



Design of experiments for response diversity

Régis Bettinger, Pascal Duchêne, Luc Pronzato, Eric Thierry

► To cite this version:

Régis Bettinger, Pascal Duchêne, Luc Pronzato, Eric Thierry. Design of experiments for response diversity. 6th International Conference on Inverse Problems in Engineering: Theory and Practice (ICIPE 2008), Jun 2008, Dourdan, France. hal-00290418

HAL Id: hal-00290418

<https://hal.science/hal-00290418>

Submitted on 25 Jun 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Design of experiments for response diversity

Régis Bettinger^{1,2}, Pascal Duchêne¹, Luc Pronzato² and Eric Thierry²

¹ IFP-Lyon, DTIMA, BP 3, 69390 Vernaison, France

² Laboratoire I3S, CNRS/UNSA, Sophia Antipolis, France

E-mail: regis.bettinger@ifp.fr

Abstract. A design method is proposed that sequentially generates observation sites for the construction of a kriging predictive model. The objective of the construction is to allow a precise inversion of the system in the following sense: with any reachable target point T in the output space one wishes to be able to associate an input vector x_T such that the system response at x_T will be close to T (which requires that the model ensures a precise prediction of the response at x_T). Intuitively, the observation sites should not be spread over the admissible input space, but should rather concentrate in areas such that, when mapped by the system, they cover the reachable output space. Two approaches are proposed that are shown on examples (one with five input and two outputs, derived from a problem in oil industry) to give satisfactory results. They are based on the maximization of a measure of dispersion of the observations in the output space and can cope with the presence of observation errors, a rather typical situation for experimentation with real physical systems.

1. Introduction

We consider an experimental design problem related to the construction of a prediction model for a multi-input/multi-output system, with application to a problem in oil industry. We assume that no parametric (knowledge-based) model is available to describe the input/output relations of the system. The objective of the construction is as follows: with any given reachable (p -dimensional) target vector T in the output space of the system one wishes to associate a (m -dimensional) input vector x_T such that the model prediction $\hat{y}(x_T)$ is as close to T as possible. The experimental design problem consists in selecting N points X_1, \dots, X_N (N fixed) in the m -dimensional input space \mathcal{X} (a compact subspace, or a finite subset, of \mathbb{R}^m) such that this inverse-prediction problem is solved with maximum accuracy after the observation of $Y_i = Y(X_i)$, $i = 1, \dots, N$.

We model the input/output relationship by kriging [1, 2], a modelling tool based on stochastic processes, which, at low computational cost, gives an estimate of the (mean squared) prediction error at any unsampled point, thereby indicating the accuracy of the model prediction in each point of the input domain, see Section 2. In practice, we use the Matlab toolbox DACE [3]. Adopting a bayesian point of view, after having collected observations at X_1, \dots, X_N we obtain the posterior distribution of $Y(x)$ conditional to $Y^N = (Y_1, \dots, Y_N)^\top$, for any $x \in \mathcal{X}$. This can be used to compute the inverse prediction for the target T ,

$$x_T = \arg \min_{x \in \mathcal{X}} \mathbb{E}\{\|Y(x) - T\|^2 | Y^N\} = \arg \min_{x \in \mathcal{X}} [\|\hat{y}_N(x) - T\|^2 + \hat{\sigma}_N^2(x)] , \quad (1)$$

with $\hat{y}_N(x) = \mathbb{E}\{Y(x) | Y^N\}$ and $\hat{\sigma}_N^2(x) = \mathbb{E}\{\|Y(x) - \hat{y}_N(x)\|^2 | Y^N\}$ respectively the model prediction and variance at x . Define

$$\rho_N(T) = \|\hat{y}_N(x_T) - T\|^2 + \hat{\sigma}_N^2(x_T) \quad (2)$$

and

$$J_{M,N}(X_1, \dots, X_N) = \max_{T \in \mathcal{Y}} \rho_N(T), \quad J_{I,N}(X_1, \dots, X_N) = \int_{\mathcal{Y}} \rho_N(T) \mu(dT) \quad (3)$$

(which depend on the observations Y^N), with \mathcal{Y} the reachable output domain and $\mu(\cdot)$ some given probability measure on \mathcal{Y} . Using a direct formulation of the inverse-prediction design problem, one might then consider the minimization of $EJ_M(X_1, \dots, X_N) = \mathbb{E}\{J_{M,N}(X_1, \dots, X_N)\}$ or $EJ_I(X_1, \dots, X_N) = \mathbb{E}\{J_{I,N}(X_1, \dots, X_N)\}$ with respect to X_1, \dots, X_N . However, this is a formidable task and we shall follow another route (the functions $J_{M,N}(X_1, \dots, X_N)$ and $J_{I,N}(X_1, \dots, X_N)$ will nevertheless be used to evaluate *a posteriori* the quality of a given design in terms of inverse prediction performance).

In order to avoid the explicit construction of x_T for every T , we reformulate the problem and simply try to obtain good model predictions over \mathcal{Y} by designing experiments that sample \mathcal{Y} as uniformly as possible, in the sense that the N observed responses $Y_i = Y(X_i)$, $i = 1, \dots, N$, are as spread as possible in \mathbb{R}^p . One can note that this enforces the exploration of the whole domain \mathcal{Y} . At the same time, one wishes that the X_i 's remain as concentrated as possible in \mathcal{X} in order to have a precise prediction model via kriging.

The design method is sequential, in the sense that the observation point X_{n+1} is chosen after Y_1, \dots, Y_n have been observed. This implies the use of an initial design (with a small number of points compared to N). We suppose that no prior information on the system is available and use a space-filling initial design. Two methods are then used (and compared) for the sequential addition of design points. Both look one-step ahead.

The first method chooses X_{n+1} that maximizes $\mathbb{E}\{\min_{i=1, \dots, n} \|Y(x) - Y_i\|^2 | Y_1, \dots, Y_n\}$ (Section 3), the conditional (posterior) expectation of the minimum squared distance between $Y(x)$ and the Y_i 's already observed. The second method (Section 4) chooses X_{n+1} that maximizes the conditional expectation of the second order Tsallis entropy [4] of a kernel density estimator formed from the Y_i 's and $Y(x)$ (following the idea that maximum entropy over a compact set is obtained for the uniform distribution, the second-order Tsallis entropy being used for it allows simple analytical calculations). The precision of the model prediction is taken into account by both approaches (through the conditional distribution of $Y(x)$), which favors the choice of X_{n+1} close to points X_i already sampled.

It may happen that for practical reasons the use of the experimental device requires that batches of k points X_{i+1}, \dots, X_{i+k} are chosen simultaneously. Moreover, for the application considered in Section 6.2 those k points must be chosen before the k previous observations are available. This k -points-at-a-time design problem is considered in Section 5.

In Section 6 the two approaches of Sections 3 and 4 are compared (in particular in terms of $J_{M,N}$ and $J_{I,N}$, see (3)) through simulations, first for a 2-inputs/1-output system (which allows easy graphical presentations of the results), second for a 5-inputs/2-outputs system (constructed from real data obtained from a problem in oil industry). The corruption of observations by errors is considered, to account for the presence of random perturbations in real (physical) experiments.

2. Prediction by kriging

The presentation is for one output only. When there are p outputs, as in the example of Section 6.2, we use p distinct models (although co-kriging could then be used, see, e.g., [5]). The system output (response) at x is modelled by

$$Y(x) = f^\top(x)\beta + Z(x) + \varepsilon(x), \quad (4)$$

where $f(x) = (f_1(x), \dots, f_k(x))^T$ is a vector of known regression functions, $\beta = (\beta_1, \dots, \beta_n)^T$ is a vector of unknown parameters, $Z(\cdot)$ is a stochastic process, stationary on \mathcal{X} , the admissible domain for the inputs x , and $\varepsilon(x)$ denotes an observation error. Those errors are assumed to be centered and independently distributed (also independently from $Z(x)$), with variance σ_ε^2 . Throughout the paper we assume that σ_ε^2 is known. The process $Z(x)$ is supposed to have zero mean, variance σ_z^2 and covariance $\text{Cov}\{Z(X_1), Z(X_2)\} = \sigma_z^2 \Omega(X_1 - X_2, \psi)$ with Ω the correlation function and ψ a vector of unknown parameters.

Suppose that n observations have been performed for the inputs (design points) X_1, \dots, X_n . The associated outputs are denoted Y_1, Y_2, \dots, Y_n and satisfy $Y_i = f(X_i)^T \beta + Z(X_i) + \varepsilon_i$. We also denote $Y^n = (Y_1 \dots Y_n)^T$, $Z^n = (Z(X_1) \dots Z(X_n))^T$, $F = (f(X_1) \dots f(X_n))^T$, $\varepsilon = (\varepsilon_1 \dots \varepsilon_n)^T$ and R the correlation matrix of Z^n . Take an arbitrary point x_0 and denote $f_0 = f(x_0)$, $Y_0 = f_0^T \beta + Z_0 + \varepsilon_0$ the observation at x_0 and r_0 the correlation between $Z(x_0)$ and Z^n (thus a n -dimensional vector). The kriging predictor at x_0 is linear, unbiased and has minimum variance, i.e., $\hat{y}(x_0) = a_0 + a^T Y^n$, where a_0 and a can be determined by minimizing the predictor variance under the zero-bias constraint. When the characteristics of the process are known (i.e., when Ω , σ_z^2 and ψ are known), direct calculations give the solution

$$\hat{y}_n(x_0) = f_0^T \hat{\beta} + r_0^T \mathcal{R}^{-1} (Y^n - F \hat{\beta}), \quad (5)$$

where

$$\hat{\beta} = (F^T \mathcal{R}^{-1} F)^{-1} F^T \mathcal{R}^{-1} Y^n, \quad (6)$$

with $\mathcal{R} = R + (\sigma_\varepsilon^2/\sigma_z^2) I_n$ and I_n the n -dimensional identity matrix. The variance of this prediction is given by

$$\hat{\sigma}_n^2(x_0) = \sigma_z^2 \left(a^T \mathcal{R} a - 2a^T r_0 + 1 \right) + \sigma_\varepsilon^2, \quad (7)$$

where a is obtained from $\hat{y}_n(x_0) = a^T Y^n$.

Supposing that the process $Z(\cdot)$ and the observation errors are jointly normal, one can easily construct the likelihood function $l(\beta, \sigma_z^2, \psi)$ for the unknown parameters β , σ_z^2 and ψ , which can then be estimated by maximum likelihood, see, e.g., [6]. Easy calculations give

$$l(\beta, \sigma_z^2, \psi) = -\frac{1}{2} \left[n \log(\sigma_z^2) + \log(\det \mathcal{R}) + \frac{(Y^n - F \beta)^T \mathcal{R}^{-1} (Y^n - F \beta)}{\sigma_z^2} \right]. \quad (8)$$

Note that the maximum likelihood estimator for β coincides with (6) when σ_z^2 and ψ are replaced by their optimal (maximum-likelihood) values. When $\sigma_\varepsilon^2 = 0$ (no measurement errors), we get $\sigma_z^2 = (1/n)(Y^n - F \hat{\beta})^T R^{-1} (Y^n - F \hat{\beta})$ and its substitution in (8) gives an expression that only depends on ψ . Also, one can easily check that $\hat{y}_n(X_i) = Y_i$ and $\hat{\sigma}_n^2(X_i) = 0$, $i = 1, \dots, n$, when $\sigma_\varepsilon^2 = 0$; that is, the predictor is a perfect interpolator with zero prediction variance at the design points when there are no observation errors.

The design points X_i will be determined sequentially, which requires using an initial design \mathcal{X}_0 that defines where the first observations are to be taken. When no prior information on the system to be modelled is available, this initial design should have a space-filling property, i.e., the initial points should be as spread as possible in the admissible domain \mathcal{X} . Two situations will be considered in Section 6. When the input domain is 2-dimensional (Section 6.1), we use a latin hypercube design with 9 points, which is easy to generate, combined with the optimization of a maximin-distance criterion in order to ensure that the points are well enough spread in \mathcal{X} , as suggested in [7]. In the case of a 5-dimensional input domain (Section 6.2), a complete factorial design with 32 points will be preferred. Indeed, we noticed that the presence of observations at the edges of \mathcal{X} gives better kriging predictions. The factorial design is then completed by 16 points added inside the domain, such that all canonical projections along 1, 2, 3 and 4 dimensions are identical.

A rather standard approach for the sequential construction of design points ensuring precise kriging predictions over the whole domain \mathcal{X} consists in locating the $(n + 1)$ -th observation at the point X_{n+1} given by

$$X_{n+1} = \arg \max_{x \in \mathcal{X}} \hat{\sigma}_n^2(x) \quad (9)$$

where $\hat{\sigma}_n^2(x)$ is given by (7). When there are no observation errors, this approach thus prevents the repetition of observations at the same location (since $\hat{\sigma}_n^2(x) = 0$ when an observation has already been made at x). This property generalizes to approaches that take the prediction variance into account, but may not be satisfied when $\sigma_\varepsilon^2 \neq 0$. In that case, one may use the re-interpolation technique suggested in [8]. A first kriging interpolator is constructed, taking the presence of measurement errors into account. Therefore, the predictor does not interpolate the data, i.e., the predictions $\hat{y}_n(X_i)$ at the design points X_i differ from the associated observations Y_i . Second, a new kriging predictor is constructed, using the predictions $\hat{y}_n(X_i)$ as if they were observations, assuming there are no observation errors. One can show that for any x the prediction at this second stage coincides with $\hat{y}_n(x)$ of the first stage, with the noticeable difference that the prediction variance for the second predictor is now zero at the observation points X_i . Although this re-interpolation technique is proposed in [8] for the case of computer experiments, we shall use it in Section 6.2 for the case when the observations are corrupted by measurement errors.

3. Sequential maximin distance

Consider the following quantity, $d(y, Y^n) = \min_{i=1, \dots, n} \|y - Y_i\|^2$, which measures how far y is from the observations already performed. Using a bayesian approach with a non-informative prior, one can consider that the posterior distribution of $Y(x)$ at some unsampled x given the observations Y^n is approximately normal, $Y(x) \sim \mathcal{N}(\hat{y}_n(x), \hat{\sigma}_n^2(x))$, where $\hat{y}_n(x)$ and $\hat{\sigma}_n^2(x)$ are respectively given by (5) and (7). We then choose next design point X_{n+1} as the one that maximizes

$$J_{1,n}(X) = \mathbb{E}\{\bar{d}(Y(X), Y^n) | Y^n\} \quad (10)$$

where $\bar{d}(Y(X), Y^n)$ is a truncated version of $d(Y(X), Y^n)$, $\bar{d}(Y(X), Y^n) = d(Y(X), Y^n) \chi_\lambda(Y^n)$ with $\chi_\lambda(Y^n)$ the indicator function of the set $[\min(Y^n) - \lambda, \max(Y^n) + \lambda]$ (this truncation permits to avoid to give excessive weight to $Y(X)$ falling outside the reachable output domain). The computation of $J_{1,n}(X)$ requires the evaluation of an integral in \mathbb{R}^p . An analytic expression is easily obtained when $p = 1$; for larger p the integrant is computed with $\lambda = 0$ over a grid (100×100 for $p = 2$ in the example of Section 6.2).

4. Sequential maximization of Tsallis entropy

This second approach follows an idea similar to that in the method above, but measures the dispersion of the observations in the output space through entropy. Since maximum entropy over a compact set is obtained for the uniform distribution, we shall maximize the entropy of the distribution of the observations in \mathcal{Y} . Entropy estimation is a vast domain and many approaches have been suggested, see, e.g., [9, 10], among which nearest-neighbor methods are promising, especially when the dimension p is large, see [11]. Here we follow a rather standard approach, where a kernel-density estimator for the distribution of the Y_i 's is plugged into the expression of the entropy. At step n , when the observations Y^n are available, consider the kernel density estimator based on Y^n and $Y(X)$

$$\hat{\phi}_n(y; Y(X)) = \frac{1}{n+1} \left[\sum_{i=1}^n \varphi_{Y_i, \sigma}(y) + \varphi_{Y(X), \sigma}(y) \right] \quad (11)$$

with $\varphi_{z,\sigma}(\cdot)$ the probability density function of the normal $\mathcal{N}(z, \sigma^2)$. Usually, the smoothing parameter σ (window width) is taken decreasing with n (at an appropriate rate to ensure suitable convergence properties for the estimator). However, the number of observations is often severely restricted for the type of applications considered here, and σ will be kept constant.

A natural candidate is Shannon entropy, which, for $\phi(\cdot)$ a probability density function (pdf) on (a subset of) \mathbb{R}^p , writes $H_1(\phi) = -\int_{\mathbb{R}^p} \phi(t) \log[\phi(t)] dt$. However, the computation of $H_1[\hat{\phi}_n(\cdot; Y(X))]$ with $\hat{\phi}_n(\cdot; Y(X))$ given by (11) is rather cumbersome and we thus consider another form of entropy that yields far simpler calculations. The second-order Tsallis entropy of the pdf $\phi(\cdot)$ writes $H_2(\phi) = 1 - \int_{\mathbb{R}^p} \phi^2(t) dt$ and can be given an analytic expression when substituting the estimator $\hat{\phi}_n(\cdot; Y(X))$ for ϕ . Indeed, for $\phi(y; z_1, \dots, z_m) = (1/m) \sum_{i=1}^m \varphi_{z_i, \sigma}(y)$ we have

$$H_2[\phi(\cdot; z_1, \dots, z_m)] = 1 - \int_{\mathbb{R}^p} \phi^2(t; z_1, \dots, z_m) dt = 1 - (1/m^2) \sum_{i,j=1}^m \varphi_{z_i, \sigma\sqrt{2}}(z_j).$$

Similarly, the criterion

$$J_{2,n}(X) = \mathbb{E}\{H_2[\hat{\phi}_n(\cdot; Y(X))] | Y^n\}, \quad (12)$$

where $Y(x)$ has the normal posterior $Y(X) \sim \mathcal{N}(\hat{y}_n(X), \hat{\sigma}_n^2(X))$, can be given an analytic expression using the property $\int_{\mathbb{R}^p} \varphi_{a,\sigma}(t) \varphi_{b,\delta}(t) dt = \varphi_{a, \sqrt{\sigma^2 + \delta^2}}(b)$. Next point X_{n+1} is then obtained by maximizing $J_{2,n}(X)$.

5. Batch sequential implementation

It may happen that practical considerations impose that batches of k points are chosen simultaneously, i.e., X_{n+1}, \dots, X_{n+k} at step n ($k = 6$ in the example of Section 6.2). Both criteria $J_{i,n}(X_{n+1}, \dots, X_{n+k})$, $i = 1, 2$, then depend on $Y_{n+1}, \dots, Y_{n+k-1}$ which are unknown at step n . This makes the problem very complicated and we use the following drastic simplification: the k next design points at step n are simply chosen as local maximizers of $J_{i,n}(X)$ (including the global one).

Moreover, it also happens that observations do not correspond to direct measurements but require some treatment of the collected data, which induces a delay k' , meaning that at step n the observations $Y_n, \dots, Y_{n-k'+1}$ are not available yet ($k' = k = 6$ in the example of Section 6.2). Here we use a drastic simplification again, and simply replace the unknown observations by their predicted values based on $Y^{n-k'}$.

One may notice that when using the criterion $J_{2,n}(\cdot)$ for choosing the location of next design points, the problems caused by batch treatment and delays can receive an exact analytic solution. Indeed, consider for instance the case of batch treatment. When $Y_{n+1}, \dots, Y_{n+k-1}$ are unknown, one may consider first $\mathbb{E}\{J_{2,n}(X_{n+1}, \dots, X_{n+k}) | Y^{n+k-2}\}$. The expectation with respect to Y_{n+k-1} can be calculated analytically, which gives an expression that depends on the prediction $\hat{y}_{n+k-2}(X_{n+k-1})$. Using the linearity of this prediction with respect to Y^{n+k-2} , similar calculations can be back-propagated and yield a criterion depending only on X_{n+1}, \dots, X_{n+k} and Y^n . Similar developments can be made to calculate expectations with respect to $Y_n, \dots, Y_{n-k'+1}$ when a delay k' is present. However, we do not follow this approach here (those developments will be presented elsewhere), and use the simplifications mentioned above for the examples in the next section.

6. Examples

We first consider a 2-inputs/1-output toy example that allows a graphical illustration of the behavior of the two methods of Sections 3, 4. We then apply those methods to a more complex

5-inputs/2-outputs model that represents the behavior of a real chemical system. For both examples, we take $f(x) = 1$ in (4) and $\Omega(t, \psi) = e^{-\psi\|t\|^2}$.

6.1. A toy example with two inputs and one output

The system is supposed to follow the equation (no observation error, $\sigma_\varepsilon = 0$)

$$Y(x_1, x_2) = 0.2e^{x_1-3} + 2.2|x_2| + 1.3x_2^6 - 2x_2^2 - 0.5x_2^4 - 0.5x_2^4 + 2.5x_1^2 + 0.7x_1^3 + \frac{3}{(8x_1 - 2)^2 + (5x_2 - 3)^2 + 1} + \sin(5x_1) \cos(3x_1^2).$$

The initial design is a 9-points maximin latin hypercube (several such initial designs have been used, showing little influence on the performance), points are then added sequentially up to $N = 20$. The designs obtained when using the criteria $J_{1,n}(\cdot)$ and $J_{2,n}(\cdot)$, see (10, 12), are presented in Figure 1–left. The dots correspond to the initial design. The associated responses are presented on Figure 1–right. The diamonds (resp. stars) correspond to the points added when using $J_{1,n}(\cdot)$ (resp. $J_{2,n}(\cdot)$). Note that the design points generated by both methods tend to concentrate in \mathcal{X} whereas the associated responses are dispersed in \mathcal{Y} .

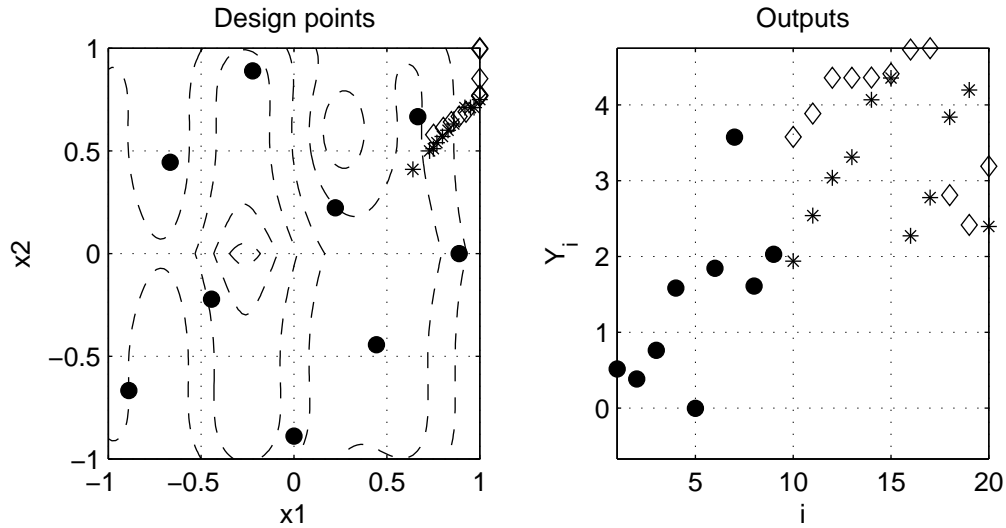


Figure 1. Comparison of the designs (left) generated by $J_{1,n}(\cdot)$ (diamonds) and $J_{2,n}(\cdot)$ (stars) and associated responses (right), the initial design is a maximin Latine hypercube with 9 points (dots); the dashed lines (left) correspond to contour plots of $Y(x)$.

We illustrate now the performance of the methods in terms of x_T and $\rho_N(T)$, see (1, 2). Figure 2–left presents the location of x_T for T varying within its reachable bounds on a regular grid of size 50 and Figure 2–right presents $\rho_N(T)$ as a function of T . The symbols are the same as in Figure 1: diamonds for the design generated with $J_{1,n}(\cdot)$, stars for $J_{2,n}(\cdot)$. The left part shows how the system can be inverted, the right part indicates the (mean-squared) error that can be expected. Note that the points x_T tend to gather around design points (compare with Figure 1–left) in order to reduce the prediction error.

6.2. An application in oil industry with a five inputs/two outputs system

We consider now a more complex system, with measurement errors, 5-inputs and 2-outputs, constructed empirically from real experimental data obtained from a chemical system. The

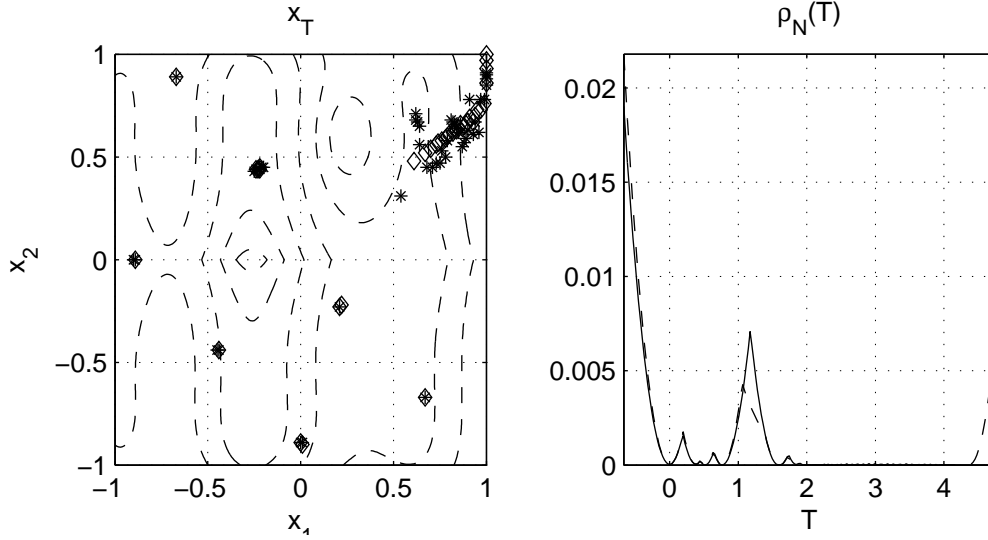


Figure 2. Location of x_T (left) and $\rho_N(T)$ (right) for the designs generated by $J_{1,n}(\cdot)$ (diamonds and solid line) and $J_{2,n}(\cdot)$ (stars and dashed line).

input domain is reduced to a grid \mathcal{G}_X that has been selected with the help of chemists and contains $K = 13 \times 17 \times 13 \times 6 \times 5 = 86190$ points. The initial design contains 48 points, see Section 2, and points are added sequentially up to $N = 500$. The standard deviations $\sigma_{\varepsilon,i}$, $i = 1, 2$, of the measurement errors for both responses are supposed to be known and constant, equal to 5% of the range for the corresponding response. We consider batch sequential design, with batches of size $k = 6$, in the presence of a delay $k' = 6$, see Section 5.

The noise-free responses (available since this is a simulated example) associated with the design points generated when using the criteria $J_{1,n}(\cdot)$ and $J_{2,n}(\cdot)$, see (10, 12), are presented in Figure 3. The dots correspond to the responses for the initial design. Note that both methods yield responses reasonably spread in \mathcal{Y} .

Finally, we measure the performance of the methods by computing the values of the criteria $J_{M,N} = \max_{T \in \mathcal{G}_Y} \rho_N(T)$ and $J_{I,N} = (1/K) \sum_{i=1}^K \rho_N(T_i)$, see (3), with $T_i \in \mathcal{G}_Y$, the image of the grid \mathcal{G}_X by the noise-free system.

For the sequential maximin-distance approach of Section 3 (criterion $J_{1,n}(\cdot)$) we get $J_{M,N} = 0.0383$ and $J_{I,N} = 0.0071$, whereas for the approach based on Tsallis entropy (criterion $J_{2,n}(\cdot)$) of Section 4) we get $J_{M,N} = 0.0462$ and $J_{I,N} = 0.0074$. It should be noticed that the second method (based on Tsallis entropy) is much faster than the first one. For that reason the first method is partly implemented in Fortran. The two methods then generate the last batch of 6 points (the computational burden increases with n) in approximately 10 minutes.

When the design points are generated according to the rule (9), thereby enforcing a space-filling property in the whole domain \mathcal{X} , we get $J_{M,N} = 0.0546$ and $J_{I,N} = 0.0091$, which shows the interest of the two approaches proposed in the paper for the inverse prediction problem under consideration.

7. Conclusion

This paper deals with the issue of designing experiments for the construction of a model with good ‘inverse-prediction’ features, that is, a model which associates with each output value T an input value x_T that predicts T well. Two methods are proposed, both of them taking the prediction accuracy into account *via* a kriging model. They have been compared first on a 2-

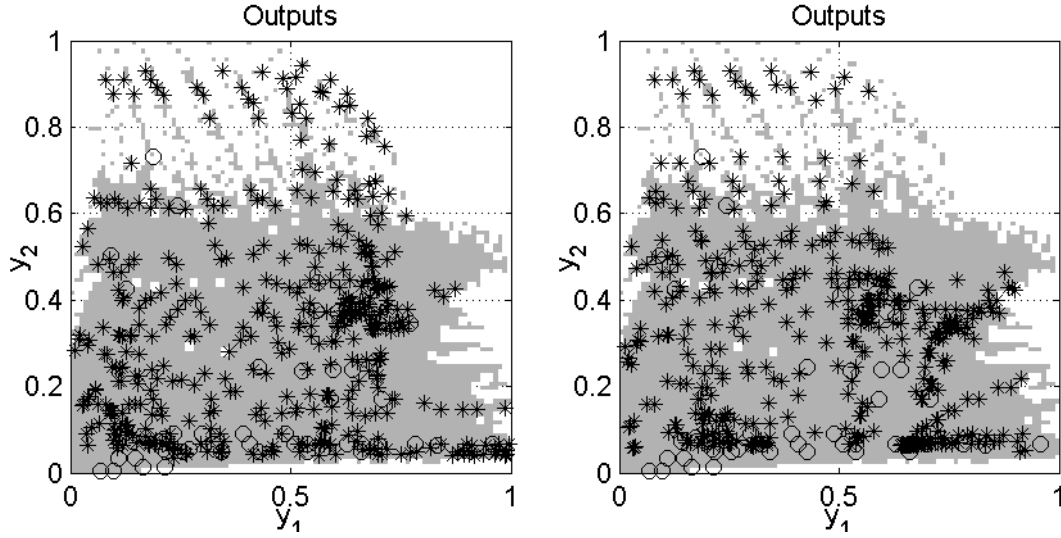


Figure 3. Dispersion of the responses associated with the designs generated by $J_{1,n}(\cdot)$ (left) and $J_{2,n}(\cdot)$ (right), responses associated with the initial design points are indicated by dots; the grey background corresponds to the image of the input grid \mathcal{G}_X by the true (simulated noise-free) system.

inputs/1-output noise-free toy example and then on a 5-inputs/2-outputs noisy system derived from a physical application. Both methods seem to give satisfactory results, the one using Tsallis entropy being more advisable for high output dimensions due to the analytic form of the design criterion. Finally, when compared with the space-filling algorithm that places design points at the maxima of the kriging variance, both approaches perform better for the inverse-prediction problem under consideration.

Acknowledgments

We would like to thank the two referees for their helpful comments.

References

- [1] Sacks J, Welch W, Mitchell T and Wynn H 1989 *Statistical Science* **4** 409–435
- [2] Santner T J, Williams B J and Notz W 2003 *The Design and Analysis of Computer Experiments* (Heidelberg: Springer)
- [3] Lophaven S, Nielsen H B and Sondergaard J 2002 Dace, a Matlab kriging toolbox Tech. Rep. IMM-REP-2002-12
- [4] Tsallis C 1988 *Journal of Statistical Physics* **52** 479–487
- [5] Cressie N 1991 *Statistics for spatial data* (New York: Wiley)
- [6] Stein M 1999 *Interpolation of Spatial Data. Some Theory for Kriging* (Heidelberg: Springer)
- [7] Morris M D and Mitchell T J 1995 *J. Statist. Plann. Inference* **43** 381–402
- [8] Forrester A I J, Keane A J and Bressloff N W 2006 *AIAA Journal* **44**(10) 2331–9
- [9] Beirlant J, Dudewicz E, Györfi L and van der Meulen E 1997 *International Journal of Mathematical and Statistical Sciences* **6** 17–39
- [10] Hall P and Morton S 1993 *Ann. Inst. Statist. Math.* **45** 69–88
- [11] Leonenko N, Pronzato L and Savani V *The Annals of Statistics* To appear