Echantillonnage : Monte Carlo et hypercube latin et indices basés sur la régression

École Chercheur Mexico

Thierry Faure

Cemagref — LISC, Aubière, France

Gien, le 12 Mai 2009

Plan

Plan

Objectifs et enjeux

Lorsque l'objectif est de modéliser le code en phase exploratoire (quand aucune simulation n'a encore été réalisée), on recherche souvent à satisfaire les deux contraintes suivantes :

- D'une part, répartir les points dans l'espace le plus possible de façon à capter les non-linéarités.
- D'autre part, faire en sorte que ce remplissage de l'espace subsiste par réduction de la dimension.

On cherche à réaliser des expérimentations en remplissant l'espace des paramètres "au mieux"

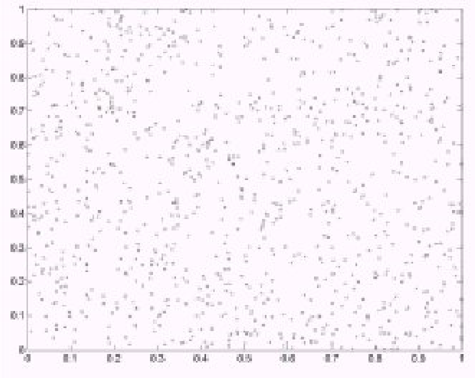
- Comment définir ce "au mieux" ?
- Minimum de points, représentativité

La qualité de la répartition spatiale est mesurée soit à l'aide de critères déterministes comme les distances minimax ou maximin (Johnson et al., 1990), soit à l'aide de critères probabilistes comme la discrépance

Plan

Exemple d'échantillonnage

tirage selon une loi uniforme de n points indépendants de Ω
plot(runif(1000),runif(1000))



Plan

Maximin et Minimax

Utilisation des critères minimax ou maximin (Johnson et al., 1990) pour disperser les points au mieux.

Maximin

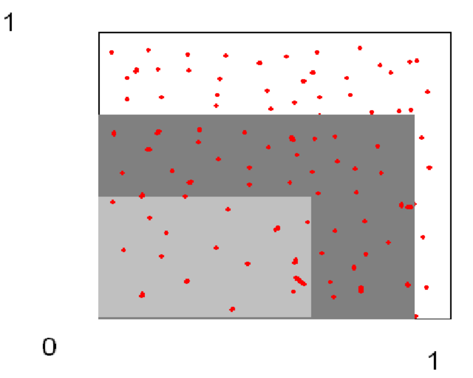
Distance minimum : C'est la plus petite distance entre deux points de l'échantillon : $min_{i \neq j} d(x_i, x_j)$, avec d la distance euclidienne
$d(x_i, x_j) = \sqrt{\sum_{l=1}^{k} (x_{il} - x_{jl})^2}$
Plan maximin : Un plan maximin est un plan qui maximise la distance minimale entre deux points de l'échantillon.

Minimax

Critère de distance maximum :C'est la plus grande distance entre un point du domaine d'étude et un point de l'échantillon. Plan minimax : Un plan minimax est un plan qui minimise la distance maximale entre un point du domaine et un point de l'échantillon.

Discrépance

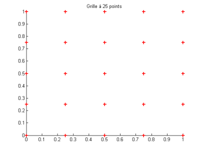
Discrépance : $D^*(X) = \sup_{P \in I^{s*}} \left| \frac{\#(P,(x(n)))}{n} - \lambda(P) \right|$ Mesure l'écart maximal de la distribution des points de la suite à la répartition uniforme.



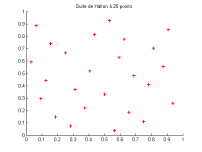
L'inégalité de Koksma-Hlawka montre que la discrépance intervient dans l'erreur de calcul d'une intégrale pour des fonctions à variation bornée.

Dispersion

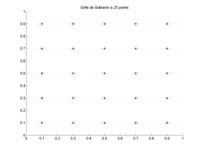
Dispersion : $\delta(x_n) = \max_{y \in I^s} \min_{i=1,\ldots,n} d(y, x_i)$



Regular Grid
(25 points)



Halton sequence
(25 points)



Sukharev grid
(25 points)

Pour les discrépances :

Grille réguliére (avec n points)	Sequence à faible discrepance
$O\left(\frac{1}{\sqrt{n}}\right)$	$O\left(\frac{log^d(n)}{n}\right)$

Plan

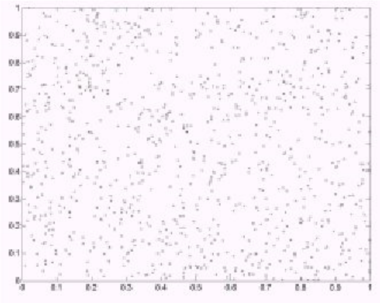
méthode de Monte carlo

Le plan classique dans les méthodes de monte-carlo consiste à générer des suites de points pseudo-aléatoires dans l'espace des facteurs. L'erreur d'estimation découle du théorème central limite. Elle est en $1/\sqrt{n}$, n étant la taille de l'échantillon.
On distingue :

- nombres pseudo-aléatoires
- nombres quasi-aléatoires

nombres pseudo-aléatoires

- tirage selon une loi uniforme de n points indépendants de Ω
- erreur probable d'estimation en $O(\sqrt(n))$
- la vitesse de convergence ne dépend pas de la dimension d de Ω



Nombres quasi-aléatoires

- Utilisation de suite à faible discrépance
- exemples : Niederreiter, Halton, Sobol

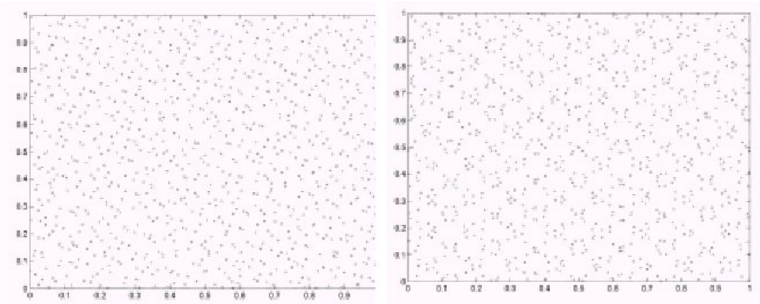


FIG.: Niederreiter et Sobol

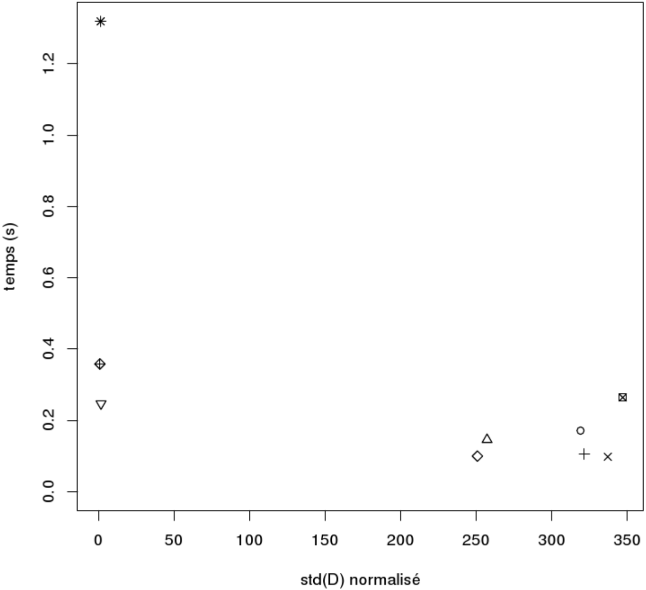


FIG.: Comparaison des méthodes d'échantillonnages

Plan

Hypercube latin

Méthode la plus populaire des “space-filling designs”
Définition :Un plan hypercube latin (LHD) avec n runs et K facteurs d’entrées (noté LHD(n,K) est une matrice $n \times K$, dans laquelle chaque colonne est une permunation aléatoire de { 1, 2, $\cdots$, K }.
Propriétés :

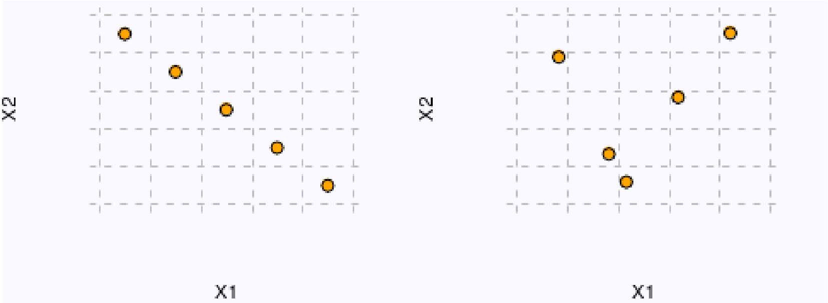
-

Quelques méthodes de construction

- Méthode standard : distribution
- Méthode de construction : Algorithm LHSA
- Itérative
- maximin L

Méthode standard

- Chaque dimension de Ω est découpée en L classes de même amplitude, découpage de Ω en L^d strates,
- Tirages de L strates telles que toutes les classes du découpage des facteurs soient présentes une fois et une seule,
- Tirage d'un point dans chacune des L strates.
- Répétition du processus.



Méthode de construction : Algorithm LHSA (Mc Kay, 1979)

Pour générer un LHS pour n $=$ 8, s $=$ 2, dans la première étape nous générons 2 permutations $\{1, 2, \cdots, 8\}$ comme (2,5,1,7,4,8,3,6) et (5,8,3,6,1,4,7,2) pour former un LHS(8, 2) qui est la matrice sur la gauche. Nous générons alors nombres aléatoires pour former un matrice 8×2. exemple ici

Indices basés sur la régression

Méthodes statistiques reliées à la régression (Venables and Ripley, 1999)
Coefficients de Corrélation utilisés pour mesurer les relations entre les entrées et les sorties du modèle.

- Coefficient PEAR (Pearson Product Moment Correlation Coefficient) $r_{Z_{i,\hat{Y}}} = \frac{\hat{cov}(\hat{Y}, Z_i)}{s_{\hat{Y}} s_{Z_i}}$ Il varie entre -1 et $+1$ et il mesure le degré d'association linéaire entre les variations de Zi et celles de $\hat{Y}$.
- Coefficient PCC : partial correlation coefficient : Mesure l'association entre les Z_i et les $\hat{Y}$ apres élimination des effets possible des facteurs d'entrées $Z_j ineqj$.
- Standardized regression coefficients (SRC) : $\hat{b}_i(\frac{s_{Z_i}}{s_{\hat{Y}}})$ avec $f(z_k) = b_0 + b_i z_{k,i} + \epsilon_{ik}$

Conclusion

- Le modèle de régression peut être étendus pour prendre en compte les intéractions . En particulier lorsque le coefficient de determinaison est faible.
- Les méthodes de régression sont efficaces pour prendre en compte les effets linéaires. D'autres méthodes sont à considérer pour capturer les non linéarité.

Utilisation

```
library(sensitivity) pcc(X, y, rank = FALSE, nboot = 0, conf = 0.95)
src(X, y, rank = FALSE, nboot = 0, conf = 0.95)
plot(x, ylim = c(-1,1), ...)
```

Bibliographie

- G. E.P. Box, et al., 2005, Statistics for experimenters, 2nd edition, Wiley.
- B. Husslage et al., 2008, Space-filling Latin hypercube designs for computer experiments, SSRN-id895464 Report No. 2008–104
- M. E. Johnson et al., 1990, Moore L.M., Ylvisaker D. Minimax and maximin distance designs. J. of Statis. Planning and Inference, 26, 131-148.
- A. Saltelli, K. Chan and E. M. Scott eds, 2000, Sensitivity Analysis, Wiley.
- A. Saltelli et al. 2008, Global sensitivity analysis - The primer, Wiley.
- Thomas J.Santner, et al., 2003, The design and analysis of computer experiments, Springer.
- Venables, W.N., Ripley, B.D., 1999. Modern Applied Statistics with S-PLUS. Springer, Berlin.

Licence