Plans d'expérience numériques d'information de Kullback-Leibler minimale

Astrid Jourdan, Jessica Franco

PLANS D'EXPÉRIENCES NUMÉRIQUES D'INFORMATION DE KULLBACK LEIBLER MINIMALE

Astrid Jourdan & Jessica Franco

E.I.S.T.I.
26 avenue des lilas
64062 Pau cedex 9

TOTAL - DGEP/GSR/TG/G&I
Avenue Larribau
64018 Pau Cedex

RESUME

Les utilisateurs de codes numériques onéreux en temps de calcul souhaitent réduire le coût en limitant le nombre de simulations suivant un choix judicieux fondé sur l'utilisation de plans d'expériences adaptés au contexte numérique et appelés « space filling designs ». Afin de remplir au mieux l'espace des paramètres, nous proposons une méthode de construction de plans dont les points sont le plus uniformément répartis dans l'hypercube unité. Pour mesurer l'écart entre la fonction de densité associée aux points du plan et celle de la loi uniforme, nous utilisons l'information de Kullback-Leibler, ce qui revient par ailleurs à utiliser l'entropie de Shannon. Celle-ci est estimée par une méthode de Monte Carlo dans laquelle la fonction de densité est remplacée par son estimation par noyaux gaussiens.

Mots-clés : space filling designs – estimation de l'entropie – méthodes à noyaux

ABSTRACT

Experimental designs are tools which can drastically reduce the number of simulations required by time-consuming computer codes. Because we don't know the true relation between the response and inputs, designs should allow one to fit a variety of models and should provide information about all portions of the experimental region. One strategy for selecting the values of the inputs at which to observe the response is to choose these values so they are spread evenly throughout the experimental region, according to "space-filling designs". In this article, we suggest a new method based on comparing the empirical distribution of the points in a design to the uniform distribution with the Kullback-Leibler information. The considered approach consists in estimating this difference or, reciprocally, the Shannon entropy. The entropy is estimated by a Monte Carlo method where the density function is replaced by its kernel density estimator.

Keywords : space filling designs – entropy estimation – kernel density estimation

Introduction

Depuis quelques années, la simulation numérique modélise des phénomènes toujours plus complexes. De tels problèmes, généralement de très grande dimension, exigent des codes de simulation sophistiqués et très coûteux en temps de calcul (parfois plusieurs jours). Dans ce contexte, le recours systématique au simulateur devient illusoire. L'approche actuellement privilégiée consiste à définir un nombre réduit de simulations organisées selon un plan d'expériences numériques et d'adapter, à partir de celui-ci, un métamodèle pour approcher le simulateur.

Dans le cadre de ces travaux, nous nous sommes intéressés à la construction des plans d'expériences en phase exploratoire *i.e.* lorsque la dépendance entre les entrées et les sorties est *a priori* inconnue. Il est alors difficile de prévoir quel type de métamodèle va convenir. Les plans élaborés lors de cette phase exploratoire doivent donc s'affranchir de toute contrainte par rapport à un type de métamodèle (régression linéaire, krigeage, réseaux de neurones) puisque celui-ci est choisi à l'issue de ces premières simulations.

Ainsi, les expériences de ces plans doivent remplir au mieux l'espace des paramètres afin d'obtenir des informations dans tout le domaine expérimental et notamment pour détecter les éventuelles irrégularités à l'intérieur du domaine de simulation. On cherche ainsi un plan dont les points seraient le plus uniformément répartis dans l'hypercube unité. Similairement à la discrépance, l'information de Kullback-Leibler (1951) permet de mesurer l'écart entre la distribution empirique

et la loi uniforme. L'idée est alors de construire, de façon empirique, des plans d'information KL minimale à l'aide d'un simple algorithme d'échange.

Information de Kullback Leibler

Supposons que les points du plan sont les *n* réalisations d'un vecteur aléatoire $X=(X_1,...,X_d)$ de fonction de densité inconnue *f* de support le cube unité $E=[0,1]^d$. L'objectif est de construire un plan dont la fonction de densité associée est la plus proche possible de celle de la loi uniforme sur E. L'information de Kullback-Leibler (information KL),

$$I(f,g)=\int_E f(x)\ln\left(\frac{f(x)}{g(x)}\right)dx$$

permet de mesurer l'écart entre les deux fonctions de densité *f* et *g* de support E. Elle est définie positive et I(*f*,*g*)=0 si et seulement si *f*=*g* p.p. L'objectif est donc de minimiser cette expression dans le cas où la fonction *g* est la densité de la loi uniforme sur E. On a alors

$$I(f)=\int_E f(x)\ln\big(f(x)\big)dx=-H[f].$$

Ainsi, minimiser l'information KL revient à maximiser l'entropie. On retrouve la définition des plans à entropie maximale couramment utilisés en planification numérique (Shewry & Wynn, 1987, Currin *et al.*, 1988). L'originalité de l'idée proposée ici vient du fait que nous sommes en phase exploratoire et que nous n'avons aucun modèle sous-jacent. La maximisation de l'entropie n'a donc pas pour objectif d'augmenter la quantité d'information (au sens de Shannon) contenue dans l'échantillon relativement à des paramètres du modèle. En revanche, il est bien connu que la loi uniforme maximise l'entropie des lois à support dans [0,1]. L'entropie du plan est donc négative et faire tendre cette entropie vers 0, revient à s'approcher d'une distribution uniforme.

Estimation de l'entropie par méthode de Monte Carlo

L'article de Beirlant *et al.* (1997) présente un état de l'art concernant les méthodes d'estimation de l'entropie et leurs propriétés. Dans le cas multidimensionnel, Joe (1989) propose l'estimation suivante. Etant donné que l'entropie s'écrit sous la forme d'une espérance,

$$H(X)=-E\big[\ln\big(f(X)\big)\big],$$

la méthode de Monte Carlo fournit un estimateur sans biais et convergent de l'entropie

$$\hat{H}(X)=-\frac{1}{n}\sum_{i=1}^{n}\ln f(X_i).$$

Cette estimation fait intervenir la fonction de densité *f* inconnue mais pouvant être estimée à partir de $X_1,...,X_n$ à l'aide d'une méthode à noyaux (Silverman, 1986),

$$\hat{f}(x)=\frac{1}{nh^d}\sum_{i=1}^{n}\mathcal{K}\left(\frac{x-X_i}{h}\right),\ \forall x\in[0,1]^d$$

La taille de la fenêtre *h* est choisie selon la règle de Scott (1992),

$$\hat{h}_j=n^{-1/(d+4)}\hat{\sigma}_j,\ j=1,...,d$$

où $\hat{\sigma}_j$ est l'écart-type empirique de la $j^{ème}$ coordonnée X_j.

Joe (1989) montre alors que l'estimateur de l'entropie est biaisé et que le biais dépend naturellement de la taille de l'échantillon *n* et de la dimension *d*, mais aussi de la fenêtre *h*. Dans le cadre de la construction d'un plan optimal, il n'est pas nécessaire d'avoir une estimation exacte de l'entropie. En revanche, il est indispensable que le biais soit fixe au cours de l'algorithme d'échange. C'est pourquoi, l'écart-type estimé est remplacé par celui de la loi uniforme sur [0,1],

$$\hat{h}=\frac{1}{\sqrt{12}}\frac{1}{n^{1/(d+4)}}.$$

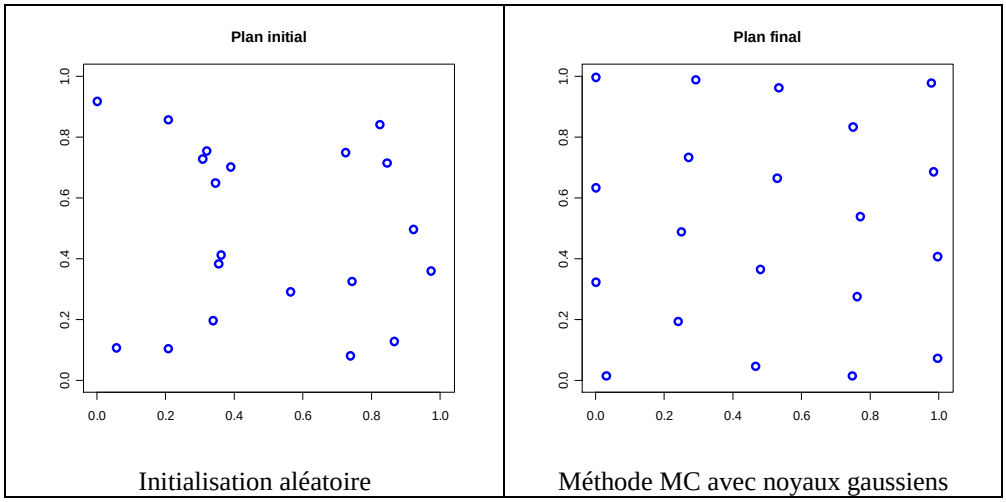
Le noyau $\mathcal{K}$ est choisi gaussien

$$\mathcal{K}(z) = \frac{(2\pi)^{-d/2}}{s^d} \exp\left[-\frac{1}{2s^2}\|z\|^2\right],$$

où s^2=d/12. Un noyau uniforme ou d'Epanechnikov ne peut convenir lorsque la dimension *d* est grande car alors la probabilité que le noyau soit non nul devient extrèment faible.

Caractéristiques des plans d'information KL minimale

- Les points du plan sont disposés au bord du domaine mais aussi à l'intérieur à la façon d'une grille régulière légèrement perturbée (plans dits quasi-périodiques). Ainsi le remplissage de l'espace est garanti tout en conservant une distribution aléatoire.
- Chaque paramètre est testé sur un grand nombre de niveaux avec toutefois des répétitions aux extrémités. De même que le critère maximin (Koelher & Owen, 1996), le critère proposé ici privilégie les bords du domaine et non l'intérieur lorsque *n* est petit.



Initialisation aléatoire | Méthode MC avec noyaux gaussiens

Résultat de l'algorithme d'échange pour un plan de dimension 2 et de taille 20

Comparaison avec les plans usuels

Afin de comparer les plans en dimension d≥2, il est nécessaire de faire appel à des critères usuels permettant de juger du bon remplissage de l'espace ainsi que de la distribution uniforme.

- La mesure de recouvrement (Cov) permet de mesurer l'écart entre les points du plan et ceux d'une grille régulière (Gunzburger *et al.*, 2004). Ce critère est nul pour une grille régulière. L'objectif est donc de le minimiser pour se rapprocher d'une grille régulière, et ainsi assurer le remplissage de l'espace, sans toutefois l'atteindre pour respecter une distribution uniforme notamment en projection sur les axes factoriels.
- Johnson *et al.* (1990) ont introduit les distances maximin et minimax afin de construire des plans répondant à la question de remplissage de l'espace. Le critère maximin (Mindist) consiste à maximiser la distance minimale entre deux points du plan.
- La discrépance permet de mesurer l'écart entre la fonction de répartition empirique des points du plan et celle de la loi uniforme. Contrairement aux deux critères précédents, la discrépance n'est pas basée sur la distance entre les points. Il existe différentes mesures de discrépance (Niederreiter, 1987, Thiémard, 2000). Nous retenons la discrépance en norme L2 (DL2).

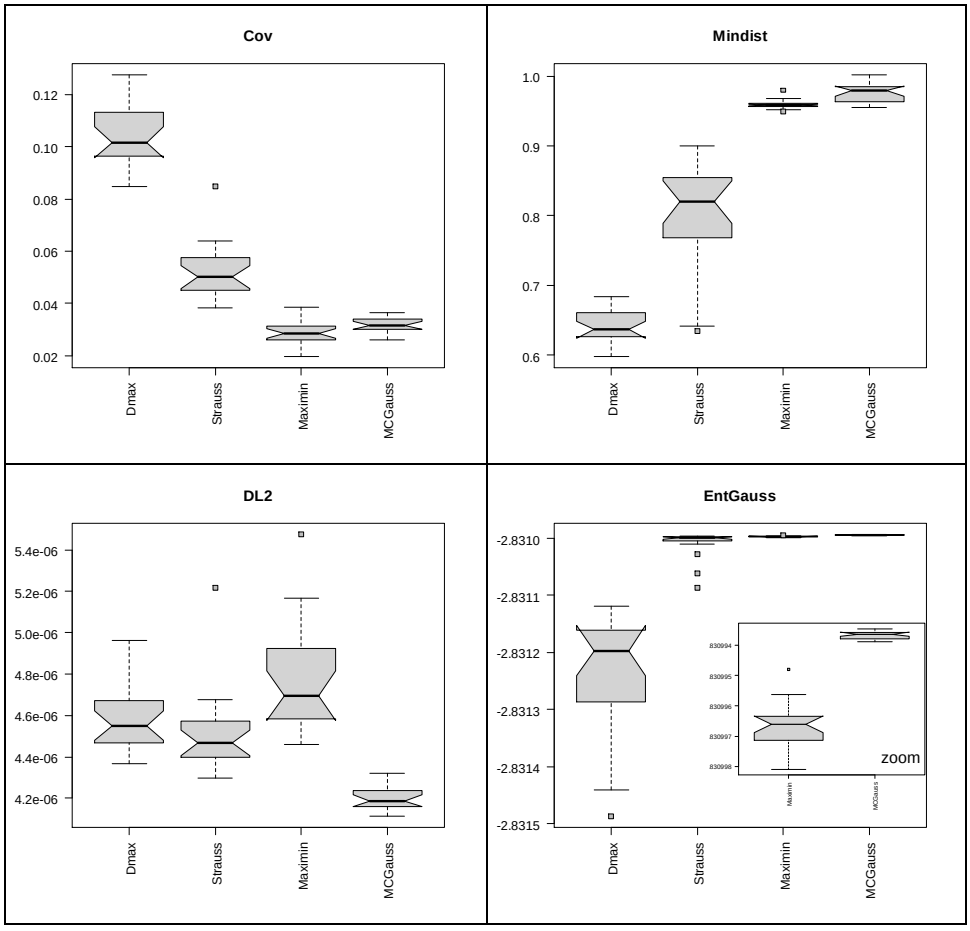
Les plans construits ici sont comparés avec les plans usuellement utilisés en expériences numériques. Nous conseillons la lecture des travaux de Koehler & Owen (1996) et Franco (2008) pour un état de l'art concernant les « space filling designs ».

- Dmax : plans à entropie maximale (Shewry & Wynn, 1987, Currin *et al.*, 1988) construits de façon à maximiser le déterminant d'une matrice de covariance. Ces plans sont ainsi très utilisés lorsque la surface de réponse est ajustée par krigeage. Ils supposent cependant un

modèle sous-jacent.

- Strauss : plans élaborés à partir d'un processus de Strauss qui considère de la répulsion entre les points de manière à remplir au mieux l'espace des paramètres (Franco, 2008)
- Maximin : plans optimaux pour le critère maximin (Johnson *et al*., 1990).
- MCGauss : plans d'information KL minimale construits par la méthode décrite précédemment.

Les suites de faibles discrépance (Halton, Sobol, ...) n'ont pas été représentées car les résultats ne sont satisfaisants que pour le critère de discrépance.



Représentation des critères usuels calculés sur 20 plans à 100 points en dimension 10

Les plans d'information KL minimale (MCGauss) donnent incontestablement les meilleurs résultats quel que soit le critère. Ils semblent concurrencer les plans maximin traditionnellement utilisés en phase exploratoire, et ceci même pour le critère de distance maximin.
On note que les résultats des plans MCGauss sont très peu dispersés. Cela signifie qu'ils dépendent peu de l'initialisation de l'algorithme d'échange. On peut donc limiter le nombre d'initialisations aléatoires dans l'algorithme et ainsi réduire le temps de calcul.

Perspectives

Malgré des conditions éloignées de l'asymptotique, un critère basé sur l'estimation de l'entropie fournit de très bons résultats. Cependant, lorsque la dimension augmente, la méthode d'estimation par Monte Carlo devient moins performante lorsque l'échantillon est de petite taille. De plus, étant donnée que la densité doit être ré-estimée à chaque échange, cette méthode requière un temps de calcul important. Il pourrait être intéressant de calculer le critère à l'aide d'une estimation de l'entropie qui évite, à la fois la méthode de Monte Carlo, et à la fois l'estimation de la fonction de densité. Par exemple, la méthode des plus proches voisins (Kozachenko et Leonenko, 1987) se prête

particulièrement bien aux grandes dimensions (Leonenko *et al*.). Une autre perspective consiste à définir le critère à partir d'une entropie différente de celle de Shannon, par exemple, l'entropie de Rényi qui peut être estimée à partir des arbres de longueur minimal (Hero *et al*., 2002), ou bien l'entropie de Tsallis qui, à l'ordre deux et dans le cas d'une méthode d'estimation de la densité par noyaux gaussiens, s'écrit de façon analytique comme une somme des noyaux et évite ainsi l'erreur d'approximation de la méthode de Monte Carlo (Bettinger *et al*., 2008).

Bibliographie

[1] Beirlant J., Dudewicz E.J., Györfi L., Van Der Meulen E.C. (1997). Nonparametric entropy estimation : an overview. Int. *J. Math. Stat. Sci*., 6(1) 17-39.

[2] Bettinger R., Duchêne P., Pronzato L., Thierry E. (2008). Design of experiments for response diversity. In *Proc. 6th International Conference on Inverse Problems in Engineering (ICIPE), Journal of Physics: Conference Series}*, Dourdan (Paris), 15-19 juin 2008, to appear. http://hal.archives-ouvertes.fr/hal-00290418/fr/

[3] Currin C., Mitchell T., Morris M., Ylvisaker D. (1988). A bayesian approach to the design and analysis of computer experiments. ORNL Technical Report 6498, available from the national technical information service, Springfield, Va. 22161.

[4] Franco J (2008). *Planification d'expériences numériques en phase exploratoire pour des codes de calculs simulant des phénomènes complexes*. Thèse présentée à l'Ecole Nationale Supérieure des Mines de Saint-Etienne

[5] Gunzburger M., Burkardt J. (2004). Uniformity measures for point sample in hypercubes. https://people.scs.fsu.edu/~burkardt/pdf/ptmeas.pdf

[6] Hero A., Bing Ma, Michel O., Gorman J. (2002). Applications of entropic spanning graphs. *Signal Processing Magazine, IEEE*, 19, 85 – 95.

[7] Joe H. (1989). Estimation of entropy and other functional of multivariate density. *Ann. Int. Statist. Math.*, 41, 683-697.

[8] Johnson M.E., Moore L.M., Ylvisaker D. (1990). Minimax and maximin distance design. *J. Statist. Plann. Inf.*, 26,131-148.

[9] Koehler J.R., Owen A.B (1996). *Computer Experiments*. Handbook of statistics, 13, 261-308.

[10] Kosachenko L.F., Leonenko N.N. (1987). Sample estimate of entropy of a random vector. *Problem of Information Transmission*, 23, 95-101.

[11] Kullback S., Leibler R.A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22 79-86.

[12] Leonenko N, Pronzato L, Savani V. A class of Rényi information estimators for multidimensional densities. *Annals of Statistics* , to appear.

[13] Niederreiter H. (1987). Point sets and sequences with small discrepancy. *Monasth. Math.*, 104, 273-337.

[14] Scott D.W. (1992). *Multivariate Density Estimation : Theory, practice and visualization*, John Wiley & Sons, New York, Chichester.

[15] Shewry M.C., Wynn H.P. (1987). Maximum Entropy Sampling. *J. Appl. Statist*., 14, 165-170.

[16] Silverman B.W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall, London.

[17] Thiémard E. (2000). *Sur le calcul et la majoration de la discrépance à l'origine*. Thèse présentée au département de mathématiques de l'école polytechnique fédérale de Lausanne