Design of computer experiments

Bertrand Iooss and Nicolas Bousquet
Starting point: uncertainties everywhere in the modeling chain!

Main problem: credibility of predictions

- Physical phenomenon
  - Input data
    - Variables of interest
      - Parameters
      - Observations
    - Stochastic uncertainties
    - Epsitemic uncertainties
  - Numerical model
  - Coding Errors
    - Computer scientist
    - Statistician
    - Mathematician
  - Model uncertainties

- Simplifications
  - Physical model
  - Numerical approximation
    - Numerical uncertainties
Similar safety and uncertainty issues in CS&E and Nature sciences

Climate Modeling: Prediction

Nuclear industry: Conception, Maintenance, risks

Oil, gas, CO2: Production optimization

Astrophysics: Understanding

Car and plane: Conception
Main stakes of uncertainty management

• **Modeling phase:**
  - **Improve** the model
  - **Explore** the best as possible different input combinations
  - Identify the predominant inputs and phenomena in order to prioritize R&D

• **Validation phase:**
  - **Reduce** prediction uncertainties
  - **Calibrate** the model parameters

• **Practical use of a model:**
  - **Safety studies:** *assess* a *risk* of failure (rare events)
  - **Conception studies:** *optimize* system *performances* and *robustness*
Typical engineering practice: One-At-a-Time (OAT) design

Main remarks:
OAT brings some information, but potentially wrong
Exploration is poor: Non monotonicity? Discontinuity? Interaction?
Leave large unexplored zones of the domain (curse of dimensionality)
Illustration of the curse of dimensionality

7 points in a 3D space

Bad space covering

Surf. circle / Surf. square ~ 3/4
Surf. circle / Surf. square ~ 3/4
Vol. sphere / Vol. cube ~ 1/2

p=2

p=3

p=10

Ratio ~ 0.0025

Vol. sphere \( (r = 0.5) = \frac{\pi^{p/2}}{\Gamma \left( \frac{p}{2} + 1 \right)} \left( \frac{1}{2} \right)^p \)

Hypercube volume >> (included and tangent) hypersphere volume

For large dimensions, all the points will be in the corner of the hypercube
Model exploration goal

**GOAL**: explore as best as possible the behaviour of the code

Put some points in the whole input space in order to « maximize » the amount of information on the model output

Contrary to an uncertainty propagation step, it depends on $p$

Regular mesh with $n$ levels $\rightarrow N=n^p$ simulations

Ex: $p=2$, $n=3$ $\rightarrow N=9$
$\quad p=10$, $n=3$ $\rightarrow N=59049$

To minimize $N$, needs to have some techniques ensuring good « coverage » of the input space

Simple random sampling (Monte Carlo) does not ensure this

Ex: $p=2$, $N=10$
Exploration in physical experimentation

Design of experiments develops strategies to define experiments in order to obtain the required information as efficiently as possible.

**Designs for real experiments**

Estimate parameters of linear regression with a minimal number of points.

*Examples:*

- Full factorial design $2^3$
- Fractional factorial design $2^{3-1}$

**Designs for numerical experiments**

*Characteristics*

- Deterministic experiments (no error),
- Large number of input variables,
- Large range of input variation domain,
- Multiple output variables,
- Strong interactions between inputs,
- High non linearity in the model

*space filling designs (uniform coverage in the input space)*

**Bibliography**
- Fisher (1917), Box et Wilson (1954), Taguchi (1960), Mitchell (1958), ů
PLAN

- Part 1: Factorial designs
  - Full factorial design
  - Fractional factorial design
  - Resolution

- Part 2: Designs for numerical experiments
  - Properties
  - Low discrepancy sequence
  - Latin Hypercube sampling (LHS)
PART 1 : FACTORIAL DESIGNS

Full factorial design

- Hypothèses :
  - \( p \) inputs (« factors »)
  - 2 levels per factor

- Full factorial design
  - **Orthogonality principle**
    Variation of each factor when the others are are successively fixed at their 2 possible values
    \[ \Rightarrow 2^p \text{ experiments.} \]

- For continuous or discrete factors

- **Problem**: Number of experiments becomes too large when \( p \) or the number of levels increase.

Ex : \( p=10 \) factors at 2 levels \( \Rightarrow 1024 \) experiments

Factorial design at 2 levels for \( p = 3 \) factors
PART 1 : FACTORIAL DESIGNS

Exploitation of a factorial design
- realisation of the experiments
- exploitation => computation of coefficients $b_{ij}$ => construction of a simplified model

$$Y = \text{Cste} + \sum_i b_i X_i + \sum_{i < j} b_{ij} X_i X_j + \sum_{i < j < k} b_{ijk} X_i X_j X_k + \ldots$$

Response  Main effects  Interactions of order 2
PART 1: FACTORIAL DESIGNS

Fractional factorial design

- Study of all the factors with a reduced number of experiments

- Fraction of a full design
  \[ \Rightarrow 2^{p-q} \text{ experiments} \]

- Selection of this fraction?
  \[ \Rightarrow \text{Choice of an alias structure} \]
  \[ \Rightarrow \text{determine which effects are confused} \]

\[ Y = Cste + \sum_{i} b_i X_i + \sum_{i<j} b_{ij} X_i X_j + \sum_{i<j<k} b_{ijk} X_i X_j X_k + ... \]

Response  Main effects  Interactions of order 2

Full factorial design \(2^3\) decomposed in 2 fractional factorial designs \(2^{3-1}\) (black and white)
ILLUSTRATION

Fractional factorial designs

Well-balanced design
All the factors (and interactions) are produced at their low and high levels the same number of times.
The columns are orthogonals.
→ Good statistical properties
Fractional factorial design

Example of a fractional factorial design for $p = 5$ factors and $q = 2$

- Design with $2^{5-2} = 8$ experiments:
  - Full design at 3 factors for $(X_1, X_2, X_3)$
  - Effects of $X_4$ confused with interaction $X_1X_2$
  - Effects of $X_5$ confused with interaction $X_1X_3$
    $\Rightarrow$ 3 alias of 1
      - $X_1X_2X_4 = 1$
      - $X_1X_3X_5 = 1$
      - $X_2X_3X_4X_5 = 1$

- Resolution $R$:
  $R = \text{minimal number of elements of the alias of 1}$
  $= \text{cardinal of the smallest alias generator}$
  \[ \text{Example: } R = III \]

A design of resolution $R$ does not confound the effects of order $s_1$ and $s_2$ with $s_1 + s_2 < R$
### Resolution of a fractional factorial design

- **Résolution III**: all the main effects are not confused.

- **Resolution IV**: a main effect cannot be confused with an interaction, but 2 interactions can be confused.

- **Resolution V**: we can pose a model with the effects and interactions of order 2 without confusion.

#### Resolution V is considered as sufficient in most of the applications.
Resolution III is considered as a minimal property.

### Table: Resolution of a fractional factorial design

<table>
<thead>
<tr>
<th>nb of variables</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>R3</td>
<td>R4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>R4</td>
<td>R4</td>
<td>R4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>R4</td>
<td>R4</td>
<td>R4</td>
<td>R4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>R5</td>
<td>R4</td>
<td>R4</td>
<td>R4</td>
<td>R5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>R5</td>
<td>R4</td>
<td>R4</td>
<td>R4</td>
<td>R5</td>
<td>R5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R5</td>
<td>R5</td>
<td>R5</td>
</tr>
</tbody>
</table>
PLAN

Part 1: Factorial designs
- Full factorial design
- Fractional factorial design
- Resolution

Part 2: Designs for numerical experiments
- Properties
- Low discrepancy sequence
- Latin Hypercube sampling (LHS)
Exploration in physical experimentation

Design of experiments develops strategies to define experiments in order to obtain the required information as efficiently as possible.

**Designs for real experiments**

Estimate parameters of linear regression with a minimal number of points

Examples:

- Full factorial design $2^3$
- Fractional factorial design $2^{3-1}$

**Designs for numerical experiments**

**Characteristics**

- Deterministic experiments (no error),
- Large number of input variables,
- Large range of input variation domain,
- Multiple output variables,
- Strong interactions between inputs,
- High non linearity in the model

- space filling designs (uniform coverage in the input space)

Biblio: Fisher (1917), Box et Wilson (1954), Taguchi (1960), Mitchell (1958), ű

Objectives

When the objectives is to discover what happens inside the model and when no model computations have been realized, we want to respect the two following constraints:

- To spread the points over the input space in order to capture non-linearities of the model output,
- To ensure that this input space coverage is robust with respect to dimension reduction.

Therefore, we look for some design which insures the « best coverage » of the input space.

Main question:

- How to define this « best »?
Space filling designs

Sparsity of the space of the input variables in high dimension

The learning design choice is made in order to have an optimal coverage of the input domain.

The space filling designs are good candidates.

Example: Sobol sequence

Two possible criteria:
1. Distance criteria between the points: minimax, maximin, ...
2. Uniformity criteria of the design (discrepancy measures)
Geometrical criteria (1/2)

- Minimax design $D_{MI}$: Minimize the maximal distance between one point of the domain and one point of the design

$$\min_D \max_x d(x, D) = \max_x d(x, D_{MI})$$

where $d(x, D) = \min_{x^{(0)} \in D} d(x, x^{(0)})$

All points in $[0,1]^p$ are not too far from a design point

$[\text{Johnson et al. 1990}]$
$[\text{Koehler & Owen 1996}]$

$=>$ One of the best design, but too expensive to find $D_{MI}$
Minimax design

\[ \delta p = 1 ; \ X_i = (2i-1)/(2N) ; \ \phi_{mM} = 1 / 2N \]

\[ \delta p > 1 : \text{sphere recovering} \]
Geometrical criteria (2/2)

- Mindist distance: \( \phi(\Xi^N) = \min_{x^{(1)}, x^{(2)} \in \Xi^N} d(x^{(1)}, x^{(2)}) \) (L₂ norm for example)

Maximin design \( \Xi^N_{\text{Mm}} \):
maximize minimal distance between two points of the design

\[
\max_{\Xi^N} \min_{x^{(1)}, x^{(2)} \in \Xi^N} d(x^{(1)}, x^{(2)}) = \min_{x^{(1)}, x^{(2)} \in \Xi^N_{\text{Mm}}} d(x^{(1)}, x^{(2)})
\]
Maximin design

\[ \delta p = 1 \;; \; X_i = (i-1)/(N-1) \;; \; \phi_{mm} = 1 / (N-1) \]

\[ \delta p > 1 : \text{sphere packing} \]

\[ L^2\text{-LHD of 9 points} \]
\[ d = 0.395284707521 \text{ and } D^2 = 10 \]

[ www.spacefillingdesigns.nl ]

[ www.packomania.com ]
Space filling measure of a design: the discrepancy

Measure of the maximal deviation between the distribution of the sample’s points to an uniform distribution

⇒ Measure of deviation from the uniformity

Geometrical interpretation:
Comparison between the volume of intervals and the number points within these intervals

\[ Q(t) \in [0,1]^p, Q(t) = [0,t_1] \times [0,t_2] \times \ldots \times [0,t_p] \]

\[ \text{disc}(D) = \sup_{Q(t) \in [0,1]^p} \left| \frac{N_{Q(t)}}{N} - \prod_{i=1}^{p} t_i \right| \]

Lower the discrepancy is, the more the points of the design \( D \) fill the all space
Link with the integration problem

\[ I = \int_{[0,1]^p} f(x) \, dx \]

Monte Carlo: \( I_N^{MC} = \frac{1}{N} \sum_{i=1}^{N} f(x^{(i)}) \)

with \((x^{(i)})_{i=1...N}\) a sequence of random points in \([0,1]^p\)

\[ E(I_N^{MC}) = I \quad \text{Var}(I_N^{MC}) = \frac{\text{Var}(N)}{N} \quad \Rightarrow \varepsilon = O \left( \frac{1}{\sqrt{N}} \right) \]

General property (Koksma-Hlawka inequality): \( \varepsilon \leq V(f) \times \text{disc}(D) \)

With a low discrepancy sequence \( D \) (quasi Monte Carlo sequence):

Well-known choice: Sobol' sequence

\[ \varepsilon = O \left( \frac{(\ln N)^p}{N} \right) \]
**L₂ discrepancy**

Several definitions, depending on considered norms and intervals

\[
D^*(\Xi^N) = \sup_{t \in [0,1]^p} \left| \frac{1}{N} \sum_{i=1}^{N} x^{(i)}_{t} - \text{Volume}(Q(t)) \right|
\]

Choice allowing computations : L₂ discrepancy

\[
D^*_2(\Xi^N) = \left[ \int_{[0,1]^p} \left( \frac{1}{N} \sum_{i=1}^{N} x^{(i)}_{t} - \text{Volume}(Q(t)) \right)^2 dt \right]^{1/2}
\]

Missing property: taking into account uniformity of the point projections

On lower-dimensional subspaces of [0,1]^p

=> Modified L₂ discrepancies

\[
D_2(\Xi^N) = \left[ \sum_{u \neq \emptyset} \int_{C^u} \left( \frac{1}{N} \sum_{i=1}^{N} x^{(i)}_{t} - \text{Volume}(Q_u(t)) \right)^2 dt \right]
\]

with \( u \subset \{1,\ldots,p\} \)

and \( Q_u(t) = \text{projection of } Q(t) \text{ on } C^u \) (unit cube of coordinates in \( u \))
Discrepancy computation in practice

- **Modified $L_2$-discrepancy** (intervals with minimal boundary 0)

- **Centered $L_2$-discrepancy** (intervals with boundary one vertex of the unit cube)

\[
\text{disc}_2(D) = \left( \frac{13}{12} \right)^p - \frac{2}{N} \sum_{i=1}^{N} \prod_{k=1}^{p} \left( 1 + \frac{1}{2} \left| x_k^{(i)} - \frac{1}{2} \right| - \frac{1}{2} \left| x_k^{(i)} - \frac{1}{2} \right|^2 \right) \\
+ \frac{1}{N^2} \sum_{i,j=1}^{N} \prod_{k=1}^{p} \left( 1 + \frac{1}{2} \left| x_k^{(i)} - \frac{1}{2} \right| + \frac{1}{2} \left| x_k^{(j)} - \frac{1}{2} \right| - \frac{1}{2} \left| x_k^{(i)} - x_k^{(j)} \right| \right)
\]

- **Symetric $L_2$-discrepancy** (intervals with boundary one « even » vertex of the unit cube)
Sobol sequence vs. Random sample vs. regular grid

[From: Kucherenko, 2010]
Example - $N = 150$ - Dimension = 8

- Sobol
- Sobol scrambling Owen
**Example - N = 150 - Dimension = 8**

**Halton**

![Graphs showing random sampling points for Halton sequences in two dimensions.](image-url)
Pathologies on 2D projections

Halton
Tests: low-discrepancy sequences vs. Monte Carlo

For integration: convergence often close to $1/N$ for $p < 40$.

Example: Ishigami function, dimension $p = 3$
Important property: robustness in terms of subprojections

Most of the times, the function $f(X)$ has low effective dimensions:
- in the truncation sense ($p_1 = \text{number of influent inputs} \Rightarrow p_1 \ll p$)
- in the superposition sense ($p_2 = \text{higher order of influent interaction} \Rightarrow p_2 \ll p$)

Then, we need SFD which keeps their space-filling properties in low-dimensional subspaces (by importance: in dimensions $p' = 1$, then $p' = 2$, ...)

- $p' = 1 \Rightarrow \text{LHS ensures good 1D projection properties}$
- $p' \geq 2$
  In their definition, the modified $L^2$-discrepancy criteria take into account subprojections

In contrary design points distance criteria are not robust at all
Latin Hypercube Sample (LHS)

Most often, only a small number of variables are influent.

**Property:** Uniform projections on margins

**Principle:** \( p \) variables, \( N \) points \( \Rightarrow \) LHS\((p,N)\)

Divide each dimension in \( N \) intervals
Take one point in each stratum

Each level is taken only one time by each variable
\( \Rightarrow \) Each column of the design is a permutation of \{1, 2, ..., \( N \)\}

[McKay et al. 1979]
Algorithm of LHS($p, N$) - Stein method

```
ran = matrix(runif(N*p), nrow=N, ncol=p)  # tirage de N x p valeurs selon loi U[0,1]
x = matrix(0, nrow=N, ncol=p)            # construction de la matrice x

for (i in 1:p) {
    idx = sample(1:N)  # vecteur de permutations des entiers {1, 2, ..., N}
    P = (idx - ran[,i]) / N  # vecteur de probabilités
    x[,i] <- quantile_selon_la_loi (P)  }
```

Example: $p = 2$, $N = 10$, $X_1 \sim U[0,1]$, $X_2 \sim N(0,1)$

(a) Simple Random Sampling

(b) Latin Hypercube Sampling
Optimisation of LHS \(\Rightarrow\) Space-filling LHS

Simple method: produce a large number (for ex 1000) of different LHS. Then, choose the best with respect to a criterion \(\phi(.)\) (« space filling »)

Example: LHS(2,16)

Maximin criterion

\[
\text{BUT: the number of LHS is huge : } (N!)^p
\]

Methods via optimization algo (ex: minimisation of \(\phi(.)\) via simulated annealing):

1. Initialisation of a design \(\Xi\) (LHS initial) and a temperature \(T\)

2. While \(T > 0\):
   1. Produce a neighbor \(\Xi_{\text{new}}\) of \(\Xi\) (permutation of 2 components in a column)
   2. Replace \(\Xi\) by \(\Xi_{\text{new}}\) with proba \(\min\left(\exp\left(-\frac{\phi(\Xi_{\text{new}}) - \phi(\Xi)}{T}\right), 1\right)\)
   3. Decrease \(T\)

3. Stop criterion \(\Rightarrow \Xi\) is the optimal solution
Examples of optimized LHS
Joining the two properties (space filling and LHS)

Example: $p = 2 - N = 16$

Maximin LHS  Low wrap-around discrepancy LHS  For comparison: Sobol sequence
Summary on the design of numerical experiments

**Goal:** Sample a high dimensional space in an « optimal » manner (obtain the maximum of information on the behaviour of the output $Z / X \in \mathbb{R}^p$)

**Problem:** a pure random sample (Monte Carlo) badly fills the space

1. « Space filling » designs are good candidates:

   - Based on a distance criterion between points (minimax, maximin, ...)
   - Based on a citerion of uniform distribution of the points *(discrepancy)*

2. Property of uniform projections on margins can be obtained via the *Latin hypercube designs* (LHS)

3. It is possible to couple 1 and 2
## Synthesis on the properties of space filling designs

<table>
<thead>
<tr>
<th></th>
<th>Patterns, alignment</th>
<th>Sequentiality</th>
<th>Dimension reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Monte Carlo</strong></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Faure, Halton, Sobol</strong></td>
<td>Yes, in high dim.</td>
<td>Yes</td>
<td>Yes, but pathologies</td>
</tr>
<tr>
<td><strong>LHS à discrép centrée faible</strong></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>LHS maximin</strong></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Bibliography

BOOKS
- Fang et al., Design and modeling for computer experiments, Chapman & Hall, 2006
- Lemieux, Monte Carlo and Quasi-Monte Carlo Sampling, Springer-Verlag, 2009
- Saltelli, Chan & Scott, Sensitivity analysis, Wiley, 2000
- Saltelli et al., Global sensitivity analysis - The primer. Wiley, 2008

PAPERS
- Damblin, Couplet & Iooss, Numerical studies of space filling designs: optimization algorithms and subprojection properties, Journal of Simulation, 7, 2013
- McKay, Beckman, and Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21, 1979