OPTIMAL DESIGNS FOR THE PROPAGATION OF UNCERTAINTY IN COMPUTER EXPERIMENTS

Laurent CARRARO [(1)], Bernard CORRE [(2)],
Céline HELBERT [(1)], Olivier ROUSTANT [(1)], Sylvain JOSSERAND [(1)]

[(1)] *Département Méthodes et Modèles Mathématiques pour l'Industrie,*
École Nationale Supérieure des Mines de Saint-Etienne
courriel : carraro@emse.fr

[(2)] *Direction Exploration Production,*
Division Techniques Géosciences
Total
courriel :bernard.corre@total.com

ABSTRACT

Response surfaces, or *meta-models*, and *design of experiments* are widely used for various experimental works. But numerous physical phenomenons are studied through complex and costly numerical simulators. In such cases, the *response* is influenced by *factors* but the link between these variables is deterministic. Nevertheless, factors are often known with uncertainty and the influence of this ignorance is important for the practician. Due to the computing time, it is not possible to obtain the uncertainty of the response through a standard Monte Carlo method and an approximation of the simulator, a meta-model, is needed. We present an optimality criterion, the MC-V, in order to evaluate the probability distribution of the response with a minimal error. We chose to apply the criterion on parts of 2 real cases derived from the petroleum industry. The simulator 2^{nd} order polynomial meta-model and the three distributions of input factors (uniform, gaussian, triangular) are among those used in this industry. Unlike the standard optimal D-criterion, the MC-V criterion considers the distribution of input factors. As a result, some points are placed near the center of the domain, where they are most likely to be observed.
Finally, the MC-V criterion provides a realistic tool to assess the quality of the designs used to propagate uncertainty in computer experiments.

Keywords: *design of experiments, computer experiments, response surface, optimality criteria, MC-V criterion, IMSE criterion*

I. INTRODUCTION

Since R.A. Fisher pioneering work, the design of experiments is used to conceive and to plan experimental procedures, in order to provide maximum information by analysing results ([1]). Such planning and analysis is problem dependent. Therefore ad hoc methods are often applied for different experimental objectives such as screening, optimization, robustness testing… These methods have benefited enormously from the development of computers, weakening numerous assumptions made in the first half of the 20^{th} century. We cannot deal

with this whole technological revolution. We will focus on two specific techniques, the Monte Carlo method and the construction of optimal designs.

Before going down this road, let us recall that the exponential growth of computing power has had another great consequence: the emergence of very complex software, simulating phenomena from the modelling of chemical reactions, climate, to evaluation of oil and gas reserves. In theory, computer simulations should simplify. Paradoxically, the more powerful computer resources are, the closer one is able to mimic reality, and the more complex the simulations. Let us take the actual case of the evaluation of oil fields. If we could take into account all the input factors of an average field, they would represent about 10^6 inputs. Obviously reducing the complexity of the inputs is important. This can be accomplished by experts, and with the addition of further information (via seismography, for example), one can even obtain a probability distribution for the input factors. But simplifying the inputs is not enough. Indeed the simulator itself should be simplified. The CPU time is enormous, from one hour to several days. Evidently, the Monte Carlo method, which requires at least 10.000 simulations, cannot be used to propagate the uncertainty of the input factors on the simulator itself.

One of the basic ideas is to replace the simulator by a simple function, the meta-model, on which it is easy to apply the Monte Carlo method. But computer experiments have at least two major differences with respect to real-life experiments. First of all, in the *real world*, two experiments made under identical experimental conditions lead to different measurements. The existence of such differences, consequences of noise, (environmental noise, measurement noise, etc.) leads to the fundamental concept of variability. But in the *virtual world*, apart from numerical instabilities, two experiments under the same conditions lead to the same result. Therefore the concept of variability must be reinvented. This has been explored by numerous authors at the end of the 80's (see [2], [3] or [4] for a recent review). From a frequentist point of view, the basis of their work is to consider the simulator as the realization of a stochastic process. Using the data provided by computer experiments, the approximation of the simulator is a kriging surface, which is utilized often in geosciences. One of the numerous advantages of this approach is that the approximation of the simulator interpolates it at the experimental points. However in practical applications, the meta-models is often a linear regression model (for example, a 2^{nd} order polynomial), which is easier to use especially for assessing uncertainty.

The second main difference between real and computer experiments is that for computer experiments, the unknown variables (as porosity, temperature, pressure, etc.) can be controlled since they are simply inputs of the computer code. Therefore, we can assess the impact of the input uncertainty on the output by propagating uncertainty through the code.

In this paper, we show how to construct designs of experiments adapted to the propagation of uncertainty with a linear regression meta-model. These designs will be obtained from a new optimality criterion, the MC-V criterion. Indeed, tabulated or optimal designs (A, D, E, …) do not consider the probability distribution of input factors. In addition, a new class of designs such Latin hypercubes may be used to take into account the perfect repeatability of computer simulations. This will lead to MC-V optimal Latin hypercube designs.

The article is divided as follows. In section II, we present the 2^{nd} order polynomial meta-model used for the simulator and its approximation. In section III, we build the MC-V criterion towards the objective of propagating uncertainty. Some properties are given in section IV, such as the equivalence between the IMSE criterion. Section V is devoted to examples of optimal designs. Computations of MC-V optimal designs and comparisons to other designs are made in 3 and 8-dimension. Finally, conclusions are given in section VI.

II. SIMULATOR META-MODELLING AND APPROXIMATION

1. Simulator meta-modelling

For practical applications, industrial simulators are often modelled statistically by linear regression:

$$Y_{\text{sim}}(x) = X(x)\beta + \varepsilon(x)$$

with:

- $x = (x_1, ..., x_k)'$, the $k \times 1$ vector of factors values
- A trend (often called « response surface »), $X(x)\beta$, given as a linear combination of basis functions (for instance polynomials):
  - $X(x) = [f_1(x), ..., f_p(x)]$, the $1 \times p$ vector of basis functions values at x
  - β, a $p \times 1$ vector of unknown parameters, to be estimated
- the error $\varepsilon(x)$, given as a random variable with normal distribution $N(0, \sigma^2)$. We suppose that in two points x and y, the errors $\varepsilon(x)$ and $\varepsilon(y)$ are independent. The unknown variance term σ^2 is to be estimated.

We choose a second order polynomial for the basis functions.

2. Simulator approximation

For given experiments $x^{(1)}, ..., x^{(n)}$, the simulator response is approximated in two steps:

a) Least squares estimation of β and σ. The estimators are:
   - $\hat{\beta} = (X'X)^{-1}X'Y$ (Gauss-Markov estimator), with:
     - $X = [X(x^{(1)})', ..., X(x^{(n)})']'$, the experimental matrix $n \times p$
     - $Y = (Y_{\text{sim}}(x^{(1)}), ..., Y_{\text{sim}}(x^{(n)}))'$, the $n \times 1$ vector of responses at experiments
   - $\hat{\sigma}^2 = \dfrac{1}{n-p}\sum_{i=1}^{n}(Y_i - X(x^{(i)})\hat{\beta})^2$

b) Simulator approximation, by replacing β and σ by their estimates in the simulator model:

$$Y_{\text{app}}(x) = X(x)\hat{\beta} + \eta(x)$$

where $\eta(x)$ is normally distributed $N(0, \tau^2)$ with $\tau^2 = \hat{\sigma}^2$, the errors $\eta(x)$ and $\eta(y)$ being independent at two different points.

Remarks

- In linear regression $\hat{\sigma}^2$ is independent of $\hat{\beta}$ (see [5]), so the $\eta(x)$ are independent of $\hat{\beta}$.
- Mathematically, it would be more correct to write the model for Y_{sim} in another form, using stochastic processes, as done in [6]. This has no impact on the results.

III. THE MC-V CRITERION

1. Motivation and definition

Notation. For uncertainty considerations, the factors are viewed as random variables. To make it appear more clearly, we will denote $\tilde{x}$ (x tilde) instead of x in this case to emphasize on the random nature of x.

Motives. As stated before, our aim is to assess uncertainty on the simulator output. In mathematical terms, it means that we have to estimate the distribution of $Y_{\text{sim}}(\tilde{x})$. We will denote this distribution by d_{sim}. Since we do not know $Y_{\text{sim}}(\tilde{x})$, we can only estimate the distribution of the proxy $Y_{\text{app}}(\tilde{x})$ which can be constructed with experiments $x^{(1)},...,x^{(n)}$. Denote d_{app} the distribution of $Y_{\text{app}}(\tilde{x})$. Therefore, one idea is to choose the experiments $x^{(1)},...,x^{(n)}$ to minimize the "distance" between the (unknown) theoretical distribution d_{sim} and the proxy distribution d_{app}. This "distance" will be our criterion.

Constraint. Then, the MC-V criterion (see below for the sense of "MC-V") could be defined as one of the distance criterion between probability distributions, as Kullback-Leibler, χ^2, Hellinger…(see for example [7]). However such criteria are based on a measure of discrepancy between the distributions densities and may not be suitable for applications. Indeed in practise, the knowledge of the factors uncertainty is unsatisfactory. Hence, a criterion based on the probability density may be too sensitive to the poorly known distribution of the inputs. Rather, we may prefer criteria based on more basic elements of distributions, such as moments.

Definition. Finally, for the aforementioned reasons, we have decided to base the MC-V criterion on the first two moments of the distribution of d_{sim} and d_{app}. As the first moments are both equal to $E(X(\tilde{x})\beta)$, we define the criterion by the difference of the second order moments of d_{sim} and d_{app}, or equivalently, the difference of variances:

$$\boxed{C(x^{(1)},...,x^{(n)}) = \text{var}(Y_{\text{app}}(\tilde{x})) - \text{var}(Y_{\text{sim}}(\tilde{x}))}$$

Actually, this expression is positive, as we see in next paragraph and there is no need to use an absolute value. It represents the increase of variance due to the modelling of $Y_{\text{sim}}(\tilde{x})$ by $Y_{\text{app}}(\tilde{x})$. We propose to denote this criterion by MC-V, for Monte Carlo – Variance: the term "Monte Carlo" reminds the user that this criterion is designed to estimate uncertainty of the output (which is currently done by the means of Monte Carlo simulations).

Finally, MC-V optimal designs are those obtained by minimizing the MC-V criterion.

IV. EXPRESSIONS OF THE MC-V CRITERION. EQUIVALENCE TO IMSE.

In this section, we assume that the experimental design $x^{(1)},...,x^{(n)}$ is known. Mathematically, it means that all computations are made conditionally to $x^{(1)},...,x^{(n)}$. We give two expressions of MC-V optimality: the first one makes a link with IMSE-optimality, the second is useful for the criterion implementation.

Firstly, let us recall what IMSE optimality is. The IMSE criterion (see [4]) is defined as the mean of the squared forecast errors:

$$IMSE = \int_{x\in\Delta} E[(\hat{Y}(x) - Y_{\text{sim}}(x))^2]d\mu(x)$$

where $\hat{Y}(x)$ denotes the forecast at point x, μ is the probability distribution of vector x and Δ is the experimental domain. In our regression framework, we have $\hat{Y}(x) = X(x)\hat{\beta}$.
When working with IMSE, the justification of the choice for measure μ is not clear. Uniform measure is widely used, arguing that there is no available information. In an uncertainty framework, one may prefer the *a priori* distribution given by experts (in oil Exploration / Production, μ may be the sub-surface distribution). However, there is no evidence as to why such a choice may lead to an accurate approximation of the output distribution. In the following proposition, we show that this is actually the case and obtain three expressions for the criterion.

Proposition

Let $X = [X(x^{(1)})',...,X(x^{(n)})']'$ *be the experiment matrix for experiments* $x^{(1)},...,x^{(n)}$. *We have:*

i. $IMSE(x^{(1)},...,x^{(n)}) = \sigma^2 + \sigma^2\{E\left[X(\tilde{x})(X'X)^{-1}X(\tilde{x})'\right]\}$

ii. $C(x^{(1)},...,x^{(n)}) = \sigma^2\{E\left[X(\tilde{x})(X'X)^{-1}X(\tilde{x})'\right]\}$

iii. $C(x^{(1)},...,x^{(n)}) = \sigma^2\{m_{X(\tilde{x})}(X'X)^{-1}m_{X(\tilde{x})}' + \text{Tr}((X'X)^{-1}\Gamma_{X(\tilde{x})})\}$, *where* $m_{X(\tilde{x})}$ *and* $\Gamma_{X(\tilde{x})}$ *denote respectively the expectation and covariance matrix of random vector* $X(\tilde{x})$.

Remarks

- The first two expressions show the equivalence of IMSE and MC-V optimality, with $IMSE(x^{(1)},...,x^{(n)}) = \sigma^2 + C(x^{(1)},...,x^{(n)})$. This equivalence justifies *a posteriori* the use of the ISME criterion for uncertainty purposes. Indeed, the IMSE is related to uncertainty issues and is not solely a "mean-squared-errors" expression. Note that IMSE optimality is related to L-optimality and A-optimality (see [4], §6.2.2.). As a consequence, MC-V optimality is related to these criteria as well.
- The second expression shows that the MC-V criterion is positive (since the matrix $X'X$ is positive definite).
- The third expression is useful for practical implementation.

Proof

- Formula (i) can be found in [4], page 171 (the equivalence of notations is the following: *J* lies for $IMSE/\sigma^2$, *f(x)* for *X(x)*, *F* for *X*, and $d\mu(x)=w(x)1_\chi(x)dx$).

- Let us prove the second one. By the conditional variance formula, we have:
  $var(Y_{\text{sim}}(\tilde{x})) = E[\text{var}(Y_{\text{sim}}(\tilde{x})\,|\,\tilde{x})] + \text{var}[E(Y_{\text{sim}}(\tilde{x})\,|\,\tilde{x})]$

$$v\mathrm{ar}(Y_{\mathrm{app}}(\tilde{x})) = E[\mathrm{var}(Y_{\mathrm{app}}(\tilde{x}) \mid \tilde{x})] + \mathrm{var}[E(Y_{\mathrm{app}}(\tilde{x}) \mid \tilde{x})]$$

Now, the two first conditional moments are equal:

$$E(Y_{\mathrm{sim}}(\tilde{x}) \mid \tilde{x}) = X(\tilde{x})\beta = E(Y_{\mathrm{app}}(\tilde{x}) \mid \tilde{x})$$

The first equality is obvious, and the second one is due to the fact that $\hat{\beta}$ is an unbiased estimator. Therefore we obtain:

$$C(x^{(1)},...,x^{(n)}) = E[\mathrm{var}(Y_{\mathrm{app}}(\tilde{x}) \mid \tilde{x})] - E[\mathrm{var}(Y_{\mathrm{sim}}(\tilde{x}) \mid \tilde{x})]$$

Since $\mathrm{var}(Y_{\mathrm{sim}}(\tilde{x}) \mid \tilde{x}) = \sigma^2$, we then have the following expression:

$$\underline{C(x^{(1)},...,x^{(n)}) = E[\mathrm{var}(Y_{\mathrm{app}}(\tilde{x}) \mid \tilde{x})] - \sigma^2}$$

Let us now compute $\mathrm{var}(Y_{\mathrm{app}}(\tilde{x}) \mid \tilde{x})$. We have:

$$\begin{aligned} Y_{\mathrm{app}}(\tilde{x}) - E(Y_{\mathrm{app}}(\tilde{x}) \mid \tilde{x}) &= X(\tilde{x})\hat{\beta} + \eta(\tilde{x}) - X(\tilde{x})\beta \\ &= X(\tilde{x})(\hat{\beta} - \beta) + \eta(\tilde{x}) \end{aligned}$$

Conditionally to $\tilde{x}$, the two terms $\eta(\tilde{x})$ and $\hat{\beta}$ are independent (see 1st remark, § II.2), so we can write:

$$\mathrm{var}(Y_{\mathrm{app}}(\tilde{x}) \mid \tilde{x}) = \mathrm{var}(X(\tilde{x})(\hat{\beta} - \beta) \mid \tilde{x}) + \mathrm{var}(\eta(\tilde{x}) \mid \tilde{x})$$

Therefore, if we denote $\mathrm{var}(\hat{\beta} \mid \tilde{x})$ the covariance matrix of $\hat{\beta}$ (conditionally to $\tilde{x}$), we obtain:

$$\underline{\mathrm{var}(Y_{\mathrm{app}}(\tilde{x}) \mid \tilde{x}) = X(\tilde{x})\,\mathrm{var}(\hat{\beta} \mid \tilde{x})X(\tilde{x})' + \hat{\sigma}^2}$$

To achieve the proof of (ii), we have to compute the expectation of the preceding expression. By Gauss-Markov theorem, we have: $\mathrm{var}(\hat{\beta} \mid \tilde{x}) = \sigma^2(X'X)^{-1}$, and $E(\hat{\sigma}^2 \mid \tilde{x}) = \sigma^2$ ($\hat{\sigma}^2$ is an unbiased estimator). Finally we obtain:

$$E[\mathrm{var}(Y_{\mathrm{app}}(\tilde{x}) \mid \tilde{x})] = \sigma^2 E\left[X(\tilde{x})(X'X)^{-1}X(\tilde{x})'\right] + \sigma^2$$

which leads to the second expression:

$$\underline{C(x^{(1)},...,x^{(n)}) = \sigma^2 E\left[X(\tilde{x})(X'X)^{-1}X(\tilde{x})'\right]}$$

The third expression (iii) in the proposition is a consequence of the second. To see this, one has to remark that if Q is a symmetric matrix and u a centred random vector,

$$E\left[uQu'\right] = \sum_{k,l} q_{k,l} E\left[u_i u_j\right] = \sum_{k,l} q_{k,l}\,\mathrm{cov}(u)_{k,l} = \mathrm{Tr}(Q\,\mathrm{cov}(u))$$

Using this formula with $Q = (X'X)^{-1}$ and $u = X(x) - m_{X(x)}$, we obtain:

$$C(x^{(1)},...,x^{(n)}) = \sigma^2\{E\left[m_{X(\tilde{x})}(X'X)^{-1}m_{X(\tilde{x})}'\right] + E\left[(X(\tilde{x}) - m_{X(\tilde{x})})(X'X)^{-1}(X(\tilde{x}) - m_{X(\tilde{x})})'\right] + 2\times 0\}$$

and finally:

$$\underline{C(x^{(1)},...,x^{(n)}) = \sigma^2\{m_{X(\tilde{x})}(X'X)^{-1}m_{X(\tilde{x})}' + \mathrm{Tr}((X'X)^{-1}\Gamma_{X(\tilde{x})})\}}$$

V. CALCULATION ALGORITHMS AND EXAMPLES.

The aim of this section is to present some examples of MC-V optimal designs. We compare them with standard A-optimal, D-optimal and FCC designs, which do not take into account the factors distribution. We also present the results that can be obtained by restricting to a

useful class of designs, named Latin hypercubes. In all this section, we assume that the experimental domain is the cube $[-1,1]^d$.

1. Some standard designs. Latin hypercubes.

A-optimal, D-optimal designs. The A-criterion and D-criterion are defined to minimize the estimation error in the statistical meta-modelling. In the present linear regression framework, the parameters estimator $\hat{\beta}$ follows the normal distribution $N(0,\sigma^2(X'X)^{-1})$. Then the confidence region around β is an ellipsoid. Minimizing the estimation error can be achieved by minimizing the surfaces of the ellipsoid, which gives the D-optimal designs, or its perimeter, which gives the A-optimal designs ([8]). Formally, the two criteria are as follows:

$$A(x^{(1)},...,x^{(n)}) = \mathrm{Tr}\left((X'X)^{-1}\right)$$
$$D(x^{(1)},...,x^{(n)}) = \det\left((X'X)^{-1}\right)$$

Remark. Another equivalent definition for the D-criterion is $D(x^{(1)},...,x^{(n)}) = \det\left(X'X\right)$. We have chosen the definition above so that optimal designs are obtained by *minimization* of the corresponding criterion.

Face-Centered Cubic (FCC) design. Among the large list of tabulated designs, the FCC is widely used in industrial applications. This design is a special case of central composite design (CC, see [8]), and is also denoted by CCF for central composite face-centered. It consists of all vertices + the center of all faces + the center of the domain. Its size is $2^d + 2d + 1$. In our 8-dimensional case, we will use a fractional FCC design: it consists of all vertices of a (resolution V) fractional 2^{8-2} design + the center of all faces + the center of the domain. Its size is $2^{8-2} + 2\times 8 + 1 = 81$.

Latin hypercubes. Some very useful designs for physical experiments purposes can be of little interest for virtual computer experiments. For instance, suppose that there are 3 variables, but that only 2 of them are really influential, say the two first ones. Then the two vertices (-1,-1,-1) and (-1,-1,1) will give *exactly* the same response since the simulator gives deterministic results, rendering further simulations counterproductive. Introduced by [9] (see also [4]) for sampling, the Latin hypercube designs prevent useless repetitions, by compelling all one-dimensional projections of the *d*-dimensional design to be different. Here, we will define Latin hypercubes in discrete space. Then, a Latin hypercube is a matrix $n\times d$. The first column is a discretization of [-1,1] in *n* parts. Each other column is a permutation of the first one. So we can write:

$$LHd = (x,\sigma_1(x),...,\sigma_{d-1}(x))$$

where x is the $n\times 1$ vector $x = (-1,-1+2/(n-1),...,-1+2(n-2)/(n-1),-1)'$ and $\sigma_1,...,\sigma_{d-1}$ are permutations on x. Consequently, all one-dimensional projections are different. In addition, note that all projections are equally spaced. This is particularly interesting when all input variables follow a uniform distribution. In other cases, the definition can be adapted according to the input distributions of factors. For instance, if all factors follow a triangular distribution, it is more logical to use for x the percentiles of the triangular distribution.

2. Presentation of the algorithms used to get optimal experimental designs

In the last section, we have recalled the definitions of A-optimality and D-optimality criteria, which are the ones often used for experimental design methodology. We have also recalled that Latin hypercube designs (LHd) are commonly used and could be suitable for computer experiments. In part II, we have built a criterion especially adapted to the problem of uncertainty propagation, the Monte Carlo-Variance optimality criteria (MC-V criteria). The criterion clearly depends on the distributions of input factors. In this third part, these factors will vary from -1 to 1 and have the same distribution.

Now, we compare the MC-V optimal designs with A-optimal, D-optimal, FCC designs, restricting or not to the class of Latin hypercubes. Recall that only MC-V optimality takes into account the factors distribution. Here, for the MCV-optimal designs, we will consider three particular cases:

- uniform distribution;
- normal distribution (conditional, knowing that the parameters are between -1 and 1);
- "triangular" distribution.

We have implemented two optimisation algorithms which can search for A, D or MC-V optimal designs.

- The first algorithm works on unrestricted experimental designs. Depending on optimisation progress, Newton or conjugate gradient method is applied. We do not define any discrete grid, therefore design points can be positioned in any place in the experimental domain. However, when there are few variables (especially for 3-dimension), we have often observed that several points could be placed at the same position in the domain. These redundant experiments are useless because computer experiments are deterministic. That is why we have chosen to impose a minimal distance between two points in the same design. This distance can be chosen by users.

- The second algorithm works on the restricted class of Latin hypercube (LH) designs. Here, an exchange algorithm is adapted from Park's algorithm ([10]). Often, experiments are placed on a uniform grid. But, in normal and triangular cases, users can choose between two different discretizations: uniform or adapted (according to the distribution of input factors, see V.1).

In what follows, we will first visualize different types of designs in 3-dimension. Then, we will study 8 dimensional experimental designs and discuss the values of A, D and MC-V criteria.

3. Optimal experimental designs in 3-dimension

In 3-dimension visualization is possible, allowing us to intuitively understand different characteristics of criteria and types of design. The Face-centered cubic (FCC) will be our reference point and will be always set in the top left-hand corner on the next 5 illustrations. All the experimental designs represented here are comprised of 15 experiments.

Now, let us look at experimental designs optimized for D, A and MC-V criteria (for uniform distribution) on figure 1. Notice that all the experiments of D optimal design are at the boundaries of the experimental domain. On the contrary, although A-optimal and MC-V

optimal designs have a majority of points at the boundaries, one or more points are placed at the centre of the domain. Indeed, we have 1 point in this position in A-optimal case and 3 points in MC-V optimal case.

Recall that the MC-V criterion is introduced to propagate the uncertainty of input factors. Therefore, MC-V optimal designs depend on these distributions. That can be verified on figure 2. In normal and triangular cases, the high density of these distributions at the centre of the domain results in a higher number of points at this place: 3 points for uniform distribution, 5 points for normal or triangular. However, a majority of points is again observed at the limits of the domain. This ensures the quality of the estimation of the meta-model used to propagate uncertainty. The others points are inside the domain, where they are most likely to be observed according to the distribution of the input factors.

Please insert figures 1 and 2 here

Figure 3 represents Latin hypercube designs optimized for D, A and MC-V (for uniform distribution) criteria. With LH constraint, experiments seem to be scattered better in the domain. However, differentiating among the D, A and MC-V optimality criteria becomes very difficult. On figure 4, we represent the LH MC-V optimal designs for the three distributions of input factors. Similar conclusions apply, but with an even better dispersal. Differentiating among the three cases is difficult too. Maybe, we can say that there are slightly more points at the centre for triangular distribution.

Please insert figures 3 and 4 here

Figure 5 is similar to figure 4, but the discretization is adapted to the distributions of input factors. The influence of normal or triangular distribution is particularly visible, with many points near the center of the domain.

Please insert figure 5 here

4. Optimal experimental designs in 8-dimension

As visualizing designs in 8-dimension is problematic, we have produced comparative tables in order to appraise the relative quality of each design. There are three tables, one for each distribution of the input factors (uniform, normal, triangular), giving criteria values and computation time. Depending on the distribution choice, only MC-V criterion and MC-V optimal designs (LHd or not) change from one table to another. Let us state that adapted and uniform discretizations are the same for uniform distribution. Finally, let us recall that optimal designs are obtained by *minimization* of the corresponding criterion (see in particular the definition of D-optimality [§ V.1]).

Let us look at table 1, concerning uniform distribution. Firstly, it appears clearly that calculation times are quicker for LHd than for other designs. However, optimisation is restricted to LHd class and the values of criteria are not as appropriate than for ordinary designs. Notice that the tabulated FCC design gives excellent results, but is comprised of more experiments (81 instead of 50).

Secondly, calculating D optimal designs (LHd or not) takes much more time than the others. Generally, computation times are much too high for this dimension. Petroleum companies need to have designs for a minimum of 8 dimensions and to use them in real cases. In addition, optimising MC-V criterion under LHd constraint seems to be not completely reliable every time. Indeed, for example, the best design for the MC-V criterion, the MC-V optimal, is surpassed by the D optimal LHd !

Apparently, the A-optimal design is nearly optimal for the MC-V criterion. We had already observed this in 3-dimension, where A and MC-V optimal design have both at least one point in the centre of the domain. On the contrary, the lack of experiments in the centre of the domain seems to penalize D-optimal design for A and MC-V criteria (and even more for normal and triangular distributions!).

Please insert table 1 here

Lastly, in normal and triangular cases (tables 2 and 3), the LHd with adapted discretizations seem to be less efficient as to criteria values.

Please insert tables 2 and 3 here

5. Industrial applications and perspectives

Computational time is too high at present for industrial use. One idea would be to try other optimization methods such as genetic or evolutionary strategies (ES) algorithms. However, this may not be enough. Another approach could simply be to select some known standard designs (FCC, Box-Benhken…) by calculating their corresponding MC-V criterion value, which is easy to do.

The crucial issues for petroleum companies are inverse modelling problems, and especially history matching. To estimate future production of a hydrocarbon reservoir, engineers have simulators at their disposal. Sometimes, they also have information of past production, which constitutes history data. With an appropriate choice of input parameters, the simulators must approximately give as an answer the real data already measured on the field. The methodology of experimental designs is commonly used to find these appropriate sets of input factors. According to our observations, it seems that both (non LH) MC-V optimal designs as well as LH A and D-optimal designs may be promising in history matching applications.

VI. CONCLUSION

We have introduced a new criterion, the MC-V, to construct optimal designs for the propagation of uncertainty through a simulator. We have obtained several expressions of the criterion and shown its equivalence with IMSE. We have studied two cases, in 3 and 8 dimensions. The visualization of the 3-dimensional MC-V optimal designs show that most of the points are placed on the boundaries of the experimental domain. This ensures a good estimation of the meta-model used to propagate uncertainty. In addition, the others points are near the center of the domain, where they are most likely to be observed according to the distribution of the input factors. For the 8-dimensional case, the MC-V optimal designs are difficult to obtain, because of optimizational issues. However, the MC-V criterion is useful

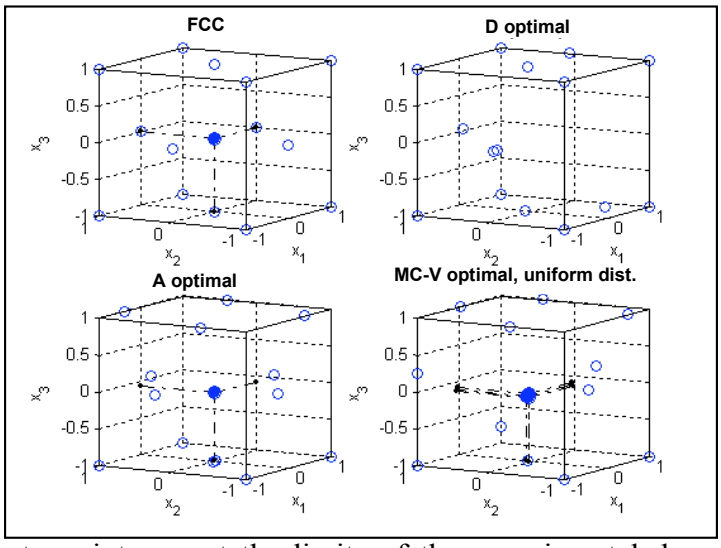
for comparison. For example, in our instance it shows that the A-optimal is preferred to the D-optimal.
In a broader sense, the MC-V criterion provides a useful tool to assess the quality of the designs used to propagate uncertainty in computer experiments.

ACKNOWLEDGMENTS
We thank Xavier Bay, Eric Touboul and André Haas for their useful remarks. We also thank Chris Yukna for his help in editing.

Figures

Figure 1:

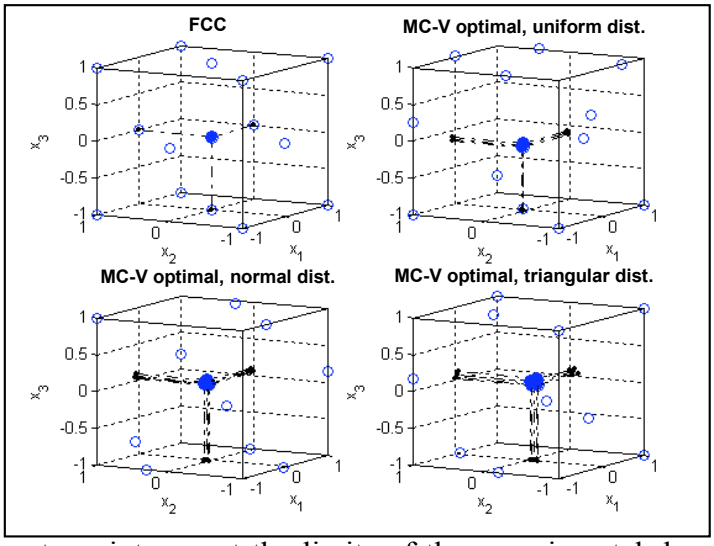
FCC, D-optimal, A-optimal, and MC-V optimal (for uniform distributions) designs with 15 points in dimension 3



On this figure, empty points are at the limits of the experimental domain. Full points are inside and have their projections drawn with a discontinued line. The abbreviation “dist.” means “distribution”.

Figure 2:

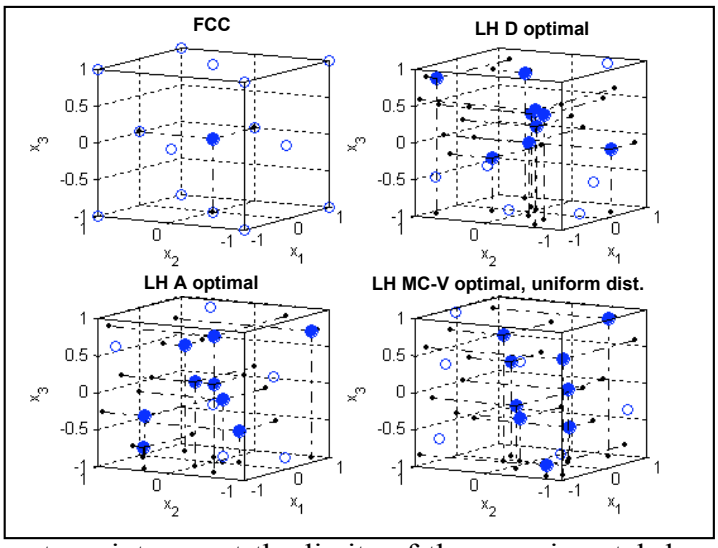
FCC, MC-V optimal for uniform, normal and triangular distributions with 15 points in dimension 3



On this drawing, empty points are at the limits of the experimental domain. Full points are inside and have their projections drawn with a discontinued line. The abbreviation “dist.” means “distribution”.

Figure 3:

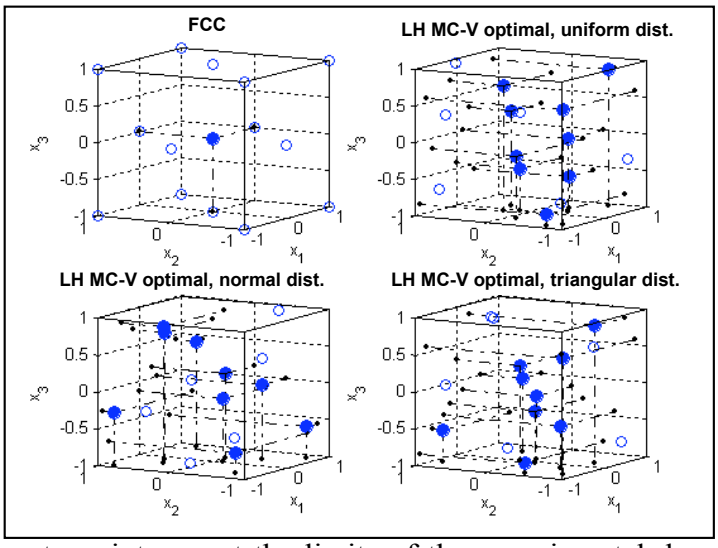
FCC, LH D-optimal, LH A-optimal, LH MC-V optimal (uniform distributions) designs with 15 points in dimension 3



On this drawing, empty points are at the limits of the experimental domain. Full points are inside and have their projections drawn with a discontinued line. The abbreviation “dist.” means “distribution”; “LH” means “Latin Hypercube”.

Figure 4:

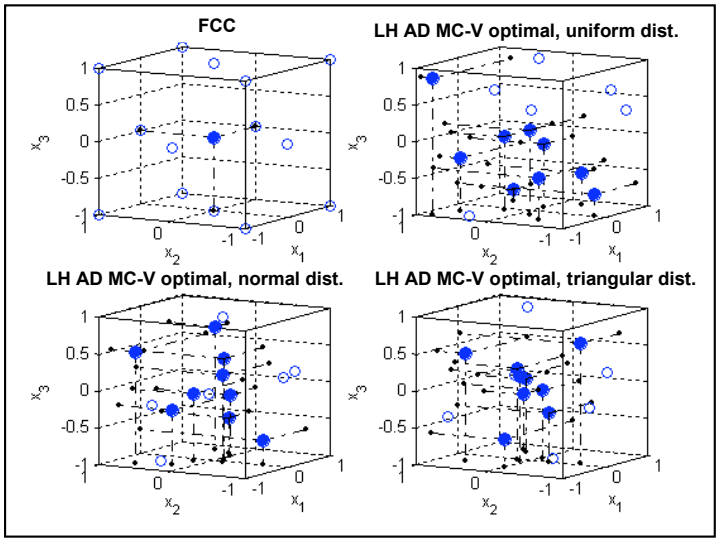
FCC, LH MC-V optimal for uniform, normal and triangular distributions with 15 points in dimension 3



On this drawing, empty points are at the limits of the experimental domain. Full points are inside and have their projections drawn with a discontinued line. The abbreviation “dist.” means “distribution”; “LH” means “Latin Hypercube”.

Figure 5:

Same a figure 4, when the discretization is adapted to the distributions of input factors



On this drawing, empty points are at the limits of the experimental domain. Full points are inside and have their projections drawn with a discontinued line. The abbreviation “dist.” means “distribution”; “LH” means “Latin Hypercube”, and “AD” means “Adapted Discretization”.

Tables

Table 1:

Computation time and criteria values for different computed designs and FCC design, when input factors have uniform distributions

Uniform distribution	Calculation time (en sec)	MC-V criterion	D criterion	A criterion
MC-V optimal design	4333	0.455	2.54E-58	3.302
D optimal design	24968	0.989	1.67E-62	5.192
A optimal design	4431	0.477	6.74E-60	3.204
MC-V optimal LHd (uniform discretization)	117	1.501	6.42E-34	13.110
MC-V optimal LHd (adapted discretization)	405	1.441	2.53E-34	12.510
D optimal LHd	9248	1.250	1.09E-35	10.825
A optimal LHd	2643	1.676	3.47E-33	14.240
Face-centered cubic (81 points)	/	0.436	8.55E-72	4.146

Table 2:

Computation time and criteria values for different computed designs and FCC design, when input factors have normal distributions

Normal distribution	Calculation time (en sec)	MC-V criterion	D criterion	A criterion
MC-V optimal design	5132	0.300	6.62E-56	3.786
D optimal design	24968	1.095	1.67E-62	5.192
A optimal design	4431	0.378	6.74E-60	3.204
MC-V optimal LHd (uniform discretization)	155	0.761	1.37E-33	13.422
MC-V optimal LHd (adapted discretization)	646	1.419	2.62E-18	29.980
D optimal LHd	9248	0.644	1.09E-35	10.825
A optimal LHd	2643	0.831	3.47E-33	14.240
Face-centered cubic (81 points)	/	0.274	8.55E-72	4.146

Table 3:

Computation time and criteria values for different computed designs and FCC design, when input factors have triangular distributions

Triangular distribution	Calculation time (en sec)	MC-V criterion	D criterion	A criterion
MC-V optimal design	4388	0.261	1.73E-54	4.136
D optimal design	24968	1.121	1.67E-62	5.192
A optimal design	4431	0.359	6.74E-60	3.204
MC-V optimal LHd (uniform discretization)	106	0.618	1.09E-33	14.098
MC-V optimal LHd (adapted discretization)	249	1.414	3.54E-13	41.630
D optimal LHd	9248	0.549	1.09E-35	10.825
A optimal LHd	2643	0.692	3.47E-33	14.240
Face-centered cubic (81 points)	/	0.221	8.54E-72	4.146

References

[1] R.A. Fisher, The Design of Experiments, Hafner, New-York, 9th edition, 1971.

[2] J. Sacks, W.J. Welch, T.J. Mitchell, H.P. Wynn, Design and Analysis of Computer Experiments, Statistical Science, 4, No. 4, 1989, p. 409-435.

[3] J. Sacks, S.B. Schiller, W.J. Welch, Designs for Computer Experiments, Technometrics, 31, No. 1, 1989, p. 41-47.

[4] T.J. Santner, B.J. Williams, W.I. Notz, The design and analysis of computer experiments, Springer, 2003.

[5] J.D. Jobson, Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design, Springer-Verlag, 1991.

[6] L. Carraro, B. Corre, C. Helbert, O. Roustant, Construction d'un critère d'optimalité pour plans d'expériences numériques dans le cadre de la quantification d'incertitudes, Revue de Statistique Appliquée, 2005.

[7] A. Borovkov, Statistique mathématique, Editions Mir Moscou, 1987.

[8] G.E.P. Box, N.R. Draper, Empirical model-building and response surfaces, Wiley, 1987.

[9] M.D. Mc Kay, W.J. Conover, R.J. Beckman, A comparison of the three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 21, 1979, p. 239-245.

[10] J.S. Park, Optimal Latin-hypercube designs for computer experiments, Journal of Statistical Planning and Inference, 39, 1994, p. 95-111.