Stat Comput
DOI 10.1007/s11222-015-9624-z

A standardized distance-based index to assess the quality of space-filling designs

François Wahl[1,2] · **Cécile Mercadier**[1] · **Céline Helbert**[1]

Received: 3 May 2014 / Accepted: 24 December 2015
© Springer Science+Business Media New York 2016

Abstract One of the most used criterion for evaluating space-filling design in computer experiments is the minimal distance between pairs of points. The focus of this paper is to propose a normalized quality index that is based on the distribution of the minimal distance when points are drawn independently from the uniform distribution over the unit hypercube. Expressions of this index are explicitly given in terms of polynomials under any L_p distance. When the size of the design or the dimension of the space is large, approximations relying on extreme value theory are derived. Some illustrations of our index are presented on simulated data and on a real problem.

Keywords Minimal distance · Maximin · Space-filling design · Computer experiments · Extreme value theory

1 Introduction

In the field of computer experiments, simulation of designs replace the real data generating process. Under a lack of information on how inputs are linked to outputs, one strategy is to spread points evenly throughout the experimental region to cover all the input space. This technic is called *space-filling* design (Sacks et al. 1989; Santner et al. 2003). Many measures are available to quantify this property, as for example distance based criteria, discrepancies or entropy (Fang et al. 2005; Pronzato and Muller 2012). In this paper we focus on the minimal distance between points of a design.

It is very common to iterate a few times the procedure for finding a design and to keep as the "best" one, the design with the maximum minimal distance (mindist). When it is maximized in the context of space-filling design, it is called the *maximin criterion*. Due to its simplicity, it is by far the most commonly used.

The minimal distance is dependent on both the number of input factors and the number of points present in the design. Thus, by itself, its value is not informative.

The aim of this paper is then to propose a quality index normalized between 0 and 1 for the minimal distance. This index allows a better comparison between designs involving different sizes or dimensions rather than looking at the order, for which mindist would be sufficient. The use of our index does not replace a maximin procedure. It is a measure of the quality of space-filling design through a natural common scale of reference. It also gives an idea of the cost of the design: more precisely as this index is based on a probability, it specifies the difficulty of obtaining the same minimal distance. The quality index will thus assist the user in her/his decision to keep the design or to generate a better one.

The paper is organized as follows. Section 2 introduces the new index. Section 3 provides probabilistic characteristics of the distance between two random points independently and uniformly sampled in the unit hypercube. Section 4 states some approximations for the previous distribution and for the distribution of the minimal distance among all the pairs of a design. Finally, in Sects. 5 and 6, we illustrate the use of the index as a measure of quality through a simulation study and on a real example in engine calibration.

✉ François Wahl
francois.wahl@univ-lyon1.fr; Francois.Wahl@ifpen.fr

Cécile Mercadier
mercadier@math.univ-lyon1.fr

Céline Helbert
celine.helbert@ec-lyon.fr

[1] Université de Lyon, CNRS UMR 5208, Université Lyon 1, Institut Camille Jordan, 43 bvd du 11 novembre 1918, 69622 Villeurbanne, France

[2] IFPEN, BP3, 69390 Solaize, France

Published online: 05 January 2016

2 New index of quality

Let $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a set of N points in a d-dimensional hypercube, assumed to be $[0, 1]^d$ without loss of generality. In the field of computer experiments, $\mathbf{x}$ is referred to as a *design*. The minimal distance between pairs of $\mathbf{x}$ is defined as

$$\delta_{\mathbf{x},p} = \delta_{\mathbf{x},p}(N, d) := \min_{1 \le i \ne j \le N} D_{p,d}(\mathbf{x}_i, \mathbf{x}_j) \ ,$$

where, for any positive real p, $D_{p,d}$ stands for the L_p distance in $\mathbb{R}^d$.

In particular $D_{1,d}$ is the Manhattan distance, $D_{2,d}$ the Euclidean distance and $D_{\infty,d}$ the Chebyshev distance. More generally, recall that for a finite value of p,

$$D_{p,d}(\mathbf{u}, \mathbf{v}) = \left(\sum_{k=1}^{d} |\mathbf{u}^{(k)} - \mathbf{v}^{(k)}|^p \right)^{1/p}$$

where $\mathbf{u}^{(k)}$ is the kth coordinate of the vector $\mathbf{u}$, and that

$$D_{\infty,d}(\mathbf{u}, \mathbf{v}) = \max_{k=1,\ldots,d} |\mathbf{u}^{(k)} - \mathbf{v}^{(k)}| \ .$$

The main idea of our quality index for a design $\mathbf{x}$ is to give a probability, and thereby a standardized index, of obtaining a minimal distance that is less than or equal to $\delta_{\mathbf{x},p}$. The *reference law* for evaluating this probability is the distribution of the minimal distance between N points independently drawn from the uniform distribution over the hypercube $[0, 1]^d$. More precisely, this means that all components of all the points of the reference design are independent and identically distributed from the uniform distribution on $[0, 1]$. Throughout this paper, we call such design a *uniform random sampling*.

To make the definition of the index rigorous, let us denote by $H_{p,N,d}$ the cumulative distribution function (c.d.f.) of the L_p minimal distance for a uniform random sampling. Then, we define the *quality index* by

$$\mathbb{I}_p(\mathbf{x}) := H_{p,N,d}(\delta_{\mathbf{x},p}) \ . \tag{1}$$

As a probability, the index $\mathbb{I}_p(\mathbf{x})$ will give a number between 0 and 1. This common scale will enable the user to compare different designs independently from their number of points. Note however that our index depends crucially on the particular chosen distance. Consequently, the minimal L_p distance $\delta_{\mathbf{x},p}$ of observed designs has to be quantified through $\mathbb{I}_p(\mathbf{x})$ for the same p.

An index close to 0 indicates that the minimal distance between the points of the observed design is easily reachable by a random uniform sampling. The reader should be aware that, in space-filling design, one will be mostly interested in values of the quality index very close to 1. In most cases, the probability of having a minimal L_p distance less than or equal to $\delta_{\mathbf{x},p}$ will be so close to 1 that direct numerical calculations will give exactly 1. In that case, the corresponding design is unlikely to come from a uniform random sampling, and may result from an optimization process. Consequently, to keep it meaningful in this situation, the index should be rewritten as

$$\mathbb{I}_p(\mathbf{x}) := 1 - 10^{-\mathbb{I}'_p(\mathbf{x})}$$

with $\mathbb{I}'_p(\mathbf{x})$ defined accordingly as

$$\mathbb{I}'_p(\mathbf{x}) = -\log_{10}(1 - \mathbb{I}_p(\mathbf{x})) \ .$$

Obtaining a tractable form or a numerical evaluation of $H_{p,N,d}$, the aforementioned c.d.f. that appears in (1), is the central topic of the following two sections. More details on the organization of the paper can now be stated. In Sect. 3, we shall evaluate $G_{p,d}$ the c.d.f. of $D_{p,d}(\mathbf{X}_i, \mathbf{X}_j)$ which denotes the random L_p distance between $\mathbf{X}_i$ and $\mathbf{X}_j$ (two points extracted from a uniform random sampling). Even if $\{D_{p,d}(\mathbf{X}_i, \mathbf{X}_j), 1 \le i \ne j \le N\}$ are clearly not independent under uniform random sampling, since each point contributes to $N - 1$ distances, independence approximation is investigated in Sect. 4.1. A strong link between the c.d.f. $H_{p,N,d}$ and $G_{p,d}$ will be then deduced. Finally, useful approximations for $H_{p,N,d}$ that are based on extreme value theory will be provided in Sect. 4.2.

3 Distribution of the distance of a pair

Recall that $G_{p,d}$ denotes the c.d.f. of the distance between a pair taken from a uniform random sampling. Let $g_{p,d}$ stands for the associated probability density function (p.d.f. for short). In dimension $d = 1$, these functions are all the same whatever the choice of p. Specifically, one gets

$$g_{p,1}(t) = (2 - 2t)\mathbf{1}_{[0,1]}(t)$$

and

$$G_{p,1}(t) = (2t - t^2)\mathbf{1}_{[0,1]}(t) + \mathbf{1}_{t>1} \ .$$

3.1 Polynomial forms on [0, 1]

The aim of the next proposition is to give exact expressions for $G_{p,d}(t)$ when t is restricted to the interval $[0, 1]$. Indeed, the whole support of these distributions is the interval $[0, d^{1/p}]$ but they have polynomial forms on the first unit interval.

Proposition 1 (Under L_p distance) *For any dimension d greater than or equal to 2, and for any finite positive real p, the c.d.f. of the distance of a pair extracted from a uniform random sampling has the expression:*

$$G_{p,d}(t)\mathbf{1}_{[0,1]}(t) = \sum_{\ell=d}^{2d} a_{p,d}^{(\ell)} t^{\ell} \mathbf{1}_{[0,1]}(t)$$

where

$$a_{p,d}^{(\ell)} = (-1)^{\ell-d}\binom{d}{\ell-d}\left(\frac{2}{p}\right)^{d} \frac{\Gamma(1/p)^{2d-\ell}\Gamma(2/p)^{\ell-d}}{\Gamma(\ell/p+1)}.$$

Proof This formula can be established by recursion. Denoting by $f_{p,d}$ the p.d.f. of the random variable $(D_{p,d}(\mathbf{X}_i, \mathbf{X}_j))^p$, we have for $d = 1$

$$f_{p,1}(t) = \frac{2}{p}\left(t^{1/p-1} - t^{2/p-1}\right)\mathbf{1}_{[0,1](t)} \ ,$$

and one can use the recurrence rule

$$f_{p,d}(t) = \int_0^t f_{p,d-1}(u) f_{p,1}(t-u)du \ .$$

The p.d.f. of interest is deduced from

$$g_{p,d}(t) = pt^{p-1} f_{p,d}(t^p) \ ,$$

and the c.d.f. comes as a primitive.

Let us introduce the beta function

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt \ .$$

The formula for $a_{p,d}^{(\ell)}$ is easily checked for $d = 2$ since

$$G_{p,2}(t) = \frac{2}{p}b_{1,1}t^2 - \frac{8}{3p}b_{2,1}t^3 + \frac{1}{p}b_{2,2}t^4$$

where $b_{i,j} = B\left(\frac{i}{p}, \frac{j}{p}\right)$. Assume that the formula is true until $d-1$. By application of the recurrence rule, we get the equality

$$f_{p,d}(t) = \sum_{\ell=d-1}^{2(d-1)} \frac{\ell}{p} a_{p,d-1}^{(\ell)} \int_0^t u^{\ell/p-1} f_{p,1}(t-u)du.$$

The primitive of $u^{\ell/p-1} f_{p,1}(t-u)$ can be expressed as a difference of two beta functions, which leads to

$$f_{p,d}(t) = \sum_{\ell=d-1}^{2(d-1)} \frac{2\ell}{p^2} a_{p,d-1}^{(\ell)} \left(t^{\frac{\ell+1}{p}-1} b_{\ell,1} - t^{\frac{\ell+2}{p}-1} b_{\ell,2}\right).$$

Since $B(x, y) = \dfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ and $\Gamma(x+1) = x\Gamma(x)$, we get, after some simplifications, the expected form for the coefficient $a_{p,d}^{(\ell)}$. □

Comments

1. We start with the observation that the first term $a_{p,d}^{(d)} t^d$ in $G_{p,d}(t)$ is exactly the volume of a hypersphere of radius t in dimension d, as soon as d is greater than or equal to 2.
2. The expression of the coefficient $a_{p,d}^{(\ell)}$ is now given under the most commonly used distance, that is the Euclidean distance:

$$a_{2,d}^{(\ell)} = (-1)^{\ell-d}\binom{d}{\ell-d}\frac{\pi^{(2d-\ell)/2}}{\Gamma(\ell/2+1)}.$$

3. The expressions of $G_{p,d}(t)$ are generally not polynomial when t is greater than 1. When $p = 2$, expressions of $G_{p,d}(t)$ involve trigonometric functions, as can be seen in Philip (2007) and (2010) who give detailed expressions for $d = 2, 3$ and 4.
4. Under L_1 distance, the support of $G_{1,d}$ is the interval $[0, d]$ and its c.d.f. is a polynomial on every unit subinterval of the form $[i-1, i]$. We have $g_{1,1}(t) = 2(1-t)\mathbf{1}_{[0,1]}(t)$ and the recurrence rule $g_{1,d}(t) =$

$$\begin{cases} \int_0^t g_{1,d-1}(u)g_{1,1}(t-u)du, & \text{if} \quad t \in [0,1], \\ \int_{t-1}^t g_{1,d-1}(u)g_{1,1}(t-u)du, & \text{if} \quad t \in [i-1,i], \\ \int_{t-1}^d g_{1,d-1}(u)g_{1,1}(t-u)du, & \text{if} \quad t \in [d-1,d], \end{cases}$$

   so that exact calculations (not shown here) can be done with any symbolic manipulation software.
5. Numerical approximations of $G_{p,d}(t)$ can still be easily obtained for any L_p distance and any dimension d: repeatedly draw a large enough number of times two points from a random uniform sampling and take their distance.

The next proposition is given for completness. It relies on the simple fact that $G_{\infty,d}(t) = (G_{1,1}(t))^d$ by independence of the components of the points from a uniform random sampling.

Proposition 2 (Under L_∞ distance) *For any dimension d greater than or equal to 2, the c.d.f. of the L_∞ distance of a pair extracted from a uniform random sampling has the expression:*

$$G_{\infty,d}(t) = (2t - t^2)^d \mathbf{1}_{[0,1]}(t) + \mathbf{1}_{(1,\infty)}(t) \ .$$

3.2 Gaussian approximation

As already mentioned, there is no closed form expression of the c.d.f. $G_{p,d}$ on the whole interval $[0, d^{1/p}]$ as soon as $d \geq 5$, see Philip (2010). Moreover, from Proposition 1, one knows that when restricted to the interval $[0, 1]$, $G_{p,d}(t)$ involves the values of $d + 1$ coefficients. As the dimension d increases, the number of terms increases and their numerical evaluations (involving high power of t) may become difficult.

A Gaussian approximation is possible under any L_p distance, when p is finite. It is worth mentioning that this approximation is true everywhere and not only on the interval $[0, 1]$. Furthermore, it allows easy computation of quantiles which is not true in the case of the exact polynomials and recursive expressions described in Sect. 3.1.

Proposition 3 *Let Φ stands for the standard normal c.d.f. For any finite value of p, as d tends to infinity,*

$$\forall t \in \mathbb{R}, \quad G_{p,d}(t) \sim \Phi\left(\frac{t^p - d\mu_p}{\sqrt{d}\sigma_p}\right) \tag{2}$$

with $\mu_p = \frac{2}{(p+1)(p+2)}$ and $\sigma_p^2 = \mu_{2p} - \mu_p^2$.

Proof (of Proposition 3) Let X and Y be two independent uniform random variables on $[0, 1]$. One can see that the scaled and powered distance $\frac{1}{d} D_{p,d}^p$ is given by the empirical mean of i.i.d. terms of the form $|X - Y|^p$. The asymptotic normality follows easily from Central Limit Theorem, and the asymptotic constants are obtained by $\mu_p = \mathbb{E}[|X - Y|^p]$ and $\sigma_p^2 = \mathrm{Var}[|X - Y|^p]$. □

Figure 1 illustrates the Gaussian approximation under L_2 distance.

The presentation is done through the density point of view since differences are much more emphasized this way. Two cases are provided: $d = 10$ (top) and $d = 20$ (bottom). On the first figure, the true density is displayed on $[0, 1]$, following the result from Proposition 1. The second curve comes from the empirical density estimation obtained from an intensive simulation. It can be considered as the true density on the whole domain. Finally, the third estimate illustrates the Gaussian approximation. Its performance increases with the value of d, as required by the theory. For the sake of simplicity, the corresponding illustration under L_1 distance is not provided. The Gaussian approximation is even more accurate under L_1 distance compare to the one achieved under L_2. Moreover, it is good even for reasonably small dimensions. To summarize, one should prefer the Gaussian approximation:

- When d is large, by simplicity of the expression ;
- When quantile estimates are needed, to avoid complex numerical solutions;

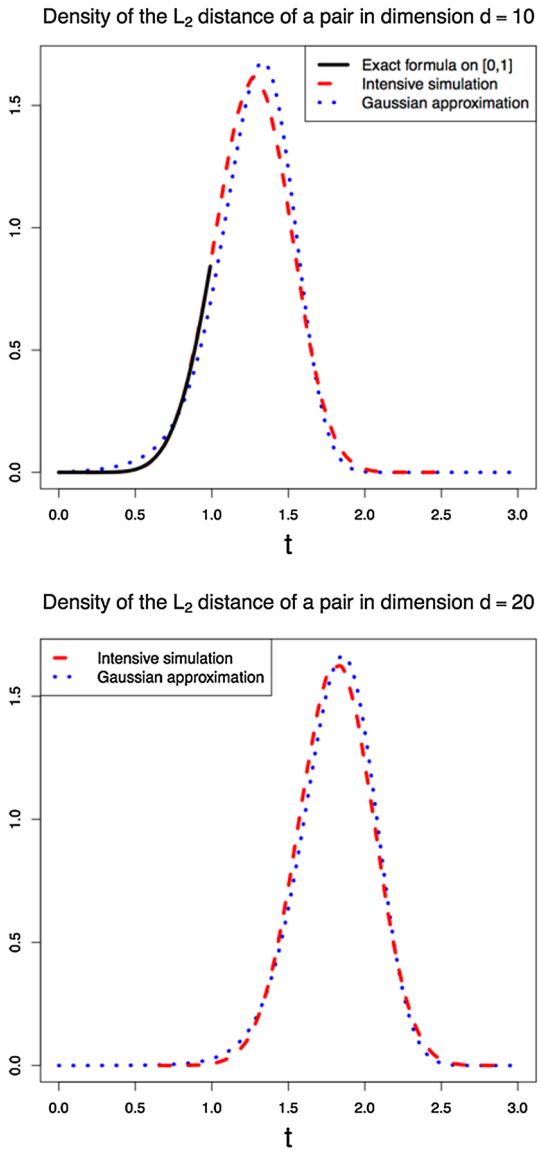

Fig. 1 Gaussian approximation in dimension $d = 10$ (*top*) and $d = 20$ (*bottom*). several plots of the density of the L_2 distance of a pair: exact expression on $[0, 1]$ from Proposition 1, estimation from an intensive simulation (10^6 repetitions) and Gaussian approximation from Proposition 3

- When one focuses on the distribution outside the interval $[0, 1]$.

4 Minimal distance between pairs of a design

The minimal distance among the set of pairs of points of a design is bounded (see Van Dam et al. 2009) and depends on both the number of points N and the dimension d. Note that in dimension $d = 1$, it is well known from the theory of uniform spacing that

$$H_{p,N,1}(t) = 1 - (1 - (N - 1)t)^N \ .$$

We refer to David (1980) for more details on this equality.

4.1 Independence approximation

In this section, we assume $d \geq 2$. The distances $\{D_{p,d}(\mathbf{X}_i, \mathbf{X}_j), 1 \leq i \neq j \leq N\}$ are clearly not independent random variables. However, one can check that the independence assumption yields a good approximation when taking the minimum. This is accurate as soon as the value of N is not too small with respect to d, e.g. $N = 30$ points in $d = 10$ dimensions.

One can thus compute the c.d.f. $H_{p,N,d}$ of the minimal distance thanks to the c.d.f. $G_{p,d}$ of the distance of a pair by

$$H_{p,N,d}(t) \underset{N\gg 1}{\sim} 1 - \left(1 - G_{p,d}(t)\right)^{N(N-1)/2} . \tag{3}$$

In terms of densities, it corresponds to

$$h_{p,N,d}(t) \underset{N\gg 1}{\sim} \widetilde{N}\, g_{p,d}(t)\, \left(1 - G_{p,d}(t)\right)^{\widetilde{N}-1} ,$$

where, for ease of notation, $\widetilde{N} = N(N-1)/2$. The formal definition (1) of the quality index is therefore simplified to

$$\mathbb{I}_p(\mathbf{x}) \underset{N\gg 1}{\sim} 1 - \left(1 - G_{p,d}(\delta_{\mathbf{x},p})\right)^{\widetilde{N}} . \tag{4}$$

To assess the quality of the approximation (3), we compare, for several values of the dimension d, the p.d.f $h_{2,100,d}$ obtained by an intensive simulation (1000 runs) and the one derived from the independence assumption. The patterns are illustrated in Fig. 2. Other number of points (but large enough) or other choice of distance would give similar

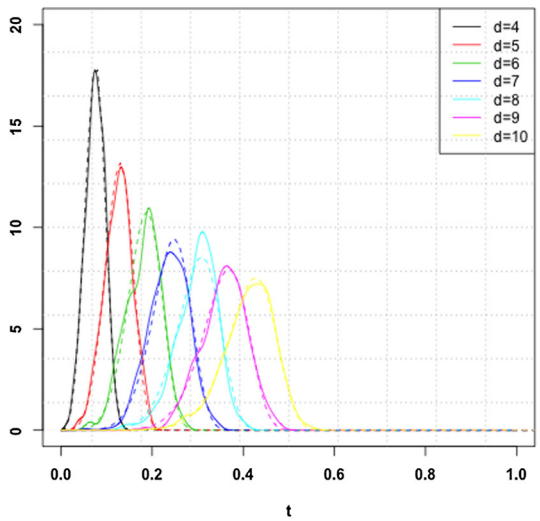

Fig. 2 Independence approximation. Density $h_{2,100,d}$ obtained from intensive simulation (*solid line*) and its approximations from Eq. (3) (*dashed line*)

results. One can see that this approximation is very accurate for any value of d: the dotted curves match very closely the solid lines. However when N is small, differences are large enough to be noticed.

4.2 Other approximations

We focus in this section on simplified expressions of the c.d.f. $H_{p,N,d}$ in specific situations: the dimension d becomes large, the size N increases, or both d and N tends to be at a high level.

4.2.1 Gumbel approximation

Note that the independence approximation (3) combined with Proposition 3 leads to

$$H_{p,N,d}(t) \simeq 1 - \left(1 - \Phi\left(\frac{t^p - d\mu_p}{\sqrt{d}\sigma_p}\right)\right)^{\widetilde{N}} . \tag{5}$$

Recall now that on the one hand the accuracy of the independence approximation (3) increases with N, whereas on the other hand, that of the Gaussian approximation (2) comes with d. Consequently, while looking at the approximation (5) one should have in mind that both N and d are large. Therefore, (5) can be approximated by

$$1 - \exp\left[-\exp\left(\alpha_{\widetilde{N}}\left\{\frac{t^p - d\mu_p}{\sqrt{d}\sigma_p} - \beta_{\widetilde{N}}\right\}\right)\right] , \tag{6}$$

with $\alpha_{\widetilde{N}} = \sqrt{2\log(\widetilde{N})}$ and

$$\beta_{\widetilde{N}} = -\alpha_{\widetilde{N}} + \frac{\left(\log\log\widetilde{N} + \log(4\pi)\right)}{2\alpha_{\widetilde{N}}} .$$

The justification comes from the fact that the minimum of independent Gaussian random variables is asymptotically Gumbel (minimum) distributed.

4.2.2 Weibull approximation

Focusing on the expansion given by (3), it is natural to wonder whether it simplifies when $G_{p,d}(t)$ is close to zero, that corresponds obviously to t close to zero. Since one should think that t stands for the minimal distance among pairs of a design, this occurs naturally when N is large with respect to d. From polynomials expressions stated in Proposition 1, $G_{p,d}(t)$ is thereby given by $a_{p,d}^{(d)} t^d$, the smallest monomial. Indeed, due to the finite precision of numbers in computers, $G_{p,d}(t)$ will numerically be reduced to its first term as soon as it is less than or equal to a relative accuracy depending on the computer. If we take the relative accuracy to be

$\epsilon = 2 \times 10^{-16}$, for usual L_2 distances it gives $t \simeq 0.05$ for $d = 10$ and $t \simeq 0.001$ for $d = 5$. This can happen when the number of points of the design N is relatively large compared to d.

Combining this first term expansion with the independence approximation (3), one gets the convergence

$$H_{p,N,d}\left(\frac{t}{\tilde{N}^{1/d}}\right) \xrightarrow[N\to\infty]{\sim} \left(1 - \exp\left(-a_{p,d}^{(d)} t^d\right)\right) \mathbf{1}_{t>0} \ ,$$

where we recall that $\tilde{N}$ denotes $N(N-1)/2$. This exponential limit is the well-known reversed Weibull distribution with shape and scale parameters respectively equal to d and $\left(a_{p,d}^{(d)}\right)^{-1/d}$.
Consequently, as soon as N is large in comparison to the value of d, the minimal distance is small, the independence approximation is reasonable and the c.d.f. is approximated by

$$H_{p,N,d}(t) \underset{0^+}{\sim} 1 - \exp\left(-a_{p,d}^{(d)} \tilde{N} t^d\right) \ . \tag{7}$$

4.2.3 Illustration on quantile estimates

Since our goal is to compare a given design to a very good one that comes from a uniform random sampling, it is important to obtain accurate estimates of high quantiles of the minimal distance distribution under uniform random sampling. Let us end this section with estimates of the quantile of order 0.99 of the distribution $H_{p,N,d}$ obtained from previous approximations. These quantiles are summarized in Table 1.

Depending on the value of N, Fig. 3 displays two cases: $N = 10$ (top) and $N = 20$ (bottom).

Table 1 Quantile summary

General formula
$H_{p,N,d}^{\leftarrow}(pr)$
Independence approximation
$G_{p,d}^{\leftarrow}(1-(1-pr)^{1/\tilde{N}})$
Independence and Gaussian approximations
$\left(d\mu_p + \sqrt{d}\sigma_p \Phi^{\leftarrow}(1-(1-pr)^{1/\tilde{N}})\right)^{1/p}$
Gumbel approximation
$\left(d\mu_p + \sqrt{d}\sigma_p \left\{\frac{1}{\alpha_{\tilde{N}}} \log(-\log(1-pr)) + \beta_{\tilde{N}}\right\}\right)^{1/p}$
Weibull approximation
$\left(-\frac{\log(1-pr)}{a_{p,d}^{(d)}\tilde{N}}\right)^{1/d}$

Computing the quantile of order pr of $H_{p,N,d}$ with previous approximations

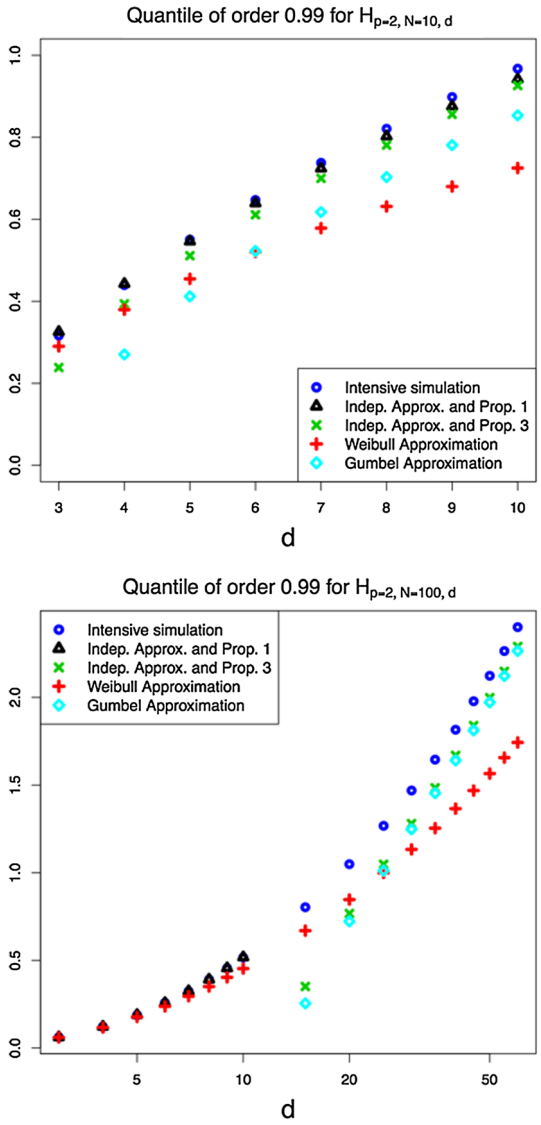

Fig. 3 Other approximations. For the size $N = 10$ (*top*) and $N = 100$ (*bottom*) of a design, quantile of order 0.99 of the minimal distance distribution is given as a function of the dimension: estimated from an intensive simulation, deduced from (3), computed from (5), (6) or (7)

In these figures, five estimates of the quantile of order 0.99 of the minimal distance distribution are available. On both figures, these estimations are plotted as functions of the dimension d. The intensive simulation result might be considered as the true value. When available, the value obtained from (3) and Proposition 1, called in the legend "Indep. Approx. and Prop. 1" is always the best estimate. On these graphs, it has been computed until $d = 10$ only. Most of the estimates from approximation (5), referred as "Indep. Approx. and Prop. 3", increase in precision with the value of d. This was also expected since it is the mixture of two accurate approximations, as already mentionned in Sects. 3.2 and 4.1 respectively. For large N, the corresponding quantiles do

not appear when d is too small. Under L_2 distance and for $N = 100$, the computation of quantiles based on (5) gives positive estimates only from $d = 14$.

It is understandable that Weibull approximation is not informative for $N = 10$. Indeed, to get small minimal distance among pairs of a design, N cannot be of the same order than d. Consequently, the main message is again that when d is small in comparison to N, the approximation is accurate. This explains why the Weibull approximation is very informative on the left part of the bottom figure.

Finally, Gumbel estimates appear more and more precise with large N and d, as suggested by the theory.

5 Numerical simulations

Our index for space-filling designs allows comparisons between different designs and can be used in two ways: first, it gives a standardized evaluation of quality, independent of the number of points and of the dimension of the input space; secondly, as a probability, it can be interpreted as a cost: if the design had been randomly generated, how many simulations would have been needed to obtain the same minimal distant? In this section, following this second point of view, we discuss the relevance of optimization procedure for design construction and compare maximin-based optimization methods with uniform random sampling. Indeed, optimization procedures are often time consuming and difficult to tune, whereas random samplings are very simple to put into practice.

In a first part, we seek reference values of the index above which the design should ensure user's satisfaction: designs of various index values are compared through different space-filling criteria. In a second part, we show that in case of high number of points and high dimensions, it can be more efficient to generate uniform random samplings until the desired index is observed rather than to run an optimization routine.

5.1 Designs comparison for different indices

Let us first visualize several designs with different indices in order to give an idea of the link between indices and patterns. For illustration purposes, four designs with 25 points in dimension 2 are plotted in Fig. 4. The top-left design with the lowest index ($\mathbb{I}_2 = 1-10^{-1}$) is deficient since it contains several empty areas. The top-right design that corresponds to $\mathbb{I}_2 = 1-10^{-5}$ is better since it spreads out the points. It is as well the case of the Sobol low-discrepancy sequence (see Niederreiter 1987) drawn at bottom-left of Fig. 4. The latter has an index equal to $\mathbb{I}_2 = 1-10^{-3.3}$. The 5×5 grid, plotted at bottom-right, corresponds to the optimal maximin design ($\mathbb{I}_2 > 1-10^{-16}$). Compared to the others, it is highly structured (see Pronzato and Muller 2012 and references therein for maximin designs in dimension 2 and 3).

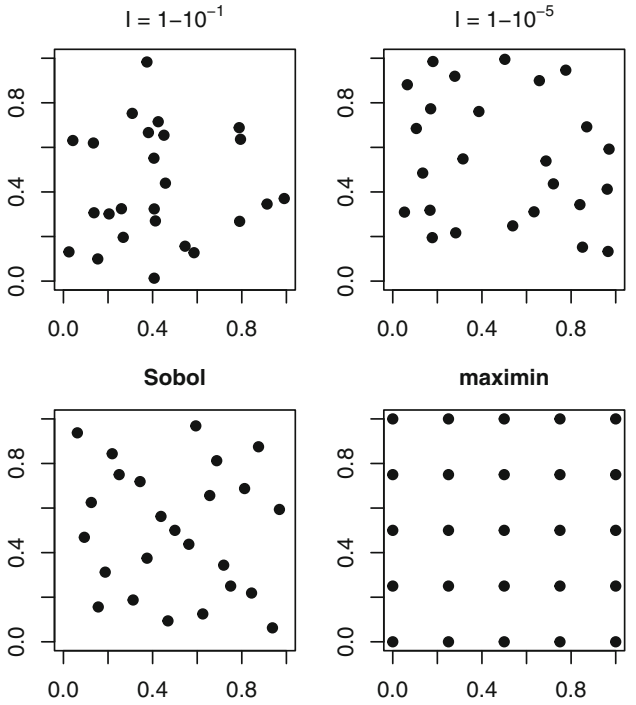

Fig. 4 Index influence on patterns. Designs on $[0, 1]^2$ with $N = 25$ points: $\mathbb{I}_2 = 1-10^{-1}$ (*top left*), $\mathbb{I}_2 = 1-10^{-5}$ (*top right*), Sobol sequence with $\mathbb{I}_2 = 1-10^{-3.3}$ (*bottom left*) and maximin optimal design with $\mathbb{I}_2 > 1-10^{-16}$ (*bottom right*)

Table 2 Index influence on criteria

	$\mathbb{I}_2$	*mindist* Mean (SD)	*discr* Mean (SD)
	$1-10^{-1}$	0.05 (10^{-5})	0.13 (0.03)
	$1-10^{-3}$	0.08 (10^{-4})	0.10 (0.02)
	$1-10^{-5}$	0.11 (10^{-3})	0.09 (0.02)
Sobol	$1-10^{-3.3}$	0.09	0.04
Maximin	$>1-10^{-16}$	0.25	0.14

Comparison of two criteria for designs with 25 points in dimension 2

Now, let us again consider designs with several values of the index. In order to collect designs with $\mathbb{I}_2$ say in $\{1-10^{-1}, 1-10^{-3}, 1-10^{-5}\}$, 10^6 simulations of uniform random samplings have been run, and empirical quantiles of the minimal distance guides the selection. We propose to evaluate on these designs, the two most frequently used criteria: the minimal distance among pairs and the centered discrepancy Fang et al. (2005). Note that discrepancy criteria measures the distance between empirical and uniform distributions. On each design that has been selected, the previous two criteria have been computed. The distribution of these measures are summarized in Table 2 (see column *mindist* and *discr* respectively). The mean and standard deviation are based on 20 repetitions of the experiment.

Let us now analyze the results from Table 2:

– As expected, *mindist* increases with the quality index whereas *discr* slowly decreases. It can be noticed that the distributions of *discr* overlap each other, the standard deviation *sd* remains at a high level whatever the value of the index.
– The maximin optimal design (the 5×5 grid) has the highest index of quality (numerically equal to 1) but does not seem appropriate. It has too numerous points at the edges, and several alignments. This is reflected by a high discrepancy.
– On the contrary, the Sobol sequence whose index is not very high ($1 - 10^{-3.3}$) has a good covering (very low discrepancy).
– It can be noticed that, because we invert the empirical c.d.f., standard deviation for *mindist* is not null but very small.

As a conclusion, it can be sufficient to look for a design with a high value of the quality index, say between $1 - 10^{-3}$ and $1 - 10^{-5}$, but not extremely high.

5.2 Efficiency of optimization methods versus uniform random sampling

In this context, what is the most efficient procedure: running an optimization routine called *strategy 1*, or generating uniform random samplings called *strategy 2*? It is a fair question because optimization routine are sometimes difficult to parametrize and to handle for the user.

We address this issue through numerical simulations. The objective is to compare the distribution of the necessary number of iterations under *strategy 1* and *strategy 2* to achieve the same desired quality.

Let N_{opt} be the number of iterations under *strategy 1* (optimization approach) that is required to produce such a design. Note that we consider an optimization routine based on simulated annealing which is famous in the context of computer experiments (see Jin et al. 2005; Auffray et al. 2012; Damblin et al. 2013). The result of the optimization routine is random. Randomness comes from: (i) the initial configuration, (ii) the exchange procedure and (iii) the reject or acceptation of the new design. The point (ii) is now described in more details: at each step of the algorithm, one point of the design is randomly chosen and its exchange with another randomly-chosen point of the domain is examined. If this exchange involves an improvement, it is accepted. Otherwise, it can be accepted with a non null probability that slowly decreases to zero.

The empirical c.d.f. of N_{opt} obtained for different dimensions or numbers of points or indices are displayed on Fig. 5. As expected, the number of iterations increases with $\mathbb{I}$ (green compared to black). It can also be seen that for the same index of quality $\mathbb{I} = \mathbb{I}_2 = 1 - 10^{-1}$, the number of iterations

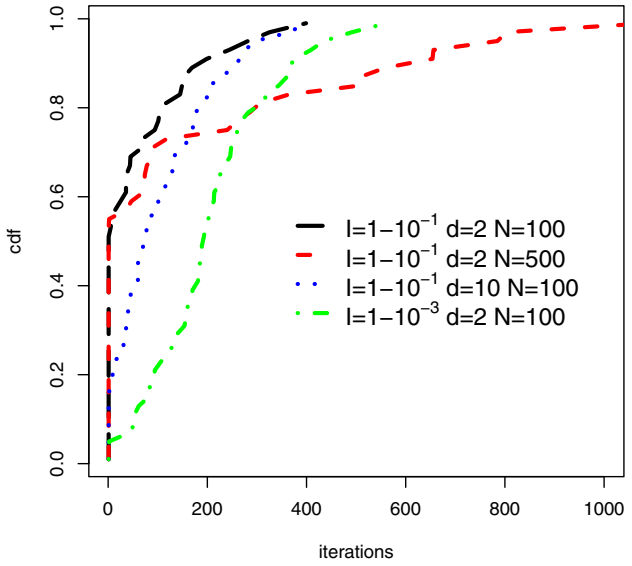

Fig. 5 Optimization efficiency. Influence of the quality index $\mathbb{I} \in \{1 - 10^{-1}, 1 - 10^{-3}\}$, the dimension $d \in \{2, 10\}$ and the number of points $N \in \{100, 500\}$ on the distribution of the number of iterations under *strategy 1*

increases with the dimension and the number of points (blue and red compared to black). Indeed, the problem consists in optimizing the d coordinates of the N points of the design.

To formalize the problem under *strategy 2*, let us write the index as $\mathbb{I} = 1 - 10^{-\mathbb{I}'}$ for some positive real $\mathbb{I}'$ and denote by $\delta_{\mathbb{I}}$ the corresponding minimal distance among pairs of such design. Let also N_{rnd} be the number of uniform random samplings that must be generated before one has a minimal distance higher than $\delta_{\mathbb{I}}$. The theoretical probability distribution of N_{rnd} is a geometric law of parameter $1 - \mathbb{I}$. The mean of the distribution is then $10^{\mathbb{I}'}$ and the quantile of order $1 - \alpha$ is

$$N_{rnd,1-\alpha} = \frac{\log(\alpha)}{\log(1 - 10^{-\mathbb{I}'})} .$$

In other words, the probability of waiting more than $N_{rnd,1-\alpha}$ simulations before observing a design of index $1 - 10^{-\mathbb{I}'}$ is α. Note that this probability is independent of both the dimension d and the number of points of the design N.

It can be seen in Fig. 6 that the number of iterations, needed to achieve the requested quality, is increasing with $\mathbb{I}$ (black corresponds to $\mathbb{I} = 1 - 10^{-1}$ and purple to $\mathbb{I} = 1 - 10^{-3}$), whatever the strategy (dashed line for *strategy 1*, referred to as optim and solid line for *strategy 2*, named random). One can see that *strategy 2* performs better than *strategy 1* for $\mathbb{I} = 1 - 10^{-1}$ (weak quality requirement). In this case, the expectation of N_{rnd} is 10 whereas the number of iterations of the optimization routine is distributed between 0 and 500 for $N = 100$. When $\mathbb{I} = 1 - 10^{-3}$, it is the opposite: the c.d.f. of N_{opt} reaches the value 1 faster than the geomet-

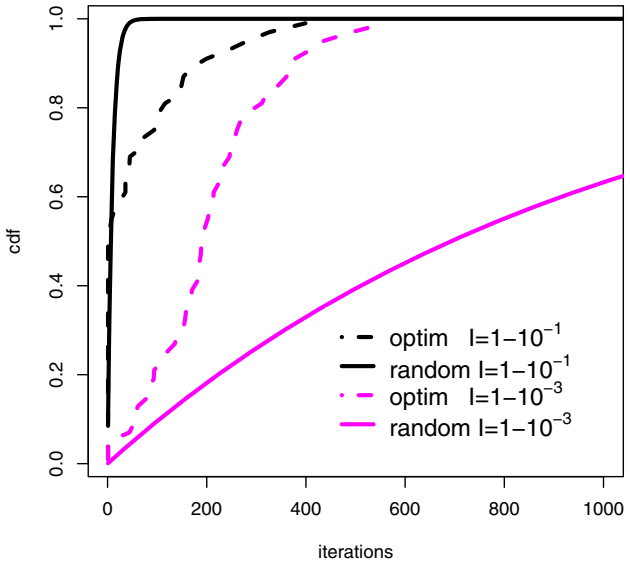

Fig. 6 Optimization versus random. Distributions of the number of optimization iterations (*dashed line*) and uniform random sampling iterations (*solid line*) for $\mathbb{I} \in \{1-10^{-1}, 1-10^{-3}\}$, with $d = 2$ and $N = 100$

ric distribution. Note that the comparison is less favorable when the dimension or the number of points is high, i.e. when the optimization problem is more complex (cf. Fig. 5). When $\mathbb{I} = 1 - 10^{-5}$, the use of an optimization method is inescapable: the expectation of the corresponding law for N_{rnd} being extremely large, equal to 10^5. This case is not represented on Fig. 6.

In conclusion, if the requested quality of the space-filling design is not too high, it is not efficient to run an optimization procedure, especially when the optimization problem is difficult (high dimension or high number of points).

5.2.1 Additional comment

Different routines exist in R, Matlab, Python etc. to generate maximin-designs. One should pay attention to the related strategy of each of them. For example in R, *maximinlhs* from package *lhs* Carnell (2012) corresponds exactly to *strategy 2* whereas *maximinSA_LHS* (see Jin et al. 2005; Damblin et al. 2013) from package *DiceDesign* Franco et al. (2015) corresponds to *strategy 1*. Note that these routines have been implemented in conjunction with Latin Hypercube Sampling, a well known method since Mckay et al. (1979), which ensures uniform distribution of the margins but has no space-filling qualities.

6 Real example

Some of the major challenges for the automotive industry are the reduction of greenhouse gas emissions, fossil fuel dependency and local pollution. In many cases, phenomenological models are not predictive enough for approximating these responses. This is why these objectives rely on what is called engine calibration, which consists in two steps. In the first one, the optimal tuning of parameters used by engine control strategies, for example for achieving low consumption or low pollution is determining experimentally. Indeed, in order to get a usable description of the engine under study, automotive industrials launch experimental studies on test benches. In the second step, the obtained experimental results are approximated by polynomials in very simple cases, and by Gaussian surrogate models otherwise (see Santner et al. 2003; Sacks et al. 1989). In this work, we are mainly interested by the first part of this procedure.

Due to their increasing number, manual tuning of engine parameters is now replaced by mathematically assisted calibration process, that is based on design of experiments. The dimension of the input space for parameters is in our example $d = 11$, among which we can find injection engine speed, load, air flow rate, injection parameters, as described in Magand et al. (2011). To avoid unnecessary complications, all the parameters are supposed to vary between 0 and 1.

To prevent the engine from going in forbidden regions, some constraints (linear and non linear) have to be added, and the final experimental area is not any longer hypercubic. During the calibration process, while exploring the physical limits of the engine, dependencies between the couple of input speed and load and the other parameters are formalized. They mostly result in linear constraints. But in some cases, these relations can be too simple. For example, the limits of the air loop parameter depend on injection parameters values as well as speed and load. Some of these limits are modeled with low degree polynomials, others with non linear functions (see Magand et al. 2011). Eventually, more than 80 constraints have to be considered in this example.

Design in constrained domains is still an active domain of research. Stinstra et al. (2003) and more recently Auffray et al. (2012) propose performing strategies in constrained regions based on a simulated annealing algorithm. Petelet et al. (2010) consider the case where linear constraints passing through the origin restrict the available domain.

Here we propose the following strategy to compute the index in a constrained domain. To calculate the volume of the constrained region by Monte-Carlo simulations, we would simply count the number N of points falling in the targeted region out of a total number of draws N_{tot}, and the estimated volume would be $V = N/N_{tot}$. Conversely, N points uniformly drawn in the experimental area should correspond to N/V points in the unit hypercube, where V is the volume of the constrained region. Then if $\delta_{\mathbf{x}}$ is the minimal distance obtained for the points in the experimental region, the probability that a uniform random design has a minimal distance larger than $\delta_{\mathbf{x}}$ in the unit hypercube is approximated by

$$(1 - G_{p,d}(\delta_{\mathbf{x}}))^{\frac{N/V(N/V-1)}{2}}$$

where $G_{p,d}$ has been defined in Sect. 3.

Under L_2 distance In our real example, we have run a large number of simulations of uniform random sampling designs and counted the points falling in the targeted constrained region. We estimate the volume of this constrained region by $V = 0.23$. With $N = 250$ points, the minimum L_2 distance between pairs of such design is around $\delta_{\mathbf{x}} = 0.33$, corresponding to an index $\mathbb{I} = 1 - 10^{-1.06}$ calculated with an equivalent number of points equal to N/V. As previously done by Jin et al. (2005) or Petelet et al. (2010), this first initial design can be improved by exchanging coordinates. After this procedure, the minimal distance becomes $\delta_{\mathbf{x}} = 0.395$ and the index equals $1 - 10^{-6}$. As the cumulative distribution function is very steep, this small change in the distance has an important impact on the index. Moreover, due to its relatively high value, the user can now be confident about the success of the procedure.

Under L_1 distance The same calculations can be done with L_1 distance for maximin designed as in Husslage (2006). We find $\delta_{\mathbf{x}} = 0.52$ in the initial design and 0.98 after exchanging coordinates. These values correspond to an index of $0.0137 = 1 - 10^{-0.006}$ for the initial design and $1 - 10^{-4}$ for the modified one. Note that these calculations are only indicative here, since maximin designs depend on the chosen distance.

In this section we have illustrated the use of the proposed index on a real example. We have also shown how to deal with difficulties occurring in concrete situations, such as constraints limiting the experimental domain.

7 Conclusion

In this paper, we propose a new index to measure the quality of space-filling designs, based on the probability distribution of the minimal distance of N points drawn uniformly in the unit hypercube of dimension d. As a probability measure, our index is normalized, and can thus be interpreted independently from both the dimension and the number of points. Its evaluation requires the knowledge of a c.d.f. that can be very well approximated by polynomial forms in most common situations.

However, when the number of points in the design is large or under high dimension, some adequate approximations should be preferred, for their simplicity, in order to evaluate the index. To sum it all up, we gather in Table 3 the different ways of evaluating the index. Therein, the variable t plays the role of the minimal L_p distance $\delta_{p,\mathbf{x}}$ among the set of pairs of points of an observed design $\mathbf{x}$.

Table 3 Index summary

General formula
$H_{p,N,d}(t)$
Independence approximation
$1 - (1 - G_{p,d}(t))^{\widetilde{N}}$
Independence and Gaussian approximations
$1 - \left(1 - \Phi\left(\frac{t^p - d\mu_p}{\sqrt{d}\sigma_p}\right)\right)^{\widetilde{N}}$
Gumbel approximation
$1 - \exp\left[-\exp\left(\alpha_{\widetilde{N}}\left\{\frac{t^p - d\mu_p}{\sqrt{d}\sigma_p} - \beta_{\widetilde{N}}\right\}\right)\right]$
Weibull approximation
$1 - \exp\left(-a_{p,d}^d \widetilde{N} t^d\right)$

Computing the value of the index $\mathbb{I}_p$, when the minimal distance is t, with several approximations provided in Sect. 4

The major conclusion of the simulation study is that, to achieve particular values of our index, it is not always necessary to run an optimization procedure. This depends on how large the desired index is and how complex the optimization procedure is.

The application of our index to an automotive industry example reveals that our tools can be easily adapted to real experimental domains.

Acknowledgments The authors are grateful to the Associate Editor and the anonymous referees for helpful suggestions which helped to greatly improve the initial text. Luc Pronzato deserves thanks for careful reading and interesting ideas on the occasion of a first version of the paper. Special thanks also go to John P. Nolan for his help on correcting our English.

References

Auffray, Y., Barbillon, P., Marin, J.: Constrained maximin designs for computer experiments. Stat. Comput. **22**, 703–712 (2012)

Carnell, R.: Latin Hypercube Sample. R package version 0.10 (2012). http://CRAN.R-project.org/package=lhs

Damblin, G., Couplet, M., Iooss, B.: Numerical studies of space-filling designs: optimization of Latin Hypercube Samples and subprojection properties. J. Simul. **7**, 276–289 (2013)

David, H.: Order Statistics. Wiley, New York (1980)

Fang, K., Li, R., Sudjianto, A.: Design and Modeling for Computer Experiments. Chapman & Hall, London (2005)

Franco, J., Dupuy, D., Roustant, O.: DiceDesign: Design of computer experiments. R package version 1.7 (2015). http://CRAN.R-project.org/package=DiceDesign

Husslage, B.: Maximin designs for computer experiments. PhD thesis Universiteit van Tilburg (2006)

Jin, R., Chen, W., Sudjianto, A.: An efficient algorithm for constructing optimal design of computer experiments. J. Stat. Plan. Inf. **134**, 268–287 (2005)

Magand, S., Pidol, L., Chaudoye, F., Sinoquet, D., Wahl, F., Castagne, M., Lecointe, B.: Use of ethanol/diesel blend and advanced calibration methods to satisfy Euro 5 emission standards without a DPF. Oil Gas Sci. Technol. **66**, 855–875 (2011)

Mckay, M.D., Beckman, R.J., Conover, W.J.: Constrained maximin designs for computer experiments. Technometrics **21**, 239–245 (1979)

Niederreiter, H.: Low-discrepancy and low-dispersion sequences. J. Number Theory **30**, 51–70 (1987)

Petelet, M., Iooss, B., Asserin, O., Loredo, A.: Latin hypercube sampling with inequality constraints. Adv. Stat. Anal. **3**, 11–21 (2010)

Philip, J.: The probability distribution of the distance between two random points in a box. TRITA MAT 10 (2007)

Philip, J.: The distance between two random points in a 4- and 5-cube. J. Chemom. **21**, 198–207 (2010)

Pronzato, L., Muller, G.: Design of computer experiments: space-filling and beyond. Stat. Comput. **23**, 681–701 (2012)

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P.: Design and analysis of computer experiments. Stat. Sci. **4**, 409–442 (1989)

Santner, T., Williams, B., Notz, W.: The Design and Analysis of Computer Experiments. Springer, New York (2003)

Stinstra, E., den Hertog, D., Stehouwer, P., Vestjens, A.: Constrained maximin designs for computer experiments. Oper. Res. **57**, 595–608 (2003)

Van Dam, E., Rennen, G., Husslage, B.: Bounds for maximin latin hypercube designs. Oper. Res. **57**, 595–608 (2009)