# Identification and Confidence Region of Parameters in non Linear and Dynamical Systems

**S. Massebeuf[1], C. Fonteix[2], I. Marc[1], F. Pla[1]**

[1] Laboratoire des Sciences du Génie Chimique – U.P.R. C.N.R.S. 6811
E.N.S.I.C., 1 rue Grandville, BP 451, F-54001 Nancy Cedex

[2] Laboratoire des Sciences du Génie Chimique – U.P.R. C.N.R.S. 6811
E.N.S.A.I.A., 2 avenue de la forêt de Haye,
BP 172, F-54505 Vandœuvre-lès-Nancy Cedex
Telephone: (33) 3 83 59 58 42; facsimile: (33) 3 83 59 58 04
e-mail: fonteix@ensaia.u-nancy.fr

***Abstract***

*When a model is elaborated, the first thing is to see if the experiments are well described by the simulation responses. If the model has not to be changed or adapted, a statistical analysis has to be done to have a satisfactory confidence region of the parameters. If the statistical reliability of the parameters is suitable, the model can be used for a more advanced study of the system, else, an optimal experiment design may be developed in order to reduce the costs of experiments. The aim of this study is to elaborate a method to be nearing the true confidence region. It is based on a diploid genetic algorithm which allows to have parameters estimation and their joint confidence region formed by a set of points. An enzymatic reaction is presented as a non linear biological application and the experimental data vs time allows to propose a dynamical model.*

## I. INTRODUCTION

Many methods exist for determining optimal parameter estimates in systems of ordinary nonlinear differential equations. In data analysis, it is also very important to have a realistic measure of the statistical confidence of the parameter estimates. This requires the calculation of at least individual confidence limits and preferably the joint confidence region associated with the parameters. Some statements must be reported: the most probable value of the parameters, the confidence region of the parameters and a probability statement about the confidence domain (Van Boekel 1996). An idea of the magnitude of such regions is of special interest when the models are based on mechanistic considerations. Indeed, the parameters in nonlinear systems often have a specific physical meaning.

The individual confidence limits at the selected probability level of $(1-\alpha)$ are often determined in the litterature but they do not take into account the variability of all parameters simultaneously. The remaining parameters are held constant at their optimal value while individual confidence limits are determined. The joint confidence region of the parameters certainly give more information about their accuracy (Emig and Hosten 1974).

Nonlinear systems are often considered nearly linear in the parameters around the optimal set of parameters. So, with the linearization of the model in its vicinity, it is developed in Taylor series retaining only the partial derivatives of first or second

order (Frederiksen 1998). In this case, the theory assumes that the mathematical model can be adequately linearized in the neighbourhood of the optimal parameter estimates. Another method is to treat confidence regions as constraints in a nonlinear programming model. It is shown that the confidence regions depend on the value of the Lagrange multiplier of the region's constraint (Del Castillo 1996). All these authors make linear assumptions, which can be checked near optimal parameter estimates, to have an ellipsoidal approximation of the confidence domain. So, parameter estimates and their joint confidence regions are only as good as the linear approximation. The calculated confidence domains are asymptotic and therefore they may underestimate the real confidence regions.

Estimation of confidence region for nonlinear models is not straightforward. The nonlinearity of the function and the experimental data nature give an asymetry to the confidence region contrary to the linear one. Recently, the nonlinear approach has been developed (Walter and Pronzato 1994) to determine the rigorous confidence region of the parameters. Without knowledge of the standard deviations of the measurement errors, the authors give the expression of the true confidence domain by using orthogonal projection on tangential plane to expectation surface of the model output.

In this article, we use the two approaches to find the confidence region of parameters of a biochemical example. After the choice of the kinetic model for the studied enzymatic reaction and its parametric identification, the uncertainty on parameters is defined with linearization of the system on the one hand and with the Walter-Pronzato's approach on the other hand. The parameters are obtained by minimization of a discrepancies function between experimental data and simulation responses, and the confidence region by statistical test optimization. So, an optimization method has been developed, based on genetic algorithms, to approximate the confidence region by a set of points.

## II. IDENTIFICATION AND CONFIDENCE REGION OF PARAMETERS METHODOLOGIES

### II.1. A diploid genetic algorithm as an optimization method

Genetic algorithms can be considered as a stochastic optimization method which is able to search in a large space and to make evolve solutions set (called population). This method is inspired by Darwin's concept by analogy with the evolution of populations (Goldberg 1989) and its popularity has increased in the last decade to locate optimal solutions in complex landscapes and to keep a set of good solutions in the last generations. In our problems, a Diploid Genetic Algorithm (DGA) is used whose principles were previously elaborated (Perrin *et al.* 1997). The diploid version is kept because its performances were found to be better compared with a haploid one (Fonteix *et al.* 1995). Each individual (which can be a possible solution of the problem) is described by a four-tuple ($a_j$, $a_j$', $D_j$, $x_j$). $a_j$ and $a_j$' represent the two alleles of the j gene, $D_j$ is the dominance of one allele over the other chosen in {0,1} values and $x_j$ represents the phenotype which is the result of the combination of the respective alleles, $a_j$ and $a_j$':

$$x_j=D_j.a_j+(1-D_j).a_j' \quad (1)$$

An initial population is randomly created by generating a set of m points from the search area. Each point is tested and evaluated. If this population is not the

solution, then selection and genetic operators are used to make it evolve. Only the better individuals will survive (elitist selection) and participate in the creation of a new generation. The reproduction of the individuals in the diploid model consists of a multi-crossover of the two chromosomes of each parent, a mutation and a homozygosity. Mutation is, for the selected individual, a randomly draw for all the genes of the two chromosomes and all the dominances. Homozygosity allows to modify a child by copying out their phenotypes on their two chromosomes. A homozygote is created with $a_j$ and $a_j$' defined by the multi-crossover method and the dominance is as $0<D_j<1$. The phenotype is deducted: $x_j=D_j.a_j+(1-D_j).a_j$' and alleles become equal: $a_j=a_j'=x_j$. If the generated child is worse than the worst parent, he is not adapted, he is eliminated and another is created to complete the generation. The population of each generation is evaluated until it satisfies the stop criterion: $f_{max}-f_{min}<\varepsilon$ where $f_{min}$ and $f_{max}$ are respectively the minimal and the maximal objective function values in the current population and $\varepsilon$ is the given precision for solution estimation. Figure 1 summarizes the working of the DGA.

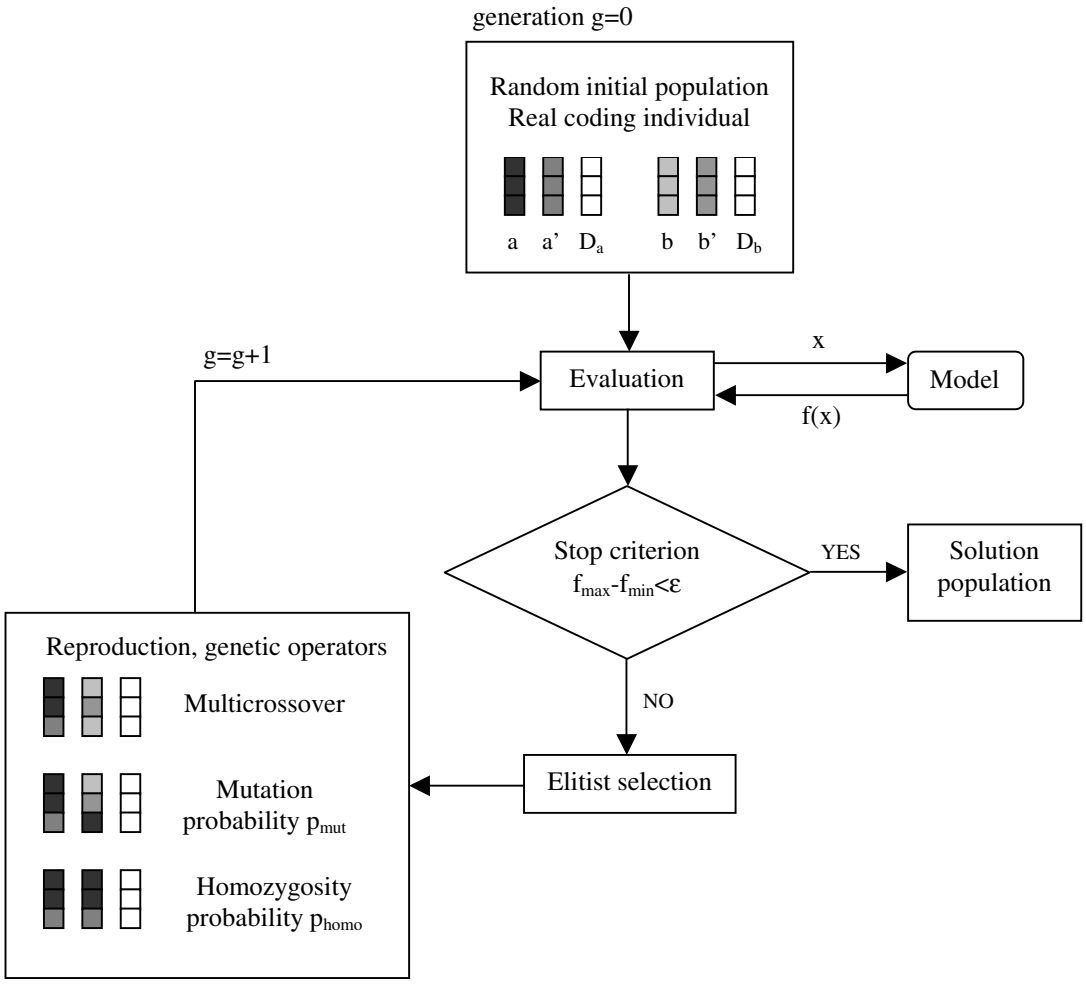


Figure 1: diagram of the diploid genetic algorithm.

## II.2. Parametric identification

When a mathematic expression is proposed to describe physical, chemical or biological aspect of a system, the goal is to obtain the best model as possible. This assertion needs a criterion definition $J(\theta)$, scalar function of model parameters $\theta$

which have to be optimised. The parameters optimal value depends on the choice of the criterion and a certain can be elaborated with the knowledge and the assumptions about statistical properties of the measurement errors. The fitness criterion depends on the discrepancies between the calculated and the measured values: $y_{jmod}(t_i,\theta)$-$y_{jexp}(t_i)$ for the kind of measure j at time $t_i$. We do not make an overview of the different used criteria in this article but we present the most useful approaches.

A usual approach is to estimate the unknown parameters such that the weighted sum of squared discrepancies is minimal:

$$J(\theta) = \sum_{j=1}^{N} \sum_{i=1}^{n_{jexp}} \omega_j . \left( y_{jmod}(t_i,\theta) - y_{jexp}(t_i) \right)^2 \quad (2)$$

where the positive weights, $\omega_j$, are based on the accuracy of the measurements, N the kinds of response number and $n_{jexp}$ the number of j measurements. This is called the weighted least squares criterion when *a priori* knowledge about the accuracy of the measurements is available.

In most practical situations, the standard deviations of the measurement errors, $\sigma_i$, are unknown and different kind of responses, which can be of different rough estimates, are measured. The fitness criteria used in nonlinear system depend on the assumptions and knowledge about the measurement errors. From the probability density function of the measurement errors $\tilde{f}$, the maximum likelihood estimates of the parameters can be derived by maximizing (or minimizing the opposite) its logarithm:

$$\text{MIN} \quad J(\theta) = -\ln\left\{ \prod_{j=1}^{N} \prod_{i=1}^{n_{jexp}} \tilde{f}\left( y_{jmod}(t_i,\theta) - y_{jexp}(t_i) \right) \right\} \quad (3)$$

If we assume that the measurement errors are stochastically independent and normally distributed $N(0,\sigma_j^2)$, the log-likelihood is then deducted:

$$J(\theta) = \sum_{j=1}^{N} \left\{ \frac{n_{jexp}}{2} \ln(2\pi.\sigma_j^{\,2}) + \frac{1}{2\sigma_j^{\,2}} \sum_{j=1}^{n_{jexp}} \left( y_{jmod}(t_i,\theta) - y_{jexp}(t_i) \right)^2 \right\} \quad (4)$$

A lot of other criteria can be elaborated with other assumptions on the measurement errors distribution (Walter 1987) but the aim of this article is not to make a state of the art in parametric identification. The studied application has only one kind of response, so the choice of the criterion is simplified (N=1 and $\omega_j$=1) because fitnesses are equivalent and are restricted to the least squares criterion. In the following parts, only the mono-response case is taken into account.

## II.3. Uncertainty on parameters

In the general case, the search of the best parameters value for the chosen criterion is not always enough. With the uncertainties on experimental data and numerical errors, uncertainties on parameters have to be estimated to validate the model. Only one response is available, so the following notations $n_{jexp}$, $y_{jmod}$, $y_{jexp}$ and $\sigma_j$ are respectively transformed into $n_{exp}$, $y_{mod}$, $y_{exp}$ and $\sigma$.

### II.3.1. Known variance

Let be denoted prediction error: $e(t_i,\theta^*)=\varepsilon(t_i)$ for $i=1..n_{exp}$ experiments and where $\varepsilon(t_i)$ are stochastically independent variable normally distributed $N(0,\sigma^2)$ with

known variance $\sigma^2$. $\theta^*$ is the true parameters vector. The maximum likelihood method gives the criterion to minimize:

$$J(\theta) = \sum_{i=1}^{n_{exp}} [e(t_i, \theta)]^2 \quad (5)$$

For the true parameters values: $J(\theta^*) = \sum_{i=1}^{n_{exp}} [\varepsilon(t_i)]^2 \approx n_{exp}.\sigma^2$ (6)

So, according to standard statistics $\frac{J(\theta^*)}{\sigma^2}$ has $\chi^2$ distribution with $n_{exp}$ degrees of freedom. The points loci $\theta$ described by: $J(\theta) \leq \sigma^2.\chi^2_\alpha(n_{exp})$ (7)

give a 100(1-$\alpha$) % confidence region for parameters.

**II.3.2. Unknown variance**

In most practical situations, the response variance is not known and prediction error is supposed as an output error:

$$e(t_i, \theta) = y_{mod}(t_i, \theta) - y_{exp}(t_i) \quad (8)$$

When $\theta$ is varying, the output $y_{mod}$ describe a hypersurface $S_m$, which is a hyperplan for a linear model in parameters and a curved hypersurface in the general case for a nonlinear one. Let $\Pi(t_i,\theta)$ be the orthogonal projection matrix on the tangential plane to $S_m$ in $y_{mod}(t_i,\theta)$ (Walter and Pronzato 1994):

$$\Pi(t_i, \theta) = \frac{\partial y_{mod}(t_i, \theta)}{\partial \theta^T} \left\{ \left[ \frac{\partial y_{mod}(t_i, \theta)}{\partial \theta^T} \right]^T \left[ \frac{\partial y_{mod}(t_i, \theta)}{\partial \theta^T} \right] \right\}^{-1} \left[ \frac{\partial y_{mod}(t_i, \theta)}{\partial \theta^T} \right]^T \quad (9)$$

$n_{exp}$=dim ($y_{exp}$) and $n_\theta$=dim ($\theta$) are denoted.

$J(\theta^*) = e^T(t_i, \theta^*)e(t_i, \theta^*)$ has $\sigma^2\chi^2(n_{exp})$ distribution because $e(t_i,\theta)$ is on a $n_{exp}$ dimension space.

$e^T(t_i, \theta^*)\Pi(t_i, \theta^*)e(t_i, \theta^*)$ has $\sigma^2\chi^2(n_\theta)$ distribution because it is the error projection on a $n_\theta$ dimension hyperplan.

$e^T(t_i, \theta^*)[I - \Pi(t_i, \theta^*)]e(t_i, \theta^*)$ has $\sigma^2\chi^2(n_{exp}\text{-}n_\theta)$ distribution and is the previous projection complementary.

$e^T\Pi e$ and $e^T[I\text{-}\Pi]e$ are independent (orthogonality) so:

$\frac{e^T(t_i, \theta^*)\Pi(t_i, \theta^*)e(t_i, \theta^*)}{e^T(t_i, \theta^*)[I - \Pi(t_i, \theta^*)]e(t_i, \theta^*)} \frac{n_{exp} - n_\theta}{n_\theta}$ has $F(n_\theta, n_{exp}\text{-}n_\theta)$ distribution (Fisher-Snedecor's distribution with $n_\theta$ and $n_{exp}\text{-}n_\theta$ degrees of freedom). So, the 100(1-$\alpha$) % confidence region of the parameters $\theta^*$ corresponds to (Walter and Pronzato 1994):

$$f(\theta) = \frac{e^T(t_i, \theta)\Pi(t_i, \theta)e(t_i, \theta)}{e^T(t_i, \theta)e(t_i, \theta)} \leq \frac{\frac{n_\theta}{n_{exp} - n_\theta} F_\alpha(n_\theta, n_{exp} - n_\theta)}{1 + \frac{n_\theta}{n_{exp} - n_\theta} F_\alpha(n_\theta, n_{exp} - n_\theta)} \quad (10)$$

The probability, that the true value of the parameters is in the domain so defined, is (1-$\alpha$) and the expression is independent of the variance. It is essential to notice that

the number of experiments has to be higher than the number of parameters to be identified: $n_{exp} > n_\theta$. If $n_{exp} = n_\theta$ (particular case depending on the experimental design), the rigorous confidence domain of the parameters cannot be defined by the Walter-Pronzato's approach but by linearization of the model (Cunha 1999).

### II.3.3. Linearization

Let $\hat{\theta}$ be denoted an estimation of the true parameters θ* by the minimization of the criterion J(θ).

$J(\theta^*) - J(\hat{\theta})$ has $\sigma^2\chi^2(n_\theta)$ distribution.

$J(\hat{\theta})$ has $\sigma^2\chi^2(n_{exp}\text{-}n_\theta)$ distribution.

So, $\frac{J(\theta^*) - J(\hat{\theta})}{J(\hat{\theta})}\frac{n_{exp} - n_\theta}{n_\theta}$ has $F(n_\theta, n_{exp}\text{-}n_\theta)$ distribution if the two previous are independent. With the first order approximation, the 100(1-α) % confidence region of the parameters θ* is given by (Draper and Smith 1981):

$$\frac{J(\theta) - J(\hat{\theta})}{J(\hat{\theta})} \leq \frac{n_\theta}{n_{exp} - n_\theta} F_\alpha(n_\theta, n_{exp} - n_\theta) \tag{11}$$

This expression needs a first estimation of the optimal parameters and is obtained with the classical assumption: linearity in parameters in the neighbourhood of the true optimal value.

## III. APPLICATION TO A BIOCHEMICAL PROCESS

### III.1. Example presentation

A classical enzymatic reaction is considered, which is a building block for many biochemical processes. The following chemical equations are given:

$$E + S \underset{k_2}{\overset{k_1}{\rightleftharpoons}} C \tag{12}$$

$$C \xrightarrow{k_3} E + P \tag{13}$$

An enzyme E with a substrate S transitorily gives a specific complex enzyme-substrate C before the researched product P. These chemical equations were proposed after the equilibrium between E, S and C is fast compared to the reaction (13. This scheme, called the Michaelis-Menten kinetics, is a simple approach in enzyme catalysis processes (Wong and Whitesides1994). The formation of P directly depends on the complex concentration and these measurements are available. The state variables in the reaction scheme are the concentrations of the enzyme [E], substrate [S] and complex [C]. A classical mathematical description of the problem is then proposed in a batch reactor (Stortelder 1998):

$$\frac{d[S]}{dt} = -k_1[E][S] + k_2[C] \quad (14)$$

$$\frac{d[C]}{dt} = k_1[E][S] - k_2[C] - k_3[C] \quad (15)$$

$$[E] + [S] = [E]_0 + [S]_0 \quad (16)$$

The initial values are $[S]_0$=1.0 mol.l$^{-1}$, $[C]_0$=0.0 mol.l$^{-1}$ and $[E]_0$=1.0 mol.l$^{-1}$. The vector of unknown, positive parameters is $\theta^T=(k_1,k_2,k_3)$.

### III.2. Parametric identification results

The parametric identification is made with the minimization of the least squares criterion by the diploid genetic algorithm described in II.1. In this simple example, maximum likelihood estimation and least squares criterion are equivalent. The obtained parameters are given in table 1:

Table1: compared estimation of the parameters.

| | Genetic algorithm | Stortelder 1998 |
|---|---|---|
| $k_1$ | 0.683 | 0.683 |
| $k_2$ | 0.311 | 0.312 |
| $k_3$ | 0.212 | 0.212 |
| $J(\theta)$ | $4.1\ 10^{-4}$ | $4.1\ 10^{-4}$ |

Stortelder uses a gradient method (Levenberg-Marquardt) to determine the unknown parameters, while we use a genetic algorithm. The results showed in table 1 are similar. An evolutionary algorithm is certainly not the best method to solve this simple reaction scheme but it gives a satisfactory result too. Figure 2 shows result simulation of the complex concentration with the obtained parameters.

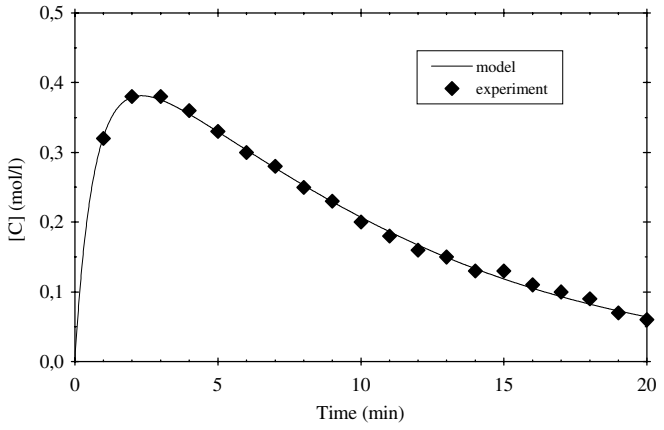


Figure 2: simulation response and experimental data of the complex concentration vs time.

### III.3. Confidence region with Walter-Pronzato's approach

We propose a procedure, based on the diploid genetic algorithm described in II.1, to plot the confidence region of the kinetic parameters. The algorithm is adapted in the evaluation phase. The fitness criterion is $f(\theta)$ defined in equation (10). Elitist selection and genetic operators are used until all the individuals of the population satisfy the statistical test defined in (10). So, the last generation contains a set of points which are in the rigorous confidence region. The more the population has points, the more the domain is well defined. 1000 points are randomly chosen and evolve to obtain the plot of the confidence region in figure 3. A classical value for $\alpha$ is 0.05 which corresponds to a reasonable risk. In this example, $n_\theta$=3 and $n_{exp}$=20.

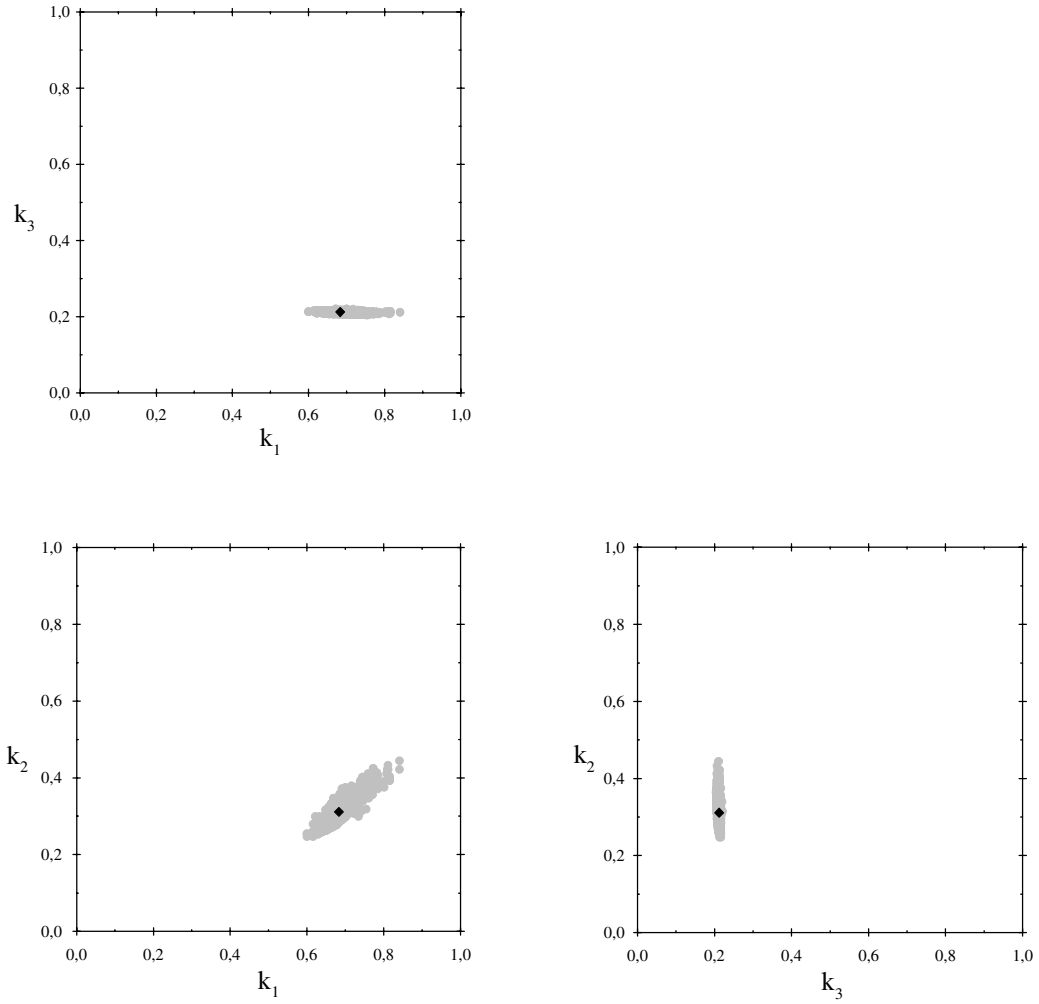


Figure 3: 95 % confidence region of the parameters with Walter-Pronzato's approach.

For a better understanding, projections of one parameter on the two others are presented. Gray points represent the domain and the black diamond the parameters optimal value. The asymetry of the confidence region is noticeable for this nonlinear system. The diploid genetic algorithm gives a set of points, that is to say an approximation of the confidence region but 1000 points define the domain with sufficient precision.

### III.4. Confidence region with linear assumption

In the same way, a procedure is developed, based on the diploid genetic algorithm, to determine the confidence region with the linear assumption. In this case, estimate the true parameters value by the minimization of the least squares criterion is the first step. In the second step, the DGA is used a second time to minimize the function defined in (11) and populations evolve until all the points satisfy the statistical test (11). For comparison, 1000 points are ploted in figure 4 and $\alpha$ is taken equal to 0.05.

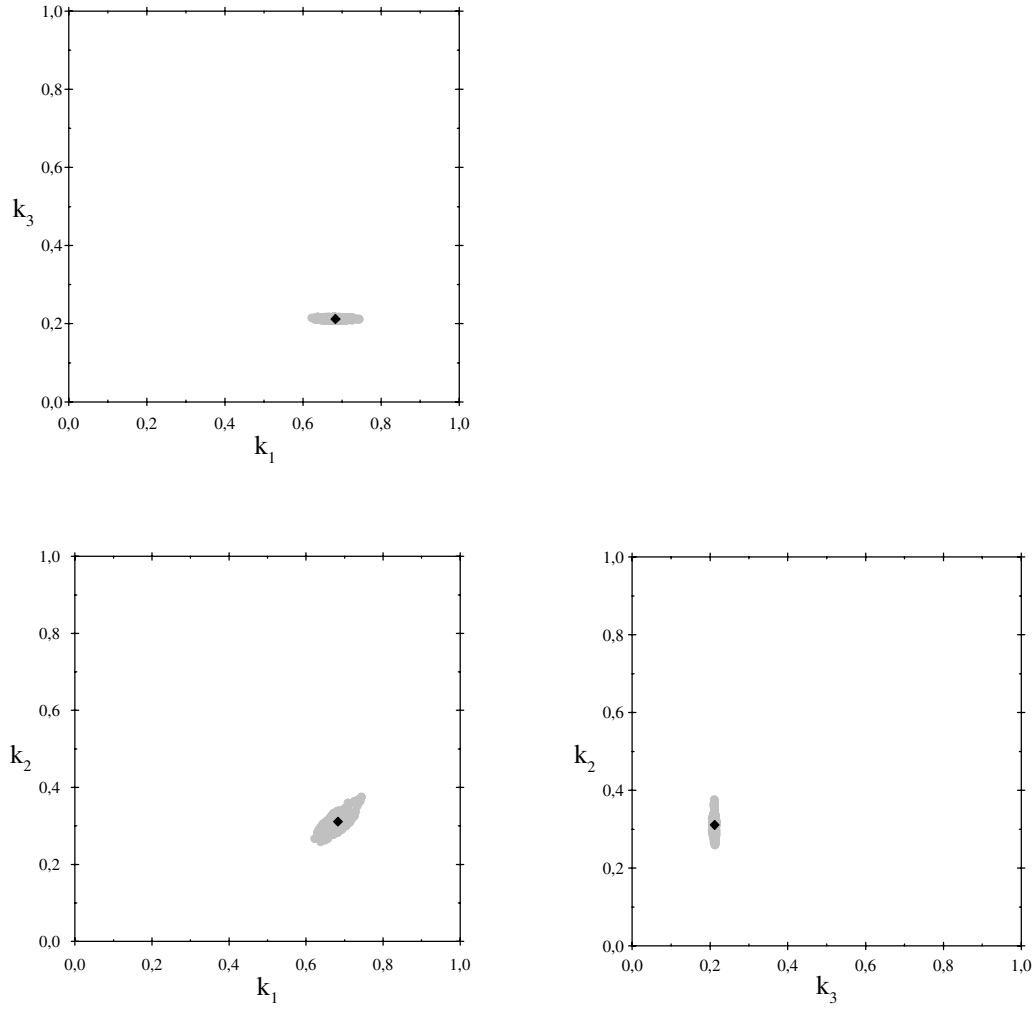


Figure 4: 95 % confidence region of the parameters with linear assumption.

Here, the symetry of the domain is noticeable. Moreover, the confidence region with the linear assumption is smaller than the one with the Walter-Pronzato's approach. So, in this case, the linearization of the model tends to underestimate the real confidence region, but the domain gives a rather good idea of it, compared to the search space. The calculation time to determine the $\Pi$ matrix is rather long so, after hypothesis had been satisfied, the linear assumption could be a good approximation. This last figure corresponds to the one found previously (Stortelder 1998).

## IV. CONCLUDING REMARKS

In this work, the determination of confidence region of parameters in nonlinear and dynamical systems was pointed out. It is a further stage in the parametric identification domain but necessary to choose a correct model. Two approaches are studied and compared to determine the confidence region. The obtention of the rigorous ones is demonstrate and an expression of it with the linear approximation is deducted. The diploid genetic algorithm allows to obtain these confidence regions with a set of points randomly generated initially which evolve and converge to the wanted domain. The linear assumption in the vicinity of the true parameter value is usually used in the litterature but can be far from the real confidence region in the case of a strongly nonlinear model. The linear case can give a first idea of the parameters uncertainty with a reasonable calculation time but cannot be substituted for the real confidence region.

When a confidence region is determined, a characterization of this one has to be made. It must be rather small for the parameter significance and does not show a correlation between the parameters. With the points obtained by the genetic algorithm, the volume can be determined as well as the orthogonal lengths of the confidence region. The parameter correlation can be defined by the ratio of the smallest length to the longest. In figure 3, we can notice a small uncertainty on $k_3$ and two correlated parameters: $k_1$ and $k_2$, which is less obvious in the ellipsoidal region in figure 4. Some techniques could be developed by analogy with the determination of the eighen value of Fischer information matrix. The volume of the domain and the correlation of the parameters could be new criteria to find an optimal experimental design.

The enzymatic reaction is a rather simple nonlinear system in its modelling but allows to show the performances of our technique based on the Walter-Pronzato's approach and on the adaptation of a genetic algorithm to determine the confidence region of the kinetic parameters. The same procedure could be used for more complex systems with an important number of parameters. For example, an emulsion polymerization process will be able to be studied to learn about its modelling.

## V. REFERENCES


Cunha, L. M. (1999). *Experimental design procedures and data analysis strategies for improved estimation of the kinetic parameters of non-linear models used in food research*, PhD Thesis, Porto.

Del Castillo, E. (1996). Multiresponse process optimization via constrained confidence regions. *Journal of Quality Technology* 28, 61-70.

Draper, N. R. and Smith, H. (1981). *Applied regression analysis*, John Wiley & sons, 2nd edition.

Emig, G. and Hosten, L. H. (1974). On the reliability of parameter estimates in a set of simultaneous nonlinear differential equations. *Chem. Eng. Sci.* 29, 475-483.

Fonteix, C., Bicking, F., Perrin, E. and Marc, I. (1995). Haploid and diploid algorithms, a new approach for global optimization: compared performances. *Int. J. of Systems Science* 26, 1919-1933.

Frederiksen, P. S. (1998). Parameter uncertainty and design of optimal experiments for the estimation of elastic constants. *Int. J. Solids Structures* 35, 1241-1260.

Goldberg, D. E. (1989*). Genetic algorithms in search, optimization and machine learning*, Addison-Wesley Publishing Company.

Perrin, E., Mandrille, A., Oumoun, M., Fonteix, C. and Marc, I. (1997). Optimisation globale par stratégie d'évolution : technique utilisant la génétique des individus diploïdes. *Recherche Opérationnelle* 31, 161-201.

Stortelder, W. (1998). *Parameter estimation in nonlinear dynamical system*, PhD Thesis, Amsterdam.

Van Boekel, M. A. J. S. (1996). Statistical aspects of kinetic modeling for food science problems. *Journal of Food Science* 61, 477.

Walter, E. (1987). *Identifiability of parametric models*, Pergamon Press, Oxford.

Walter, E. and Pronzato, L. (1994). *Identification de modèles paramétriques à partir de données expérimentales*, Masson, Paris.

Wong, C. H. and Whitesides, G. M. (1994). *Enzymes in synthetic organic chemistry*, Pergamon.