Analyse des plans d'expérience SFD de grande dimension par l'arbre de longueur minimale

Olivier Vasseur, Jessica Franco, Sidonie Lefebvre, Michelle Sergent

ANALYSE DES PLANS D'EXPERIENCES SFD DE GRANDE DIMENSION PAR L'ARBRE DE LONGUEUR MINIMALE.

Olivier Vasseur[a], Jessica Franco[b], Sidonie Lefebvre[a] , Michelle Sergent[c]

a. ONERA/DOTA, Chemin de la Hunière et des Joncherettes, 91761 Palaiseau Cedex.
b. Total EP/GSR/TG/G&I, CSTJF Avenue Larribau 64018 PAU Cedex.
c. Université Paul Cézanne Aix Marseille III, LMRE, 13397 MARSEILLE Cedex 20.

Résumé

Dans le domaine de l'expérimentation numérique, lorsque les relations entre la réponse et les entrées du code de calcul sont complexes, les plans d'expériences Space Filling Designs (SFD) sont utilisés pour l'exploration du code ou la construction de métamodèles. Il est alors nécessaire que les points de ces plans soient répartis au mieux dans l'espace d'étude. Nous présentons dans un premier temps les avantages qu'offre l'utilisation de critères basés sur l'Arbre de Longueur Minimale (ALM) pour qualifier différents types de répartition de points dans des espaces de grande dimension. Les résultats sont ensuite illustrés par divers types de plans d'expériences (hypercubes latins, suite à faible discrépance, plans de Strauss et WSP) pour des dimensions d'espace variant entre 10 et 55. En conclusion, l'intérêt des plans de Strauss et WSP est mis en évidence.

Summary

In the field of computer experiments, when the relations between the answer and the entries of the computer code are complex, the Space Filling Designs (SFD) are used to study the response evenly throughout the region or to build metamodels. It is necessary that the points of these plans are distributed as well as possible in the space of study. We present the advantages of the use of criteria based on Minimal Spanning Tree (MST) to qualify various types of distribution of points in high dimension spaces. The results are then illustrated by various types of design experiments (hypercubes Latins, low discrepancy suites, Strauss plans and WSP) for dimensions of space between 10 and 55. Ultimately, the interest of the plans of Strauss plans and WSP is pointed out.

Mots clés

Plans d'expériences, Space Filling Desing, Arbre de Longueur Minimale.

Introduction

Le développement de codes qui modélisent ou simulent des phénomènes complexes sont de plus en plus réalistes et même si la puissance des ordinateurs augmente sans cesse, les temps de calcul demeurent importants et limitent ainsi le recours aux techniques de Monte-Carlo. Le développement de métamodèles permet alors de remplacer le simulateur par un outil plus "simple" construit à partir du simulateur complexe. Ces métamodèles ou surfaces de réponse sont en général des fonctions obtenues à l'aide de méthodes d'interpolation ou d'approximation à partir d'un nombre limité d'exécutions du simulateur sur des jeux de paramètres constituant le plan d'expériences numériques. Du fait des caractéristiques non linéaires et/ou non paramétriques des codes de modélisation ou simulation, il est nécessaire de répartir les points dans l'espace le plus

uniformément possible de façon à capter au mieux le comportement du simulateur. C'est le mode de répartition des points que cherchent à proposer les plans d'expériences Space Filling Designs (SFD).
L'étude de l'uniformité d'une distribution de points est difficile en grande dimension et nécessite l'utilisation de plusieurs critères tels la discrépance et les critères basés sur des calculs de distances entre les points. Dussert, Rasigni G., Rasigni M.et Palmari (1986) ont montré que la construction d'un arbre de longueur minimale (ALM) sur un ensemble de points en dimension 2 permettait de qualifier la répartition des points (ordonnée, aléatoire, amas...). Le critère proposé par Franco, Vasseur, Corre, Sergent (2007) qui s'appuie sur la théorie des graphes et des arbres de longueur minimale (ALM) en particulier permet de qualifier la répartition des points d'un plan d'expériences dans un espace multidimensionnel i.e. de classer ces plans selon leur structure, ce que ne permettent pas les autres critères couramment utilisés, comme cela a été montré pour des dimensions assez faibles (<10).
Nous présentons ici les premiers résultats obtenus en grande dimension (dimension supérieure à 20) que nous comparons à ceux obtenus pour une faible dimension. Nous mettons alors en évidence, dans le cas des grandes dimensions, les défauts des plans basés sur les suites à faible discrépance et l'intérêt des plans de Strauss (cf. Franco, 2008) et WSP[1] dont l'algorithme assure que les points de l'espace sont choisis de telle façon qu'ils sont à la fois au moins à une distance minimale (Dmin) de chaque point déjà inclus dans le plan et également, aussi près que possible du centre de l'hypercube unité (cf. Sergent, 1989 et 1997). Les résultats obtenus en dimension 55 seront présentés lors de la conférence.

Méthodologie de qualification des plans d'expériences par l'Arbre de Longueur Minimale (ALM).

Propriétés de l'ALM

Un arbre de longueur minimale est un arbre-maximal pondéré (à chaque arête est affecté un poids appelé longueur) dont la longueur (somme des longueurs des arêtes), parmi tous les arbres-maximaux associés à la même fonction de pondération, est minimale (cf. Figure 1). L'arbre de longueur minimale n'est pas unique contrairement à l'histogramme des longueurs de branche.

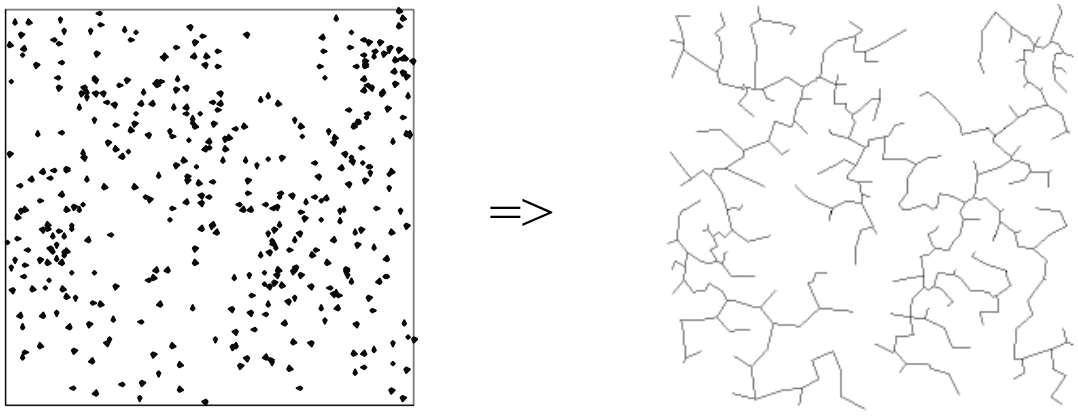


Figure 1 : Construction d'un ALM sur un ensemble de 400 points en dimension 2.

Chaque distribution de points est alors caractérisée par la longueur moyenne m et l'écart-type σ des

[1] WSP : **W**ootton **S**ergent **P**han-Tan-Luu

longueurs de branche. Wallet et Dussert (1998) ont comparé différentes méthodes d'analyse topographique sur des simulations de répartitions de points en dimension 2 et mis en évidence que celle utilisant l'ALM (moyenne m et écart-type σ des longueurs de branches) présentait les meilleures performance en terme de discrimination des structures, de stabilité et d'erreurs. Dans ce plan (m, σ), les distributions de points sont réparties dans des zones différentes en fonction de leur structure (cf. Figure 2).

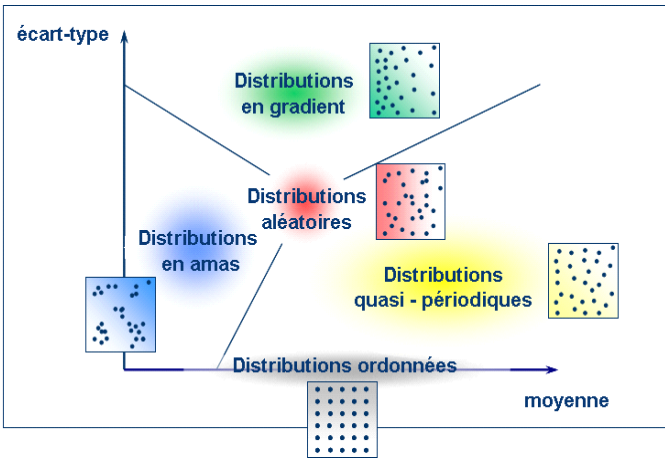
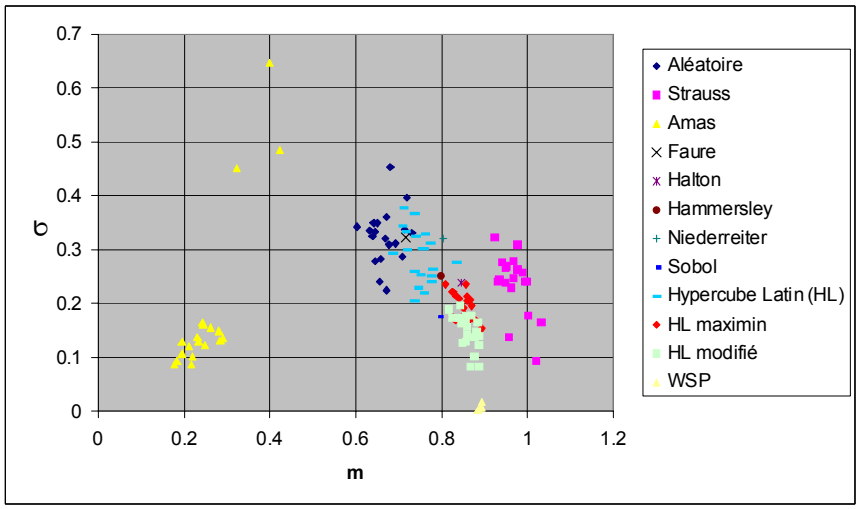


Figure 2 : Représentation de la répartition des distributions dans le plan (m, σ)

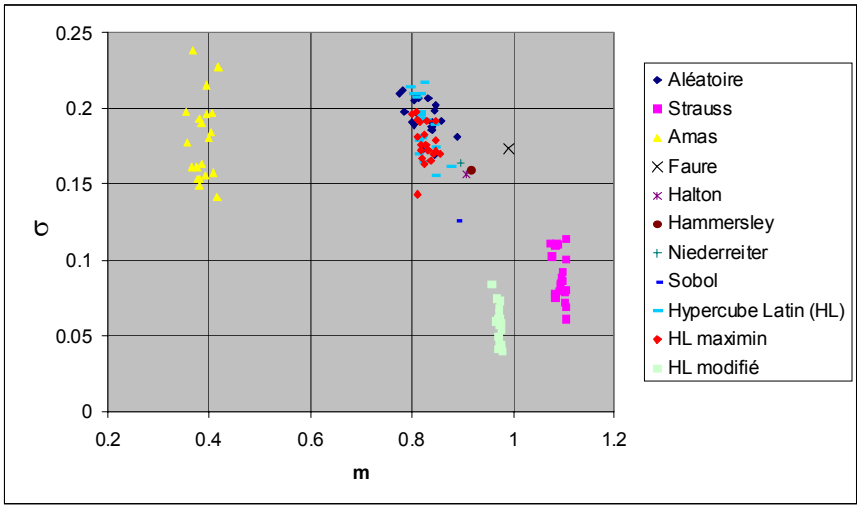
Franco, Vasseur, Corre, Sergent (2007) ont montré que la répartition des distributions définissait également des zones en fonction de leur structure pour des dimensions supérieures à 2 bien que les valeurs m et σ de chaque structure évoluent en fonction de la dimension.
Si aucune connaissance a priori n'est disponible en ce qui concerne le simulateur ou modèle sur lequel on souhaite construire un métamodèle, il est préférable que les points d'un plan d'expériences SFD soient répartis dans l'espace selon des distributions de points quasi-périodiques (i.e. ne sont ni des distributions en gradient, en amas ou aléatoires) en offrant ainsi un bon compromis entre des plans factoriels et les plans aléatoires en évitant les alignements et les zones de vide.

Utilisation de l'ALM pour la qualification des plans SFD.

La construction de l'ALM à partir des répartitions de points de différents plans d'expériences SFD a permis de qualifier quelques plans pour des dimensions variant de 2 à 10.
Nous prenons en compte dans ce travail plusieurs suites à faible discrépance, les plans de Strauss et différents plans construits à partir d'hypercube latin (HL, HL maximin et HL modifié en optimisant la distance euclidienne entre les points du plan) et nous évaluons l'ensemble de ces plans d'expériences SFD pour plusieurs dimensions (2, 5, 10, 20 puis 55).
Sur la Figure 3, on peut constater que l'ensemble des plans d'expériences SFD (et par conséquent les suites à faible discrépance) se situe dans la zone des distributions quasi-périodiques.



(a)

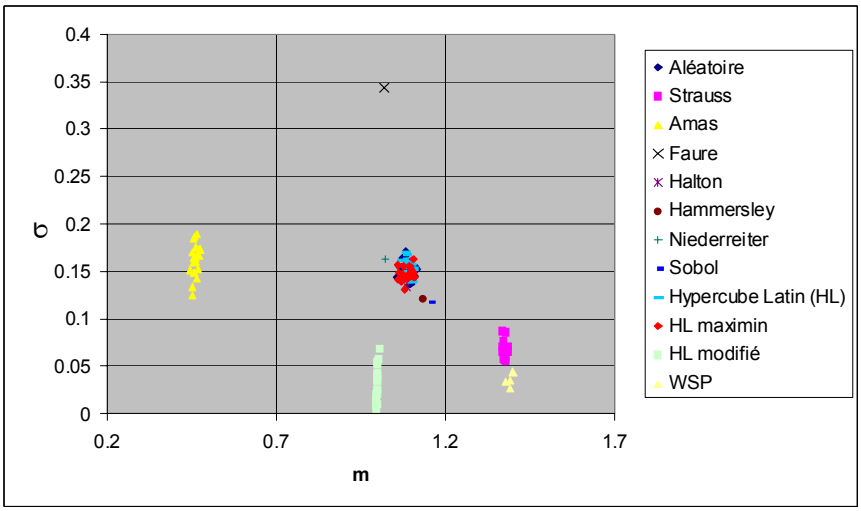


(b)

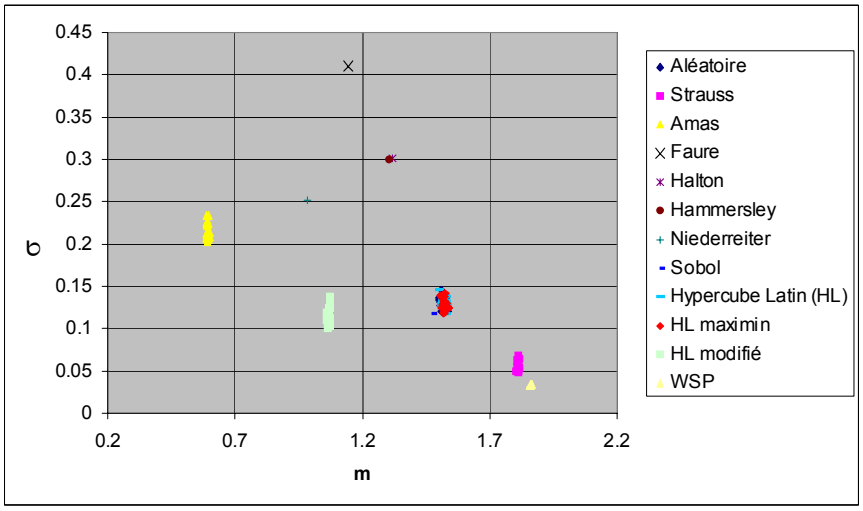
Figure 3 : Représentation de la répartition des distributions dans le plan (m, σ) en dimension 2 avec 20 points (a) et en dimension 5 avec 100 points (b).

En revanche, pour les dimensions 10 et 20 (cf. Figure 4), seuls les plans de Strauss et les WSP présentent longueur moyenne de branche supérieure et un écart-type plus faible que ceux des plans aléatoire et un écart-type plus faible que celui des plans aléatoires, ce qui correspond aux caractéristiques des répartitions quasi-périodiques.

On peut noter que la répartition de la majorité des plans a tendance à être regroupée avec les plans issus de distribution aléatoire. En dimension 20, les répartitions de points issus des suites à faible discrépance ont tendance à évoluer vers la zone en gradient ou en amas. Seule la suite de Sobol est au voisinage des plans aléatoires. L'analyse des longueurs de branche de l'ALM permet de mieux comprendre les résultats obtenus en grande dimension pour ces différentes répartitions de points. Ainsi, la suite de Niederreiter produit un arbre pour lequel les longueurs de branche sont régulièrement réparties dans l'intervalle [0.63, 1.57]. Au contraire la suite de Faure produit un arbre dont les valeurs des longueurs de branche sont réparties notamment sur deux valeurs (~0.35 et ~1.38). Dans ces deux cas, l'écart-type a une valeur importante. Les critiques portant sur les suites à faible discrépance ont souvent souligné (cf. Par exemple Tan et Boyle, 1997) les caractéristiques spécifiques obtenues par projection des points dans certains sous-espaces (plans). Les insuffisances des plans issus de suites à faible discrépance sont ici mises en évidence indépendamment de toute projection, l'ALM étant construit directement sur les points de l'espace d'étude.



(a)



(b)

Figure 4 : Représentation de la répartition des distributions dans le plan (m, σ) en dimension 10 avec 200 points (a) et en dimension 20 avec 400 points (b).

Ainsi en grande dimension, seuls les plans de Strauss et les WSP s'avèrent être les plans d'expériences SFD les plus performants en terme d'exploration de l'espace puisqu'ils présentent un écartement important entre les points sans être tout a fait régulier (σ faible mais non nul). Les résultats concernant ces différents plans en dimension 55 seront présentés lors des 41èmes JdS.

Conclusion

La construction de l'ALM sur les points d'un plan d'expériences permet de qualifier le type de répartition des points que fourni le plan par l'examen de la moyenne et de l'écart-type des longueurs des branches d'une part et de s'affranchir d'analyses par projection dans différents sous-espaces (plans) d'autre part. En grande dimension la plupart des plans d'expériences convergent vers la zone "aléatoire". Les suites classiques à faible discrépance ne permettent pas d'obtenir de bonnes répartitions de points dans l'espace en grande dimension. En revanche, les plans de Strauss et les plans WSP présentent de bonnes propriétés de répartition, même en grande dimension. Les résultats présentés dans ce document seront complétés lors des 41èmes JdS par ceux obtenus en dimension 55.

Les résultats obtenus avec ce critère ont permis de qualifier sur ces différents types de plans et d'évaluer leur évolution en fonction de la dimension. Ces études seront poursuivies par l'analyse d'autres plans d'expériences tels les hypercubes latins minimisant la discrépance.

Bibliographie

[1] Dussert C., Rasigni G., Rasigni M., Palmari J. (1986) Minimal spanning tree: A new approach for studying order and disorder, *Physical Review B*, 34, 5, 3528-3531.

[2] Franco J., Vasseur O., Corre B., Sergent M. (2007) Un nouveau critère basé sur les arbres de longueur minimale pour déterminer la qualité de la répartition spatiale des points d'un plan d'expériences numériques, *Congrès Chimiométrie 2007,* Lyon.

[3] Franco J., Vasseur O., Corre B., Sergent M. (soumis) Minimum Spanning Tree : A new approach to assess the quality of the design of computer experiments, *Chemometrics and Intelligent Laboratory Systems.*.

[4] Franco J. (2008) Planification d'expériences numériques en phase exploratoire pour la simulation des phénomènes complexes, Thèse de doctorat, Ecole Nationale Supérieure des Mines de Saint-Etienne.

[5] Sergent M. (1989). Contribution de la Méthodologie de la Recherche Expérimentale à l'élaboration de matrices uniformes : Application aux effets de solvants et de substituants, Thèse de doctorat, Université Aix Marseille III.

[6] Sergent M., Phan-Tan-Luu R., Elguero J. (1997) Statistical Analysis of Solvent Scales. Part 1, Anales de Quimica Int. Ed., 93, 3-6.

[7] Tan K.S., Boyle P.P. (1997) Applications of Scrambled Low Discrepancy Sequences To Exotic Options, *International AFIR Colloquium Proceedings*, Australia.

[8] Wallet F., Dussert C. (1998) Comparison of spatial point patterns and processes characterization methods, *Europhysics Lett.*, 42, 493-498.