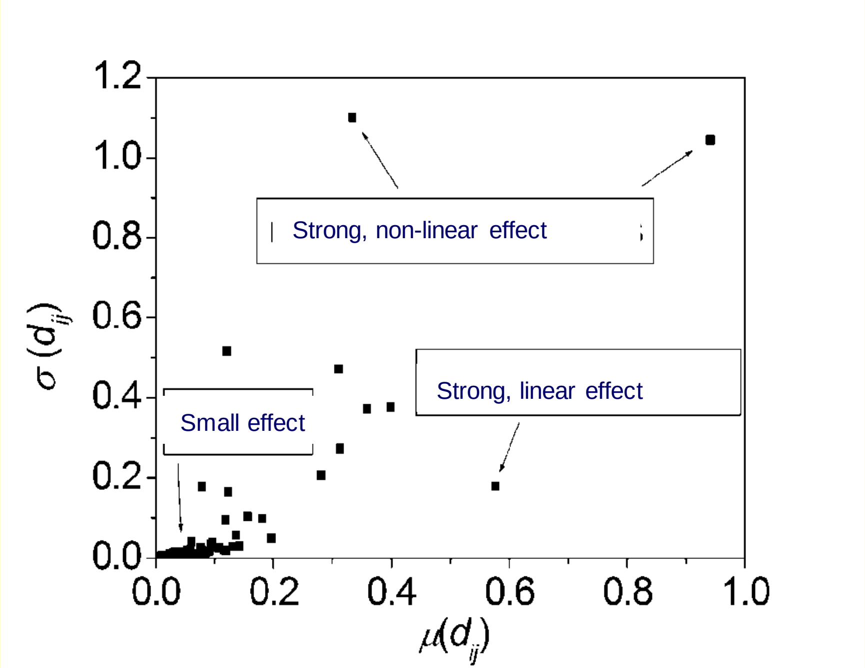
Global uncertainty analysis

IX. Morris Method

d_{ij} shows the effect of parameter p_j when the other parameter values are fixed

$$d_{ij} = \frac{Y_i\left(p_1^z, p_2^z, \ldots, p_j^z + \Delta, \ldots, p_N^z\right) - Y_i\left(\mathbf{p}^{z-1}\right)}{|\Delta|}$$

In every step, we change one parameter value (the other values are fixed) We calculate many times the value of d_{ij} and we determine the average and standard deviation of these d_{ij} values

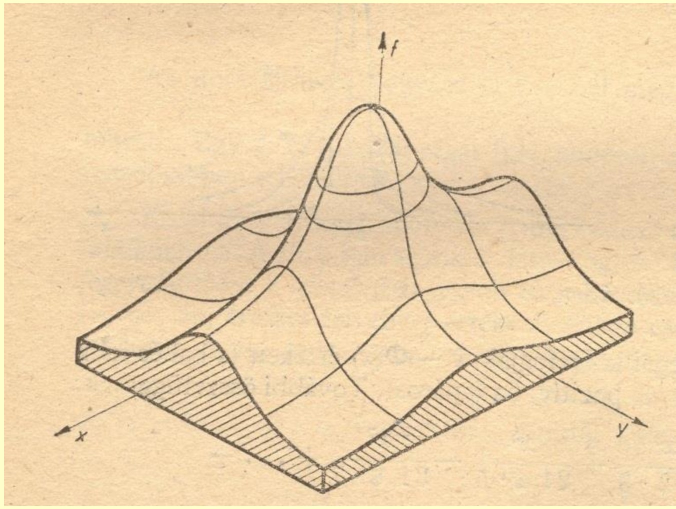


Global uncertainty analysis

X. Monte Carlo Method

Playing roulette in Monte Carlo :o)

We create lots of samples (thousands) based on the distribution function of the parameters



For every parameter sets we calculate the model results

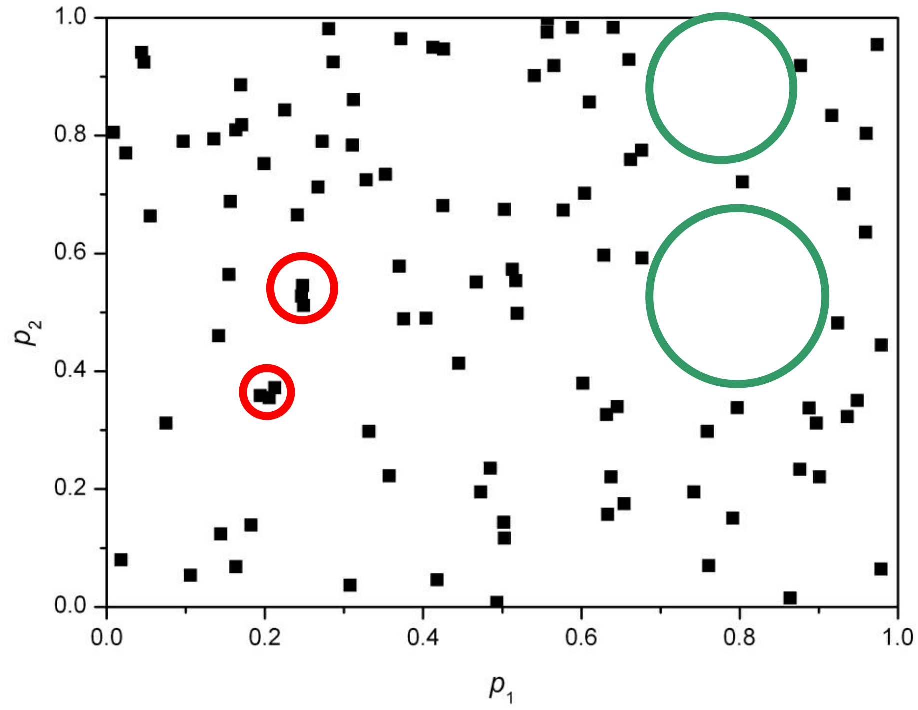
We calculate the average, standard deviation, distribution of the results of the model, create histograms, determine the marginal distributions…etc.

Very expensive (computational time)

Difficult to separate the effect of the parameters

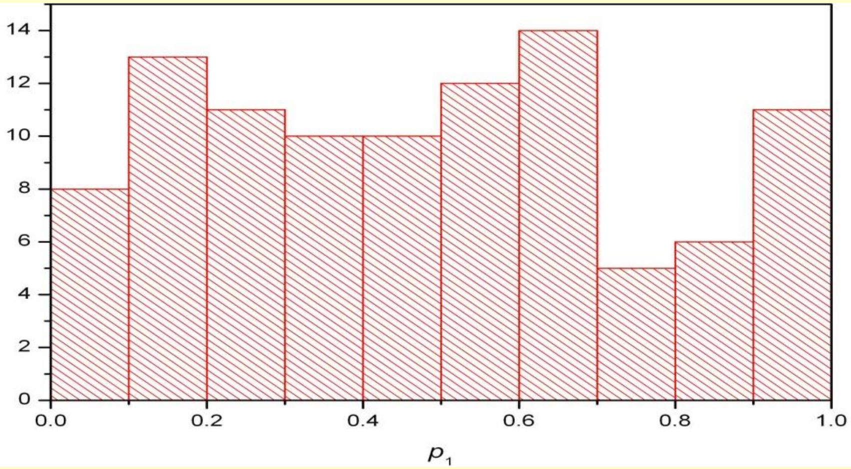
Global uncertainty analysis

X. Monte Carlo Method



100 (ξ,γ) points
ξ and γ random numbers
Uniform distribution in [0,1]

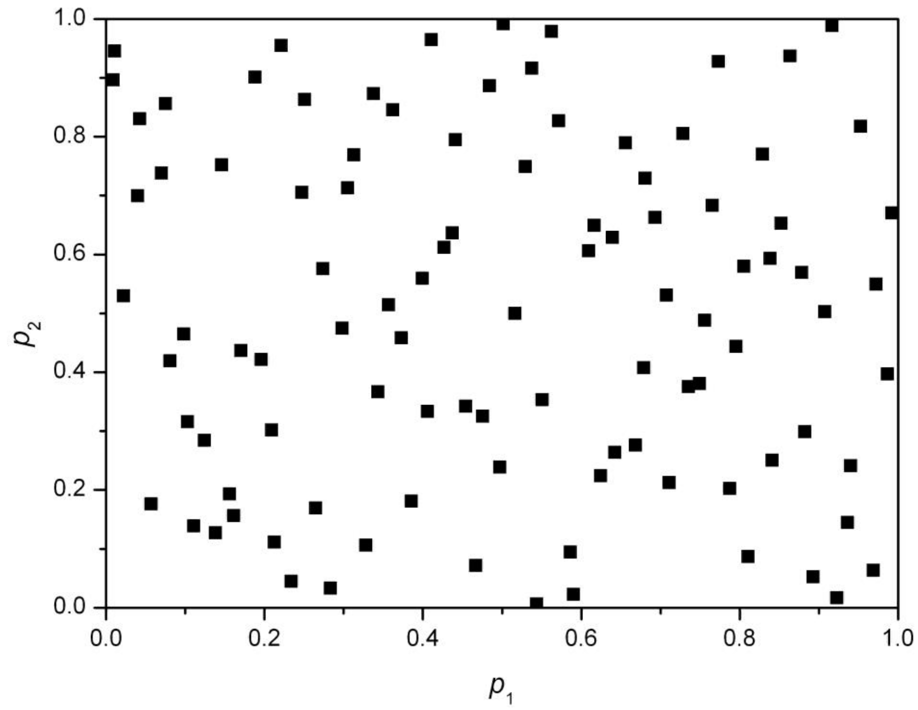
Clustering and
White places!



Histogram of p_1

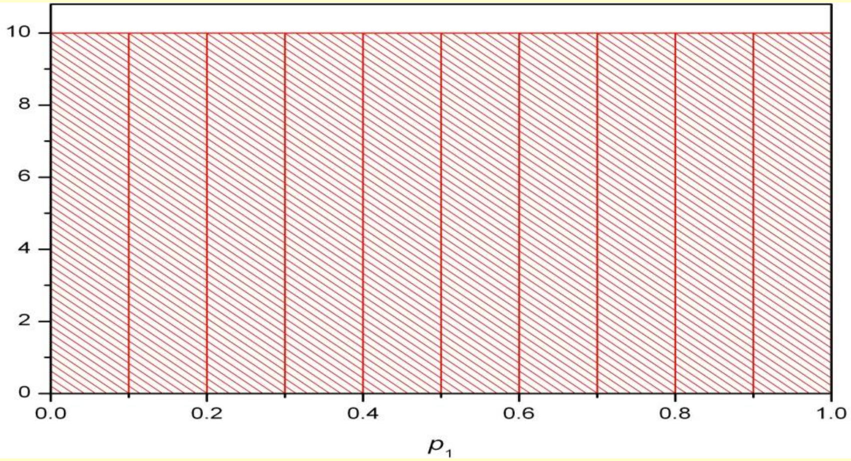
Global uncertainty analysis

X. Monte Carlo Method



100 (ξ,γ) points
Latin Hypercube Sampling

More uniform distribution



Histogram of p_1

Global uncertainty analysis

XI. Sensitivity indices(Sobol)

Saltelli, A., *Comput. Phys Commun.*, **145**, 280(2002)

Expected value of Y_i

$$E(Y_i)=\iint\ldots\int f_i(p_1,p_2,\ldots,p_N)P(p_1,p_2,\ldots,p_N)\mathrm{d}\,p_1\,\mathrm{d}\,p_2\ldots\mathrm{d}\,p_N$$

Variance of Y_i

$$V(Y_i)=\iint\ldots\int\left(f_i(p_1,p_2,\ldots,p_N)-E(Y_i)\right)^2P(p_1,p_2,\ldots,p_N)\mathrm{d}\,p_1\,\mathrm{d}\,p_2\ldots\mathrm{d}\,p_N=$$

$$=\iint\ldots\int f_i^2(p_1,p_2,\ldots,p_N)P(p_1,p_2,\ldots,p_N)\mathrm{d}\,p_1\,\mathrm{d}\,p_2\ldots\mathrm{d}\,p_N-E^2(Y_i)$$

Variance of Y_i when the value of p_j is fixed: $V(Y_i|p_j)$

Expected value of this: $E(V(Y_i|p_j))$

Variance of Y_i caused by p_j $V(E(Y_i|p_j))=V(Y_i)$- $E(V(Y_i|p_j))$

First order sensitivity index:

$$S_{j(i)}=\frac{V\left(E\left(Y_i\middle|p_j\right)\right)}{V(Y_i)}$$

Global uncertainty analysis

XI. Sensitivity Indices (Sobol)

Variance of Y_i caused by p_j and p_k $V(E(Y_i|p_j,p_k))$

Second order sensitivity index:

$$S_{kj(i)} = \frac{V\left(E\left(Y_i \middle| p_k, p_j\right)\right) - V\left(E\left(Y_i \middle| p_k\right)\right) - V\left(E\left(Y_i \middle| p_j\right)\right)}{V\left(Y_i\right)}$$

Any higher order sensitivity index can be calculated similarly

Denote the three parameters of a model: *a, b, c*

Total index: $S_{a(i)}^{\mathrm{tot}} = S_{a(i)} + S_{ab(i)} + S_{ac(i)} + S_{abc(i)}$

If parameter *j* independent of the other parameters $S_{j(i)} = S_{j(i)}^{\mathrm{tot}}$

Global uncertainty analysis

XI. Sensitivity Indices (Sobol)

- Global method
- Uses random samples to calculate the integrals fast
- First and second order effects of the parameters
- Calculates the total effect
- Uses the *pdf* of the parameters
- Very expensive (computational time)
  (for 50 parameters ~ 25000 runs)

Uncertainty Analysis

Local uncertainty analysis

changing only one parameter at time

based on partial derivatives

fast calculation (time and computational time)

Filtering methods

changing lots of parameters at time

slower calculation

Morris Method

Global uncertainty analysis

changing every parameters at time according to the joint probability density function

very expensive

Monte Carlo Analysis with Latin Hypercube Sampling,

Sensitivity Indices

Thank you for your attention!