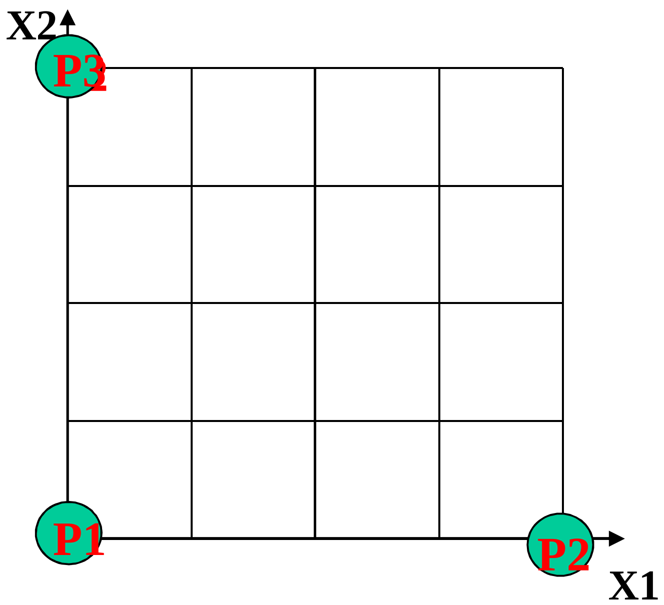
Design of computer experiments

Fabrice Gamboa
Bertrand Iooss

28/02/2013

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)

Model exploration goal

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

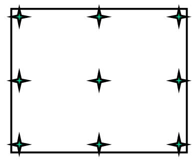
Put some points in the whole input space in order to « maximize » the amount of information on the model output

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

Regular mesh with n levels → $N = n^p$ simulations

Ex: p =2, n =3 → N =9

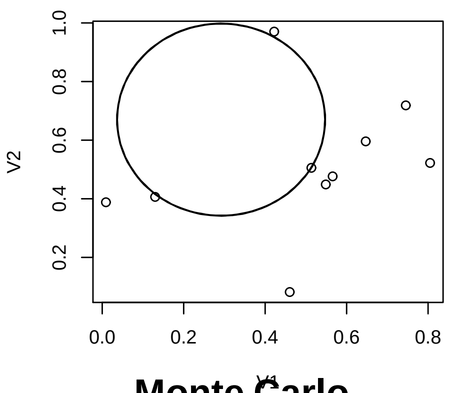
p = 10, n=3 → N = 59049



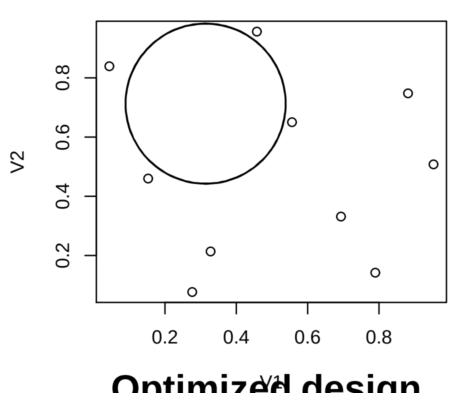
To minimize N, needs to have some techniques ensuring good « coverage » of the input space

Simple random sampling (Monte Carlo) does not ensure this

Ex: p = 2
N = 10



Monte Carlo



Optimized design

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 some design which insures the « best coverage » of the input space

<u>Main question:</u>

- How to define this « best » ?

Exploration in physical experimentation

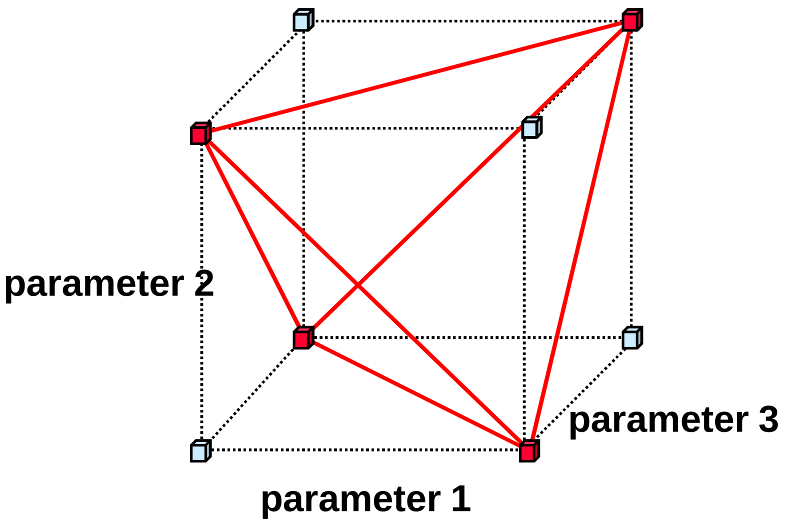
Design of experiments develops strategies to define experiments in order to obtain the required information as efficiently as possible

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)

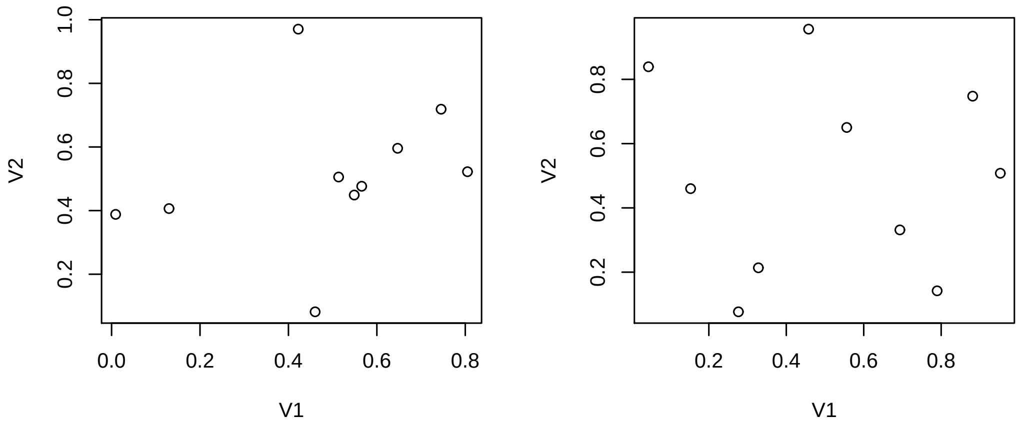
Space filling designs

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.

Simple Random Sample (SRS)



Space Filling Design (SFD)

<u>Example:</u> Sobol sequence

<u>Two possible criteria:</u>

1. Distance criteria between the points: minimax, maximin, …
2. Uniformity criteria of the design (discrepancy measures)

Geometrical criteria (1/2)

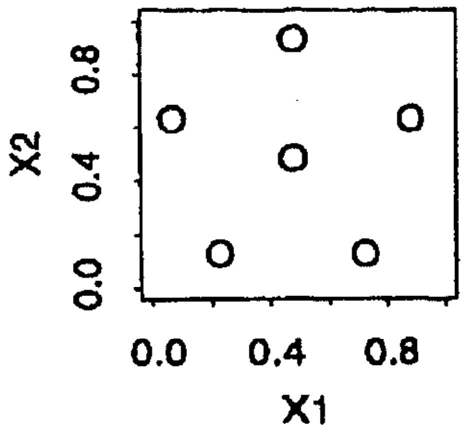
• Minimax design D_{MI} : Minimize the maximal distance between one point of the domain and one point of the design

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

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

[Johnson et al. 1990]
[Koehler & Owen 1996]

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



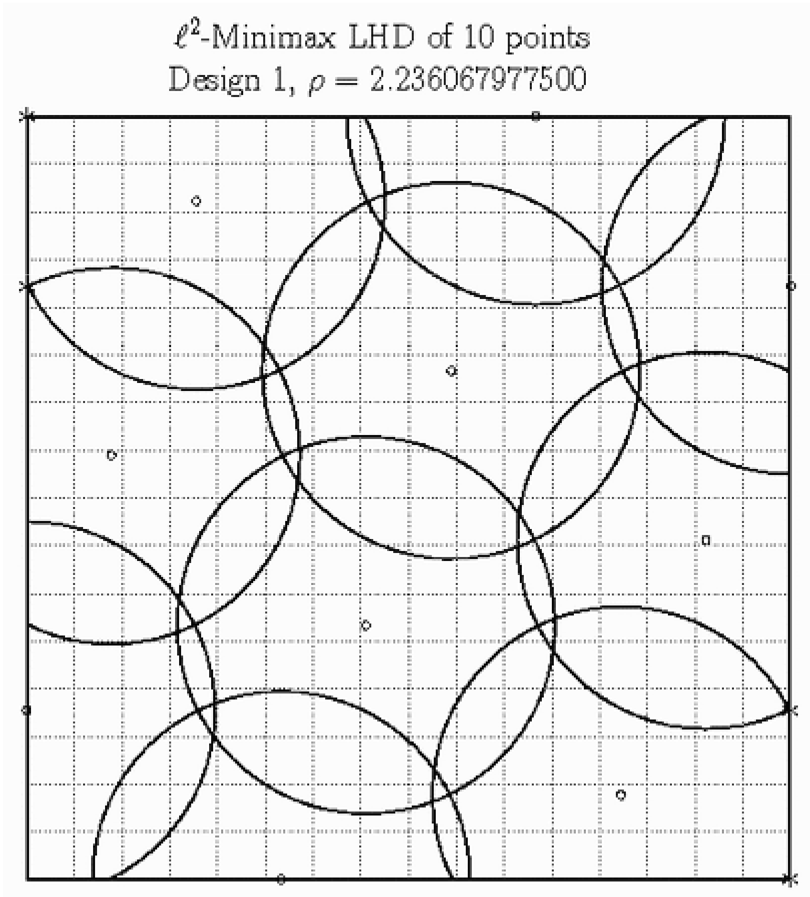
(a) Minimax

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

Minimax design

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

- $p > 1$: sphere recovering



[www.spacefillingdesigns.nl]

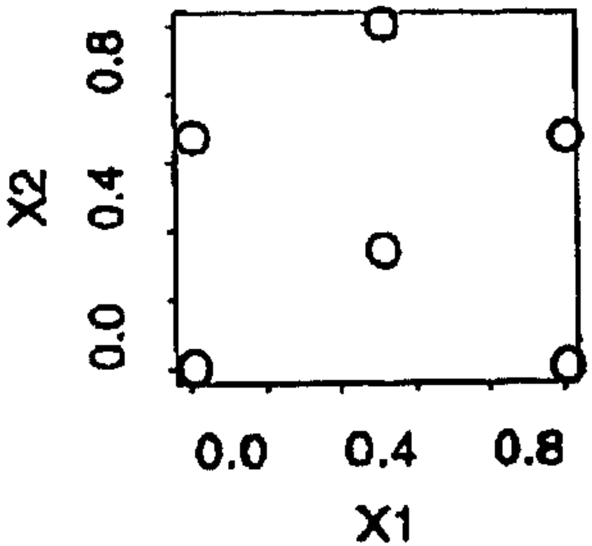
Geometrical criteria (2/2)

- **Mindist distance:** $\phi(\Xi^N) = \min_{x^{(1)},x^{(2)} \in \Xi^N} d(x^{(1)}, x^{(2)})$ (L_2 norm for example)

⟹ Maximin design Ξ^N_{Mm} :

maximize minimal distance between two points of the design

$$\max_{\Xi^N} \min_{x^{(1)},x^{(2)} \in \Xi^N} d(x^{(1)}, x^{(2)}) = \min_{x^{(1)},x^{(2)} \in \Xi^N_{\mathrm{Mm}}} d(x^{(1)}, x^{(2)})$$

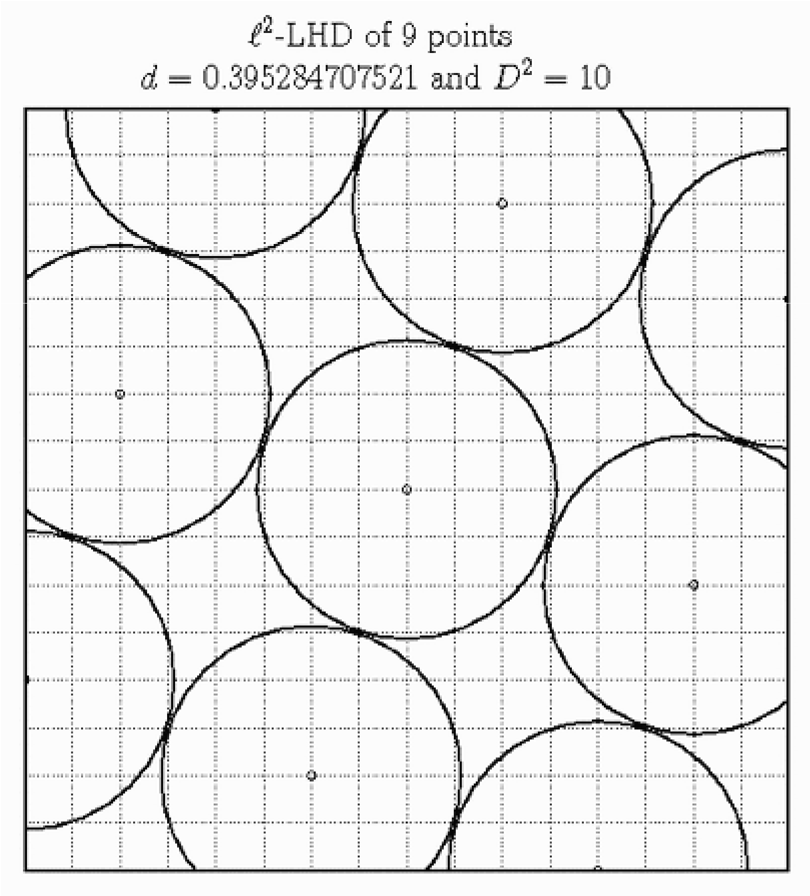


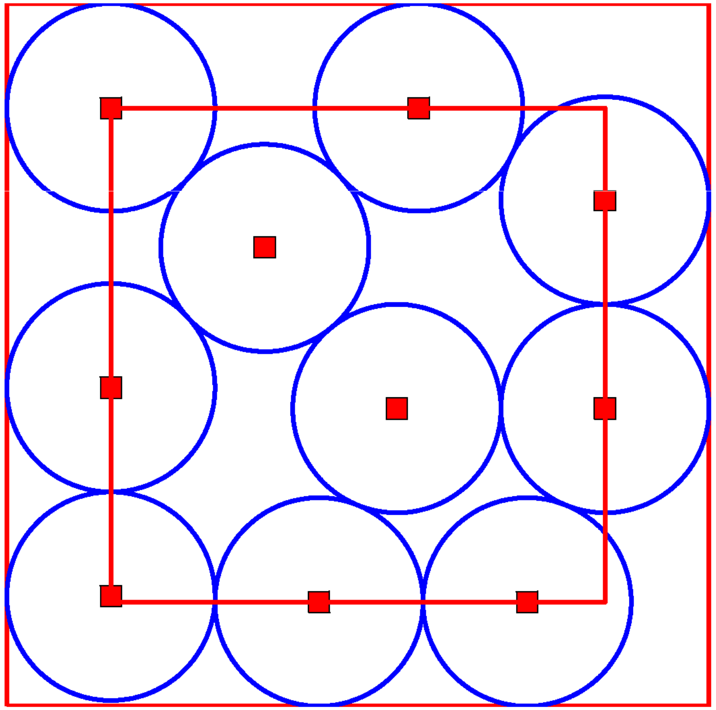
- …

Maximin design

- $p = 1$; $X_i = (i-1)/(N-1)$; $\phi_{mM} = 1 / (N-1)$

- $p > 1$: sphere packing



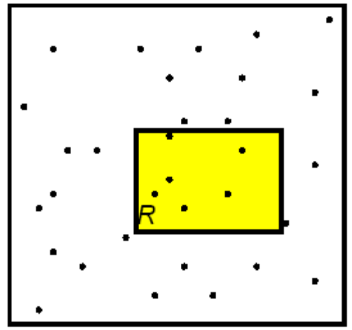


[www.spacefillingdesigns.nl] *[www.packomania.com]*

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

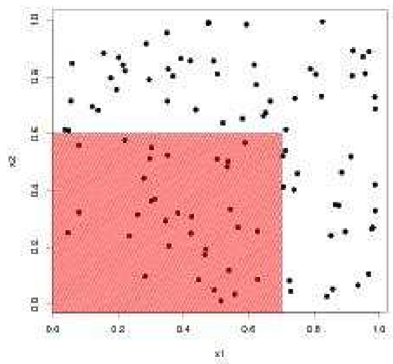


<u>Geometrical interpretation:</u>

Comparison between the volume of intervals and the number points within these intervals

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

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



Lower the discrepancy is, the more the points of the design D fill the all space

Link with the integration problem

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

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

$$\text{with } \left(x^{(i)}\right)_{i=1\ldots N} \text{ a sequence of random points in } [0,1[^p$$

$$\mathrm{E}\left(I_N^{\text{MC}}\right) = I \;;\; \mathrm{Var}\left(I_N^{\text{MC}}\right) = \frac{\mathrm{Var}(N)}{N} \Rightarrow \varepsilon = O\left(\frac{1}{\sqrt{N}}\right)$$

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

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

$$\varepsilon = O\left(\frac{(\ln N)^p}{N}\right)$$

Well-known choice: Sobol' sequence

L_2 discrepancy

Several definitions, depending on considered norms and intervals

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

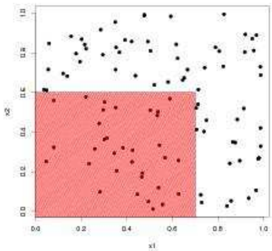
Choice allowing computations : L^2 discrepancy ***[Hickernell 1998]***

L^2 discrepancy at origin : $$D_2^*\left(\Xi^N\right) = \left[\int_{[0,1[^p} \left[\frac{1}{N} \sum_{i=1}^{N} 1_{\mathbf{x}^{(i)}\in Q(\mathbf{t})} - \mathrm{Volume}(Q(\mathbf{t})) \right]^2 d\mathbf{t} \right]^{1/2}$$

<u>Missing property:</u> taking into account uniformity of the point projections On lower-dimensional subspaces of [0,1[^{p}

=> Modified L_2 discrepancies

$$D_2\left(\Xi^N\right) = \left[\sum_{u\neq\emptyset} \int_{C^u} \left[\frac{1}{N} \sum_{i=1}^{N} 1_{\mathbf{x}_u^{(i)}\in Q_u(\mathbf{t})} - \mathrm{Volume}(Q_u(\mathbf{t})) \right]^2 d\mathbf{t} \right]$$

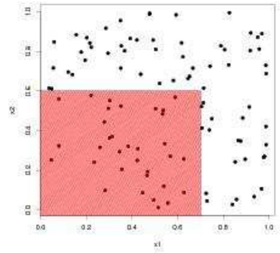


with $u \subset \{1,\ldots,p\}$

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

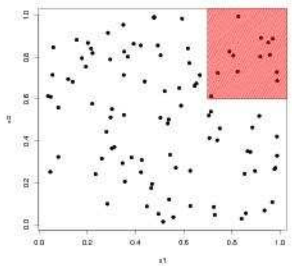
Discrepancy computation in practice

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

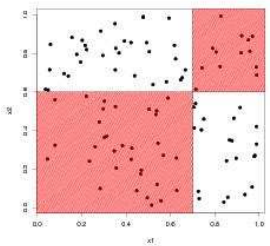


- **Centered L_2-discrepancy** (intervals with boundary one vertex of the unit cube)

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

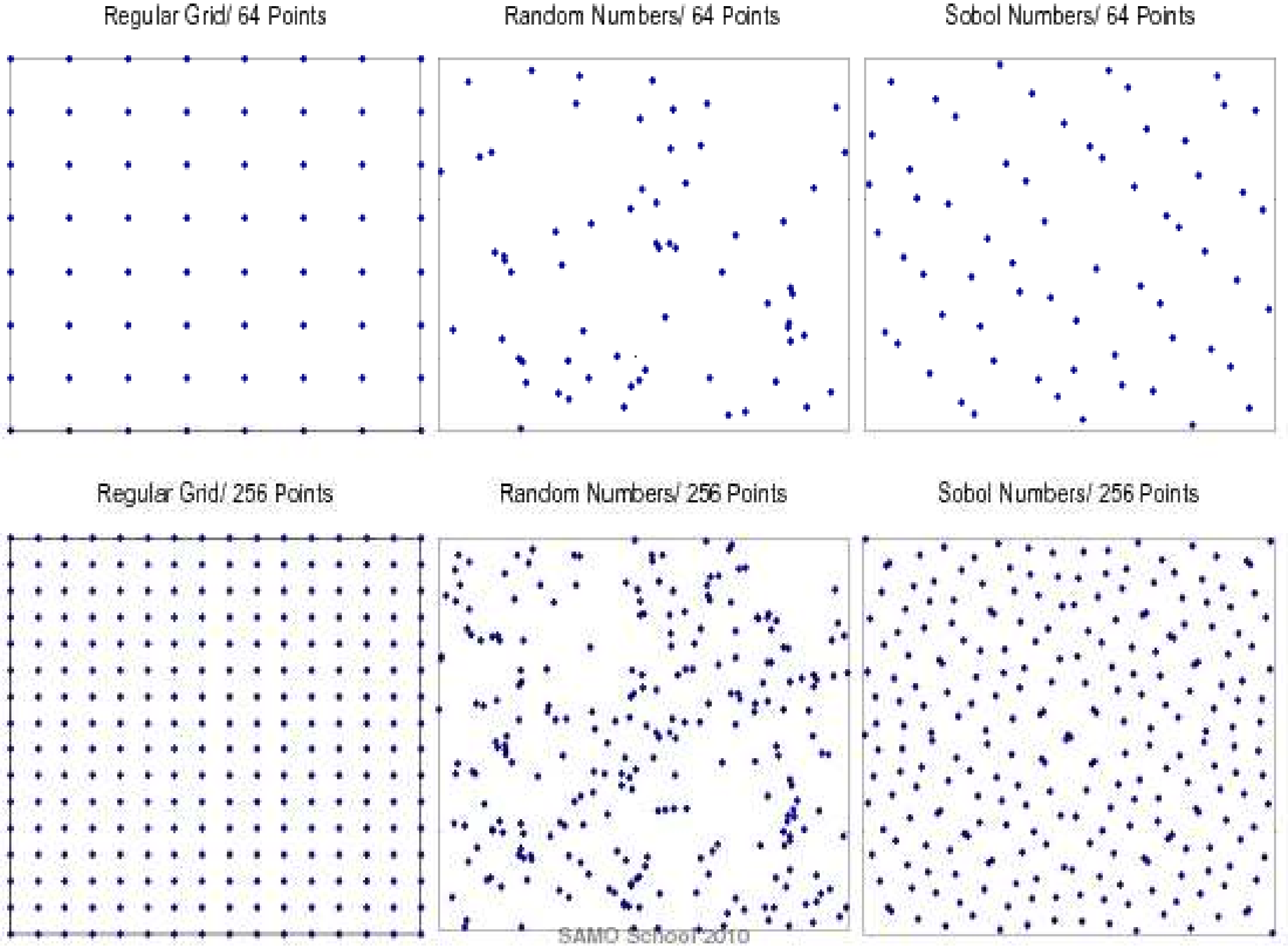


- **Symetric L_2-discrepancy** (intervals with boundary one « even » vertex of the unit cube)



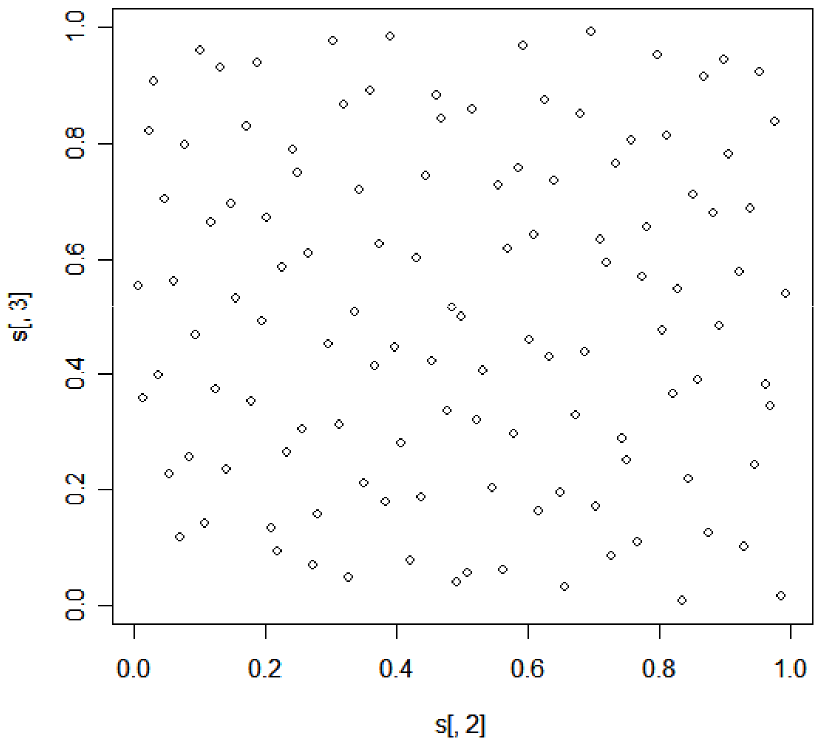
Sobol'sequence vs. Random sample vs. regular grid

[From: Kucherenko, 2010]

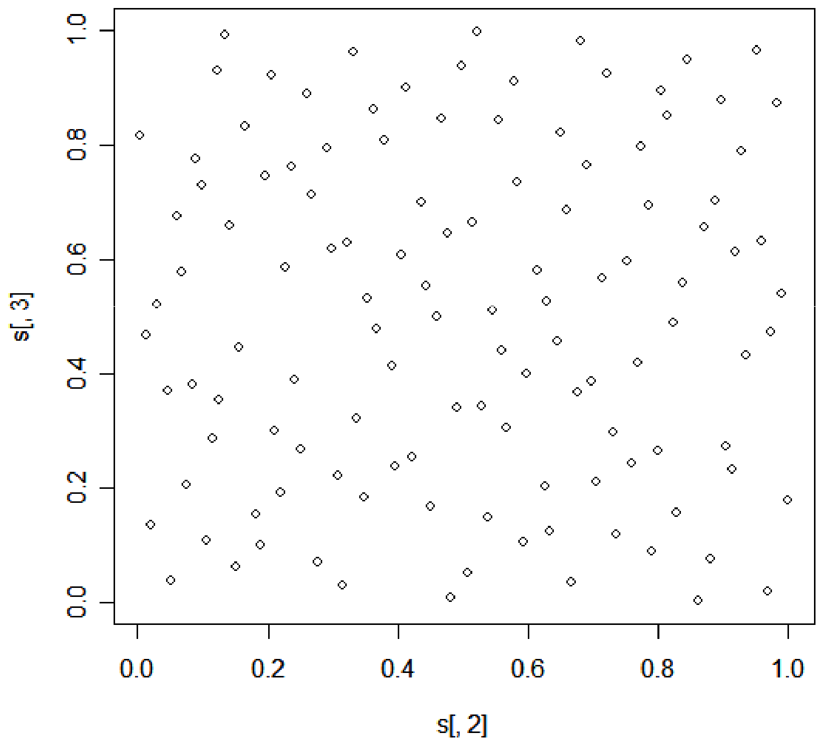


Example - N = 150 - Dimension = 8

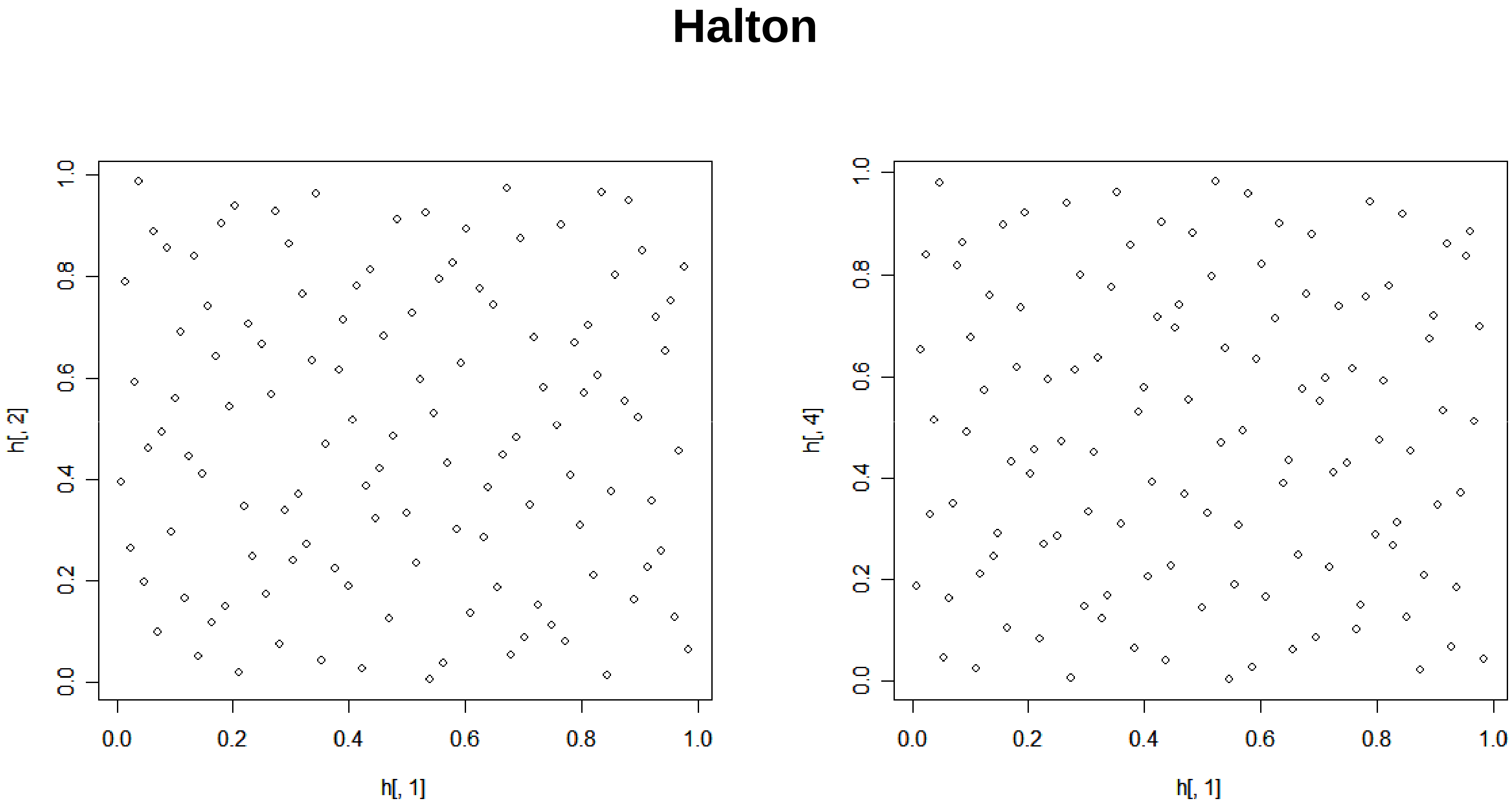
Sobol



Sobol scrambling Owen

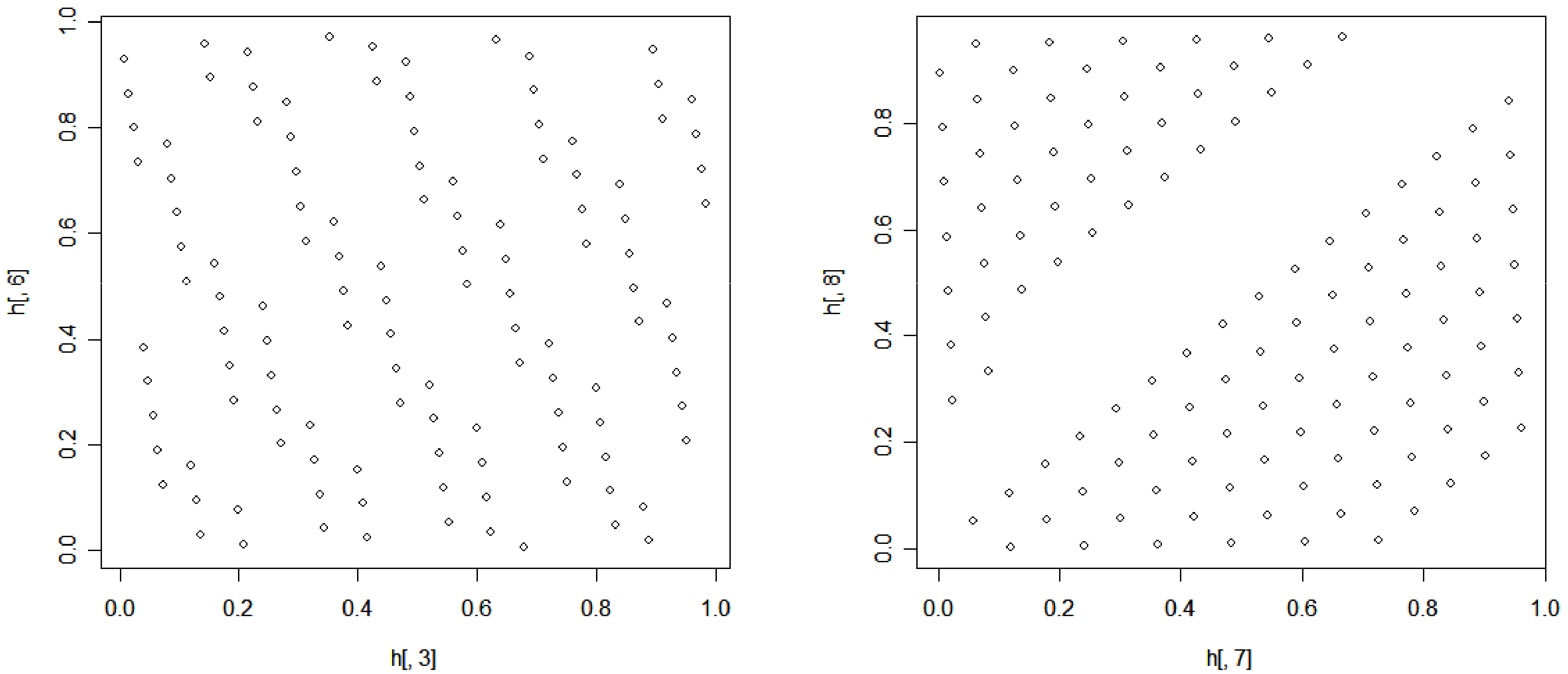


Example - N = 150 - Dimension = 8



Pathologies on 2D projections

Halton

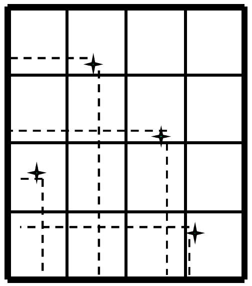


Important property: robustness in terms of subprojections

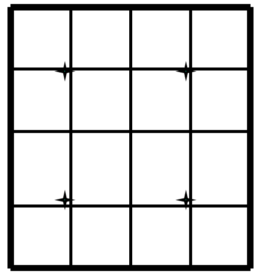
Most of the times, the function f($\mathbf{X}$) has low <u>effective dimensions:</u>
- in the truncation sense (p_1 = number of influent inputs) $\Rightarrow$ $\boldsymbol{p_1 \ll p}$
- in the superposition sense (p_2 = higher order of influent interaction) $\Rightarrow$ $\boldsymbol{p_2 \ll p}$

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 $\Rightarrow$ LHS ensures good 1D projection properties



good



bad

- $p' \geq 2$

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)

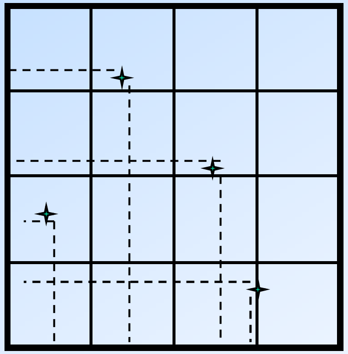
[McKay et al. 1979]

Most often, only a small number of variables are influent

Property: Uniform projections on margins

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

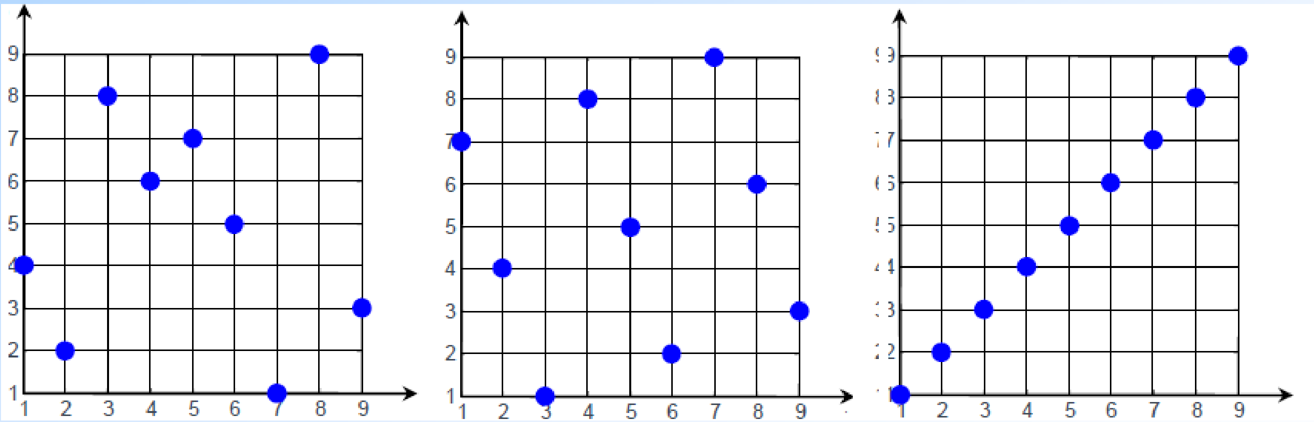
Divide each dimension in N intervals
Take one point in each stratum



Exemple : $p=2$, $N=4$

Each level is taken only one time by each variable

$\Rightarrow$ **Each column of the design is a permutation of $\{1,2,..,N\}$**

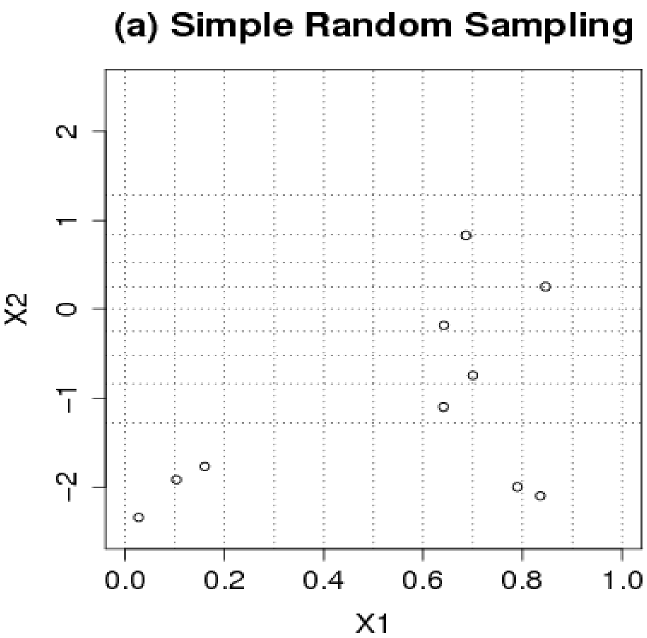


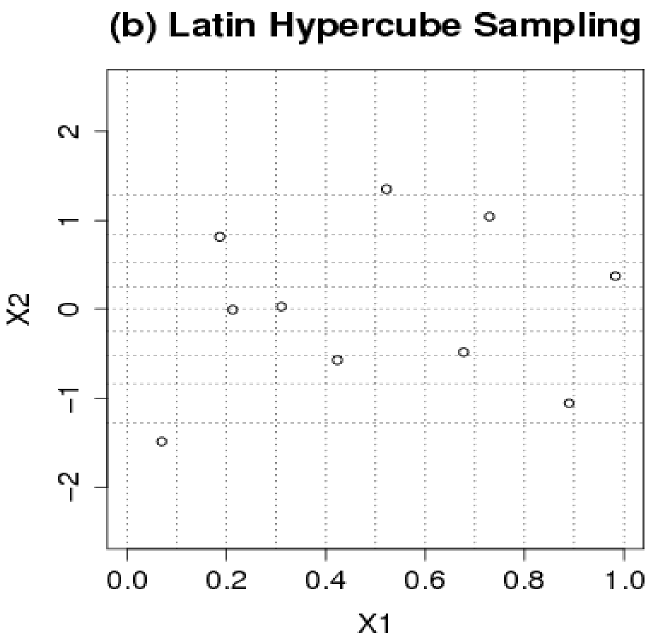
Algorithm of LHS(p,N) – Stein method

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





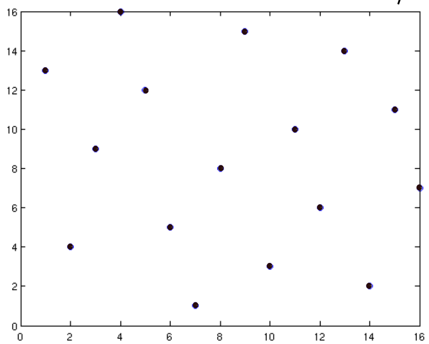
Optimisation of LHS => Space-filling LHS

[Park 1993; Morris & Mitchell 1995]

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

Example : LHS(2,16)

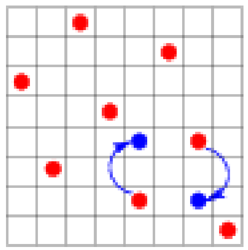
Maximin criterion



BUT: the number of LHS is huge : $(N!)^p$

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

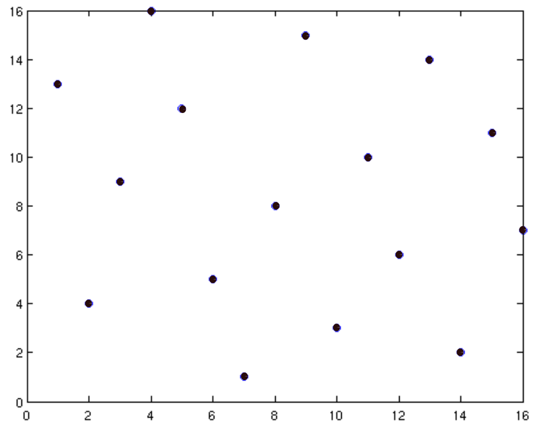
1. Initialisation of a design Ξ (LHS initial) and a temperature T
2. While $T > 0$:
   1. Produce a neighbor Ξ_{new} of Ξ (permutation of 2 components in a column)
   2. replace Ξ by Ξ_{new} with proba $\min\left(\exp\left[-\frac{\phi(\Xi_{new})-\phi(\Xi)}{T}\right],1\right)$
   3. decrease T
3. Stop criterion => Ξ is the optimal solution



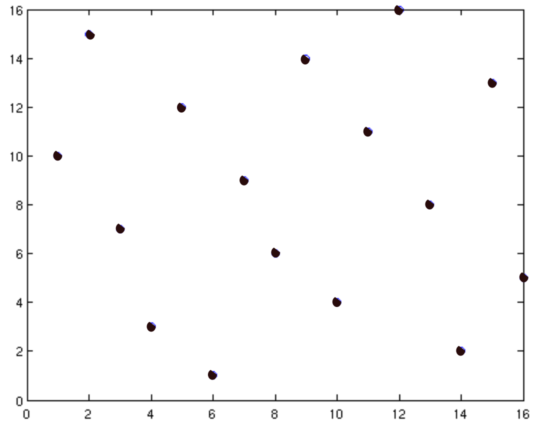
Examples of optimized LHS

Joining the two properties (space filling and LHS)

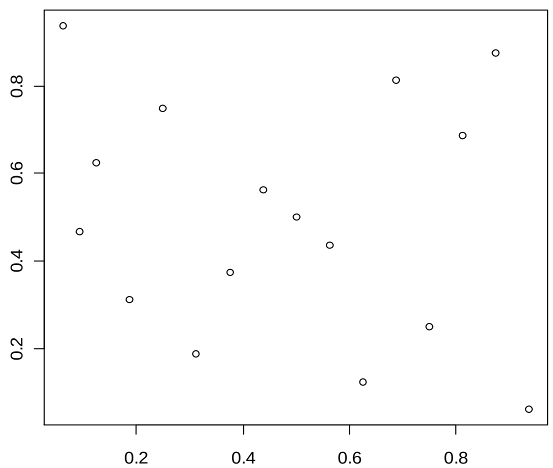
Example: $p = 2$ - $N = 16$



Maximin LHS



Low wrap-around discrepancy LHS



For comparison: Sobol sequence

Summary on the design of numerical experiments

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

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

Bibliography

– Fang et al., *Design and modeling for computer experiments*, Chapman & Hall, 2006

– J.C. Helton, J.D. Johnson, C.J. Salaberryet C.B. Storlie: Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliability Engineering and System Safety, 91:1175–1209, 2006.

– Kleijnen, *The design and analysis of simulation experiments*, Springer, 2008

– Koehler & Owen, Computer experiments, 1996

– A. Saltelli, K. Chan & E.M. Scott, *Sensitivity analysis,* Wiley, 2000

– A. Saltelli et al., Global sensitivity analysis - The primer. Wiley, 2008.