Journal de la Société Française de Statistique
Vol. 158 No. 1 (2017)

A repulsion-based method for the definition and the enrichment of optimized space filling designs in constrained input spaces

Titre: Définition et enrichissement de plans d'expériences optimisés dans des domaines contraints à partir d'une méthode de répulsion

Guillaume Perrin[1] and Claire Cannamela[1]

Abstract:

Due to increasing available computational resources and to a series of breakthroughs in the solving of nonlinear equations and in the modeling of complex mechanical systems, simulation nowadays becomes more and more predictive. Methods that could quantify the uncertainties associated with the results of the simulation are therefore needed to complete these predictions and widen the possibilities of simulation. One key step of these methods is the exploration of the whole space of the input variables, especially when the computational cost associated with one run of the simulation is high, and when there exists constraints on the inputs, such that the input space cannot be transformed into a hypercube through a bijection. In this context, the present work proposes an adaptive method to generate initial designs of experiments in any bounded convex input space, which are distributed as uniformly as possible on their definition space, while preserving good projection properties for each scalar input. Finally, it will be shown how this method can be used to add new elements to an initial design of experiments while preserving very interesting space filling properties.

Résumé :

Profitant de l'essor considérable des puissances de calcul disponibles et de progrès importants en modélisation des phénomènes physiques, le rôle de la simulation n'est plus seulement descriptif, mais prédictif. Pour garantir cette capacité prédictive, il est alors nécessaire de développer des méthodes permettant d'associer à tout résultat numérique une précision, qui intègre les différentes sources d'incertitudes. Un des enjeux de ces méthodes de quantification des incertitudes concerne l'optimisation de l'exploration du domaine de variation des entrées de modélisation. Cette tâche peut s'avèrer difficile, en particulier lorsque le coût numérique associé à une simulation est élevé, ou lorsque le domaine d'entrée présente un certain nombre de contraintes, si bien qu'il ne peut plus être transformé en un hypercube via une bijection. Dans ce contexte, ce travail présente une méthode basée sur des répulsions permettant la définition de plans d'expériences optimisés dans des domaines contraints, dont les projections sur chaque paramètre d'entrée présentent de bonnes propriétés statistiques. Enfin, on montre que cette méthode permet également l'enrichissement de plans d'expériences déjà définis, tout en préservant un bon remplissage global du domaine de définition des entrées.

Keywords: optimal design, design of experiments, computer experiment, Latin Hypercube Sampling, surrogate model

Mots-clés : plans d'expériences optimisés, simulation numérique, plans hypercubes latins, metamodèle

AMS 2000 subject classifications: 62K05, 62K20, 62L05, 05B15

[1] CEA/DAM/DIF, F-91297, Arpajon, France.
E-mail: guillaume.perrin2@cea.fr; claire.cannamela@cea.fr

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

1. Introduction

For the last decades, the use of simulation has kept increasing for the analysis of always more complex physical phenomena. As these simulations can be very expensive to run, surrogate models, such as polynomials, neural networks (Bishop, 1995; Bhadeshia, 1999), support vector machines (Vapnik, 1998; Scholkopf and Smola, 2002; Hortado, 2002), kernel interpolation methods (Sacks et al., 1989; Santner et al., 2003), can be associated with these computer codes to maximize the knowledge about these physical phenomena. Indeed, built from a limited number of calls to the expensive computer code, which are generally referred as the *initial design of experiments* (DoE), surrogate models are a very interesting tool to perform uncertainty or sensitivity analyses.

To be more precise, for $d > 0$, let $f : E \mapsto \mathbb{R}$ be a computer code function, which inputs have been normalized, such that E is a compact subset of $[0,1]^d$. In addition, for $N \geq 2$, $1 \leq n \leq N$, and $\mathbf{x}^{(n)} \in E$, let $[x] = [\mathbf{x}^{(1)} \; \cdots \; \mathbf{x}^{(N)}]$ be an initial DoE where function f is to be evaluated. The set of code evaluations, $\left\{f(\mathbf{x}^{(n)}),\; 1 \leq n \leq N\right\}$ defines therefore the maximal available information for the construction of a surrogate model, and it is well known that the relevance of such a surrogate models is strongly dependent on the capability of $[x]$ to explore space E.

In that prospect, the elements of the initial DoE have to be distributed as uniformly as possible on their definition space, in order to capture the potential non-linearity of the computer code. Two kinds of quantities are generally introduced to quantify the good coverage of the input space by the initial DoE: statistical criteria based on some measurements of uniformity (also called discrepancies, see Fang and Lin, 2003; Fang et al., 2006) and geometrical criteria based on the classical Euclidian norm. Whereas the statistical criteria are generally associated with the case when the inputs are independent, such that E is the d-dimensional hypercube, the geometrical criteria can be used in constrained input spaces. In particular, the minimax criterion (Johnson et al., 1990), $C^{mM}([x])$, such that:

$$C^{\mathrm{mM}}([x]) = \max_{\mathbf{x} \in E} \left\{ \min_{1 \leq n \leq N} \left\| \mathbf{x}^{(n)} - \mathbf{x} \right\|_d \right\}, \tag{1}$$

where $\|\cdot\|_d$ is the Euclidian norm in $\mathbb{R}^d$, is widely used as a robust evaluation of the exploratory properties of $[x]$. The smaller is this criterion, and the more relevant $[x]$ is likely to be. If kernel interpolation methods (such as the Gaussian Process Regression of Kriging) are considered, the use of this criterion is also justified by the fact that this minimax criterion appears explicitly in an upper bound (that increases with respect to $C^{\mathrm{mM}}([x])$) on the pointwise error between f and its kernel approximation given $[x]$ (Schaback, 1995). In the same manner, minimax designs have general asymptotic optimality properties for the D- and G-optimality criteria (Johnson et al., 1990).

Nevertheless, despite its wide use, very little algorithmic developments have been carried out to directly compute minimax designs. This can be explained by the fact that its evaluation requires the identification of the supremum of a non-convex expression over an infinite set, which becomes too costly when d increases. In that context, the mindist criterion,

$$C^{\mathrm{mindist}}([x]) = \min_{1 \leq m \neq n \leq N} \left\| \mathbf{x}^{(n)} - \mathbf{x}^{(m)} \right\|_d, \tag{2}$$

proposes an interesting alternative to $C^{\mathrm{mM}}([x])$. Indeed, it is also an indicator of the good coverage

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

of E: given two DoE $[x^{(1)}]$ and $[x^{(2)}]$, if $C^{\text{mindist}}([x^{(1)}]) \geq C^{\text{mindist}}([x^{(2)}])$, $[x^{(1)}]$ is likely to better explore E than $[x^{(2)}]$, as the distance between at least two elements of $[x^{(2)}]$ is smaller than the distances between each pair of elements of $[x^{(1)}]$. Moreover, its computation is much faster than the one of $C^{\text{mM}}([x])$ and it can be shown (Auffray et al., 2012) that if $[x^*]$ maximizes C^{mindist}, then $C^{\text{mM}}([x^*]) \leq C^{\text{mindist}}([x^*])$. Hence, the maximization of the mindist criterion can also be used to upper bound the pointwise distance between f and its kernel-based interpolation.

Considering that, in many applications, only a small number of inputs have a real influence on the studied quantities of interest, proposing initial DoE that present good projection properties in low-dimensional spaces (sparsity principle) is also of great interest. Focusing on the formerly introduced mindist criterion, this means that not only $C^{\text{mindist}}([x])$ has to be high, but $C^{\text{mindist}}([x^{\text{sub},p}])$ has also to be high for each matrix $[x^{\text{sub},p}]$ that gathers p lines of $[x]$, with $1 \leq p \leq d$. When only the one-dimensional projection properties are considered ($p = 1$), a relevant way to build space-filling designs is to optimize a statistical or geometrical criterion within the well known class of Latin Hypercube Samples noted LHS (McKay et al., 1979). Very interesting developements can be found in Morris and Mitchell (1995) and Jin et al. (2005) for this purpose. Some authors have generalized these works to improve the space-filling properties of projections in higher dimensions than one (see Tang, 1993, Park, 1994,Osuolale et al., 2003, Moon et al., 2011). Recently, Joseph et al. (2015) have proposed a method to optimize the space-filling properties on all sub-projections spaces. Such designs are called maximum projection designs and will be referred as MaxPro designs in the following. In the same manner, a greedy approach has been proposed in Draguljić et al. (2012) to iteratively compute designs that have good space filling properties in lower dimensional projections (named non-collapsing designs).

The majority of the literature on the construction of space-filling designs considers input spaces that are hyperrectangles. However, we can find some contributions to the construction of designs for a bounded convex domain. Stinstra et al. (2003) have used non-linear programming solvers to construct maximin designs in explicit constrained regions. Auffray et al. (2012) have proposed maximin designs in any bounded connected domain based on simulated annealing scheme with a particular Metropolis-within-Gibbs algorithm. Stinstra et al. (2010) have developed a nearly uniform design in any convex space. The approach of Draguljić et al. (2012) can also be applied in this input space type. Interesting results for bounded convex input regions can also be obtained using iterative clustering-based approaches. To this end, Flury (1990) has used the k-means clustering whereas Lekivetz and Jones (2015) have proposed a hierarchical clustering. Recently, these methods have been improved to better focus on the minimax minimization (Mak and Joseph, 2016).

However, in most of these methods adapted to non-hyperrectangle input spaces, there is no guarantee to get good space filling properties in lower dimensional projections. Hence, this work aims at generalizing the notion of optimized LHS to the case when there exists constraints on the inputs such that the input space is a bounded convex domain (the simplex or the hypersphere are two examples of such constrained input spaces). Following on the works of Dupuy et al. (2015) and Franco et al. (2008), the proposed method is based on an adaptive procedure that combines several kinds of repulsion between each elements of the DoE. The main interest of our repulsion-based procedure is to produce DoE that allow the exploration of the whole input space, which can be constrained, without accumulating points at the boundaries, with a relatively low complexity.

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

Section 2 presents the main theoretical and numerical arguments on which the repulsions are based, and compares the efficiency of the method to already existing DoE generators when the input space is a hypercube. Section 3 generalizes these concepts to the case when the input space is any bounded convex domain. In particular, the generation of DoE in the multidimensional simplex and hypersphere is analyzed in details. At last, Section 4 shows to what extent such a repulsion-based approach allows the addition of new elements to already existing DoE while preserving global good space filling properties.

2. A repulsion-based method to construct optimized LHS

Let d be the number of scalar inputs involved in the considered computer code, and N be the number of experiments that we want to generate to build an initial DoE. We then denote by $\mathscr{M}_{d,N}([0,1])$ the set of all the $(d \times N)$-dimensional matrices with values in $[0,1]$. Hence, we denote by $\mathscr{H}_{d,N}^{\mathrm{LHS}} \subset \mathscr{M}_{d,N}([0,1])$ the set of all the $(d \times N)$-dimensional Latin Hypercube Samples (LHS) $[x]$, such that each row of $[x]$ has one component in each interval $[(n-1)/N, n/N]$, $1 \leq n \leq N$. It is assumed that the good coverage of $[0,1]^d$ by any element $[x]$ in $\mathscr{H}_{d,N}^{\mathrm{LHS}}$ can be assessed through an easy-to-compute (geometrical or statistical) space filling criterion, $\mathscr{C}$, to be minimized. Hence, if $[x^{(1)}]$ and $[x^{(2)}]$ are two elements of $\mathscr{H}_{d,N}^{\mathrm{LHS}}$, $[x^{(1)}]$ will be said to be better that $[x^{(2)}]$ to cover $[0,1]^d$ if $\mathscr{C}([x^{(1)}]) \leq \mathscr{C}([x^{(2)}])$.

Therefore, the idea of this section is to propose an adaptive method to search $[x^\star]$ as one solution of the following optimization problem:

$$[x^\star] = \arg \min_{[x] \in \mathscr{H}_{d,N}^{\mathrm{LHS}}} \mathscr{C}([x]). \tag{3}$$

Notations

We denote by $\langle \cdot, \cdot \rangle_N$ and $\|\cdot\|_N$ the classical Euclidean inner product and norm respectively, such that for all $\mathbf{x}$ and $\mathbf{y}$ in $[0,1]^N$,

$$\langle \mathbf{x}, \mathbf{y} \rangle_N = \sum_{n=1}^{N} x_n y_n, \quad \|\mathbf{x}\|_N = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_N}. \tag{4}$$

For all $\mathbf{x}$ in $[0,1]^N$, we then define $\mathbf{s}(\mathbf{x})$ the vector gathering the elements of $\mathbf{x}$, which have been sorted in an ascending order, and $\{j_1(\mathbf{x}), \ldots, j_N(\mathbf{x})\}$ the corresponding indices, such that:

$$s_1(\mathbf{x}) = x_{j_1(\mathbf{x})} < \ldots < s_N(\mathbf{x}) = x_{j_N(\mathbf{x})}. \tag{5}$$

In addition, the N columns of any element $[x]$ in $\mathscr{M}_{d,N}([0,1])$ will be written $\left\{\mathbf{x}^{(n)},\ 1 \leq n \leq N\right\}$, whereas its d rows will be written $\left\{\mathbf{x}_{(i)},\ 1 \leq i \leq d\right\}$, such that:

$$[x] = [\mathbf{x}_{(1)} \ ; \ \ldots \ ; \ \mathbf{x}_{(d)}] \ = \ [\mathbf{x}^{(1)} \ \cdots \ \mathbf{x}^{(N)}]. \tag{6}$$

Therefore, for all $1 \leq i \leq d$ and all $1 \leq n \leq N$, the element that corresponds to the i^{th} row and the n^{th} column can be written:

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

$$[x]_{i,n} \; = \; x_{(i),n} \; = \; x_i^{(n)}. \tag{7}$$

2.1. *Optimization of the space filling properties of LHS in the hypercube*

By construction, if $[x] = [\widehat{x}] + [\delta x]$, where, for all $1 \leq i \leq d$ and $1 \leq n \leq N$, $[\widehat{x}]_{i,n} = (n-0.5)/N$ and $[\delta x]_{i,n}$ is uniformly chosen in $[-0.5/N, 0.5/N]$, then $[x]$ is in $\mathscr{H}_{d,N}^{\text{LHS}}$. Moreover, it can be noticed that permuting several elements in each row of $[x]$ leads to another element of $\mathscr{H}_{d,N}^{\text{LHS}}$. Hence, as generating LHS is relatively easy, searching a "good" (with respect to criterion $\mathscr{C}$) LHS can be done by searching the best LHS among a large number of DoE that have been drawn at random within the class of Latin Hypercube arrangements. However, noticing that $N!$ elements of $\mathscr{H}_{d,N}^{\text{LHS}}$ can be computed for each draft of $[\delta x]$, it appears that such a procedure is rather inefficient, which has encouraged the development of adaptive methods to optimize the space filling properties of an initial element of $\mathscr{H}_{d,N}^{\text{LHS}}$ from a series of well-chosen permutations (see Damblin et al., 2013 for more details about these methods).

The problem with such methods based on permutations is that they are limited to the hypercube case. Hence, another approach is presented in this section, which is based on the introduction of two repulsions: first, a component by component repulsion to project any $(d \times N)$-dimensional matrices with values in $[0,1]$ in $\mathscr{H}_{d,N}^{\text{LHS}}$, and second, a repulsion based on the Euclidian distance between points to optimize the space filling properties of the DoE.

Although such an approach gives very interesting results in the case of the hypercube, it is recalled that the motivation for the following developments is their generalization to any bounded convex domains.

2.1.1. *Projection in $\mathscr{H}_{d,N}^{LHS}$*

This section aims at presenting two pseudo-projections: the first one is only based on the sorting of the rows of the elements of $\mathscr{M}_{d,N}([0,1])$; the second one introduces an iterative procedure, which will be very useful in the context of constrained input spaces.

Rank-based projection

If $[x] = \left[\mathbf{x}_{(1)} \; ; \; \dots \; ; \; \mathbf{x}_{(d)}\right]$ is an element of $\mathscr{M}_{d,N}([0,1])$, the matrix $[\Pi x]$, where for all $1 \leq i \leq d$ and for all $1 \leq n \leq N$,

$$[\Pi x]_{i,n} = (j_n(\mathbf{x}^{(i)}) - 0.5)/N, \tag{8}$$

is in $\mathscr{H}_{d,N}^{\text{LHS}}$ and can be seen as a pseudo-projection of $[x]$ on $\mathscr{H}_{d,N}^{\text{LHS}}$.

Repulsion-based projection

For all $\mathbf{x}$ in $[0,1]^N$, let $\mathscr{C}^{\text{LHS}}$ and B^{LHS} be the two following criteria:

$$\mathscr{C}^{\text{LHS}}(\mathbf{x}) = \max\left\{s_2(\mathbf{x}) - s_1(\mathbf{x}), \dots, s_N(\mathbf{x}) - s_{N-1}(\mathbf{x}), 1 - s_N(\mathbf{x}) + s_1(\mathbf{x})\right\}, \tag{9}$$

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

$$B^{\mathrm{LHS}}(\mathbf{x}) = \begin{cases} 1 \text{ if for all } 1 \leq n \leq N,\ (n-1)/N \leq s_n(\mathbf{x}) \leq n/N, \\ 0 \text{ otherwise.} \end{cases} \tag{10}$$

By construction, the minimal value of $\mathscr{C}^{\mathrm{LHS}}$ is $1/N$, which corresponds to a configuration when all the components of $\mathbf{x}$ are separated by the same distance $1/N$. Such a configuration is not unique. For all $\mathbf{x}$ in $[0,1]^N$, let us now consider the repulsion vector $\mathbf{R}(\mathbf{x}) = (R_1(\mathbf{x}), \ldots, R_N(\mathbf{x}))$, such that for all $1 \leq n \leq N$:

$$R_n(\mathbf{x}) = \sum_{1 \leq m \neq n \leq N} \widetilde{R}(x_n, x_m) + \widehat{R}(x_n, x_m), \tag{11}$$

$$\widetilde{R}(x_n, x_m) = \frac{x_n - x_m}{|x_n - x_m|^3}, \quad \widehat{R}(x_n, x_m) = \frac{x_m - x_n}{|x_n - x_m|(1 - |x_n - x_m|)^2}. \tag{12}$$

Whereas $\widetilde{R}(x_n, x_m)$ is the classical $1/r^2$ repulsion, $\widehat{R}_{nm}(x_n, x_m)$ is a symmetry-based regularization term, to take into account the fact that the components of $\mathbf{x}$ have to stay in the bounded interval $[0,1]$. In addition, let $\{\mathbf{x}_{(k)}\}_{k \geq 1}$ be the sequence defined by the algorithm 1, where, for all $\mathbf{x}$ in $[0,1]^N$:

$$\delta(\mathbf{x}) = \begin{cases} 0 \quad \text{if} \quad R_1(\mathbf{x}) = \ldots = R_N(\mathbf{x}), \\ -\langle \Delta\mathbf{x}, \Delta\mathbf{R}(\mathbf{x}) \rangle_N \,/\, \|\Delta\mathbf{R}(\mathbf{x})\|_N^2 \quad \text{otherwise, with :} \end{cases} \tag{13}$$

$$\begin{cases} \Delta\mathbf{x} = (s_2(\mathbf{x}) - s_1(\mathbf{x}), \ldots, s_N(\mathbf{x}) - s_{N-1}(\mathbf{x}), 1 - (s_N(\mathbf{x}) - s_1(\mathbf{x}))), \\ \Delta\mathbf{R}(\mathbf{x}) = (R_2(\mathbf{s}(\mathbf{x})) - R_1(\mathbf{s}(\mathbf{x})), \ldots, R_N(\mathbf{s}(\mathbf{x})) - R_{N-1}(\mathbf{s}(\mathbf{x})), R_1(\mathbf{s}(\mathbf{x})) - R_N(\mathbf{s}(\mathbf{x}))). \end{cases} \tag{14}$$

```
Initialization: x_(1) = x ∈ [0,1]^N ;
for k ≥ 1 do
    x_(k+1) = x_(k) + δ(x_(k)) × R(x_(k)) ;
    x_(k+1) = x_(k+1) mod(1).
end
```

Algorithm 1: Repulsion-based projection

The introduction of algorithm 1, which depends on the particular function $\mathbf{x} \mapsto \delta(\mathbf{x})$, is justified by the following proposition.

Proposition 1. *The sequence* $\{\mathscr{C}^{LHS}(\mathbf{x}_{(k)})\}_{k \geq 1}$ *converges to its minimal value* $1/N$ *when* k *tends to infinity.*

□ **Proof:**

1. We define:

$$\mathscr{C}^\star(\mathbf{x}) = \sum_{n=1}^{N} \left(\Delta x_n - \frac{1}{N} \right)^2 = \sum_{n=1}^{N} \Delta x_n^2 - \frac{1}{N}, \quad \mathbf{x} \in [0,1]^N. \tag{15}$$

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

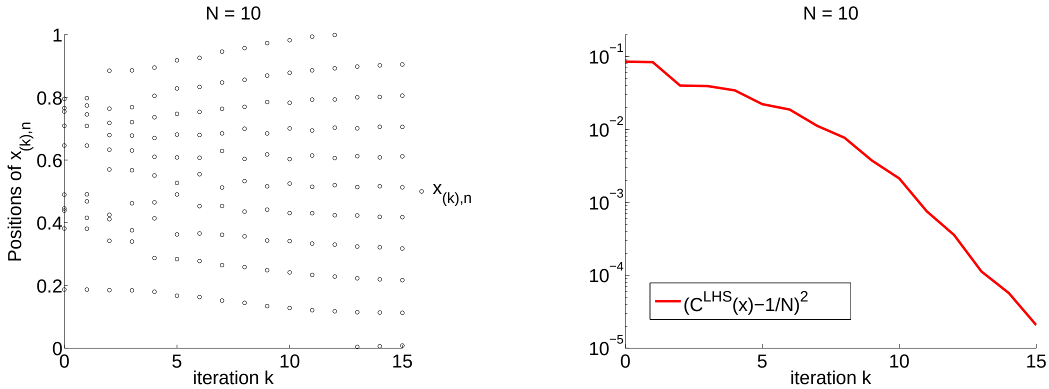


FIGURE 1. *Illustration of the efficiency of the proposed iterative algorithm to generate designs in* $[0,1]$, *which elements are separated by a distance that is as close as wanted to* $1/N$. *Left: evolution with respect to the number of iterations k of the positions of the elements* $x_{(k),n}$, *starting from a given design* $\mathbf{x}_{(1)}$ *which elements have been chosen randomly and independently in* $[0,1]$. *Right: evolution of the associated distance between criterion* $\mathscr{C}^{LHS}$ *and 1/N.*

2. For all $\mathbf{x}$ in $[0,1]^N$, $\mathscr{C}^\star(\mathbf{x}) \geq \left(\mathscr{C}^{\mathrm{LHS}}(\mathbf{x}) - \frac{1}{N}\right)^2 \geq 0$.
3. For all $k \geq 1$, $\mathscr{C}^\star(\mathbf{x}_{(k+1)}) \leq \mathscr{C}^\star(\mathbf{x}_{(k)})$. Indeed:
   – By construction of parameter $\delta(\mathbf{x}_{(k)})$,

$$\begin{aligned} &\mathscr{C}^\star(\mathbf{x}_{(k)} + \delta(\mathbf{x}_{(k)}) \times \mathbf{R}(\mathbf{x}_{(k)})) - \mathscr{C}^\star(\mathbf{x}_{(k)}) \\ &= \delta(\mathbf{x}_{(k)}) \times \left(2\left\langle \Delta\mathbf{x}, \Delta\mathbf{R}(\mathbf{x})\right\rangle_N + \delta(\mathbf{x}_{(k)}) \times \|\Delta\mathbf{R}(\mathbf{x})\|_N^2\right) \leq 0. \end{aligned} \tag{16}$$

   – For all $\mathbf{x}$ in $\mathbb{R}^N$, $\mathscr{C}^\star(\mathbf{x}) \geq \mathscr{C}^\star(\mathbf{x} \bmod(1))$.
4. Sequence $\left\{\mathscr{C}^*(\mathbf{x}_{(k)})\right\}_{k\geq 1}$ is decreasing and is greater than 0, it converges. By construction, its limit is characterized by

$$R_1(\mathbf{x}) = \ldots = R_N(\mathbf{x}), \tag{17}$$

   such that $\left\{\mathscr{C}^\star(\mathbf{x}_{(k)})\right\}_{k\geq 1}$ converges to zero.
5. It comes that $\left\{\mathscr{C}^{\mathrm{LHS}}(\mathbf{x}_{(k)})\right\}_{k\geq 1}$ converges to its minimal value $1/N$.

□

Hence, the sequence defined by Algorithm 1 allows us to generate designs in $[0,1]$, which elements are separated by a distance that is as close to $1/N$ as wanted. As an illustration, for $N = 10$, Figure 1 compares the positions of the elements of $\mathbf{x}_{(k)}$ with respect to k after 15 iterations, and shows the associated values of criterion $\mathscr{C}^{\mathrm{LHS}}$. Noticing that the LHS property considers each row of $[x]$ in $\mathscr{M}_{d,N}([0,1])$ independently, we now can introduce the algorithm 2 to define $[\Pi^{\mathrm{iter}}x]$, another pseudo-projection of $[x]$ on $\mathscr{H}_{d,N}^{\mathrm{LHS}}$. Two examples of such projected designs, $[\Pi x]$ and $[\Pi^{\mathrm{iter}}x]$, are shown in Figure 2, for $d = 2$ and $N = 10$.

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

Initialization: $[x] = \left[\mathbf{x}_{(1)} \,;\, \dots \,;\, \mathbf{x}_{(d)}\right] \in \mathscr{M}_{d,N}([0,1])$;
for $i = 1, \dots, d$ **do**
  while $B^{LHS}(\mathbf{x}_{(i)}) = 0$ **do**
    $\mathbf{x}_{(i)} = \mathbf{x}_{(i)} + \delta(\mathbf{x}_{(i)}) \times \mathbf{R}(\mathbf{x}_{(i)})$;
    $\mathbf{x}_{(i)} = \mathbf{x}_{(i)} \bmod(1)$;
  end
end
$[\Pi^{\text{iter}} x] = [x]$.

Algorithm 2: Multidimensional repulsion-based projection

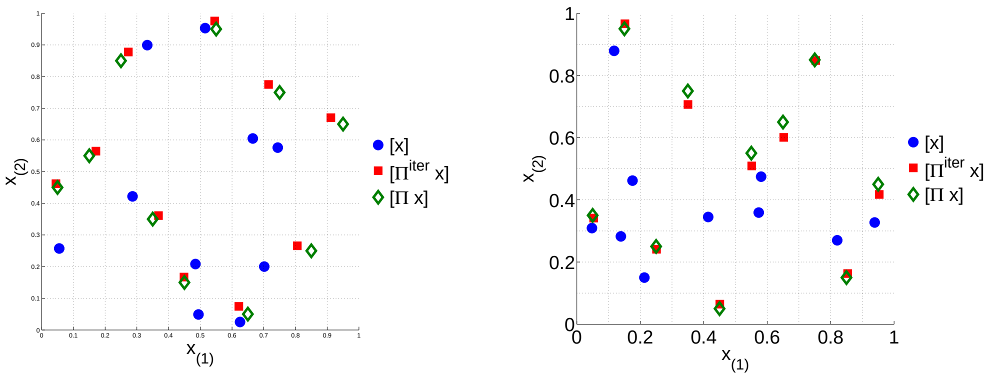


FIGURE 2. *Comparison between two random designs* $[x]$ *in* $\mathscr{M}_{d,N}([0,1])$ *and their projections in the class of LHS by the rank-based,* $[\Pi x]$*, and the repulsion-based,* $[\Pi^{iter} x]$*, algorithms defined in Section 2.1.1.*

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

2.1.2. *Optimizing the space filling properties of LHS*

The idea of this part is to show how the two LHS projectors of Section 2.1.1 can be coupled to an Euclidian distance-based repulsion between each element of the DoE to create very interesting space filling designs (SFD). Such SFD indeed allow a good global coverage rate of $[0,1]^d$ with respect to a given criterion $\mathscr{C}$, while properly covering the variation domain of each scalar product. To this end, for all $[x] = [\mathbf{x}^{(1)} \ \cdots \ \mathbf{x}^{(N)}]$ in $\mathscr{M}_{d,N}([0,1])$, we define $\mathfrak{R}^{(n)}([x])$ the distance-based repulsion vector that is seen by element $\mathbf{x}^{(n)}$, $1 \leq n \leq N$, such that:

$$\mathfrak{R}^{(n)}([x]) = \widetilde{\mathfrak{R}}^{(n)}([x]) + \widehat{\mathfrak{R}}^{(n)}([x]), \tag{18}$$

$$\widetilde{\mathfrak{R}}^{(n)}([x]) = \sum_{m \neq n} \frac{\mathbf{x}^{(n)} - \mathbf{x}^{(m)}}{\left\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\right\|_d^3}, \quad \widehat{\mathfrak{R}}^{(n)}([x]) = \sum_{m \neq n} \frac{\mathbf{x}^{(n)} - \widehat{\mathbf{x}}^{(n,m)}}{\left\|\mathbf{x}^{(n)} - \widehat{\mathbf{x}}^{(n,m)}\right\|_d^3}, \tag{19}$$

$$\widehat{\mathbf{x}}^{(n,m)} = \mathbf{x}^{(n)} + \frac{(\mathbf{x}^{(n)} - \mathbf{x}^{(m)})\left(\left\|\mathbf{e}^{(n,m),1} - \mathbf{e}^{(n,m),2}\right\|_d - \left\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\right\|_d\right)}{\left\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\right\|_d}, \quad 1 \leq m \leq N, \tag{20}$$

where $\mathbf{e}^{(n,m),1}$ and $\mathbf{e}^{(n,m),2}$ are the two intersection points between the boundary of the d-dimensional hypercube $[0,1]^d$ and the line passing by $\mathbf{x}^{(n)}$ and $\mathbf{x}^{(m)}$. In the same manner than in Eq. (12), the term $\widehat{\mathfrak{R}}^{(n)}([x])$ is a regularization term to make the elements of $\mathscr{M}_{d,N}([0,1])$ stay in $\mathscr{M}_{d,N}([0,1])$. Contrary to many classical regularizations for the hypercube, which are based on periodicity concepts (see Bornert et al., 2008 for further details), such a regularization can easily be computed for any bounded convex domain $\mathscr{B}^d$, at a controlled numerical cost. Indeed, Algorithm 3 allows a quick identification of points $\mathbf{e}^{(n,m),1}$ and $\mathbf{e}^{(n,m),2}$, with $L = \max_{\mathbf{x},\mathbf{y} \in \mathscr{B}^d} \|\mathbf{x} - \mathbf{y}\|$ and ε a required precision.

for $1 \leq n \neq m \leq N$ **do**
    $\mathbf{x}^{\mathrm{IN},1} = \mathbf{x}^{(n)}$, $\mathbf{x}^{\mathrm{OUT},1} = \mathbf{x}^{(m)} + L \times \left(\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\right) / \left\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\right\|_d$;
    $\mathbf{x}^{\mathrm{IN},2} = \mathbf{x}^{(m)}$, $\mathbf{x}^{\mathrm{OUT},2} = \mathbf{x}^{(n)} + L \times \left(\mathbf{x}^{(m)} - \mathbf{x}^{(n)}\right) / \left\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\right\|_d$;
    for $i = 1,2$ **do**
        while $\left\|\mathbf{x}^{IN,i} - \mathbf{x}^{OUT,i}\right\|_d \geq \varepsilon$ **do**
            $\widehat{\mathbf{x}} = (\mathbf{x}^{\mathrm{IN,i}} + \mathbf{x}^{\mathrm{OUT,i}})/2$;
            if $\widehat{\mathbf{x}} \in \mathscr{B}^d$ **then** $\mathbf{x}^{\mathrm{IN,i}} = \widehat{\mathbf{x}}$;
            ;
            else $\mathbf{x}^{\mathrm{OUT,i}} = \widehat{\mathbf{x}}$;
            ;
        end
        $\mathbf{e}^{(n,m),i} = \widehat{\mathbf{x}}$;
    end
end

Algorithm 3: Dichotomy-based identification of points $\mathbf{e}^{(n,m),1}$ and $\mathbf{e}^{(n,m),2}$

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

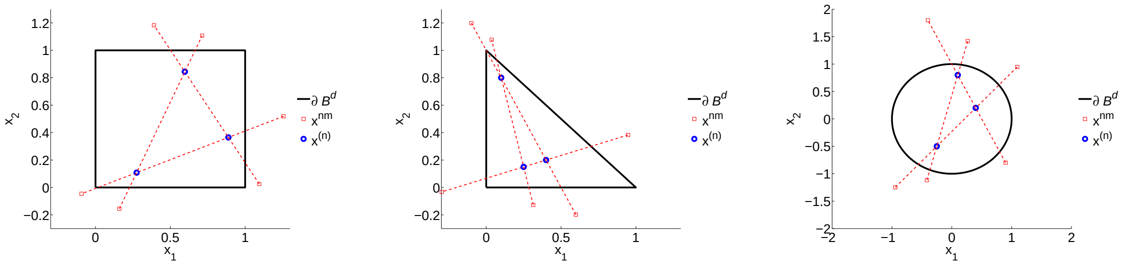


FIGURE 3. *Illustration of the proposed regularization for the distance-based repulsion, to make the elements of a bounded convex domain* $\mathscr{B}^d$ *stay inside it, in the cases of the hypercube (*$\mathscr{B}^d = [0,1]^2$*), the simplex (*$\mathscr{B}^d = \{\mathbf{x} \in [0,1]^2,\ x_1 + x_2 \leq 1\}$*) and the hypersphere (*$\mathscr{B}^d = \{\mathbf{x} \in [-1,1]^2,\ x_1^2 + x_2^2 \leq 1\}$*).*

As an illustration, Figure 3 shows an example of such a regularization in the case of the hypercube, the simplex and the hypersphere, for $d = 2$ and $N = 3$.

In order to optimize LHS with respect to criterion $\mathscr{C}$, we propose in this work to combine the former distance-based repulsion and one of the two LHS projectors introduced in Section 2.1.1, which is now denoted by Π^{LHS} for the sake of clarity. Given a particular design $[x]$, we then introduce the algorithm 4 to construct $[x^{\mathrm{opt}}]$, where $\delta^{(q)}$ and Q are numerical parameters to be optimized. Whereas Q has to be adapted to the affordable computational resources, the choice for parameter $\delta^{(q)}$ is not trivial, as there is no reason for criterion $\mathscr{C}$ to be continuous with respect to $\delta^{(q)}$. Moreover, if $\delta^{(q)}$ is too small, then there is almost no difference between $[x^{(q)}]$ and $[x^{(q-1)}]$, such that $[y^{(q)}] = [y^{(q-1)}]$, whereas a too big value for $\delta^{(q)}$ makes difficult the convergence of Algorithm 4 to a relevant design. From the analysis of a series of numerical examples with different values of d and N, it is clear that the value of parameter $\delta^{(q)}$ has to be adapted to the form of the domain and to the values of d and N. However, it is hard to infer the relation between the optimal value of $\delta^{(q)}$ (which can change from one iteration of the algorithm to another one) and these parameters. Nevertheless, it appears that values of $\delta^{(q)}$ such that:

$$\delta^{(q)} \times \max_{1 \leq n \leq N,\ 1 \leq i \leq d} |\mathfrak{R}_i^{(n)}([x^{(q-1)}])| \; = \; 10^{-2}, \;\; 2 \leq q \leq Q, \tag{21}$$

lead to relatively good results for values of Q between 10^2 and 10^3. Indeed, linking the value of $\delta^{(q)}$ to the maximal value of $\mathfrak{R}_i^{(n)}([x^{(q-1)}])$ decreases the risk of divergence of the algorithm when two points are getting too close.

The interest of such a combination between the proposed distance-based repulsion and the LHS projection is illustrated in Figure 4 for the hypercube, with $[\Pi^{\mathrm{LHS}} x] = [\Pi x]$ and the following criterion $\mathscr{C}$:

$$\mathscr{C}([x]) = \sum_{n=1}^{N} \left\| \widetilde{\mathfrak{R}}^{(n)}([x]) \right\|_d^2. \tag{22}$$

2.2. *Performance of the proposed method*

Space filling criteria

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

1 Initialization: $[x^{(1)}] = [x]$, $[y^{(1)}] = [\Pi^{\mathrm{LHS}} x^{(1)}]$;
2 **for** $2 \leq q \leq Q$ **do**
3 $[x^{(q)}] = [x^{(q-1)}] + \delta^{(q)} \times [\mathfrak{R}^{(1)}([x^{(q-1)}]) \ \cdots \ \mathfrak{R}^{(N)}([x^{(q-1)}])]$;
4 $[x^{(q)}] = [x^{(q)}] \bmod(1)$;
5 $[y^{(q)}] = [\Pi^{\mathrm{LHS}} x^{(q)}]$;
6 **end**
7 $[x^{\mathrm{opt}}] = \arg\min_{1\leq q\leq Q} \mathscr{C}\left([y^{(q)}]\right)$.

Algorithm 4: Repulsion-based optimization of LHS

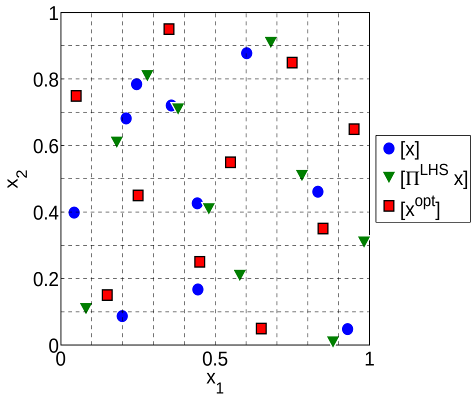


FIGURE 4. *Illustration of the interest of the combination between a LHS projection, Π^{LHS}, and a distance-based repulsion to define optimized LHS from any element $[x]$ in the hypercube $\mathscr{M}_{2,10}([0,1])$. Even if the points of $[x^{opt}]$ are almost in the centers of the boxes, projection Π^{LHS} corresponds here to the repulsion-based projection.*

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

To assess the quality of a given $(d\times N)$-dimensional DoE $[x]=[\mathbf{x}^{(1)}\ \cdots\ \mathbf{x}^{(N)}]$ to cover $[0,1]^d$, we consider in this paper three well-known quantitative indicators. First, the centered L^2-discrepancy, $D([x])$, is used to evaluate the uniformity quality of $[x]$ (see Fang et al., 2006 for further details about the discrepancy measures). Then, two geometrical criteria are introduced: the mindist criterion, $C^{\text{mindist}}([x])$, which is defined by Eq. (2), and an approximation of the minimax criterion,

$$C^{\text{mM}}([x],P)=\max_{1\le p\le P}\left\{\min_{1\le n\le N}\left\|\mathbf{x}^{(n)}-\mathbf{y}^{(p)}\right\|_d\right\},\quad \lim_{P\to+\infty}C^{\text{mM}}([x],P)=C^{\text{mM}}([x]),\tag{23}$$

where minimax criterion $C^{\text{mM}}([x])$ is defined by Eq. (1), and for $P>1$, $\left\{\mathbf{y}^{(p)},\ 1\le p\le P\right\}$ is a set of P vectors that have been randomly and uniformly chosen in $\mathscr{M}_{d,N}([0,1])$. In addition, for each DoE $[x]=[\mathbf{x}_{(1)}\ ;\ \dots\ ;\ \mathbf{x}_{(d)}]$, we denote by $D_{\text{2D}}([x])$, $C_{\text{2D}}^{\text{mindist}}([x])$ and $C_{\text{2D}}^{\text{mM}}([x],P)$ the mean values of the discrepancy, the mindist and the minimax criteria associated with each 2D-subprojection $[\mathbf{x}^{(i)}\ ;\ \mathbf{x}^{(j)}]$, $1\le i<j\le d$. These criteria are introduced in order to characterize the space filling quality of $[x]$ in terms of 2D-subprojections:

$$D_{\text{2D}}([x])=\frac{2}{d(d-1)}\sum_{1\le i<j\le d}D([\mathbf{x}^{(i)}\ ;\ \mathbf{x}^{(j)}]),\tag{24}$$

$$C_{\text{2D}}^{\text{mindist}}([x])=\frac{2}{d(d-1)}\sum_{1\le i<j\le d}C^{\text{mindist}}([\mathbf{x}^{(i)}\ ;\ \mathbf{x}^{(j)}]),\tag{25}$$

$$C_{\text{2D}}^{\text{mM}}([x],P)=\frac{2}{d(d-1)}\sum_{1\le i<j\le d}C^{\text{mM}}([\mathbf{x}^{(i)}\ ;\ \mathbf{x}^{(j)}],P).\tag{26}$$

Among these six criteria, criteria $D([x])$ and $C^{\text{mM}}([x],P)$ and their 2D projections have to be minimized, whereas criterion $C^{\text{mindist}}([x])$ and its 2D projection have to be maximized. Therefore, to make the comparison between several SFD generators easier, all the following numerical results will focus on the six criteria $D([x])$, $C^{\text{mM}}([x],P)$, $1/C^{\text{mindist}}([x])$, $D_{\text{2D}}([x])$, $C_{\text{2D}}^{\text{mM}}([x],P)$ $1/C_{\text{2D}}^{\text{mindist}}([x])$, such that if $[x^1]$ and $[x^2]$ are two elements of $\mathscr{M}_{d,N}([0,1])$, $[x^1]$ will be said to be better than $[x^2]$ to cover the $d-$dimensional hypercube if the six criteria associated with $[x^{(1)}]$ are lower than the ones associated with $[x^{(2)}]$:

$$\begin{cases}D([x^1])\le D([x^2]),\\ C^{\text{mM}}([x^1],P)\le C^{\text{mM}}([x^2],P),\\ 1/C^{\text{mindist}}([x^1])\le 1/C^{\text{mindist}}([x^2]),\\ D_{\text{2D}}([x^1])\le D_{\text{2D}}([x^2]),\\ C_{\text{2D}}^{\text{mM}}([x^1],P)\le C_{\text{2D}}^{\text{mM}}([x^2],P),\\ 1/C_{\text{2D}}^{\text{mindist}}([x^1])\le 1/C_{\text{2D}}^{\text{mindist}}([x^2]).\end{cases}\tag{27}$$

Parametric analyses

In this section, the nine following LHS generators are compared in terms of space filling properties.

- **Generator 1:** the routine "randomLHS", which is given by the R-package *lhs* is used to generate LHS in a random manner without regard to optimization.

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

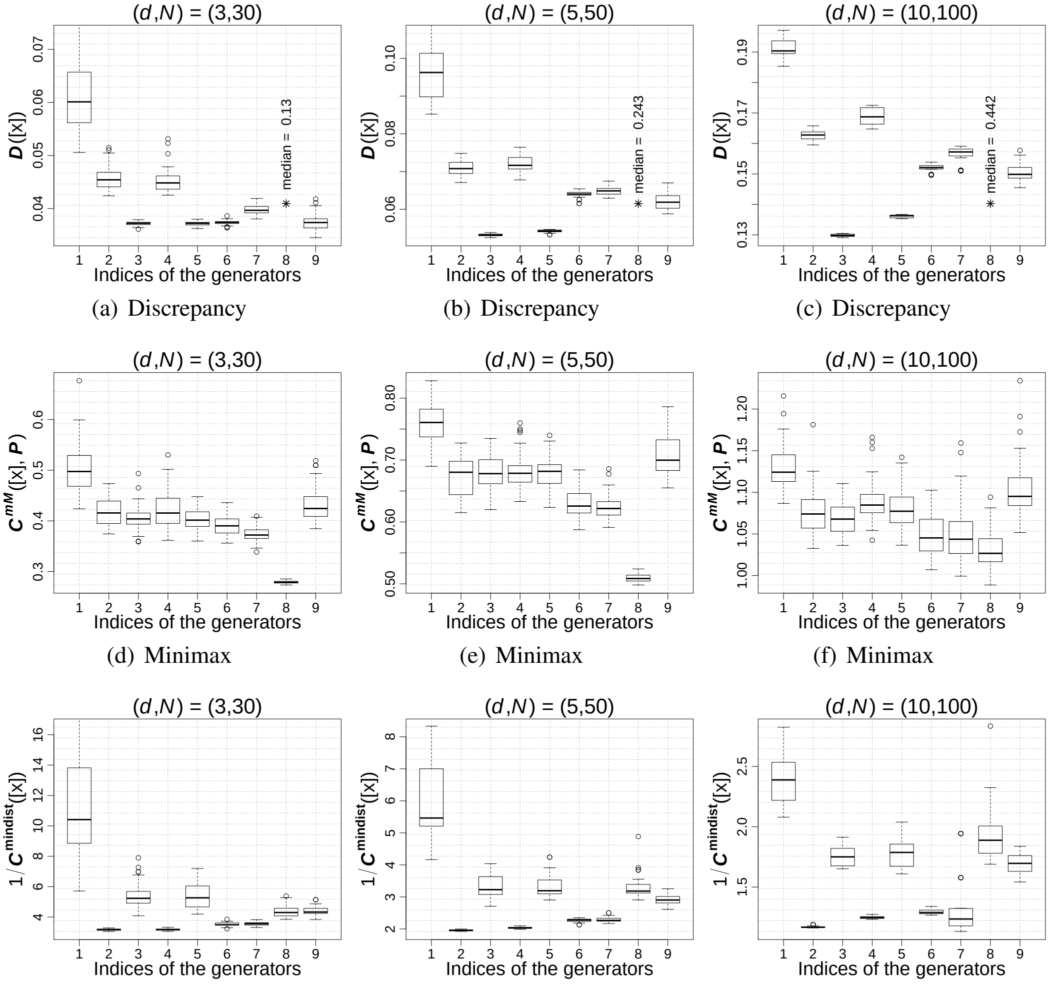
- **Generator 2, 3, 4 and 5:** the routines "maximinSA_LHS", "discrepSA_LHS", "maximinESE_LHS", and "discrepESE_LHS", which are provided by the R-package *DiceDesign*, propose a Simulated annealing (SA) and a Enhanced Stochastic Evolutionnary (ESE) algorithms for LHS optimization according to the mindist criterion or the discrepancy respectively.
- **Generators 6** and **7** correspond to Algorithm 4 associated with the criterion given by Eq. (22), with $Q = 500$, and the LHS pseudo-projections Π and Π^{iter} described in Section 2.1.1 respectively.
- **Generators 8** is associated with the clustering-based approaches. The different strategies presented in Likas et al. (2003), Lekivetz and Jones (2015), and Mak and Joseph (2016) have been tested, and at each iteration, the results associated with the smallest minimax value are kept. Contrary to the Generators 1 to 7, we lay stress on the fact that the clustering-based approaches are not constrained to have good projection properties in low-dimensional spaces.
- **Generators 9** corresponds finally to the MaxPro procedure, which has been presented in Introduction, and which is supposed to maximize the space-filling properties on any low-dimensional projections.

The parameters associated with all these generators have been optimized in order to get the best results for a computational time that is globally constant for each method. Although the minimax criterion is said to be the most representative criterion of the space filling properties of a design, we remind that, in practice, its direct minimization is almost unfeasible for values of d higher than five. This explains that most of the proposed generators are focusing on the mindist maximization, which is a much easier problem to solve.

As the first motivation for this work is to define initial DoE to build a surrogate model, rather small values of d (between 1 and 20) and N (between $5d$ and $30d$) will be studied in this work. First, for $d \in \{3,5,10\}$, and $N = 10d$, 50 different LHS are generated from the nine former generators. For each generated LHS $[x]$, we then compute the L^2-discrepancy, $D([x])$, the pseudo-minimax criterion with $P = 10^5$, $C^{\text{mM}}([x],P)$, the mindist criterion, $C^{\text{mindist}}([x])$, and their 2D mean projections. The results are then compared in figures 5 and 6, where it can be seen that the proposed method gives very interesting results in terms of space filling in the hypercube with a reasonable complexity. Indeed, the results of the proposed approach can be seen as a good compromise between the discrepancy minimization and the mindist maximization, while it is almost always the best in terms of minimax minimization among the generators that are constrained to have good space filling properties in low-dimensional spaces. Without surprise, the clustering-based designs are the ones that minimize the minimax criterion, but as they are not constrained in terms of space filling properties in low-dimensional spaces, their discrepancy values are very high, and the 2D-mean projections of the considered criteria are not satisfying. On the contrary, it can be seen that the results associated with the MaxPro procedure are very interesting when looking at the 2D-mean projections of the considered criteria, while leading to higher values for the geometrical criteria in the whole space. At last, it can be noticed that, without surprise, the influence of the choice of the pseudo projection, Π and Π^{iter} is low.

If we focus on the LHS generators 2, 3 and 6, we can verify in Figure 7 that the good space filling properties of the proposed method are found again when the dimension of the input space, d, increases, or when for a given value of d, the number of elements of the DoE, N, increases.

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238



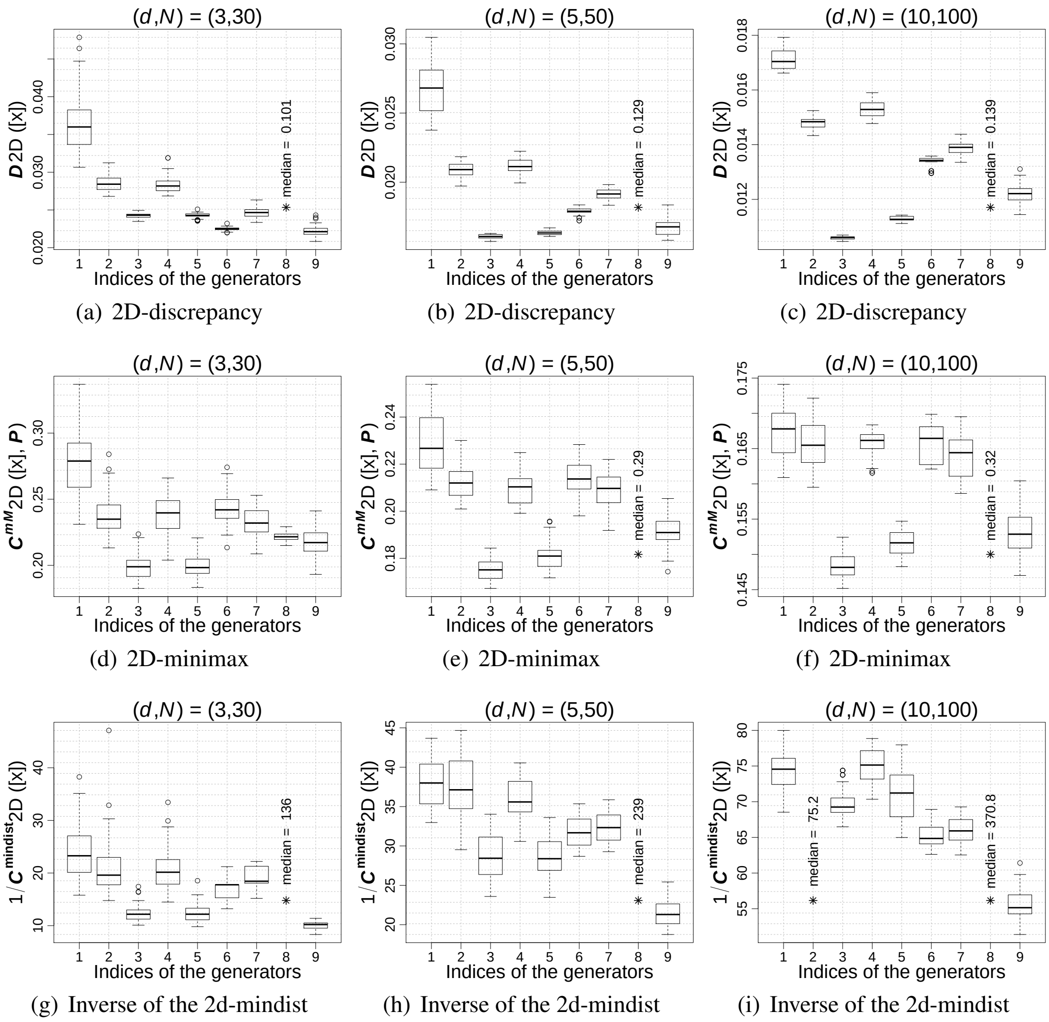
(a) Discrepancy (b) Discrepancy (c) Discrepancy

(d) Minimax (e) Minimax (f) Minimax

(g) Inverse of the mindist (h) Inverse of the mindist (i) Inverse of the mindist

FIGURE 5. *Comparison in terms of space filling properties of nine LHS generators with respect to the L^2-discrepancy (first row), the minimax criterion (second row), and the mindist criterion (third row) for (d,N) equal to $(3,30)$ (first column), $(5,50)$ (second column) and $(10,100)$ (third column). 1: randomLHS / 2: maximinSA_LHS / 3: discrepSA_LHS / 4: maximinESE_LHS / 5: discrepESE_LHS / 6: the proposed method associated with the pseudo-projection Π / 7: the proposed method associated with the pseudo-projection Π^{iter} / 8: clustering-based approach / 9: MaxPro. Boxplots are produced from 50 runs with different initial DoE. When the values of a criterion are too high compared to the others, the associated boxplot is replaced by an asterix, and the median of these values is indicated.*

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238



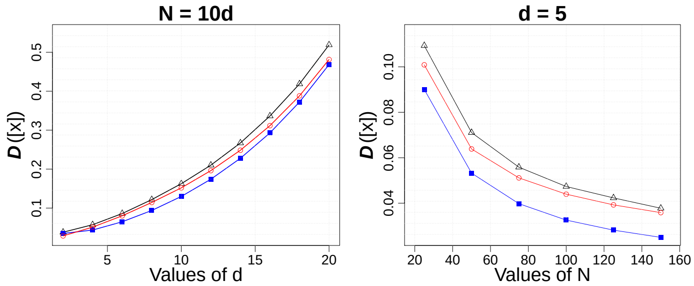
(a) 2D-discrepancy (b) 2D-discrepancy (c) 2D-discrepancy

(d) 2D-minimax (e) 2D-minimax (f) 2D-minimax

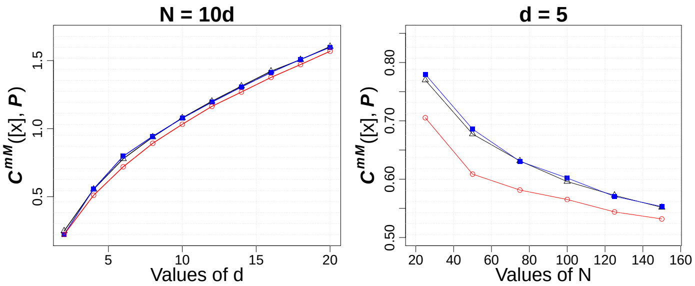
(g) Inverse of the 2d-mindist (h) Inverse of the 2d-mindist (i) Inverse of the 2d-mindist

FIGURE 6. *Comparison in terms of space filling properties of seven LHS generators with respect to the 2D-L^2-discrepancy (first row), the 2D-minimax criterion (second row), and the 2D-mindist criterion (third row) for (d,N) equal to $(3,30)$ (first column), $(5,50)$ (second column) and $(10,100)$ (third column). 1: randomLHS / 2: maximinSA_LHS / 3: discrepSA_LHS / 4: maximinESE_LHS / 5: discrepESE_LHS / 6: the proposed method associated with the pseudo-projection Π / 7: the proposed method associated with the pseudo-projection Π^{iter} / 8: clustering-based approach / 9: MaxPro. Boxplots are produced from 50 runs with different initial DoE. When the values of a criterion are too high compared to the others, the associated boxplot is replaced by an asterix, and the median of these values is indicated.*

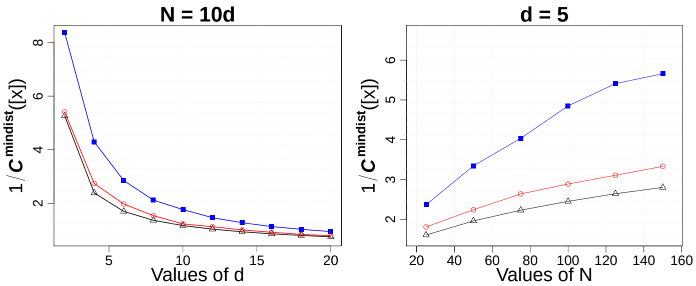
© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238



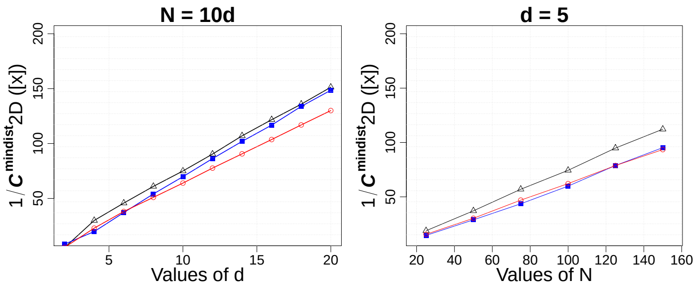
(a) Discrepancy (b) Discrepancy



(c) Minimax (d) Minimax



(e) Inverse of the mindist (f) Inverse of the mindist



(g) Inverse of the 2D-mindist (h) Inverse of the 2D-mindist

FIGURE 7. *Quantification of the influence of an increase of the dimensions d and N of the input space on the space filling properties for three particular LHS generators with respect to the* L^2*-discrepancy (first row), the minimax criterion (second row), the mindist criterion (third row), and the 2D-mindist criterion (fourth row). Black curves with upper triangles: maximinSA_LHS. Blue curves with filled squares: discrepSA_LHS. Red curves with circles: the proposed method associated with the pseudo-projection* Π*. Each curve corresponds to the mean value of a space filling criterion, which has been computed from 50 independent generated LHS.*

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

3. Adaptation to the case of constrained input space

As presented in Introduction, this work aims at constructing optimized DoE in the cases when there exists constraints on the input space. To this end, this section presents a generalization of the repulsion-based method described in Section 2, which can be used to generate optimized initial DoE when the input space is a bounded convex domain.

3.1. Generalizing the notion of LHS for bounded convex domain

Let $\mathscr{B}^d \subset \mathbb{R}^d$ be a bounded convex domain, and $\mathbf{X} = (X_1, \dots, X_d)$ be a random vector that is uniformly distributed on $\mathscr{B}^d$. By construction, the probability density function (PDF), $f_{\mathbf{X}}$, of $\mathbf{X}$, is given by:

$$f_{\mathbf{X}} : \begin{cases} \mathbb{R}^d & \to & \mathbb{R} \\ \mathbf{x} & \mapsto & \mathbb{I}_{\mathscr{B}^d}(\mathbf{x}) / \mathscr{V}(\mathscr{B}^d) \end{cases}, \tag{28}$$

where $\mathscr{V}(\mathscr{B}^d)$ is the volume of $\mathscr{B}^d$, and $\mathbb{I}_{\mathscr{B}^d}(\mathbf{x})$ is equal to 1 if $\mathbf{x}$ is in $\mathscr{B}^d$ and 0 otherwise. In the same manner, for all $1 \le i \le d$, we denote by f_{X_i} and F_{X_i} the PDF and the cumulative density function (CDF) of X_i, respectively:

$$f_{X_i}(x) = \int_{\mathbb{R}^{d-1}} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}^{(-i)}, \;\; F_{X_i}(x) = \int_{-\infty}^{x} f_{X_i}(t)dt, \;\; \mathbf{x}^{(-i)} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d), \tag{29}$$

such that $F_{X_i}(X_i)$ is a random value that is uniformly distributed on $[0,1]$.

Hence, to define a DoE $[x] = \left[\mathbf{x}_{(1)}; \dots; \mathbf{x}_{(d)}\right]$ in $\bigotimes_{n=1}^{N} \mathscr{B}^d$ with good projection properties for each scalar input, the idea is to apply Algorithm 2 on $[\mathbf{F}^1(\mathbf{x}_{(1)}); \dots; \mathbf{F}^d(\mathbf{x}_{(d)})]$, where:

$$\mathbf{F}^i(\mathbf{x}_{(i)}) = (F_{X_i}(x_{(i),1}), \dots, F_{X_i}(x_{(i),N})), \;\; 1 \le i \le d. \tag{30}$$

To this end, the algorithm 5 can be introduced as a generalization of Algorithm 2, for a given precision ε and a maximum number of iterations J.

In Algorithm 5, vector-valued repulsion function $\mathbf{R}$ is defined by Eqs. (11) and (12), criterion $\mathscr{C}^{\text{LHS}}$ is introduced in Eq. (9), $\mathscr{T}(\cdot)$ is a regularization function to make the vectors of $\mathscr{B}^d$ stay in $\mathscr{B}^d$, and $\delta^\star(\mathbf{x}_{(i)})$ is a constant that is chosen such that:

$$\delta^\star(\mathbf{x}_{(i)}) = \begin{cases} 0 \;\; \text{if} \;\; R_1(\mathbf{F}^i(\mathbf{x}_{(i)})) = \dots = R_N(\mathbf{F}^i(\mathbf{x}_{(i)})), \\ -\left\langle \Delta\mathbf{F}^i, \Delta\mathbf{R}^{\star,i} \right\rangle_N / \left\| \Delta\mathbf{R}^{\star,i} \right\|_N^2 \;\;\; \text{otherwise, with} \end{cases} \tag{31}$$

$$\Delta\mathbf{F}^i = \left(F_2^i(\mathbf{x}_{(i)}) - F_1^i(\mathbf{x}_{(i)}), \dots, 1 - (F_N^i(\mathbf{x}_{(i)}) - F_1^i(\mathbf{x}_{(i)}))\right), \tag{32}$$

$$\begin{aligned} \Delta\mathbf{R}^{\star,i} = & (f_{X_i}(x_{(i),2}) R_2(\mathbf{F}^i(\mathbf{x}_{(i)})) - f_{X_i}(x_{(i),1}) R_1(\mathbf{F}^i(\mathbf{x}_{(i)})), \dots, \\ & f_{X_i}(x_{(i),1}) R_1(\mathbf{F}^i(\mathbf{x}_{(i)})) - f_{X_i}(x_{(i),N}) R_N(\mathbf{F}^i(\mathbf{x}_{(i)}))). \end{aligned} \tag{33}$$

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

Initialization: $[x] = \left[\mathbf{x}_{(1)};\ldots;\mathbf{x}_{(d)}\right] \in \bigotimes_{n=1}^{N} \mathscr{B}^d$;
for $i = 1,\ldots,d$ **do**
  $j = 0$;
  while $|\mathscr{C}^{LHS}(\mathbf{F}^i(\mathbf{x}_{(i)})) - 1/N| \geq \varepsilon,\ j \leq J$ **do**
    $j = j+1$;
    $\widetilde{\mathbf{x}}_{(i)} = \mathbf{x}_{(i)} + \delta^\star(\mathbf{x}_{(i)}) \times \mathbf{R}(\mathbf{F}^i(\mathbf{x}_{(i)}))$;
    $\mathbf{x}_{(i)} = \mathscr{T}(\widetilde{\mathbf{x}}_{(i)})$,
  end
end
$[\Pi_{\mathscr{B}^d} x] = \left[\mathbf{x}_{(1)};\ldots;\mathbf{x}_{(d)}\right]$.

Algorithm 5: Repulsion-based algorithm for the construction of DoE in $\mathscr{B}^d$ with good projection properties for each scalar input.

Indeed, noticing that, for all $1 \leq n \leq N$, if $\delta^\star(\mathbf{x}_{(i)}) \times R_n(\mathbf{F}^i(\mathbf{x}_{(i)}))$ is small compared to $x_{(i),n}$, then:

$$
\begin{aligned}
F_n^i(\mathbf{x}_{(i)} + \delta^\star(\mathbf{x}_{(i)}) \times \mathbf{R}(\mathbf{F}^i(\mathbf{x}_{(i)}))) &= F_{X_i}(x_{(i),n} + \delta^\star(\mathbf{x}_{(i)}) \times R_n(\mathbf{F}^i(\mathbf{x}_{(i)}))) \\
&\approx F_{X_i}(x_{(i),n}) + \delta^\star(\mathbf{x}_{(i)}) \times \frac{dF_{X_i}}{dx_i}(x_{(i),n}) R_n(\mathbf{F}^i(\mathbf{x}_{(i)})), \\
&\approx F_{X_i}(x_{(i),n}) + \delta^\star(\mathbf{x}_{(i)}) \times f_{X_i}(x_{(i),n}) R_n(\mathbf{F}^i(\mathbf{x}_{(i)})),
\end{aligned} \tag{34}
$$

it comes from Eq. (15) that:

$$
\begin{aligned}
&\mathscr{C}^\star(\mathbf{F}^i(\mathbf{x}_{(i)} + \delta^\star(\mathbf{x}_{(i)}) \times \mathbf{R}(\mathbf{F}^i(\mathbf{x}_{(i)})))) - \mathscr{C}^\star(\mathbf{F}^i(\mathbf{x}_{(i)})) \\
&\approx \delta^\star(\mathbf{x}_{(i)}) \left(2 \left\langle \Delta\mathbf{R}^{\star,i}, \Delta\mathbf{F}^i \right\rangle_N + \delta^\star(\mathbf{x}_{(i)}) \times \left\| \Delta\mathbf{R}^{\star,i} \right\|_N^2 \right) \leq 0.
\end{aligned} \tag{35}
$$

There is, however, no reason for $\mathscr{C}^\star(\mathbf{F}^i(\mathbf{x}_{(i)}))$ to be lower than $\mathscr{C}^\star(\mathbf{F}^i(\mathscr{T}(\widetilde{\mathbf{x}}_{(i)})))$ for any transformation $\mathscr{T}(\cdot)$, which can limit the convergence of the former algorithm, and explains the presence of a limitation on the maximum number of iterations. For instance, a possibility for $\mathscr{T}(\cdot)$ is given by:

$$
\mathscr{T}(\widetilde{\mathbf{x}}_{(i)}) = \begin{cases} \widetilde{\mathbf{x}}_{(i)} \text{ if } \widetilde{\mathbf{x}}_{(i)} \in \mathscr{B}^d, \\ \mathbf{x}_{(i)} \text{ otherwise}, \end{cases} \tag{36}
$$

but with respect to the properties of $\mathscr{B}^d$, better choices, which can be based on permutations between near elements of the same design, can be proposed to allow a better convergence of $\mathscr{C}^\star$ and $\mathscr{C}^{\mathrm{LHS}}$. As an illustration of such an approach, Figures 8 and 9 compare two random designs (which were generated according to the algorithms presented in Section 3.3) $[x]$ in the 2-dimensional simplex, $\{\mathbf{x} = (x_1, x_2) \in [0,1]^2 \mid 0 \leq x_1 + x_2 \leq 1\}$, and in the 2-dimensional hypersphere, $\{\mathbf{x} = (x_1, x_2) \in [0,1]^2 \mid 0 \leq x_1^2 + x_2^2 \leq 1\}$, and their images, $[\Pi_{\mathscr{B}^d} x]$, by the algorithm defined by Eq. (5). For all $1 \leq i \leq d$, it is reminded that for the d-dimensional simplex:

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

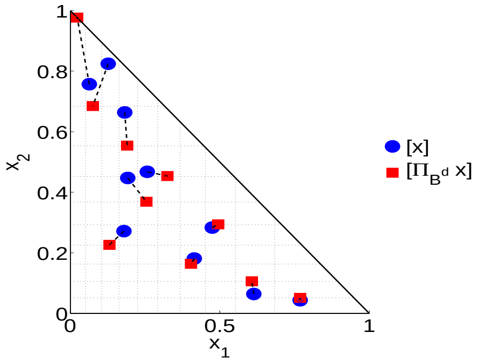


FIGURE 8. *Application of the Algorithm 5 for the definition of generalized LHS in the* 2*-dimensional simplex, for* $\varepsilon = 10^{-5}$. $[x]$ *is a DoE gathering 10 elements that are randomly chosen in the* 2*-dimensional simplex, and* $[\Pi_{\mathscr{B}^d} x]$ *is its image by Algorithm 5.*

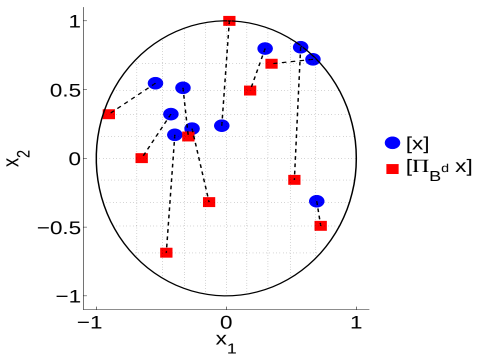


FIGURE 9. *Application of Algorithm 5 for the definition of generalized LHS in the* 2*-dimensional hypersphere, for* $\varepsilon = 10^{-5}$. $[x]$ *is a DoE gathering 10 elements that are randomly chosen in the* 2*-dimensional hypersphere, and* $[\Pi_{\mathscr{B}^d} x]$ *is its image by Algorithm 5.*

$$f_{X_i}(x) = d(1-x)^{d-1}, \quad x \in [0,1], \tag{37}$$

whereas for the d-dimensional hypersphere, we have:

$$f_{X_i}(x) = \frac{2\Gamma(N/2+1)}{\Gamma(1/2)\Gamma(N+1)}(1-x^2)^{(N-1)/2}, \quad x \in [-1,1], \tag{38}$$

where $\Gamma(\cdot)$ is the Gamma function.

In figures 8, 9 and 10, the dotted lines correspond to the $1/N, \dots (N-1)/N$ quantiles associated with F_{X_i}.

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

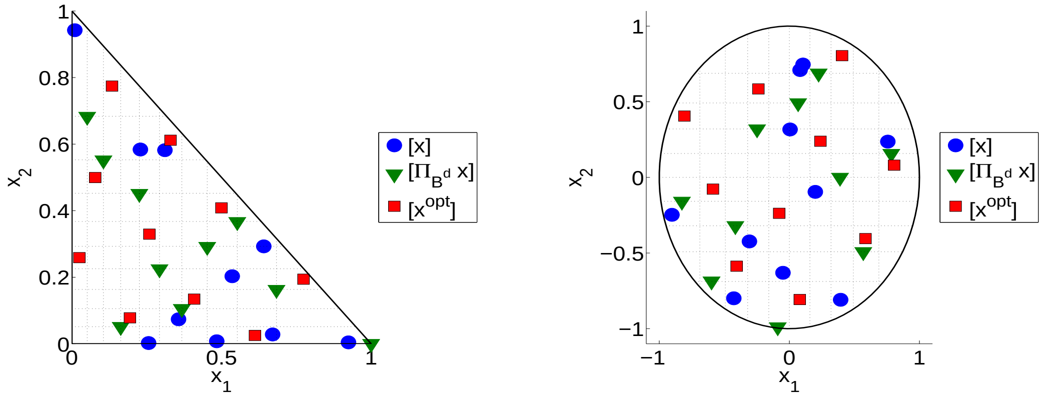


FIGURE 10. *Optimization of the space filling properties of an initial DoE in the simplex and the hypersphere for $d = 2$ and $N = 10$: $[x]$ is an initial DoE that is randomly chosen in $\mathscr{B}^d$; $[\Pi_{\mathscr{B}^d} x]$ is its projection in a subspace of $\mathscr{B}^d$ that preserves good projection properties for each scalar input; $[x^{opt}]$ is the final DoE that is optimized according to Algorithm 4.*

3.2. Construction of space filling DoE in bounded convex domains

In the same manner than in Section 2, it is supposed that the good coverage of $\mathscr{B}^d$ can be assessed from the evaluation of a criterion $\mathscr{C}$ to be minimized. Therefore, replacing Π^{LHS} by $\Pi_{\mathscr{B}^d}$ in Algorithm 4, it is possible to construct initial DoE that are optimized with respect to $\mathscr{C}$, while preserving good projection properties for each scalar input. Such a procedure is once again illustrated in the case of the simplex and the hypersphere in Figure 10, when criterion $\mathscr{C}$ is defined by Eq. (22).

3.3. Efficiency of the proposed method for the definition of space filling DoE in bounded convex domains

The idea of this section is to illustrate, for several values of d and N, the interest of the former algorithms on two examples of commonly used constrained input spaces: the simplex and the hypersphere. The notion of discrepance being difficult to generalize for these constrained input space, only the mindist and the minimax criteria will be computed to assess the space filling properties of a given DoE. In addition, to evaluate the quality of the projection properties for each scalar input of any $(d \times N)$-dimensional DoE $[x] = [\mathbf{x}_{(1)}; \dots; \mathbf{x}_{(d)}]$, the following criterion is also computed:

$$\text{distLHS}([x]) = \max_{1 \leq i \leq d} |\mathscr{C}^{\text{LHS}}\left(\mathbf{F}^i(\mathbf{x}_{(i)})\right) - 1/N|, \tag{39}$$

where function $\mathscr{C}^{\text{LHS}}$ and vector $\mathbf{F}^i(\mathbf{x}_{(i)})$ are defined by Eqs. (9) and (30) respectively. Indeed, the lower criterion distLHS is, the nearer from an uniform grid $[\mathbf{F}^1(\mathbf{x}_{(1)}); \dots; \mathbf{F}^d(\mathbf{x}_{(d)})]$ is, and so the better are the projection properties for each scalar input.

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

Generation of DoE in the simplex

In this part, we are interested in the generation of N-dimensional DoE in the d-dimensional simplex:

$$\mathscr{S}^d = \left\{ \mathbf{x} \in [0,1]^d \mid \sum_{i=1}^{d} x_i \leq 1 \right\}. \tag{40}$$

Three methods are commonly used to generate such DoE in $\mathscr{S}^d$ (Devroye, 1986; Pillards and Cools, 2003).

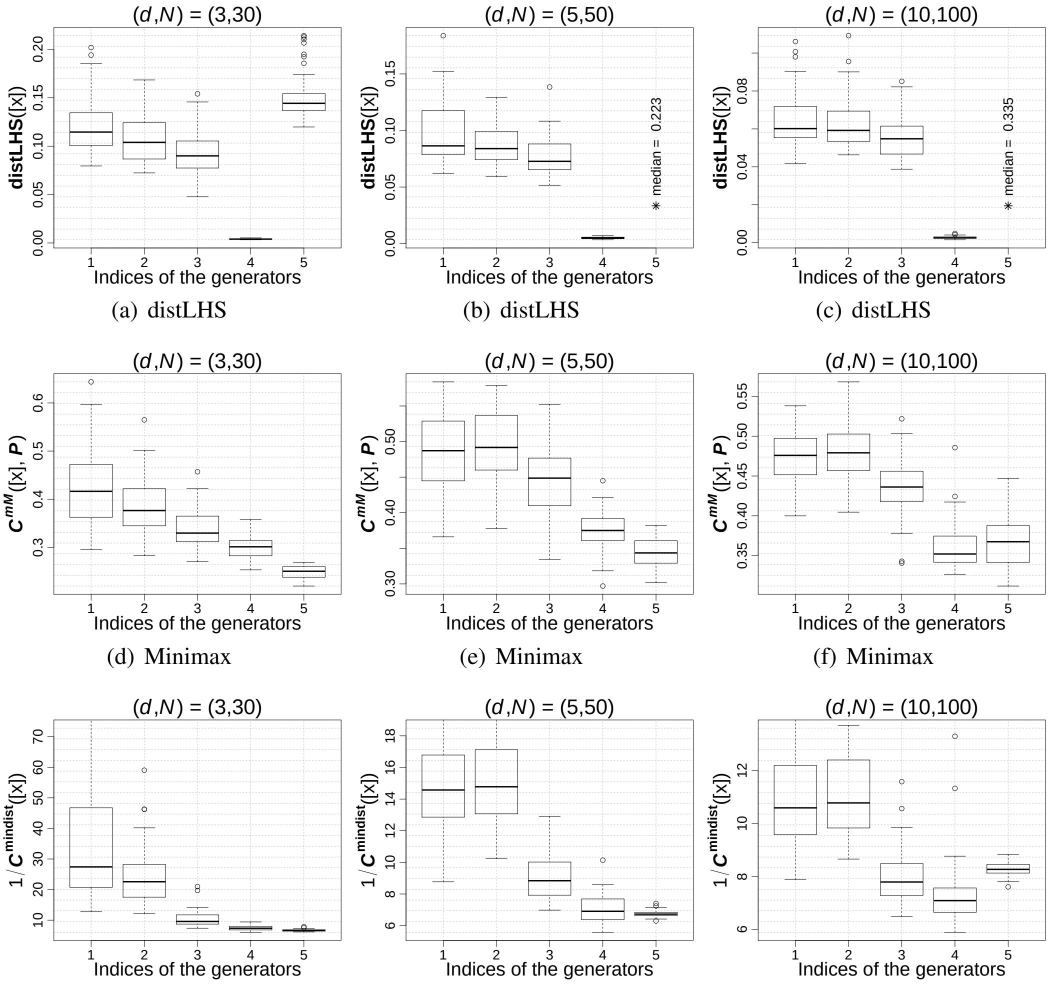
– The Dirichlet approach: if $G_1, \dots, G_{d+1}$ are $d+1$ Gamma-distributed random variables of parameters $(1,1)$, then the random vector $(x_1, \dots, x_d)$, where, for all $1 \leq i \leq d$, $x_i = G_i / \sum_{j=1}^{d+1} G_j$, is uniformly distributed on $\mathscr{S}^d$.
– The sorting approach: if $Y_1, \dots, Y_d$ are d independent random variables that are uniformly distributed on $[0,1]$, and if $Z_1 = Y_{s(1)} \leq \dots \leq Z_d = Y_{s(d)}$, where s is a bijective application from $\{1, \dots, d\}$ to $\{1, \dots, d\}$, then it can be shown that the vector $(x_1 = Z_2 - Z_1, x_2 = Z_3 - Z_2, \dots, x_d = 1 - Z_d)$ is uniformly distributed on $\mathscr{S}^d$.
– The square root approach: if $Y_1, \dots, Y_d$ are d independent random variables that are uniformly distributed on $[0,1]$, and if we define, for all $1 \leq i \leq d$, $Z_i = \prod_{j=i}^{d} Y_j^{1/j}$, then $(x_1 = Z_2 - Z_1, x_2 = Z_3 - Z_2, \dots, x_d = 1 - Z_d)$ is uniformly distributed on $\mathscr{S}^d$.

In the following, the Dirichlet approach will be considered as the non-optimized reference as the Gamma realizations on which it is based will be drawn independently, whereas the sorting and the square root approaches will be optimized in the sense that they will correspond to two transformations of an optimized LHS (according to the *maximinSA_LHS* routine with the same parameters than in Section 2) to the d-dimensional simplex. As for the hypercube case, a clustering-based approach is also introduced. It is seen as a reference for the minimax and mindist optimization without constraints on the space filling properties in the low-dimensional spaces.

As in Section 2, 50 different DoE are generated from each former generators and from the proposed method for $d = 3, 5, 10$, and $N = 10d$. For each DoE, we then compute the mindist, the pseudo minimax, the distLHS and their 2D-mean projections, and compare the results in Figures 11 and 12. In these figures, it can be seen that the methods based on the transformation of an optimized LHS give better results than the non-optimized Dirichlet approach. However, as these transformations are not linear, they do not preserve the projection properties for each scalar input, such that the distLHS criterion is always relatively high. The same phenomenon is observed for the clustering-based approaches, for which the distLHS criterion is always very high. On the contrary, for all presented values of d and N, the proposed approach allows the generation of DoE with very low values for distLHS, that is to say with very good projection properties for each scalar input, while improving the mindist, the 2D-mindist and the minimax criteria. Although the results are not shown in this paper for the sake of concision, the conclusions would have been the same for other values of d and N. Therefore, the proposed approach appears to be a very interesting tool to generate optimized DoE in the d-dimensional simplex.

Generation of DoE in the hypersphere

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238



(a) distLHS (b) distLHS (c) distLHS

(d) Minimax (e) Minimax (f) Minimax

(g) Inverse of the mindist (h) Inverse of the mindist (i) Inverse of the mindist

FIGURE 11. *Comparison in terms of space filling properties of five generators in the simplex with respect to the distLHS criterion (first row), the minimax criterion (second row) and the mindist criterion (third row), for $d = 3$ (left column), $d = 5$ (middle column) and $d = 10$ (right column), and $N = 10d$. 1: Dirichlet approach / 2: Sorting approach / 3: square root approach / 4: the proposed method / 5: clustering-based approach.*

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

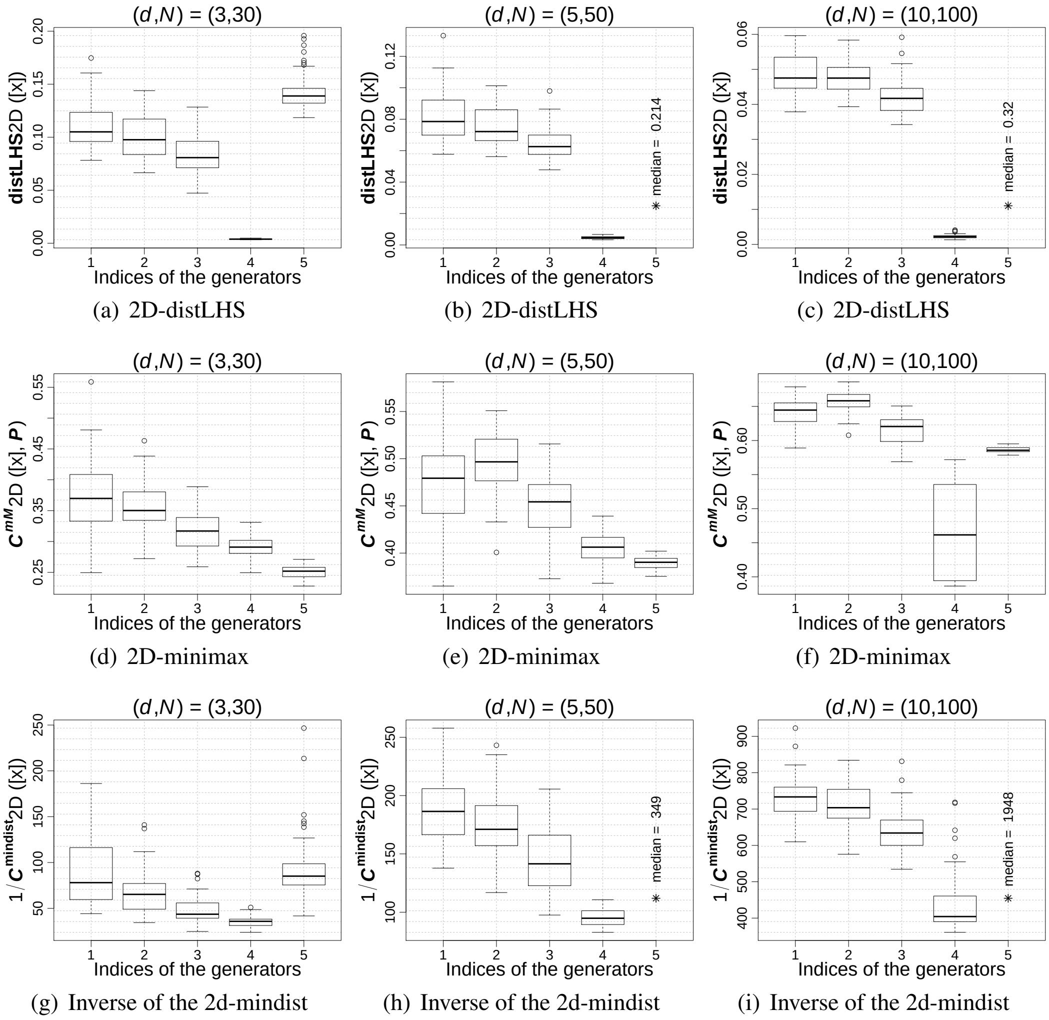


FIGURE 12. *Comparison in terms of space filling properties of five generators in the simplex with respect to the 2D-distLHS criterion (first row), the 2D-minimax criterion (second row) and the 2D-mindist criterion (third row), for $d = 3$ (left column), $d = 5$ (middle column) and $d = 10$ (right column), and $N = 10d$. 1: Dirichlet approach / 2: Sorting approach / 3: square root approach / 4: the proposed method / 5: clustering-based approach.*

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

As a second example, we are now interested in the generation of N-dimensional DoE in the d-dimensional hypersphere:

$$\mathscr{P}^d = \left\{ \mathbf{x} \in [-1,1]^d \mid \sum_{i=1}^{d} x_i^2 \leq 1 \right\}. \tag{41}$$

Such a generation is generally based on the consideration that if Y is a random variable that is uniformly distributed on $[0,1]$, and if ξ is a d-dimensional random vector which components are independent and normally distributed, then $\mathbf{x} = Y^{1/d} \times \xi / \|\xi\|_d$ is uniformly distributed on $\mathscr{P}^d$. This approach is called "normal approach" in the following, as it is based on the symmetry properties of the multivariate normal distribution. The results associated with designs based on the transformation of optimized LHS would give essentially the same results, but are not shown in this paper. Clustering-based designs are also considered in the following comparisons, as these approaches are supposed to work well in any input space from the minimax minimization point of view.

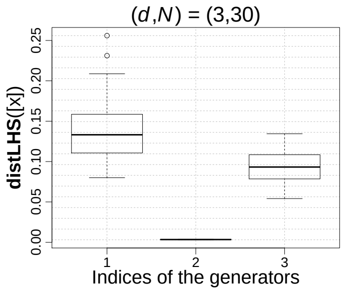
As previously, 50 different DoE are generated from the proposed method, the normal approach, and the clustering-based approach, for different values of d and N. We then compare the corresponding values of the mindist, the pseudo minimax, the distLHS and their 2D-mean projections in Figures 13 and 14. As forecast, the clustering-based approaches lead to interesting space filling properties when considering the whole input space, but are not good when focusing on the distLHS criterion or the 2D-mean projections of the considered criteria. On the contrary, these figures allow us to quantify the high interest of the proposed method to generate DoE in the hypersphere, in terms of mindist maximization and minimax and distLHS minimization, while preserving relatively good space filling properties in 2D subspaces.

4. Application to the enrichment of existing space filling designs

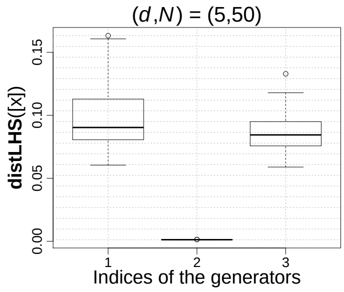
In many scientific applications, it can be useful to increase the dimensions N (and eventually d) of the initial DoE (Draguljić et al., 2012; Muehlenstaedt et al., 2014). Given an already existing (d_1, N_1)-dimensional DoE (which can be optimized or not), the question then arises of how adding the new elements for the final DoE to present good space filling characteristics and good projection properties for each scalar input. Based on the method introduced in Sections 2 and 3, an answer can be given to this question by fixing all the components of the already existing DoE but letting move the added new points in Algorithm 4. To improve the efficiency of the method, for each iteration of the distance-based repulsion, it is advised to perform the LHS projection twice: first, with the moving points alone, and second, with all the points (including the fixed points). The interest of this approach is illustrated for the hypercube only, even if it can be used for any bounded convex input space.

On the one hand, the value of d is fixed, and we want to increase the number of elements of the DoE. Hence, we first generate a (d, N_1)-dimensional DoE, and a (d, N_2)-dimensional DoE, $[y]$, from the non-optimized "randomLHS" routine that was introduced in Section 2 (N_2 can be equal or not to N_1). In addition, we denote by $[x; y]$ the concatenation of $[x]$ and $[y]$ and by $\mathscr{O}$ the proposed optimization procedure, such that $[\mathscr{O}([z_1; z_2] \mid z_2)]$ corresponds to the result of Algorithm

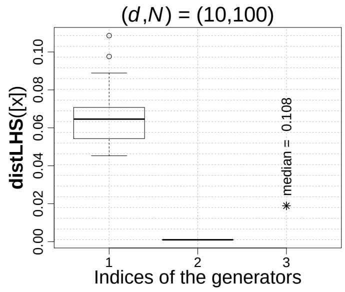
© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238



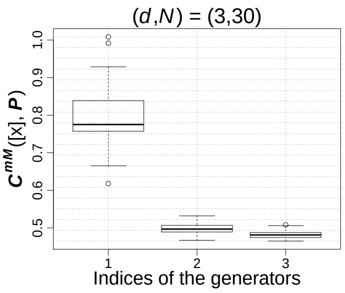
(a) distLHS



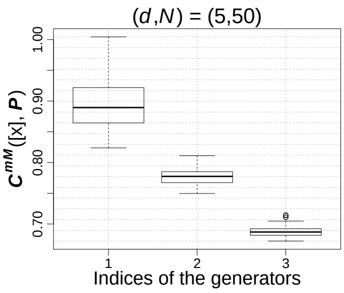
(b) distLHS



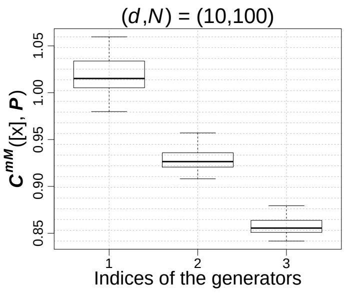
(c) distLHS



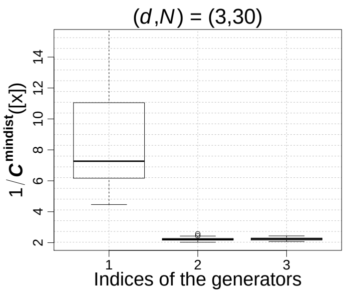
(d) Minimax



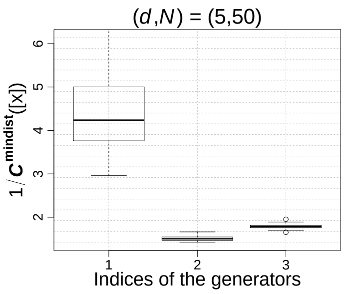
(e) Minimax



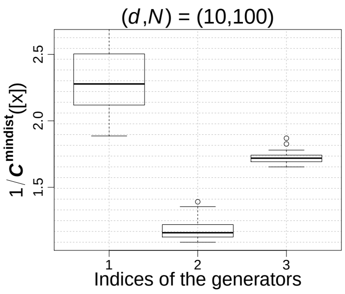
(f) Minimax



(g) Inverse of the mindist



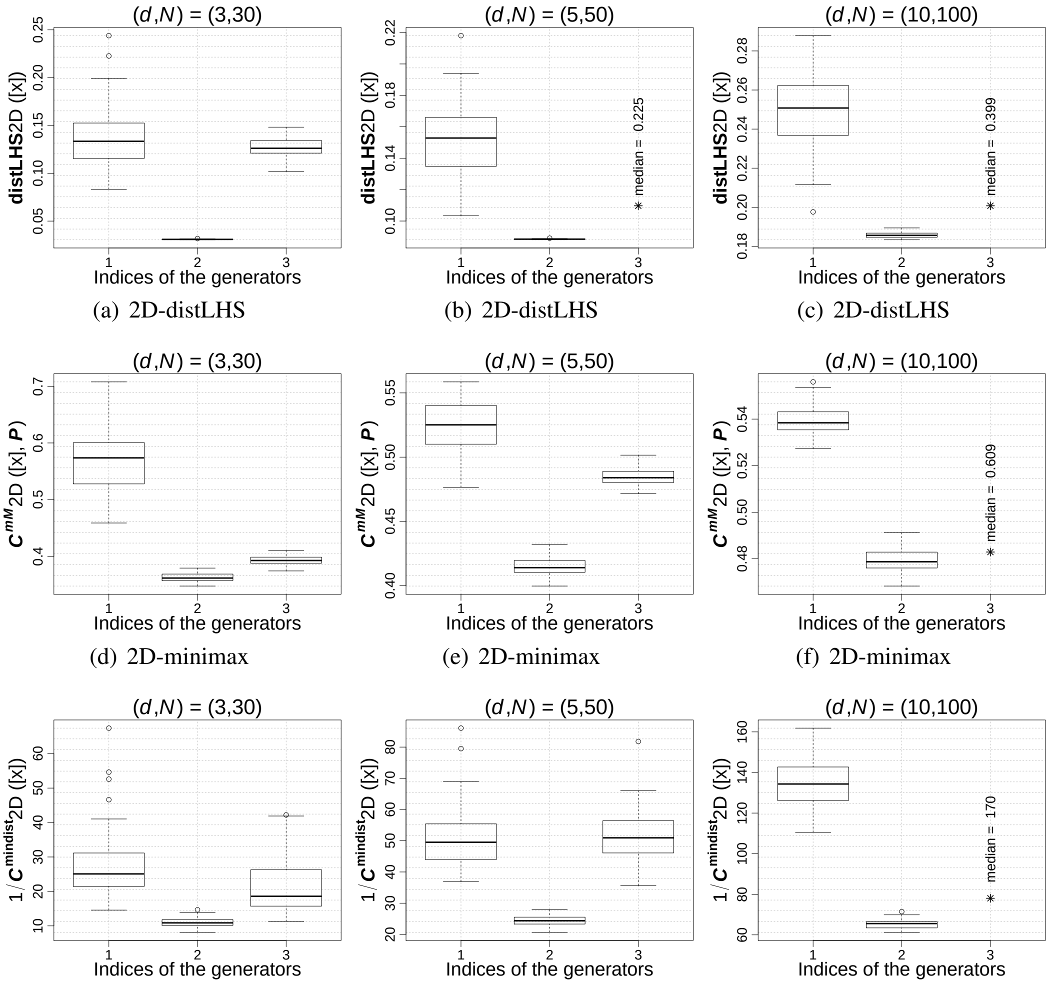
(h) Inverse of the mindist



(i) Inverse of the mindist

FIGURE 13. *Comparison in terms of space filling properties of three generators in the hypersphere with respect to the distLHS criterion (first row), the minimax criterion (second row) and the mindist criterion (third row), for $d = 3$ (left column), $d = 5$ (middle column) and $d = 10$ (right column), and $N = 10d$. 1: "normal approach" / 2: the proposed method / 3: clustering-based approach.*

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238



(a) 2D-distLHS (b) 2D-distLHS (c) 2D-distLHS

(d) 2D-minimax (e) 2D-minimax (f) 2D-minimax

(g) Inverse of the 2d-mindist (h) Inverse of the 2d-mindist (i) Inverse of the 2d-mindist

FIGURE 14. *Comparison in terms of space filling properties of three generators in the hypersphere with respect to the 2D-distLHS criterion (first row), the 2D-minimax criterion (second row) and the 2D-mindist criterion (third row), for* $d = 3$ *(left column),* $d = 5$ *(middle column) and* $d = 10$ *(right column), and* $N = 10d$*. 1: "normal approach" / 2: the proposed method / 3: clustering-based approach.*

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

4, which was initialized by the concatenation of $[z_1]$ and $[z_2]$, and in which the elements of $[z_2]$ are fixed. In particular, $[x^{\text{opt}}] = [\mathscr{O}([x] \mid \emptyset)]$ is introduced as the unconstrained optimization of $[x]$ by Algorithm 4. We now compare the space filling properties of the five following designs:

- $[x;y]$: the final DoE is constituted of the concatenation of $[x]$ and $[y]$.
- $[x^{\text{opt}};y]$: the final DoE is constituted of the concatenation of $[x^{\text{opt}}]$ and $[y]$.
- $[\mathscr{O}([x;y] \mid \emptyset)]$: the final DoE corresponds to the unconstrained optimization of $[x;y]$ by Algorithm 4.
- $[\mathscr{O}([x;y] \mid x)]$: the final DoE corresponds to the optimization of $[x;y]$ by Algorithm 4, in which the elements of $[x]$ are fixed.
- $[\mathscr{O}([x^{\text{opt}};y] \mid x^{\text{opt}})]$: the final DoE corresponds to the optimization of $[x^{\text{opt}};y]$ by Algorithm 4, in which the elements of $[x^{\text{opt}}]$ are fixed.

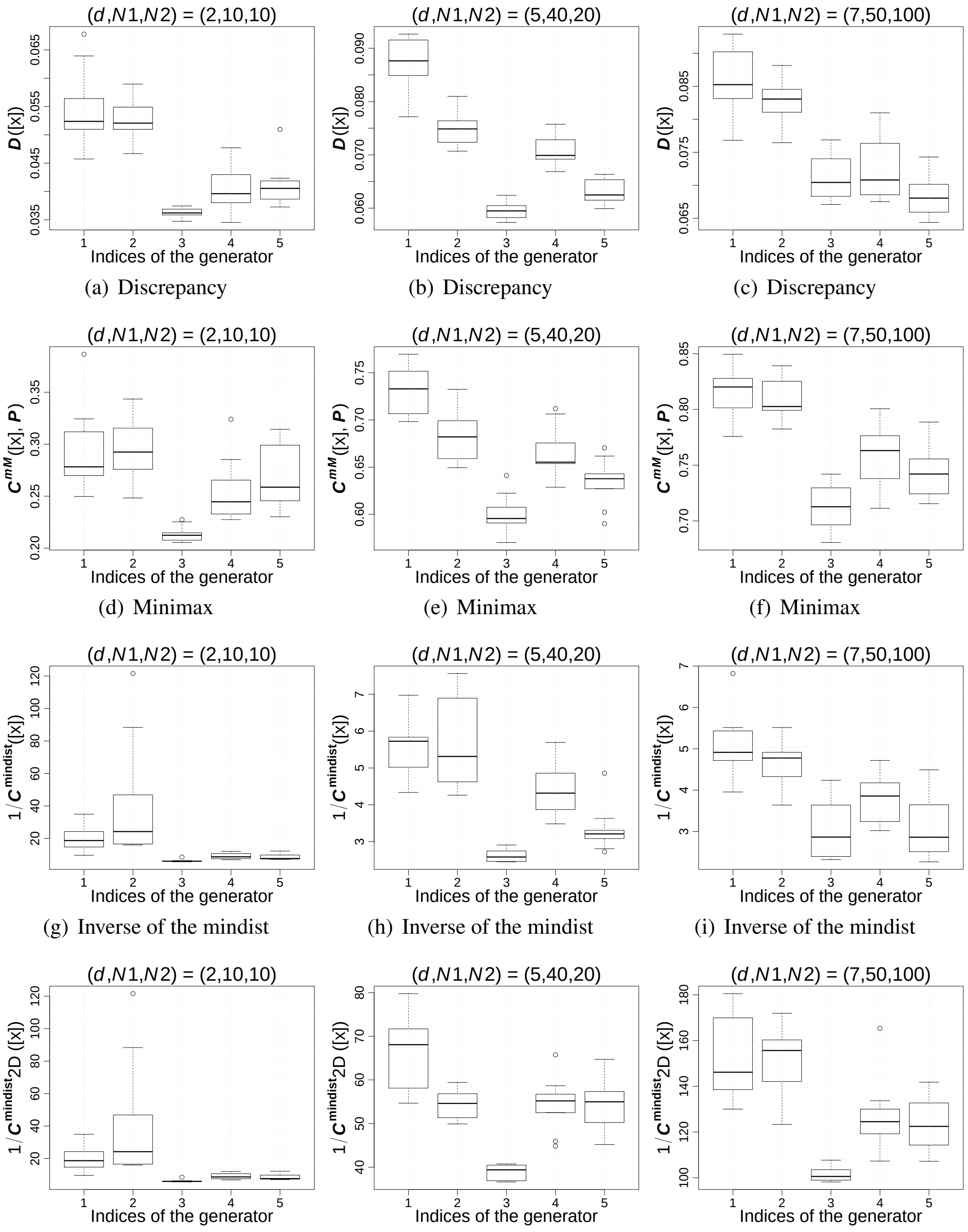
The corresponding results, in terms of minimization of the minimax and of the discrepancy, and maximization of the mindist and the 2D-mindist criteria are shown in Figure 15, for (d,N_1,N_2) in $\{(2,10,10),(5,40,20),(7,50,100)\}$. In this figure, it can be seen that, for all considered values of d,N_1,N_2, the results for $[\mathscr{O}([x;y] \mid x)]$ are much better than the ones for $[x;y]$, and in the same manner, the results for $[\mathscr{O}([x^{\text{opt}};y] \mid x^{\text{opt}})]$ are much better than the ones for $[x^{\text{opt}};y]$. This justifies the interest of the proposed method for the addition of new elements to already existing DoE, for a fixed value of d. Moreover, it can be noticed that the higher N_2 is, compared to N_1, the closer to $[\mathscr{O}([x;y] \mid \emptyset)]$ are the space filling properties of $[\mathscr{O}([x;y] \mid x)]$ and $[\mathscr{O}([x^{\text{opt}};y] \mid x^{\text{opt}})]$. At last, we verify in these figures that the better the already existing DoE is, the more chance there is for the space filling properties of the final DoE to be good.

On the other hand, we now suppose that the value of N is fixed, and we are interested in increasing the dimension of the input space. To this end, let $[x]$ be a (d_1,N)-dimensional DoE and $[y]$ be a (d_2,N)-dimensional DoE, which have been generated by the "randomLHS" routine. Even if increasing d or N lead to two different kinds of concatenation, the previous notations are used again to simplify the reading. The three following configurations are now introduced:

- $[x^{\text{opt}};y^{\text{opt}}]$: the final DoE is the concatenation of the unconstrained optimization of $[x]$ and $[y]$ separately.
- $[\mathscr{O}([x;y] \mid \emptyset)]$: the final DoE corresponds to the unconstrained optimization of $[x;y]$ by Algorithm 4.
- $[\mathscr{O}([x^{\text{opt}};y] \mid x^{\text{opt}})]$: the final DoE corresponds to the optimization of $[x^{\text{opt}};y]$ by Algorithm 4, in which the elements of $[x^{\text{opt}}]$ are fixed.

Finally, in figure 16, we compare the results associated with these three configurations, for $(d_1,d_2,N) \in \{(2,1,30),(2,2,50),(4,3,100)\}$. In this figure, we notice that the results associated with the proposed method, $[\mathscr{O}([x^{\text{opt}};y] \mid x^{\text{opt}})]$, are much better than the ones corresponding to the concatenation of two DoE that have been optimized independently, $[x^{\text{opt}};y^{\text{opt}}]$, from the four considered space filling criteria points of view. Moreover, it can once again be seen that the higher d_2 and N are, the more likely is the improvement of the space filling properties of the final DoE. In particular, in the case $(d_1,d_2,N) = (2,2,50)$, the results of the configuration $[\mathscr{O}([x^{\text{opt}};y] \mid x^{\text{opt}})]$ are very close to the results of $[\mathscr{O}([x;y] \mid \emptyset)]$, which corresponds to the limit case, as this configuration allows the optimization of the position of all the elements of the final DoE. At last, it can be noticed that the configuration $[\mathscr{O}([x;y] \mid \emptyset)]$ is not optimal for the minimization of $1/C_{2D}^{\text{mindist}}$ when $d_1 = 2$. Indeed, for this particular value of d_1, and for this particular criterion, starting from an optimized 2D DoE can be better than a direct global optimization.

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238



(a) Discrepancy (b) Discrepancy (c) Discrepancy

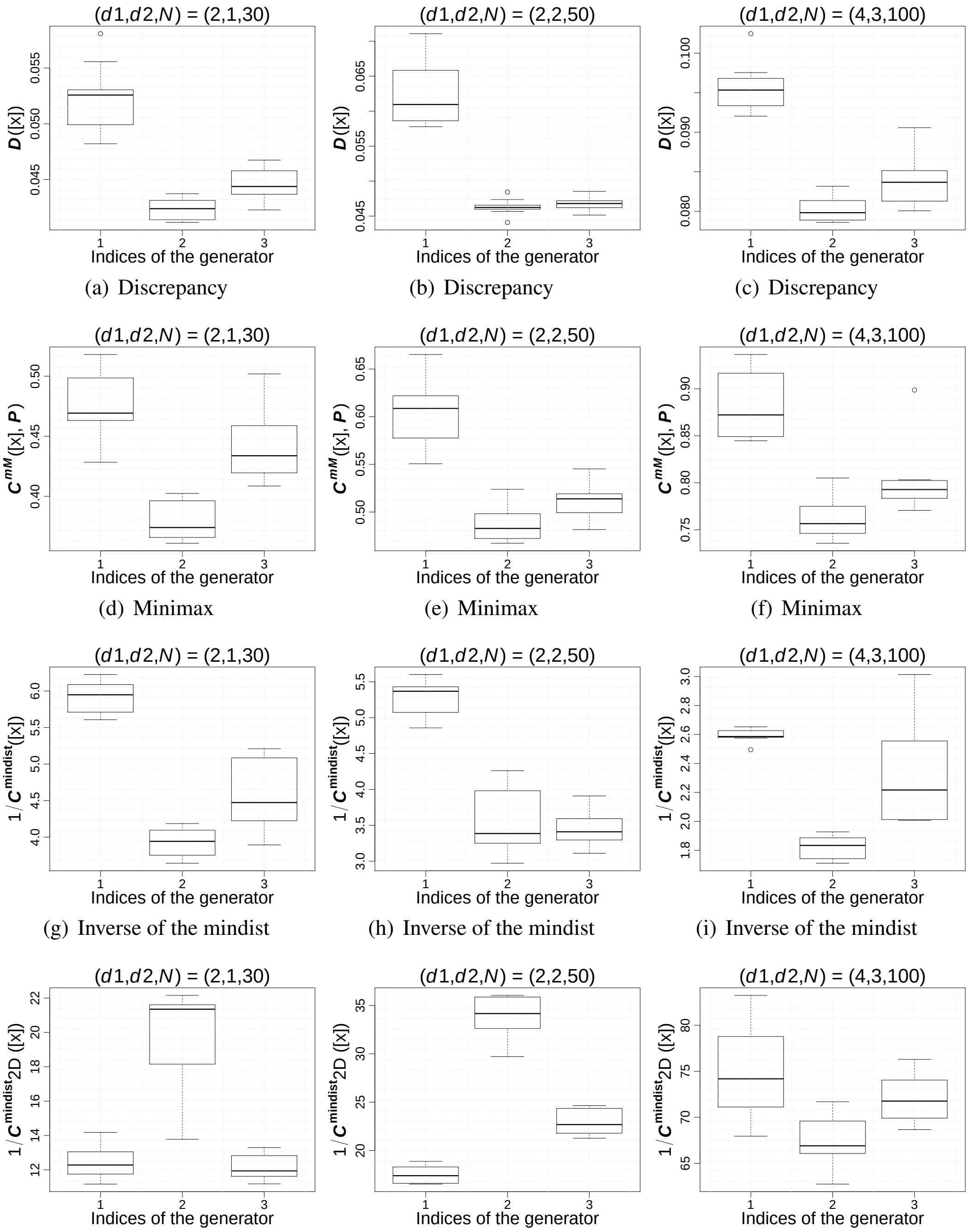
(d) Minimax (e) Minimax (f) Minimax

(g) Inverse of the mindist (h) Inverse of the mindist (i) Inverse of the mindist

(j) Inverse of the 2D-mindist (k) Inverse of the 2D-mindist (l) Inverse of the 2D-mindist

FIGURE 15. *Efficiency of the proposed method for the enrichment of already existing DoE with respect to the L^2-discrepancy (first row), the minimax criterion (second row), the mindist criterion (third row), and the 2D-mindist criterion (fourth row) for* $(d,N_1,N_2) = (2,10,10)$ *(left column),* $(d,N_1,N_2) = (5,40,20)$ *(middle column) and* $(d,N_1,N_2) = (7,50,100)$ *(right column). 1:* $[x;y]$ */ 2:* $[x^{opt};y]$ */ 3:* $[\mathscr{O}([x;y] \mid \emptyset)]$ */ 4:* $[\mathscr{O}([x;y] \mid x)]$ */ 5:* $[\mathscr{O}([x^{opt};y] \mid x^{opt})]$*. Boxplots are produced from 50 runs with different initial DoE.*

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238



(a) Discrepancy (b) Discrepancy (c) Discrepancy

(d) Minimax (e) Minimax (f) Minimax

(g) Inverse of the mindist (h) Inverse of the mindist (i) Inverse of the mindist

(j) Inverse of the 2D-mindist (k) Inverse of the 2D-mindist (l) Inverse of the 2D-mindist

FIGURE 16. *Efficiency of the proposed method for the enrichment of already existing DoE with respect to the L^2-discrepancy (first row), the minimax criterion (second row), the mindist criterion (third row), and the 2D-mindist criterion (fourth row), for $(d_1,d_2,N) = (2,1,30)$ (left column), $(d_1,d_2,N) = (2,2,50)$ (middle column) and $(d_1,d_2,N) = (4,3,100)$ (right column). Boxplots are produced from 50 runs with different initial DoE. 1: $[x^{opt};y^{opt}]$ / 2: $[\mathscr{O}([x;y] \mid \emptyset)]$ / 3: $[\mathscr{O}([x^{opt};y] \mid x^{opt})]$.*

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

5. Conclusions and prospects

This work considers the challenging problem of defining initial designs of experiments (DoE) that allow the exploration of the whole input space, while preserving good space filling properties for each scalar input. To this end, an adaptive method based on two kinds of repulsion has been proposed. This method, which gives very interesting results for the generation of DoE in the hypercube, can be generalized for the generation of optimized space filling designs in any bounded convex domain. To do so, we need, first, to be able to generate an initial DoE in the domain of interest, second, to know if an element is in the domain of interest, and third, to compute (even approximately) the PDF and the CDF of each scalar input. In particular, the relevance of the method has been quantitatively assessed in the cases of the simplex and the hypersphere with respect to several commonly used space filling measures. This method can also be used to add new points to an already existing DoE, while preserving very interesting global space filling properties, which can be very useful for the iterative improvement of surrogate models for instance.

From a numerical point of view, the algorithm being based on repulsion between each elements of the DoE, we point out that the numerical complexity of the proposed algorithm increases almost linearly with respect to the dimension d of the input space, but increases almost quadratically with respect to the number N of elements of the DoE. However, the step by step improvement of the DoE being global, the proposed algorithm appears to converge relatively quickly with respect to the number of iterations Q, such that for values of d lower than 20 and values of N lower than 1000, relevant designs are attainable at a reasonable computational cost. In addition, the only parameter that the user has to choose is Q (and eventually the proportionality factor in Eq. (21)), which makes the proposed method easy to control in practice.

At last, this work is only focused on the LHS structure, which only considers the projection properties of each scalar input. Taking into account higher subprojections properties stays an interesting issue for future works. In the same manner, an important restriction for the method is the assumption that the input space is convex. The adaptation of the proposed method for the generation of DoE in non-convex input spaces, as it is done in Golchi and Loeppky (2015) and Pratola et al. (2015), is also an open question.

References

Auffray, Y., Barbillon, P., and Marin, J. M. (2012). Maximin design on non hypercube domains and kernel interpolation. *Statistics and Computing*, 22(3):703–712.

Bhadeshia, H. K. (1999). Neural Networks in Materials Science. *ISIJ International*, 39(10).

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Bornert, M., Bretheau, T., and Gilormini, P. (2008). *Homogenization in Mechanics of Materials*. Wiley-Iste.

Damblin, G., Couplet, M., and Ioss, B. (2013). Numerical studies of space filling designs: optimization of Latin hypercube samples and subprojection properties. *Journal of Simulation*, 7:276–289.

Devroye, L. (1986). *Non-uniform random variate generation*. Springer.

Draguljić, D., Santner, T. J., and Dean, A. M. (2012). Noncollapsing Space-Filling Designs for Bounded Nonrectangular Regions. *Technometrics*, 54(2):169–178.

Dupuy, D., Helbert, C., and Franco, J. (2015). \pkgDiceDesign and \pkgDiceEval: Two R Packages for Design and Analysis of Computer Experiments. *Journal of Statistical Software*, 65(11):1–38.

Fang, K. T., Li, R., and Sudjianto, A. (2006). *Design and modeling for computer experiments*. Chapman & Hall, Computer Science and Data Analysis Series, London.

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238

Fang, K. T. and Lin, D. K. (2003). Uniform experimental designs and their applications in industry. *Handbook of Statistics*, 22:131–178.

Flury, B. A. (1990). Principal Points. *Biometrika*, 77:34–41.

Franco, J., Corre, B., and Dupuy, D. (2008). Strauss processes: A new space-filling design for computer experiments. *In Proceedings of Joint Meeting of the Statistical Society of Canada and the SFdS*, (11).

Golchi, S. and Loeppky, J. L. (2015). Space Filing Designs for Constrained Domains. 7:1–25.

Hortado, J. (2002). *Structural reliability - Statistical learning perspectives*, volume 17. Lectures notes in applied and computational mechanics, Springer.

Jin, R., Chen, W., and Sudjianto, A. (2005). An efficient algorithm for constructing optimal design of computer experiments. *Journal of Statistical Planning and Inference*, 134(1):268–287.

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–148.

Joseph, V. R., Gul, E., and Ba, S. (2015). Maximum projection designs for computer experiments. *Biometrika*, 102(2):371–380.

Lekivetz, R. and Jones, B. (2015). Fast Flexible Space-Filling Designs for Nonrectangular Regions. *Quality and Reliability Engineering International*, 31(5):829–837.

Likas, A., Vlassis, N., and J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461.

Mak, S. and Joseph, V. R. (2016). Minimax designs using clustering. pages 1–24.

McKay, M., Beckman, R., and Conover, W. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245.

Moon, H., Dean, A. M., and Santner, T. J. (2011). Algorithms for generating maximin Latin hypercube and orthogonal designs. *Journal of Statistical Theory and Practice*, 5:81–88.

Morris, M. and Mitchell, T. (1995). Exploratory designs for computationnal experiments. *Journal of Statistical Planning and Inference*, 43:381–402.

Muehlenstaedt, T., Gore, W. L., Fruth, J., Roustant, O., and Nationale, E. (2014). Computer experiments with functional inputs and scalar outputs by a norm-based approach. pages 1–22.

Osuolale, K., Yahya, W., and Adeleke, B. (2003). Optimal orthogonal-array-based latin hypercubes. *Journal of Applied Statistics*, 30(5):585–598.

Park, J. S. (1994). Optimal Latin-hypercube designs for computer experiments. *Journal of Statistical Planning and Inference*, 39:95–111.

Pillards, T. and Cools, R. (2003). *Transforming low-discrepancy sequences from a cube to a simplex*. Katholieke Universiteit Leuven.

Pratola, M. T., Harari, O., Bingham, D., and Flowers, G. E. (2015). Design and Analysis of Experiments on Non-Convex Regions. *Technometrics*, 1706(November):1–37.

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, 4:409–435.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The design and analysis of computer experiments*. Springer, New York.

Schaback, R. (1995). Error estimates and condition numbers for radial basis function interpolation. *Adv. Comput. Math.*, 3(3):251–264.

Scholkopf, B. and Smola, A. J. (2002). *Learning With Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Boston.

Stinstra, E., den Hertog, D., Stehouwer, P., and Vestjens, A. (2003). Constrained maximin designs for computer experiments. *Technometrics*, 45(4):340–346.

Stinstra, E., den Hertog, D., Stehouwer, P., and Vestjens, A. (2010). Uniform designs over general input domains with applications to target region estimation in computer experiments. *Computational Statisticsand Data Analysis*, 51(1):219–232.

Tang, B. (1993). Orthogonal array-based Latin hypercubes. *Journal of the American Statistical Association*, 88:1392–1397.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley-Interscience, New York.

© Société Française de Statistique et Société Mathématique de France (2017) ISSN: 2102-6238