



# SEQUENTIAL EXPERIMENTAL DESIGN FOR MISSPECIFIED NONLINEAR MODELS

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

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

## ABSTRACT

In design of experiments for nonlinear regression model identification, the design criterion depends on the unknown parameters to be identified. Classical strategies consist in designing sequentially the experiments by alternating the estimation and design stages. These strategies consider previous observations (collected data) only while estimating the unknown parameters during the estimation stages. This paper proposes to consider the previous observations not only during the estimation stages, but also by the criterion used during the design stages. Furthermore, the proposed criterion considers the robustness requirement: an unknown model error (misspecification) is supposed to exist and is modeled by a kernel-based representation (Gaussian process). Finally, the proposed sequential criterion is compared with a model-robust criterion which does not consider the previously collected data during the design stages, with the classical D-optimal and L-optimal criteria.

**Index Terms**— Sequential design of experiments, Gaussian process, Nonlinear regression, Robust design, Parameters identification.

## 1. INTRODUCTION

This paper addresses the problem of designing experiments for parameters identification for nonlinear regression models. The Design of Experiments (DOE) analysis is expected to entail a regression whose response function is nonlinear in the parameters.

Let  $t(x)$  be a target function which we desire to approximate by a nonlinear regression model  $\eta(\theta, x)$  where  $\theta$  is the parameters vector. Suppose that  $\theta^*$  is the parameters vector that will best approximate the target function  $t(x)$ :

$$\theta^* = \arg \min_{\theta} \int_{\mathcal{X}} (t(x) - \eta(\theta, x))^2 dx \quad (1)$$

where  $\mathcal{X}$  is the experimental domain.  $\theta^*$  is unknown and has to be estimated.

Suppose that a set of  $n$  collected data  $\{(x_i, y_i) \in \mathcal{X} \times \mathbb{R}, i = 1, \dots, n\}$  has already been collected. The  $x_i$ 's form the initial design denoted by  $\xi_n = [x_1, \dots, x_n]^T$ . The  $y_i$ 's are noisy observations of the target ( $y_i = t(x_i) + e_i$ ), where the observation errors  $e_i$  are normal and i.i.d. At this stage, the unknown parameters vector  $\theta^*$  may be estimated from the  $n$  collected observations using a nonlinear least square estimator as follows:

$$\hat{\theta}_n = \arg \min_{\theta} \sum_{i=1}^n (y_i - \eta(\theta, x_i))^2 \quad (2)$$

Suppose that we desire to refine the parameters estimation by adding a new design point  $x_{n+1}$  and its corresponding observation value  $y_{n+1}$  to the collected data. Then, the problem of sequential design of

experiments is to choose the next design point  $x_{n+1}$  that will refine the parameters estimation.

Classical sequential strategy was first proposed by Box and Hunter [1]. Afterwards, many works were progressed following the same classical strategy. Using the results in [2], Titterton and his collaborators [3] were able to show that the usual asymptotic analysis based on least squares estimation is still valid for sequential design in nonlinear models. As pointed out by several authors ([4], [5], [6]), the most attractive feature of sequential DOE is its ability to optimally utilize the dynamics of the learning process associated with experimentation and parameters identification.

The classical sequential strategy discussed in the literature considers the previously collected data while computing  $\hat{\theta}_n$  during the estimation stages. This paper suggests to also consider the collected data while deriving the design criterion during the design stages.

Moreover, classical experimental design criteria consider that the target function  $t(x)$  is perfectly represented by the regression model  $\eta(\theta^*, x)$ . This will introduce a bias in the parameters estimation. An important work that solves such drawbacks was done by Yue and Hickernell [7]. In our paper, the model error (misspecification) is considered and modeled by a kernel-based representation (Gaussian process).

The paper is organized as follows. Section 2 presents two classical design of experiments criteria (D-optimality, L-optimality). These two criteria are then used in other sections for comparison purposes with the new proposed criterion. Although the meaning of the proposed criterion is quite natural, its derivation is a challenging task. Therefore, section 3 presents the proposed criterion and its associated mathematical developments. In section 4, the new approach is applied on a nonlinear regression example. The obtained designs are compared with other designs obtained from the same criterion without taking into consideration the previously collected data during the design stages, the D-optimal and L-optimal criteria.

## 2. CLASSICAL DESIGN OF EXPERIMENTS CRITERIA

Presenting the L-optimal criterion will be helpful for the proposed criterion derivation because it is based on this criterion. The D-optimal criterion is presented for comparison purpose in section 4.

### 2.1. D-optimality

Having a set of  $n$  collected data, the D-optimality in nonlinear problems consists in choosing the next design point  $x_{n+1}^*$  which minimizes the determinant of the inverse of the Fisher matrix:

$$x_{n+1}^* = \arg \min_{x_{n+1} \in \mathcal{X}} \det \left( \nabla M_{n+1}^\top \nabla M_{n+1} \right)^{-1} \quad (3)$$

where  $\nabla M_{n+1}$  is a  $(n+1) \times d$  matrix where each row is equal to  $(\nabla \eta(\hat{\theta}_n, x_i))^\top$  which is the gradient of  $\eta(\hat{\theta}_n, x)$ .

## 2.2. L-optimality

Let  $x_{n+1}$  be a candidate value for the new design point and  $y_{n+1}$  the observation made at this design point. Then,  $\hat{\theta}_{n+1}$  the nonlinear least square estimation of  $\theta$ , is computed using  $n + 1$  points:

$$\hat{\theta}_{n+1} = \arg \min_{\theta_{n+1}} \sum_{i=1}^{n+1} (y_i - \eta(\theta_{n+1}, x_i))^2 \quad (4)$$

The L-optimality criterion attempts to choose the new design point  $x_{n+1}$  that minimizes the average prediction error over the entire experimental domain. The prediction error is defined as the Integral Quadratic Error (IQE):

$$\text{IQE}(x_1, \dots, x_{n+1}, e_1, \dots, e_{n+1}) = \int_{\mathcal{X}} |t - \hat{t}|^2 dx \quad (5)$$

where  $\hat{t}(x) = \eta(\hat{\theta}_{n+1}, x)$  is the target function estimation.

The IQE depends on the known design points  $\xi_n$ , the new design point  $x_{n+1}$  and the observation errors which are unknown. Thus, an expectation over the observation errors is taken in order to ensure a good performance averagely over their realizations.

Taking the total expectation of the IQE in (5), the Integral Quadratic Risk (IQR) can be written as follows:

$$\text{IQR}(x_{n+1}) = E_{(e_{n+1})} \left[ \int_{\mathcal{X}} |t - \hat{t}|^2 dx \right] \quad (6)$$

where  $e_{n+1}$  is the observations error vector. The L-optimality consists in choosing the design point  $x_{n+1}^*$  that minimizes (6):

$$x_{n+1}^* = \arg \min_{x_{n+1} \in \mathcal{X}} [\text{IQR}(x_{n+1})] \quad (7)$$

Solving the optimization problem in (7) requires an analytic expression of  $\hat{t}$  and therefore an analytic expression of  $\hat{\theta}_{n+1}$  in (4).

Suppose that for a sufficient number of observations  $\hat{\theta}_n$ ,  $\theta^*$  and  $\hat{\theta}_{n+1}$  are approximately the same. Then, a first order Taylor series expansion may be used to linearize the model around the the estimated parameters:

$$\eta(\theta^*, x) \approx \eta(\hat{\theta}_n, x) + \nabla \eta(\hat{\theta}_n, x)^\top (\theta^* - \hat{\theta}_n) \quad (8)$$

Therefore, the error in (4) may be approximated by:

$$\begin{aligned} y_i - \eta(\theta_{n+1}, x_i) &\approx \eta(\hat{\theta}_n, x_i) + \nabla \eta(\hat{\theta}_n, x_i)^\top (\theta^* - \hat{\theta}_n) \\ &+ e_i - \eta(\hat{\theta}_n, x_i) \\ &- \nabla \eta(\hat{\theta}_n, x_i)^\top (\theta_{n+1} - \hat{\theta}_n) \end{aligned} \quad (9)$$

Replacing the error by its approximation, (4) is written as follows:

$$\hat{\theta}_{n+1} = \arg \min_{\theta_{n+1}} \sum_{i=1}^{n+1} \left( \nabla \eta(\hat{\theta}_n, x_i)^\top (\theta^* - \theta_{n+1}) + e_i \right)^2 \quad (10)$$

For simplification and computation purposes, the above equation is written in vectorial form as follows:

$$\hat{\theta}_{n+1} = \arg \min_{\theta_{n+1}} \|\nabla M_{n+1} (\theta^* - \theta_{n+1}) + e_{n+1}\|^2 \quad (11)$$

Therefore, the solution of (11) is given by:

$$\hat{\theta}_{n+1} = \left( \nabla M_{n+1}^\top \nabla M_{n+1} \right)^{-1} \nabla M_{n+1}^\top e_{n+1} + \theta^* \quad (12)$$

The L-optimality IQR is then rewritten in an explicit form:

$$\text{IQR}(x_{n+1}) = E_{(e_{n+1})} \int_{\mathcal{X}} \left[ |\psi_{n+1}(x) \eta(\hat{\theta}_n, x) e_{n+1}|^2 \right] dx \quad (13)$$

where  $\psi_{n+1}(x) = (\nabla \eta(\hat{\theta}_n, x))^\top (\nabla M_{n+1}^\top \nabla M_{n+1})^{-1} \nabla M_{n+1}^\top$ . This expression can be written in a simplified form:

$$\text{IQR} = \sigma_e^2 \text{tr} \left( \mathbf{I}_{\eta\eta} \left( \nabla M_{n+1}^\top \nabla M_{n+1} \right)^{-1} \right) \quad (14)$$

where  $\mathbf{I}_{\eta\eta} = \int_{\mathcal{X}} \eta(\hat{\theta}_n, x) \eta(\hat{\theta}_n, x)^\top dx$  which can be computed analytically. One can see that the last IQR expression is suitable for implementation and optimization.

## 3. THE PROPOSED CRITERION DERIVATION

Now, what if the target function is not perfectly represented by the regression model? A model error (misspecification) exists. Therefore, the target function  $t$  is:

$$t(x) = \eta(\theta^*, x) + r(x) \quad (15)$$

where the misspecification  $r(x)$  is an unknown function. We choose to model it by a Gaussian process [6]. A Gaussian process is a random field defined by its mean and covariance function:

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

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  $x_{n+1}$  will guarantee a good level of performance (on average) over the set of potential misspecifications.

Now, the estimated parameters  $\hat{\theta}_{n+1}$  depends on the model error  $r(x)$ , thus equation (12) becomes:

$$\hat{\theta}_{n+1} = \psi_{n+1} z_{n+1} + \theta^* \quad (16)$$

where  $z_{n+1}$  is the *observations-model* errors vector generated by a Gaussian process  $z(x)$ : mean 0 and covariance  $c(x, x') + \sigma_e^2 \delta(x - x')$ .

The chosen statistical representation for  $r(x)$  allows to take expectation of the IQE in (5) over the model and observation errors. The IQR of the Model-Robust criterion is written as follows:

$$\text{IQR}(x_{n+1}) = E_{(e_{n+1}, r)} \int_{\mathcal{X}} |t - \hat{t}|^2 dx \quad (17)$$

Having a set of  $n$  previously collected observations will provide important information about the random variables  $e_n$  and  $r(x)$ . Hence, introducing this information in the design criterion will improve the criterion performance and refine the parameters estimation. This idea was first discussed in [8] showing its workability in linear situations. This paper will adopt this idea for nonlinear situations. Therefore, the criterion to be used in the design stages is as follows [8]:

$$\text{IQR}(x_{n+1}) = E_{(e_{n+1}, r)/CD} \int_{\mathcal{X}} |t - \hat{t}|^2 dx \quad (18)$$

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

By expanding the previous equation, the IQR can be written as follows:

$$\begin{aligned} \text{IQR}(x_{n+1}) = & \int_{\mathcal{X}} \left\{ \underset{(e,r)/CD}{E} [r^2(x)] \right\} dx \\ & + \int_{\mathcal{X}} \left\{ \underset{(e,r)/CD}{E} [\psi_{n+1}^\top(x) z_{n+1} z_{n+1}^\top \psi_{n+1}(x)] \right\} dx \quad (19) \\ & - 2 \int_{\mathcal{X}} \left\{ \underset{(e,r)/CD}{E} [r(x) \psi_{n+1}^\top(x) z_{n+1}] \right\} dx \end{aligned}$$

The expectation calculation in equation (19) becomes a bit more complicated because of considering previous information. Having a set of collected data will provide information about the random variables  $z_n$ , which is introduced in form of constraints. Let  $\mathbf{y}_n$  be the  $n \times 1$  vector of known observations  $y_i$ . Then,  $\mathbf{y}_n = \eta(\boldsymbol{\theta}^*, \boldsymbol{\xi}_n) + \mathbf{z}_n$ . According to (8), this equation can be written as follows:

$$\mathbf{y}_n = \eta(\boldsymbol{\theta}^*, \boldsymbol{\xi}_n) + \nabla \eta(\hat{\boldsymbol{\theta}}_n, \mathbf{x})^\top (\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_n) + \mathbf{z}_n \quad (20)$$

Generally,  $n > d$ . Therefore, the model matrix  $\nabla \mathbf{M}_n$  may be divided into two sub-matrices:  $\nabla \mathbf{M}_B$  a  $d \times d$  reversible matrix and  $\nabla \mathbf{M}_{\bar{B}}$  a  $(n-d) \times d$  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, (20) can be written as follows:

$$\begin{aligned} \mathbf{y}_{\bar{B}} = & \eta(\boldsymbol{\theta}^*, \boldsymbol{\xi}_n) + \nabla \mathbf{M}_{\bar{B}} \psi_B \mathbf{y}_B - \nabla \mathbf{M}_{\bar{B}} \psi_B \eta(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\xi}_n) \\ & - \nabla \mathbf{M}_{\bar{B}} \psi_B \mathbf{z}_B + \mathbf{z}_{\bar{B}} \quad (21) \end{aligned}$$

Let  $\mathbf{N}_k = [\nabla \mathbf{M}_{\bar{B}} \psi_B, \mathcal{I}_k]$  ( $\mathcal{I}_k$  is the identity matrix) be the constraints matrix, then equation (21) can be written in matrix form as follows:

$$\begin{aligned} \mathbf{N}_k [\mathbf{y}_B; \mathbf{y}_{\bar{B}}] = & \mathbf{N}_k [\mathbf{z}_B - \eta(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_B); \mathbf{z}_{\bar{B}} - \eta(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_{\bar{B}})] \\ = & \mathbf{c}_k \quad (22) \end{aligned}$$

The constraints matrix dimension is  $k \times n$ , which means that there are  $k$  constraints over  $z_n$ . The constraints vector  $\mathbf{c}_k$  is computed from the residue  $z_n$ . The Probability Density Function (PDF) of the constraints is:

$$P(\mathbf{c}_k) \propto \exp \left[ -\frac{1}{2} \mathbf{c}_k^\top \sum_C^{-1} \mathbf{c}_k \right] \quad (23)$$

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

$$\sum_C = \mathbf{N}_k \sum_Z \mathbf{N}_k^\top \quad (24)$$

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 IQR in (19) may be computed. In the following, a detailed explanation of the computation procedure of (19) is given. The expectation in (19) is taken over the observations error and model error. Therefore, it is required to compute  $\underset{(e_{n+1},r)/CD}{E} (X^2)$  and  $\underset{(e_{n+1},r)/CD}{E} (XY)$  where  $X$  and  $Y$  can be the model error  $r(x)$  or the model-observations error  $z(x)$ . Therefore, one has to compute the conditional mean and variance of  $X$  or  $Y$  and the jointly conditional variance of  $X$  and  $Y$ . Using Bayes rules:

$$\begin{aligned} P(X/c_k) = & \frac{P(X, \mathbf{c}_k)}{P(\mathbf{c}_k)} \\ \propto & \exp \left[ -\frac{1}{2} \frac{(X - m_{X/c_k}(x))^2}{\sigma_{X/c_k}^2(x)} \right] \quad (25) \end{aligned}$$

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

$$P(X, \mathbf{c}_k) \propto \exp \left[ -\frac{1}{2} [X \ \mathbf{c}_k^\top] S^{-1} [X; \mathbf{c}_k] \right] \quad (26)$$

where  $S$  is the covariance matrix of  $[X; \mathbf{c}_k]$  constructed from equation (24). The identification of (25) with (26) gives the mean and variance of  $X/c_k$ :

$$\begin{aligned} m_{X/c_k}(x) = & \frac{-\sum_{i=2}^{k+1} (S^{-1})_{1,i} \times \mathbf{c}_{i-1}}{(S^{-1})_{1,1}} \\ \sigma_{X/c_k}^2(x) = & \frac{1}{(S^{-1})_{1,1}} \quad (27) \end{aligned}$$

As can be seen from the second and third terms of the IQR (19), one has to compute the jointly conditional variances of  $r(x), z(x_i)$  and  $z(x_i), z(x_j)$ . The way of computing this jointly conditional variance is different from the one discussed above:

$$\sigma_{XY/c_k}^2(x) = (S_{XY}^{-1})_{1,2} \quad (28)$$

where  $S_{XY}$  is the  $2 \times 2$  matrix in the upper left corner of  $S$ :

$$S_{XY} = (S_1^{-1})_{1 \rightarrow 2, 1 \rightarrow 2} \quad (29)$$

and  $S_1$  is the covariance matrix of  $[X; Y; \mathbf{c}_k]$ .

Finally, the integrals in equation (19) are calculated using numerical integration. The computational burden is thus tractable.

#### 4. ILLUSTRATIVE EXAMPLE

Consider the following nonlinear model:

$$\eta(\boldsymbol{\theta}, x) = \frac{1}{1 + \exp(-\theta_1 - \theta_2 x)} + \frac{1}{1 + \exp(-\theta_1 + \theta_2 x)} \quad (30)$$

The target function  $t(x) = \eta(\boldsymbol{\theta}, x) + p(x)$ , where  $p(x)$  is a polynomial of degree  $m$  that represents the deviation from the nonlinear model (misspecification). Let  $\theta_1 = 1, \theta_2 = 6$  and  $m = 6$ .

The Gaussian kernel is used because it is the most used kernel for the Gaussian process covariance [9]:

$$c(x, x') = s^2 \exp \left[ -\left( \frac{x - x'}{\lambda} \right)^2 \right], \quad \forall (x, x') \in \mathcal{X}^2 \quad (31)$$

where,  $s^2$  (Gaussian process variance) and  $\lambda$  (correlation distance) are the Gaussian process parameters. The kernel is used with the Gaussian process parameters values  $s^2 = 1$  and  $\lambda = 0.6$ . The approach used to choose these values is based on a maximin efficiency criterion [10].

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

The proposed approach is applied by varying the number of added points in the design from 1 to 20. The IQE PDF are computed by Monte-Carlo method with 100 sequences of noise where the observations error variance  $\sigma_e^2 = 0.05$ .

Figure 1 shows the performance (in terms of IQR) of the complete proposed approach (eq.(18)), the proposed approach without considering the collected data as prior information (eq.(17)), L-optimal design (eq.(6)) and D-optimal design (eq.(3)).

The results show the advantage of considering the model errors (faster convergence to the minimum IQR of the proposed approaches over the other two approaches) and the advantage of computing all the probabilities conditionally to the collected data.

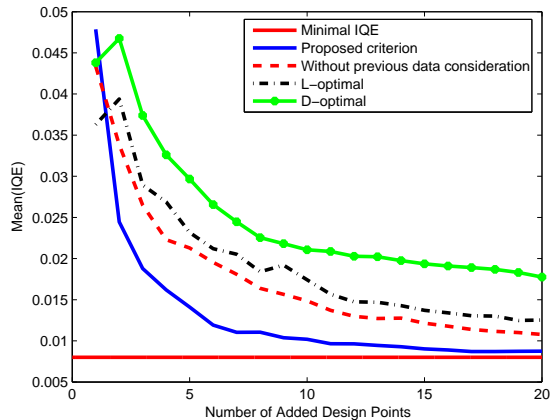


Fig. 1. Comparison among the four criteria

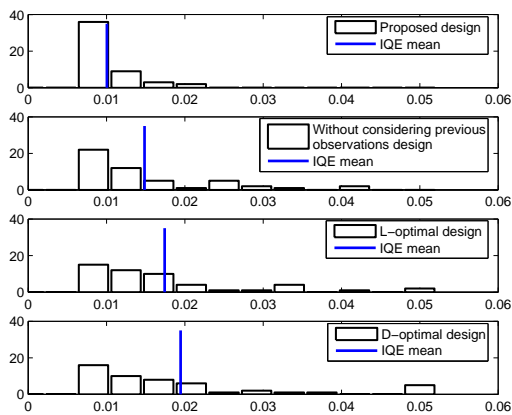


Fig. 2. IQE histograms

Another illustration is the comparison of the IQE histograms for a fixed number of added design points. Figure 2 gives the IQE histograms of the three designs where 10 design points are added. The corresponding IQE means are shown in Table 1.

Figure 3 is an example of the target function with the approximated model (30) where its parameters are estimated using the design points (6 added design points) obtained with the L-optimal and the proposed design criteria. It can be seen that the proposed criterion gives a better performance over the classical L-optimality.

	D-optimal	L-optimal	Proposed without / $CD$	Proposed with / $CD$
$\langle IQE \rangle$	0.0211	0.0174	0.0149	0.0102

Table 1. IQE Means

## 5. CONCLUSION

This paper has proposed a sequential model-robust DOE criterion for nonlinear regression problems. The proposed criterion takes into consideration the previously collected data when designing experiments during the design stages. Moreover, the proposed criterion considers the problem of model robustness by taking into account

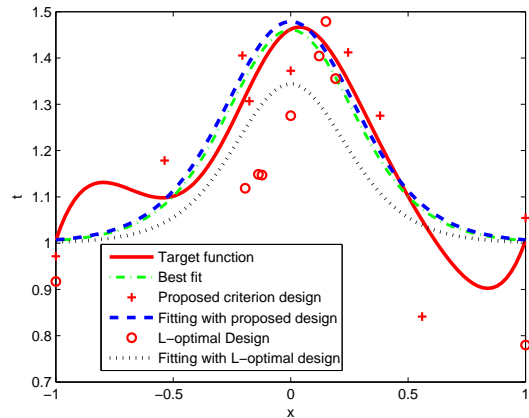


Fig. 3. Model fitting using the proposed criterion and L-optimal

a model error which is modeled by a Gaussian process. Finally, an illustrative example has shown that the proposed criterion will give better performance over criteria that do not consider previously collected data during the design stages.

## 6. REFERENCES

- [1] G. E. P. Box and W. G. Hunter, "Sequential design of experiments for nonlinear models," in *In Processing of the IBM Scientific Computing Symposium on Statistics*, Octobre 21-23, 1965, pp. 113-137.
- [2] T. L. Lai and C. Z. Wei, "Least squares estimates in stochastic regression models with application to identification and control of dynamic systems," *Ann. Statist.*, vol. 10, pp. 154-166, 1982.
- [3] D. M. Titterton I. Ford and C. F. J. Wu, "Inference and sequential design," *Biometrika*, vol. 72, pp. 545-551, 1985.
- [4] H. Chernoff, "Approaches in sequential design of experiments," in *A Survey of Statistical Design and Linear Models (Edited by J. N. Srivastava)*, pp. 67-90, 1975.
- [5] S. D. Silvey, *Optimal design*, Chapman and Hall, London, 1980.
- [6] L. Pronzato E. Walter, *Identification of Parametric Models from Experimental Data*, Springer, Communications and Control Engineering Series, Londres, 1997.
- [7] R. X. Yue and F. J. Hickernell, "Robust designs for fitting linear models with misspecification," *Statistica Sinica*, vol. 9, pp. 1053-1069, 1999.
- [8] H. El Abiad, L. Le Brusquet, M. Roger, and M. Davoust, "Model-Robust Sequential Design of Experiments for Identification Problems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii (US), April 15-April 20, 2007, vol. 2, pp. 441-444.
- [9] C. K. I. Williams, "Prediction with Gaussian processes: From linear regression to linear prediction and beyond," Tech. Rep. NCRG/97/012, Department of Computer Science and Applied Mathematics, Aston University, Birmingham, UK, Oct. 1997.
- [10] 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.