Kullback-Leibler designs

Astrid Jourdan[a], Jessica Franco[b]

ABSTRACT
Space filling designs are commonly used for selecting the input values of time-consuming computer codes. In this paper, the Kullback-Leibler information is used to spread the design points evenly throughout the experimental region. A comparison with the most common designs used for computer experiments shows the high performance of the Kullback-Leibler designs.

KEYWORDS : space filling designs, entropy estimation, kernel density estimation

INTRODUCTION

Engineers and scientists use mathematical models and numerical solutions to describe physical systems. The computer codes are generally time consuming and one strategy consists of replacing the computer model by a "metamodel" for any kind of applications (sensitivity analysis, optimization,…). In this paper, we suppose that no information is available about the relationship between the computer output and the input parameters (exploratory phase). The objective is then to run some simulations according to a space filling design, which should

- provide information about all parts of the experimental region, and then, enable one to spot possible irregularities of the computer response,
- allow one to adapt a variety of statistical models (kriging models, neural network models, ...).

In order to fill up the space in uniform fashion with the design points, we propose a new criterion based on the Kullback-Leibler information for design construction. As with the discrepancy method, the KL information measures the difference between the empirical distribution of the design points and the uniform distribution. The idea is to minimize this difference by using an exchange algorithm.

KULLBACK-LEIBLER DESIGNS

Suppose that the design points $X_1,...,X_n$, are n independent observations of the random vector $X=(X^1,...,X^d)$ with absolutely continuous density function f concentrated on the unit cube $[0,1]^d$. The aim is to select the design points in such a way as to have the density function "close" to the uniform density function. The Kullback-Leibler (KL) information measures the difference between two density functions f and g, and is equal to the opposite of the Shannon entropy if g is the uniform density function. Then, minimizing the KL information, I_{KL}, makes f converge towards the uniform density and amounts to maximizing the entropy H,

$$I_{KL}(f) = \int f(x)\ln\big(f(x)\big)dx = -IE\big[\ln\big(f(X)\big)\big] = -H\big[f\big]$$

The entropy is estimated by a Monte Carlo method (Beirlant *et al*. 1997)

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

where the unknown density function f is replaced by its kernel density estimate (Silverman 1986),

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

The quality of a kernel estimate depends essentially on the value of its bandwidth h (smoothing parameter). In our application, the bandwidth is chosen using Scott's rule (1992). Since the bias of the estimation depends on the bandwidth (Joe 1989), h needs to be fixed during the exchange algorithm. Hence the standard deviation estimates in Scott's rule are replaced with the standard deviation of the uniform distribution on [0,1],

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

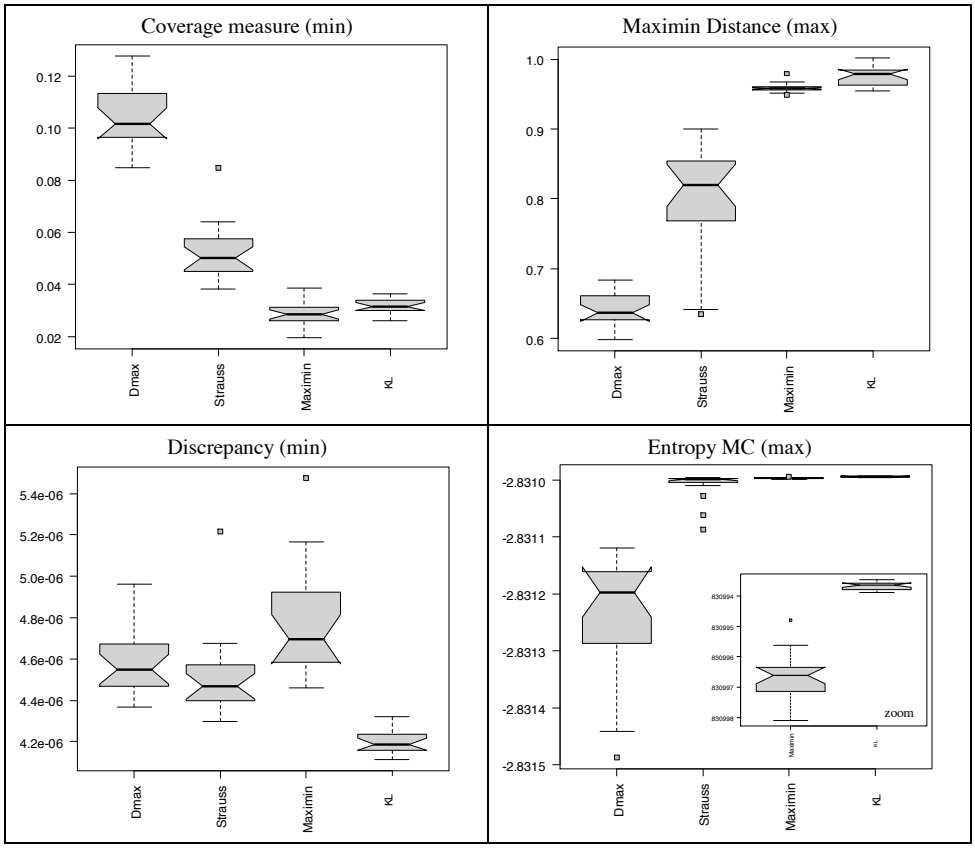
[a] Department of mathematics, E.I.S.T.I, 26 avenue des Lilas, 64062 Pau cedex 9, France
[b] Total- DGEP/GSR/TG/G&I, avenue Larribau, 64018 Pau Cedex , France

The choice of kernel function K is much less important for the behavior of the estimate than the choice of h. Most of the common kernels (uniform, Epanechnikov, triangle, ...) have a bounded support (unit sphere), so that, in our application, the probability that the kernel values are not zero is extremely low. So, the chosen kernel function K is the multivariate Gaussian distribution,

$$K(z) = \frac{(2\pi)^{-d/2}}{s^d} \exp\left[-\frac{1}{2s^2} \|z\|^2 \right], \text{ where } s^2 = \frac{d}{12}.$$

DESIGN COMPARISON

Whatever the initialization, the exchange algorithm converges toward designs with the same characteristics. The points lie on the edge of the experimental region but also in the interior like a scrambled regular grid (quasi-periodical distribution). Such distribution assures that the points are spread evenly in the unit cube and that many levels are tested for each parameter. A comparison with the most common space filling designs shows that KL designs are indisputably the best designs with regard to the usual criteria, even in high dimensions. They compete with maximin designs which are widely used in the exploratory phase.



Criteria for 20 designs of size 100 with dimension 10

REFERENCES

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.

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

Joe H., 1989. Estimation of entropy and other functional of multivariate density. Ann. Int. Statist. Math., 41, 683-697.

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

Koehler J.R., Owen A.B, 1996. Computer Experiments. Handbook of statistics, 13, 261-308.

Kullback S., Leibler R.A., 1951. On information and sufficiency. Ann. Math. Statist., 22 79-86.

Scott D.W., 1992. Multivariate Density Estimation : Theory, practice and visualization, John Wiley & Sons, New York, Chichester.

Silverman B.W., 1986. Density estimation for statistics and data analysis. Chapman & Hall, London.