A radial scanning statistic for selecting space-filling designs in computer experiments

O. Roustant[(1)], J. Franco[(2)], L. Carraro[(3)], A. Jourdan[(4)]

(1) Ecole des Mines de St-Etienne – (2) Total – (3) Telecom Saint-Etienne – (4) EISTI Pau

MODA-9, Bertinoro, Italy

Ecole Nationale Supérieure des Mines SAINT-ETIENNE

Framework

- Scientific framework
  - Keywords : Computer experiments, space-filling designs, dimension reduction, goodness-of-fit

- DICE Consortium (Sept. 2006 – Dec. 2009)
  - 5 industrial companies (TOTAL, Renault, IRSN, EDF, Onera : energy, automotive, nuclear and aerospatial engineering)
  - 4 academic partners (Ecole des Mines de St-Etienne, Univ. Aix-Marseille, Univ. Joseph Fourier, Univ. Paris 11)

  - Objective : to study costly simulators

  - Web site : http://www.dice-consortium.fr/

Supplementary material

➔ *R package "DiceDesign", available at http://cran.r-project.org/*

Outline

- Motivations
- The radial scanning statistic
- Applications
  - defect detection of SFDs
  - selection of SFDs
- Future research

Introduction - Framework

- Framework
  - **First investigation** of a costly **deterministic** simulator

$$y = f_{sim}(x_1, x_2, \ldots, x_d)$$

  - Cubic region $x=(x_1,\ldots,x_d) \in \Omega=[-1,1]^d$, d=1, 2, …,10, … 20,…50,…

- Objective
  - To study the phenomenon modeled by f_{sim} with few runs

Introduction - Assumptions

- Assumptions
  - ***Complexity of the phenomenon***

    ⇒ Non-linearities of f_{sim}
  - ***The effective dimension is << d***

    ⇒ Only few factors are influent (sparsity)

$$f_{sim}(x) = g(x_{i1}, \ldots, x_{ik}), \quad k << d$$

⇒ Only few principal components are influent

$$f_{sim}(x) = g(b_1'x, \ldots, b_k'x), \quad k << d$$

- Remark
  - The latter is standard for dimension reduction (as for Sliced Inverse Regression, see [Li, 1991])

Introduction - Consequences

- Consequences for designs
  - ***Complexity of the phenomenon***
    - ⇒ Non-linearities of f_{sim} ⇒ space-filling designs
  - ***The effective dimension is << d***
    - ⇒ Only few factors are influent (sparsity)

$$f_{sim}(x) = g(x_{i1}, \ldots, x_{ik}), \quad k << d$$

⇒ space-filling of projections onto factorial subspaces

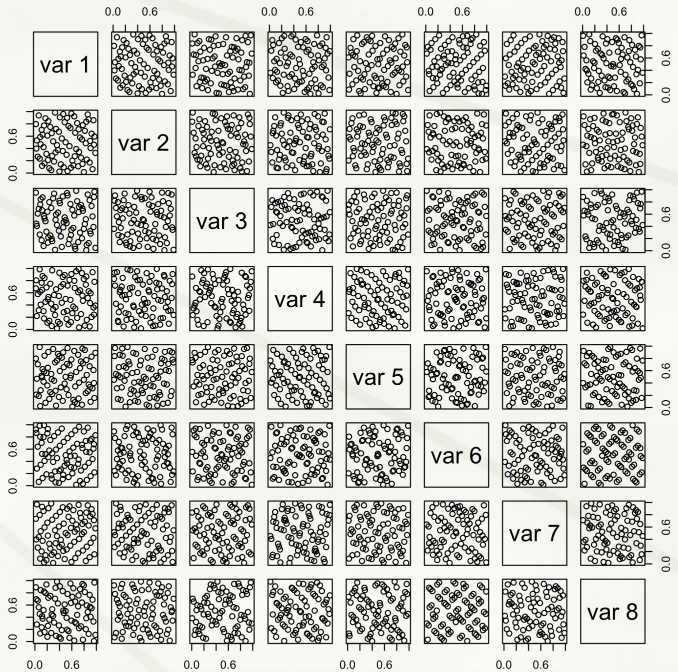
- ⇒ Only few principal components are influent

$$f_{sim}(x) = g(b_1'x, \ldots, b_k'x), \quad k << d$$

⇒ space-filling of projections onto oblique subspaces

Introductory example

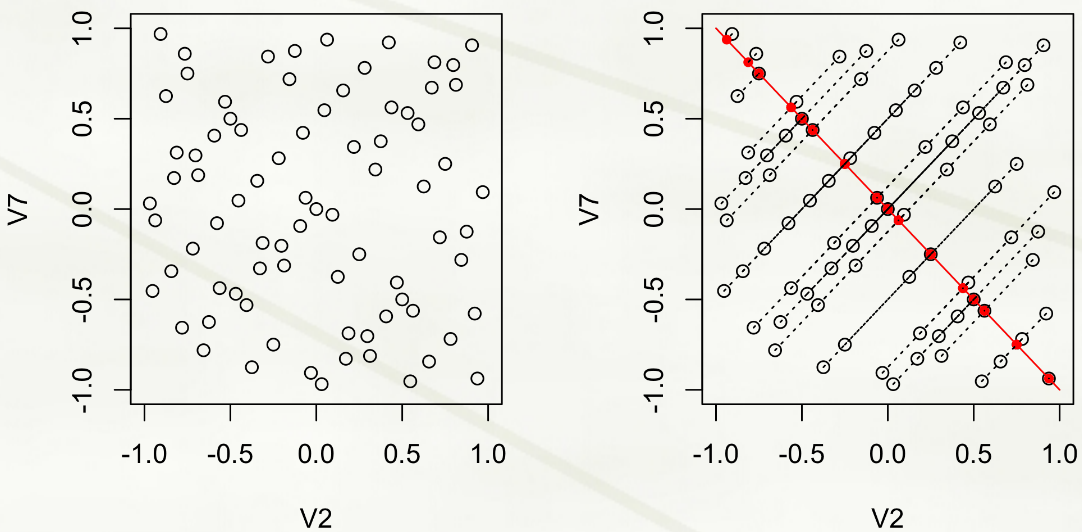
- A 8D example
  - $f_{sim}(x) = g(x_2, x_7)$, d=8
  - Common approach in practice : use a space-filling design (SFD), for instance a 80 points Sobol sequence



Introductory example

- A 8D example
  - $f_{sim}(x) = g(x_2, x_7)$, d=8
  - Space-filling design : 80 points Sobol sequence

⇒ Space-filling is not so good in the subspace Vect(x_2, x_7)
⇒ If the code is a function of $x_2 – x_7$, only 16 different points !
loss of information



Introductory example

OBJECTIVE

⇒ To detect automatically such defects

⇒ To select space-filling designs that are still space-filling in projection onto oblique subspaces

⇒ *more constraining than for LHDs or OAs*

⇒ *to be precised in next slide*

Objective

- Ideal objective:
  - There is a need to check good properties of projections onto any subspace spanned by b_1'x, …, b_k'x

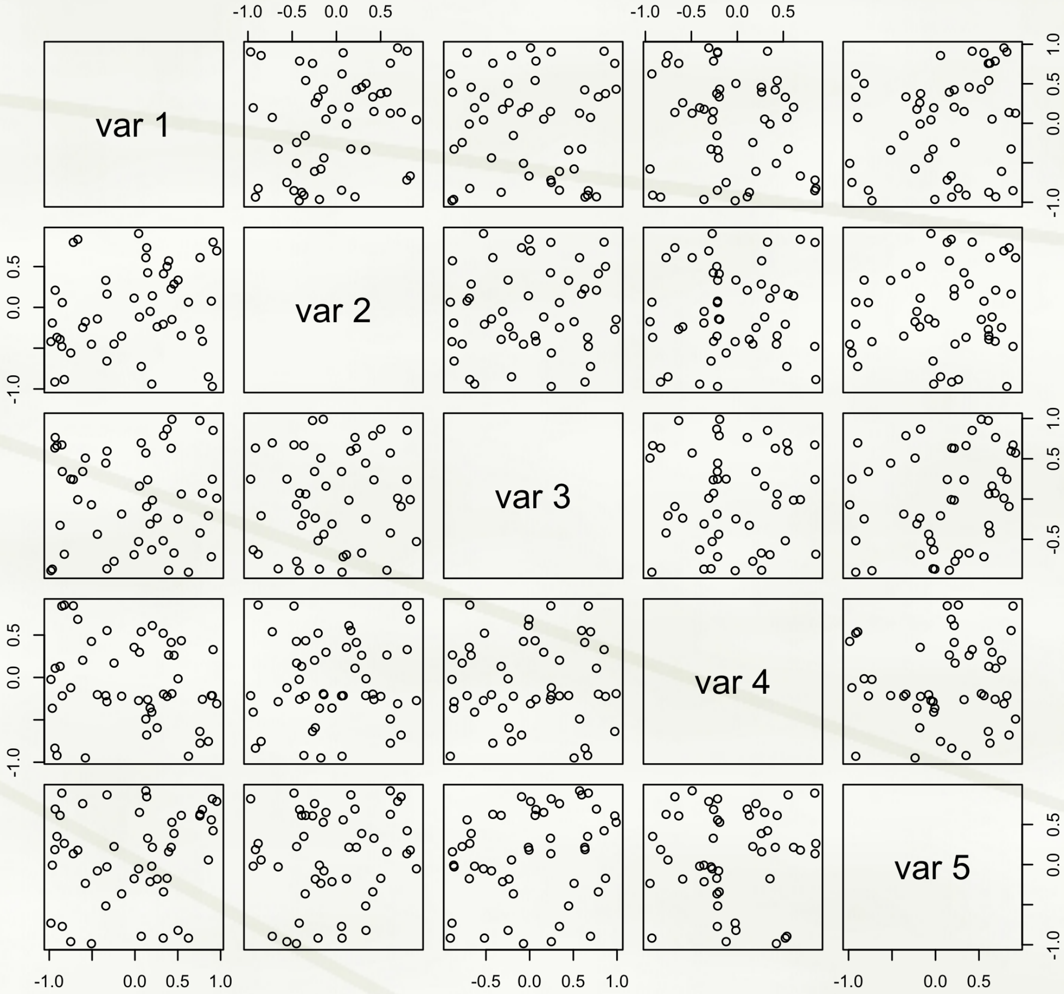
- Revised ideal objective:
  - To check space-filling properties of the projections onto any (oblique) 1-dimensional axis [spanned by one b'x]

- Realization:
  - To check space-filling properties of the projections onto any 1-dimensional axis of the form :
    - $\beta_i x_i + \beta_j x_j$ 2D radial scanning statistic
    - $\beta_i x_i + \beta_j x_j + \beta_k x_k$ 3D radial scanning statistic

A good benchmark

- Uniform designs
  - Advantage: Space-filling onto projections
  - Main defaults : clusters, holes

A good SFD ?
advantages without the drawbacks of uniform designs, … /…

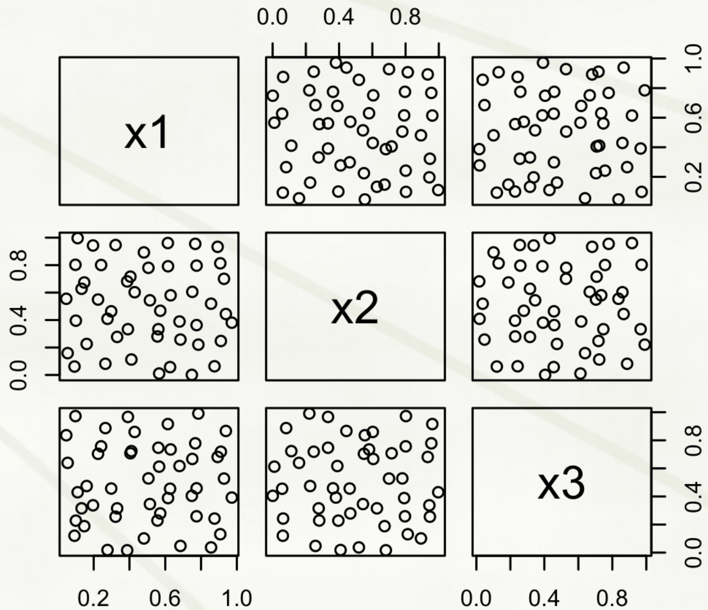


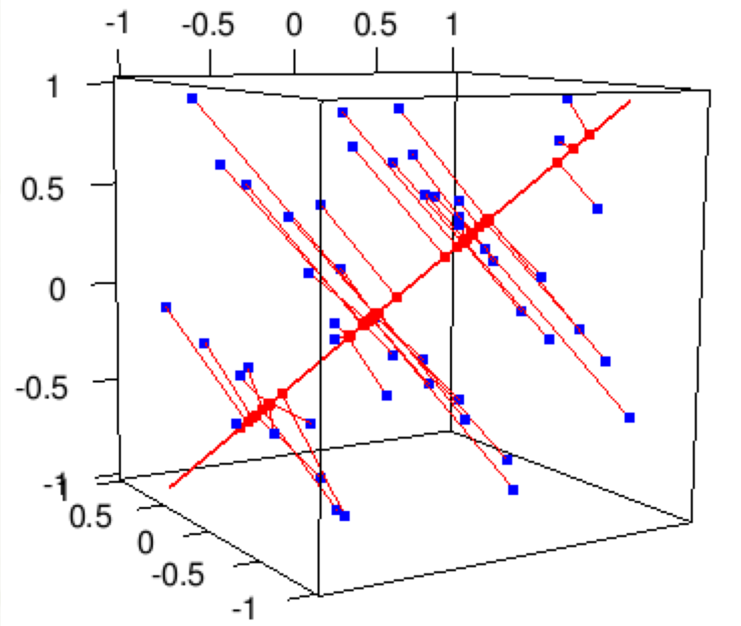
A good benchmark

A good SFD ?

… and without other drawbacks !

A 3D randomized OA





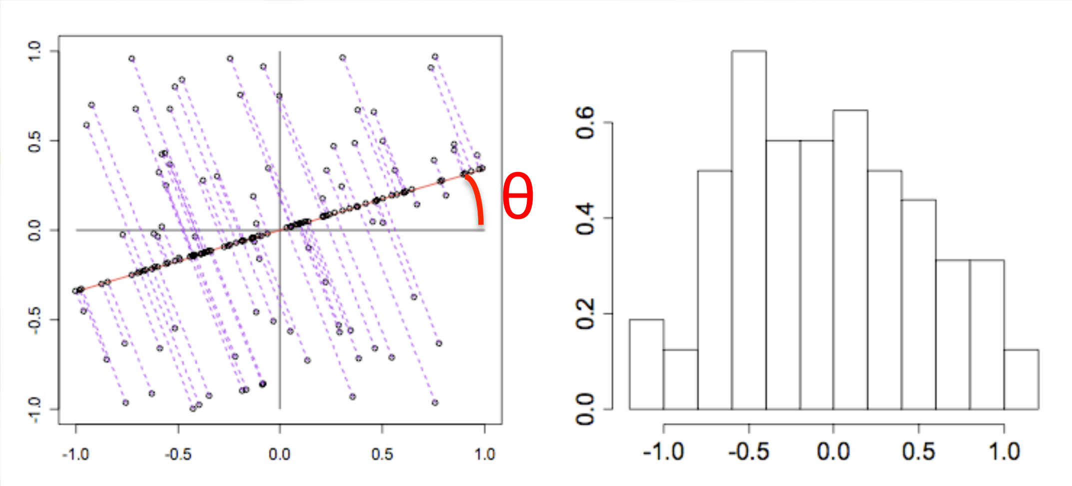
2nd part

- The radial scanning statistic

The idea

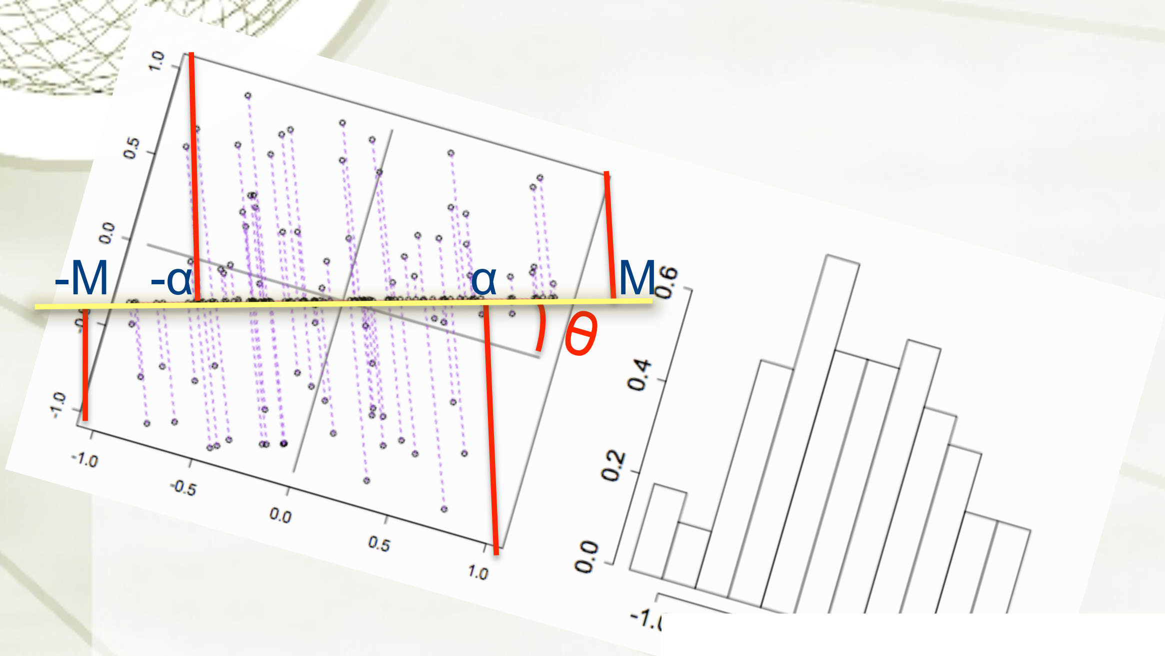
- Radial scanning – Case of a 2D subspace
  - Scan angularly the domain
  - For each radial direction, project orthogonally the design points

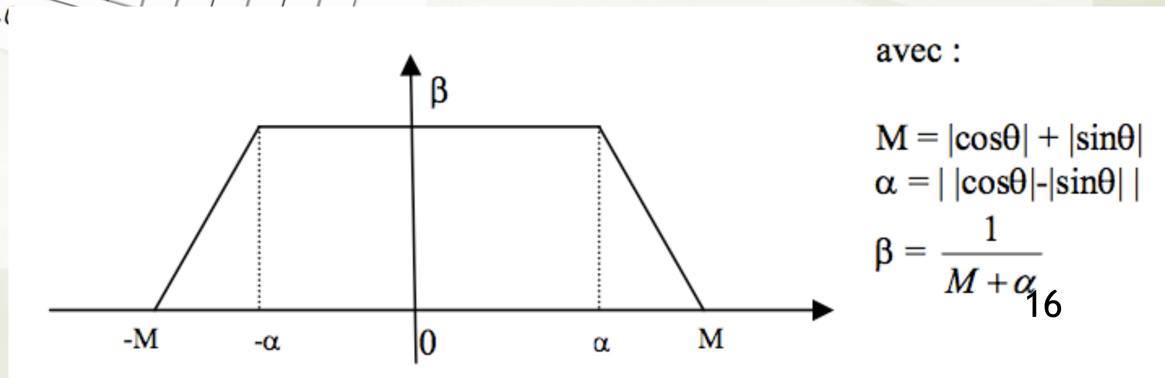
Are the projected points « **well** » distributed ?



Distribution of a sum of uniforms

- The distribution of b'X is not uniform
  - Depends on the projections of the corners of $[-1,1]^d$ onto Vect(b)





Distribution of a sum of uniforms

• Proposition (*Laplace, 18th Century ! – see a modern proof and discussion in [Elias and Shiu, 1987]*)

Proposition 1 *If X is a random vector uniformly distributed over the hypercube* $\Omega = [-1,1]^d$, *and Z is the projection of X onto the straight line generated by a unitary vector a such that* $a_j \neq 0, \forall j \in \{1,\ldots,d\}$, *then the cdf of Z is given by:*

$$F_Z(z) = \frac{1}{\prod_{j=1}^{d} 2a_j} \sum_{\mathbf{s}\in\{-1,1\}^d} \varepsilon(\mathbf{s}) \frac{(x+ <\mathbf{s},\mathbf{a}>)_+^d}{d!} \qquad (2)$$

where $\varepsilon(\mathbf{s}) = \prod_{j=1}^{d} s_j$, $<.,.>$ *is the usual scalar product and* $y_+ = \max(y,0)$. *As a result, for a given axis, Z admits a piecewise linear density whose nodes correspond to the projections of the domain corners.*

Mathematical formulation

- Assumption
  - H_0 : « $x^{(1)}, \ldots, x^{(n)}$ are a sample of the Uniform distribution »
- Formulation
  - Let b be one direction in $[-1,1]^d$
  - Let F_b be the distribution of projections b'X, with X ~U($[-1,1]^d$)
  - Question :

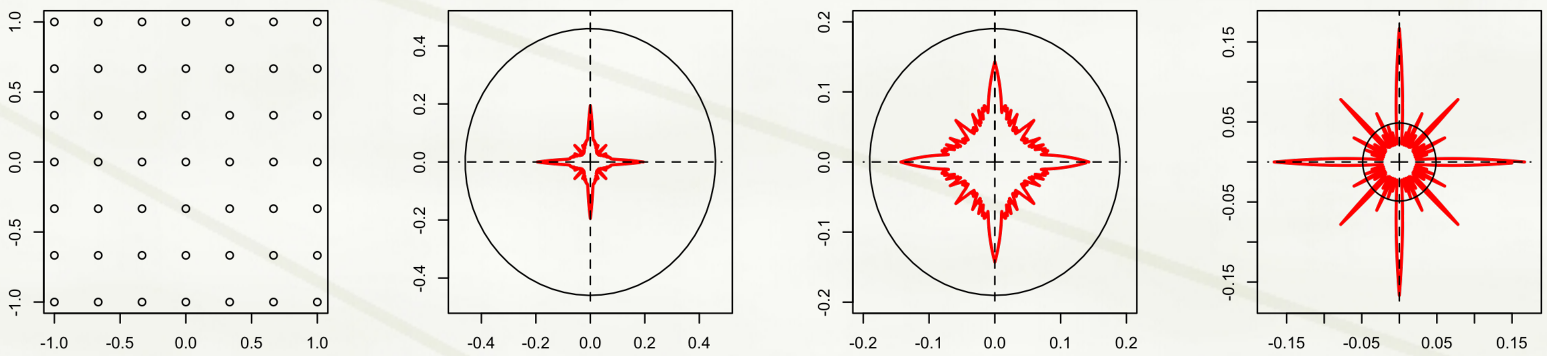
    « *Are the projections $b'x^{(1)}, \ldots, b'x^{(n)}$ drawn from F_b ?* »

  - Equivalently : « *Are $F_b(b'x^{(1)}), \ldots, F_b(b'x^{(n)})$ uniformly distributed?* »

    ⇒ Goodness-of-fit problem for the uniform distribution

Selecting a goodness-of-fit statistic

- Objective : to select a GOF statistic for the uniform distrbution that detects alignments and clustering
  - Statistics based on CDFs (KS, CVM) usually fail, statistics based on « spacings » do it, such as Greenwood ([L'ecuyer, Simard, 2007])

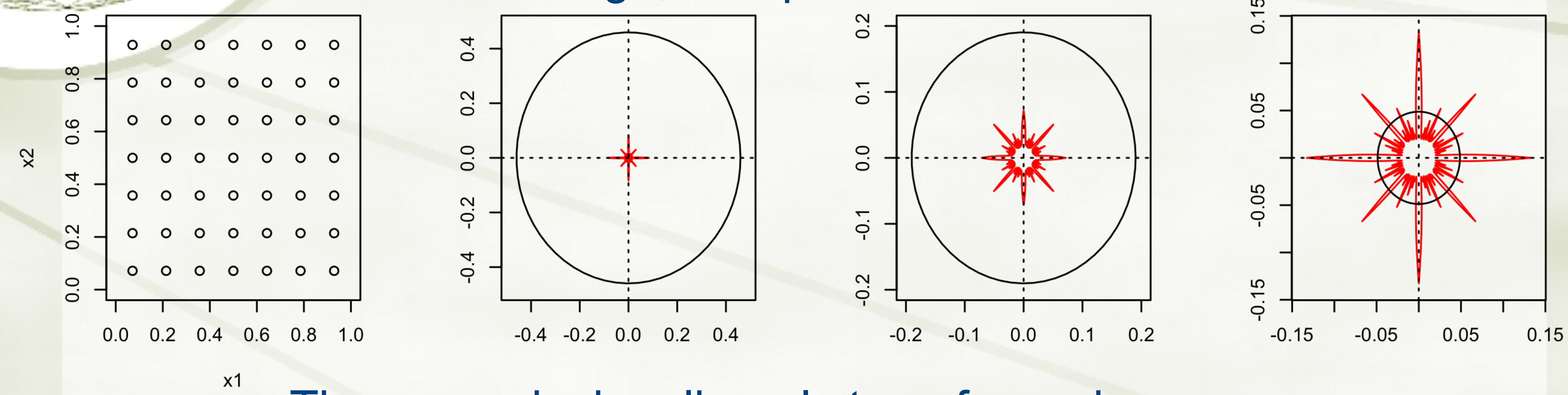


2D RSS based on : CVM (left), KS (middle), Greenwood (right)

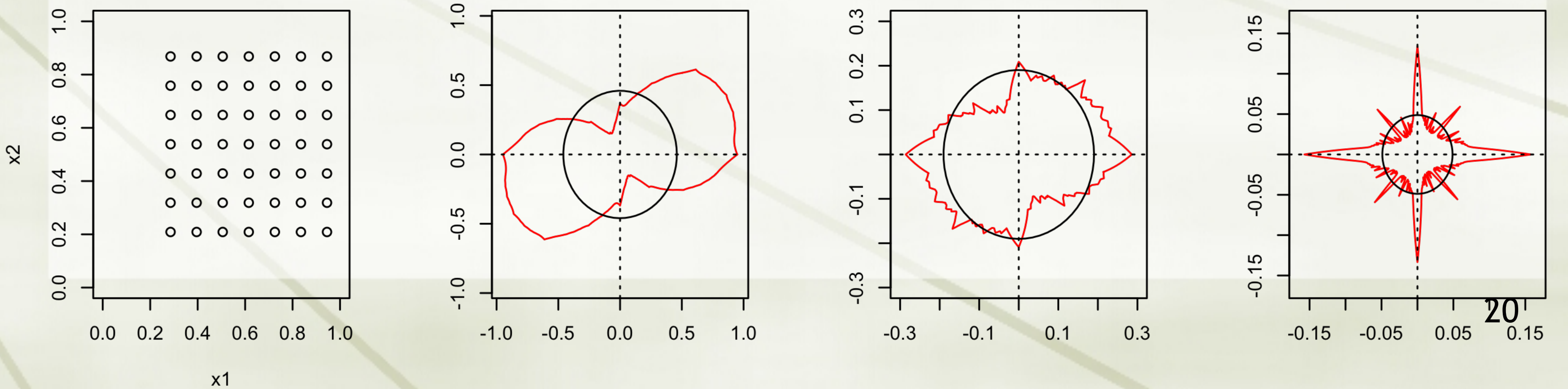
Robustness to domain misspecifications

- Another feature of the statistics based on spacings
  - ➔ much less sensitive to domain misspecifications

  - The same design, with points centered in each cell



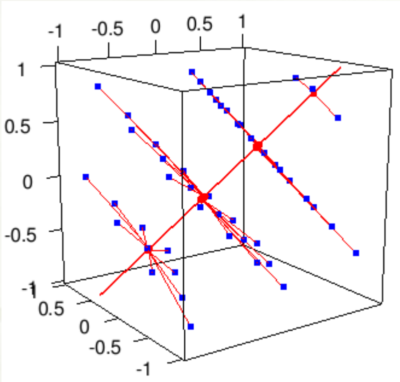
  - The same design, linearly transformed

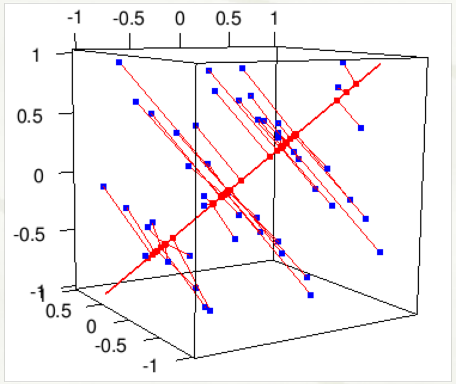


Connected work

- Spectral test
  - Context : random numbers generators testing
  - Statistic : maximal distance of points contained in oblique parallel planes

➔ Detects perfect alignements (right), but not clusters (below)





- References

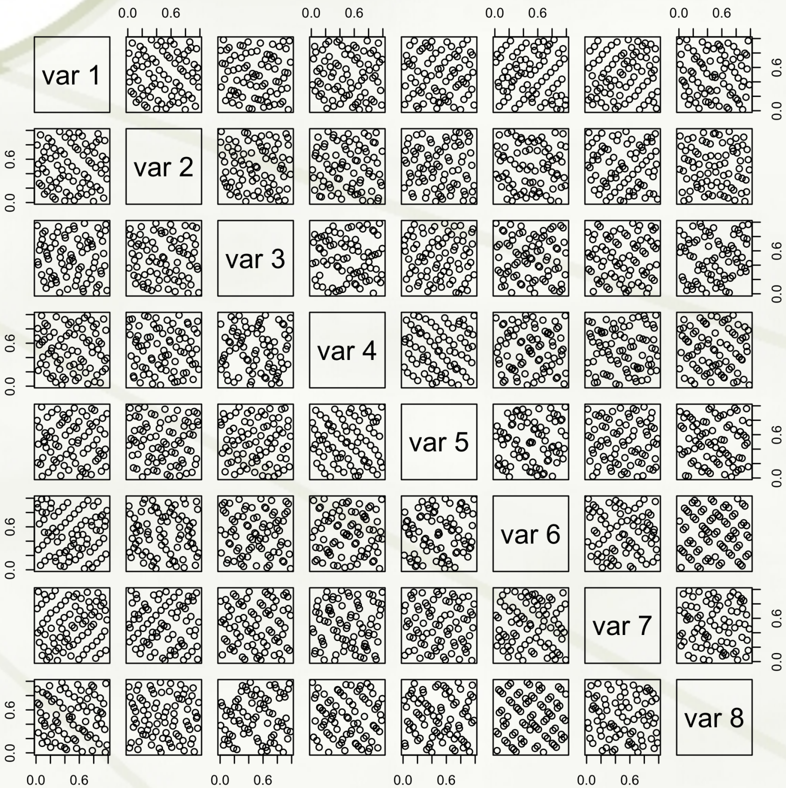
  - Knuth, D. (1997). *The Art of Computer Programmig, Volume 2: Seminumerical Algorithms, 3rd edition. Addison-Wesley.*
  - Ripley, B. (1987). *Stochastic Simulation. Wiley.*

3rd part

- Applications

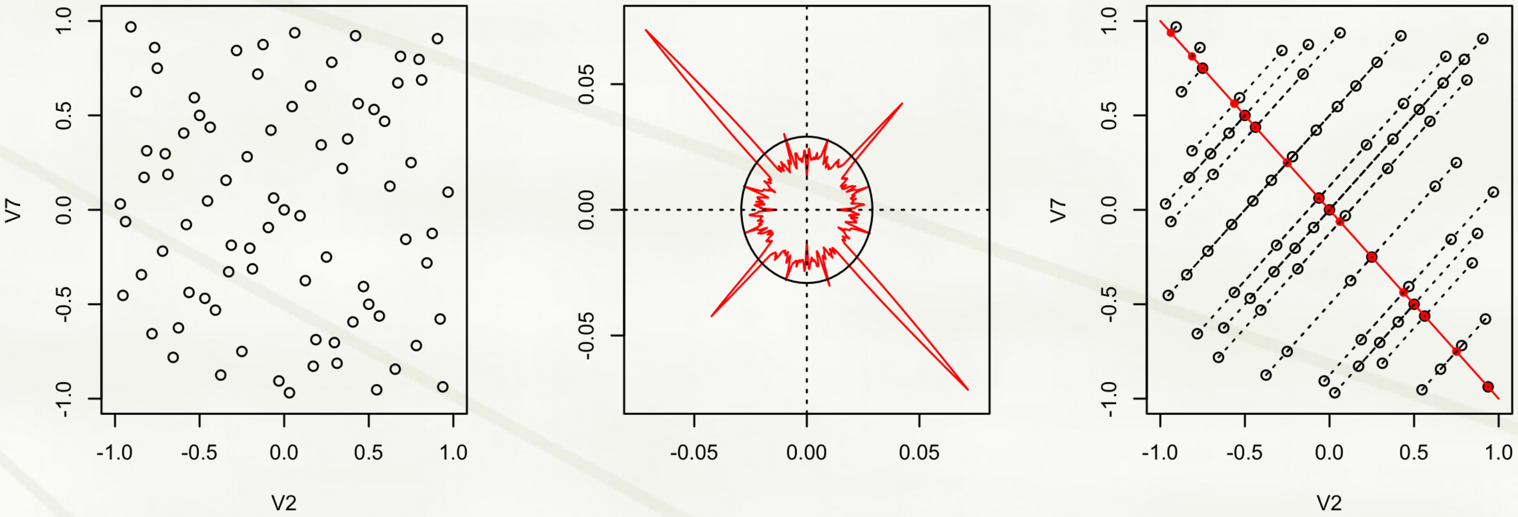
Defect detection

- Let us take again the 8D example
    - Space-filling design : 80 points Sobol sequence



Defect detection

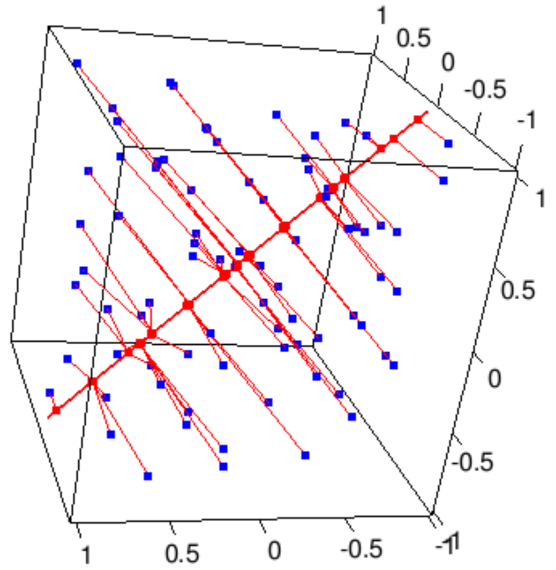
- 8D example (following) – 2D RSS
  - the worst case is for the pair of dimensions (2,7)
  - in this 2D factorial subspace, the worst direction is ≈ (1,-1)



Defect detection

- 8D example (following) – 3D RSS
  - the worst case is for the triplet of dimensions (1,2,7)
  - in this 3D factorial subspace, the worst direction is ≈ (0,1,-1)

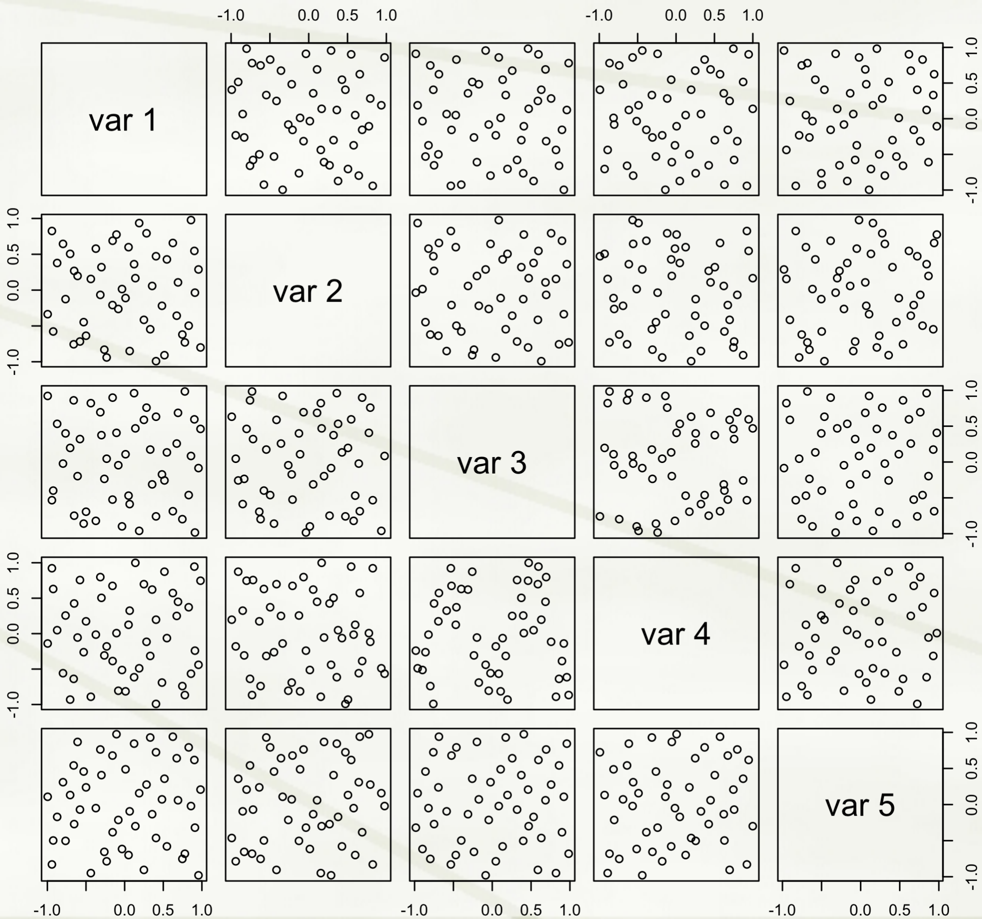
(same problem detected)



Defect detection

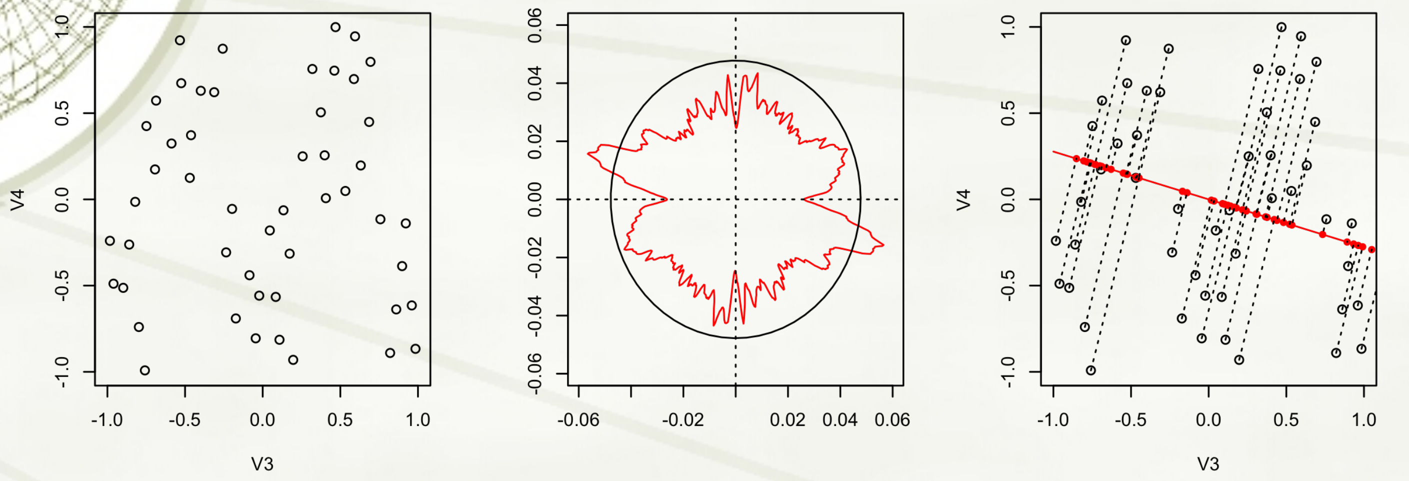
- Low Discrepancy Sequences (LDS)
  - Sobol

  Owen scrambling



Defect detection

- Low Discrepancy Sequences – 2D and 3D RSS

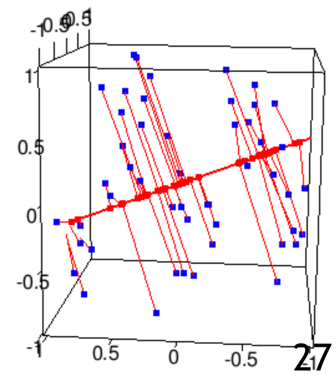


In 2D

- worst subspace : (3,4)
- worst direction ≈ (0.96, -0.26)

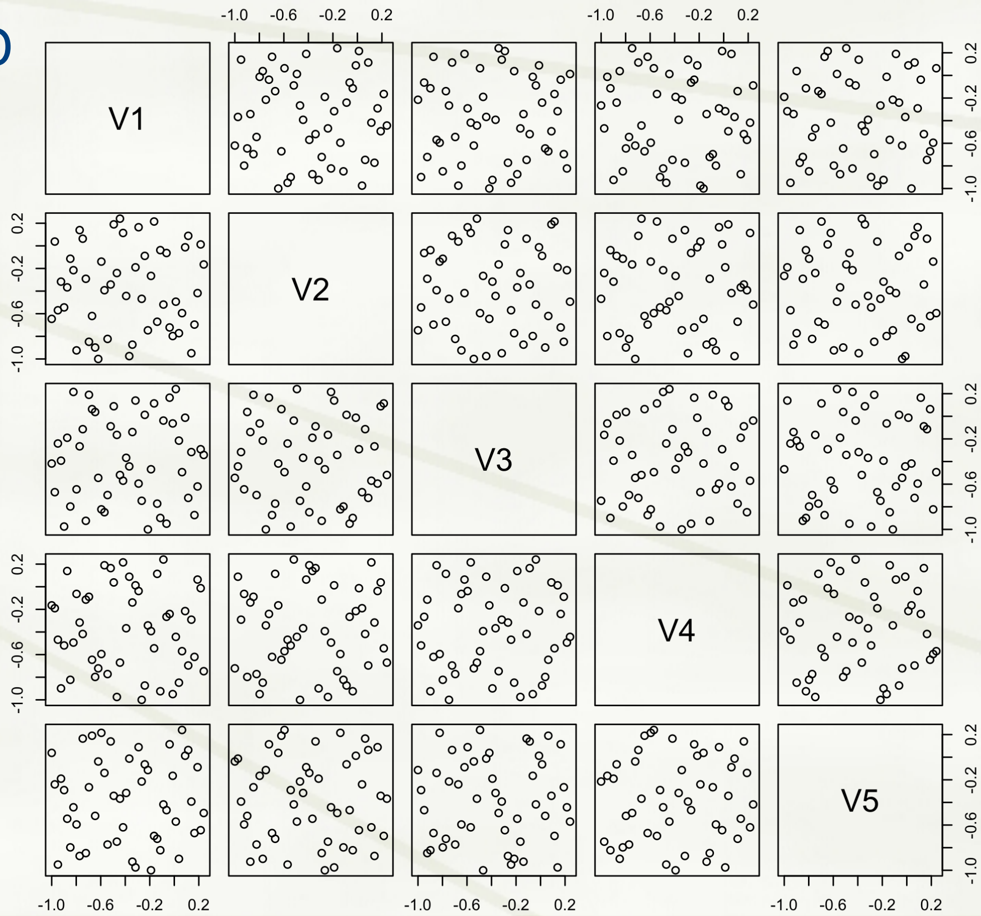
In 3D

- worst subspace : (1,3,4)
- worst direction ≈ (-0.14, 0.93, -0.33)



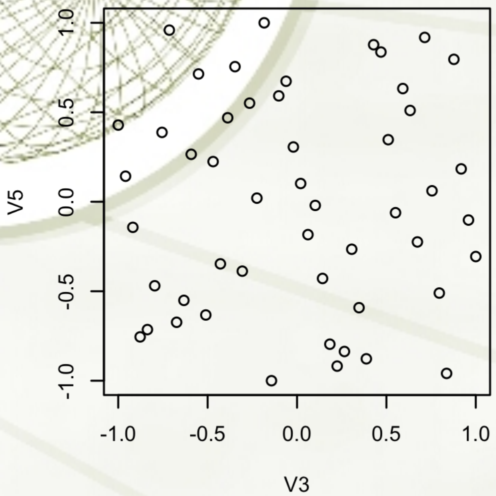
Defect detection

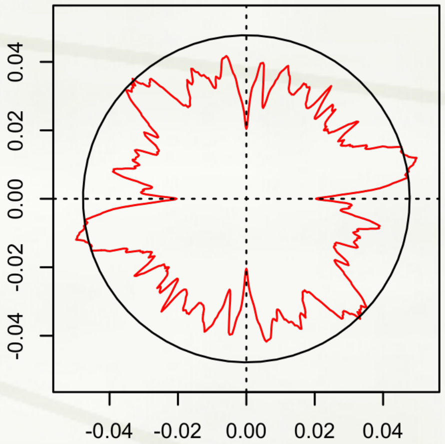
- Latin hypercube (LH) designs
  - Maximin LHD

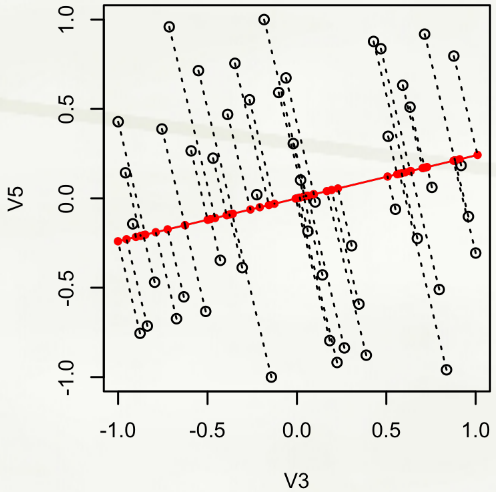


Defect detection

- Maximin LHD





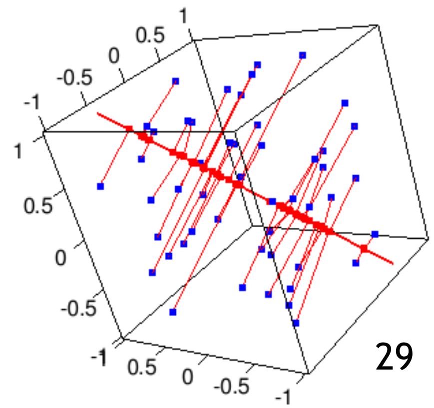


In 2D

- worst subspace : (3,5)
- worst direction ≈ (0.97, 0.23)

In 3D

- worst subspace : (3,4,5)
- worst direction ≈ (-0.31, 0.69, 0.64)



Selection of SFDs

- Comparison of 8D SFDs of size 80

Table 1 Worst value of Greenwood statistic for 8-dimensional SFDs of size 80

Design type[a]	Statistic value[b]
Uniform	0.039 (0.003)
Maximin Latin hypercube	0.048
Audze-Eglais Latin hypercube	0.037
Halton sequence	0.244
Faure sequence	0.161
Sobol sequence	0.101
Sobol sequence, with Owen scrambling	0.041 (0.006)
Sobol sequence, with Faure-Tezuka scrambling	0.088 (0.010)
Sobol sequence, with Owen + Faure-Tezuka scrambling	0.041 (0.006)
Strauss	0.040 (0.004)

[a] LHDs are taken from `http://www.spacefillingdesigns.nl`. Halton and Sobol sequences are computed with the R package randtoolbox (`http://www.r-project.org`). [b] For stochastic designs, the first number is the mean of the results over 100 simulations, and the second (into brackets) their standard deviation.

Future research

- Statistical issue
- Higher dimensions

Decisional issues

- Multiple testing framework
  - multiple pairs (triplets) of dimensions
  - multiple angles
  - strong correlation !

- Partial solution
  - consider a global statistic over directions, such as: sup/inf

  ➔ the multiple testing issue over dimensions remains…

Future research

- Radial scanning in higher dimensional subspaces
  - Visualization is no longer possible
  - Computational cost
    - ➔ Optimization techniques
    - ➔ Normal approximations ? (away from factorial subspaces)

- Other ideas
  - Radial scanning of 2D (or higher) subspaces

Acknowledgements

We wish to thank A. Antoniadis, the members of the DICE Consortium (http://www.dice-consortium.fr), the participants of ENBIS-DEINDE 2007, as well as two referees for their useful comments. We also thank Chris Yukna for his help in editing.

THANK YOU FOR YOUR ATTENTION