The Use of a Generalized Fisher Equation for Global Optimization in Chemical Kinetics

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

The use of a generalized Fisher equation for global optimization in chemical kinetics

Alejandro F Villaverde,[*][†] John Ross,[‡] Federico Morán,[¶] Eva Balsa-Canto,[†] and

Julio R Banga[*][†]

Bioprocess Engineering Group, IIM-CSIC, Vigo, Spain, Department of Chemistry, Stanford University, Stanford, CA, USA, and Departamento de Bioquímica y Biología Molecular I, Universidad Complutense Madrid, Madrid, Spain

E-mail: afvillaverde@iim.csic.es; julio@iim.csic.es

---
[*]To whom correspondence should be addressed
[†]CSIC
[‡]Stanford University
[¶]Universidad Complutense

Abstract

A new approach for parameter estimation in chemical kinetics has been recently proposed.[1] It makes use of an optimization criterion based on a Generalized Fisher Equation (GFE). Its utility has been demonstrated with two reaction mechanisms, the chlorite-iodide and Oregonator, which are computationally stiff systems. In this paper the performance of the GFE-based algorithm is compared to that obtained from minimization of the squared distances between the observed and predicted concentrations obtained by solving the corresponding initial value problem (we call this latter approach "traditional" for simplicity). Comparison of the proposed GFE-based optimization method with the "traditional" one has revealed their differences in performance. This difference can be seen as a trade-off between speed (which favors GFE) and accuracy (which favors the traditional method). The chlorite-iodide and Oregonator systems are again chosen as case studies. An identifiability analysis is performed for both of them, followed by an optimal experimental design based on the Fisher Information Matrix (FIM). This allows to identify and overcome most of the previously encountered identifiability issues, improving the estimation accuracy. With the new data, obtained from optimally designed experiments, it is now possible to estimate effectively more parameters than with the previous data. This result, which holds for both GFE-based and traditional methods, stresses the importance of an appropriate experimental design. Finally, a new hybrid method that combines advantages from the GFE and traditional approaches is presented.

1. Introduction

A new optimization approach for estimating rate coefficients in chemical reaction systems has been recently presented.[1] Its novelty lies in the use of a Generalized Fisher Equation (GFE) as the optimization criterion. The GFE was first proposed by Nagylaki[2] as an extension of Fisher's equation for population dynamics. The Fisher equation was originally proposed in 1930[3] and it has become known as the fundamental law of population genetics. Given a set of N different species, or alleles, it relates the time derivative of the average of the intrinsic rates of growth (r) to

its variance:

$$\frac{d\bar{r}}{dt} = \overline{\Delta r^2} > 0 \tag{1}$$

In 1992 Nagylaki[2] generalized this equation for time-varying rates of growth, obtaining the GFE as:

$$\frac{\overline{dr(t)}}{dt} = \overline{\Delta r^2(t)} + \frac{\overline{dr_u(t)}}{dt} \tag{2}$$

where $r_u(t) = (1/x_u(t))(dx_u(t)/dt)$, with $x_u(t)$ being the relative abundance of allele u. Thus, the average of the derivatives of the rates of growth is

$$\frac{\overline{dr_u(t)}}{dt} = \sum_{u=1}^{N} \frac{dr_u(t)}{dt} \frac{x_u(t)}{\sum_{u=1}^{N} x_u(t)} \tag{3}$$

while the average rate of growth is

$$\overline{r(t)} = \sum_{u=1}^{N} r_u(t) \frac{x_u(t)}{\sum_{u=1}^{N} x_u(t)} = \frac{1}{\sum_{u=1}^{N} x_u(t)} \sum_{u=1}^{N} \frac{dx_u(t)}{dt} \tag{4}$$

and the variance of the intrinsic rate of growth is

$$\overline{\Delta r^2(t)} = \sum_{u=1}^{N} \left(r_u(t) - \overline{r(t)} \right)^2 \frac{x_u(t)}{\sum_{u=1}^{N} x_u(t)} \tag{5}$$

Vlad et al[4] suggested that Eq. (2) could be applied in other fields different from population genetics, such as chemical kinetics. In this case, alleles are replaced by chemical species; $x_u(t)$ is the concentration of species u, and $r_u(t)$ has units of [1/time]. We have demonstrated the feasibility of this approach in a previous work,[1] where two practical applications of the GFE were proposed: (i) as an indication of high levels of measurement noise, and (ii) as the optimization criterion in a model calibration procedure. In the second application, the objective function was formulated as follows: let us suppose that we have an analytical description of a reaction mechanism, which is modeled by a set of dynamical equations. These equations express the time derivatives of the

concentrations as a function of: (i) the concentrations $x(t)$, and (ii) the nominal rate coefficients k^*: $dx(t)/dt = f(k^*, x)$. Let us assume that the species concentrations can be measured and their values $x(t)$ are known. If instead of the nominal rate coefficients (k^*) we use an estimation of them (k_e), the estimated time derivatives of the concentrations $(dx^e(t)/dt)$ will differ from the real ones:

$$\frac{dx^e(t)}{dt} = f(k^e, x(t)) \neq \frac{dx(t)}{dt} = f(k^*, x(t)) \tag{6}$$

The mismatch between both sets of derivatives increases with the mismatch between nominal and estimated rate coefficients. With the estimated derivatives, $dx^e(t)/dt$, and the given experimental concentrations, $x(t)$, we can obtain the left-hand side (s_1) and right-hand side (s_2) of the GFE (Eq. (2)), using numerical differentiation in the calculations. If there is a mismatch in the derivatives, there will also be a mismatch between s_1 and s_2. Therefore, if all the terms in the GFE are calculated with the experimental concentrations, $x(t)$, and the analytical derivatives, $dx^e(t)/dt$, there will be a mismatch unless the estimation of the rate coefficients matches the nominal values $(k^e = k^*)$. From the point of view of optimization, this is an inverse problem, similar to most parameter estimation problems arising in dynamic systems modeling. [5] The expected discrepancy, f_{GFE}, can be written as:

$$f_{GFE} = \frac{\int_{t_1}^{t_2} \left| s_1(t)^2 - s_2(t)^2 \right| dt}{\int_{t_1}^{t_2} \left| s_1(t)^2 + s_2(t)^2 \right| dt} \tag{7}$$

This measure can be used as an optimization criterion for the inverse problem. Other similar measures may be chosen without significantly altering the results.

The feasibility of an optimization approach based on Eq. (7) has been demonstrated [1] with two test cases: the chlorite-iodide [6] and the Oregonator [7] reaction systems, which are described in Table 1 and Table 2, respectively.

If the assumed reaction mechanism is incomplete or incorrect, the proposed method will still try to find the best possible fit. For example, given an assumed rate law, the algorithm will return the values of the rate coefficients that enable the closest reproduction of experimental data. The

problem of the possible existence of unmeasured species is not explicitly considered, although it constitutes an interesting topic deserving more research. Information-theoretical methods[8] are a promising tool for addressing it.

For further reading on model calibration, we refer the interested reader to a general introduction to the problem;[9] to a review of the parameter estimation and optimal experimental design steps;[10] and to another recent review[11] that discusses identifiability and parameter estimation aspects for biochemical models.

In the present work we have carried out the following developments:

1. We compare the performance of the proposed GFE-based optimization method to that of a "traditional" approach, where the objective function is the squared difference between predicted and observed values of the concentrations time series.

2. The two systems selected as case studies in the previous work[1] exhibited identifiability issues. As a result, 4 out of 9 parameters in the chlorite-iodide system and 5 out of 9 in the Oregonator system had estimation errors larger than 50% (it should be noted that, although apparently large, these results were consistent with the variability found by experiments reported in the literature). Although this problem was acknowledged,[1] only a basic evaluation of the parameter sensitivities was carried out and no further action was adopted. Here we obtain contour plots for the parameters, which allows us to visualize their multimodal character, thus helping to predict and explain identifiability issues.

3. Once this has been accomplished, the next step is to carry out optimal experimental design (OED). The simulated experiments used in[1] were not very informative; they presented few dynamical features. We use the Fisher Information Matrix[5] (FIM) as a criterion for obtaining a new set of simulated experiments that maximize the information content of the data. We use the new experiments to carry out new model calibrations, overcoming the previous identifiability issues.

4. Finally, we propose a new hybrid optimization method that combines advantages from the

two approaches, GFE-based and "traditional".

2. Methodologies

2.1. Model Calibration

The objective of model calibration or parameter estimation is, given a model structure and a set of experimental data, to determine the unknown rate coefficients (collected in the parameter vector **p**) so as to fit the experimental results in the best possible way. This is performed by minimizing a cost function which measures the goodness of the fit.

Mathematically, the problem is formulated as a non-linear programming problem (NLP) where the decision variables are the unknown parameters, subject to dynamic and algebraic constraints that define the admissible region for their values and the cost function is a certain weighted measure of the distance among the experimental data and the model predictions.[12]

Selection of the cost function depends on the information available for the case under consideration. Here two different cost functions will be used: i) the one based on the GFE,[1] Eq. (7), and ii) the "traditional" weighted least squares function:

$$J(\mathbf{p}) = f_{Trad} = \sum_{i=1}^{nexp} \sum_{j=1}^{n} \sum_{k=1}^{m} Q_{ijk} \left(x_{ijk}(\mathbf{p}) - \tilde{x}_{ijk} \right)^2 \tag{8}$$

where *nexp* regards the number of experiments, n is the number of observables, m the number of sampling times or measurements per observable and experiment, $\tilde{x}_{ijk}$ are the measurement values, $x_{ijk}(\mathbf{p})$ the model corresponding predictions, and Q_{ijk} is a weighting coefficient which may be used to, for instance, incorporate information about the experimental noise when available or scale different contributions. For the illustrative examples considered here, $Q_{ijk} = \frac{1}{max_{i,k}(x_j(k))^2}$ so that all species contribute equally to the objective function value.

It should be remarked that weighted least squares computation requires the solution of the system dynamics in Eq. (6) by means of a suitable initial value problem solver.

As shown in the Results section, the NLPs for the systems selected as case studies have multiple local minima. This means that standard local NLP solvers tend to converge to sub-optimal solutions. Hence, in order to solve the calibration problem in a robust and efficient way, there is a need of global optimization methods, as it has been illustrated by several authors. [14,15]

Global optimization [16,17] methods can be divided into deterministic and stochastic. Deterministic global optimization methods guarantee that the solution is the global optimum, but the computational effort they require can make them unfeasible for large-scale problems. Stochastic global optimization [18] methods, on the other hand, do not guarantee the global optimality of the solution, but they are frequently capable of finding good solutions in reasonable computation times. A particularly efficient class of stochastic global optimization methods are the so-called metaheuristic approaches, which combine mechanisms for exploring the search space and exploiting previously obtained knowledge. In computer science, the term *heuristic* is applied to experience-based techniques that arise from an expert knowledge of the task to be solved. The term *metaheuristic* [19] is applied to more sophisticated techniques, which may be viewed as a heuristic superimposed on another heuristic. Metaheuristic methods are general purpose algorithms that can be applied to different optimization problems. Some of them can be classified as "evolutionary algorithms", because they mimic – to a certain extent – the process of natural evolution; the Genetic Algorithms (GA) are a popular example. Scatter search [20] is one of those evolutionary optimization methods, which arose in the context of integer optimization, but has been adapted to continuous problems in recent years. It is a population-based method, that is, there is always a set of reference points which are chosen according to their quality (objective function value) and diversification (region of the parameter space). Scatter search explores the parameter space in a systematic way, guiding the search to promising areas, and has proved very useful in solving challenging parameter estimation problems. [21] To carry out the model calibration we use an advanced implementation of the scatter search metaheuristic algorithm, called "enhanced Scatter Search" (eSS) [22] (SSmGO Toolbox, release R2009B, November 16, 2009).

2.2. Identifiability Analysis

A model is identifiable if it is theoretically possible to estimate an accurate value of its parameters from a set of observations.[5,9,10] In general, we will distinguish between structural and practical identifiability. Practical identifiability analysis enables the evaluation of the quality of the parameter estimates after model calibration. Possibly the simplest approach to perform such analysis is to draw contours of the model calibration cost function by pairs of parameters for the given set of available experimental data.

Since the Hessian of the objective function $J(\mathbf{p})$ is usually positive-definite in the vicinity of the minimum, the associated confidence region may then be approximated by a hyper-ellipsoid of dimension equal to the number of parameters. Cost contours will be then projections of the hyper-ellipsoid into a two-dimensional space. Many methods can be employed to draw cost contours in a two-dimensional space. Basically they consist of evaluating the cost function in a grid of parameter values in the vicinity of the optimum.

Visual inspection of the resulting plots helps to detect rapidly typical practical identifiability problems, such as strong correlation between parameters, the lack of identifiability for some parameters or the presence of sub-optimal solutions.

Parameters are said to be correlated when cost contours correspond to eccentric ellipses and will be uncorrelated for the case of circular cost contours. Lack of indentifiability arises when cost contours tend to infinity. Multimodality appears when cost contours present multiple valleys.

2.3. Optimal Experimental Design (OED)

Poor practical identifiability has to do with the type of experimental scheme being used and the quality of the corresponding experimental data in terms of experimental noise. The purpose of optimal experimental design is to devise the necessary dynamic experiments in such a way that the parameters are estimated from the resulting experimental data with the best possible statistical quality, which is usually a measure of the accuracy and/or decorrelation of the estimated parameters. In this way, the model and a close-to-optimal solution for the parameters are being used to

design new more informative experiments which in general will result in better practical identifia-

bility properties. [23]

In the context of the present work, the purpose of OED is to generate dynamical features in the evolution of the species concentrations that lead to recognition of the rate coefficients. This is possible if the set of experimental data is rich enough, that is, it contains sufficient information.

The information content of a given experiment can be measured by means of the Fisher Infor-mation Matrix[5] (FIM), which is defined as:

$$\text{FIM} = \underset{\mathbf{Y}_m|\mu}{E} \left\{ \left[\frac{\partial J(\mathbf{p})}{\partial \mathbf{p}} \right] \left[\frac{\partial J(\mathbf{p})}{\partial \mathbf{p}} \right]^T \right\} \tag{9}$$

where E represents the expectation for a given value of the parameters μ presumably close to the optimal solution $\mathbf{p}^*$.

Mathematically the OED problem can be formulated as a dynamic optimization problem where the objective is to find a set of inputs, usually time-varying variables, for example, input flow con-centrations, together with initial conditions or experiment durations, so as to maximize or minimize a cost function which is a scalar measure of the FIM.

There are several criteria that can be used as cost functions, differing in the information they provide about the FIM. Since the FIM is related to the confidence hyper-ellipsoid for the parame-ters, the different criteria provide different informations about the hyper-ellipsoids shape and size. For example, the D-criterion seeks to maximize the determinant of the FIM. This, in turn, mini-mizes the confidence ellipsoids volume, and therefore the geometric mean of the parameters error. Another possible choice is the E-criterion, which maximizes the minimum eigenvalue of the FIM. This minimizes the length of the largest axis, or equivalently, the largest parameter error. The modified E-criterion, on the other hand, minimizes the condition number of the FIM, that is, the ratio of the largest to the smallest eigenvalue. Since the eigenvalues measure the axes length, this criterion seeks to make the ellipsoids as spherical as possible, thus distributing the parameter errors equally. There are still other possible criteria. In the present work we have chosen to minimize the largest error, which entails using the E-criterion. Its mathematical definition is:

$$E = \max \lambda_{min} (\text{FIM}) \qquad (10)$$

where λ_{min} is the minimum eigenvalue.

2.4. Hybrid Optimization Method

The comparison between GFE-based and traditional optimization can be seen as a trade-off between speed (which favors GFE) and accuracy (which favors the traditional method). This suggests the combined use of both methods in order to profit from both strengths. With this aim, we design a two-phase hybrid optimization procedure as follows:

- In a first stage, the GFE-based objective function is used. The maximum computation time of this stage, τ_1, must be set to a small value, but sufficiently large to allow for several iterations of the optimization algorithm. This phase allows for a fast obtaining of a good solution, which will then be refined in the second phase.

- The solutions obtained in the first phase are passed as an initial point to the second stage. Here the traditional objective function is used, and the maximum computation time of this stage is set as $\tau_2 > \tau_1$.

The combined method provides a good initial guess to the traditional optimization procedure, which can then reach an accurate solution in less time.

3. Results and Discussion

3.1. Model Calibration

We have calibrated the models using the same simulated experiments as in the previous work.[1] The external stimuli consist of initial concentrations and, for the chlorite-iodide reaction system, an additional tank parameter. Their values are kept constant, as detailed in the captions of Table 1 and

Table 2. The number of data points used in the calculations is 10^5. The E-criterion value of these experiments (Eq. (10)) is $E_{opt} = 3.2209 \cdot 10^7$ for the chlorite-iodide system, and $E_{opt} = 2.1447 \cdot 10^2$ for the Oregonator system.

The calculations were carried out in a computer with an Intel Xeon Quad-core processor (2.50 GHz, 3 GB/core) in a Matlab environment. First, we test the ability of the GFE-based and traditional methods to yield good predictions in short computation times. Hence, we limit the computation time to 15 minutes; this value was chosen because it allowed for the completion of several iterations of the enhanced scatter search algorithm, for both the chlorite-iodide and Oregonator systems. Due to the metaheuristic (and hence stochastic) nature of the search algorithm, the results are in some sense random. Thus, in order to ensure a fair comparison, we run the algorithm ten times for each method and compare the best results from both methods. In each run, the algorithm's pseudorandom number generator is initialized with a different number, or "seed". Results are shown in Table 3 for the chlorite-iodide, and in Table 4 for the Oregonator. For both systems it is concluded that, if the computation time is restricted to a few minutes, the GFE-based technique gives better estimates than the traditional one.

Then we repeat the calculations, letting the algorithm run for a long time. It should be noted that the random fluctuation in the algorithm's performance decreases with time. That is, for small computation times, different random seeds can give rise to relatively large variations in performance; however, for long time intervals these variations become small. We have tested this for the chlorite-iodide system, launching 10 different realizations of the algorithm with different seeds. With the GFE-based optimization algorithm, after 48 hours all of the 10 programs return an identical value of the objective function, $3.9 \cdot 10^{-8}$. With the traditional optimization algorithm, all of the programs yield very small values of the objective function, between 10^{-18} and 10^{-21}. All of the programs yield qualitatively similar parameter values, and the estimation does not improve significantly if the algorithm is allowed to run for a longer time. Hence, for comparing the performances in the long run we restrict the computation time to 48 hours and perform only one optimization per method instead of ten. The results after 48 hours are also in Table 3 and Table 4. They show

that the initial advantage of the GFE-based technique is lost if the algorithms run for a long time period.

In order to further compare the algorithms' speed we plot the evolution of the objective function value versus time. This is shown on the convergence curves of Figure 1 for the chlorite-iodide system and Figure 2 for the Oregonator system. Since each algorithm uses a different objective function, f_{GFE} and f_{Trad}, it would not make sense to compare their evolutions directly. Hence in the first two subfigures of Figure 1 and Figure 2 we take the best solutions found by the two algorithms over time, and evaluate f_{Trad} for both of them. That is, the curves labeled 'GFE' show the f_{Trad} values of the points obtained with the GFE-based optimization as a function of the computation time. In this way we can determine which algorithm has found the best solution at a given instant, according to a single criterion. At first sight, the reason why the GFE-based optimization curves are not always decreasing in these plots may not be obvious. This is because they do not show the f_{GFE} values found by the optimization algorithm, which are shown in the third subfigures and which, as expected, always decrease in time. Instead, they plot the value of f_{Trad}, and it should be noted that, given two parameter vectors p_A and p_B, the fact that $f_{GFE}(p_A) < f_{GFE}(p_B)$ does not necessarily entail that $f_{Trad}(p_A) < f_{Trad}(p_B)$. Hence, even though the GFE-based algorithm is always minimizing f_{GFE}, it is not always minimizing f_{Trad}.

The results summarized in Table 3, Table 4, Figure 1, and Figure 2 show that the GFE-based optimization procedure is faster than the traditional one. This is particularly important for stiff systems, such as the chlorite-iodide and Oregonator. These are computationally demanding problems; their integration requires a very small step size, otherwise it can become numerically unstable. Chemical reaction systems frequently belong to this class.

On the other hand, the traditional optimization procedure is capable of obtaining more accurate predictions in the long run. Therefore, there is a trade-off between speed and accuracy. This can be summarized as:

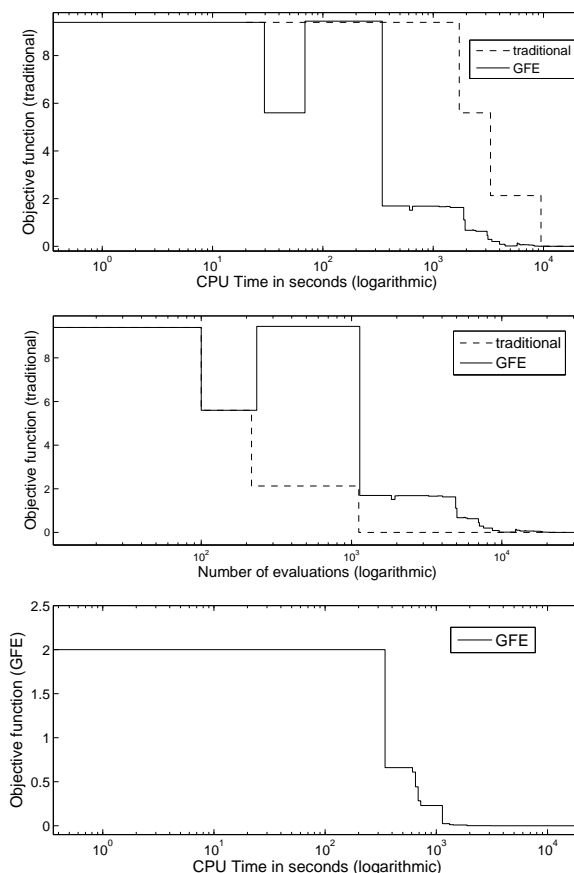- GFE-based: less accurate, faster algorithm.

Figure 1: **Convergence Curves, chlorite-iodide.** Top plot: Convergence curves if the traditional objective function in Eq. (8) is evaluated at each point found by both the GFE-based and traditional optimization procedures. Medium plot: conv. curves plotted as a function of the number of evaluations instead of time. Bottom plot: conv. curve of the GFE-based optimization algorithm when the GFE objective function value is plotted.

- Traditional: more accurate, slower algorithm.

We have already remarked that the original experiments used a large number of data points (10^5), which amounts to taking measurements with a sampling period of 10^{-2} seconds for the chlorite-iodide system and 10^{-3} seconds for the Oregonator system. This is not a problem for simulated data, but is hardly feasible for an experimental setup. Therefore we want to limit ourselves to a smaller, more realistic set of measurements. Thus, we reduce the number of available "measurements" (or simulated values) to 1000 data points for each system (one per second for the chlorite-iodide; 10 per second for the Oregonator). We repeat the optimizations allowing them to run for 15 minutes and 48 hours, and show the results in Table 3 for the chlorite-iodide and Table 4
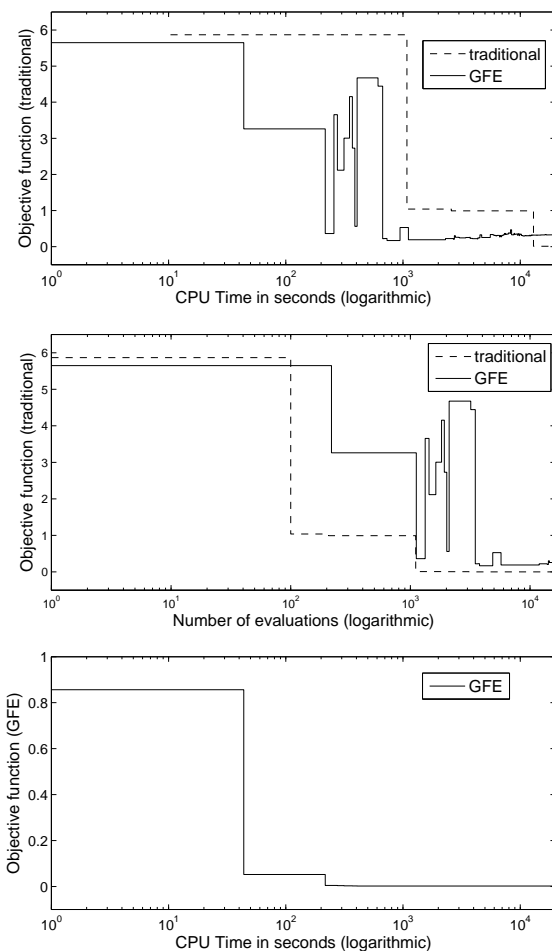
Figure 2: **Convergence Curves, Oregonator.** Top plot: Convergence curves if the 'traditional' objective function in Eq. (8) is evaluated at each point found by both the GFE-based and traditional optimization procedures. Medium plot: curves plotted as a function of the number of evaluations instead of time. Bottom plot: conv. curve of the GFE-based optimization algorithm when the GFE objective function value is plotted.

for the Oregonator. From them it can be concluded that the performance of the GFE-based method

is not significantly degraded when used with a realistic number of measurements. This is due to

the fact that, by reducing the number of data points, the time required for calculating the objective

function value of a vector is reduced; hence, more candidate solutions can be evaluated in less

time.

In the results presented until now there are a number of parameters that fail to be accurately

estimated: for the chlorite-iodide, k_7 and, if the computation time is restricted to 15 minutes, also

k_8, k_9 and k_{10}. For the Oregonator, all of the parameters are eventually estimated in the long term

(48 hours); however, for short computation times only poor estimations are obtained for k_{-4} and

k_{-5}. This fact was already reported by Ross et al,[1] although no solution was provided. In the

following sections, we take a closer look at the identifiability issues, re-design the experiments in

order to overcome them, and finally re-calibrate the models using the new data.

3.2. Identifiability Analysis

Figure 3 shows the cost contour plots corresponding to some pairs of parameters of the chlorite-

iodide reaction system, while Figure 4 does the same for the Oregonator system. They have been

obtained with the nominal values and the same simulated experiments as those originally used for

model calibration.[1] Each 2D plot shows how the (traditional) objective function value is affected

by changes in 2 of the model parameters (rate coefficients, k_i). The 7 remaining parameters are

kept constant and equal to their nominal value. As the objective function value increases, the color

changes from blue to yellow to red.

Both examples show poor identifiability. As mentioned in section 2.2, lack of indentifiability

arises when cost contours tend to infinity. For instance, Figure 3 clearly reveals lack of identifia-

bility of some parameters such as k_7 (Figure 3.D, vertical axis) or k_{10} (Figure 3.A, vertical axis)

with cost contours tending to infinity in their direction. On the contrary k_3 (Figure 3.C, horizon-

tal axis) or k_9 (Figure 3.E, vertical axis) seem to be clearly identifiable since any change in their

estimations produces a change in the objective function value. On the other hand, Figure 4.B or

Figure 4.C show noisy cost contours that present multiple valleys, revealing difficulties to find the

global optimum for some parameters.

To perform the identifiability analysis the MATLAB based toolbox AMIGO[13] (Advanced

model identification using global optimization, www.iim.csic.es/~amigo) was used in this
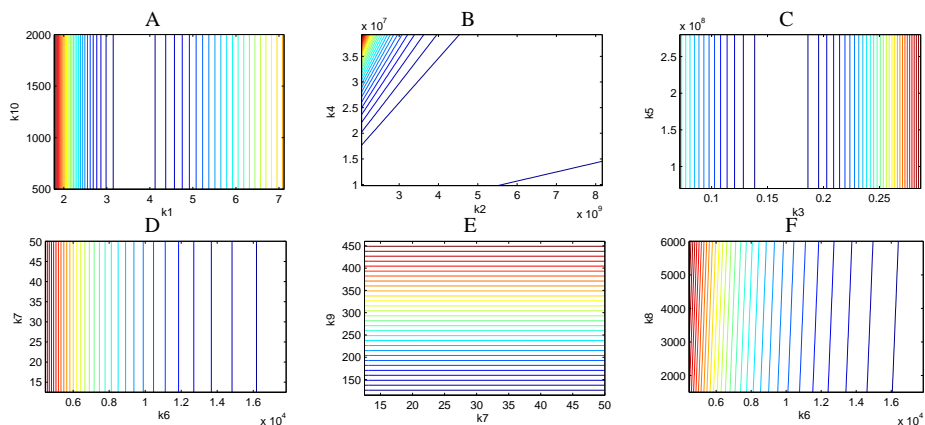
work.

Figure 3: **Contour plots for some pairs of parameters from the chlorite-iodide system.** Each 2D plot shows how the (traditional) objective function value is affected by changes in 2 of the model parameters (rate coefficients, k_i). The 7 remaining parameters are kept constant and equal to their nominal value.

3.3. Optimal Experimental Design (OED)

The AMIGO[13] toolbox solves the optimal experimental design problem by means of the combination of the control vector parameterization approach and global optimization methods, allowing for the design of sequential and parallel experiments within the experimental constraints and several FIM related cost functions.

The OED specifications are summarized in Table 5. To measure the information content we have chosen the E-criterion (Eq. (10)) which maximizes the minimum eigenvalue (λ) of the FIM, thus minimizing the length of the largest axis of the confidence ellipsoids. In this way, we seek to minimize the largest of the parameter errors, maximizing the distance from the non identifiable case.

Beginning with the chlorite-iodide system, we perform optimal experimental design in order to maximize the information content of the simulated experiments. Our aim is to obtain better artificial data, thus allowing the subsequent optimization procedures to estimate the parameters more accurately. Additionally, we require that the new simulated experiments are feasible, that is, they can be realistically expected to be carried out in a laboratory. Hence we restrict the number of data points (or 'simulated measurements') to be obtained with a sampling time of 1 second, for a total duration of 1000 seconds. Furthermore, only 3 external stimuli (control variables) can be
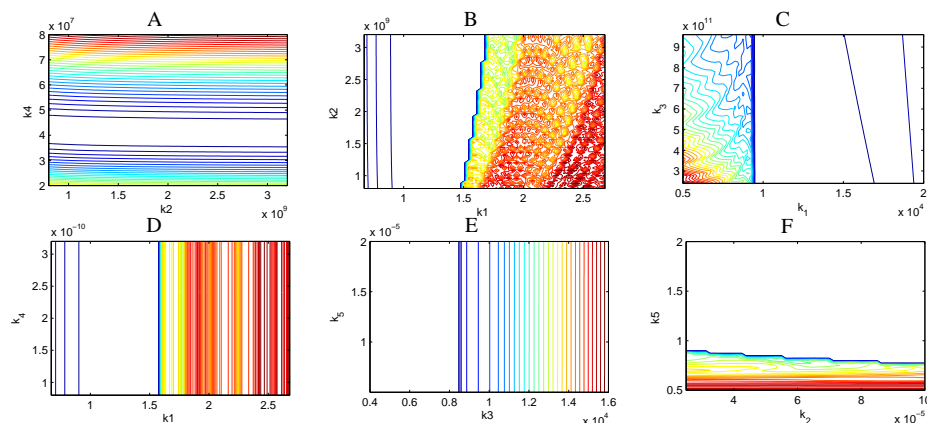
Figure 4: **Contour plots for some pairs of parameters from the Oregonator system.** Each 2D plot shows how the (traditional) objective function value is affected by changes in 2 of the model parameters (rate coefficients, k_i). The 7 remaining parameters are kept constant and equal to their nominal value.

changed during the experiment: the input flow concentrations $[ClO_2]_0$, $[I]_0$ and $[k_0]$. These stimuli are allowed to have up to 3 different values during the experiments (3 steps). In contrast, in [1] the values of the stimuli were kept constant throughout the experiment. This additional variation, shown in Figure 6, allows for a richer dynamical behavior of the species. The achieved value of the E-optimality criterion is $E_{opt} = 9.5586 \cdot 10^{10}$, which is an improvement with respect to the original experiments [1] (which, for the same number of data points, yielded $E_{opt} = 3.2209 \cdot 10^7$).
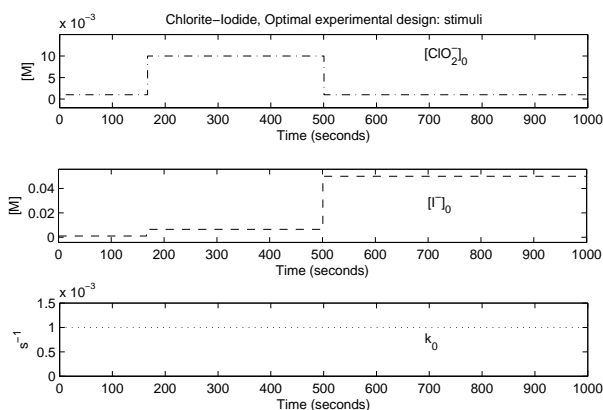

Figure 5: **Optimal Experimental Design, chlorite-iodide.** Time evolution of the external stimuli that maximize the information content of the experiments.
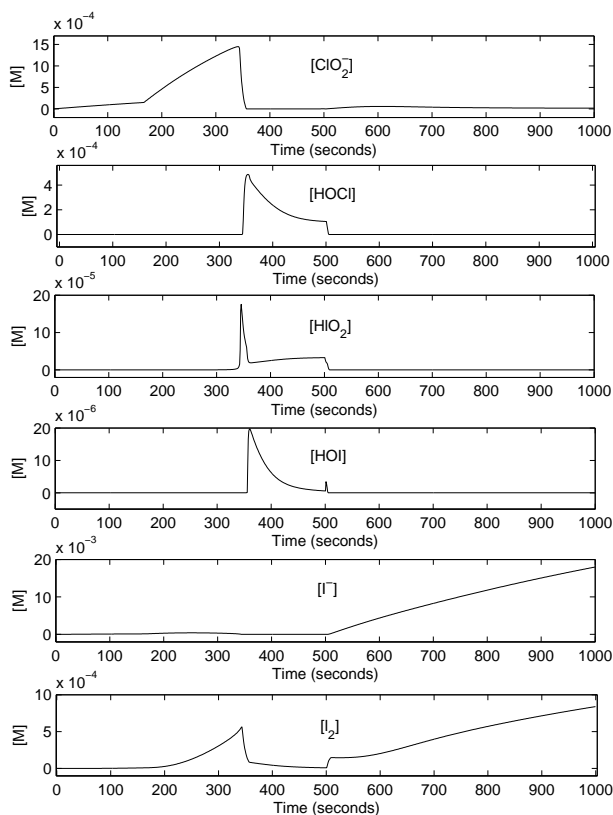
Figure 6: **Optimal Experimental Design, chlorite-iodide.** Time evolution of the species concentrations that result from the external stimuli shown in Figure 5.

The same procedure is carried out for the Oregonator system. Here the total duration of the experiment is set to 100 seconds, and a sampling time of 1 second is initially chosen. The only external stimulus or control variable allowed to change during the experiment, in three different steps, is $[BrO_3^-]$, as shown in Figure 8. Another system parameter is fixed: $[HOBr] = 0.01$ [M]. Figure 8 shows the resulting experimental profiles. The achieved value of the E-optimality criterion is $E_{opt} = 1.0933 \cdot 10^5$, improving what was obtained with the old experiments ($E_{opt} = 2.1447 \cdot 10^2$).

3.4. Calibrations after Optimal Experimental Design

We now re-calibrate the models, using the simulated experiments designed in the previous section.

Table 6 shows the results for the chlorite-iodide system. As happened with the old experiments, the GFE-based procedure is faster than the traditional one; hence it yields better results for reduced
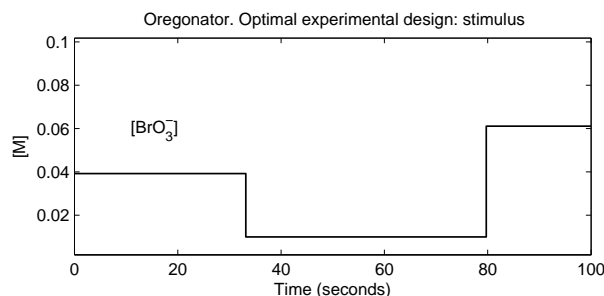
Figure 7: **Optimal Experimental Design, Oregonator.** Time evolution of the external stimulus, $[BrO_3^-]$, that maximizes the information content of the experiments.

computational time (15 minutes). The newly devised experiments improve the system's identifiability: it is now possible to obtain reasonably accurate estimates[1] of *all* of the parameters in *less than 15 minutes*. Compare to the previous experiments (Table 3), from which only 5 parameters could be estimated in the same time. If the algorithm is allowed to run for 48 hours, the new experimental design also outperforms the old one. It should also be noticed that the GFE-based procedure hardly improves the solution after the first few minutes. The new experiments also improve the performance of the traditional method: it provides estimations of *all* of the parameters with less than $< 1\%$ error, while with the previous experiments only 7 of the 9 parameters could be estimated.

Table 7 shows similar results for the Oregonator system[2]. Again, the new experimental design improves performance, especially when the computation time is restricted to 15 minutes (columns labeled as "B" in Table 4). As with the chlorite-iodide, the GFE-based procedure yields better results in the short term, while the traditional procedure produces better results in the long term.

3.5. Hybrid Optimization Method

The performance improvement of the hybrid two-phase optimization method is demonstrated for the chlorite-iodide system. We set $\tau_1 = 1$ minute and $\tau_2 = 59$ minutes, and compare the results with

---

[1]6 parameters are retrieved with an error $< 1\%$, 1 with an error $< 5\%$, and 2 with errors between 16% and 33%

[2]Despite obtaining the experimental design with a sampling period of $\Delta T = 1$ seconds, for optimization purposes we decrease it to 0.1 seconds. Thus, more artificial data are generated and the optimization performance can be improved. This was done because the $\Delta T = 1$ setting was found to be, in practice, very restrictive for the Oregonator.
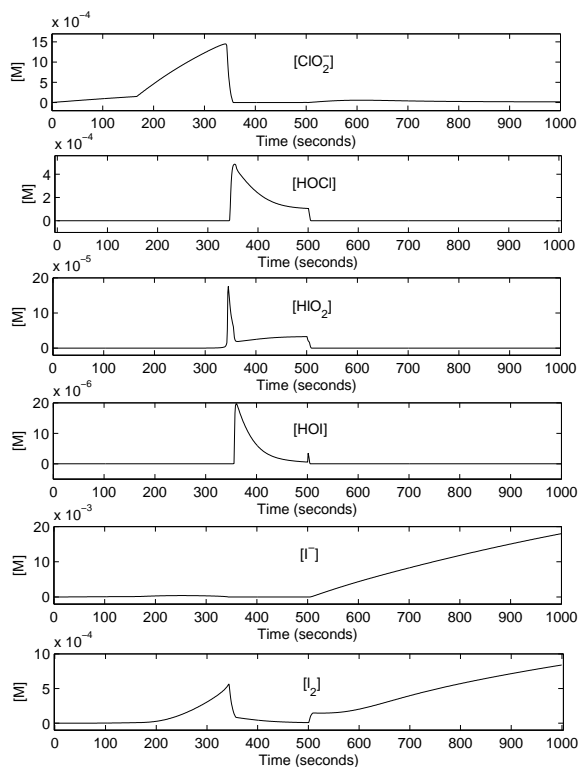
Figure 8: **Optimal Experimental Design, Oregonator.** Time evolution of the species concentrations that result from the external stimuli shown in Figure 7.

the two other methods considered in this paper: the GFE-based and the traditional optimization procedures. Results are shown in Figure 9 and Table 8. During the first minute, the hybrid method uses the GFE-based criterion and thus achieves a faster improvement than the traditional method. Then, it switches from the GFE-based to the traditional criterion. Since a good parameter vector has already been found, when the hybrid method starts using the traditional criterion it has a better starting point than the purely traditional method. Hence, in the subsequent computations the hybrid method is able to find a more refined solution than the traditional one. The GFE-based method, on the other hand, is less accurate than the traditional, and as a result it does not improve significantly its solution after the first minutes. Therefore the hybrid method outperforms the other two.
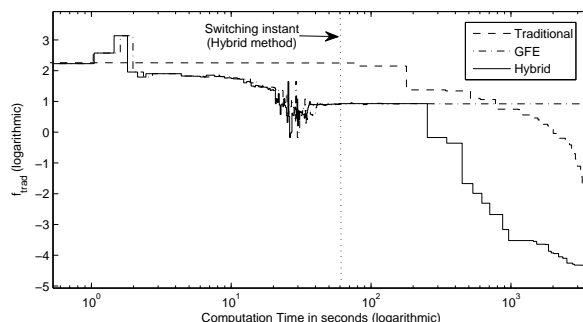
20

Figure 9: **Convergence Curves, Hybrid Optimization.** The multi-stage method is compared to the GFE-based and traditional methods for the chlorite-iodide. In the three cases, the value of the traditional objective function is plotted.

4. Conclusion

In a previous work[1] we have presented an optimization criterion based on a generalized Fisher equation. Here we have evaluated its performance by comparing it with the traditional criterion of weighted least squares, concluding that their differences may be expressed as a trade-off: whereas the GFE method is computationally faster, the traditional method is more accurate. Hence, we have presented an hybrid optimization method that exploits the strengths of both approaches.

Additionally, we have demonstrated the utility of integrating identifiability analysis with optimal experimental design. By applying these techniques to the reaction mechanisms used as case studies in the previous work[1] we have been able to design new simulated experiments that provide more informative data. The previous identifiability issues, caused by a lack of dynamical features in the experimental data, have been overcome. As a result, we are now able to obtain more accurate estimations of the kinetic parameters.

All the tools used to perform this study have been published elsewhere[13,22] and the software packages are freely available for academic use www.iim.csic.es/~gingproc/software.html.

The problem of deducing information from experiments is ever present. The methods discussed here are a step forward in that direction.

Acknowledgement

References

(1) Ross, J.; Fernández Villaverde, A.; Banga, J. R.; Vázquez, S.; Morán, F. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, 107:12777-12781.

(2) Nagylaki, T. *Introduction to Theoretical Population Genetics*; Springer: Berlin, Germany, pp 5-27, 1992.

(3) Fisher, R. A. *The Genetical Theory of Natural Selection*; Clarendon: Oxford, 1930; reprinted 1999, Oxford Univ. Press, Oxford, pp 22-47.

(4) Vlad, M. O.; Szedlacsek, S. E.; Pourmand, N.; Cavalli-Sforza, L. L.; Oefner, P.; Ross, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, 102:9848-9853.

(5) Walter, E.; Pronzato, L. *Identification of Parametric Models from Experimental Data*, Springer: Berlin, Germany, 1997.

(6) Citri, O.; Epstein, R.; *J. Phys. Chem.* **1987**, 91:6034-6040.

(7) Field, R. J. *J. Chem. Phys.* **1975**, 63:2289-2296.

(8) Sayyed-Ahmad, A.; Tuncay, K.; Ortoleva, P.J. *J. Phys. Chem. A* **2003**, 107:10554-10565.

(9) Jaqaman, K.; Danuser, G. *Nat. Rev. Mol. Cell Biol.* **2006**, 7(11):813–819.

(10) Banga, J. R.; Balsa-Canto, E. *Essays Biochem.* **2008**, 45:195.

(11) Ashyraliyev, M.; Fomekong-Nanfack, Y.; Kaandorp, J.A.; Blom, J.G. *FEBS Journal* **2009**, 276:886-902.

(12) Schittkowski, K. *Numerical data fitting in dynamical systems: a practical introduction with applications and software*, Springer: Netherlands, 2002.

(13) Balsa-Canto, E.; Alonso, A. A.; Banga, J. R. *BMC Syst. Biol.* **2010**, 4:11.

(14) Esposito, W. R.; Floudas, C. A. *Ind. & Eng. Chem. Res.* **2000**, 39:1291–1310.

(15) Moles, C. G.; Mendes, P.; Banga, J. R. *Genome Res.* **2003**, 13:2467.

(16) Torn, A.; Žilinskas, A. *Global Optimization*, Springer, New York, 1989.

(17) Horst, R.; Pardalos, P. M. *Handbook of global optimization* Kluwer Academic Pub, 1995.

(18) Zhigljavsky, A.; Žilinskas, A. *Stochastic global optimization*, Springer Verlag, 2008.

(19) Glover, F.; Kochenberger, G. A. *Handbook of metaheuristics* Springer, 2003.

(20) Glover, F.; Laguna, M.; Martí, R. *Control and Cybernetics* **2000** 39:653–684.

(21) Rodriguez-Fernandez, M.; Egea, J. A.; Banga, J. R. *BMC Bioinformatics* **2006**, 7:483.

(22) Egea, J. A.; Martí, R.; Banga, J. R. *Comp. & Oper. Res.* **2010**, 37: 315-324.

(23) Balsa-Canto, E.; Alonso, A. A.; Banga, J. R. *IET Systems Biology* **2008**, 2:163–172.

Table 1: Model reactions and nominal rate coefficients of the Citri-Epstein mechanism for the chlorite-iodide reaction.[6] The model's dynamical variables are the following concentrations: x_1 = [Cl(III)], which is a common variable for [$HClO_2$] and [ClO_2^-], x_2 = [HOCl], x_3 = [HIO_2], x_4 = [HOI], x_5 = [I^-], x_6 = [I_2]. The CSTR condition is modeled by adding to each differential equation the term: $k_0 \cdot ((x_i)_0 - x_i)$. The control inputs are k_0 and the six $(x_i)_0$, which were assigned the following values in the original simulated experiments: $k_0 = 10^{-2}\,[s^{-1}]$, $(x_1)_0 = 3.187 \cdot 10^{-3}\,[M]$, $(x_5)_0 = 8.563 \cdot 10^{-3}\,[M]$, $(x_2)_0 = (x_3)_0 = (x_4)_0 = (x_6)_0 = 0$. The model's dynamical equations can be thus deduced directly from this information.

Reaction	Nominal Rate Coefficients
$H^+ + Cl(III) + I^- \rightarrow HOCl + HOI$	$k_1 = 3.5586\,[M^{-1}s^{-1}]$
$H^+ + HOI + I^- \rightarrow I_2 + H_2O$	$k_2 = 4.0836 \cdot 10^9\,[M^{-1}s^{-1}]$
$I_2 + H_2O \rightarrow H^+ + HOI + I^-$	$k_3 = 0.1434\,[s^{-1}]$
$HClO_2 + HOI \rightarrow HOCl + HIO_2$	$k_4 = 1.9620 \cdot 10^7\,[M^{-1}s^{-1}]$
$HOCl + I^- \rightarrow HOI + Cl^-$	$k_5 = 1.4 \cdot 10^8\,[M^{-1}s^{-1}]$
$H^+ + HIO_2 + I^- \rightarrow 2\,HOI$	$k_6 = 8.9125 \cdot 10^3\,[M]$
$2\,HOI \rightarrow H^+ + HIO_2 + I^-$	$k_7 = 25\,[M^{-1}s^{-1}]$
$2\,HIO_2 \rightarrow HOI + IO_3^- + H^+$	$k_8 = 3 \cdot 10^3\,[M^{-1}s^{-1}]$
$HIO_2 + HOI \rightarrow I^- + IO_3^- + 2\,H^+$	$k_9 = 230\,[M^{-1}s^{-1}]$
$HOCl + HIO_2 \rightarrow Cl^- + IO_3^- + 2\,H^+$	$k_{10} = 10^3\,[M^{-1}s^{-1}]$

Table 2: Model reactions and nominal rate coefficients of the Oregonator reaction.[7] The model's dynamical variables are the following concentrations: x_1 = [$HBrO_2$], x_2 = [Br^-], and x_3 = [2 Ce(IV)]. Other concentrations ([HOBr] and [BrO_3^-]) are fixed, playing the role of control inputs. In the original simulated experiments the following values were assigned: [HOBr] = 0.01[M], [BrO_3^-] = 0.06[M]. The model's dynamical equations[7] can be thus deduced directly from this information.

Reaction	Nominal Rate Coefficients
$BrO_3^- + Br^- \leftrightarrow HBrO_2 + HOBr$	$k_1 = 1.34\,[M^{-2}s^{-1}]$
	$k_{-1} = 10^4\,[M^{-1}s^{-1}]$
$HBrO_2 + Br^- \leftrightarrow 2\,HOBr$	$k_2 = 1.6 \cdot 10^9\,[M^{-2}s^{-1}]$
	$k_{-2} = 5 \cdot 10^{-5}\,[M^{-1}s^{-1}]$
$BrO_3^- + HBrO_2 \leftrightarrow 2\,HBrO_2 + 2\,Ce(IV)$	$k_3 = 8 \cdot 10^3\,[M^{-1}s^{-1}]$
	$k_{-3} = 4.8 \cdot 10^{11}\,[M^{-2}s^{-1}]$
$2\,HBrO_2 \leftrightarrow BrO_3^- + HOBr$	$k_4 = 4 \cdot 10^7\,[M^{-1}s^{-1}]$
	$k_{-4} = 1.6 \cdot 10^{-10}\,[M^{-2}s^{-1}]$
$2\,Ce(IV) \leftrightarrow Br^-$	$k_5 = 1\,[M^{-1}s^{-1}]$
	$k_{-5} = 10^{-5}\,[M^{-2}s^{-1}]$

Table 3: Comparison of GFE-based and traditional parameter estimation, chlorite-iodide. The estimated parameters are rate coefficients (k_i). The table entries are the percentage errors. f_{GFE} is the value of the GFE-based objective function, Eq. (7), while f_{Trad} is the value of the traditional objective function, Eq. (8). Two different time limits (t_{max}) and sampling times (ΔT) were used. Columns A: $t_{max} = 15$ minutes, $\Delta T = 10^{-2}$ seconds. Columns B: $t_{max} = 15$ minutes, $\Delta T = 1$ seconds. Columns C: $t_{max} = 48$ hours, $\Delta T = 10^{-2}$ seconds. Columns D: $t_{max} = 48$ hours, $\Delta T = 1$ seconds.

	GFE				Traditional			
	A	B	C	D	A	B	C	D
k_2	0.0%	0.0%	0.0%	0.0%	4.6%	7.1%	0.0%	0.0%
k_3	0.0%	0.0%	0.0%	0.0%	20.6%	30.9%	0.0%	0.0%
k_4	0.4%	0.0%	0.4%	0.0%	205.6%	199.0%	0.0%	0.0%
k_5	0.0%	0.0%	0.0%	0.0%	242.0%	538.1%	0.0%	0.0%
k_6	0.3%	0.5%	0.3%	0.5%	97.4%	102.3%	0.0%	0.0%
k_7	155.5%	163.2%	37.5%	296.5%	85.4%	48.7%	257.2%	100.3%
k_8	133.9%	233.3%	233.3%	233.3%	219.2%	127.6%	0.0%	0.0%
k_9	78.6%	84.3%	95.7%	86.3%	49.1%	291.2%	2.5%	46.4%
k_{10}	1123.6%	64.5%	70.0%	69.7%	37.6%	240.7%	0.4%	8.0%
f_{GFE}	$3.9 \cdot 10^{-8}$	$3.9 \cdot 10^{-8}$	$3.9 \cdot 10^{-8}$	$3.9 \cdot 10^{-8}$	–	–	–	–
f_{Trad}	$7.4 \cdot 10^{-2}$	$8.6 \cdot 10^{-2}$	$7.5 \cdot 10^{-2}$	$8.6 \cdot 10^{-2}$	$5.8 \cdot 10^{4}$	$3.8 \cdot 10^{5}$	$6.1 \cdot 10^{-21}$	$1.4 \cdot 10^{-18}$

Table 4: Comparison of GFE-based and traditional parameter estimation, Oregonator. The estimated parameters are rate coefficients (k_i). The table entries are the percentage errors. f_{GFE} is the value of the GFE-based objective function, Eq. (7), while f_{Trad} is the value of the traditional objective function, Eq. (8). Two different time limits (t_{max}) and sampling times (ΔT) were used. Columns A: $t_{max} = 15$ minutes, $\Delta T = 10^{-3}$ seconds. Columns B: $t_{max} = 15$ minutes, $\Delta T = 0.1$ seconds. Columns C: $t_{max} = 48$ hours, $\Delta T = 10^{-3}$ seconds. Columns D: $t_{max} = 48$ hours, $\Delta T = 0.1$ seconds.

	GFE				Traditional			
	A	B	C	D	A	B	C	D
k_{-1}	13.50%	61.0%	12.7%	18.9%	314.4%	486.1%	0.0%	0.0%
k_2	6.13%	0.4%	6.4%	0.2%	444.6%	513.8%	0.0%	0.0%
k_{-2}	78.76%	10.4%	14.4%	25.1%	208.5%	190.2%	0.0%	0.0%
k_3	17.88%	12.6%	13.1%	5.3%	250.1%	278.7%	0.0%	0.0%
k_{-3}	10.36%	25.0%	19.3%	78.2%	1358.3%	1015.9%	0.0%	0.0%
k_4	54.51%	14.0%	10.0%	70.8%	90.9%	45.1%	0.0%	0.0%
k_{-4}	326.95%	1500.%	1732.7%	9.7%	895.2%	672.6%	0.0%	0.0%
k_5	24.01%	58.1%	14.8%	15.8%	18.6%	16.2%	0.0%	0.0%
k_{-5}	247.64%	321.3%	1530.0%	1900.0%	1900.0%	1900.0%	0.0%	0.0%
f_{GFE}	$1.9 \cdot 10^{-3}$	$2.6 \cdot 10^{-2}$	$4.3 \cdot 10^{-4}$	$4.9 \cdot 10^{-2}$	–	–	–	–
f_{Trad}	$3.4 \cdot 10^{2}$	$1.2 \cdot 10^{3}$	$7.6 \cdot 10^{2}$	$1.2 \cdot 10^{4}$	$1.4 \cdot 10^{2}$	$1.0 \cdot 10^{2}$	$5.4 \cdot 10^{-20}$	$1.1 \cdot 10^{-20}$

Table 5: Optimal Experimental Design. The observables are the dynamical variables listed in Table 1 and Table 2. The control inputs not listed in this table were fixed to the values in Table 1 and Table 2.

	Chlorite-Iodide	Oregonator
# stimuli (control inputs)	3	1
Input limits	$10^{-3} < [ClO_2^-]_0 < 10^{-2}$	$10^{-2} < [BrO_3^-] < 10^{-1}$
	$10^{-3} < [I^-]_0 < 5 \cdot 10^{-2}$	
	$10^{-3} < k_0 < 10^{-1}$	
Stimulation type	3 steps	3 steps
Experiment duration	1000 seconds	100 seconds
Sampling time, ΔT	1 second	1 second
Optimality criterion, Eq. (10)	$E_{opt} = 9.5586 \cdot 10^{10}$	$E_{opt} = 1.0933 \cdot 10^5$
Computation time	48 hours	48 hours

Table 6: Comparison of GFE-based and traditional parameter estimation with optimally designed 'experiments', chlorite-iodide. The estimated parameters are rate coefficients (k_i). The table entries are the percentage errors. f_{GFE} is the value of the GFE-based objective function, Eq. (7), while f_{Trad} is the value of the traditional objective function, Eq. (8). Two different time limits (t_{max}) were used. The sampling time was $\Delta T = 1$ second in all cases. Columns A: $t_{max} = 15$ minutes. Columns B: $t_{max} = 48$ hours.

	GFE		Traditional	
	A	B	A	B
k_2	0.8%	0.8%	72.6%	0.7%
k_3	0.9%	0.9.%	16.2%	0.0%
k_4	0.7%	0.7%	68.5%	0.7%
k_5	0.3%	0.3%	77.8%	0.7%
k_6	0.4%	0.4%	124.4%	0.0%
k_7	16.6%	19.4%	96.0%	0.7%
k_8	0.8%	0.8%	39.8%	0.0%
k_9	4.6%	4.4%	77.5%	0.1%
k_{10}	32.6%	32.6%	6.9%	0.0%
f_{GFE}	$5.6 \cdot 10^{-1}$	$5.6 \cdot 10^{-1}$	–	–
f_{Trad}	8.3	8.3	3.9	$3.8 \cdot 10^{-5}$

1
2
3
4
5
6
7
8
9
10
11
12

Table 7: Comparison of GFE-based and traditional parameter estimation with optimally designed 'experiments', Oregonator. The estimated parameters are rate coefficients (k_i). The table entries are the percentage errors. f_{GFE} is the value of the GFE-based objective function, Eq. (7), while f_{Trad} is the value of the traditional objective function, Eq. (8). Two different time limits (t_{max}) were used. The sampling time was $\Delta T = 0.1$ seconds in all cases. Columns A: $t_{max} = 15$ minutes. Columns B: $t_{max} = 48$ hours.

	GFE		Traditional	
	A	B	A	B
k_{-1}	20.3%	55.4%	3.8%	0.0%
k_2	0.7%	0.0%	459.2%	0.0%
k_{-2}	9.4%	13.2%	366.6%	0.0%
k_3	7.0%	11.5%	4.9%	0.0%
k_{-3}	77.3%	10.7%	116.2%	0.0%
k_4	9.9%	63.5%	22.2%	0.0%
k_{-4}	79.1%	258.8.%	87.5%	0.3%
k_5	27.6%	94.9%	3.2%	0.0%
k_{-5}	1897.0%	79.9%	1544.5%	0.0%
f_{GFE}	$1.7 \cdot 10^{-4}$	$4.1 \cdot 10^{-6}$	–	–
f_{Trad}	6.2	$1.1 \cdot 10^{3}$	$4.1 \cdot 10^{-2}$	$2 \cdot 10^{-18}$

Table 8: Comparison of the Hybrid, GFE-based and Traditional methods, chlorite-iodide. The estimated parameters are rate coefficients (k_i). The table entries are the percentage errors. Simulated data was obtained from the optimally designed experiments (sampling time $\Delta T = 1$ second). f_{Trad} is the value of the traditional objective function, Eq. (8). Two different time limits (t_{max}) are considered. Columns A: $t_{max} = 15$ minutes. Columns B: $t_{max} = 1$ hour.

	Hybrid		Traditional		GFE	
	A	B	A	B	A	B
k_2	1.0%	0.7%	72.6%	134.1%	0.8%	0.8%
k_3	0.3%	0.0%	16.2%	0.6%	0.9%	0.9%
k_4	0.7%	0.7%	68.5%	128.6%	0.7%	0.7%
k_5	0.6%	0.4%	77.8%	138.2%	0.3%	0.3%
k_6	0.4%	0.3%	124.4%	124.0%	0.4%	0.4%
k_7	290.75%	276.5%	96.0%	223.4%	16.6%	19.1%
k_8	0.2%	0.2%	39.8%	2.4%	0.8%	0.8%
k_9	12.4%	20.8%	77.5%	334.1%	4.6%	4.5%
k_{10}	0.4%	0.0%	6.9%	0.1%	32.6%	32.6%
f_{Trad}	$6.8 \cdot 10^{-4}$	$4.5 \cdot 10^{-5}$	3.9	$1.1 \cdot 10^{-2}$	8.3	8.3