Introduction à la génération des plans d'expériences.

J-B. Blanchard[1,a]

[1]Den-Service de thermo-hydraulique et de mécanique des fluides (STMF), CEA, Université Paris-Saclay, F-91191, Gif-sur-Yvette, France

Résumé Cette note introduit la notion de plan d'expériences en discutant leur intérêt dans le cadre de l'expérimentation réelle et de la simulation numérique. Ces techniques sont particulièrement utiles pour l'analyse de sensibilité et la construction de métamodèles. Plusieurs méthodes sont présentées (suivant les connaissances a priori et les hypothèses sous-jacentes au problème considéré) et sont appliquées à des cas simples pour illustrer leurs intérêts et limitations.

1 Introduction

Le concept de plan d'expériences n'est pas propre aux sciences modernes et a toujours été indissociable de la caractérisation d'un problème. Si l'on souhaite, par exemple, mesurer la masse de 4 objets m_i ($i = 1,\dots,4$) avec une balance à deux fléaux, la méthode évidente est de mesurer chacune des masses individuellement. En supposant que l'erreur de mesure suit une loi normale $\varepsilon \sim \mathcal{N}(0,\sigma^2)$, on obtient 4 mesures $y_i = m_i + \varepsilon_i$ ($i = 1,\dots,4$) donnant une estimation des masses $\hat{m_i} \sim \mathcal{N}(m_i,\sigma^2)$. Une méthode plus sophistiquée consiste à faire les quatre mesures suivantes :

$$\begin{aligned} y_1 &= m_1 + m_2 + m_3 + m_4 + \varepsilon_1 \\ y_2 &= -m_1 + m_2 - m_3 + m_4 + \varepsilon_2 \\ y_3 &= -m_1 - m_2 + m_3 + m_4 + \varepsilon_3 \\ y_4 &= m_1 - m_2 - m_3 + m_4 + \varepsilon_4 \end{aligned}$$

L'estimation de chaque masse se fait en combinant ces mesures, par exemple $\hat{m_4} = \frac{1}{4}\sum_i y_i = m_4 + \sum_i \frac{\varepsilon_i}{4}$. On constate alors que l'estimation résultante suit la loi $\hat{m_i} \sim \mathcal{N}(m_i, \frac{\sigma^2}{4})$ donnant une bien meilleure précision pour autant de mesure.

Cet exemple montre l'intérêt de bien définir le problème et notre but, afin de choisir la stratégie la plus à même de fournir le maximum d'information avec le moins d'effort possible. Dans cette note nous allons introduire différentes méthodologies pour générer des plans d'expériences. Le terme d'**observation** sera utilisé pour parler des résultats d'un **processus**, que ce soit un code ou une expérience.

2 Plans d'expériences pour modèles linéaires

Cette partie introduit le cas de la régression linéaire pour lequel on peut écrire la quantité d'intérêt sous la forme $y(\mathbf{x}) = \sum_{i=1}^{n_r} \beta_i h_i(\mathbf{x}) = \mathbf{h}^T(\mathbf{x})\boldsymbol{\beta}$, où $\mathbf{x} \in \mathbb{R}^p$, les $\{\beta_i\}_{i\in[0,n_r]}$ sont les coefficients de la régression et les $\{h_i\}_{i\in[0,n_r]}$, les régresseurs, sont des fonctions de base (polynômes, fonctions trigonométriques…), n_r étant le nombre de régresseurs [1].

a. e-mail : jean-baptiste.blanchard@cea.fr

Quand $n_r \leq N$ (le nombre d'observations), la détermination de ces coefficients revient à la minimisation de $||\mathbf{y} - \mathbf{H}\boldsymbol{\beta}||^2$ pour $\mathbf{H}$ la matrice des régresseurs, qu'on peut écrire $\mathbf{H} = (\mathbf{h}^T(\mathbf{x}_1)\dots\mathbf{h}^T(\mathbf{x}_N))^T$. La forme générale de la solution est alors $\hat{\mathbf{y}} = \mathbf{H}\hat{\boldsymbol{\beta}} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$.

2.1 Plans factoriels et fractionnaires

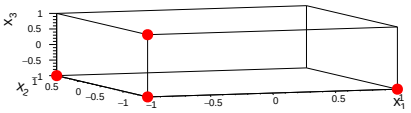
Suivant les hypothèses faites sur le modèle défini ci-dessus, il existe une zoologie très vaste des plans optimaux (ici au sens du plus petit nombre de points à calculer) pour obtenir une mesure des coefficients du premier ordre [2]. Parmi ces derniers, on trouve :

- le plan OAT (*One-At-a-Time*, *c.f.* fig. 1a), si on suppose aucune interaction entre les variables d'entrée. Il correspond à une variation dans chacune des dimensions, ne couvrant que partiellement l'espace des entrées, mais ne nécessitant que $p+1$ estimations.
- le plan factoriel complet (*c.f.* fig. 1b), si on suppose que des interactions peuvent exister entre toutes les entrées. Il correspond à la variation simultanée de toutes les dimensions, engendrant la génération de 2^p calculs, nombre divergeant rapidement avec p.
- les plans factoriels fractionnaires (*c.f.* fig. 1c), suivant les hypothèses sur les interactions. Il existe un grand nombre de dénomination correspondant aux différents cas : R_{III}, R_{IV}, R_V… [2].

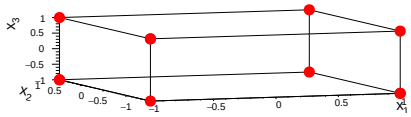
2.2 Plans optimaux

Les plans d'expériences discutés précédemment sont dits optimaux au sens de minimiser le nombre de calcul pour obtenir un résultat. Il est aussi possible de réfléchir à des plans dits optimaux pour un but plus précis dépendant du résultat voulu. Avec le formalisme du début, comme

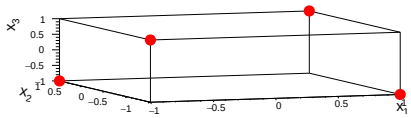
$$\hat{\boldsymbol{\beta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}, \text{ alors } \mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1} = \frac{\sigma^2}{N}\mathbf{M_N}^{-1},$$



(a) OAT



(b) Factoriel complet



(c) Factoriel fractionnaire (R_{III})

FIGURE 1: Exemple de plans “optimaux” pour mesurer les effets du premier ordre d’un modèle linéaire, dépendant de trois entrées dont les valeurs sont normalisées, sans interaction (a), avec toutes les interactions possibles (b) et avec interactions possibles deux-à-deux (c).

où $\mathbf{M_N}$ est la matrice d’information (ou de Fisher, $\in \mathbb{R}^{p\times p}$). Il existe une grande variété de plans optimaux, maximisant une fonction scalaire de $\mathbf{M_N}$ [3], parmi lesquels les plans

- A-optimaux qui maximisent -trace($\mathbf{M_N}^{-1}$) pour minimiser la somme quadratiques des axes de l’ellipsoïde de confiance des paramètres $\hat{\boldsymbol{\beta}}$ obtenus.
- E-optimaux qui maximisent $\lambda_{min}(\mathbf{M_N}^{-1})$ pour minimiser l’axe principal de l’ellipsoïde de confiance des paramètres $\hat{\boldsymbol{\beta}}$ obtenus.
- D-optimaux qui maximisent $\log(\det(\mathbf{M_N}^{-1}))$ pour minimiser le volume de l’ellipsoïde de confiance des paramètres $\hat{\boldsymbol{\beta}}$ obtenus.

L’exemple donné en section 1 est un cas de plan optimisé.

3 Plans d’expériences remplissant l’espace de définition

Parallèlement aux méthodes introduites en section 2, les plans discutés ci-après ont pour but de représenter au mieux l’espace de définition des entrées. Ce type de plan est souvent utilisé quand aucune hypothèse n’est faite sur le modèle ou sur l’importance ou non d’une région (lors d’une analyse de sensibilité ou pour faire l’apprentissage du comportement d’un processus par un modèle de substitution).

La section 3.1 introduit des propriétés génériques de ces plans, que ces derniers soient générés par des méthodes quasi Monte-Carlo (illustrées en section 3.2.1 par les suites à faibles discrépances) ou par des méthodes Monte-Carlo (illustrées en section 3.2.2).

3.1 Propriétés attendues et recherchées

3.1.1 Couverture optimale de l’espace

Il existe plusieurs critères pour décrire la couverture spatiale (le remplissage) d’un plan d’expériences [4], comme

- le *minimax*, pour minimiser la distance maximale entre un point du domaine et un point du plan.
- le *maximin*, pour maximiser la distance minimale entre les points du plan.
- la *discrépance*, pour quantifier de combien un échantillon diffère d’une distribution purement uniforme.

3.1.2 Robustesse en sous-projection

Les processus étudiés ont souvent une dimension effective (k) plus faible que leur dimension d’entrée (p), par exemple si certains paramètres sont négligeables. La robustesse en sous-projection est la faculté de conserver une bonne couverture dans le sous-espace $[0,1]^k$ et dépend principalement de la nature du plan considéré [5].

3.1.3 Séquentialité

La séquentialité est la propriété pour un plan de pouvoir recycler les calculs précédents pour augmenter le nombre de points au fur et à mesure (afin d’améliorer la précision d’un estimateur par exemple).

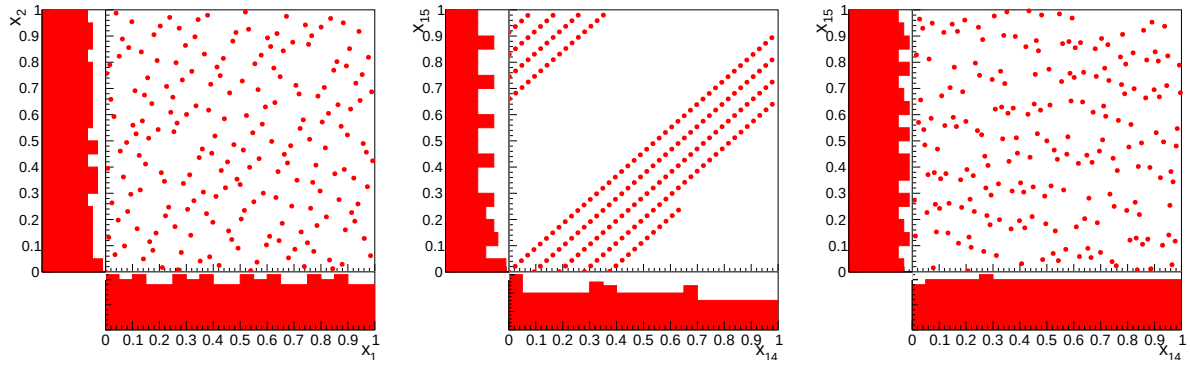
3.2 Génération de ces plans

3.2.1 Suite à faible discrépance

Cette partie introduit une classe de méthode quasi Monte-Carlo générant des plans déterministes, appelées suites à faible discrépance car leur but est de remplir l’espace des entrées en minimisant cette dernière tout en respectant la propriété de séquentialité.

La fig. 2 regroupe les projections de plan de 200 points tiré dans $[0,1]^{15}$ (donc à 15 dimension). Si le résultat projeté sur (x_1,x_2) est bon pour une suite de Halton [4] (*c.f.* fig. 2a), la projection sur (x_{14},x_{15}) montre une corrélation qui n’était pas demandée (*c.f.* fig. 2b). Cette dernière est due à la manière dont fonctionne la séquence : elle repose sur l’utilisation de nombre premier comme base ($\{b_i\}_{i=1,..,15}$) pour déterminer le position des point dans chacune des dimensions. Le remplissage se fait dans un certain ordre dépendant de ces nombres premiers. Ainsi pour ne pas avoir cette structure, il faudrait demander un nombre de point de l’ordre de $N \sim b_{14} \times b_{15}$ ($b_{14} = 43$ et $b_{15} = 47$).

Il existe d’autres séquences, comme la séquence de Sobol [6], montrant un meilleur comportement dans les mêmes conditions (*c.f.* fig. 2c), mais cette dernière est aussi sensible au fléau de la dimension.



(a) Halton, x_2 *vs* x_1 (b) Halton, x_{15} *vs* x_{14} (c) Sobol, x_{15} *vs* x_{14}

FIGURE 2: Exemple de projection de plan d'expériences à 200 points, tirés selon 15 lois uniformes $[0,1]$, à partir d'une suite de Halton (a et b) et de Sobol (c) en ne regardant que les deux premières variables (a) ou les deux dernières (b et c).

3.2.2 Plan d'expériences aléatoires

Dans cette partie, les plans aléatoires sont illustrés par la fig. 3 dans le cas où les deux entrées considérées suivent une loi uniforme $[0,1]$.

La méthode la plus évidente consiste à tirer de manière aléatoire [1] des valeurs suivant chacune des dimensions. Ces plans (appelés SRS pour *Simple Random Sampling*) sont très simples à générer et ont une bonne séquentialité (tant la graine du générateur pseudo-aléatoire utilisé pour augmenter l'échantillon est changée). Toutefois, les estimations des paramètres du processus faites à partir de ce genre de plan ont des incertitudes relativement grandes. La fig. 3a montre aussi la faible robustesse en sous-projection , même 1D : les distributions marginales souffrent de grandes fluctuations.

Une deuxième méthode consiste à scinder l'espace des probabilités de chacune des dimensions en N intervalles équiprobables et de tirer aléatoirement dans chaque intervalle de chaque dimension. L'association des tirages faits suivant chaque dimension pemet obtenir le plan d'expériences [7]. Ces plans, dits stratifiés (appelés LHS pour *Latin Hypercube Sampling*), ont une bonne robustesse en sous-projection 1D (*c.f.* fig. 3b) au dépend de la séquentialité : il n'est pas possible d'ajouter des points à un plan LHS déjà construit tout en respectant le caractère stratifié du plan.

Finalement, il est possible de construire des plans couvrants (dits *space-filling*) en utilisant un critère tel que ceux définis en section 3.1.1. À partir du plan LHS de la fig. 3b, on peut construire un plan LHS maximin, (*c.f.* fig. 3c) ayant une meilleur discrépance. Ce type de plan est avantageux en terme de précision pour le calcul d'estimateur Monte-Carlo reposant sur une mesure d'intégrale [8].

1. on utilise ici un générateur *pseudo-aléatoire* : un algorithme donnant une suite de nombre semblant aléatoire caractérisé seulement par la graine de départ (appelée *seed*).

4 Plans d'expériences adaptatifs

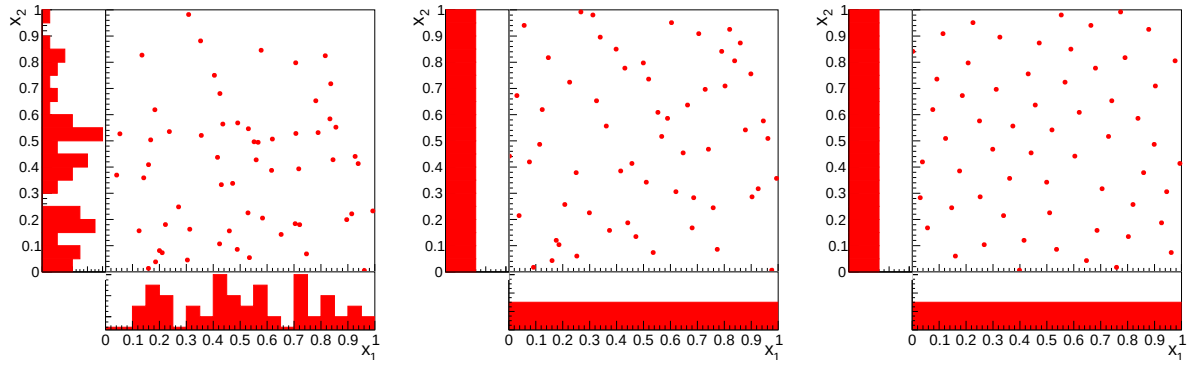
Les plans introduits jusqu'ici peuvent être définis dans un premier temps, avant de procéder à l'estimation de chaque points par le processus. Il est possible de créer des plans dont les observations sont définies au fur et à mesure des estimations, surtout si le processus est très coûteux en ressources.

A partir d'un premier plan, servant de base d'apprentissage pour construire un modèle de substitution, on peut définir un critère pour déterminer le point le plus adapté. La fig. 4 est un exemple sur un processus jouet inconnu (ligne bleue), utilisant un modèle de krigeage qui fournit une estimation (ligne rouge) de la réponse du processus ainsi qu'une incertitude (bande rouge) sur cette dernière [9]. On peut alors choisir la prochaine observation, comme la valeur maximale de l'incertitude de prédiction du modèle de krigeage (représentée aussi en noir dans les fenêtres inférieures). La fig. 4a représente la première construction du modèle krigeage à partir d'un plan LHS (points noirs). Les étapes suivantes sont illustrées par les différentes figures :

— Fig. 4b : le point autour de 2.5 a été calculé par le processus et intégré à la base d'apprentissage du krigeage, tout comme le suivant (autour de 6, en vert). Le modèle de krigeage réévalué donne une bonne description pour les temps inférieurs à 10 (les incertitudes de prédiction sont de l'ordre de 70), mais décrit mal le reste (différence bleu-rouge).

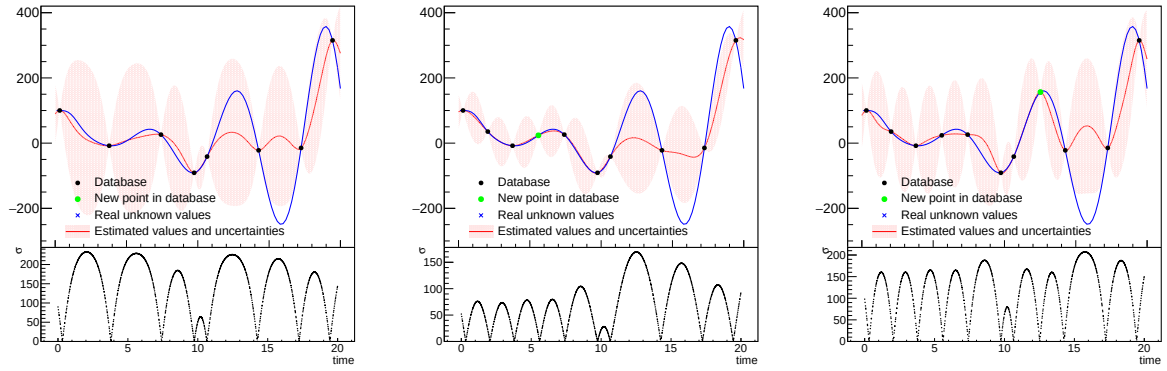
— Fig. 4c : le point autour de 12.5 a été estimé. Le modèle de krigeage a été réévalué et les incertitudes prédites pour la partie inférieure à 10 ont été largement augmentées.

L'arrêt de cette procédure repose généralement sur un critère de qualité du modèle de substitution (donc sur la bonne description des observations par le modèle). Cette méthode, fastidieuse pour le cas simple de la fig. 4, permet



(a) SRS (b) LHS (c) LHS maximin

FIGURE 3: Exemple de plan d'expériences à 60 points, tiré selon deux lois uniformes $[0, 1]$, par la méthode SRS (a), LHS (b) et LHS maximin (c). Dans chaque cas le nuage de points est entouré des projections sur x_1 et x_2.



(a) étape 0 (b) étape 2 (c) étape 3

FIGURE 4: Création d'un plan adaptatif, à partir d'un plan LHS (points noirs en fig. 4a) et d'un modèle de krigeage (rouge) pour estimer les nouvelles observations à réaliser (points verts). Le comportement réel du processus est représenté en bleu.

de se focaliser sur les régions les plus importantes de l'espace des entrées.

5 Conclusion

Le concept de plan d'expériences a été introduit dans cette note en présentant plusieurs méthodologies, dépendant du but de l'étude considérée et des hypothèses sur le modèle sous-jacent.

Références

1. Wikistat, "Régression linéaire simple — wikistat," 2016. [En ligne ; Page disponible le 21-janvier-2016].
2. R. L. Plackett and J. P. Burman, "The design of optimum multifactorial experiments," *Biometrika*, pp. 305–325, 1946.
3. A. Atkinson, A. Donev, and R. Tobias, *Optimum experimental designs, with SAS*, vol. 34. Oxford University Press, 2007.
4. H. Niederreiter, *Random number generation and quasi-Monte Carlo methods*, vol. 63. Siam, 1992.
5. G. Damblin, M. Couplet, and B. Iooss, "Numerical studies of space-filling designs : optimization of latin hypercube samples and subprojection properties," *Journal of Simulation*, vol. 7, no. 4, pp. 276–289, 2013.
6. I. M. Sobol', "On the distribution of points in a cube and the approximate evaluation of integrals," *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, vol. 7, no. 4, pp. 784–802, 1967.
7. R. L. Iman and W. J. Conover, "A distribution-free approach to inducing rank correlation among input variables," *Communications in Statistics - Simulation and Computation*, vol. 11, no. 3, pp. 311–334, 1982.
8. M. D. McKay, R. J. Beckman, and W. J. Conover, "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, no. 2, pp. 239–245, 1979.
9. J.-B. Blanchard, "Introduction aux modèles de substitution," *I3P book of notice*, 2020.