Model Selection and Optimal Sampling in High-Throughput Experimentation

ARTICLE *in* ANALYTICAL CHEMISTRY · JULY 2004

Impact Factor: 5.64 · DOI: 10.1021/ac035542o · Source: PubMed

4 AUTHORS, INCLUDING:

Johan Westerhuis
University of Amsterdam
123 PUBLICATIONS **3,922** CITATIONS

Anal. Chem. **2004,** *76,* 3171−3178

Model Selection and Optimal Sampling in High-Throughput Experimentation

Johan A. Westerhuis,[†] Hans F. M. Boelens,[†] David Iron,[‡] and Gadi Rothenberg*,[†]

Chemical Engineering Department and KdV Institute, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands

The practical difficulties encountered in analyzing the kinetics of new reactions are considered from the viewpoint of the capabilities of state-of-the-art high-throughput systems. There are three problems. The first problem is that of model selection, i.e., choosing the correct reaction rate law. The second problem is how to obtain good estimates of the reaction parameters using only a small number of samples once a kinetic model is selected. The third problem is how to perform both functions using just one small set of measurements. To solve the first problem, we present an optimal sampling protocol to choose the correct kinetic model for a given reaction, based on *T*-optimal design. This protocol is then tested for the case of second-order and pseudo-first-order reactions using both experiments and computer simulations. To solve the second problem, we derive the information function for second-order reactions and use this function to find the optimal sampling points for estimating the kinetic constants. The third problem is further complicated by the fact that the optimal measurement times for determining the correct kinetic model differ from those needed to obtain good estimates of the kinetic constants. To solve this problem, we propose a Pareto optimal approach that can be tuned to give the set of best possible solutions for the two criteria. One important advantage of this approach is that it enables the integration of a priori knowledge into the workflow.

The introduction of high-throughput experimentation (HTE) to the chemical laboratory is revolutionizing experimental chemistry. It is changing the value of the basic scientific unit operation—the laboratory experiment. This change has both practical and psychological consequences that must be addressed if chemists are to make good use of the HTE revolution.[1−3]

In the 19th and the 20th centuries, owing to the high costs in manpower and facilities, each single experiment had to be justified. Moreover, a laboratory's experimental capacity often played an important role in determining a project's success. Relatively few experiments were performed, and chemists got used to a modus operandi wherein manual data analysis was performed alongside each project and further eased with the introduction of personal computers into many laboratories in the 1980s.[4]

This is completely changing now. HTE is past the "science" barrier—it is an enabling technology.[5,6] HTE tools are commercially available from a number of companies such as Symyx,[7] hte,[8] and Chemspeed.[9] Robotic systems can take over many routine laboratory handlings—platforms for parallel experimentation with 20−200 batch or flow reactors cost as little as $10 000 (these systems have been on the market for 3−5 years, proven and tested). Moreover, prices are likely to decrease as competition and sales volume increase. HTE systems are especially useful for the "discovery + optimization" approach, wherein many test experiments are run simultaneously (discovery stage), with the most promising leads used for further study (optimization stage).[10−12]

The bottleneck is now shifting from performing the experiments to data analysis and management. Information must be sifted to distinguish between "valuable data" and "garbage data".[13] Specifically, a HTE system should be able to do the following: (i) analyze data automatically, (ii) respond on-line to analysis feedback, and (iii) avoid performing unnecessary measurements (that give little or no information on the reaction).

Catalysis is one particular field where HTE techniques can have a profound effect. In so-called "combinatorial catalysis", many reactions are run in parallel under different conditions to find the

* Corresponding author Fax: +31 20 525 5604. E-mail: gadi@science.uva.nl.
[†] Chemical Engineering Department.
[‡] KdV Institute.

(1) Wilson, R. C.; Hill, D. R.; Gibbs, P. R. In *High-throughput Synthesis*; Sucholeiki, I., Ed.; Marcel Dekker: New York, 2001; pp 271−282.
(2) Czarnik, A. W. *Acc. Chem. Res.* **1996**, *29*, 112−113.
(3) Lutz, M. W.; Menius, J. A.; Choi, T. D.; Laskody, R. G.; Domanico, P. L.; Goetz, A. S.; Saussy, D. L. *Drug Discovery Today* **1996**, *1*, 277−286.
(4) For example, this is what *Fundamentals of Analytical Chemistry* had to say about FT-IR in 1992: "Fourier-transform infrared spectrometers offer the advantages of unusually high sensitivity, resolution, and speed of data acquisition...offsetting these advantages is their high cost, because a moderately sofisticated dedicated computer is needed to decode the output data" (this was then a 386-processor PC): Skoog, D. A.; West, D. M.; Holler, F. J. *Fundamentals of Analytical Chemistry*, 6th ed.; HBJ: New York, 1992.
(5) Reetz, M. T. *Angew. Chem., Int. Ed.* **2001**, *40*, 284−310.
(6) Pescarmona, P. P.; van der Waal, J. C.; Maxwell, I. E.; Maschmeyer, T. *Catal. Lett.* **1999**, *63*, 1−11.
(7) www.symyx.com.
(8) www.hte-company.de.
(9) www.chemspeed.com.
(10) Holzwarth, A.; Denton, P.; Zanthoff, H.; Mirodatos, C. *Catal. Today* **2001**, *67*, 309−318.
(11) Jandeleit, B.; Schaefer, D. J.; Powers, T. S.; Turner, H. W.; Weinberg, W. H. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2495−2532.
(12) Maxwell, I. E. *Nature* **1998**, *394*, 325−326.
(13) Caruthers, J. M.; Lauterbach, J. A.; Thomson, K. T.; Venkatasubramanian, V.; Snively, C. M.; Bhan, A.; Katare, S.; Oskarsdottir, G. *J. Catal.* **2003**, *216*, 98−109.

10.1021/ac035542o CCC: $27.50
Published on Web 05/01/2004

best catalyst for a given system.[14] Preferably, all these reactions would be monitored in real time to estimate kinetic constants for each reaction. Measuring the concentration of reactants and products in real time is not trivial and usually requires expensive instruments (e.g., GC, HPLC, or spectrometric analyzers). The system must often share one analyzer between many reactors, making continuous analysis of all of the reactions impossible. In this case, the bottleneck in obtaining information has shifted from performing many reactions to analyzing many samples of these reactions.

Previously, we tackled the problem of sharing the analyzer time. We developed optimal sampling strategies to obtain maximum information from only a small number of measurements.[15,16] To this end, we introduced the concepts of the *information function* **f** and the *information gain ratio* χ_r, with which the sampling of an array of first-order reactions could be optimized.[17] Subsequently we extended this concept to a cascade of first-order reactions of the type A → B → C.[18]

In those studies, we assumed that the kinetic model of the reaction was known a priori. In practice, however, there is often only a general idea of the reaction mechanism, with no clear microkinetic scheme. In many studies in catalysis, the objective is to carry out reactions under slightly different chemical or physical conditions, which frequently changes the primary reaction pathway (i.e., gives different observed kinetics). Therefore, our challenge was to develop a data analysis approach suitable for HTE, which could be used to select between different kinetic models as well as minimize the number of samples. Here, we present a sampling approach for situations where the kinetic model of the reaction is unknown. We provide the theoretical background required to perform this analysis and simulate various scenarios where two kinetic models (a second-order reaction and a pseudo-first-order reaction) have to be distinguished. Finally, we compare the simulations to experimental results from a set of carbon−sulfur cross-coupling experiments performed under various conditions.

RESULTS AND DISCUSSION

General Considerations. This section will cover three subjects. First, we will introduce a *T*-optimal sampling design to select the best model for a reaction based on as few measurements as possible (a sampling design contains the times at which the measurements should be made). Then, we will present a *D*-optimal sampling design for second-order models in which the reaction rate constant for a second-order model can be estimated with minimum error, also based on as few measurements as possible. Unfortunately, we will see that the best sampling design for model selection leads to a bad estimation of the rate constant and vice versa. Finally, we will combine the *T*-optimality and the *D*-optimality criteria using a Pareto approach to obtain a sampling design, which is good for both model selection and for reaction rate estimation.

Optimal Sampling for Model Selection. In many complex reactions (e.g., catalytic cycles), the microkinetic model is unclear. To select a kinetic model for such a reaction, usually many measurements of reactants or products are obtained and fitted on a set of plausible models. A "good" model fits the data well and no systematic differences between the measurements and the model are present, while for a "bad" model there will be systematic differences. The more samples taken from the reaction, the better the selection of the best model can be made. In a high-throughput setup, however, the number of measurements must often be kept to a minimum.[17] Here we will describe a strategy that predicts when to measure the reaction to be able to decide between two models for a specific reaction with as few measurements as possible. We focus on two models, but the theory is general and can be extended to a larger set of kinetic models.

Let us consider the reaction A + B → P. Reactant A can be measured using an analytical technique at any time during the reaction. For this reaction, the starting concentration of B ($[B]_0$) is higher than the starting concentration of A ($[A]_0$). It is not known according to which kinetic model this reaction proceeds. The objective is to decide which kinetic model applies to the disappearance of reactant A using as few measurements as possible.

Two models are compared, one of which is assumed to be the true kinetic model $\eta_c(\mathbf{t},k_c)$, and an alternative model $\eta_w(\mathbf{t},k_w)$. Here the vector **t** contains the times at which a measurement should be performed, and k_c and k_w are the kinetic constants for the true and for the alternative model, respectively. In this example, we will assume that the reaction follows a second-order rate law, i.e., that the second-order model is the true kinetic model, and compare it with an alternative pseudo-first-order model. Measuring A at any combination of sampling times will give an estimate for both reaction rate constants, $\hat{k}_c$ and $\hat{k}_w$. Note that besides the initial concentrations of A and B, at least two measurements of the concentration of A should be performed to distinguish between the two models (if only one additional measurement is available, both models will fit the data perfectly and no selection can be made). The true model will fit the data perfectly up to the measurement error in A. The wrong model will also show systematic differences, called the lack of fit (or the fit error). If the fit error is larger than the measurement error, then it is possible to distinguish the wrong model from the true model. The larger the fit error, the more confidence one has in selecting the true model. For the best possible selection between the true and the alternative model, the fit error *T*, given by eq 1, of the wrong model should be maximized.[19] At the maximum *T*, the sampling design **t** is called *T*-optimal. Here $\eta_c(\mathbf{t},k_c)$ are the concentrations of A at the times in the sampling design **t** according to the true model η_c with the "true" reaction rate constant k_c. Likewise, $\eta_w(\mathbf{t},\hat{k}_w)$ contains the concentrations of A predicted by the wrong model with the estimate $\hat{k}_w$. The summation is over all sampling times in the sampling design **t**.

(14) Senkan, S. *Angew. Chem., Int. Ed.* **2001**, *40*, 312−329.
(15) Rothenberg, G.; Boelens, H. F. M.; Iron, D.; Westerhuis, J. A. *Catal. Today* **2003**, *81*, 359−367.
(16) Rothenberg, G.; Boelens, H. F. M.; Iron, D.; Westerhuis, J. A. *Chim. Oggi* **2003**, *21*, 80−83.
(17) Boelens, H. F. M.; Iron, D.; Westerhuis, J. A.; Rothenberg, G. *Chem. Eur. J.* **2003**, *9*, 3876−3881.
(18) Iron, D.; Boelens, H. F. M.; Westerhuis, J. A.; Rothenberg, G. *Anal. Chem.* **2003**, *75*, 6701−6707.
(19) Atkinson, A. C.; Donev, A. N. *Optimal Experimental Designs*; Clarendon Press: Oxford, U.K., 1992.

$$T = \sum_t (\eta_c(\mathbf{t},k_c) - \eta_w(\mathbf{t},\hat{k}_w))^2 \quad (1)$$

Our objective is to find the pair of measurement times that enables us to distinguish the first-and the second-order model in the best way. Using the assumed "true" model η_c and the initial estimate of k_c, we can calculate the "true" concentration of A at all sampling times. The initial estimate of k_c might be available from earlier measurements or from prior knowledge. For each pair of sampling times, the concentrations of A are then fitted to the wrong model η_w. This will give slightly different $\hat{k}_w$ values for all designs. T is the sum of the error between the "true" concentration as predicted by the correct model and the wrongly estimated concentration by the wrong model over all sampling times in the design **t**. Dividing T by the number of samples and taking the square root makes it a quantity directly comparable to the measurement error in analyzing A. This is the root-mean-squared fit error (rmse) of the wrong model. Below is a protocol for model selection using a *T*-optimal design: (1) Select "true" model and assign initial k_c value. (2) Calculate concentration profile of analyte. (3) Define set of sampling designs. (4) Estimate $\tilde{k}_w$ and calculate T of alternative model for each sampling design. (5) Use sampling design with T_{max} to perform measurements. (6) Fit measurements to both models using nonlinear regression. (7) Estimate kinetic parameters for both models and calculate T values. (8) Select model with lowest fit error T as "true" model. (9) If fit error is not satisfactory go back to step 2.

To get a **t** design for *T*-optimal sampling, one of the kinetic models is selected as the correct one (η_c). This is an advantage, as it allows the user to introduce a priori knowledge in the data processing method. If no such knowledge is available, it is best to select the simplest kinetic model.

Model Selection versus Parameter Estimation. How is the *T*-optimal sampling design related to the *D*-optimal sampling design for parameter estimation?[17] In *T*-optimality, after the measurements have been obtained according to a sampling design, both models can be fitted to the data. If the assumed true model is indeed the model that best describes the measured concentration, the kinetic parameter $\hat{k}_c$ can be estimated. However, the measurements are usually not performed at the optimal times to estimate $\hat{k}_c$ with minimum variance. Thus, the *T*-optimal design that is optimal for selecting the correct model need not be optimal for estimating the kinetic constants. The optimal sampling times of both criteria do not necessarily coincide. It is, however, possible to perform measurements at suboptimal sampling times (with regard to the model selection) in order to obtain an improved estimate of $\hat{k}_c$.

These suboptimal times can be selected using a Pareto optimal (PO) approach.[20,21] A sampling design is Pareto optimal when no other sampling design exists in which the scores on both criteria are simultaneously improved. There will be PO designs in which the performance for one criterion will improve and accordingly the performance on the other will deteriorate. Figure 1 shows the Pareto optimal design concept for a set of sampling designs

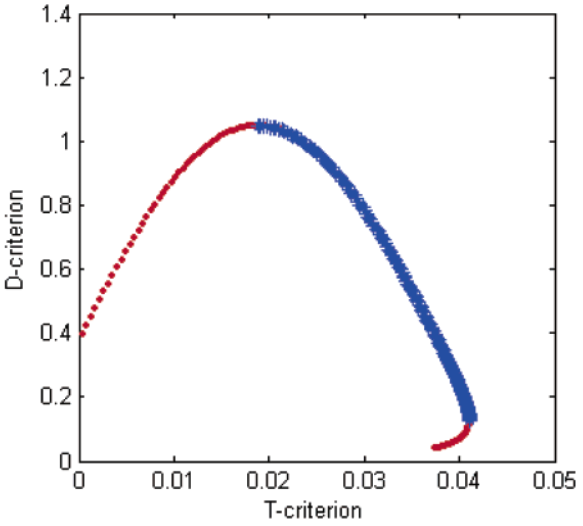


Figure 1. PO plot of the *T* and *D* criteria. PO sampling designs (∗) can be found at the upper right corner of the figure since both criteria are maximized. Sampling designs that are not PO (●) can be improved in both criteria simultaneously.

where both the *D* criterion and the *T* criterion have to be optimized. Here the ∗ symbols represent PO sampling designs, while the ● symbols denote designs that are not PO. The nice feature is that, given the range of sampling points that are PO, the user may decide to emphasize more the selection of the correct model (the *T* criterion) or the estimation of the rate constant (the *D* criterion). This decision can thus be made dependent on prior knowledge, e.g., the measurement error in A. From Figure 1 it is clear that one criterion can be improved considerably without giving up too much of the quality of the other criterion.

Optimal Sampling of a Second-Order Reaction. In an earlier paper, we developed an optimal sampling approach for the estimation of the rate constant k_1 in a pseudo-first-order reaction.[17] Here we extend this approach also to second-order reactions. The reaction A + B → P follows a second-order rate law when $[A]_0 \approx [B]_0$. The concentration of A during the reactions can be obtained from eq 2, in which $[A]_t$ is the concentration of A at time t during

$$[A]_t = \frac{\delta_0[A]_0}{[B]_0 e^{k\delta_0 t} - [A]_0} \quad (2)$$

the reaction. $[A]_0$ and $[B]_0$ are the initial concentrations of the reactants. We will assume that $[A]_0 < [B]_0$, that $\delta_0 = [B]_0 - [A]_0$, and that the concentration of A can be obtained at any point in time during the reaction.

For this system, the information vector $\mathbf{f}_2$ (eq 3) relates the sampling time to the variance of the estimate of the second-order

$$\mathbf{f}_2 = \frac{-\delta_0^2[A]_0[B]_0\mathbf{t}e^{k_2^0\delta_0\mathbf{t}}}{([B]_0 e^{k_2^0\delta_0\mathbf{t}} - [A]_0)^2} \quad (3)$$

reaction rate constant (k_2). Sampling at times that pertain to high $\mathbf{f}_2$ values will yield better estimates of k_2 (having low variance). If $[A]_0$ and $[B]_0$ are known, one additional measurement of A will

(20) Hendriks, M.; de Boer, J. H.; Smilde, A. K.; Doornbos, D. A. *Chemom. Intell. Lab. Syst.* **1992**, *16*, 175−191.
(21) de Boer, J. H.; Smilde, A. K.; Doornbos, D. A. *Acta Pharm. Technol.* **1988**, *34*, 140−143.

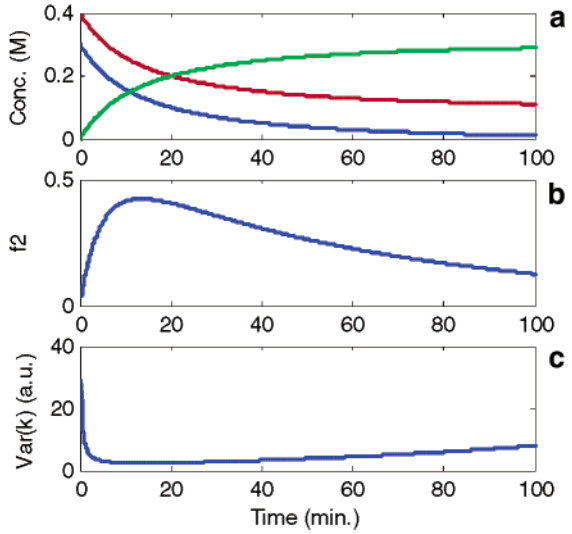


Figure 2. Second-order reaction data with $[A]_0 = 0.3$ M, $[B]_0 = 0.4$ M, and $k_2 = 0.2$ $M^{-1}min^{-1}$. (a) Concentration profiles of A, B, and P. (b) The norm of the information vector $\mathbf{f}_2$. (c) The variance in the estimation of k_2 due to the measurement error in concentration.

suffice to obtain an estimate of k_2 (the full derivation of the information vector $\mathbf{f}_2$ for second-order reactions is given in the Appendix, vide infra).

The information vector is calculated using the known values of $[A]_0$ and $[B]_0$ and an initial estimate of the reaction rate constant k_2^0. Given the information vector $\mathbf{f}_2$, the variance of the estimate of the rate constant k_2 is given by eq 4, where $||\mathbf{f}_2||^2$ is the norm

$$\mathrm{var}(k_2) = \frac{1}{||\mathbf{f}_2||^2}\,\mathrm{var}([A]) \qquad (4)$$

of $\mathbf{f}_2$ and var([A]) is the variance from the analytical method used to monitor A. This variance is assumed to be independent of the concentration of A.

Equation 4 shows that the variance of the estimate of k_2 is minimized when the norm of $\mathbf{f}_2$ is maximized.[19] Figure 2a shows the concentration profiles for the second-order reaction where $[A]_0 = 0.3$ M, $[B]_0 = 0.4$ M, and $k_2 = 0.2$ $M^{-1}\cdot min^{-1}$. Figure 2b shows the information vector $\mathbf{f}_2$ and the variance in the estimation of k_2 as a function of the sampling time when only one additional measurement is performed. At the maximum of $\mathbf{f}_2$, the variance in k_2 is minimal. The decrease in variance of k_2 is especially fast in the first part of the reaction. The increase in variance after the minimum at 13.6 min is much slower. Thus, performing the measurement after 10 min or after 20 min has little effect the estimation of k_2.

Two special cases can be considered for this reaction. The first is when a large excess of B is present (i.e., when $[B]_0 \gg [A]_0$). In that case, the reaction is known to be a pseudo-first-order reaction and first-order kinetics will apply. The second case is when $[B]_0 = [A]_0$. This is the case similar to the reaction $2A \rightarrow P$. For this second-order reaction, the information vector simplifies into eq 5

$$\mathbf{f}_2^{\,*} = \frac{-([A]_0)^2\mathbf{t}}{(1 + k_2^{\,0}\mathbf{t}[A]_0)^2} \qquad (5)$$

Table 1. Optimal Sampling Times for Second-Order Reactions at Varying $[B]_0$[a]

$[B]_0$	optimal sampling time for second-order reaction/min
0.3	16.67
0.5	11.10
1.0	5.37
3.0	1.71
30.0	0.17

[a] $[A]_0$ is 0.3 M and $k_2 = 0.2$ $M^{-1}\cdot min^{-1}$. Note that for $[B]_0 = 30$ M the results equal the first-order model results.

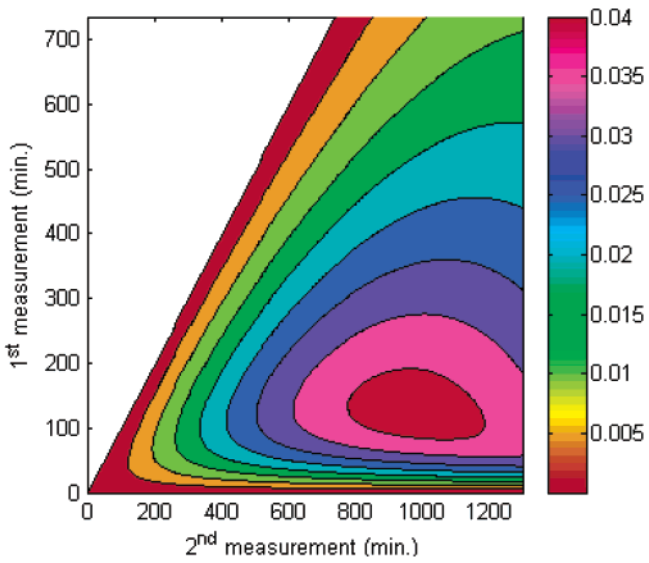


Figure 3. Fit error of alternative first-order model at any combination of sampling times, considering $[A]_0$ is known. The maximum fit error for the first-order model is found at 127 and 975 min for the two samples.

(see Appendix for the derivation). The norm of $\mathbf{f}_2^{\,*}$ will be maximal at $1/(k_2^0[A]_0)$, and the best sampling time to obtain k_2 for this reaction with minimum variance is at $t = 1/(k_2^0[A]_0)$. Table 1 gives the optimal sampling times for some situations in which $[B]_0$ is increased while $[A]_0$ equals 0.3 M and $k_2 = 0.2$ $M^{-1}\cdot min^{-1}$.

Testing the Theory Using Simulations. Here we examine the above theory by simulating the reaction $A + B \rightarrow P$. We assume that there is a slight excess of B ($[A]_0 = 1.00$ mM, $[B]_0 = 1.70$ mM) and that the reaction will follow a second-order rate law. A *T*-optimal design is calculated at which the largest difference between the "correct" second-order model and an alternative first-order model is obtained. The true kinetic constant k_2 is taken as 2×10^{-6} $M^{-1}\cdot min^{-1}$. The reaction is monitored over 1300 min, and the simulations are based on noiseless data.

We consider two situations. The first assumes that $[A]_0$ and $[B]_0$ are known and that an accurate estimate of the kinetic constant is available. Every combination of two sampling times is tested. At each combination, the concentration of A is calculated based on k_2 and the sampling times. The concentrations are fitted on the alternative first-order model and the differences between the actual concentrations and the concentrations estimated by the first-order model are squared and summed to obtain *T*. Figure 3 shows a contour plot of the fit error of the alternative model for any combination of sampling times. Sampling first at 127 min and

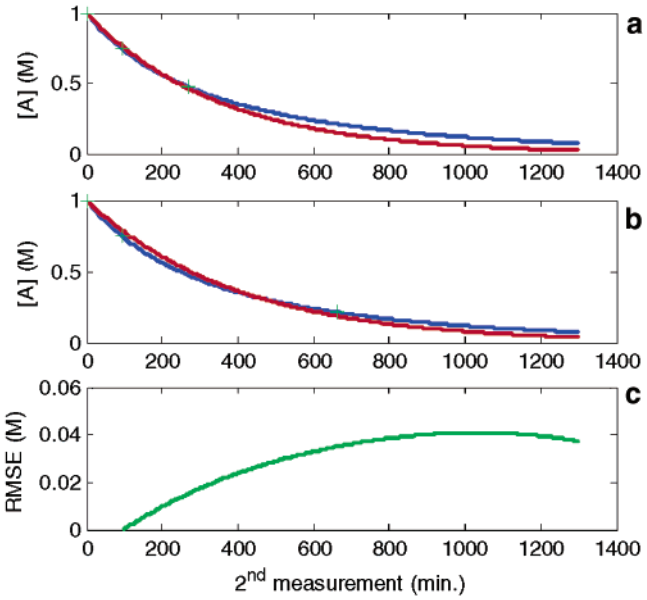


Figure 4. Top and middle: Predicted concentration profiles of A according to second-order (blue) and first-order (red) models. The + symbols denote the measured concentration at the specific sampling times. Bottom: Fit error (rmse) of the first-order model for the second measurement when the first measurement is performed at 93 min.

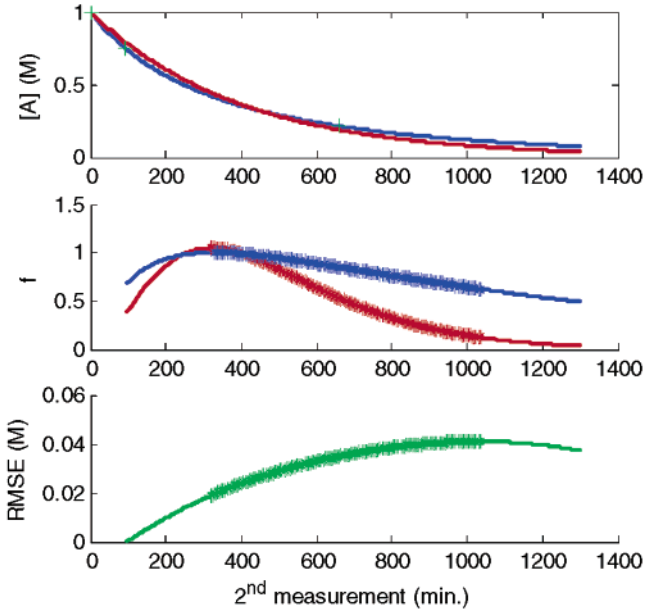


Figure 5. Top: Fitted concentration profiles of A by second-order (blue) and first-order (red) models. Middle: The norm of the information vectors $\mathbf{f}_1$ (red) and $\mathbf{f}_2$ (blue). Bottom: Fit error for the wrong model. The * symbols denote the PO times for the second measurement when the first measurement in performed at 93 min.

then at 975 min gives the largest fit error for the alternative first-order model. This is the *T*-optimal design for this specific case.

In the second situation, it is assumed that no accurate estimate of k_2 is available. This makes it impossible to find a *T*-optimal design because $\hat{k}_2$ must be obtained first. Therefore, an initial measurement has to be performed early in the reaction to get an estimate of k_2. Based on $[A]_0$ and this initial measurement, we obtain $\hat{k}_2$. As we are simulating error-free conditions, the estimated value equals the real value for k_2, 2×10^{-6}. Figure 4 shows the results obtained when the initial measurement is performed after 93 min. The top and the middle sections pertain to second measurements performed at 270 and 660 min, respectively. The predicted concentration profiles for the second-order (blue) and first-order (red) models are shown. Since the reaction is second order, the measured concentrations lie exactly on the blue curve. The two concentration profiles are very similar. The bottom section shows the fit error between the "true" concentrations and the concentrations estimated by the "wrong" first-order model. When the first measurement is performed after 93 min, the optimal sampling time for the second measurement is ~1034 min.

Why is this important for high-throughput studies? If you use a robot system that has several reactors but only one analyzer, it is usually less expensive to repeat a reaction with the same conditions than to take additional samples of the same reaction. Thus, an initial measurement to estimate k_2 should preferably be performed at the *D*-optimal time for that reaction. Although the *D*-optimal time is unknown (because k_2 is unknown), Figure 2 shows that for second-order reactions there is a broad range of sampling times that will provide good estimates of k_2. After this initial measurement to estimate k_2, the system can decide whether to keep the reaction running and do a second measurement at a later time or to start a new reaction and calculate the optimal sampling times for model selection.

As mentioned above, the optimal sampling times for model selection often differ from the optimal sampling times for estimating the kinetic parameters. Here it is useful to employ a Pareto approach to balance the optimality of model selection and the optimality of parameter estimation. Figure 5 shows the application of the Pareto approach to the situations shown in Figure 4. In the top section, both the first- and the second-order profiles are given for the situation where measurements were done at $t = 0$, 93, and 660 min. The estimated *k* values for both models are used to calculate the information vectors $\mathbf{f}_1$ and $\mathbf{f}_2$. The norms of both $\mathbf{f}$ vectors are presented in the middle plot. The values of $\mathbf{f}_1$ and $\mathbf{f}_2$ were scaled to have the same maximum value. The optimal value for the second-order $\mathbf{f}_2$ (red) comes slightly later than $\mathbf{f}_1$ for the first-order model. However, these times are much earlier than the optimal sampling time for model selection (1034 min, bottom plot). Note that the decrease in $\mathbf{f}_2$ after the maximum is much slower than for the first-order reaction. This is related to the slower decrease of reactant A in the second-order case. Pareto optimal sampling times for all three criteria ($\mathbf{f}_1$, $\mathbf{f}_2$, and *T*) are represented by * symbols. If the second measurement is performed at, for example, 600 min, the model selection ability is only slightly decreased while the parameter estimation performance is increased considerably.

Testing the Theory Using Experiments. Next, we compared the above mathematical model with results obtained from a set of controlled experiments. We used the carbon−sulfur coupling of 3-chlorophenylhydrazonopropane dinitrile (**A**) with β-mercaptoethanol (**B**) to give the adduct **C** (Scheme 1).[22,23] The reaction was followed using UV−visible spectroscopy. Figure 6 shows concentration profiles of **A** in five repeated reactions. The

(22) Bijlsma, S.; Boelens, H. F. M.; Hoefsloot, H. C. J.; Smilde, A. K. *Anal. Chim. Acta* **2000**, *419*, 197−207.
(23) Bijlsma, S.; Louwerse, D. J.; Smilde, A. K. *J. Chemom.* **1999**, *13*, 311−329.

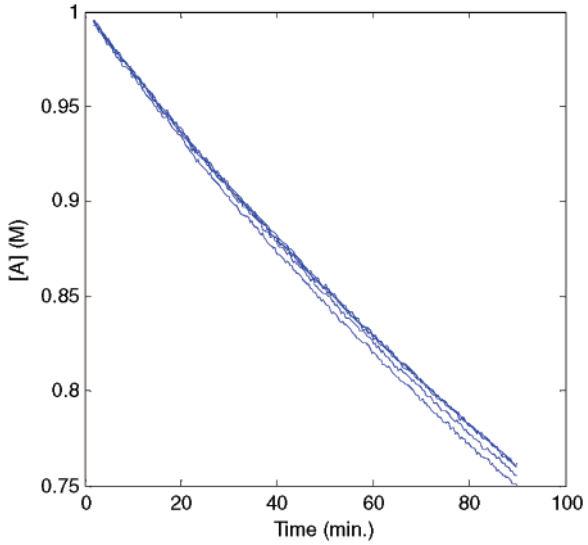
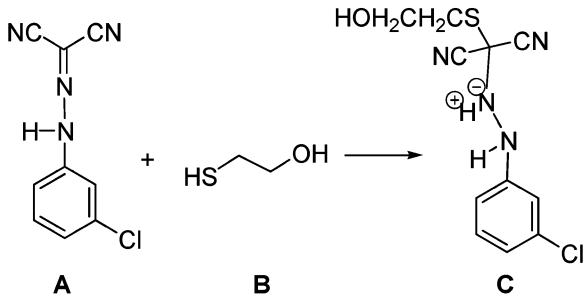


Figure 6. Five concentration profiles of 3-chlorophenylhydrazono-propane dinitrile in the reaction with β-mercaptoethanol. The ratio of initial concentrations is 1:1.7 mol/mol.

Scheme 1



repeatability is excellent so it is justified to assume that systemic effects may be excluded.

For each of the reactions, we used all of the 20 250 combinations of the two sampling times to obtain concentration estimates for A. These were then fitted to a first-order model and to a second-order model. In contrast to the simulations, even the correct second-order model exhibited fit errors due to the measurement error. The fit errors for both models were averaged over the five reactions. Figure 7 shows the difference between the fit errors of the two models for each pair of sampling times. These differences in fit errors are related to the *T*-optimality defined in eq 1 (note that the reaction is only followed for 90 min while in the simulation we reached 98% of conversion only after 1300 min). T_{max} between the first- and second-order models for this reaction was obtained for measurements performed at 30 and 90 min. This is comparable to the simulated results in Figure 2. The value of the difference for these sampling points is 0.003, which is also comparable to the values found in the simulation for those sampling times.

The fit error of the correct second-order model is almost constant, ~0.0005. It is almost independent of the sampling times. There is no systematic error in the fit of the second-order model. The fit errors are only affected by the measurement error of the analysis method. The fit error is much larger for the alternative first-order model. The alternative model is wrong in this case. The magnitude of the fit error for the wrong model is dependent on the sampling times, with $T_{max} = 0.0035$ found at 30 and 90 min.

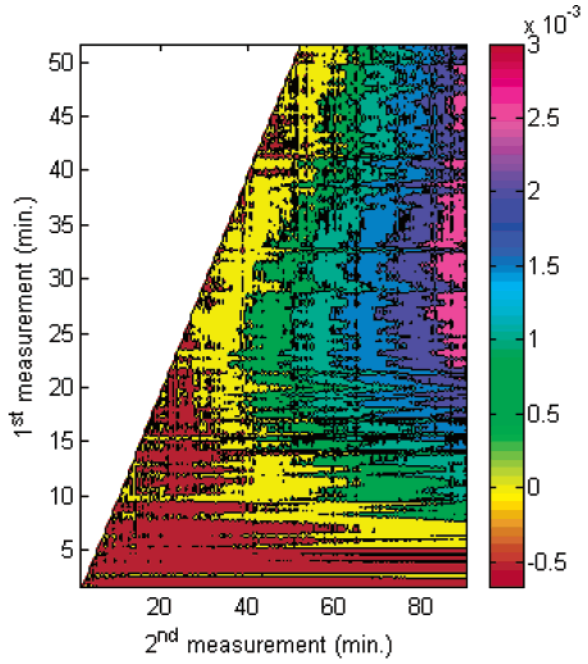


Figure 7. Difference in fit error of a first-order model and a second-order model. The maximum fit error can be found in the region around 30 and 90 min for the two measurements. The difference in fit error is ~0.003 mM.

Even though the above analysis was based on only two measurements, it was possible to distinguish between a second-order and a first-order model. However, if the measurements had been performed at other times, the difference between the two models would have been much smaller and model selection would have been impossible. When the models are more similar (e.g., for a second-order reaction with [A]/[B] = 1:10), more measurements will be necessary to distinguish between the models.

The objective of a "smart" sampling strategy is to reduce the number of samples for each reaction. However, model selection might not always be necessary and the system itself should recognize situations when model selection is required or not. If model selection is not necessary, samples can be placed for optimal parameter estimation of the true model.

Which Situations Would Benefit from Model Selection? This is best illustrated with an example: Consider the reaction A + B → P. This reaction can always be modeled with a second-order model, but if there is a sufficient excess of B, it may be pseudo first order. If $[B]_0$ is only slightly higher than $[A]_0$ then second-order model should be used, but when $[B]_0$ increases to, for example, $50[A]_0$ (see Table 1) the simpler pseudo first order may be sufficient. The definition of "slightly higher" here also depends on the accuracy of the experimental measurements and on the number of samples that you are willing to measure. If, for example, a fast HPLC is used to measure the concentrations at an accuracy of ±5%, then any changes in the concentration of B that are <5% will be insignificant, unless you take many samples. In that case, it is impossible to distinguish between the second-order and the pseudo-first-order models and the latter can (and should) be used without model selection. Model selection is only meaningful, in this case, in the "gray region" (see Figure 8) that corresponds to $[B]_0/[A]_0$ ratios of 1:5−20:1.

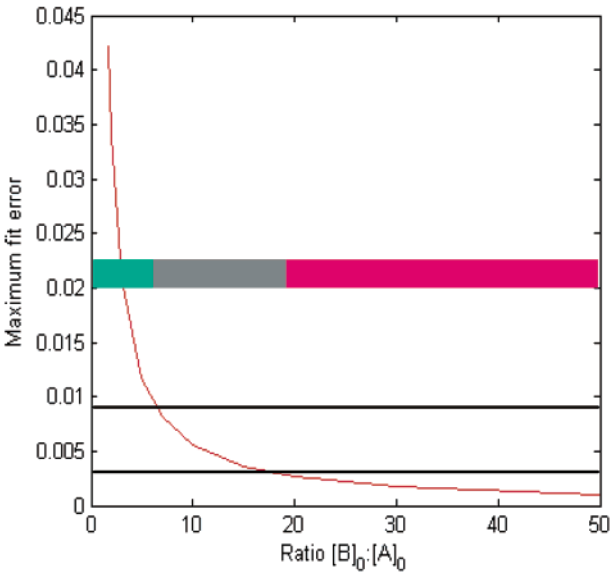


Figure 8. Fit error of first-order model for varying $[B]_0/[A]_0$ ratios. The lower limit (0.003) represents the measurement error of the analysis method. Below this limit (the red zone) no distinction between the first-order and second-order models can be made, and the first one can be chosen without using model selection. The higher limit equals 3 times the measurement error. Above this limit (the green zone) a second-order model can be chosen without using model selection. The gray zone is where model selection should be applied.

CONCLUSIONS

In situations where you do not know which rate law applies to your reaction, and the number of samples you may take is limited, *T*-optimal design can be used to predict when to sample in order to choose between two different kinetic models. As we demonstrate here using both experiments and simulations, this method is effective in selecting the correct option in the case of second-order pseudo-first-order models. In practice, the optimal sampling times for model selection and those for kinetic parameter estimation rarely coincide. To deal with this problem, a Pareto optimal approach can be used, allowing a small decrease in the model selection ability in order to improve the quality of kinetic parameter estimation.

EXPERIMENTAL SECTION

All chemicals were commercially available (99% pure) and used without further purification. KH_2PO_4 buffers were purchased from Acros (pro analysi 0.2 M). UV−visible spectra were recorded using a Hewlett-Packard 8453 spectrophotometer (quartz cuvettes, 1.00-cm path length). Data processing was performed using MATLAB.[24] A detailed description of the sample preparation methods and the experimental apparatus was published previously.[25]

A stock solution of 3-chlorophenylhydrazonopropane dinitrile (**A**) (1.034 M in 0.1 N NaOH) was prepared. For each experiment, part of this stock solution was then diluted to 51.71 μM, buffered to pH 4.4 with KH_2PO_4, and mixed in the quartz cuvette with a β-mercaptoethanol (**B**) solution in a KH_2PO_4 buffer to give a final ratio of 1:1.7 mol/mol. UV−visible spectra of the reaction mixtures were recorded at a wavelength range from 300 to 500 nm. The reaction was sampled every 20 s for 90 min. Temperature was maintained at 25 °C. The reaction was performed six times. The results of one reaction were used to estimate the pure spectra of the reactants and the product. Using these pure spectra the concentration profiles of **A** were obtained for the remaining five reactions. In 90 min, conversions up to 25% were obtained. The first few measurements of each reaction were not used in the analysis, as these spectra were rather noisy due to mixing effects.

(24) MATLAB version 6.1, 2001, is commercially available from MathWorks.
(25) Bijlsma, S.; Boelens, H. F. M.; Smilde, A. K. *Appl. Spectrosc.* **2001**, *55*, 77−83.

APPENDIX

Derivation of the Information Vector for a Second-Order Model. Equation 2 is expanded in a Taylor series around an initial estimate k_0 of the second-order reaction rate constant k_2 sought for. Only the first term of the series is considered. We assume that **t** is the sampling design and contains sampling times. $[A]_t$ contains measurements of A obtained at the sampling times in **t**.

$$[A]_t = \frac{\delta_0[A]_0}{[B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0} + (k - k_0)\frac{-\delta_0^2[A]_0[B]_0 \mathbf{t} e^{k_0\delta_0 \mathbf{t}}}{([B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0)^2}$$

$$[A]_t - \frac{\delta_0[A]_0}{[B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0} = (k - k_0)\frac{-\delta_0^2[A]_0[B]_0 \mathbf{t} e^{k_0\delta_0 \mathbf{t}}}{([B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0)^2}$$

$$[A]_t - \frac{\delta_0[A]_0}{[B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0} = \frac{\delta_0^2[A]_0[B]_0 \mathbf{t} e^{k_0\delta_0 \mathbf{t}} k_0}{([B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0)^2} - \frac{\delta_0^2[A]_0[B]_0 \mathbf{t} e^{k_0\delta_0 \mathbf{t}} k}{([B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0)^2}$$

$$[A]_t - \frac{\delta_0[A]_0}{[B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0} - \frac{\delta_0^2[A]_0[B]_0 \mathbf{t} e^{k_0\delta_0 \mathbf{t}} k_0}{([B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0)^2} = - \frac{\delta_0^2[A]_0[B]_0 \mathbf{t} e^{k_0\delta_0 \mathbf{t}}}{([B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0)^2} k \quad (6)$$

This problem can be regarded as a least-squares problem $\mathbf{y} = \mathbf{f}_2 k + \mathbf{e}$ in which k is estimated in a least-squares sense.

$$\mathbf{y} = [A]_t - \frac{\delta_0[A]_0}{[B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0} - \frac{\delta_0^2[A]_0[B]_0 \mathbf{t} e^{k_0\delta_0 \mathbf{t}} k_0}{([B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0)^2}$$

$$\mathbf{f}_2 = \frac{-\delta_0^2[A]_0[B]_0 \mathbf{t} e^{k_0\delta_0 \mathbf{t}}}{([B]_0 e^{k_0\delta_0 \mathbf{t}} - [A]_0)^2} \quad (7)$$

Here **y** contains the measured concentration profile, $\mathbf{f}_2$ contains the theoretical second-order concentration profile for A considering $k = k_0$ and is considered errorless. The scalar k is the parameter to be estimated; i.e., what is the value of k that minimizes the difference between $[A]_t$ and the theoretical second-order curve under the assumption that $[A]_0$ and $[B]_0$ are given and an initial estimate k_0 is available. The least-squares solution to this linear regression problem is given as follows:

$$\hat{k} = (\mathbf{f}_2{}^T\mathbf{f}_2)^{-1}\mathbf{f}_2{}^T\mathbf{y} = \frac{\mathbf{f}_2{}^T\mathbf{y}}{||\mathbf{f}_2||^2} \quad (8)$$

The variance of the estimate of the rate constant $\hat{k}$ is as follows:

$$\mathrm{var}(\hat{k}) = \frac{1}{||\mathbf{f}_2||^2}\mathrm{var}([\mathrm{A}]) \quad (9)$$

Here var([A]) represents the measurement error of the analytical technique used to monitor A. Var([A]) is assumed to be constant for different concentrations of A. The variance of the estimate $\hat{k}$ is minimized when maximizing the norm of $\mathbf{f}_2$. This result is known as the *D*-optimal design.[19] When the initial concentrations of A and B are equal, the situation is similar to the reaction 2A → P. In that case, the concentration of A at time *t* simplifies to

$$[\mathrm{A}]_t = \frac{[\mathrm{A}]_0}{1 + k\mathbf{t}[\mathrm{A}]_0} \quad (10)$$

The information vector $\mathbf{f}_2$* for this situation is derived in the same way as above. Using a similar derivation as above, the information vector for this situation equals

$$\mathbf{f}_2{}^* = \frac{-([\mathrm{A}]_0)^2\mathbf{t}}{(1 + k_0\mathbf{t}[\mathrm{A}]_0)^2} \quad (11)$$

Here the norm of $\mathbf{f}_2$* will be maximal at $1/(k_0{}^*[\mathrm{A}]_0)$. Thus, the best sampling time to obtain *k* with minimum variance is at $t = 1/(k_0{}^*[\mathrm{A}]_0)$.

Received for review December 31, 2003. Accepted March 18, 2004.

AC035542O