Rapid estimation of chemical kinetics by implicit calibration. I

Veli-Matti Taavitsainen\textsuperscript{1*} and Heikki Haario\textsuperscript{2}

\textsuperscript{1}Espoo-Vantaa Institute of Technology, Vanha Maantie 6, FIN-02600 Espoo, Finland
\textsuperscript{2}ProfMath Oy, Finland

SUMMARY

In this study, methods for estimating kinetic parameters from on-line measurements of batch reactions are developed. The basic idea is to combine non-linear parameter estimation with implicit calibration between measured spectra and concentrations given by a known kinetic model. Several alternatives for the calibration step are discussed and the three most promising methods have been studied more closely. The basic ideas and identifiability questions are elucidated by simulation studies with simple kinetic models. The simulations show that the implicit calibration method works also with uncalibrated chromatograms. Two real esterification reactions with spectroscopic data are studied. In the first one, concentration data are available and our method gives results quite similar to those obtained in more traditional ways. The latter reaction, however, was a fast one and consequently off-line calibration based on known mixtures failed. In contrast, our method gave almost identical parameter values to the ones obtained by an independent study on the same reaction based on a complicated on-line chromatographic measurement system. In this case the results were compared with a new way of using pure component spectra and generalized ridge regression with stoichiometric constraints. Our studies show that the most important factors in obtaining reliable results are proper design of the batch experiments, proper weighting of the initial measurements of each batch and proper measure for goodness of fit.

Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: kinetics; non-linear parameter estimation; spectroscopy; chromatography; reaction modelling

1. INTRODUCTION

In developing processes or process units for new chemical syntheses, it is essential to know the chemical kinetics of all reactions. Traditional methods for kinetics determination are based on concentration data given by various sampling techniques and chemical analyses. These have several drawbacks. Developing a reliable analytical method for some of the reactive species may take a long time. Some species may be very unstable and only on-line methods are applicable. Even with the on-line methods, preparation of standards for calibration purposes may be impossible, e.g. for some intermediate compounds.

In this paper we propose new techniques which are based on combining the calibration and kinetic parameter estimation into a single estimation task called implicit calibration. This study is an ongoing project related to our work [1]. In this study we investigate the applicability, limitations and
alternatives to using prior knowledge of the kinetics and measurements. The simulation studies are extended to chromatographic data, the aim being to solve complex kinetics problems using both chromatographic and spectroscopic data. Two new real cases are studied which show the potential and possible difficulties of the method. In one case the traditional methods fail but implicit calibration gives approximately correct values for the kinetic parameters.

Several authors have studied basically the same problem with slight modifications of the same basic idea to relate the principal components of the spectra to concentrations obtained by the kinetic model. Furusjö and Danielsson [2] study acid-catalysed hydrolysis and nucleophilic substitution reactions and use, in our terminology (see Section 2.3.2), implicit calibration by indirect principal component regression (PCR). Bijlsma et al. [3] study first-order consecutive reactions and use a method which resembles what we would call direct implicit calibration. However, their basis of fit is in the score space of spectra, not in the absorbances themselves. This has the drawback that positivity constraints cannot be applied, because the solution does not directly represent the pure component spectra. Kubista et al. [4] study chemical equilibria by almost the same method. Their mathematical model is a set of algebraic, not differential, equations. All these studies are based on minimizing the residual sum of squares using unconstrained regression. We will show that an $R^2$-based minimization criterion works better in indirect implicit calibration.

It should be noted that our goal is not to achieve a new analytical technique or a new method for process monitoring, all cases where the emphasis would be on estimating the concentrations. Rather, we are seeking techniques for speeding up model-based process development, and not only speeding up modelling, but even to make it feasible in cases where traditional methods fail; consequently, the emphasis is on estimating the model parameters.

2. KINETICS ESTIMATION

2.1. Traditional kinetics estimation

For the convenience of the reader who is not familiar with non-linear kinetic parameter estimation, we give a short introduction to the basic ideas in this section. The goal of kinetic parameter estimation is simply to determine the values of unknown parameters such that the modelled responses fit the measured responses in the best possible way. Of course, ‘in the best possible way’ is not a unique concept. By responses we mean here the dependent variables, typically concentrations, of the model.

Determining kinetic parameters (rate constants, frequency factors, activation energies, etc.) leads into non-linear parameter estimation, except for the simplest cases, and therefore to iterative procedures. Take, for example, the simple reaction $A + B \leftrightarrow C$ with non-equal initial concentrations, i.e. $C_{A_0} \neq C_{B_0}$. This reaction has the following explicit solution for $C_A$:

$$C_A = \frac{C_{B_0} - C_{A_0}}{(C_{B_0}/C_{A_0}) \exp[kt(C_{B_0} - C_{A_0})]} - 1$$

(1)

If the initial concentrations are known, this can be transformed into a linear regression problem by taking reciprocals of both sides of Equation (1), rearranging the terms and finally taking logarithms of both sides. Such transformations always change also the error structure of the problem. Moreover, such transformations do not even exist if the kinetic mechanism is complicated enough. Another important fact is that explicit solutions similar to Equation (1) are available only for the simplest kinetic models. Fortunately, there are many good algorithms and computer codes for numerical solutions of chemical rate equations (e.g. LSODE (Fortran), ODE15S (Matlab)). In this study we have used both Matlab’s ODE15S and Fortran code LSODE which also solve so-called stiff ordinary
differential equations (ODEs). Stiff equations are quite common in chemistry, and kinetic models with fast equilibrium reactions combined with slower reactions produce stiff systems of ODEs.  ‘Traditional’ kinetic parameter estimation is based on the following procedure.

1. Find an initial guess for the unknown kinetic parameters.
2. Solve numerically the ODEs describing the kinetic mechanism. This gives the estimated (modelled) concentrations.
3. Calculate the errors, i.e. the differences between the measured and estimated concentrations at measurement time points.
4. Calculate the sum of squares of the errors (or any other measure of goodness of fit).
5. Improve the guessed values of the unknown kinetic parameters according to some minimization algorithm (e.g. Nelder and Mead’s simplex, Marquardt algorithm).
6. If significant improvement cannot be attained, stop; otherwise go back to step 2.

There are several special questions to be answered: how to deal with multiple responses (multivariate Y variables); how to weight the observations; what is the best criterion for goodness of fit (e.g. SS or $R^2$); which optimizer to use; and so on. A good presentation is given in Reference [5].

2.2 Curve resolution-based estimation

The traditional kinetic parameter estimation is a two-step procedure: 1, calculate the concentrations from analytical signals by calibration; 2, determine the kinetic parameters by non-linear parameter estimation. Recently, new approaches have been suggested for step 1. The idea of these approaches is to replace the calibration step by curve resolution techniques that take into account the special nature of kinetic data [6]. Commercial software, ConcIRT, has also been developed for this approach [7]. As already stated, the implicit calibration approach combines the two steps into a single one and can also be considered as a special case of curve resolution where also the stoichiometry and differential aspects are taken into account. Intuitively, it is clear that the implicit calibration approach uses more prior knowledge than curve resolution and should give better results if the prior knowledge (the kinetic mechanism) is correct.

It should be noted that there is a major difference between the curve resolution and implicit calibration approaches. The former is a valuable tool when the kinetic mechanism is not known, helping us in finding the possible reaction schemes. The latter, in turn, provides a tool for finding a quantitative kinetic model when the kinetic mechanism is known.

2.3. Combining kinetics estimation and calibration (implicit calibration)

2.3.1. A simple example. The basic idea and possible problems are best illustrated by considering the reaction $A \rightarrow B$. Owing to its simplicity, it is easy to see the relevant points. The task is to estimate $k$ from uncalibrated spectroscopic signals. The reaction gives us the simple model

$$C_A(t) = C_{A_0}e^{-kt}$$ (2)

Assume that we can measure a single uncalibrated signal (absorbance) $y$ with small errors, i.e.

$$y = bC_A + e$$ (3)

Combining Equations (2) and (3) at measurement times $t_0, t_1, \ldots, t_n$ yields the following model for calibration and kinetics:
Now it is quite obvious that both unknown parameters \( b \) and \( k \) could be determined from the concentration profile of a single batch with known \( C_{A0} \). (By a batch we mean a reaction with one set of initial values where the measurements have been made at discrete time points \( t_0, t_1, \ldots, t_n \).) If the value of \( b \) were known, \( C_{A0} \) could be unknown, but they could not be unknown simultaneously unless several batches were measured.

Now suppose that the signal were an absorbance at a single wavelength of a spectrum with large baseline errors varying between different time points. In this case the combined error structure, calibration and kinetic model would be

\[
y_i = a_i + bC_{A0}e^{-kt_i} + e_i, \quad i = 0, 1, \ldots, n
\]

where \( a_i \) stands for the baseline error at time \( t_i \). It is obvious that we could not estimate the unknowns, since this system of equations has more unknowns than equations. In the case of large baseline errors, if we could measure two absorbances \( y_{i1} \) and \( y_{i2} \) at different wavelengths, the situation would change. Namely, taking the difference of these two would eliminate the baseline term \( a_i \). Similarly, a linear baseline drift error in the absorbances could be eliminated by taking second-order differences. This, in turn, would require at least three responses per observation. It is obvious that the more complicated the error structure, the more measured response variables we need per observation.

Thus, in principle, estimation of the kinetic parameters is possible, though the spectra would be contaminated by errors. Unfortunately, this is not always the case in practice, since the errors are confounded with the effect of changing values of the unknown kinetic parameters. This explains why pretreatment of spectra by various spectral corrections sometimes gives better results than ‘autocorrective’ methods such as PLS or PCR with ‘extra’ dimensions. We consider these aspects in more detail in subsequent sections, but first let us illustrate the calibration alternatives by this simple example.

Suppose we knew the parameters \( C_{A0} \) and \( k \) of the kinetic model. With only one measured response (absorbance) this leads to ordinary straight line calibration. If we denote \( C_{A0}e^{-kt} \) by \( C_i \) and consider these concentrations errorless, the natural calibration model is \( y_i = a + b C_i \), where \( a \) represents a possible common baseline error for every observation. In the implicit calibration, however, we do not know the kinetic parameters (here \( k \)); instead, what we have is an initial guess for them. Therefore the concentrations \( C_i \) contain errors, though of a systematic nature. Besides, the precision of measured values \( y_i \) can be very high. Thus a reciprocal calibration model \( C_i = \alpha + \beta y_i \) might be equally justified. Later we will call models similar to \( y_i = a + b C_i \) direct implicit calibration and models similar to \( C_i = \alpha + \beta y_i \) indirect implicit calibration.

All cases above lead to the simultaneous estimation of kinetic and calibration parameters, i.e. calibration is implicit in the overall estimation task. In the next section these ideas are extended to a more general case.

2.3.2. The general case. Let us now consider the general case where the kinetic model consists of more than one species, \( (C_i \) is a vector) and the measurements are spectra or chromatograms \( (y_i \) is a vector too). Thus the kinetic model is a multiresponse model relating time and unknown kinetic parameters to concentrations. The related calibration problem is also a multivariate calibration problem. The algorithm to solve the general case is identical to the algorithm in section 2.1, except for step 3.

As in the simple example above, the basic difference with respect to the traditional kinetics estimation, with measured concentrations, is the introduction of unknown calibration parameters into
the problem. Analogously with the simple case, there are two alternatives to solve the calibration problem: either the concentrations or the absorbances are considered independent variables. In the general case it is not efficient to estimate all unknowns, both the parameters of the calibration model and the parameters of the kinetic model, in a single non-linear estimation task. Owing to the linear nature of the calibration model and the typically large number of unknown calibration parameters, it is more reasonable to solve the linear problem inside the iteration of the kinetic parameters. The implicit calibration approach could also be seen as curve resolution with kinetic constraints, but in our approach it is more natural to think that the parameter estimation is the ‘master’ and the calibration the ‘slave’, not vice versa.

In the sequel, if the absorbances $y_i$ are treated as independent variables in the calibration step, we call it the indirect mode of calibration (as the fit is indirectly related to the measurements). If the concentrations $C_i$ are treated as independent variables in the calibration step, we call it the direct mode of calibration (as the fit is directly related to the measurements).

Furthermore, we might want to set some parameters to zero or to impose other restrictions on the solution vector: positivity, smoothness, linear equations and so on. This gives us a multitude of calibration alternatives. It is impossible to test all alternatives, but we have tried a few of the most appealing ones. The basic choices to be made at step 3 are:

- mode of calibration (direct or indirect);
- calibration method (OLS, PCR, PLS, RR, etc.);
- constraints (positivity, smoothness, stoichiometry, etc.).

Although, in general, the least squares criterion is the most common measure of goodness of fit in all regression-type parameter estimation problems, it is a bit problematic in implicit calibration. In indirect calibration the components whose concentrations are small have a negligible effect on the sum of squares (SS) of the residuals. Suitable weights would be a natural solution, but finding such weights is not easy, since concentrations are not measured. We have chosen the geometric mean of the $R^2$ values of the component profiles. The arithmetic mean could be used as well, though in our simulations the geometric mean worked slightly better. The important aspect is that one uses individual $R^2$ values for each response, not the overall $R^2$ value obtained by treating responses as a single long vector. The latter is actually equivalent to using residual SS.

If the dimension of the calibration model is too low or some of the components do not give any signal in the measurement, it is possible to get negative $R^2$ values. They are simply neglected in computing the geometric mean. Because in the direct calibration mode the model is fitted to the absorbances, an analogous scaling/weighting does not exist. Therefore in direct calibration we have used SS as the measure of goodness of fit.

There are two special cases that allow explicit solutions for the calibration step: the case of a known calibration model and the case of known pure component spectra.

2.3.3. The calibration model is known. This case is practically similar to the traditional kinetics estimation. The calibration model simply means a change of variables (and units) in the model or in the data, depending on the mode of calibration. In the indirect calibration mode the spectral data are transformed into concentration data according to the calibration model (in this case typically a PLS model) prior to parameter estimation. In the direct calibration mode the concentrations are transformed into spectral measurements according to the calibration model (in this case typically the Lambert–Beer model) at each iteration step. The main point is that calibration does not bring any new unknowns to the kinetic parameter estimation.
2.3.4. The pure component spectra (or chromatograms) are known. If the pure component spectra are known, the calibration actually becomes unnecessary. If we assume that the measured spectra are linear combinations of pure component spectra (the law of Lambert and Beer), each measured spectrum gives a system of equations

\[ y_{\text{meas}} = Y_{\text{pure}} c \]  \hspace{1cm} (6a)

where \( y_{\text{meas}} \) is one observed spectrum as an \( n_{\text{wl}} \times 1 \) column vector, \( Y_{\text{pure}} \) is an \( n_{\text{wl}} \times p \) matrix of the known pure component spectra (or chromatograms) and \( c \) is the \( p \times 1 \) vector of unknown concentrations in one sample. Here \( n_{\text{wl}} \) is the number of wavelengths (or retention times) and \( p \) is the number of components.

In this case it is natural to demand positivity of the solution. As the measurements are made during a batch reaction, it is equally natural to demand the solution to satisfy the stoichiometric equations of the reaction in question. This requirement imposes that the initial concentrations have to be known or otherwise the initial spectra should be subtracted from all spectra (as is done in Reference [7]). This, in turn, would hinder the use of positivity constraints. There is, however, a better way to solve this problem.

For measurements taken from a complete batch of \( n \) measurements, the system of equations can also be presented as a matrix or multiresponse system of equations

\[ Y_{\text{meas}} = Y_{\text{pure}} C \]  \hspace{1cm} (6b)

where \( Y_{\text{meas}} \) is the \( n_{\text{wl}} \times n \) matrix of observed spectra, \( Y_{\text{pure}} \) is an \( n_{\text{wl}} \times p \) matrix of the known pure component spectra and \( C \) is the \( p \times n \) matrix of unknown concentrations in one sample. Again it is natural to demand positivity and stoichiometry from the solution matrix. Now the stoichiometric constraints can be imposed on the time differences of the solution matrix. Therefore the initial concentrations need not be known and positivity can still be demanded. Of course, if the initial concentrations are known, they can be put as equality constraints into Equation (6b). After obtaining the concentrations, the estimation of the kinetic parameters is done according to the algorithm in Section 2.1.

In both cases (Equations (6a) and (6b)) the constraints can be given as exact constraints or approximate constraints. In the former case the system is solved by a Lagrange technique (see e.g. Reference [8]) and in the latter case by generalized ridge regression (GRR [9]; a short description is given in Section 5). For ideal data, both techniques can give the correct solution exactly. We have chosen GRR, whose use in this context is a novel approach to our knowledge. In Section 5.4 we will see how this method works in practice with non-ideal data.

Next we consider some other important aspects of kinetic parameter estimation.

2.4. Data preprocessing

In kinetic studies both on-line and sample-based analytical methods can be used. Traditional kinetics estimation requires off-line calibration, since the original measurement signals (typically absorbances) must be converted into concentrations. Many analytical instruments use a simple straight line calibration (e.g. HPLC), but also multivariate methods are used, especially with on-line instruments. In addition, with spectroscopic data one must decide whether to use different spectral corrections (MSC, orthogonal PLS, differential spectra, smoothing, etc.) or a calibration technique with ‘internal corrections’. For example, PLS regression can automatically correct for several types of mild errors in the spectra if the calibration data are representative with respect to the spectral errors.
2.5. Basis of fitting

In the traditional approach the goodness of fit is calculated in concentration units. Thus an individual residual error has the form

$$\frac{y_{\text{measured}}}{C_0} - \frac{y_{\text{modelled}}}{C_1}$$

where the former term is constant and the latter is a function of the unknown kinetic parameters. In the case of the implicit calibration approach the situation is different and we have basically two alternatives: fitting in concentration units or in absorbance units. This could also be expressed in other words, since the former means fitting in measured variables and the latter in non-measured state variables. Instead of absorbance, we could have any analytical signal, and instead of concentrations, we could have any state variable of the kinetic system, mole fraction, mass fraction, temperature, enthalpy, etc. It should be noted that in the implicit calibration approach the concentrations are calculated by the kinetic model, not measured.

3. SIMULATION STUDIES

It is clear that the validity of results from simulation studies is always limited. Still, simulation is a valuable tool to check whether a method works at all. It provides an answer to the question of whether to continue with real examples. It also provides an opportunity to test new fields of applications before any experimentation, to test the sensitivity to different error structures, different experimental designs and other influential aspects.

In our previous study [1] we simulated an equilibrium reaction with spectroscopic-type data with baseline drift and white noise-type errors. Now we have studied the simplest reaction \( A \rightarrow B \). We also studied the second-order reaction \( A + B \rightarrow C \), which, however, behaved almost identically. The motivation to study these reactions is that they were used in illustrating the basic ideas in Section 2. Moreover, it is easier to understand the results with the simplest cases. In the present paper we have tested the alternatives of implicit calibration with HPLC-type chromatographic data. Special

Figure 1. Simulated overlapping chromatographic peaks of species A and B with errors in baseline and peak shapes.
emphasis is given to identifiability and choice of initial concentrations of the batches. The structure of
the simulations is given below and the results of implicit calibration are given in Section 5.

3.1. $A \rightarrow B$

As was seen in the previous section, in this reaction it easy to see from the structure of the model that
kinetic parameter estimation is possible with well-separated peaks. Some experimentation showed
that it really does not matter what kind of calibration technique is used. Therefore we proceed directly
to the more interesting case of overlapping peaks for $A$ and $B$. In addition, we introduce not only
white noise but also errors in the peak shapes (location, height, width and skewness) and baseline drift
errors. The relative error in shape parameters was set to 10 per cent. The baseline-type errors are less
interesting in the sense that quite effective methods exist for correcting for them. The simulated
chromatographic data of a single batch are shown in Figure 1.

3.2. $A + B \leftrightarrow C + D$

This basic equilibrium reaction is more interesting than simple irreversible reactions. It was studied
already in Reference [1] with simulated spectroscopic data and two real cases. Here we have tested it
with simulated chromatographic data with overlapping peaks and errors in peak shapes. Typical
simulated chromatograms are shown in Figure 2. They are linear combinations of four unimodal
chromatograms with different absorbance coefficients and consequently different peak heights. The
peak of $A$ overlaps the peak of $C$, and likewise for $B$ and $D$. Peak shapes vary randomly between
measurement times with 10% relative error in the shape parameters (location, width, height and
skewness).

Before examining the simulations more closely, let us try to determine whether the problem is
solvable in principle. If we could find an explicit and simple enough solution like Equation (5), the
question would be easily answered. For simplicity, let us denote the concentration of $A$ by $\alpha (t)$ or
simply by $A$. The kinetics leads to the simple differential equation

\[ \frac{dA}{dt} = k \]
\[
\frac{dA}{dt} = -k_1A(B_0 - A_0 + A) + k_2(C_0 + A_0 - A)(D_0 + A_0 - A)
\]

where \(A_0, B_0, C_0\) and \(D_0\) stand for the initial concentrations. The solution of the general equation is the following rather formidable set of equations:

\[
A(t) = \frac{-1}{2} \left( \frac{1}{k_1 - k_2} \right) \left[ \beta + 2A_0(k_2 - k_1) + \alpha \tan \left( -\frac{\tan^{-1}(\beta/\alpha)}{2} + \frac{1}{2} \alpha t \right) \right]
\]

where

\[
\alpha = \sqrt{\alpha_1 + \alpha_2}
\]

\[
\alpha_1 = (-4C_0D_0 - 2C_0A_0 - 2A_0D_0 - 4B_0A_0 - 2B_0D_0 - 2B_0C_0)k_1k_2
\]

\[
\alpha_2 = -(A_0 + B_0)^2k_1^2 - (C_0 + D_0)^2k_2^2
\]

\[
\beta = (A_0 + B_0)k_1 + (C_0 + D_0)k_2
\]

An interesting detail in this solution is that although the argument of the tangent is a complex number, its value is real. Consequently, it is applicable only for software that can handle complex numbers.

The other components can be derived from this by stoichiometric relations. Although Equation (8) is not very handy for solving the kinetic differential equations (they are solved more easily numerically; see Section 5.2), it is useful for showing how the kinetic parameters relate to each other. In the case of a single uncalibrated signal, i.e. \(y = bA(t)\), a thorough inspection shows that all unknowns \(b, k_1\) and \(k_2\) are estimable. Because the estimability is not so obvious with this much more complex equation, it was also verified by calculating the rank of the Jacobian matrix of the calculated observation vector \([bA(t_i)]_{i=1,2,...,n}\) with respect to \(b, k_1\) and \(k_2\). The rank is 3 when evaluated with the known values of \(b, k_1\) and \(k_2\). This shows that the model is identifiable, i.e. all parameters are estimable [10]. Of course, at least three time points are needed (\(n \geq 3\)).

As in the previous simpler case, if more than one signal is measured per observation, some errors of a systematic nature (baseline drifts, etc.) can be tolerated.

4. EXPERIMENTAL

4.1. Esterification of ethanol with acetic acid

Here we consider the equilibrium reaction \(A + B \leftrightarrow C + D\) with \(A \equiv \) ethanol, \(B \equiv \) acetic acid, \(C \equiv \) ethyl acetate and \(D \equiv \) water. This reaction was studied already in Reference [1] with highly non-ideal data. Now we use data presented in Reference [11]. The temperature and the amount of catalyst were different from the values used in Reference [1] and therefore the results cannot be straightforwardly compared.

The details of these experiments are published in Reference [11], so only some essential points are given here. The experiments consisted of nine batches with different initial concentrations of ethanol and acetic acid. Initial concentrations of ethyl acetate and water were zero in all batches. Samples and IR spectra were taken at 5–10 min intervals. In addition, a separate set of mixtures of the four chemical components was prepared for developing a multivariate calibration model. Two different estimates of concentrations of the samples were available: chromatographic analyses and values.
given from the spectra by a calibration model. Therefore it was possible to make reliable comparisons between traditional methods and the implicit calibration approach.

4.2. **Esterification of methanol with formic acid**

Here we consider the same equilibrium reaction \(A + B \leftrightarrow C + D\) with \(A \equiv \text{methanol}, B \equiv \text{formic acid}, C \equiv \text{methyl formate} \text{ and } D \equiv \text{water.}\) From the implicit calibration point of view this is a more interesting case because of the faster reaction. Traditional analyses are difficult to carry out, because the reaction starts immediately already at room temperature. The spectroscopic data were produced at Lappeenranta University of Technology with the same instrumentation as in Reference [11].

This reaction was also studied at Kemira Agro Oyj research centre, where a reliable kinetic model was developed by sophisticated on-line chromatography and traditional non-linear parameter estimation. Thus we could compare this kinetic model with the one obtained by implicit calibration. A more detailed comparison of these data with the data produced at Kemira Agro Oyj will be given in a subsequent paper.

5. **RESULTS**

In all cases the methods tested were: I, indirect calibration by PCR; II, indirect calibration by PLS; III, direct calibration by GRR. Therefore, in the sequel, PCR and PLS always refer to indirect calibration and GRR to direct calibration. GRR [9] stands for the following general least squares problem:

\[
\min_{X \geq 0} \left\{ \|AX - Y\|^2 + w_1 \|B_1 X\|^2 + w_2 \|XB_2\|^2 \right\}
\]

The matrices \(B_1\) and \(B_2\) are linear constraints for rows and columns of the solution matrix \(X\). Typically they are differential operators related to the smoothness of the solution. If they are identity matrices, and positivity is not demanded, the solution is the ordinary RR solution. The weights \(w_1\) and \(w_2\) are typically chosen by cross-validation, but in the implicit calibration this would be too slow.

With PCR and PLS, several dimensions were tested. In theory, the dimension (rank) could be determined from the number of independent reactions, the number of absorbing species and the number of independent sets of initial concentrations [12]. In practice, however, the rank cannot be chosen in that way for at least two reasons. First, the spectra are always contaminated with systematic errors caused by non-idealities. Second, the relevant rank is not the rank of the spectral matrix but the rank of the Jacobian matrix of the function to be minimized (or maximized) with respect to the kinetic parameters.

With GRR, different ridge weights were tested, with and without positivity constraints. Because it is not reasonable to report all possible combinations, we discuss some typical results and comment on the general behaviour.

5.1. **Simulation studies of the reaction \(A \rightarrow B\)**

Let us recall that the simulations were based on unimodal chromatograms with errors in peak location, height, width and skewness. White noise and baseline drifts were added too. The studies based on simulated spectra are discussed in Reference [1].

The simple reaction \(A \rightarrow B\) posed no difficulties and therefore we discuss it only briefly. In the models of \(A \rightarrow B\) there is only one unknown kinetic parameter and thus the least squares minimum or \(R^2\) maximum can be easily found graphically. The graphical determination also allows immediate
Figure 3. Geometric mean of $R^2$ values for concentration profiles of A and B as a function of rate constant $k$. Implicit calibration with PCR with two dimensions (top left), four dimensions (top right), five dimensions (bottom left) and eight dimensions (bottom right).

Figure 4. Residual sum of squares as a function of rate constant $k$. Implicit calibration with GRR with positivity constraints and ridge weights $w_1 = 0.1$, $w_2 = 0.1$ (top left), $w_1 = 1$, $w_2 = 0.1$ (top right), $w_1 = 0.1$, $w_2 = 100$ (bottom left) and $w_1 = 1$, $w_2 = 100$ (bottom right).
inspection of the model identifiability, i.e. estimability of all unknowns. A clear peak or trough is a sign of a well-identified model.

All tested methods worked well even with quite high error levels (for direct calibration, baseline drift-type errors must be corrected prior to estimation). Moreover, all methods fail if the error level is high enough. Some examples are shown in Figures 3 and 4. In both cases the true value of the rate constant $k$ is $0.35$.

The results with PLS were almost identical, but an interesting feature was found in calibration with PLS: sometimes, with higher dimensions, strange peaks appear in the $R^2$ curves. This seems to be numerical noise and comes from the fact that without any errors the $X$ and $Y$ matrices in the calibration are both only of rank 2. The results by direct calibration with positivity and smoothness-constrained GRR with different values of ridge weights $w_1$ and $w_2$ are shown in Figure 4. The matrix $B_1$ was an identity matrix and the matrix $B_2$ was a matrix of fourth-order differences corresponding to fourth-order derivatives. Note the different $y$-scale, i.e. SS instead of $R^2$. In this case the chromatograms were baseline-corrected, because direct methods cannot automatically correct such errors.

It can be seen that increasing the weight $w_2$ does not change the SS curve much. The weight $w_1$ has some influence. In general, the method is not too sensitive to the wrong dimension or ridge weights (for more discussion on selecting the ridge weights, see Section 5.2).

5.2. Simulation studies of the reaction $A + B \leftrightarrow C + D$

Next we consider more closely some of the numerous alternatives in steps 3 and 4 of the implicit calibration algorithm in the case of the equilibrium reaction. Now there are two unknown kinetic parameters. Again the least squares minimum or $R^2$ maximum can be easily found graphically by contour surfaces. The graphical representation also provides more information about the identifiability than mere numerical optimization.

5.2.1. The mode of calibration. Let us first consider the most difficult case, a single batch. The methods tested were indirect calibration with PCR and PLS, direct calibration with GRR, and some preliminary tests with unimodal regression. Several simulations were carried out in order to estimate the variability of the results. The true values for the kinetic parameters in these simulations were $k_1 = 0.35$ and $K = 0.55$, where $k_1$ is the forward rate constant and $K = k_2/k_1$ is the inverse of the equilibrium constant of the reaction

$$A + B \xrightarrow{k_1/k_2} C + D$$

Indirect calibration gave rather poor results. Figure 5 illustrates a typical result.

Although the optimal estimates with maximal $R^2$ value ($\hat{k}_1 \approx 0.38$ and $\hat{K} \approx 0.32$) are not very far from the true values, the shape of the contour lines suggests low reliability. Indeed, that is the case, and repeated simulations showed that sometimes the estimates can be quite far from the true values, e.g. we had a case where $\hat{k}_1 \approx 0.95$ and $\hat{K} \approx 0.053$. Changing PCR calibration into PLS calibration does not change the situation, and the only, rather obvious, difference seems to be that in PLS the number of dimensions needed is typically one less than that in PCR.

Direct calibration by GRR with positivity and smoothness constraints succeeds slightly better. In Figure 6, for instance, the minimum residual sum of squares (SS) is obtained with $\hat{k}_1 \approx 0.34$ and $\hat{K} \approx 0.53$, which are very close to the true values.
The contours of SS show that even in this case the equilibrium constant $K$ is not well determined. Yet, repeated simulations showed that direct calibration never failed completely with the noise level of these simulations.

We also did some preliminary tests with unimodality constraints using the code of Bro and Sidiripoulos [13]. It worked well but was very slow for implicit calibration purposes. We also tried our own version of GRR with additional unimodality constraints that speeded up the computations three fold to five fold, but the present code must be tailored separately for each problem. There is slight evidence that the most restrictive method, i.e. direct calibration by unimodal positive regression, gives the most uniquely determined parameters, but this topic has to be studied further.

5.2.2. The measure of goodness of fit. In Section 2.3.2 we discussed the problems of using residual sum of squares as the measure of goodness of fit in indirect calibration. As an example, Figure 7 shows the SS contours for the same data and the same calibration method (PCR) as in Figure 5 (note the different scale for $K$).

The change is significant and $K$ is not at all estimable. Interestingly, the simulation studies also showed that $R^2$ is a systematically better measure for goodness of fit than SS in indirect calibration. The main reason for this is that the $R^2$ value is a mean of individual $R^2$ values of the responses and therefore gives more weight to lower concentrations, which are quite significant in determining the kinetic parameters. In direct calibration such a weighting problem does not exist and the SS criterion works quite well, as can be seen from Figure 6.

5.2.3. Weighting. In the implicit calibration approach we can have two different kinds of weights: weights for residuals in calculating the criterion for goodness of fit after calibration, and weights used
in the calibration. As already noted, the $R^2$ criterion provides a kind of natural weighting scheme for the former. For the latter, giving more weight to the known initial concentrations of each batch is a natural demand. This improved the results in all simulations. In the real data cases the weighting of initial concentrations seemed rather to be a necessity for obtaining good results.

Figure 6. Contour lines of residual sum of squares. Implicit calibration with GRR with positivity and smoothness constraints. The $X$-axis represents the forward rate constant $k_1$ and the $Y$-axis represents the equilibrium constant $K$.

Figure 7. Contour lines of residual sum of squares for concentration profiles of A, B, C and D. Implicit calibration with PCR with four dimensions. The $X$-axis represents the forward rate constant $k_1$ and the $Y$-axis represents the equilibrium constant $K$. 

5.2.4. Choosing the right dimension or ridge regression weights. In indirect calibration, using the right dimension in PCR or PLS is an essential point. If the dimension is too low, the calibration fails even with the correct values of the kinetic parameters; on the other hand, if it is too high, the calibration fits into concentration profiles given with any values of the kinetic parameters. With errorless data the right dimension for PCR calibration is determined by the number of independent reactions and independent initial conditions of the different batches. For PLS the situation should be, at least approximately, the same. Errors in peak shapes and in the baseline bring up a need for extra dimensions, and the theoretically correct dimension cannot be applied. Our experience shows that the correct dimension is such that a dimension which is one lower than the correct dimension drastically reduces the SS or $R^2$ criterion. Therefore finding the right dimension by trial and error is not so difficult after all.

The direct calibration seems to be less sensitive to the ridge weights than the indirect calibration is to the dimension of the calibration model. Because the model itself, being smooth in time, forces the solution to be smooth row-wise, the matrix $B_1$ can be an identity matrix. It is not so clear that the solution would be smooth column-wise and therefore we have tested differential matrices of different orders. However, the simulations show (see e.g. Figure 4) that the solution is not sensitive to column-wise smoothing. Moderate weighting seems to stabilize the solution, otherwise the positivity constraints seem to be the most essential constraints in direct calibration.

5.2.5. Design of experiments. It is well known from the theory of non-linear estimation [5] that experimental design has a great influence on the reliability of the estimates. Therefore we have made some preliminary studies on this topic by simulations with different experimental conditions. Since we have limited our studies to isothermal experiments, design variables are measurement times and initial concentrations.
How the time points are chosen matters, because this has an effect on the confidence regions of the kinetic parameters. As an example, let us consider the equilibrium reaction where \( t_i, i=1,2,3 \) is \([0 \ 0 \ 3 \ 4], b=1, k_1=0.1, K=1\) and the initial concentrations are \([1 \ 1 \ 0 \ 0]\). This gives a Jacobian whose rank is 3 but whose condition number is 2022.5, which means that confidence regions will be

Figure 9. Contour lines of geometric mean of \( R^2 \) values of responses of implicit calibration with principal component regression of two batches with initial concentrations \([1 \ 1 \ 0 \ 0]\) and \([0 \ 0 \ 1 \ 1]\). The X-axis represents the forward rate constant \( k_1 \) and the Y-axis represents the equilibrium constant \( K \).

Figure 10. Contour lines of geometric mean of \( R^2 \) values of responses of implicit calibration with principal component regression of two batches with initial concentrations \([1 \ 1 \ 0 \ 0]\) and \([0 \ 0 \ 1 \ 1]\). The X-axis represents the forward rate constant \( k_1 \) and the Y-axis represents the equilibrium constant \( K \).
‘banana-shaped’ as in Figure 7. If the time points are \([0 \ 30 \ 400]\), the condition number is 1.8059, which is rather optimal. If the time points are \([30 \ 40 \ 50]\), the condition number is numerically infinite, the rank of the Jacobian is 1 and consequently the model is no longer identifiable.

On the other hand, if the measurements are made by on-line instruments, we can have a dense enough set of time points. Therefore, although choosing optimal measurement times is often important in traditional kinetic parameter estimation, it is seldom worth bothering with in the implicit calibration approach.

Next let us consider the choice of initial concentrations in different batches as well as the number of these batches. We shall illustrate this in an example with two batches with different initial conditions, keeping the relative error of peak shapes and noise on the same level as in previous examples. Figure 8 shows the \(R^2\) and SS contours of the indirect PCR solution in the case where the initial concentrations of A, B, C and D are \([1 \ 1 \ 0 \ 0]\) and \([1.5 \ 0.5 \ 0 \ 0]\). Figure 9 shows the case where the initial conditions are \([1 \ 1 \ 0 \ 0]\) and \([2 \ 3 \ 2 \ 3 \ 2 \ 3 \ 2 \ 3]\).

The difference is not big, but the initial conditions \([2 \ 3 \ 2 \ 3 \ 2 \ 3 \ 2 \ 3]\) of Figure 9 give a slightly better shape for the \(R^2\) surface. However, a much more substantial improvement is made if the initial conditions of the two batches are \([1 \ 1 \ 0 \ 0]\) and \([0 \ 0 \ 1 \ 1]\), giving the really beautiful peak in Figure 10 (note the different spacing of contour levels). In practice, of course, it may be impossible to start from a mixture of products only. Note also that in the first two cases the rank of the concentration matrix containing both batches is 3 but in the last one is only 2. Thus a higher rank is not always better.

Determining the optimal initial conditions for a given number of batches is a special case in optimal design of experiments and will be studied more closely in our forthcoming papers.

5.2.6. Computational time. The computational time depends on the solution of the kinetic differential equations, on the solution of the calibration step and on the required number of iterations (or grid evaluations). In indirect calibration methods (PCR, PLS) the computational time of
calibration is usually negligible compared with the solution time of the differential equations. In direct calibration with GRR, and especially with unimodal regression, the computational time of the calibration can be considerable. All computations were made with Matlab® and Fortran subroutines. Using Fortran for solving the differential equations increases the speed threefold to fivefold. Actually, in the reaction \( A + B \rightarrow C + D \) the Fortran-based numerical solution is as fast as the solution based on the analytical algebraic equation in vectorized form. Thus in most cases it is not worth elaborating on finding complex analytical solutions, since effective numerical solutions are fast enough. Besides, analytical solutions are not available for more complex kinetics. For instance, the case \( A + B \rightarrow C + D \) with a strong enough acid, e.g. formic acid, requires an autocatalytic form of the rate equations when an explicit analytical solution such as equation (8) does not exist.

As a summary of the simulation studies here and in Reference [1] it can be stated that the implicit calibration approach works well for all kinetic mechanisms studied, with both spectroscopic and chromatographic types of data with overlapping peaks and moderate error levels. Noise and baseline drift-like errors deteriorate the performance less than errors in peak shapes. In addition, these errors can quite often be corrected prior to kinetic parameter estimation.

5.3. Experimental studies: esterification of ethanol with acetic acid

This was a good test case, because the reaction is rather slow and therefore traditional methods can also be applied for comparative purposes. In this case we could make a comparison of implicit calibration results with (1) concentrations given by the on-line spectra applied to an off-line PLS model and (2) concentrations from HPLC analyses of samples taken during the reaction. Actually there is not much difference whether one uses the PLS-predicted or the chromatographically analysed data, since the off-line calibration predicted the results of the HPLC analyses very well. For comparative purposes the concentration data were used for traditional kinetic parameter estimation.
which yielded values that were within the confidence limits given by a software package for mathematical modelling and estimation, MODEST® [14].

At the first trial, only direct calibration with GRR (with positivity and smoothness constraints) gave reasonable, though not very good, contour lines. Next we tried off-line spectral corrections (fourth-
order baseline and MSC correction). This gave approximately the same results as the traditional kinetic parameter estimation, though the equilibrium constant $K$ was poorly determined. Then we turned back to the original uncorrected data but gave more weight in calibration for the initial concentrations. This is a natural procedure, since the initial concentrations are given by weighing known amounts of A and B, not by chemical analyses. This improved the results of all three methods tested: indirect PCR and PLS and direct GRR. We discovered some interesting features that were not observed in the simulations.

PCR with three or fewer dimensions did not fit well enough for any purposes. With four dimensions the result is as shown in Figure 11. In PCR, increasing the number of dimensions did not improve the situation from this; rather, it became worse.

Figure 15. Concentration profiles of methanol, formic acid, water and methyl formate. Solution by implicit indirect PCR calibration (broken line) and by GRR from measured pure component spectra with positivity and stoichiometric constraints (full line). Time in hours on X-axis and concentration on Y-axis.
PLS does not work at all with one dimension. With two dimensions the result is as shown in Figure 12, and with three dimensions the figure is very similar. From four dimensions onward the figures resemble very much those given by PCR. This suggests that one step in changing PCR dimensions is too ‘coarse’ in the sense that it changes from lack of fit to overfit between dimensions three and four. As a summary, in the present case, PLS worked slightly better than PCR.

Direct calibration by GRR works quite similarly (though not as well as) to PLS (Figure 13). If we compare these three methods, PLS and GRR give values for the kinetic parameters which are within the confidence limits of the traditional estimation by MODEST, whereas PCR gives too low a value for $K$. If one looks more closely at the three contour plots, it can be seen that the equilibrium constant is poorly determined by all methods, and the good results by PLS or GRR may be accidental. On the other hand, the rate constant is fairly well determined by all methods. This result matches well with the simulations, which showed that the equilibrium constant is more difficult to estimate. The simulations showed that performance could be greatly enhanced by proper selection of the initial concentrations. Although we have altogether nine batches, all of them have zero initial concentrations of ethyl acetate or water. Our next example will show how much improvement can be achieved by adding products into the initial conditions. According to the simulations, it would be even better to have batches with zero initial concentrations for ethanol and acetic acid.

5.4. Experimental studies: esterification of methanol with formic acid

This reaction is a more interesting case than the previous one in two ways. First, the kinetics is much faster. A set of known calibration mixtures for IR was also made (see Section 4.2), but the reaction turned out to be too fast to get a reliable calibration model. Second, it is extremely difficult to take stable samples for chemical analyses, e.g. HPLC, for traditional kinetic estimation purposes. An alternative, however, is to use the measured pure component spectra as discussed in Section 2.3.4. Moreover, the kinetics of this reaction had been determined at several temperatures at Kemira Agro Oyj by an advanced on-line HPLC system specially designed for fast reactions. A detailed description
of this system will be given in a forthcoming paper. The values of the kinetic parameters at a reference temperature of 393.15 K are $k_1 = 0.11$ and $k_2 = 0.20$. They turned out to be very close to the values given by implicit calibration. Note the different parametrization: $k_1$ and $k_2$ instead of $k_1$ and $K$. The values correspond to rate constants at the reference temperature; the model converts them to rate constants at the temperature of our experiments according to the Arrhenius equation.

![Figure 17. Contour lines of geometric means of $R^2$ values. Implicit calibration by PLS. The X-axis represents the forward rate constant $k_1$ and the Y-axis represents the backward rate constant $k_2$.](image1)

![Figure 18. Contour lines of residual sum of squares. Implicit calibration by GRR with positivity and smoothness constraints. The X-axis represents the forward rate constant $k_1$ and the Y-axis represents the backward rate constant $k_2$.](image2)
Let us first consider the alternative of using pure component spectra. The concentrations are solved by constrained regression (GRR) with pure component spectra as the 'X matrix' and positivity and stoichiometry as constraints on the time differences of the solution matrix, i.e. concentration matrix. The spectrum of water was out of measurement range at two separate wavelength ranges which were not used in the calculations. Figure 14 shows the four pure component spectra.

The measured spectra did not obey the law of Lambert and Beer very well. For instance, if the initial concentrations were not forced to the known values, they were far from the true values. In addition, without stoichiometric constraints the solution did not obey the stoichiometry of the reaction. Thirdly, some strange ‘wiggles’ can be seen on the concentration curves (Figure 15, batch 3). The solution is shown in Figure 15 together with the implicit calibration solution given by indirect PCR calibration. All batches fit rather poorly with the implicit calibration solution. The errors are of a systematic nature and it is obvious that kinetic parameter estimation using these concentrations would give quite different values from the implicit calibration method. However, in spite of the poor compatibility of the two solutions, the time when the equilibrium is reached is approximately the same for both methods.

Next let us consider the implicit calibration approach. In this case it did not pose any difficulties, and all calibration alternatives worked approximately equally well, though with slightly different optimal values of the kinetic parameters. Figures 16–18 show the contours of estimation by indirect PCR and PLS calibration and direct GRR calibration respectively.

PCR calibration gives $k_1 = 0.10$ and $k_2 = 0.21$, which are very close to the values obtained at Kemira Agro Oyj. PLS and GRR gave $k_1 = 0.09$, $k_2 = 0.18$ and $k_1 = 0.09$, $k_2 = 0.29$ respectively, which are quite close to the same values too. The values are within the confidence limits given by the MODEST software [14].

Since the solution matrix of GRR can be interpreted as the matrix of the pure component spectra, it is interesting to compare this with the measured pure component spectra. This is shown in Figure 19.

---

Figure 19. Measured and estimated pure component spectra. Estimation by direct calibration with GRR with positivity and smoothness constraints. Wavelength on X-axis and absorbance on Y-axis.
The result is not bad, but, especially in the case of water, systematic differences between measured and calculated spectra can be seen.

6. DISCUSSION

Modelling and simulation have become an essential part of industrial chemical process development, including process design, scale-up, control, safety and fault detection. Many software packages are available for both dynamic and steady state simulation and they include thermodynamic databases and thermodynamic properties estimation. Reaction kinetics is, however, often the bottleneck of reliable simulation studies and still quite often has to be experimentally determined. For complex or fast reaction mechanisms this is far from being a simple task. The proposed implicit calibration method seems to be able to overcome many difficulties: neither sampling nor calibration based on known mixtures or pure components is needed. The measurements correspond to the true state of the reaction at each time step. All intermediates can be detected provided that the instrument is sensitive to them. Thus reaction mechanisms that could not determined by traditional methods can be handled, and even in cases where traditional methods work, the implicit calibration approach can significantly speed up the kinetics estimation procedure.

Even in the case where the analytic signal is not sensitive to all chemical species, the kinetic parameters may be estimable. In indirect calibration mode this would typically induce negative $R^2$ values for some species, which should be omitted in calculating the geometric mean. However, a multiresponse system may identifiable by a subset of responses [15]. Owing to stoichiometric relationships, this is obvious with simple kinetics and known initial concentrations.

Nevertheless, a multitude of details that affect the results have to be settled. These include preprocessing of spectra, mode of implicit calibration, the multivariate calibration method itself, the measure of goodness of fit, use of prior knowledge in constraining the solution, weighting of measurements, the dimension of the calibration model or the weights for the constraints, and finally the design of the experiments. This is an awesome list, and if all choices were always to be done, the method certainly would not speed up the kinetics estimation. However, our studies show that some general guidelines can be given.

The most important element seems to be the design of the experiments. With well-designed experiments the method seems to be rather insensitive to the other choices in the list above. Therefore the design of kinetic experiments is one of the main topics that we shall focus on in our future research. At this point it can be stated that a good design is a set of batch reactions with a ‘representative’ set of initial concentrations. If the design is not so good, the simulation studies support the choice of the most restrictive calibration method. This, in turn, suggests the use of direct calibration mode with at least positivity constraints. If prior knowledge allows the unimodality assumption, then using it as a constraint may further improve the results. Furthermore, both simulation studies and real examples support the use of $R^2$ (i.e. geometric or some other mean of the $R^2$ values of different response curves) in indirect calibration mode and SS in direct calibration mode as the measure of goodness of fit. Strong baseline shifts should be corrected in connection with the direct calibration mode, but otherwise weighting of initial concentrations seem to be more important than spectral corrections.

An important topic, reliability of the estimates, needs a lot more research. The surface of the measure of goodness of fit gives some information about the reliability and clearly shows if the problem is not identifiable, but how to transform this information into confidence limits or how to visualize multidimensional surfaces is a complicated matter. The theory of non-linear parameter estimation gives some guidelines. Unfortunately, cases like ours are not much studied where the model contains ‘nuisance’ parameters of the implicit calibration or the estimation method is not
maximum likelihood or least squares. Bootstrapping and other alternatives will be tried out in our future studies.

7. CONCLUSION

In this paper we have studied alternative ways to estimate unknown kinetic parameters directly from on-line spectral or chromatographic measurements. The basic idea is to solve the calibration problem inside the iteration loop of the corresponding non-linear estimation problem. Simulation studies with very simple kinetics show why the method works in principle. They, together with two real examples of esterification reactions, show that the method also works in practice with non-ideal data. The other esterification reaction gives an example of a case where traditional methods based on simple sampling or off-line calibration with known mixtures fail owing to the fast reaction. Although many details remain partly open, the method has already proved to be a valuable tool for kinetics estimation, and some recommendations on how the problem should be solved can already be given.

ACKNOWLEDGEMENTS

We thank Mr Esko Tirronen from Kemira Agro Oyj for giving us the kinetic parameter values of the esterification of formic acid with methanol. We also thank Professor P. Paatero and his co-workers from Lappeenranta University of Technology for the on-line measurements of the esterification reactions in this paper.

REFERENCES