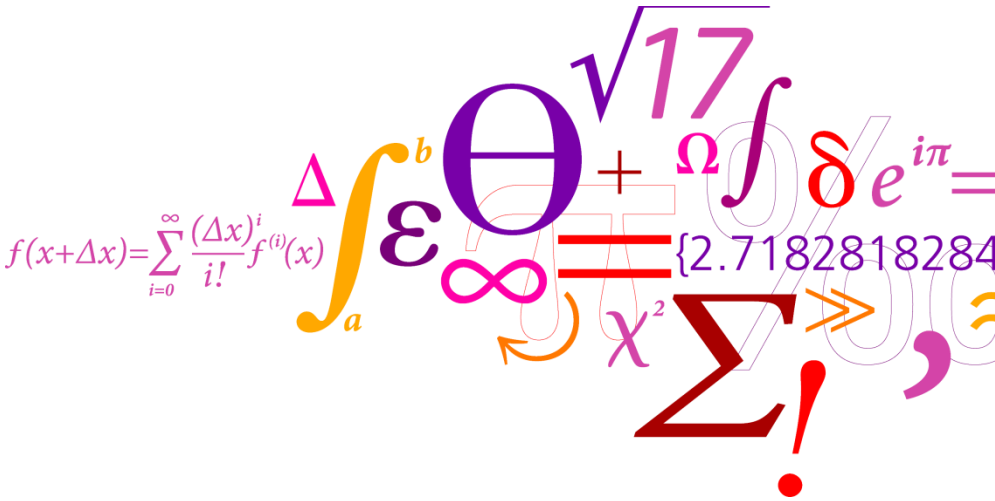
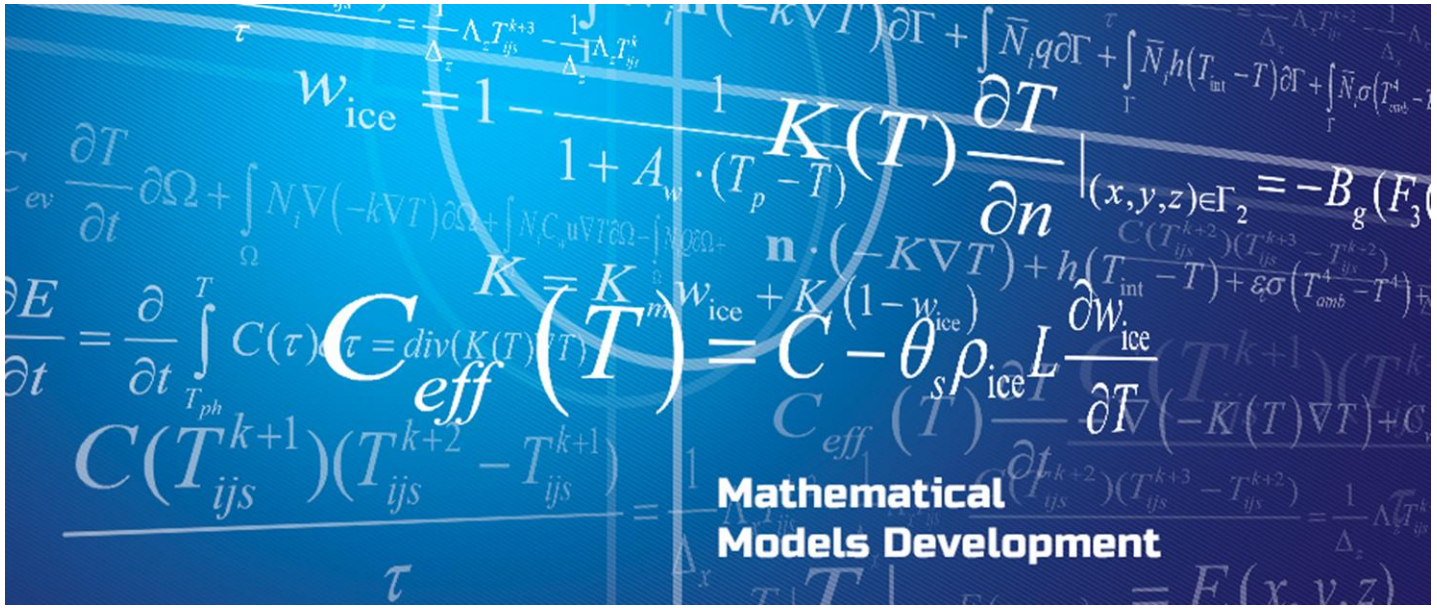
Model prediction uncertainty using Monte Carlo method

Gürkan Sin, Associate Professor
PROSYS research center
DTU Chemical Engineering
gsi@kt.dtu.dk



DTU Kemiteknik
Institut for Kemiteknik

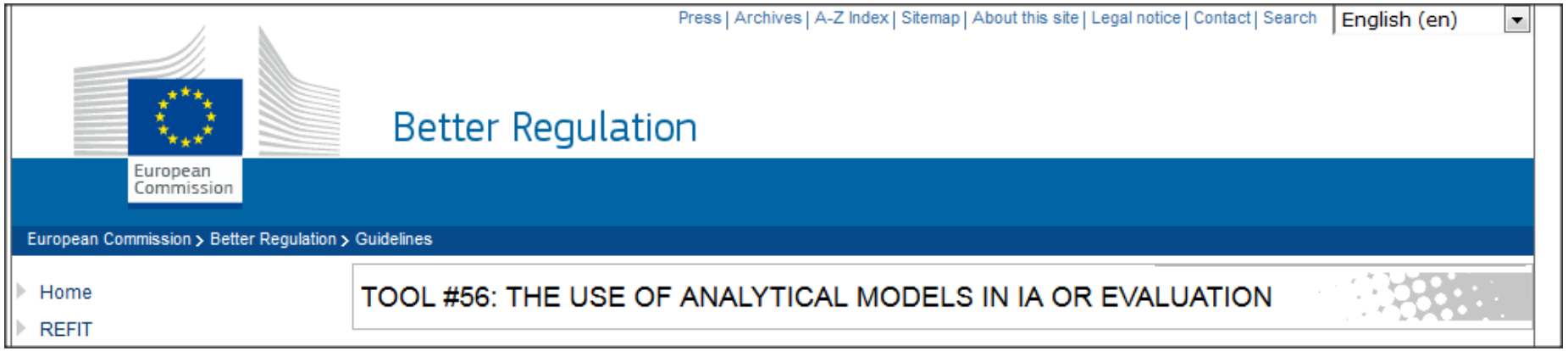
Problem setting for uncertainty



To simulate/understand complex phenomena when experiments are too expensive or impractical,

To make predictions and support decisions

Problem setting



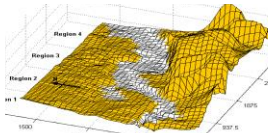
"All models are simplifications, but good models provide insights and understanding if used correctly."

"It is important to ensure that the right model is selected and used in a manner to deliver policy relevant results of the requisite quality."

Computer models may calculate several output values (scalars or time/space dependent functions) that can depend on a high number of input parameters and physical variables.

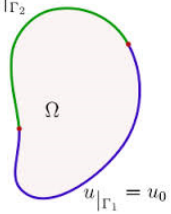
Parameters

Input data



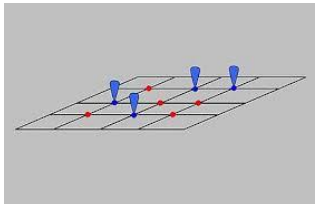
Boundary conditions

$$\left.\frac{\partial u}{\partial n}\right|_{\Gamma_2} = g$$



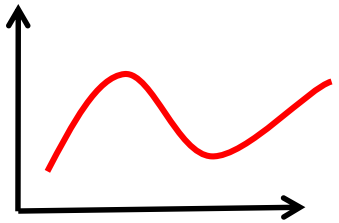
$$u_{|\Gamma_1} = u_0$$

Assumptions



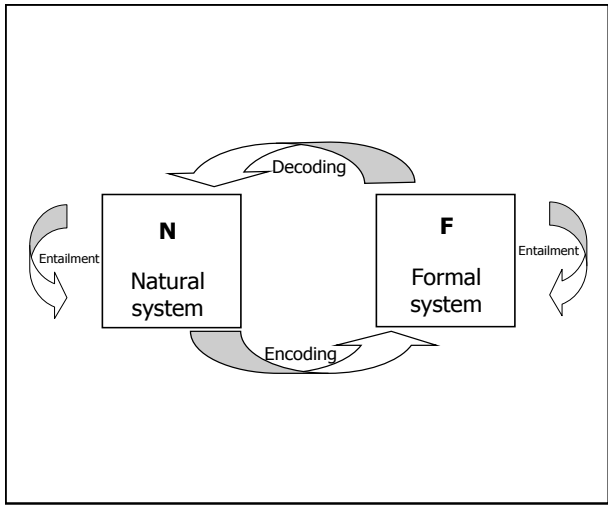
M

Computer Model



Time Series

Hazard maps



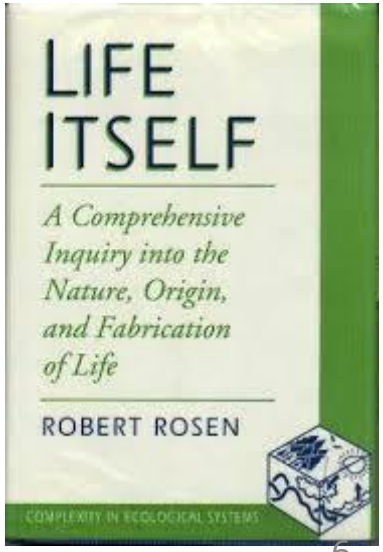
Modelling is a craftmanship

…

(R. Rosen, Life itself 1991)

Modelling is an encoding process from the phenomenon to the formal system: theories, laws, subjective choices, values, assumptions

Different assumptions can lead to different predictions



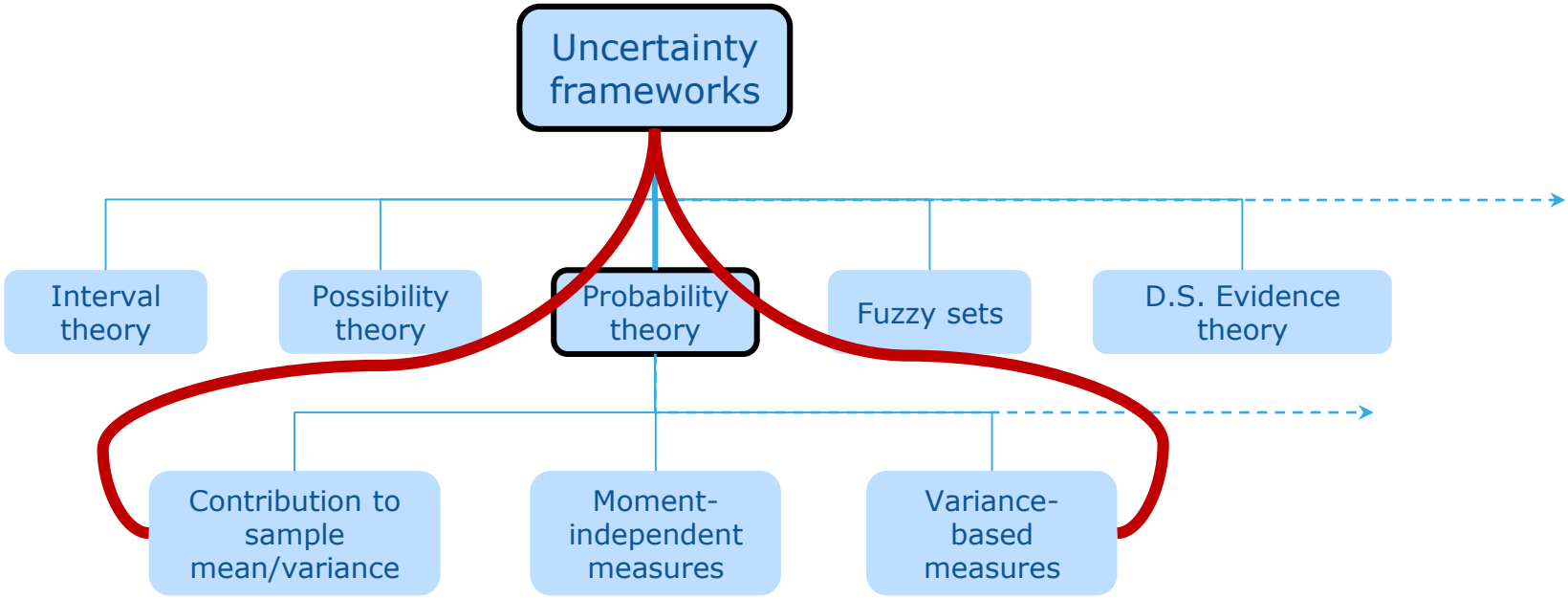
Types of uncertainty in the encoding process:

- Uncertainty in data
- Spatial resolution of data maps
- Uncertainty in parameter values and coefficients
- Modelling assumptions,

but are rarely taken into account.

Need to quantify uncertainties.

Quantifying uncertainty: context



Monte Carlo Method – definition

General definition

"Monte Carlo Method" refers in general to solution techniques that uses random numbers and probability statistics to investigate problems (and hopefully come up with approximate solutions.)

Statistical definition (Halton, 1970)

Representing the solution of a problem as a hypothetical population, and using a random sequence of numbers to construct a sample of the population, from which statistical estimates of the parameter is possible!

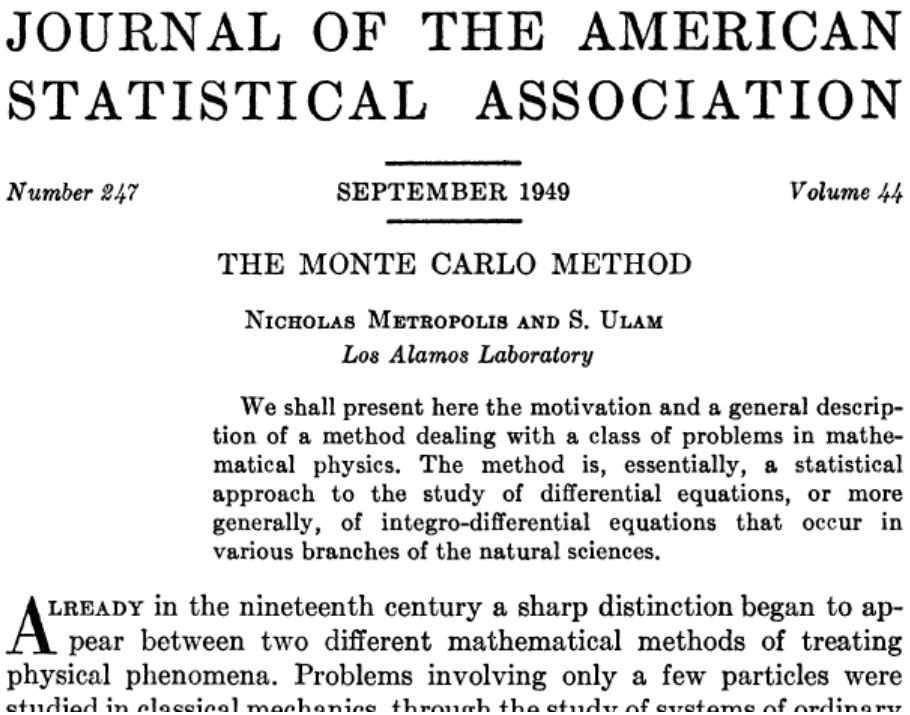
Monte Carlo Method – motivation

To find numerical solutions to problems, which are too complicated or impossible to solve analytically .

Example 1 – Volume of 20-dimensional unit cube

Traditional approach: divide each axis (x_1, x_2….x_{20}) into 10 intervals and calculate the 10^{20} lattice points (UNDOABLE)

Practical approach consider say 10^4 points at random of this ensemble & analyze them (DOABLE)



JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 247 SEPTEMBER 1949 *Volume 44*

THE MONTE CARLO METHOD

NICHOLAS METROPOLIS AND S. ULAM
Los Alamos Laboratory

We shall present here the motivation and a general description of a method dealing with a class of problems in mathematical physics. The method is, essentially, a statistical approach to the study of differential equations, or more generally, of integro-differential equations that occur in various branches of the natural sciences.

ALREADY in the nineteenth century a sharp distinction began to appear between two different mathematical methods of treating physical phenomena. Problems involving only a few particles were studied in classical mechanics, through the study of systems of ordinary

Monte Carlo Method – today

One finds MC methods used across many disciplines from economics to nuclear physics, climate change, computer graphics, and of course chemical engineering among many others...

The way MC methods are applied varies from field to field reflecting on the nature of the problem being solved…

(say integration in multi-dimensional space versus sensitivity analysis, etc)

Monte Carlo for multidimensional integration

We consider the integral of a function $f(u_1,\ldots u_d)$, depending on d variables $u_1, \ldots, u_d$ over the unit hypercube $[0, 1]^d$. We assume that f is square-integrable. As a short-hand notation we will denote a point in the unit hypercube by $x = (u1, \ldots, ud)$ and the function evaluated at this point by $f(x) = f(u1, \ldots, ud)$, then the multidimensional integration operation is given by:

$$I = \int f(x)dx = \int f(u_1 \ldots u_d)d^d x$$

The Monte Carlo estimate for this integral is given by:

$$E = \frac{1}{N}\sum_1^N f(x_N) \qquad \lim_{N\to\infty} \frac{1}{N}\sum_1^N f(x_N) = I$$

The law of large numbers ensures that the MC estimate converges to true value

Error of the Monte Carlo estimate for integration

Since most of the time finite N is used, there will be error in Monte Carlo integration of multidimensional functions. This Monte Carlo integration error scales like 1/sqrt(N). Hence the average monte carlo integration error is given by $\sigma(f)/\sqrt{N}$ where *σ(f)* is the standard deviation of the error which can be approximated using sample variance:

$$\sigma^2(f) \approx s^2(f) = \frac{1}{N-1}\sum_{1}^{N}\left(f(x_N) - E\right)^2$$

Monte Carlo integration for uncertainty analysis

For notational simplicity, we consider the following simple model:

$$\mathbf{y}=f\left(\mathbf{x}\right)$$

where the function **f** represents the models under study, **x:[**x1;… xd] is a vector of model inputs, and **y [**y1;… yn] is a vector of model predictions.

The goal of an uncertainty analysis is to determine the uncertainty in the elements of **y** that results from uncertainty in the elements of **x.** Given uncertainty in the vector **x** characterised by distribution functions, D=[D1,….Dd] where D1 is the distribution function associated with x1, the uncertainty in y is given by:

$$\mathrm{var}\left(y\right)=\int\left(\left(y\right)-f\left(\mathbf{x}\right)\right)^{2}d\mathbf{x}$$

$$E\left(y\right)=\int f\left(\mathbf{x}\right)d\mathbf{x}$$

Both of which requires integration of $f(\mathbf{x})$ over the joint multidimensional probability distribution space of **x.** Hence the use of monte carlo simulations.

Workflow for Monte Carlo method for uncertainty analysis

- Step 1. Input uncertainty definition

Identify which inputs (parameters) have uncertainty. Define a range/distribution for each uncertainty input, e.g. normal distribution, uniform distribution, etc. The output from parameter estimators (e.g. bootstrap) can be used as input here.

- Step 2. Sampling from input space

Define sampling number, *N*, (e.g. 50, 100, etc) and sample from input space using an appropriate sampling technique. Most common sampling techniques are random sampling, Latin Hypercube sampling, etc. The output from this step is a sampling matrix, $X_{Nxm,}$ where N is the sampling number and m is the number of inputs.

- Step 3. Perform Monte Carlo simulations

Perform N simulations with the model using sampling matrix from step 2. Record the outputs in an appropriate matrix form to be processed in the next step.

- Step 4. Review and analyse the results

Plot the outputs and review the results. Calculate mean, standard deviation/variance, and percentiles (e.g. 95%) for the outputs. Analyse the results within the context of parameter estimation quality and model prediction uncertainty. Iterate the analysis, if necessary, by going to step 1 or step 2.

Practical experience

Exercise details

Consider the following problem:

$$y = 4x_1^2 + 3x_2$$

with

$$x_1 \sim U(-1, 1)$$

$$x_2 \sim U(-1, 1)$$

Calculate the uncertainty on y. Use the matlab scripts, provided to you. Define your own sampling number. Repeat if necessary your calculations. Consider correlation. Consider different distribution. Etc.

Step 1 input uncertainty characterisation

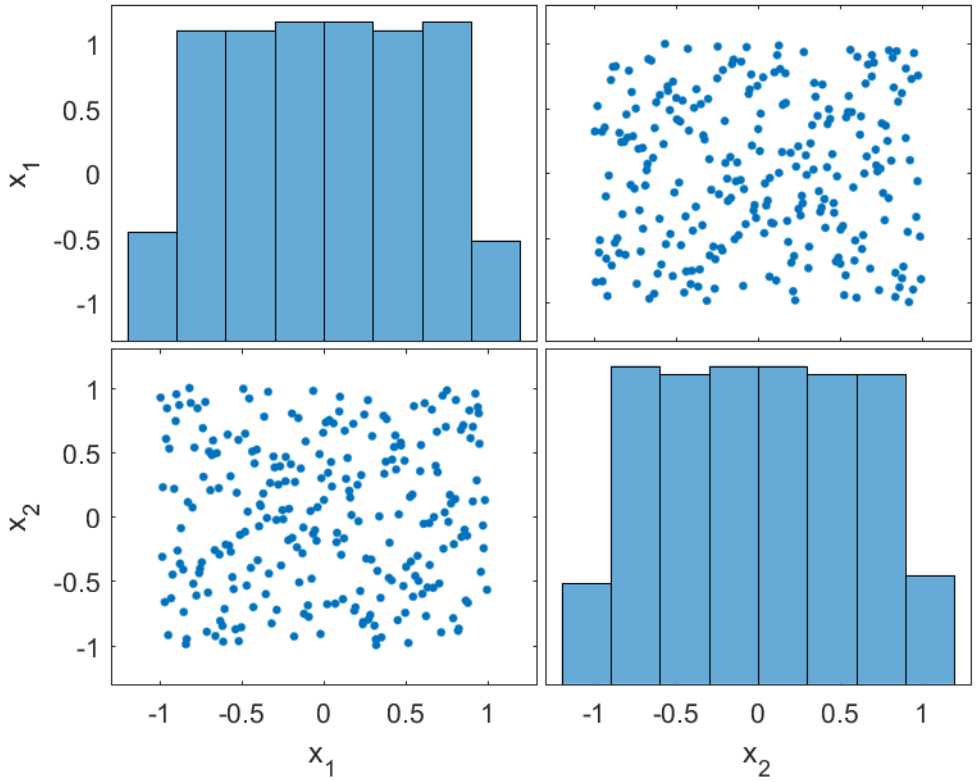
The input uncertainty is already defined for this hypothetical example. Both uncertainty parameters follow a uniform distribution with certain upper and lower bounds:

$$x_1 \sim U(-1, 1)$$

$$x_2 \sim U(-1, 1)$$

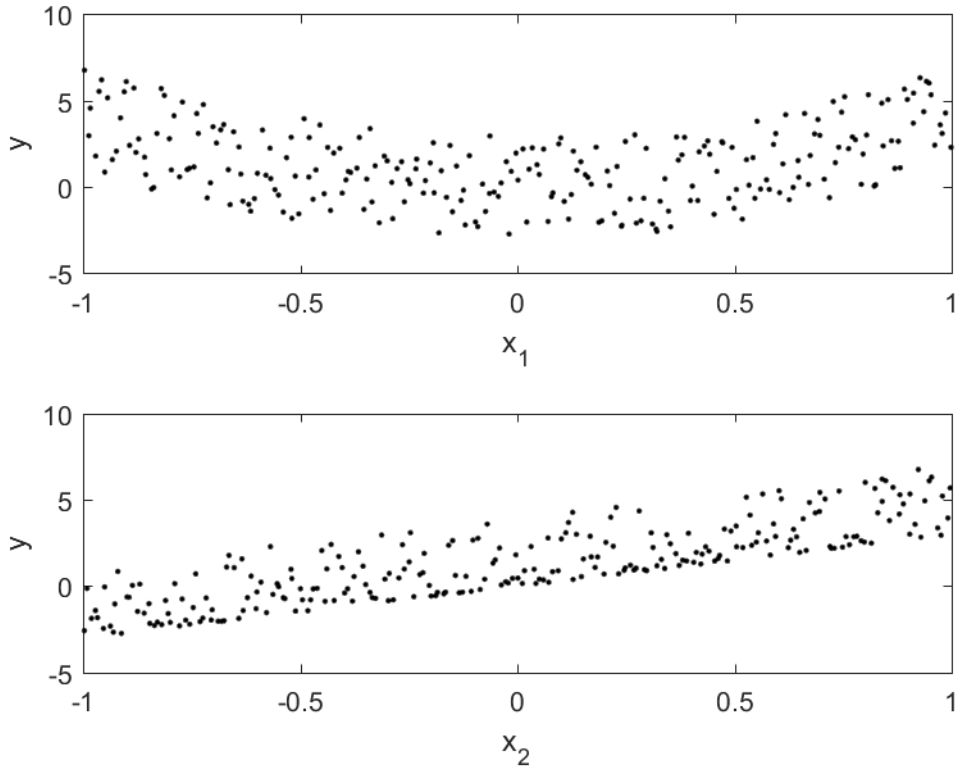
Step 2 sampling from input domain

In this example, since we define the parameters to follow a uniform distribution, the input uncertainty space is represented by a multivariate uniform distribution. We use Latin Hypercube Sampling (LHS) technique to sample from this space. The output from this step is a sampling matrix, $X_{Nxm,}$ where N is the sampling number and m is the number of inputs



Step 3 & 4 perform monte carlo simulations and evaluate the results

- In this step, we perform N model simulations using sampling matrix from step 2 (X_{Nxm}) and record the model outputs in a matrix form to be processed in the next step.



Monte Carlo simulations (N=250) of model outputs – sortd wrt x

Step 3 & 4 perform monte carlo simulations and evaluate the results

- Check the convergence statistics.

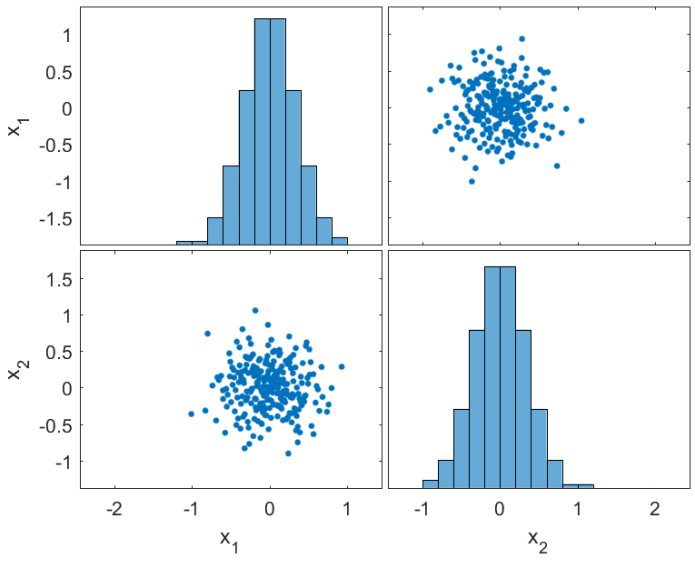
VARIANCE AND MONTE CARLO INTEGRATION ERROR (MCERR)

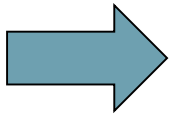
mean	Variance	MCerr
1.3335	4.6673	0.13663

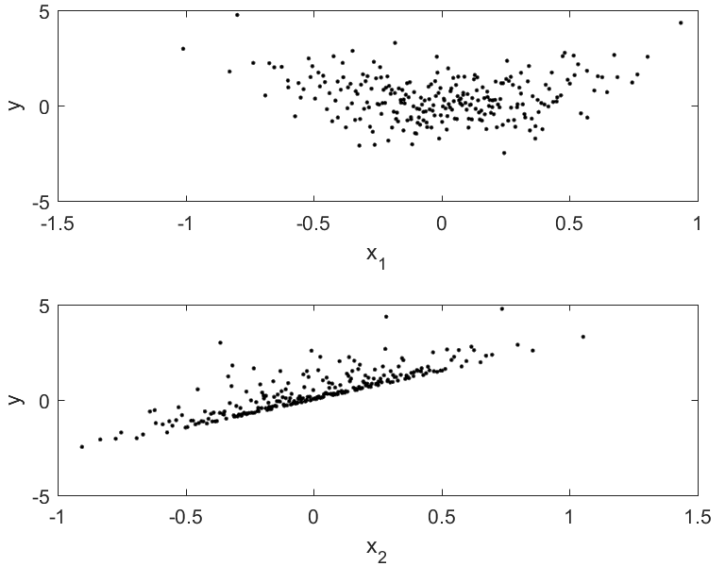
- What else are important in this analysis?

Uncertainty analysis versus INPUT uncertainty

- What if the inputs are normally distributed (equivalent to variance of prev example)?







$x_1 \sim N(0, 1/3)$

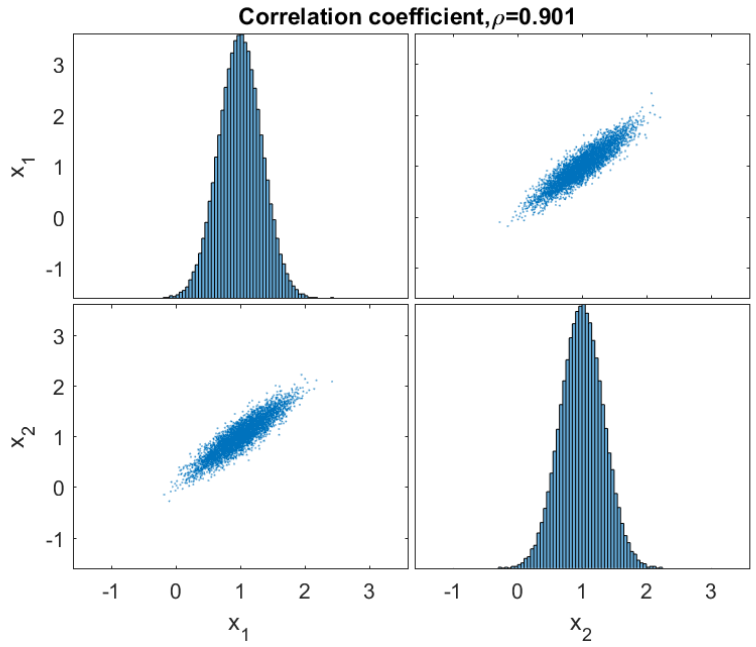
$x_2 \sim N(0, 1/3)$

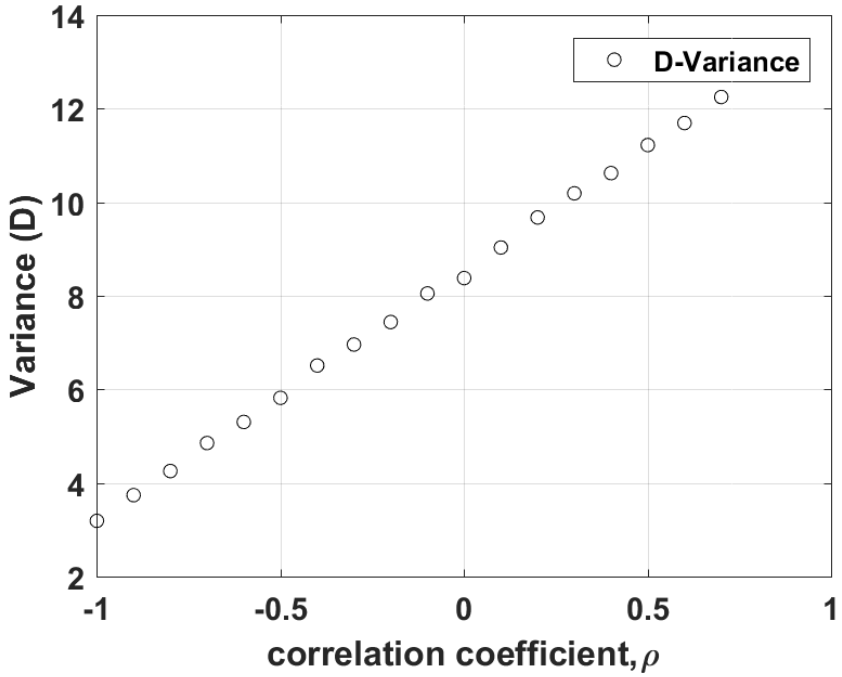
mean	Variance	MCerr
0.44487	**1.3213**	0.0727

Uncertainty analysis versus INPUT uncertainty

- What about the effect of correlation in the input space? In the presence of correlation, covariance matrix of the inputs becomes multivariate normal distribution with SIGMA:

$$\boldsymbol{x} \sim MVN\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1/3 & \rho_{12}\,1/9\\ \rho_{12}\,1/9 & 1/3\end{bmatrix}\right)$$

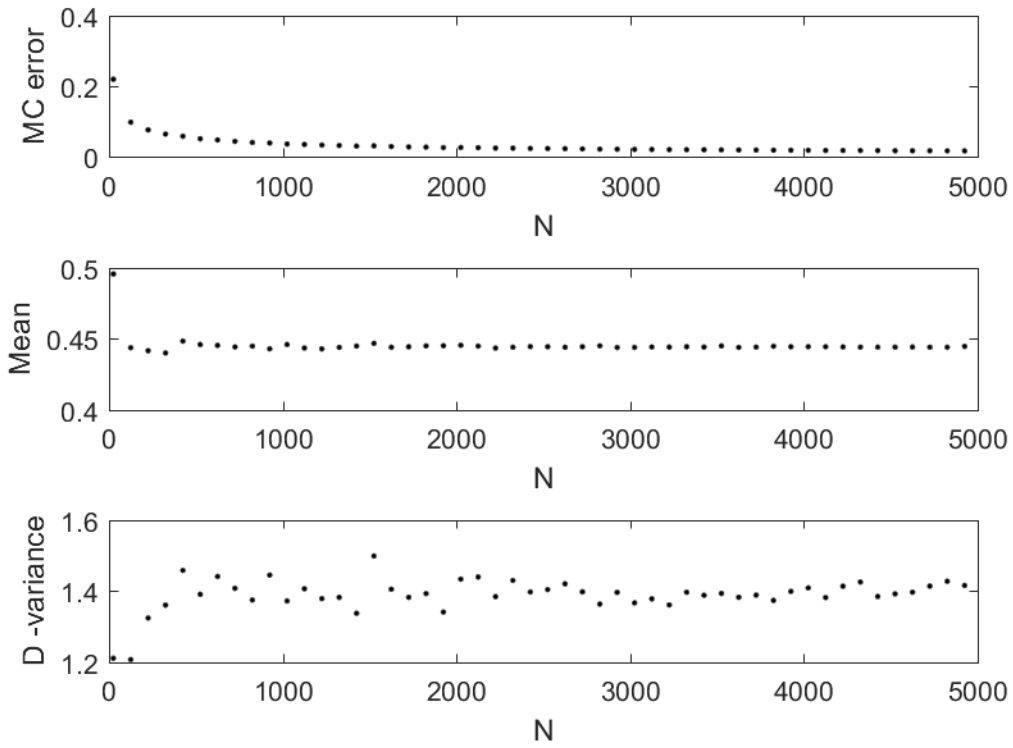




VARIANCE AND CORRELATION BETWEEN INPUTS: VERY IMPORTANT. BOTH MAGNITUDE AND DIRECTION OF CORRELATION IMPACT HUGELY THE CALCULATED OUTPUT VARIANCE. BEWARE.

Uncertainty analysis versus SAMPLING number

- Convergence statistics versus sampling number



VARIANCE AND MONTE CARLO INTEGRATION ERROR (MCERR):CACULATION OF THE MEAN CONVERGES QUICKLY. VARIANCE TAKES SOME MORE N.

Best practice in uncertainty analysis

When performing uncertainty analysis, the most important issue is the framing and the corresponding definition of the input uncertainty sources. Hence, the outcome from an uncertainty analysis should not be treated as absolute but dependent on the framing of the analysis.
A detailed discussion of these issues can be found elsewhere (e.g. Sin et al., 2009; Sin et al., 2010).

Another important issue is the covariance matrix of the parameters (or correlation matrix), which should be obtained from a parameter estimation technique. Assuming the correlation matrix is negligible may lead to over or under estimation of the model output uncertainty. Hence, in a sampling step the appropriate correlation matrix should be defined for inputs (e.g. parameters) considered for the analysis.

Regarding the sampling number, mean calculation converges very quickly as function of N. In either case, one needs to check the reproducibility by hypothesis testing (two means obtained from two iterations statistically different or not).

Work on your own further with the example(s)

- Open matlab if you have it accessible
- Load the example_uncertainty_uniform.m script
- Go throught the code.
- Try different assumptions:
  - Specify normal distribution for the inputs and run! What do you obtain? Mean versus variance? Pattern?
  - specify different sample size?
  - Try different sampling. E.g. Random sampling and see what you obtain. Etc etc

exercise

- Quantify the uncertainty in the model predictions of AOO example from lecture L1.3 (use the uncertainty obtained by bootstrap method for the parameters).

Further reading

- Metropolis, Nicholas, and Stanislaw Ulam. "The monte carlo method." *Journal of the American statistical association* 44.247 (1949): 335-341.
- Sin, G., Gernaey, K. V., Neumann, M. B., van Loosdrecht, M. C., & Gujer, W. (2009). Uncertainty analysis in WWTP model applications: a critical discussion using an example from design. Water Research, 43(11), 2894-2906.
- Sin, G., Gernaey, K. V., Neumann, M. B., van Loosdrecht, M. C., & Gujer, W. (2011). Global sensitivity analysis in wastewater treatment plant model applications: prioritizing sources of uncertainty. Water research, 45(2), 639-651.
- Prunescu, RM, Blanke, M, Jakobsen, JG & Sin, G 2015, 'Dynamic Modeling and Validation of a Biomass Hydrothermal Pretreatment Process - A Demonstration Scale Study' AIChE Journal., 10.1002/aic.14954