TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Mechanical Engineering, Institute of Fluid Mechanics, Chair of Turbomachinery and Jet-Propulsion

- Extension of Latin Hypercube samples -

Robin Schmidt, Matthias Voigt, Konrad Vogeler

Dresden, 11.10.2013

DRESDEN concept
Exzellenz aus Wissenschaft und Kultur

Motivation

STATEMENT

- Monte-Carlo Method with LHS is a multi-purpose way to tackle probabilistic analyses

QUESTION

- What determines the selection of the number of realizations?

Motivation

ANSWER

- confidence intervals of the resulting statistical values (not considering probability of failure)

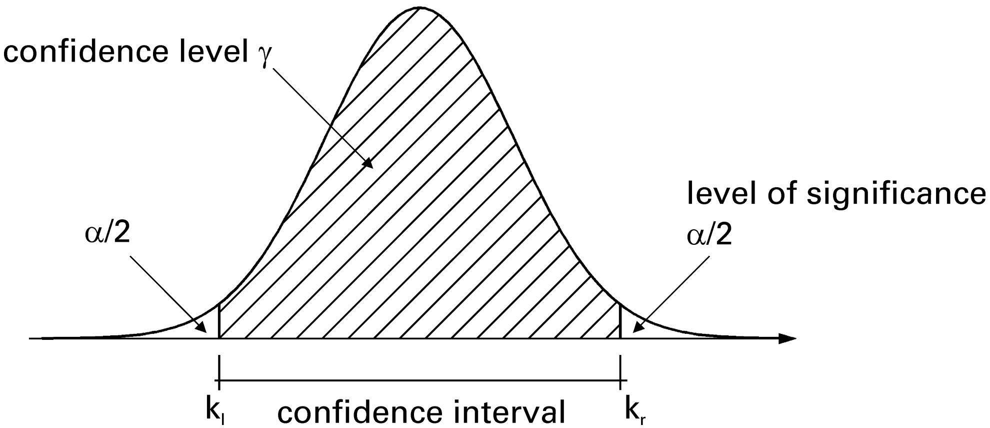


Figure 1: Confidence interval, schematic

Motivation

Example: upper confidence interval for the mean as a function of n_{sim}; determined by 1000 reps.

$$y_1 = \frac{(b_{u,1} + b_{u,2} - 1)}{\sqrt{2/12}}$$

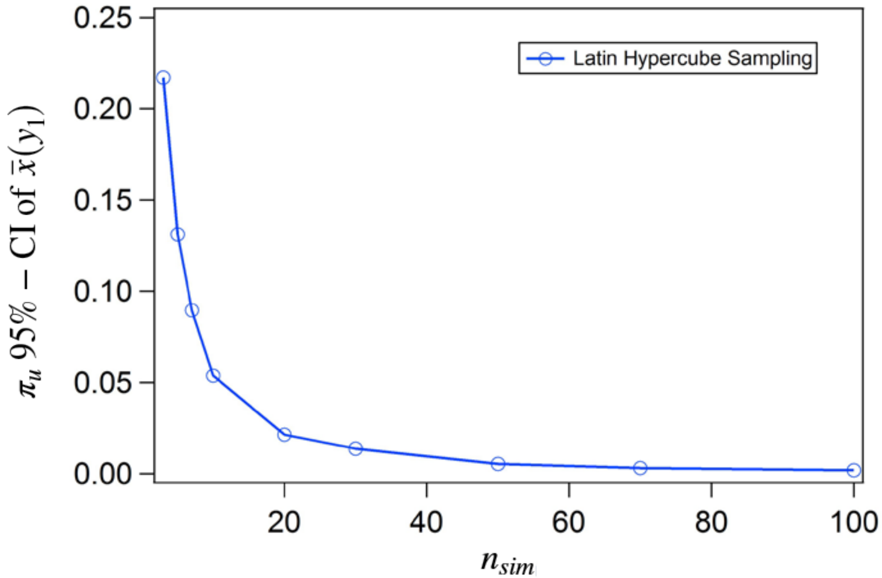


Figure 2: 95%-CI of mean as a function of n_{sim}

CONCLUSION

- Size of the confidence intervals of the resulting values are determined by the number of realizations
- statistical quality of the Monte-Carlo Simulation (MCS) with LHS can be conservatively determined after its execution (calculation of confidence intervals of the resulting values)

Motivation

OBJECTIVE TARGET for MCS + LHS

- start with a small number of realizations with sufficient (statistically reasonable) quality; especially for time-consuming deterministic calculations
- improve quality by adding more realizations
- assessment based on the confidence intervals of the statistical measures

Overview

Motivation

Latin Hypercube sampling (LHS)

State of the art

Replicated Latin Hypercube sampling (rLHS)

Evaluation

Outlook

LHS

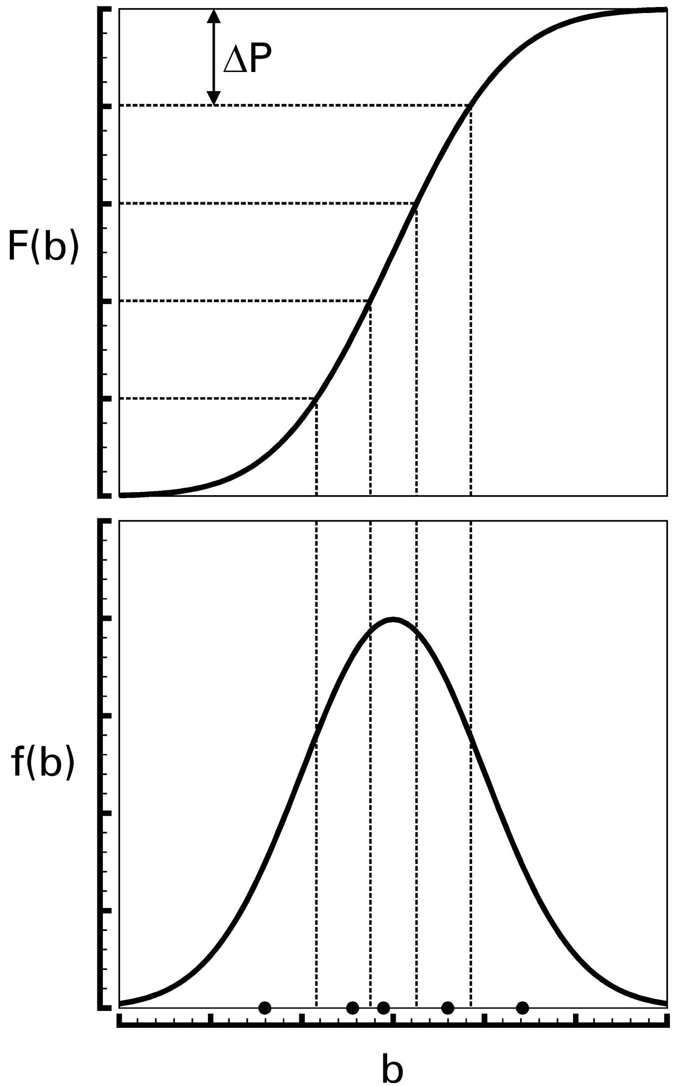


Figure 3: LHS schematic

CHARACTERISTIC:

- each realization represents equal probability ΔP

APPROACH:

- define number of realizations n_{sim}
- determine $\Delta P=1/n_{sim}$ wide intervals on F(b)
- select one value at random from each interval

PROPERTIES:

- good representation of cdf with “few” realizations
- more stable analysis outcomes than random sampling
- easier implementation than stratified sampling methods

State of the art

Extension of LHS:

Pleming et al.,2005 – Replicated LHS

multiplication of a basis value; reduplication of the intervals; intervals must not be completely filled; uses Restricted Pairing (RP) for correlation control

Sallabery et al.,2008 – Extension of LHS with correlated variables

focus on the correlation setting; reduplication of original lhs; algorithm for the calculation of the intervals to be filled in the multidimensional space with given rank correlation matrix

Calculation of Confidence Interval of LHS:

Iman,1981 – Replicated LHS

repeatedly execution of LHS to generate replicates; replicates are not mergeable; rule for the calculation of confidence intervals with the use of replicates

rLHS

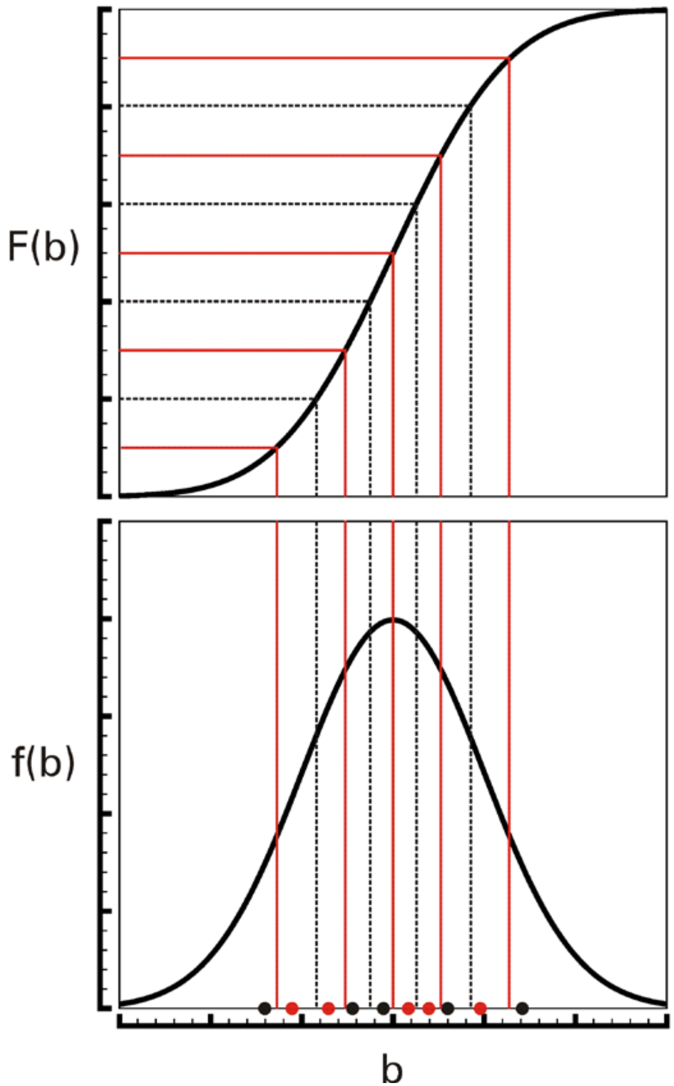


Figure 4: rLHS schematic

INITITAL POSITION

- define *basis* and *level*
- use “classic” LHS with $n_{sim,start}$=*basis* realizations

APPROACH

- Use “small” *basis* and reach the desired $n_{sim,end}$ by replication *level* times

IMPLEMENTATION

- reduplicate the intervals on F(b) if necessary
- select one value at random from each free interval
- per replication step only *basis* values are added

IMPLEMENTATION

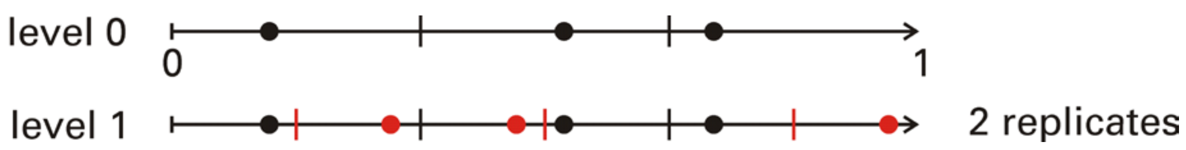


Figure 5: Reduplication 1

Example: U[0,1]; LHS with *basis* = 3:

- *level* 1 – n_{sim}=6: reduplication of original intervals, fill free intervals

IMPLEMENTATION

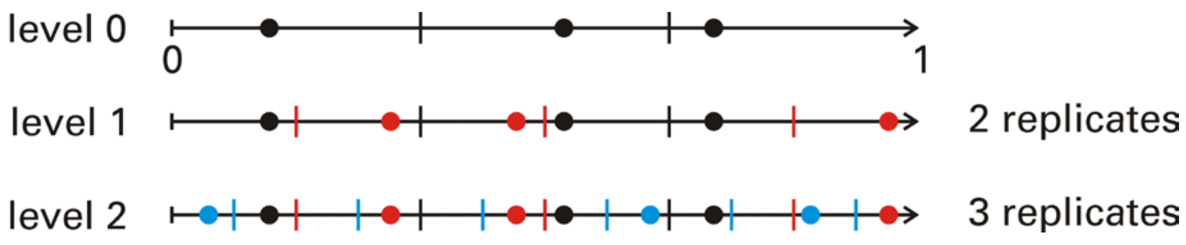


Figure 6: Reduplication 2

Example: U[0,1]; LHS with *basis* = 3:

- *level* 1 – n_{sim}=6: reduplication of original intervals, fill free intervals
- *level* 2 – n_{sim}=9: reduplication of intervals of level 1, determine D* as the largest negative distance between continuous and discrete cdf for each original interval (level 0)

$$D^* = \min_{1 \leq i \leq N} \left(F(y_i) - \frac{i}{N} \right)$$

place a random number per original interval in a free interval with respect to D*

IMPLEMENTATION

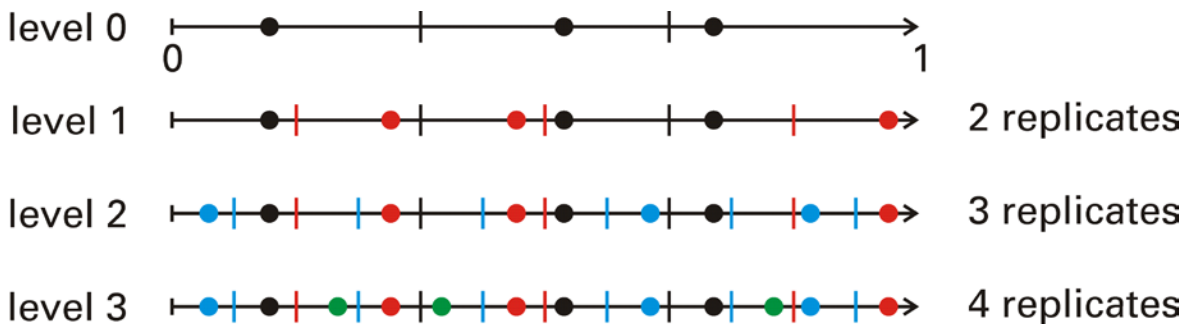


Figure 7: Reduplication 3

Example: U[0,1]; LHS with ***basis*** = 3:

- *level 1* – n_{sim}=6
- *level 2* – n_{sim}=9
- *level 3* – n_{sim}=12: if there are free intervals no reduplication is done and one value at random is selected per original interval from each free interval (in higher levels D* is used)

rLHS – correlation control

ALGORITHM

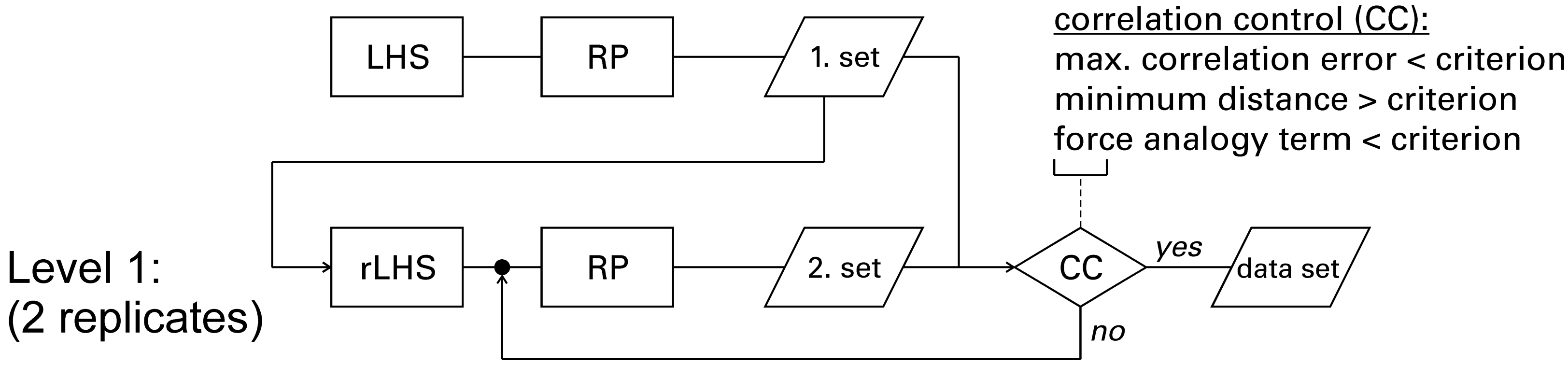


Figure 8: CC-Algorithm

- start with “classic” LHS
- use Restricted Pairing (RP) [IMAN ET AL.,1980; IMAN, 1981] for correlation definition

Overview

Motivation

Latin Hypercube sampling (LHS)

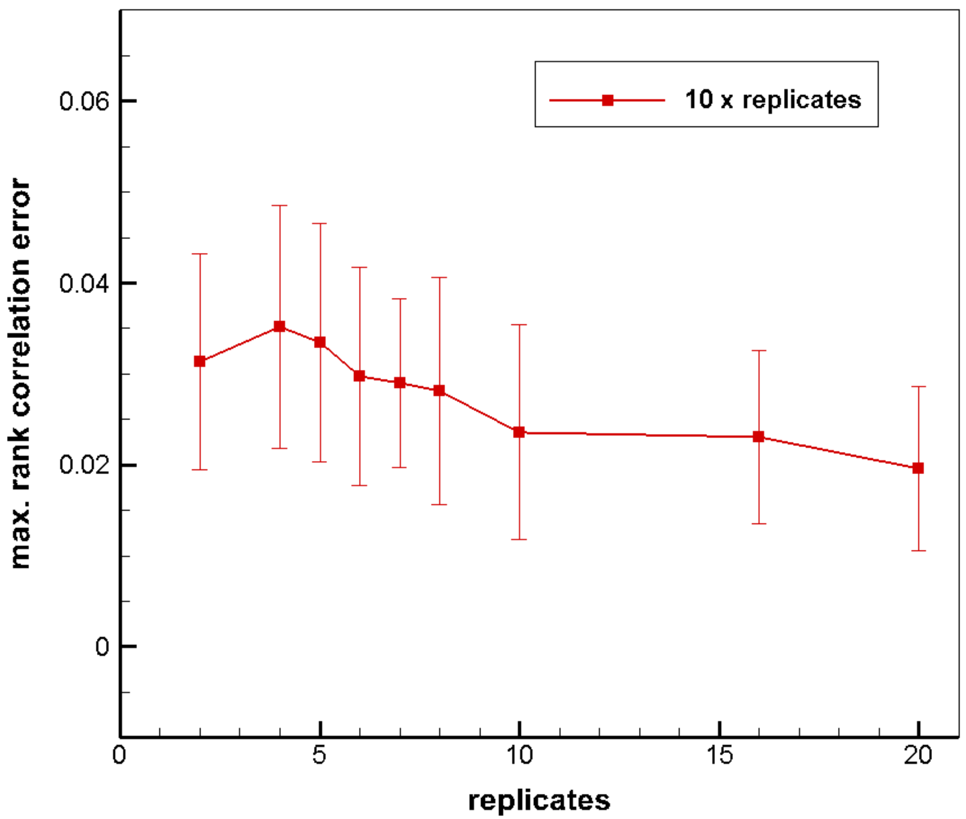
State of the art

Replicated Latin Hypercube sampling (rLHS)

Evaluation

Outlook

Correlation coefficient vs. number of replicates



Generation of 4 variables:
2 x uniform, U[0,1]
2 x normal, N(0,1)

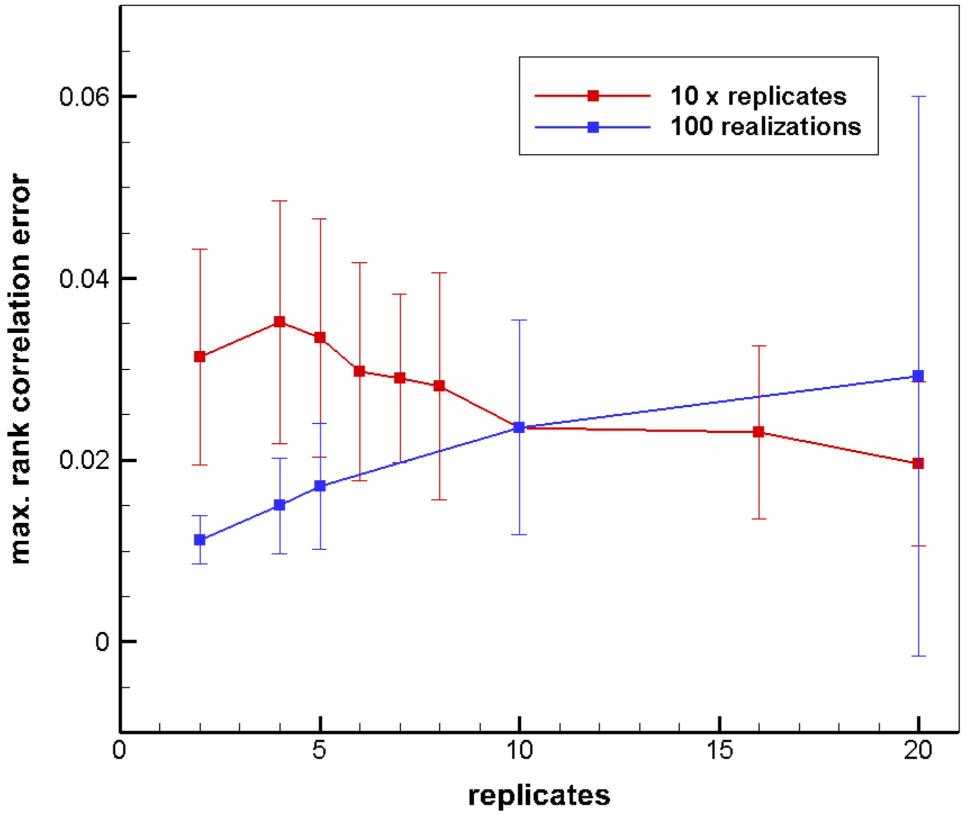
with:

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Figure 9: Max. correlation error 1

Evaluation

Correlation coefficient vs. number of replicates



Generation of 4 variables:
2 x uniform, U[0,1]
2 x normal, N(0,1)

with:

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Figure 10: Max. correlation error 2

Evaluation

K-S-value vs. number of replicates

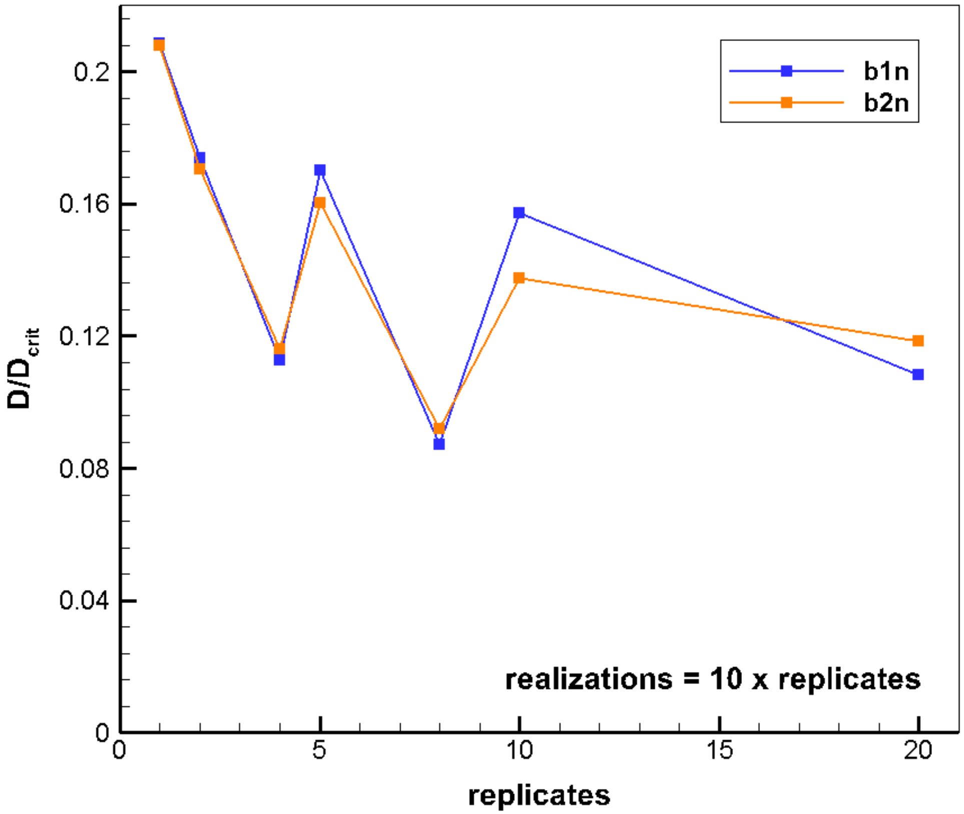


Figure 11: K-S-value

Generation of 2 variables:
2 x normal, N(0,1)

K-S-test of goodness of fit:

$$b_1, b_2, \ldots, b_n$$

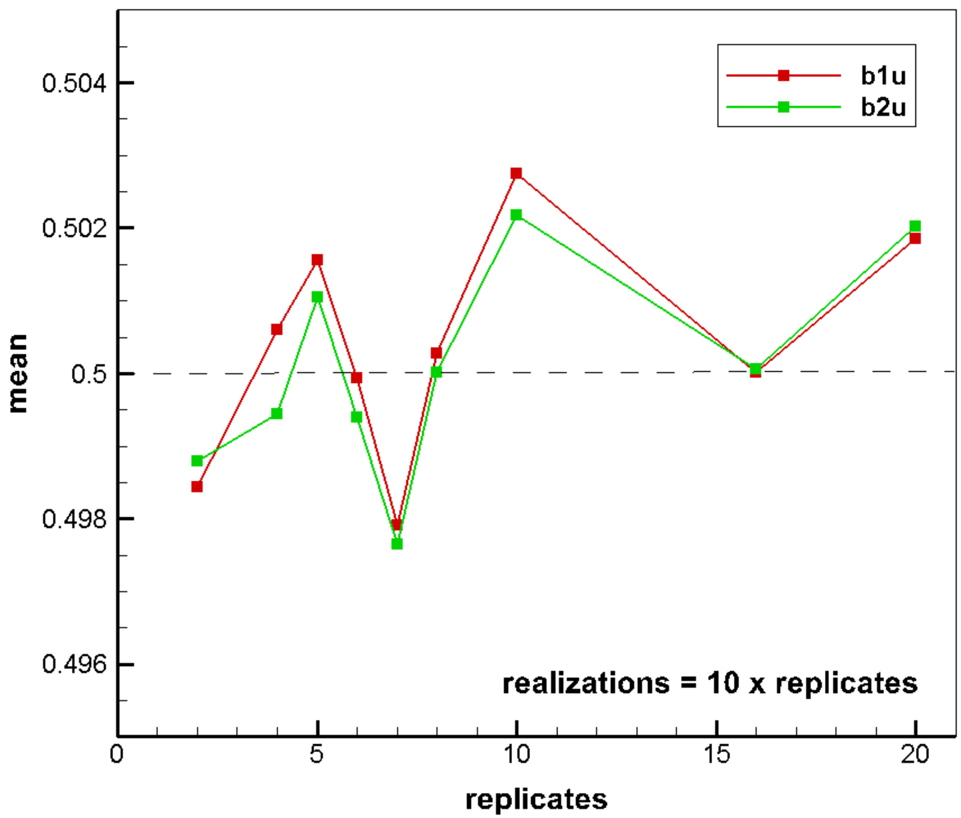
$$E_n = \frac{n(i)}{n}$$

$$D = \max_{1 \leq i \leq N} \left| F(y_i) - \frac{i}{N} \right|$$

$$D_{\text{crit}} = \frac{1.358}{\sqrt{n_{\text{sim}}}}$$

Evaluation

Statistical measures of the input values vs. number of replicates

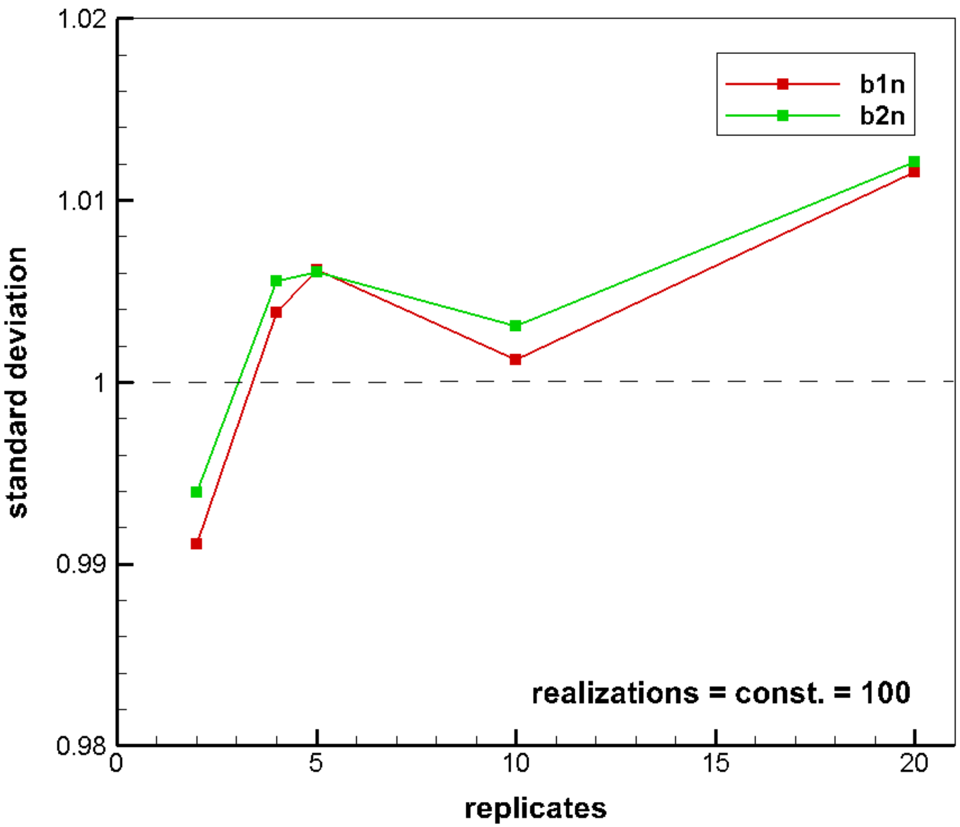


Generation of 2 variables:
2 x uniform, U[0,1]

Figure 12: Mean vs. replicates

Evaluation

Statistical measures of the input values vs. number of replicates



Generation of 2 variables:
2 x normal, N(0,1)

Figure 13: Stdev vs. replicates

Evaluation

95% - confidence interval (CI) of result values

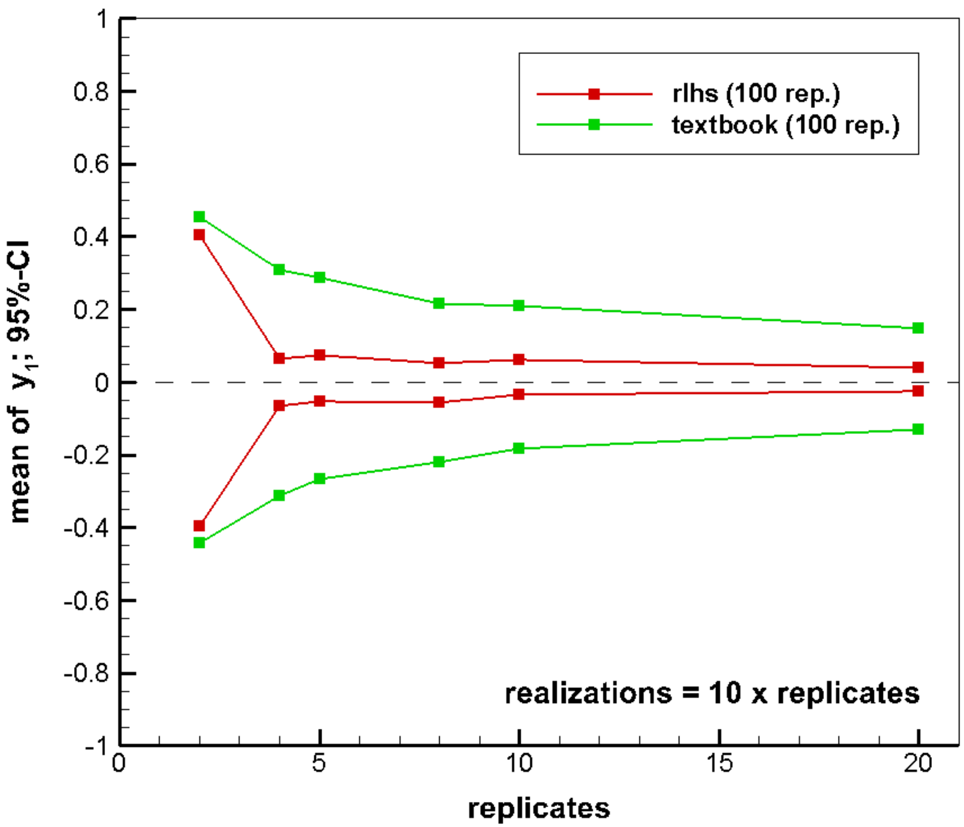


Figure 14: CI rLHS

$$y_1 = \frac{(b_{u,1} + b_{u,2} - 1)}{\sqrt{2/12}}$$

Calculation of 95%-CI Textbook:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n_{\text{sim}}}}$$

Replicated Latin Hypercube sampling (rlhs) [HELTON,2003]:

$$\bar{x} \pm t_{r-1;1-\alpha/2} \cdot SE$$

$$SE = \sqrt{\sum_{m=1}^{r} \frac{[\bar{x}_m - \bar{x}]^2}{r(r-1)}}$$

r…number of replicates

Evaluation

95% - confidence interval (CI) of result values

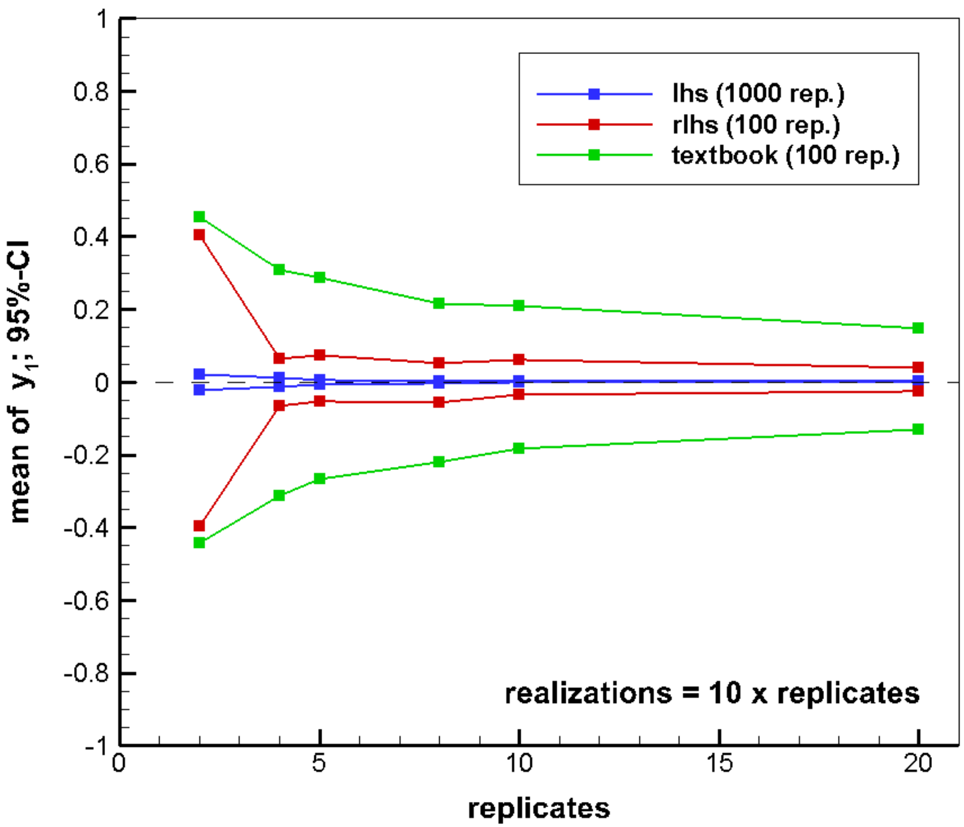


Figure 15: CI LHS

$$y_1 = \frac{(b_{u,1} + b_{u,2} - 1)}{\sqrt{2/12}}$$

Calculation of 95%-CI Textbook:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n_{\text{sim}}}}$$

Replicated Latin Hypercube sampling (rlhs) [HELTON,2003]:

$$\bar{x} \pm t_{r-1;1-\alpha/2} \cdot SE$$

$$SE = \sqrt{\sum_{m=1}^{r} \frac{[\bar{x}_m - \bar{x}]^2}{r(r-1)}}$$

r…number of replicates

Evaluation

95% - confidence interval (CI) of result values

$$y_1 = \frac{(b_{u,1} + b_{u,2} - 1)}{\sqrt{2/12}}$$

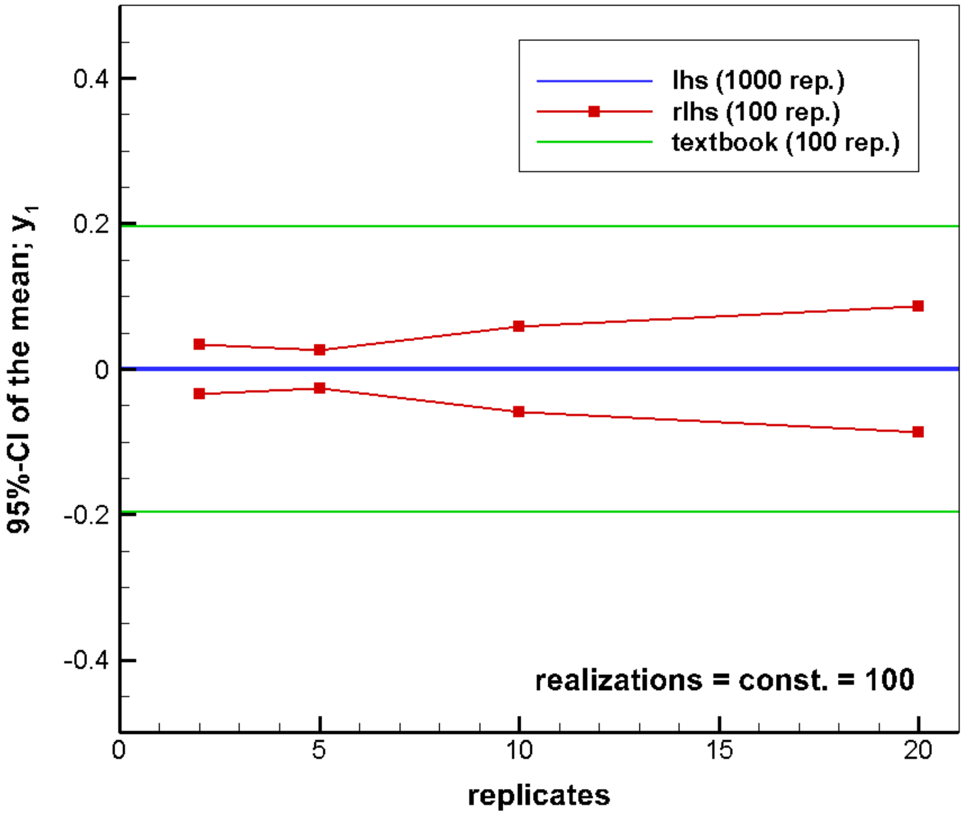


Figure 15: CI vs. replicates

Outlook

- Implementation of a replicated Latin Hypercube sampling with the ability to increase the sample size and to induce or keep a desired correlation among input parameters

- Analysis of the algorithm (influence of number of replicates) with more degrees of freedom
- Test of the performance against “classic” LHS in terms of:
  - difference in maximum correlation error
  - difference of the cdf
  - difference of the statistical measures
- Analysis of the CI calculation
  - influence of free intervals
  - deviation from experimentally determined confidence intervals

References

Pleming et al.,2005, Replicated Latin Hypercube Sampling, AIAA 2005-1819, pp. 1-18

Sallabery et al.,2008, Extension of Latin hypercube samples with correlated variables, Reliability Engineering and System Safety, *93*, pp. 1047-1059

Iman, R. L. and Conover, W., 1980, Small sample sensitivity analysis techniques for computer models with an application to risk assessment, Communication in Statistics - Theory and Methods, 9(17), pp. 1749–1842.

R. L. Iman, 1981, Statistical Methods for Including Uncertainties Associated with the Geologic Disposal of Radioactive Waste which Allow for a Comparison with Licensing Criteria, Uncertainties Associated with the Regulation of Geologic Disposal of High-Level Radioactive Waste, Gatlinburg, TN

Helton et al., 2003, Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, Reliability Engineering and System Safety, *81*, pp. 23-69

Restricted Pairing (RP)

Sample of size m from n input variables:

$$m \times n \text{ matrix} \qquad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \dots & \dots & & \dots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{pmatrix}$$

Desired correlation structure:

$$n \times n \text{ matrix} \qquad \mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ t_{21} & t_{22} & \dots & t_{2n} \\ \dots & \dots & & \dots \\ t_{n1} & t_{n2} & \dots & t_{nn} \end{pmatrix}$$

(1) Mixing or rearrangement of the realizations after LHS (problem: perfect linear correlations)
(2) Calculation of correlation matrix **C** (rank correlation of **B**)
(3) Calculation (possible with Cholesky factorization) of the lower triangular matrix **Q** such that:

$$\mathbf{C} = \mathbf{Q}\mathbf{Q}^T$$

(4) Calculation of **P** such that:

$$\mathbf{T} = \mathbf{P}\mathbf{P}^T$$

T and **C** have to be a symmetric, positive-definite matrix

(5) Calculation of **S** such that:

$$\mathbf{S} = \mathbf{P}\mathbf{Q}^{-1}$$

(6) Calculation of **R** such that:

$$\mathbf{R} = \mathbf{B}\mathbf{S}^T$$

(7) **R** will approximate **T**, the column of **B** must be sorted so that they follow the same ranking of values, as the columns in the matrix **R**