

# MODEL-ROBUST SEQUENTIAL DESIGN OF EXPERIMENTS FOR IDENTIFICATION PROBLEMS

*Hassan El Abiad, Laurent Le Brusquet, Morgan Roger, Marie-Ève Davoust*

Department of Signal Processing and Electronic Systems, Supélec, Gif-sur-Yvette, France  
E-mails : Hassan.Elbiad, Laurent.LeBrusquet, Morgan.Roger, Marie-Eve.Davoust@supelec.fr

## ABSTRACT

A new criterion for sequential design of experiments for linear regression model is developed. Considering the information provided by previous collected data is a well-known strategy to decide for the next design point in the case of nonlinear models. The paper applies this strategy for linear models. Besides, the problem is addressed in the context of robustness requirement: an unknown deviation from the linear regression model (called model error or misspecification) is supposed to exist and is modeled by a kernel-based representation (Gaussian process). The new approach is applied on a polynomial regression example and the obtained designs are compared with other designs obtained from other approaches that do not consider the information provided by previously collected data.

**Index Terms**— Sequential design of experiments, Gaussian process, linear regression, robust design.

## 1. INTRODUCTION

Consider the following linear regression model:

$$t(x) = \sum_{j=1}^d \theta_j \phi_j(x) = \phi^\top(x) \theta \quad (1)$$

where  $\theta$  is the regression parameters vector and  $\phi$  is the regression vector. The design of experiments (DOE) problem consists in choosing the design points  $x_i$  in order to best estimate  $\theta$ .

Suppose that: i) a set of  $n$  collected data  $\{(x_i, y_i) \in \mathcal{X} \times \mathbb{R}, i = 1, \dots, n\}$  is given, where  $\mathcal{X}$  is the experimental domain. The  $x_i$ 's form the design denoted by  $\xi_n = [x_1, \dots, x_n]^\top$ . The  $y_i$ 's are noisy observations of the targets  $t_i = t(x_i)$ :

$$y_i = t_i + e_i \quad (2)$$

where the observations errors  $e_i$  are normal and i.i.d., and ii) the least squares estimation has been chosen to estimate the regression parameters.

Then, the problem of sequential DOE is to choose the next experiment  $x_{n+1}$  to carry out in order to refine the parameters estimation.

Moreover, if the target function  $t(x)$  is not perfectly represented by the regression model in (1) (i.e. a model error or misspecification exists), then standard DOE methods that do not take into consideration the misspecification will introduce a bias in the parameters estimation [1]. An important work that solves such drawbacks was done by Yue and Hickernell [1].

In this paper, a new sequential DOE criterion for linear models with misspecification is derived. The new criterion takes into consideration the information provided by previously collected data. Based on previous works that deal with robustness ([1] and [2]), the misspecification is modeled by a kernel-based representation (here a Gaussian process as in [2]).

The paper is organized as follows. Next section presents some previous works on model-robust DOE, and discusses the misspecification model selection. Afterwards, the proposed sequential DOE criterion is derived. Finally, the new proposed criterion is applied on a polynomial regression example showing the benefits of this approach over other approaches that do not take into consideration the information provided by previously collected data.

## 2. MODEL-ROBUST DOE

Design robustness with respect to the model error was first discussed by [3] who studied the effect of taking a one-degree polynomial regression model when the target is two-degree. After that, many authors ([4], [5] for instance) have further discussed and developed this idea with different assumptions about the misspecification. Two important works on model-robust DOE ([1] and [2]) consider the following linear model with misspecification:

$$t(x) = \sum_{j=1}^d \theta_j \phi_j(x) + r(x) \quad (3)$$

The unknown model error (misspecification)  $r(x)$  is modeled by a Gaussian process as in [2] (see [6] for a Gaussian process tutorial). A Gaussian process is a random field defined by its

mean and covariance function:

$$\begin{aligned} E\{r(x)\} &= 0, & \forall x \in \mathcal{X} \\ E\{r(x)r(x')\} &= c(x, x'), & \forall (x, x') \in \mathcal{X}^2 \end{aligned}$$

The relevance of modeling the misspecification as a Gaussian process rises because for some classes of covariance functions, Gaussian processes span a rather large space (infinite-dimensional). Therefore, this type of representation matches the robustness requirement: the design point  $x_{n+1}$  that we look for must lead to good estimation performance over a wide range of possible misspecifications.

Let  $\mathbf{y}_n$  be the  $n \times 1$  vector of known observations  $y_i$ ,  $\mathbf{r}_n$  the  $n \times 1$  vector of the misspecifications  $r(x_i)$ , and  $\mathbf{e}_n$  the  $n \times 1$  vector of the observations error  $e(x_i)$ . Then,

$$\mathbf{y}_n = \Phi_n \boldsymbol{\theta} + \mathbf{r}_n + \mathbf{e}_n = \Phi_n \boldsymbol{\theta} + \mathbf{z}_n \quad (4)$$

where,  $\Phi_n$  is the  $n \times d$  model matrix whose  $i$ -th line is  $\phi^\top(x_i)$ , and  $\mathbf{z}_n$  is the *observations-model* errors vector generated by a Gaussian process  $z(x)$  with mean 0 and covariance  $c(x, x') + \sigma_e^2 \delta(x - x')$ .

Let  $\hat{t}$  be the estimator of  $t$  obtained by a least squares estimation of  $\boldsymbol{\theta}$ :

$$\hat{t}(x) = \boldsymbol{\psi}^\top(x) \mathbf{y}_{n+1} \quad (5)$$

where

$$\boldsymbol{\psi}(x) = \Phi_{n+1} (\Phi_{n+1}^\top \Phi_{n+1})^{-1} \phi(x)$$

is a  $(n+1) \times 1$  vector.  $\Phi_{n+1}$  is the  $(n+1) \times d$  extended model matrix, and  $\mathbf{y}_{n+1}$  is the  $(n+1) \times 1$  extended observations vector.

Given  $\hat{t}$  in (5), the goal is to find the design point  $x_{n+1}$  which maximizes the estimator performance (e.g. minimizes the prediction error over all the experimental domain and over a wide range of possible misspecifications). Therefore, the Integral Quadratic Error (IQE) is defined as (classical  $L_2$ -norm):

$$\text{IQE}(x_1, \dots, x_{n+1}, e_1, \dots, e_{n+1}, r) = \int_{\mathcal{X}} |t - \hat{t}|^2 dx \quad (6)$$

The observation errors are unknown, so an expectation over these errors is taken to ensure a good performance averagely over their realizations. Similarly, the misspecification is unknown but our idea of robustness states that the chosen design point must guarantee a good level of performance (on average) over the set of potential misspecifications. A statistical representation for  $r$  allows to take expectation, hence the idea of modeling it by a Gaussian process. Taking the total expectation of the IQE in (6), the Integral Quadratic Risk (IQR) can be written as follows:

$$\text{IQR}(x_{n+1}) = E_{(e_1, \dots, e_{n+1}, r)} \left[ \int_{\mathcal{X}} |t - \hat{t}|^2 dx \right] \quad (7)$$

The optimal design point  $x_{n+1}^*$  is the minimizer of (7):

$$x_{n+1}^* = \arg \min_{x_{n+1} \in \mathcal{X}} [\text{IQR}(x_{n+1})] \quad (8)$$

Using equation (5), the IQR is rewritten as follows:

$$\text{IQR} = E_{(e_1, \dots, e_{n+1}, r)} \int_{\mathcal{X}} |r(x) - \boldsymbol{\psi}^\top(x) \mathbf{z}_{n+1}|^2 dx \quad (9)$$

where  $\mathbf{z}_{n+1} = [\mathbf{z}_n; z_{n+1}] = [\mathbf{r}_n; r(x_{n+1})] + [\mathbf{e}_n; e_{n+1}]$  is the extended *observations-model* error vector (";" is a vertical concatenation). Expanding the previous equation, one finds:

$$\begin{aligned} \text{IQR} &= \int_{\mathcal{X}} \left\{ E_{(e_1, \dots, e_{n+1}, r)} [r^2(x)] \right\} dx \\ &+ \int_{\mathcal{X}} \left\{ E_{(e_1, \dots, e_{n+1}, r)} [\boldsymbol{\psi}^\top(x) \mathbf{z}_{n+1} \mathbf{z}_{n+1}^\top \boldsymbol{\psi}(x)] \right\} dx \\ &- 2 \int_{\mathcal{X}} \left\{ E_{(e_1, \dots, e_{n+1}, r)} [r(x) \boldsymbol{\psi}^\top(x) \mathbf{z}_{n+1}] \right\} dx \end{aligned} \quad (10)$$

As all the variables are Gaussian, the expectation in equation (10) can be computed [2] and optimized according to the criterion in equation (8).

### 3. CONSTRUCTION OF THE PROPOSED DOE CRITERION

This part is the core of the presented work. Having a set of collected data  $(x_i, y_i)$  provides information about the random variables  $(e_i, r(x))$ . Thus, the criterion to be optimized in equation (7) has been modified:

$$\text{IQR}(x_{n+1}) = E_{(e_1, \dots, e_{n+1}, r)/CD} \left[ \int_{\mathcal{X}} |t - \hat{t}|^2 dx \right] \quad (11)$$

where  $/CD$  means that all the probability density functions are calculated conditionally to the already collected data.

The expectation calculation in equation (11) becomes a bit more complicated by taking into consideration previous information. It is detailed in this section.

By writing the relation in (4) satisfied by the collected data, a set of  $k = n - d$  independent linear constraints on the random variables  $\mathbf{z}_n$  appears. The constraints are computed as follows:

The model matrix  $\Phi_n$  in equation (4) is not necessarily square because in DOE the number of  $x_i$ 's is usually greater than the number of parameters  $d$ . Then, the model matrix is divided into two sub-matrices  $\Phi_B$  and  $\Phi_{\bar{B}}$ .  $\Phi_B$  is a  $d \times d$  reversible matrix and  $\Phi_{\bar{B}}$  is a  $k \times d$  matrix. Also,  $\mathbf{y}_n$  and  $\mathbf{z}_n$  are divided into  $\mathbf{y}_n = [\mathbf{y}_B; \mathbf{y}_{\bar{B}}]$  and  $\mathbf{z}_n = [\mathbf{z}_B; \mathbf{z}_{\bar{B}}]$  respectively. Therefore, equation (4) is spread in two equations:

$$\mathbf{y}_B = \Phi_B \boldsymbol{\theta} + \mathbf{z}_B \quad (12)$$

and

$$\mathbf{y}_{\bar{B}} = \Phi_{\bar{B}} \boldsymbol{\theta} + \mathbf{z}_{\bar{B}} \quad (13)$$

Therefore, the unknown parameters are calculated from (12):

$$\boldsymbol{\theta} = \Phi_B^{-1} (\mathbf{y}_B - \mathbf{z}_B) \quad (14)$$

Injecting equation (14) in equation (13) gives:

$$\mathbf{y}_{\bar{B}} - (\Phi_{\bar{B}} \Phi_B^{-1}) \mathbf{y}_B = -(\Phi_{\bar{B}} \Phi_B^{-1}) \mathbf{z}_B + \mathbf{z}_{\bar{B}} \quad (15)$$

Let  $M_k = [-\Phi_{\bar{B}} \Phi_B^{-1}, \mathcal{I}_k]$  ( $\mathcal{I}_k$  is the identity matrix) be the constraints matrix, then equation (15) can be written in matrix form as follows:

$$M_k[\mathbf{z}_B; \mathbf{z}_{\bar{B}}] = M_k[\mathbf{y}_B; \mathbf{y}_{\bar{B}}] = \mathbf{c}_k \quad (16)$$

The constraints matrix dimension is  $k \times n$ , which means that there are  $k$  constraints over  $\mathbf{z}_n$ . The corresponding constraint vector  $\mathbf{c}_k$  is the vector containing the constraints values (previous information on  $\mathbf{z}_n$ ), its PDF is:

$$P(\mathbf{c}_k) \propto \exp \left[ -\frac{1}{2} \mathbf{c}_k^\top \sum_C^{-1} \mathbf{c}_k \right] \quad (17)$$

where  $\sum_C$  is the covariance matrix of  $\mathbf{c}_k$  given by:

$$\sum_C = M_k \sum_Z M_k^\top \quad (18)$$

where  $\sum_Z$  is the covariance matrix of  $\mathbf{z}_n$ . Because of the linearity and the jointly Gaussian character of the constraints, all the random variables remain Gaussian and therefore the expression in (11) may be computed. It is expanded in three terms as in (10) with the difference that the expectations are computed while taking into consideration the constraints. The expectation in the three terms are calculated using the Gaussian character of the variables.

The expectation in the first term is given by:

$$E_{/c_k}[\text{I}] = \left( m_{r/c_k}(x) \right)^2 + \sigma_{r/c_k}^2(x) \quad (19)$$

where,  $/c_k$  means that the probability density function is calculated conditionally to the constraints values.  $m_{r/c_k}(x)$  and  $\sigma_{r/c_k}^2(x)$  are respectively the mean and variance of  $r(x)/c_k$ . Using Bayes rules:

$$\begin{aligned} P(r(x)/c_k) &= \frac{P(r(x), \mathbf{c}_k)}{P(\mathbf{c}_k)} \\ &\propto \exp \left[ -\frac{1}{2} \frac{(r(x) - m_{r/c_k}(x))^2}{\sigma_{r/c_k}^2(x)} \right] \end{aligned} \quad (20)$$

The probability of  $(r(x), \mathbf{c}_k)$  is given by:

$$P(r(x), \mathbf{c}_k) \propto \exp \left[ -\frac{1}{2} [r(x) \ \mathbf{c}_k^\top] S_1^{-1} [r(x); \mathbf{c}_k] \right] \quad (21)$$

where  $S_1$  is the covariance matrix of  $[r(x); \mathbf{c}_k]$ . The identification of (20) with (21) gives the mean and variance of  $r(x)/c_k$ :

$$\begin{aligned} m_{r/c_k}(x) &= \frac{-\sum_{i=2}^{k+1} (S_1^{-1})_{1,i} \mathbf{c}_{i-1}}{(S_1^{-1})_{1,1}} \\ \sigma_{r/c_k}^2(x) &= \frac{1}{(S_1^{-1})_{1,1}} \end{aligned} \quad (22)$$

The first term computation (19) is fast because it just needs the inversion of the covariance matrix.

In the second term,  $[\mathbf{z}_{n+1} \mathbf{z}_{n+1}^\top]$  is a  $(n+1) \times (n+1)$  symmetric matrix. Therefore, the expectation in the second term is given by:

$$E_{/c_k}[\text{II}] = \text{tr}\{\psi(x) \psi^\top(x) E_{/c_k}[\mathbf{z}_{n+1} \mathbf{z}_{n+1}^\top]\} \quad (23)$$

Similarly to the first term computation,

$$\begin{cases} E_{/c_k}[z_i^2] = \left( m_{z_i/c_k}(x) \right)^2 + \sigma_{z_i/c_k}^2(x) \\ E_{/c_k}[z_i z_j] = m_{z_i/c_k}(x) m_{z_j/c_k}(x) + \sigma_{z_i z_j/c_k}^2(x) \end{cases} \quad (24)$$

where  $z_i$  is the  $i$ -th element of  $\mathbf{z}_{n+1}$ . The mean and variance of  $z_i/c_k$  are computed from (22) with the difference that  $S_1$  is replaced by the covariance matrix of  $[z_i; \mathbf{c}_k]$ . By following the same mathematical procedure as in first term calculation, the variance of  $z_i z_j/c_k$  is found to be:

$$\sigma_{z_i z_j/c_k}^2(x) = (S_{z_i z_j}^{-1})_{1,2} \quad (25)$$

where  $S_{z_i z_j}$  is the  $2 \times 2$  matrix in the upper left corner of  $S_2^{-1}$ :

$$S_{z_i z_j} = (S_2^{-1})_{1 \rightarrow 2, 1 \rightarrow 2} \quad (26)$$

where  $S_2$  is the covariance matrix of  $[z_i; z_j; \mathbf{c}_k]$ .

The expectation in the third term can be written as follows:

$$E_{/c_k}[\text{III}] = -2\psi^\top(x) E_{/c_k}[r(x) \mathbf{z}_{n+1}] \quad (27)$$

with,

$$E_{/c_k}[r(x) z_i] = m_{r/c_k}(x) m_{z_i/c_k}(x) + \sigma_{r z_i/c_k}^2(x) \quad (28)$$

where  $\sigma_{r z_i/c_k}^2(x)$  is computed using (25) and (26) with the difference that  $S_2$  is replaced by the covariance matrix of  $[r(x); z_i; \mathbf{c}_k]$ .

Finally, the criterion in (11) is computed by numerical integration of (19), (23) and (27). The computational burden is thus tractable.

#### 4. ILLUSTRATIVE EXAMPLE

Consider the quadratic model with cubic contamination ([4] and [5]) to be the illustrative example. That is, the regression model is a two-degree polynomial whereas the target is three-degree.

Here the target is precisely:

$$t(x) = x^3 + x^2 + x + 1 \quad (29)$$

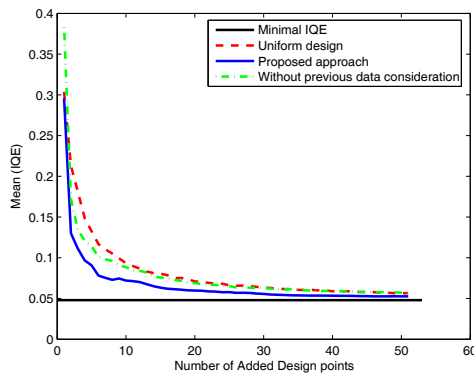
The Gaussian kernel is used because it is the most used kernel for the Gaussian process covariance [6]:

$$c(x, x') = s^2 \exp \left[ -\left( \frac{x - x'}{\lambda} \right)^2 \right], \quad \forall (x, x') \in \mathbb{X}^2 \quad (30)$$

where,  $s$  (Gaussian process variance) and  $\lambda$  (correlation distance) are the Gaussian process parameters. The corresponding space is infinite-dimensional. The kernel is used with the Gaussian process parameters values  $s = 0.8$  and  $\lambda = 0.4$ . The approach used to choose these values is based on a maximin efficiency criterion (see [2] for more details).

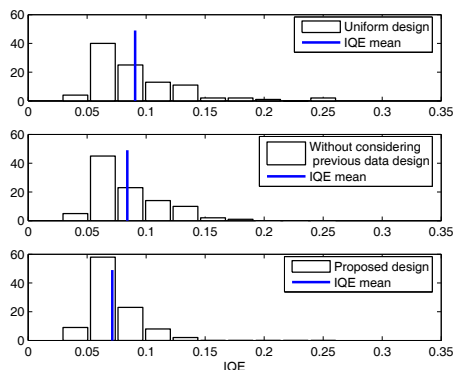
The design  $\xi_n = [-0.8, -0.4, 0, 0.4, 0.8]$  is taken to be the initial design. The centered interval  $[-1; 1]$  is taken to be the experimental domain  $\mathcal{X}$ .

The proposed approach is applied by varying the number of added points in the design from 1 to 50. The IQE PDF are computed by Monte-Carlo method with 100 sequences of noise where the observations error variance  $\sigma_e^2 = 0.05$ .



**Fig. 1.** Comparison among the three criterions

Figure 1 shows the performance (in terms of IQE) of the proposed approach, uniform design, and the IQE obtained by considering equation (7) (i.e. without taking into consideration the collected data as prior information). The results shows a faster convergence to the minimum IQE of the proposed approach over the other two approaches.



**Fig. 2.** IQE histograms

Another way of comparison is to compare the IQE histograms of the three designs for a fixed number of added de-

sign points. Figure 2 compares the IQE histograms of the three designs where 10 design points are added. The corresponding IQE means are shown in Table 1 (WOPI means the design without considering previous information). It is clear that the proposed approach is the most efficient because it gives the smallest IQE.

	Uniform	WOPI	Proposed
$\langle \text{IQE} \rangle$	0.0907	0.0845	0.0713

**Table 1.** IQE Means

## 5. CONCLUSION

This paper has proposed a sequential model-robust DOE criterion for linear models with misspecification. Modeling the misspecification with a Gaussian process, a sequential criterion that finds the design points sequentially while taking into consideration previously collected data is derived. Adding more than one design point at a time sequentially would increase the criterion optimization complexity but it is still applicable. Finally, a comparison between the proposed approach designs, and approaches that do not take into consideration previous information has shown that the proposed approach will give a better performance (in terms of IQE).

## 6. REFERENCES

- [1] R. X. Yue and F. J. Hickernell, “Robust designs for fitting linear models with misspecification,” *Statistica Sinica*, vol. 9, pp. 1053–1069, 1999.
- [2] M. Roger, L. Le Brusquet, and G. Fleury, “A criterion for model-robust design of experiments,” in *IEEE International Workshop on Machine Learning for Signal Processing*, Sao Luis (Brazil), September 29–October 1st, 2004, pp. 33–42.
- [3] G. E. P. Box and N. R. Draper, “A basis for the selection of a response surface design,” *Journal of the American Statistical Association*, vol. 54, pp. 622–654, 1959.
- [4] Z. Fang and D. P. Wiens, “Robust regression designs for approximate polynomial models,” *Journal of Statistical Planning and Inference*, vol. 117, pp. 305–321, 2003.
- [5] G. Montepiedra and V. V. Fedorov, “Minimum bias designs with constraints,” *Journal of Statistical Planning and Inference*, vol. 63, pp. 97–111, 1997.
- [6] C. K. I. Williams, “Prediction with Gaussian processes: From linear regression to linear prediction and beyond,” Tech. Rep. NCRG/97/012, Departement of Computer Science and Applied Mathematics, Aston University, Birmingham, UK, Oct. 1997.