ELSEVIER

Chemical Engineering Science 61 (2006) 5404–5420

Chemical Engineering Science

www.elsevier.com/locate/ces

Incremental identification of kinetic models for homogeneous reaction systems

Marc Brendel[a,1], Dominique Bonvin[b], Wolfgang Marquardt[a,*]

[a]*Lehrstuhl für Prozesstechnik, RWTH Aachen University, D-52064 Aachen, Germany*
[b]*Laboratoire d'Automatique, EPFL, CH-1015 Lausanne, Switzerland*

Received 16 February 2005; received in revised form 10 March 2006; accepted 5 April 2006
Available online 3 May 2006

Abstract

An incremental approach for the identification of stoichiometries and kinetics of complex homogeneous reaction systems is presented in this paper. The identification problem is decomposed into a sequence of subproblems. First, the reaction fluxes for the various species are estimated on the basis of balance equations and concentration measurements stemming from isothermal experiments. This task represents an ill-posed inverse problem that requires appropriate regularization. Using target factor analysis, suitable reaction stoichiometries can then be identified. In a further step, the reaction rates are estimated without postulating a kinetic structure. Finally, the kinetic laws, i.e., the dependencies of the reaction rates on concentrations, are constructed by selecting the best model structure from a set of model candidates. This incremental approach is shown to be both efficient and flexible for utilizing the available process knowledge. The methodology is illustrated on the industrially relevant acetoacetylation of pyrrole with diketene.

Keywords: Chemical reactors; Kinetics; Stoichiometry; Identification; Inverse problem; Modeling

1. Introduction

Mathematical modeling of chemical and biochemical processes plays an increasing role in today's competitive industries. Such models are typically needed for various tasks including process design, process analysis, optimization of process conditions and in an increasing manner also for model-based control. The description of reaction kinetics often represents the most challenging part in the modeling of (bio-)chemical reactors. A reliable description is rarely available a priori. For example, it is well known that reaction kinetics cannot necessarily be derived even if the stoichiometries are known (Connors, 1990). In some cases, even the stoichiometric model of the reaction system under investigation is not fully known. A reliable kinetic model (i.e., a model including both stoichiometries and reaction kinetics) must then be identified from experimental data obtained in laboratory experiments or during process operation itself.

These identification steps are normally carried out in systems where the relevant phenomena can be observed in isolation, preferably not in interaction with other physical phenomena such as interfacial transfer processes. For the investigation of reaction kinetics in liquid phase, a stirred batch or semi-batch reactor is used in the majority of cases. Integral and differential methods are typically used to derive the kinetics (Froment and Bischoff, 1990; Holland and Rayford, 1989). Assuming some kinetic model structure, the unknown rate constants can be determined numerically or even graphically. Experimental conditions are chosen to support the analysis. In particular, if several reactions occur simultaneously, these methods require a suitable experimental strategy to separate the effects of the individual reactions in a sequence of experimental runs.

For complex reaction systems, dynamic parameter estimation problems are often formulated to estimate the unknown parameters (Bard, 1974). All known information is combined to produce a dynamic model of the experiment, which consists of several submodels. Depending on the process considered,

* Corresponding author.
E-mail address: marquardt@lpt.rwth-aachen.de (W. Marquardt).
[1] Present address: Degussa AG, Process Technology, 63457 Hanau, Germany.

0009-2509
doi:10.1016/j.ces.2006.04.028

such submodels may represent heat and mass balances, mass transfer models, thermodynamics and, crucial for reaction systems, stoichiometries and reaction kinetics describing how and to which extent the various chemical species interact. The model is then numerically integrated and fitted to experimental data. The data fit adjusts the unknown model parameters in a way to minimize the deviation between model prediction and the noise-corrupted measurement data. To this end, weighted least squares, maximum likelihood or Bayesian approaches are most commonly employed (Bard, 1974). Alternatively, an error-in-variables approach can handle errors in both dependent and independent data coordinates. If there is no unique model structure, but rather a set of candidate models to account for the chemical reactions, fitting is performed for each candidate, and the most prospective candidate model is identified by appropriate model discrimination techniques (Akaike, 1974; Stewart et al., 1998). We refer to this approach as *simultaneous model identification*. The method is capable of handling reaction systems with arbitrary complexity including simultaneous reactions and formal kinetics. Variable experimental settings such as a variable feed rate can be accounted for in the model. Commercial parameter estimation software is readily available in a number of implementations and the method leads to statistically optimal estimation of the unknown parameters.

However, there are some disadvantages involved with simultaneous model identification. If an incorrect model structure is assumed (i.e., the stoichiometric model or some of the kinetic laws), an erroneous overall model prediction is obtained. Furthermore, the model error is difficult to attribute to a particular submodel. If, on the other hand, a set of potential model candidates is available, parameter fitting must be performed for each model candidate. The complexity grows if candidates are available for several submodels. In conjunction with the computationally expensive dynamic parameter estimation, computational cost may become prohibitive. From a numerical perspective, suitable parameter initialization is often difficult, thereby giving rise to convergence problems. Most parameter estimation packages are tailored to a small amount of data, whereas modern optical measurement techniques such as IR (Alsmeyer et al., 2002) or Raman spectroscopy (Bardow et al., 2003) produce a vast amount of data which may overstrain the capabilities of the packages. In summary, the simultaneous identification approach is often not fully satisfactory to identify complex reaction systems at moderate effort.

Alternatively, the simultaneous identification problem can be decomposed into several subproblems. Motivated by the complexity of one-step identification of hybrid models, i.e., models consisting of both a physically motivated and a data-driven part, Tholudur and Ramirez (1999) have used a two-step approach for the identification of kinetics: reaction rates are first identified, assuming known curve characteristics, and are subsequently correlated with the independent state variables using a feed-forward neural net approximation. Van Lith et al. (2002) have combined an extended Kalman filter for the estimation of states and rates with subsequent fuzzy submodel identification. Chang and Hung (2002) correlate polymerization rates and known states using neural networks. Yeow et al. (2003, 2004) presented an approach to convert time–concentration data into concentration–reaction rate data and to perform algebraic regression on this data set. These approaches aim at decomposing the identification process in two steps to reduce complexity. However, the methods either need to postulate specific model assumptions, generally unknown in real systems or are restricted to individual reactions, which limits their general applicability. A more systematic, multi-step approach for the identification of hybrid reaction models has been proposed recently (Brendel et al., 2003). The stepwise identification of kinetic phenomena has also been examined by Bardow and Marquardt (2004a) in the context of the identification of diffusive mass transfer.

In this paper, a unifying *incremental identification* concept is presented for the stepwise identification of structured submodels in complex reaction systems. The advantages of simultaneous identification, such as general applicability, are largely retained. At the same time, the complexity is considerably reduced by problem decomposition, which results in a more efficient and robust analysis and supports the modeling process. The hierarchical structure of any process model is exploited for the stepwise identification of submodels. Contrary to existing work, the approach is applicable to arbitrarily complex homogenous systems.

The time-variant reaction fluxes for the various species are first estimated from noisy concentration data using the filter-based approach of Mhamdi and Marquardt (1999) that relies on material balances. In cases where the reaction stoichiometries are unknown, *target factor analysis* (TFA) is applied to identify a stoichiometric model (Bonvin and Rippin, 1990). Subsequently, the individual reaction rates can be calculated from the estimated reaction fluxes using the knowledge of reaction stoichiometries. The reaction rates and estimated concentration data are then correlated to either identify the unknown parameters in a given kinetic model or select a suitable model structure from a set of candidates. The approach is especially suited for nowaday's data-rich measurement techniques such as IR or Raman spectroscopy, where concentration data can be obtained almost continuously in situ. It is applicable to all ideally mixed reactor types operated under transient conditions. Continuous flow experiments with plug flow reactors can also be treated.

The paper is organized as follows: first, the principal ideas behind the incremental identification approach are presented in Section 2. Section 3 describes the base case when data are available for all species involved in the reaction network, while Section 4 extends the theory to the case of incomplete measurements. The approach is illustrated on the industrially relevant acetoacetylation of pyrrole with diketene in Section 5. Finally, the conclusions are summarized in Section 6.

2. The incremental identification concept

The incremental identification approach mirrors the steps taken when developing a model for a given process. For clarification, dynamic model development of an isothermal reactor

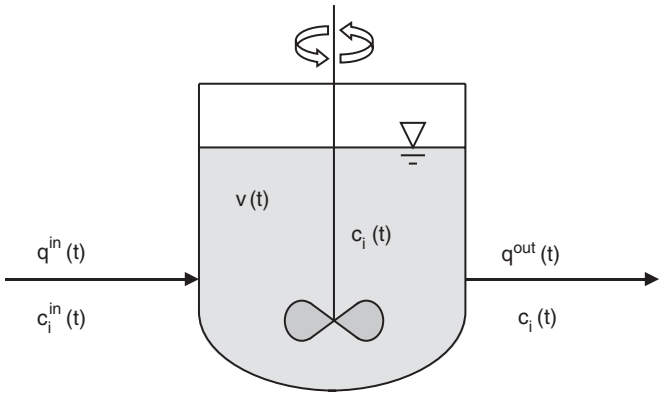

Fig. 1. Generic reactor schematic.

is considered first and the methodology is then transferred to model identification.

2.1. Reactor modeling

Consider the generic, ideally mixed, homogeneous and isothermally operated reactor depicted in Fig. 1, for which a dynamic model is to be generated. The mole balance equations are set up first

$$\frac{\mathrm{d}}{\mathrm{d}t}(v(t)\mathbf{c}(t)) = q^{\mathrm{in}}(t)\mathbf{c}^{\mathrm{in}}(t) - q^{\mathrm{out}}(t)\mathbf{c}(t) + \mathbf{f}^{\mathrm{r}}(t). \tag{1}$$

Here, $v(t)$ denotes the reactor volume at time t, $q^{\mathrm{in}}(t)$ is the volumetric feed rate and $\mathbf{c}(t)$ and $\mathbf{c}^{\mathrm{in}}(t)$ are molar concentration vectors[2] of the species in the reactor and the feed, respectively. In the balance equation, the reaction fluxes $\mathbf{f}^{\mathrm{r}}(t)$ for the various species, i.e., the net molar amounts consumed or produced per unit time by all reactions are unknown.

The stoichiometric relations describing the reaction network may be cast into the $n_R \times n_S$ stoichiometric matrix $\mathbf{N}$ defined as

$$\mathbf{N} = \begin{bmatrix} \nu_{11} & \cdots & \nu_{1n_S} \\ \vdots & \ddots & \vdots \\ \nu_{n_R1} & \cdots & \nu_{n_Rn_S} \end{bmatrix}, \tag{2}$$

where ν_{ji}, $j = 1, 2, \ldots, n_R$, $i = 1, 2, \ldots, n_S$, are the stoichiometric coefficients for the ith species in the jth reaction and n_S is the number of species involved in the n_R reactions.

Using the stoichiometric matrix $\mathbf{N}$, a constitutive equation is set up to express the reaction fluxes in terms of the n_R-dimensional reaction rate vector $\mathbf{r}(t)$,

$$\mathbf{f}^{\mathrm{r}}(t) = v(t)\mathbf{N}^{\mathrm{T}}\mathbf{r}(t). \tag{3}$$

The reaction rates can then be described by a set of constitutive equations as functions of the concentrations $\mathbf{c}(t)$ and the reaction parameters $\boldsymbol{\theta}$:

$$\mathbf{r}(t) = \mathbf{m}(\mathbf{c}(\mathbf{t}), \boldsymbol{\theta}). \tag{4}$$

[2] Unless otherwise indicated, a vector is defined as a column vector. In Eq. (1), the dimension of the vectors $\mathbf{c}(t)$, $\mathbf{c}^{\mathrm{in}}(t)$ and $\mathbf{f}^{\mathrm{r}}(t)$ is $n_S \times 1$.

This way, a dynamic model of the reactor has been set up, which is capable of predicting the reactor behavior over time, once the various terms in (1)–(4) are known.

2.2. Reactor model identification

It is assumed next that the reactor model is unknown and needs to be identified from experimental data. In particular, the number of occurring reactions, the stoichiometric model of the network and the kinetic laws describing the chemical conversion are unknown and need to be determined from data.

Measurements over time are supposed to be available for the reactor volume v (l) and the concentrations c_i (mol/l) of some or all species involved in the reaction network. A detailed analysis of the measurements required is given below. The flow rates q^{in} (l/min) and q^{out} (l/min) as well as the feed concentrations c_i^{in} (mol/l) are set by the experimental procedure and are therefore known as functions of time. Measurements taken are corrupted with noise.

The incremental identification of reaction kinetics is schematically depicted in Fig. 2. The method exploits the hierarchical model structure sketched in Section 2.1, providing stepwise identification of quantities as they are used in the modeling process. Incremental identification includes the following steps, as marked in Fig. 2:

(1) The reaction fluxes $\hat{f}_i^{\mathrm{r}}(t)$ are estimated individually from concentration c_i and c_i^{in}, volume and flow rate measurements using mole balances only (Eq. (1); Section 3.1).
(2) If the reaction stoichiometric model $\mathbf{N}$ is unknown, TFA is used to test possible stoichiometries and to determine the number of occurring reactions (Eq. (3); Section 3.2).
(3) With the stoichiometric information, the reaction rates $\hat{\mathbf{r}}(t)$ are then calculated from the fluxes $\hat{\mathbf{f}}^{\mathrm{r}}(t)$ (Eq. (3); Section 3.3).
(4) Kinetic laws are obtained by regressing the time-variant estimates of concentrations $\hat{\mathbf{c}}(t)$ and rates $\hat{\mathbf{r}}(t)$ with candidate kinetic structures (Eq. (4); Section 3.4).

In an adaptive modeling context (Marquardt, 2002), the incremental approach allows the utilization of as much information as can be safely provided by first-principle modeling or sound empirical approaches. The process of identification then reduces to modeling uncertainties, i.e., unknown parameters in a given structure or the model structure itself. For the identification of reaction kinetics, the approach permits to determine relevant reaction kinetics directly, independently of the other reactions, i.e., models for irrelevant reactions need not be included in the identification process.

Each of the identification steps provides additional information regarding the reaction system, which facilitates the selection of feasible model candidates for the following steps. If no suitable model structure can be established, some data-driven function approximation may replace the structured model (Brendel et al., 2003; Brendel, 2005).

Decomposition of the identification procedure results in a sequence of decoupled identification problems. Decoupling is

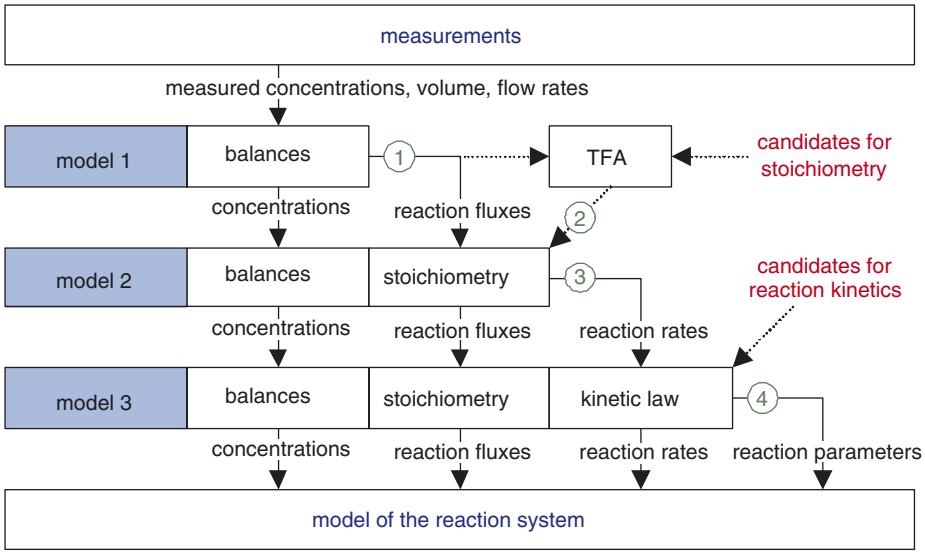

Fig. 2. Schematic of the incremental identification approach.

achieved both vertically and horizontally. Vertical decomposition results from exploiting the hierarchical model structure in Fig. 2, i.e., subsequent estimation of reaction fluxes, stoichiometry, reaction rates and reaction parameters. Horizontal decomposition denotes the fact that a given identification step can be performed individually for each component, e.g. reaction flux for species i independent of the other reaction fluxes, stoichiometry for reaction j independent of the other reactions, reaction rate for reaction j independent of the other reaction rates, and parameters for reaction j independent of the other reactions. Due to the decoupling, the number of possible model candidates in each step is drastically reduced. In addition, kinetics identification is restricted to the solution of purely *algebraic regression problems* as process dynamics are considered in the flux estimation and can be omitted subsequently. This leads to a drastic increase in efficiency and robustness, compared to conventional simultaneous parameter estimation. For illustration, a reaction system with n_R reactions involved is considered. For each of the reactions, a set of l_j, $j = 1, \ldots, n_R$, feasible candidates for the description of kinetic laws is available. The most suitable model for the description of the reaction system needs to be identified from experimental data. Using simultaneous identification, $l_{\text{sim}} = \prod_{j=1}^{n_R} l_j$ dynamic parameter estimation problems need to be solved, where the most suitable model is subsequently determined using some model discrimination criterion. In comparison, the incremental identification approach only requires the solution of $l_{\text{inc}} = \sum_{j=1}^{n_R} l_j$ algebraic regression problems in addition to the linear flux estimation problem in the first step.

The steps involved in the incremental identification process are discussed in more detail subsequently. To retain clarity of description, the focus is on isothermal, single-phase systems. The concept can however be readily extended to cover multiphase systems and non-isothermal cases, including the identification of temperature-dependent reaction parameters (Brendel, 2005).

The detailed explanation of the identification steps is presented in the following two sections. The basic steps are covered in Section 3, when measurements are available for all species participating in the reaction network. If measurements are available only for a subset of species, some extensions to the identification scheme are required. They are presented in Section 4, including identifiability criteria for the unknown rates and reconstruction of unmeasured species.

3. Incremental identification—the base case

Consider a homogeneous, chemical reaction system with n_R reactions involving n_S species $i \in \mathscr{S}$, where $\mathscr{S}$ is the set of reacting species. The reactions take place in a generic, well mixed and isothermal reactor with feed and effluent streams, as depicted in Fig. 1. All species are assumed to be measured for this base case.

3.1. Reaction flux estimation

The reaction fluxes need to be calculated for each species from concentration data, flow rate and volume measurements.

3.1.1. Problem formulation

The time evolution of the number of moles of species i, n_i (mol), is given by

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = f_i^{\text{in}} - f_i^{\text{out}} + f_i^{\text{r}}, \quad i = 1, \ldots, n_S, \tag{5}$$

where f_i^{in} and f_i^{out} (mol/min) are the molar flow rates of species i into and out of the reactor, and f_i^{r} (mol/min) is the reaction flux of species i. The molar flow rates f_i^{in} and f_i^{out} are calculated from

$$f_i^{\text{in}} = q^{\text{in}} c_i^{\text{in}}, \quad f_i^{\text{out}} = q^{\text{out}} c_i, \tag{6}$$

where q^{in} and q^{out} denote the total volumetric feed and effluent streams, respectively. The concentration of species i in the reactor is expressed by c_i, whereas c_i^{in} represents the feed concentration. The molar concentrations c_i are defined according to

$$n_i = c_i v, \tag{7}$$

where v is the reactor volume. Integration of (5) yields

$$n_i(t) = n_i(t_0) + \int_{\tau=t_0}^{\tau=t} \left[f_i^{\text{in}}(\tau) - f_i^{\text{out}}(\tau) \right] \mathrm{d}\tau + \int_{\tau=t_0}^{\tau=t} f_i^{\text{r}}(\tau)\, \mathrm{d}\tau. \tag{8}$$

Following the approach of Mhamdi and Marquardt (1999), the generic model of a dynamic system with unknown inputs is formulated as

$$\frac{\mathrm{d}y_i(t)}{\mathrm{d}t} = f_i^{\text{r}}(t), \quad y_i(t_0) = 0, \tag{9}$$

where $f_i^{\text{r}}(t)$ is considered as an unknown input to a dynamic system.

Eqs. (8) and (9) give

$$y_i(t) \equiv \int_{\tau=t_0}^{\tau=t} f_i^{\text{r}}(\tau)\, \mathrm{d}\tau = n_i(t) - n_i(t_0) - \int_{\tau=t_0}^{\tau=t} \left[f_i^{\text{in}}(\tau) - f_i^{\text{out}}(\tau) \right] \mathrm{d}\tau, \tag{10}$$

which indicates that the reaction fluxes $f_i^{\text{r}}(t)$, $i \in \mathscr{S}$, can be estimated independently for each species.

The unknown input must be determined on the basis of the noisy measurement

$$\tilde{y}_i(t) = y_i(t) + \varepsilon_{y_i}(t). \tag{11}$$

Here, the superscript $(\tilde{\cdot})$ is used to denote a noisy quantity and ε_y represents the measurement noise contained in $\tilde{y}$.

To determine $y_i(t)$, measurement data need to be available for the volume $v(t)$, the concentrations $c_i(t)$ and $c_i^{\text{in}}(t)$, and the volumetric feed and effluent flow rates $q^{\text{in}}(t)$ and $q^{\text{out}}(t)$. Each of the measured quantities represents a noise-corrupted instance $\tilde{z}(t) = z(t) + \varepsilon_z(t)$ of the true quantity $z(t)$. The measurement noise terms $\varepsilon_z(t)$ contribute to the errors $\varepsilon_{y_i}(t)$. These errors $\varepsilon_{y_i}(t)$ usually do not show a normal distribution even if $\varepsilon_z(t)$ can be assumed to be normal.

3.1.2. Calculation of regularized flux estimates

The estimation of $f_i^{\text{r}}(t)$ from (9) represents a classical ill-posed inverse problem (Engl et al., 1996). Since the measurement $\tilde{y}_i(t)$ is noisy, the error in the estimate $\hat{f}_i^{\text{r}}(t)$ of $f_i^{\text{r}}(t)$ can be arbitrarily large if no stabilizing regularization of the solution is considered.

For the solution of ill-posed problems, a variety of methods exist in the literature. Mhamdi and Marquardt (1999, 2003) have used Tikhonov–Arsenin filtering for the estimation of $f_i^{\text{r}}(t)$. The quality of the estimation is greatly influenced by the choice of the regularization parameter that expresses the trade-off between noise reduction and bias in the estimate. Adequate regularization parameters can be determined for example by the *L-curve criterion* (Hansen, 1998). Smoothing splines (Craven and Wahba, 1979) constitute an alternative to Tikhonov–Arsenin filtering. Splines are piecewise polynomial functions that possess certain smoothness and differentiability properties at the nodes. *General cross validation* (GCV) is often used to select a suitable regularization parameter (Craven and Wahba, 1979). Further alternatives can be found in the theory on kernel smoothers (Härdle, 1990), wavelet decomposition (Abramovich and Silverman, 1998) or neural networks (MacKay, 1992). Taking a different point of view, estimation of unknown inputs can be regarded as a dynamic optimization problem applied to (9) (Binder et al., 2002).

3.1.3. Reduction of measurement noise

If all species are measured, the measurement noise contained in $\tilde{y}_i$ can be reduced by means of data reconciliation based on atomic balances (Bonvin and Rippin, 1990).

In particular, consider n_Q data samples at the time instants t_q, $t_0 \leqslant t_q \leqslant t_{n_Q-1}$. The $n_Q \times n_S$ data matrix

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \tilde{y}_1(t_0) & \cdots & \tilde{y}_{n_S}(t_0) \\ \vdots & \ddots & \vdots \\ \tilde{y}_1(t_{n_Q-1}) & \cdots & \tilde{y}_{n_S}(t_{n_Q-1}) \end{bmatrix} \tag{12}$$

represents the molar changes due to the chemical reactions for each species $i \in \mathscr{S}$ at time instants $t_0 \leqslant t_q \leqslant t_{n_Q-1}$. For the n_S species, the $n_A \times n_S$ atomic matrix reads

$$\mathbf{M} = \begin{bmatrix} m_{11} & \cdots & m_{1n_S} \\ \vdots & \ddots & \vdots \\ m_{n_A1} & \cdots & m_{n_An_S} \end{bmatrix}, \tag{13}$$

where m_{ij}, $i = 1, 2, \ldots, n_A$, $j = 1, 2, \ldots, n_S$, is the number of atoms of the ith type in the jth chemical species.

During reaction, the number of atoms of each type is conserved. Hence, the error-free data matrix $\mathbf{Y}$ must necessarily obey

$$\mathbf{Y}\mathbf{M}^{\text{T}} = \mathbf{0}. \tag{14}$$

This property can be used to project the noisy matrix $\tilde{\mathbf{Y}}$ onto the null space of $\mathbf{M}$. With the $n_S \times n_S$ projection matrix $\mathbf{P}_{\mathbf{M}}$ associated with the null space of $\mathbf{M}$ (Björck, 1996),

$$\mathbf{P}_{\mathbf{M}} = \mathbf{I} - \mathbf{M}^{\dagger}\mathbf{M}, \tag{15}$$

the reconciled data matrix $\tilde{\mathbf{Y}}'$ results from

$$\tilde{\mathbf{Y}}' = \tilde{\mathbf{Y}}\mathbf{P}_{\mathbf{M}}. \tag{16}$$

Matrix $\mathbf{I}$ in Eq. (15) is the $n_S \times n_S$ identity matrix. The reconciled data matrix is used now to address the inverse problem in Eq. (9).

3.2. Identification of a stoichiometric model

The time-varying reaction fluxes are now available for each measured species. The focus is turned to the reaction pathways, i.e., the stoichiometric relations between the various reactions.

If the stoichiometric matrix $\mathbf{N}$ is known, the reaction rates can be readily calculated from the determined reaction fluxes (Section 3.3). Often, however, the exact stoichiometric model is unknown and needs to be identified from the data. To this end, TFA can be used to determine the number of independent reactions and the corresponding stoichiometries, without knowledge of reaction kinetics (Bonvin and Rippin, 1990; Amrhein et al., 1999). The basic idea is as follows: educated guesses or chemical intuition on possible reaction pathways generally provides candidate reactions for the system under investigation. Then, TFA allows testing each candidate reaction individually for compatibility with the measured data.

To identify a stoichiometric model from the available measurements, TFA requires (i) a target stoichiometric matrix $\mathbf{N}_{\text{tar}}$, the rows of which constitute possible stoichiometric candidates, and (ii) the experimental data in the form of the data matrix $\tilde{\mathbf{Y}}$ or its reconciled version $\tilde{\mathbf{Y}}'$. When data from different experiments are available, $\tilde{\mathbf{Y}}$ is obtained as $\tilde{\mathbf{Y}} = [\tilde{\mathbf{Y}}_1^{\mathrm{T}}, \ldots, \tilde{\mathbf{Y}}_{n_E}^{\mathrm{T}}]^{\mathrm{T}}$, where $\tilde{\mathbf{Y}}_i$ is the data matrix for experiment i, $i = 1, \ldots, n_E$. The target stoichiometric matrix $\mathbf{N}_{\text{tar}}$ is $n_{R_c} \times n_S$, where n_{R_c} is the number of candidate reaction stoichiometries and n_S the number of species involved in the reaction network.

Bonvin and Rippin (1990) have used the method of factor analysis (Malinowski, 1991) to determine the number of reactions and to derive an observed stoichiometric space from $\tilde{\mathbf{Y}}$. The validity of the target stoichiometries proposed is then individually tested on the observed stoichiometric space. Good results were achieved using a recursive TFA approach, allowing stepwise identification of stoichiometries and thereby reducing considerably the effect of measurement noise on the data. The procedure results in a $n_R \times n_S$, $n_R \leqslant n_{R_c}$ stoichiometric matrix $\mathbf{N}$, compatible with the data observed in the system. For details on the implementation, the reader is referred to the original work (Bonvin and Rippin, 1990; Amrhein et al., 1999).

TFA is capable of handling non-isothermal systems, unmeasured species and unknown elements in the target matrices. A necessary condition for the use of the TFA technique is that the number of measured species must exceed the pseudo-rank of data matrix $\tilde{\mathbf{Y}}$, i.e., the number of reactions that can be seen in the measurements (Bonvin and Rippin, 1990).

The application of the TFA technique requires special attention when linearly dependent stoichiometries occur, such as reversible reactions. Indeed, TFA will accept any linear combination of stoichiometries present in the system. This may lead to difficulties, for example, in discriminating between sequential and parallel reaction mechanisms. Moreover, the matrix $\mathbf{N}_{\text{tar}}$ must have full rank. Dependent reactions violate this condition, which requires the reduction of $\mathbf{N}_{\text{tar}}$ to a full-rank matrix. These topics are detailed in the example below.

Example 3.1. Consider a simple reaction system with components A, B and C, where A is known to convert to species B and C. The correct reaction path is unknown. Candidate stoichiometries for the system can be set up as

$$A \rightarrow B, \tag{17a}$$

$$B \rightarrow C, \tag{17b}$$

$$A \rightarrow C. \tag{17c}$$

For the reaction of A to B and C, the following three possible cases can be distinguished:

(1) Only reactions (17a) and (17b) occur (sequential reactions).
(2) Only reactions (17a) and (17c) occur (parallel reactions).
(3) Reactions (17a)–(17c) occur (sequential and parallel reactions).

Assume that all three reactions represent feasible stoichiometric candidates (case 3). Thus, the corresponding stoichiometric target matrix reads

$$\mathbf{N}_{\text{tar}} = \begin{bmatrix} -1 & +1 & 0 \\ 0 & -1 & +1 \\ -1 & 0 & +1 \end{bmatrix}. \tag{18}$$

Due to the linear dependence of the stoichiometric candidates, $\text{rank}(\mathbf{N}_{\text{tar}}) = 2$. To apply TFA, $\mathbf{N}_{\text{tar}}$ is reduced to

$$\mathbf{N}_{\text{tar}}^{\text{red}} = \begin{bmatrix} -1 & +1 & 0 \\ 0 & -1 & +1 \end{bmatrix} \tag{19}$$

by omission of reaction (17c).

In general, any of the linearly dependent stoichiometries can be omitted to construct a full-rank matrix with $\text{rank}(\mathbf{N}_{\text{tar}}^{\text{red}}) = \text{rank}(\mathbf{N}_{\text{tar}})$. Matrix $\mathbf{N}^{\text{red}}$ then results from the application of TFA to $\mathbf{N}_{\text{tar}}^{\text{red}}$.

In the example system, the target stoichiometric model $\mathbf{N}_{\text{tar}}^{\text{red}}$ (19) will always be accepted, regardless of the true mechanism (either case 1, 2 or 3), as it is not possible to discriminate between the three cases from experimental data at this point. Likewise, any other target matrix constructed from two linearly independent stoichiometries of reaction system (17) will be accepted.

Generally, for such target matrices representing linear combinations of the stoichiometries inherent to the system, the accepted stoichiometric model is normally not identical to the true stoichiometric model but rather represents a (known) linear combination. In consequence, the rates $r(t)$ identified with these stoichiometric matrices do not correspond to the real rates occurring in the system, but constitute linear combinations of the true rates. Hence, they are referred to as *pseudo-rates* $r^{\Sigma}(t)$. As will be shown in Section 3.4.2, even these pseudo-rates can be used to identify the kinetic laws.

3.3. Reaction rate estimation

According to Eq. (3), the reaction flux of species i at time $t_q \in \mathbf{t} = [t_0, \ldots, t_{n_Q-1}]$ can be expressed in terms of the individual (pseudo-)reaction rates

$$f_i^{\mathrm{r}}(t_q) = v(t_q) \sum_{j=1}^{n_R} \nu_{ji} r_j(t_q), \quad i = 1, \ldots, n_S, \quad q = 0, \ldots, n_Q - 1, \tag{20}$$

where $r_j(t_q)$ (mol/(l min)) is the rate of the jth reaction at time t_q. In matrix form, Eq. (20) reads

$$\mathbf{F}^{\mathrm{r}} = \mathbf{V}\mathbf{R}\mathbf{N}, \tag{21}$$

where $\mathbf{F}^{\mathrm{r}}$ is the $n_Q \times n_S$ reaction flux matrix with elements $f_i^{\mathrm{r}}(t_q)$, $i = 1, \ldots, n_S$, $q = 0, \ldots, n_Q - 1$, and $\mathbf{R}$ is the $n_Q \times n_R$ reaction rate matrix. The $n_Q \times n_Q$ diagonal matrix $\mathbf{V} = \mathrm{diag}\{\boldsymbol{v}\}$ represents the volume measurements $\boldsymbol{v} = [v(t_0), \ldots, v(t_{n_Q-1})]$.

The estimation of the reaction rates $\hat{\mathbf{R}}$ from the reaction fluxes $\hat{\mathbf{F}}^{\mathrm{r}}$ may be formulated as a weighted least-squares problem, i.e.,

$$\hat{\mathbf{R}} = \arg\min \mathrm{tr}(\boldsymbol{\Phi}),$$
$$\boldsymbol{\Phi} = \left(\hat{\mathbf{F}}^{\mathrm{r}} - \mathbf{V}\mathbf{R}\mathbf{N}\right) \boldsymbol{\Psi}_{\mathrm{F}}^{-1} \left(\hat{\mathbf{F}}^{\mathrm{r}} - \mathbf{V}\mathbf{R}\mathbf{N}\right)^{\mathrm{T}}, \tag{22}$$

with $\boldsymbol{\Psi}_{\mathrm{F}}$ representing the $n_S \times n_S$ covariance matrix of the noise on the individually estimated reaction fluxes. The solution is given as (Bard, 1974)

$$\hat{\mathbf{R}} = \mathbf{V}^{-1} \hat{\mathbf{F}}^{\mathrm{r}} \boldsymbol{\Psi}_{\mathrm{F}}^{-1} \mathbf{N}^{\mathrm{T}} \left(\mathbf{N} \boldsymbol{\Psi}_{\mathrm{F}}^{-1} \mathbf{N}^{\mathrm{T}}\right)^{-1}. \tag{23}$$

The stoichiometries contained in $\mathbf{N}$ are valid for both reversible and irreversible reactions. In the former case, the reactions may proceed in both directions, causing the estimated reaction rates to take either only positive, only negative or positive as well as negative values (see Section 3.4.3). In the latter case, the absence of a reverse reaction restricts the estimated rates to positive values (assuming the stoichiometry describes the correct reaction direction). Such *prior* knowledge of the existence of an irreversible reaction can be incorporated as a constraint in the reaction rate estimation problem. Eq. (22) is then extended to

$$\hat{\mathbf{R}} = \arg\min \mathrm{tr}(\boldsymbol{\Phi}),$$
$$\boldsymbol{\Phi} = \left(\hat{\mathbf{F}}^{\mathrm{r}} - \mathbf{V}\mathbf{R}\mathbf{N}\right) \boldsymbol{\Psi}_{\mathrm{F}}^{-1} \left(\hat{\mathbf{F}}^{\mathrm{r}} - \mathbf{V}\mathbf{R}\mathbf{N}\right)^{\mathrm{T}}$$
$$\text{s.t.} \quad \mathbf{0} \leqslant \mathbf{r}_j, \quad j \in \mathscr{I}, \tag{24}$$

where $\mathbf{r}_j$ is the jth reaction rate vector in matrix $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_{n_R}]$ and $\mathscr{I}$ is the set of irreversible reactions. The bounded least-squares problem (24) can be solved by quadratic programming (Gill and Murray, 1978; Stoer, 1971).

Eq. (23) yields an unbiased estimate $\hat{\mathbf{R}}$ of $\mathbf{R}$ for the case of normally distributed noise on $\hat{\mathbf{F}}^{\mathrm{r}}$. Due to the nature of the flux estimation algorithm (Section 3.1), the noise in $\hat{\mathbf{F}}^{\mathrm{r}}$ generally does not exhibit normal distribution. Furthermore, only rough estimates of the bounds of the noise on $\hat{\mathbf{F}}^{\mathrm{r}}$ can be given (Bardow and Marquardt, 2004b). The covariance matrix $\boldsymbol{\Psi}_{\mathrm{F}}$ is pragmatically chosen to be the $n_S \times n_S$ identity matrix $\mathbf{I}$. Then, Eq. (23) simplifies to

$$\hat{\mathbf{R}} = \mathbf{V}^{-1} \hat{\mathbf{F}}^{\mathrm{r}} \mathbf{N}^{\dagger}, \tag{25}$$

where $\mathbf{N}^{\dagger} = \mathbf{N}^{\mathrm{T}}(\mathbf{N}\mathbf{N}^{\mathrm{T}})^{-1}$ is the Moore–Penrose inverse of $\mathbf{N}$. This choice of covariance matrix is used throughout this paper.

3.4. Identification of kinetic laws

So far, a set of estimated reaction rates $\hat{r}_j$, $j = 1, \ldots, n_R$, and measured concentration transients $\tilde{c}_i$, $i = 1, \ldots, n_S$, are available for various time instants t_q, $q = 0, \ldots, n_Q - 1$. Since the measurements $\tilde{c}_i(t_q)$ may contain a significant level of noise, non-parametric smoothing algorithms (e.g. Craven and Wahba, 1979; Härdle, 1990; Abramovich and Silverman, 1998) may be used to obtain smooth estimates $\hat{c}_i(t_q)$. Now, a correlation between those quantities is established according to

$$\hat{r}_j(t_q) = m_j(\boldsymbol{\theta}_j, \hat{\mathbf{c}}(t_q)), \quad j = 1, \ldots, n_R, \quad q = 0, \ldots, n_Q - 1, \tag{26}$$

where $\boldsymbol{\theta}_j$ is the set of unknown parameters in model m_j, and $\hat{\mathbf{c}}(t_q) = [c_1(t_q), \ldots, c_{n_S}(t_q)]$. The correlation constructs a general predictive model m_j for reaction j, given the set of concentrations.

3.4.1. Regression problem

Given a model structure, the unknown parameters in the model need to be determined such that the model prediction comes close to the available data. A variety of criteria defining such goodness of fit and the corresponding parameter estimation methods are available in the literature, see e.g. (Bard, 1974). In the univariate regression problem considered here, the model may exhibit nonlinearity in the parameters. In addition, since the error level on the concentration data $\hat{c}$ is generally much smaller than the error on the estimated rates $\hat{r}$, a simple least-squares approach seems adequate. The parameter estimates result from

$$\hat{\boldsymbol{\theta}}_j = \arg\min \frac{1}{2} \sum_{q=0}^{n_Q-1} (\hat{r}_j(t_q) - m_j(\boldsymbol{\theta}_j, \hat{\mathbf{c}}(t_q)))^2, \tag{27}$$

with the estimated parameter vector $\hat{\boldsymbol{\theta}}_j$ consisting of one single parameter, e.g. in the case of elementary kinetics, or a set of parameters, in the case of formal kinetics.

For a set of candidate models under consideration, parameter estimation is performed for each candidate. Subsequently, the particular model that best describes the data needs to be identified. Model discrimination techniques rank the models according to their probability for correctly predicting the next experimental observation (Verheijen, 2003).

3.4.2. Dependent reactions

The case of dependent reactions has already been introduced earlier in the text: Section 3.2 dealt with the issue of stoichiometric matrix reduction for the case of dependent reactions. Full rank of the stoichiometric matrix is also required for the

identification of kinetic laws. In case of a rank-deficient stoichiometric model, since the rates are not uniquely identifiable from the estimated reaction fluxes, full-rank reduction of the stoichiometric matrix is required.

The rates identified with the $n_R \times n_S$ reduced stoichiometric matrix $\mathbf{N}^{\mathrm{red}}$ do not necessarily correspond to the true rates in the system (even if the remaining stoichiometries in $\mathbf{N}^{\mathrm{red}}$ actually correspond to occurring reactions), but rather constitute linear combinations of the true rates, the so-called $pseudo-rates\ r_j^{\Sigma}$. It must be stressed that a fully predictive model is already obtained from the reduced matrix $\mathbf{N}^{\mathrm{red}}$ and adequate state-dependent description of the pseudo-rates. Knowledge of the relation between identifiable pseudo-rates and true rates is however beneficial to derive physically motivated kinetic laws or exploit the structure in data-driven model approaches.

For $\mathbf{N}^{\mathrm{red}}$, the relation between fluxes and pseudo-rates is expressed as

$$\mathbf{R}^{\Sigma} = \mathbf{V}^{-1}\mathbf{F}^{\mathrm{r}}(\mathbf{N}^{\mathrm{red}})^{\dagger} \tag{28}$$

and

$$\mathbf{F}^{\mathrm{r}} = \mathbf{V}\mathbf{R}^{\Sigma}\mathbf{N}^{\mathrm{red}}, \tag{29}$$

whereas, for the full, rank-deficient $n_R^{\mathrm{rdf}} \times n_S$ stoichiometric matrix $\mathbf{N}$, (21) is valid. A criterion for the relation between $n_Q \times n_R^{\mathrm{rdf}}$ matrix $\mathbf{R}$ and $n_Q \times n_R$ matrix $\mathbf{R}^{\Sigma}$ is derived from (21) and (29) as

$$\mathbf{R}^{\Sigma} = \mathbf{R}\mathbf{C}^{\Sigma} \tag{30}$$

with the $n_R^{\mathrm{rdf}} \times n_R$ aggregation matrix

$$\mathbf{C}^{\Sigma} = \mathbf{N}\left(\mathbf{N}^{\mathrm{red}}\right)^{\dagger}. \tag{31}$$

The aggregation matrix $\mathbf{C}^{\Sigma}$ can be decomposed into block-diagonal structure (Pothen and Fan, 1990), where the individual blocks representing the relations between $\mathbf{R}^{\Sigma}$ and $\mathbf{R}$ can be solved independently. An example of incremental identification of a system with dependent reactions is presented in Section 4.4.

3.4.3. Reversible reactions

A special case of dependent reactions is encountered when reversible reactions are present in the system. Due to the linear dependence of the stoichiometries of the forward (1) and reverse (2) reactions, only one of the stoichiometries can be included in the stoichiometric matrix. The estimated reaction rate $\hat{\mathbf{r}}_1^{\Sigma} = [\hat{r}_1^{\Sigma}(t_0), \ldots, \hat{r}_1^{\Sigma}(t_{n_Q-1})]^{\mathrm{T}}$ will then describe both the forward and reverse reactions according to $\mathbf{r}_1^{\Sigma} = \mathbf{r}_1 - \mathbf{r}_2$. Unknown model parameters in the kinetic laws describing $\mathbf{r}_1$ and $\mathbf{r}_2$ are estimated from regression of $\hat{\mathbf{r}}_1^{\Sigma}$ with the estimated concentration trajectories $\hat{\mathbf{c}}$. Hence, the case of reversible reactions integrates in the framework proposed above, thereby allowing multiple reversible reactions in complex system stoichiometries.

4. Incremental identification—extensions

While concentration measurements have been assumed to be available for all reacting species in Section 3, the focus is now set on the case where data are present for a subset of species only. The full and measured set of species will be denoted $\mathscr{S}$ and $\mathscr{S}_{\mathrm{m}} \subset \mathscr{S}$ in the following. In addition to those sets, $\mathscr{S}_{\mathrm{u}}$ is introduced to denote the set of the $n_{S_{\mathrm{u}}}$ unmeasured species. The relations $\mathscr{S}_{\mathrm{m}} \cup \mathscr{S}_{\mathrm{u}} = \mathscr{S}$ and $n_{S_{\mathrm{m}}} + n_{S_{\mathrm{u}}} = n_S$ obviously hold. The incremental identification steps for this case will be sketched in Section 4.1, where the previously discussed steps are largely transferable. However, two issues demand special attention: the question regarding which rates $r_j(t)$ are identifiable from the data available, and the reconstruction of the concentrations $c_i(t)$ of unmeasured species that are required in the identification of kinetic laws. These topics are further detailed in Sections 4.2 and 4.3, respectively. In Section 4.4, an illustrative example on handling systems with unmeasured species and dependent reactions is presented.

4.1. Incremental identification for incomplete measurements

First, the reaction fluxes $\hat{f}_i^{\mathrm{r}}$ are estimated for each species $i \in \mathscr{S}_{\mathrm{m}}$ with the techniques discussed in Section 3.1. Due to the presence of unmeasured species, not all entries in the data matrix $\tilde{\mathbf{Y}}$ are known. Note that the reduction of measurement errors in $\tilde{\mathbf{Y}}$ using the knowledge of the atomic matrix, performed in Section 3.1.3 on the full set $\mathscr{S}$, is not applicable any more.

The recursive TFA approach introduced in Section 3.2 can be applied next to test possible stoichiometries. The case of unmeasured species is detailed in the original paper (Bonvin and Rippin, 1990) and not commented further here. Note that $n_{S_{\mathrm{m}}} > n_R$ is generally required.

Once the stoichiometric model has been determined, the reaction rates $\hat{r}_j$, $j = 1, \ldots, n_R$, are estimated from the reaction fluxes. Since reaction flux estimates are available for the measured species only, some of the reaction rates may not be identifiable. An analysis of identifiability and a procedure for calculating the corresponding rates are presented in Section 4.2.

For the set $\mathscr{S}_{\mathrm{m}}$, noisy concentration transients are available from measurements. To construct kinetic laws for the description of reaction kinetics, concentration data for the (possibly) rate-influencing species are essential. For the set $\mathscr{S}_{\mathrm{u}}$ of unmeasured species, however, the concentration data are not readily accessible. Using the stoichiometries and known initial concentrations of the unmeasured species, some or all of the unmeasured concentration data can be reconstructed from the available measurements. Identifiability criteria and the reconstruction of the concentrations of unmeasured species are described in Section 4.3.

Finally, kinetic laws can be obtained by correlation of the time-variant estimates of identifiable rates $\hat{\mathbf{r}}(t)$ and concentrations $\hat{\mathbf{c}}(t)$, as discussed in Section 3.4 above.

4.2. Identifiability and estimation of reaction rates

From the general stoichiometric matrix $\mathbf{N}$, the $n_R \times n_{S_{\mathrm{m}}}$ stoichiometric sub-matrix $\mathbf{N}_{\mathrm{m}}$ of measured species and the

$n_R \times n_{S_u}$ stoichiometric sub-matrix $\mathbf{N}_u$ of unmeasured species can be obtained as

$$\mathbf{N}_m = \mathbf{N}\mathbf{Q}_m, \quad \mathbf{N}_u = \mathbf{N}\mathbf{Q}_u, \tag{32}$$

where the $n_S \times n_{S_m}$ matrix $\mathbf{Q}_m$ and the $n_S \times n_{S_u}$ matrix $\mathbf{Q}_u$ are introduced to single out the columns corresponding to the measured and unmeasured species, respectively. The elements of each column of $\mathbf{Q}_m$ and $\mathbf{Q}_u$ consist of zeros and a single one.

4.2.1. All reaction rates identifiable

The $n_Q \times n_R$ reaction rate matrix $\mathbf{R}$ generated by the n_R independent reactions has full rank, i.e., $\text{rank}(\mathbf{R}) = n_R$. The stoichiometric matrix $\mathbf{N}_m$ projects $\mathbf{R}$ onto the $n_Q \times n_{S_m}$ matrix $\mathbf{F}_m^r$ according to

$$\mathbf{F}_m^r = \mathbf{V}\mathbf{R}\mathbf{N}_m \tag{33}$$

(cf. (21)). Since $\text{rank}(\mathbf{F}_m^r) = \min\{\text{rank}(\mathbf{R}), \text{rank}(\mathbf{N}_m)\}$, complete information on $\mathbf{R}$ is preserved if and only if

$$\text{rank}(\mathbf{N}_m) = n_R. \tag{34}$$

This identifiability condition allows all n_R rates to be estimated uniquely from the available fluxes. In this case, the reaction rates are obtained from the $n_Q \times n_{S_m}$ matrix $\mathbf{F}_m^r$ as (cf. Eq. (25))

$$\hat{\mathbf{R}} = \mathbf{V}^{-1}\hat{\mathbf{F}}_m^r\mathbf{N}_m^\dagger. \tag{35}$$

4.2.2. Subset of reaction rates identifiable

For the case where $\text{rank}(\mathbf{N}_m) < n_R$, only a part of the occurring reaction rates can be obtained from the measured species $i \in \mathscr{S}_m$. Which of those rates can be obtained, is examined in the following.

For the set of measurable reaction fluxes $\mathbf{F}_m^r$, i.e., the fluxes estimated from the measured concentration data, (33) is valid. For $\text{rank}(\mathbf{N}_m) = n_{R_m} < n_R$, the matrix of reaction rates $\mathbf{R}$ of rank n_R is projected onto the matrix of measured reaction fluxes $\mathbf{F}_m^r$ with

$$\text{rank}(\mathbf{F}_m^r) = \min\{\text{rank}(\mathbf{R}), \text{rank}(\mathbf{N}_m)\} = n_{R_m}. \tag{36}$$

Hence, at most n_{R_m} (pseudo-)reaction rates may be identified from the measurable reaction fluxes.

A criterion for identifiability of the rates is derived from an analysis of the difference between the true rates $\mathbf{R}$ and those ($\mathbf{R}_{inv}$) obtained from direct inversion of (33), assuming (34) is satisfied and regardless of (36):

$$\mathbf{R}_{inv} = \mathbf{V}^{-1}\mathbf{F}_m^r\mathbf{N}_m^\dagger. \tag{37}$$

Substitution of $\mathbf{F}_m^r$ from (33) gives the difference between $\mathbf{R}_{inv}$ and $\mathbf{R}$ as

$$\mathbf{R}_{inv} - \mathbf{R} = \mathbf{R}[\mathbf{N}_m\mathbf{N}_m^\dagger - \mathbf{I}], \tag{38}$$

where $\mathbf{I}$ is the $n_R \times n_R$ identity matrix. The (symmetric) difference matrix $\mathbf{\Delta}^r$ for the identifiable rates is defined as

$$\mathbf{\Delta}^r \equiv \mathbf{N}_m\mathbf{N}_m^\dagger - \mathbf{I}. \tag{39}$$

Obviously, a reaction rate r_j is theoretically identifiable if the corresponding column $\mathbf{\Delta}_j^r$ of $\mathbf{\Delta}^r = [\mathbf{\Delta}_1^r, \ldots, \mathbf{\Delta}_{n_R}^r]$ is represented by the null vector. A simplified identifiability criterion can be derived from

$$\boldsymbol{\delta}^r = \left[\delta_1^r, \ldots, \delta_{n_R}^r\right] \tag{40}$$

with elements

$$\delta_j^r = \|\mathbf{\Delta}_j^r\|_1. \tag{41}$$

The n_{R_r} zero elements of the identifiability vector $\boldsymbol{\delta}^r$ indicate identifiability of the corresponding rates from the available measurements.

The $n_Q \times n_{R_r}$ matrix $\hat{\mathbf{R}}$ of identifiable reaction rates is now obtained in analogy to Eq. (35) as

$$\hat{\mathbf{R}} = \mathbf{V}^{-1}\hat{\mathbf{F}}^r\mathbf{N}_m^\dagger\mathbf{Q}_r, \tag{42}$$

where the $n_R \times n_{R_r}$ matrix $\mathbf{Q}_r$ is chosen such as to hide the non-identifiable reaction rates. The elements of each column of $\mathbf{Q}_r$ consist of zeros and a single one.

4.3. Concentration estimation

Concentration data for the (possibly) rate-influencing species are essential to construct kinetic laws for the description of reaction kinetics. The required concentration data might not be readily accessible in the set $\mathscr{S}_m$. Consider the simple example

$$A \rightarrow B \tag{43}$$

with species B measured. The reaction flux f_B^r and the reaction rate can be calculated from the available concentration data. However, to identify the reaction parameter k in a given kinetic law (e.g. $r = kc_A$), an estimate of the concentration transient c_A of species A is required.

In such cases, concentration estimates for some or even all of the unmeasured species have to be obtained from the data available. In the following, general conditions for the reconstruction of concentration transients for unmeasured species are derived, and equations are given to obtain them from the set of measured concentration trajectories using the reaction stoichiometries. To this end, some additional notation needs to be introduced: while $\mathscr{S}_m$ is the set of measured and $\mathscr{S}_u$ the set of unmeasured species, $\mathscr{S}_c$ describes the n_{S_c} unmeasured species that can be reconstructed from the data. The relation $\mathscr{S}_c \subseteq \mathscr{S}_u \subseteq \mathscr{S}$ applies.

4.3.1. Reconstruction of reaction fluxes

In Section 3.1, we have shown that the unknown reaction fluxes can be calculated for each component from measured concentration data. Conversely, concentration transients can be calculated for those species where the reaction flux and the initial amount present in the system are known. Such reaction fluxes for unmeasured species can be calculated from the reaction rates estimated from concentration measurements by means of the stoichiometric model.

Case 1: *All reaction fluxes identifiable*. If $\text{rank}(N_m) = n_R$, all reaction rates are identifiable from the measured species. This implies identifiability of the reaction fluxes for all species involved in the reaction network, i.e., $\mathscr{S}_c = \mathscr{S}_u$.

The $n_Q \times n_{S_c}$ reaction flux matrix $\mathbf{F}_c^r$ of the unmeasured but reconstructable species is related to the $n_Q \times n_{S_m}$ matrix $\mathbf{F}_m^r$ of estimated reaction fluxes by

$$\mathbf{F}_c^r = \mathbf{F}_m^r \mathbf{T}_N, \tag{44}$$

with $\mathbf{T}_N$ being a $n_{S_m} \times n_{S_c}$ matrix to be determined next. In analogy to Eq. (33),

$$\mathbf{F}_c^r = \mathbf{V}\mathbf{R}\mathbf{N}_c \tag{45}$$

is valid for the set $\mathscr{S}_c$. With Eqs. (33), (44), (45) and $\mathbf{N}_c = \mathbf{N}_u$ we obtain

$$\mathbf{T}_N = \mathbf{N}_m^\dagger \mathbf{N}_u, \tag{46}$$

which is called reconstruction matrix subsequently.

Case 2: *Subset of reaction fluxes identifiable*. For the case where $\mathrm{rank}(N_m) < n_R$, identification is feasible for a subset of the reaction fluxes only. From the (partly erroneous) rates $\mathbf{R}_{inv}$ (37), fluxes $\mathbf{F}_{inv}^r = \mathbf{V}\mathbf{R}_{inv}\mathbf{N}$ can be calculated. In analogy to (38), a comparison between $\mathbf{F}_{inv}^r$ and the true fluxes $\mathbf{F}^r$ (cf. Eq. (21)) is employed to analyze the set $\mathscr{S}_c$, i.e., the unmeasured species for which the fluxes can be calculated correctly from the available data. The difference $\mathbf{F}_{inv}^r - \mathbf{F}^r$ can be written as

$$\mathbf{F}_{inv}^r - \mathbf{F}^r = \mathbf{V}[\mathbf{R}_{inv} - \mathbf{R}]\mathbf{N} = \mathbf{V}\mathbf{R}\boldsymbol{\Delta}^r\mathbf{N} \tag{47}$$

using (38) and (39). Defining

$$\boldsymbol{\Delta}^f \equiv \boldsymbol{\Delta}^r\mathbf{N}, \tag{48}$$

a flux is identifiable if the corresponding column of $\boldsymbol{\Delta}^f$ is the null vector. In analogy to (40), a simplified identifiability criterion can be derived from

$$\boldsymbol{\delta}^f = \left[\delta_1^f, \ldots, \delta_{n_S}^f\right], \tag{49}$$

with elements

$$\delta_i^f = \|\boldsymbol{\Delta}_i^f\|_1, \tag{50}$$

where the column vectors $\boldsymbol{\Delta}_i^f$ compose the $n_R \times n_S$ matrix $\boldsymbol{\Delta}^f = [\boldsymbol{\Delta}_1^f, \ldots, \boldsymbol{\Delta}_{n_S}^f]$. A zero element δ_i^f indicates the flux of species i as identifiable from the available measurements. The set $\mathscr{S}_c$ contains the n_{S_c} unmeasured species whose reaction fluxes (and thus the corresponding concentration transients) can be estimated from the available data, i.e., $\mathscr{S}_c = \{i | \delta_i^f = 0, i \in \mathscr{S}_u\}$.

To obtain the matrix $\mathbf{F}_c^r$ of calculable reaction fluxes from (44), Eq. (46) is replaced by the $n_{S_m} \times n_{S_c}$ matrix

$$\mathbf{T}_N = \mathbf{N}_m^\dagger \mathbf{N}_c, \quad \mathbf{N}_c = \mathbf{N}\mathbf{Q}_c, \tag{51}$$

where the $n_S \times n_{S_c}$ matrix $\mathbf{Q}_c$ is chosen so as to hide the already measured and non-identifiable reaction fluxes. The elements of each column of $\mathbf{Q}_c$ consist of zeros and a single one.

4.3.2. Calculation of concentration estimates

Assuming known initial reactor and feed concentrations $\mathbf{c}_c(t_0)$ and $\mathbf{c}_c^{in}(t)$ for the set $\mathscr{S}_c$, the calculation of corresponding concentration transients is sketched in short.

Eq. (8) applied to sets $\mathscr{S}_m$ and $\mathscr{S}_c$ yields expressions for the mole vectors $\mathbf{n}_m(t)$ and $\mathbf{n}_c(t)$ as functions of the sets of reaction fluxes $\mathbf{f}_m^r(t)$ and $\mathbf{f}_c^r(t)$, respectively. Insertion of (44) and reformulation leads to

$$\begin{aligned}\mathbf{n}_c(t) = \mathbf{n}_c(t_0) &+ \int_{\tau=t_0}^{\tau=t} \left[\mathbf{f}_c^{in}(\tau) - \mathbf{f}_c^{out}(\tau)\right] \mathrm{d}\tau \\ &+ \mathbf{T}_N^T \left[\mathbf{n}_m(t) - \mathbf{n}_m(t_0)\right] \\ &- \mathbf{T}_N^T \int_{\tau=t_0}^{\tau=t} \left[\mathbf{f}_m^{in}(\tau) - \mathbf{f}_m^{out}(\tau)\right] \mathrm{d}\tau.\end{aligned} \tag{52}$$

Using Eqs. (6) and (7), the expression

$$\begin{aligned}\mathbf{c}_c(t) = \mathbf{T}_N^T\mathbf{c}_m(t) &+ \frac{v_0}{v(t)}\left[\mathbf{c}_c(t_0) - \mathbf{T}_N^T\mathbf{c}_m(t_0)\right] \\ &+ \frac{1}{v(t)}\int_{\tau=t_0}^{\tau=t}\left[q^{in}(\tau)\mathbf{c}_c^{in}(\tau) - q^{out}(\tau)\mathbf{c}_c(\tau)\right]\mathrm{d}\tau \\ &- \frac{1}{v(t)}\mathbf{T}_N^T\int_{\tau=t_0}^{\tau=t}\left[q^{in}(\tau)\mathbf{c}_m^{in}(\tau) - q^{out}(\tau)\mathbf{c}_m(\tau)\right]\mathrm{d}\tau\end{aligned} \tag{53}$$

is obtained to relate the concentrations in the sets $\mathscr{S}_m$ and $\mathscr{S}_c$.

This equation allows the calculation of concentration transients $\hat{\mathbf{c}}_c(t)$ for the species in the set $\mathscr{S}_c$ from smoothed estimates $\hat{\mathbf{c}}_m(t)$ (cf. Section 3.4). The expression may be simplified for a variety of cases, e.g. for

- batch reactors:

$$\hat{\mathbf{c}}_c(t) = \mathbf{T}_N^T\hat{\mathbf{c}}_m(t) + \frac{v_0}{v(t)}\left[\mathbf{c}_c(t_0) - \mathbf{T}_N^T\mathbf{c}_m(t_0)\right] \tag{54}$$

- or for semi-batch reactors with constant feed:

$$\begin{aligned}\hat{\mathbf{c}}_c(t) = \mathbf{T}_N^T\hat{\mathbf{c}}_m(t) &+ \frac{v_0}{v(t)}\left[\mathbf{c}_c(t_0) - \mathbf{T}_N^T\mathbf{c}_m(t_0)\right] \\ &+ \frac{q_{in}t}{v(t)}\left[\mathbf{c}_c^{in} - \mathbf{T}_N^T\mathbf{c}_m^{in}\right].\end{aligned} \tag{55}$$

4.4. Unmeasured species in dependent reaction systems

In the case of dependent reactions, the pseudo-rate concept introduced in Section 3.4 is equally applicable to sets $\mathscr{S}_m \subset \mathscr{S}$. The relation between rates and pseudo-rates is expressed as $\mathbf{R}^\Sigma = \mathbf{R}\mathbf{C}^\Sigma$ (30). Here, Eq. (31) is replaced by

$$\mathbf{C}^\Sigma = \mathbf{N}_m(\mathbf{N}_m^{red})^\dagger, \tag{56}$$

where $\mathbf{N}_m$ is rank deficient and $\mathbf{N}_m^{red}$ is the corresponding full-rank matrix.

An illustrative example is introduced to point out the treatment of unmeasured species and dependent reactions. Consider a batch reactor with five species involved in up to four reactions. The reaction mechanism is given as

$$A \rightarrow B, \tag{57a}$$

$$B \rightarrow C, \tag{57b}$$

$$A \rightarrow C, \tag{57c}$$

$$C + D \rightarrow E. \tag{57d}$$

The full stoichiometric matrix is

$$\mathbf{N} = \begin{bmatrix} -1 & +1 & 0 & 0 & 0 \\ 0 & -1 & +1 & 0 & 0 \\ -1 & 0 & +1 & 0 & 0 \\ 0 & 0 & -1 & -1 & +1 \end{bmatrix}. \tag{58}$$

Let us assume that species A–C are available from measurements, i.e., matrix $\mathbf{N}_\mathrm{m}$ is

$$\mathbf{N}_\mathrm{m} = \begin{bmatrix} -1 & +1 & 0 \\ 0 & -1 & +1 \\ -1 & 0 & +1 \\ 0 & 0 & -1 \end{bmatrix} \tag{59}$$

with $\mathrm{rank}(\mathbf{N}_\mathrm{m}) = 3 < n_R = 4$. As a consequence, criterion (40) yields $\boldsymbol{\delta}^\mathrm{r} = [1, 1, 1, 0]$, categorizing only the rate corresponding to (57d) as identifiable, whereas the rates belonging to (57a) to (57c) cannot be determined uniquely. To meet the full-rank condition, the contribution due to a linear-dependent reaction has to be deleted in $\mathbf{N}_\mathrm{m}$. A feasible reduced matrix $\mathbf{N}_\mathrm{m}^\mathrm{red}$ is found by dropping the third reaction:

$$\mathbf{N}_\mathrm{m}^\mathrm{red} = \begin{bmatrix} -1 & +1 & 0 \\ 0 & -1 & +1 \\ 0 & 0 & -1 \end{bmatrix}, \tag{60}$$

Criterion (40), applied to $\mathbf{N}_\mathrm{m}^\mathrm{red}$, suggests that all three corresponding pseudo-rates are identifiable from the data present ($\boldsymbol{\delta}^\mathrm{r} = [0, 0, 0]$). The aggregation matrix $\mathbf{C}^\Sigma$ (56) is calculated as follows:

$$(\mathbf{C}^\Sigma) = \begin{bmatrix} +1 & 0 & 0 \\ 0 & +1 & 0 \\ +1 & +1 & 0 \\ 0 & 0 & +1 \end{bmatrix}, \tag{61}$$

from which the independent pseudo-rates can be expressed in terms of the reaction rates:

$$\left.\begin{aligned} \mathbf{r}_1^\Sigma &= \mathbf{r}_1 + \mathbf{r}_3 \\ \mathbf{r}_2^\Sigma &= \mathbf{r}_2 + \mathbf{r}_3 \end{aligned}\right\} \text{ block 1,} \qquad \left.\begin{aligned} \mathbf{r}_3^\Sigma &= \mathbf{r}_4 \end{aligned}\right\} \text{ block 2,} \tag{62}$$

where $\mathbf{r}_j^\Sigma$, $j = 1, \ldots, 3$, are the rates corresponding to the full-rank stoichiometric matrix $\mathbf{N}_\mathrm{m}^\mathrm{red}$. Assuming model structures for the rates $\mathbf{r}_i$, $i = 1, \ldots, 4$, the chemical reactions present and their parameters can now be identified by algebraic regression (27) using the estimated rates $\hat{\mathbf{r}}_j^\Sigma$, $j = 1, \ldots, 3$.

Concentration transients of the unmeasured species $\mathscr{S}_\mathrm{u} = \{D, E\}$ are required for such regression. Evaluation of (49) results in $\boldsymbol{\delta}^\mathrm{f} = [0, 0, 0, 0, 0]$. Hence, both unmeasured species can be calculated from available data, given initial concentrations $c_{i0}, i \in \mathscr{S}_\mathrm{u}$. The resulting reconstruction matrix $\mathbf{T_N}$ is calculated from (51) as

$$\mathbf{T_N} = \begin{bmatrix} +1 & -1 \\ +1 & -1 \\ +1 & -1 \end{bmatrix}. \tag{63}$$

The concentration estimates finally result from Eq. (54).

5. Illustrative example: acetoacetylation of pyrrole

The incremental approach for reaction model identification is illustrated for the acetoacetylation of pyrrole with diketene (Ruppen, 1994; Ruppen et al., 1998). The system has a main desired reaction and several undesired side reactions that impair selectivity. To validate the incremental identification approach, simulated data are used. This way, the results of the identification process can easily be compared to the model assumptions made for generating the data. The simulation is based on the experimental work of Ruppen (1994), who developed a kinetic model of the reaction system.

The reaction system is introduced next. The identification of the reaction system is carried out for some experimental scenario employing the incremental approach suggested in this paper.

5.1. Reaction system and experimental conditions

The reaction system comprises the reactions

$$\mathrm{P} + \mathrm{D} \xrightarrow{\mathrm{K}} \mathrm{PAA}, \tag{64a}$$

$$\mathrm{D} + \mathrm{D} \xrightarrow{\mathrm{K}} \mathrm{DHA}, \tag{64b}$$

$$\mathrm{D} \rightarrow \mathrm{OL}, \tag{64c}$$

$$\mathrm{PAA} + \mathrm{D} \xrightarrow{\mathrm{K}} \mathrm{G}. \tag{64d}$$

In addition to the desired reaction (64a) of diketene (D) and pyrrole (P) to 2-acetoacetyl pyrrole (PAA), there are three undesired side reactions (64b)–(64d). These include the dimerization and oligomerization of diketene to dehydroacetic acid (DHA) and oligomers (OLs) as well as a consecutive reaction to the by-product G.

The reactions take place in an isothermal laboratory-scale semi-batch reactor, to which a diluted solution of diketene is added continuously. Reactions (64a), (64b) and (64d) are catalyzed by pyridine (K), the concentration of which continuously decreases during the run due to addition of diluted diketene feed. Reaction (64c), which is assumed to be promoted by other intermediate products, is not catalyzed. Hence, the reaction rates are described by the constitutive equations

$$r_a(t) = k_a c_\mathrm{P}(t) c_\mathrm{D}(t) c_\mathrm{K}(t), \tag{65a}$$

$$r_b(t) = k_b c_\mathrm{D}^2(t) c_\mathrm{K}(t), \tag{65b}$$

$$r_c(t) = k_c c_\mathrm{D}(t), \tag{65c}$$

$$r_d(t) = k_d c_\mathrm{PAA}(t) c_\mathrm{D}(t) c_\mathrm{K}(t), \tag{65d}$$

where k_a, k_b, k_c and k_d represent the rate constants.

The reaction fluxes $\mathbf{F}^\mathrm{r} = [\mathbf{f}_\mathrm{D}^\mathrm{r}, \mathbf{f}_\mathrm{P}^\mathrm{r}, \mathbf{f}_\mathrm{PAA}^\mathrm{r}, \mathbf{f}_\mathrm{DHA}^\mathrm{r}, \mathbf{f}_\mathrm{OL}^\mathrm{r}, \mathbf{f}_\mathrm{G}^\mathrm{r}]$ can be related to the reaction rates $\mathbf{R} = [\mathbf{r}_a, \mathbf{r}_b, \mathbf{r}_c, \mathbf{r}_d]$ by

$$\mathbf{F}^\mathrm{r} = \mathbf{VRN}$$

Table 1
Values of rate constants

	k_a (l^2/mol^2 min)	k_b (l^2/mol^2 min)	k_c (1/ min)	k_d (l^2/mol^2 min)
Value	0.053	0.128	0.028	0.000

with the stoichiometric matrix

$$\mathbf{N} = \begin{bmatrix} -1 & -1 & +1 & 0 & 0 & 0 \\ -2 & 0 & 0 & +1 & 0 & 0 \\ -1 & 0 & 0 & 0 & +1 & 0 \\ -1 & 0 & -1 & 0 & 0 & +1 \end{bmatrix} \tag{66}$$

for the set of species $\mathscr{S} = \{\mathrm{D}, \mathrm{P}, \mathrm{PAA}, \mathrm{DHA}, \mathrm{OL}, \mathrm{G}\}$.

The catalyst is not affected by the chemical reactions occurring. Its dilution during the run of the experiment can be modeled as

$$c_{\mathrm{K}}(t) = \frac{v_0}{v(t)} c_{\mathrm{K}0}, \tag{67}$$

where $c_{\mathrm{K}0}$ is the initial concentration of catalyst in the reactor. Under the assumption that no volume change is induced by the reactions occurring, the reactor volume is modeled as

$$\frac{\mathrm{d}v(t)}{\mathrm{d}t} = q^{\mathrm{in}}, \quad v(t_0) = v_0, \tag{68}$$

with constant volumetric feed flow rate q^{in}.

The material balance for species $i \in \mathscr{S}$ reads as

$$\frac{\mathrm{d}c_i(t)}{\mathrm{d}t} = \frac{q^{\mathrm{in}}}{v(t)}\left[c_i^{\mathrm{in}} - c_i(t)\right] + \frac{f_i^{\mathrm{r}}(t)}{v(t)}, \quad c_i(t_0) = c_{i0}, \tag{69a}$$

where $c_{\mathrm{D}}^{\mathrm{in}}$ is the constant concentration of diketene in the feed. For all other species, $c_i^{\mathrm{in}} = 0,\ i \neq \mathrm{D}$. The initial conditions c_{i0} are known.

To assess the performance of the incremental identification approach and allow a comparison of the identified and simulated kinetics, concentration trajectories are generated using the model described above and the rate constants given in Table 1. The rate constant of the fourth reaction is set to $k_d = 0$, i.e., this reaction is assumed not to occur in the network.

Concentration data are assumed to be available for the set of species $\mathscr{S}_{\mathrm{m}} = \{\mathrm{D}, \mathrm{PAA}, \mathrm{DHA}, \mathrm{OL}, \mathrm{G}\}$. For P, no measurements exist. The measured concentrations are assumed to stem from a data-rich in situ measurement technique such as Raman spectroscopy, taken with the sampling period $t_s = 10$ s over the batch time $t_f = 60$ min. Thus, a total of $n_Q = 361$ data points for each species result. The data are corrupted with normally distributed white noise. The standard deviation σ_i differs for each species i, depending on its calibration range. The calibration ranges of the species can be taken from Table 2, where concentration data are expected in the range $0 \leqslant c_i \leqslant c_i^{\max}, i \in \mathscr{S}_{\mathrm{m}}$. The same relative, normally distributed error $\alpha_\sigma = 1.0\%$ within the component specific calibration range $[0, c_i^{\max}]$ is assumed for each species. Hence, the standard error on the data is assumed to follow the relation

$$\sigma_i = \alpha_\sigma c_i^{\max}, \quad i \in \mathscr{S}_{\mathrm{m}}. \tag{70}$$

Table 2
Concentration ranges for calibration

	c_{D} (mol/l)	c_{PAA} (mol/l)	c_{DHA} (mol/l)	c_{OL} (mol/l)	c_{G} (mol/l)
Min	0.00	0.00	0.00	0.00	0.00
Max	0.38	0.45	0.63	0.52	0.05

The time-varying reactor volume $v(t)$ is measured with negligible error. In addition, error-free data on q^{in} and $c_{\mathrm{D}}^{\mathrm{in}}$ exist. The concentration of catalyst K can be calculated from the volume and the initial concentration of catalyst according to Eq. (67).

In the experimental setup, the initial concentrations are $c_{\mathrm{D},0} = 0.14$ mol/l, $c_{\mathrm{P},0} = 0.30$ mol/l, $c_{\mathrm{PAA},0} = 0.08$ mol/l and $c_{\mathrm{DHA},0} = 0.01$ mol/l. Negligible amounts of both the OLs and the by-product G are supposed to be present in the reactor at $t_0 = 0$, i.e., $c_{\mathrm{OL},0} = 0.01$ mol/l and $c_{\mathrm{G}},0 = 0.01$ mol/l. The initial reactor volume is set to $v_0 = 0.5$ l, the volumetric feed rate is $q^{\mathrm{in}} = 5.0$ ml/min and its concentration of diketene amounts to $c_{\mathrm{D}}^{\mathrm{in}} = 6.0$ mol/l.

5.2. Identification procedure

For the identification process, a candidate stoichiometric matrix

$$\mathbf{N}_{\mathrm{tar}} = \begin{bmatrix} -1 & -1 & +1 & 0 & 0 & 0 \\ -2 & 0 & 0 & +1 & 0 & 0 \\ -1 & 0 & 0 & 0 & +1 & 0 \\ -1 & 0 & -1 & 0 & 0 & +1 \end{bmatrix} \tag{71}$$

is available, corresponding to the stoichiometries of reactions (64a)–(64d). However, the number and type of actually occurring reactions remain unknown and need to be identified from the data. Furthermore, for each reaction, a set of kinetic law candidates is available, corresponding to feasible power-law kinetic combinations. The number of model candidates is 10 for reactions (64a) and (64d) and six for reactions (64b) and (64c) (see Table 3 for a summary). Theoretical identifiability of the individual model structures is ensured by construction.

The noisy data sets generated by simulation are depicted in Fig. 3, together with the true concentration transients.

In a first step, the reaction fluxes $f_i^{\mathrm{r}}(t),\ i \in \mathscr{S}_{\mathrm{m}}$, are calculated using smoothing splines to solve the ill-posed inverse problem in Eq. (9). A suitable regularization parameter is obtained by means of GCV. Fig. 4 shows the resulting reaction fluxes for the set $\mathscr{S}_{\mathrm{m}}$. For species P, no reaction flux can be estimated at this point.

Next, the stoichiometries of the reaction network have to be determined, based on the candidate stoichiometries (71). The recursive TFA approach is applied to check the validity of the proposed stoichiometries and to identify the number of reactions occurring. The method successively accepts reactions (64b), (64a) and (64c) (in this order). Reaction (64d) does not take place in the simulation and is correctly not accepted. The

Table 3
Kinetic model candidates

Reaction a: $P + D \xrightarrow{K} PAA$	Reaction b: $D + D \xrightarrow{K} DHA$	Reaction c: $D \xrightarrow{K} OL$	Reaction d: $PAA + D \xrightarrow{K} G$
$m_a^{(1)} = k_a^{(1)}$	$m_b^{(1)} = k_b^{(1)}$	$m_c^{(1)} = k_c^{(1)}$	$m_d^{(1)} = k_d^{(1)}$
$m_a^{(2)} = k_a^{(2)} c_D$	$m_b^{(2)} = k_b^{(2)} c_D$	$m_c^{(2)} = k_c^{(2)} c_D$	$m_d^{(2)} = k_d^{(2)} c_D$
$m_a^{(3)} = k_a^{(3)} c_P$	$m_b^{(3)} = k_b^{(3)} c_D^2$	$m_c^{(3)} = k_c^{(3)} c_D^2$	$m_d^{(3)} = k_d^{(3)} c_{PAA}$
$m_a^{(4)} = k_a^{(4)} c_K$	$m_b^{(4)} = k_b^{(4)} c_D c_K$	$m_c^{(4)} = k_c^{(4)} c_D c_K$	$m_d^{(4)} = k_d^{(4)} c_K$
$m_a^{(5)} = k_a^{(5)} c_D c_P$	$m_b^{(5)} = k_b^{(5)} c_D^2 c_K$	$m_c^{(5)} = k_c^{(5)} c_D^2 c_K$	$m_d^{(5)} = k_d^{(5)} c_D c_{PAA}$
$m_a^{(6)} = k_a^{(6)} c_P c_K$	$m_b^{(6)} = k_b^{(6)} c_K$	$m_c^{(6)} = k_c^{(6)} c_K$	$m_d^{(6)} = k_d^{(6)} c_{PAA} c_K$
$m_a^{(7)} = k_a^{(7)} c_D c_K$			$m_d^{(7)} = k_d^{(7)} c_D c_K$
$m_a^{(8)} = k_a^{(8)} c_D c_P c_K$			$m_d^{(8)} = k_d^{(8)} c_D c_{PAA} c_K$
$m_a^{(9)} = k_a^{(9)} c_D c_P^2$			$m_d^{(9)} = k_d^{(9)} c_D c_{PAA}^2$
$m_a^{(10)} = k_a^{(10)} c_D^2 c_P$			$m_d^{(10)} = k_d^{(10)} c_D^2 c_{PAA}$

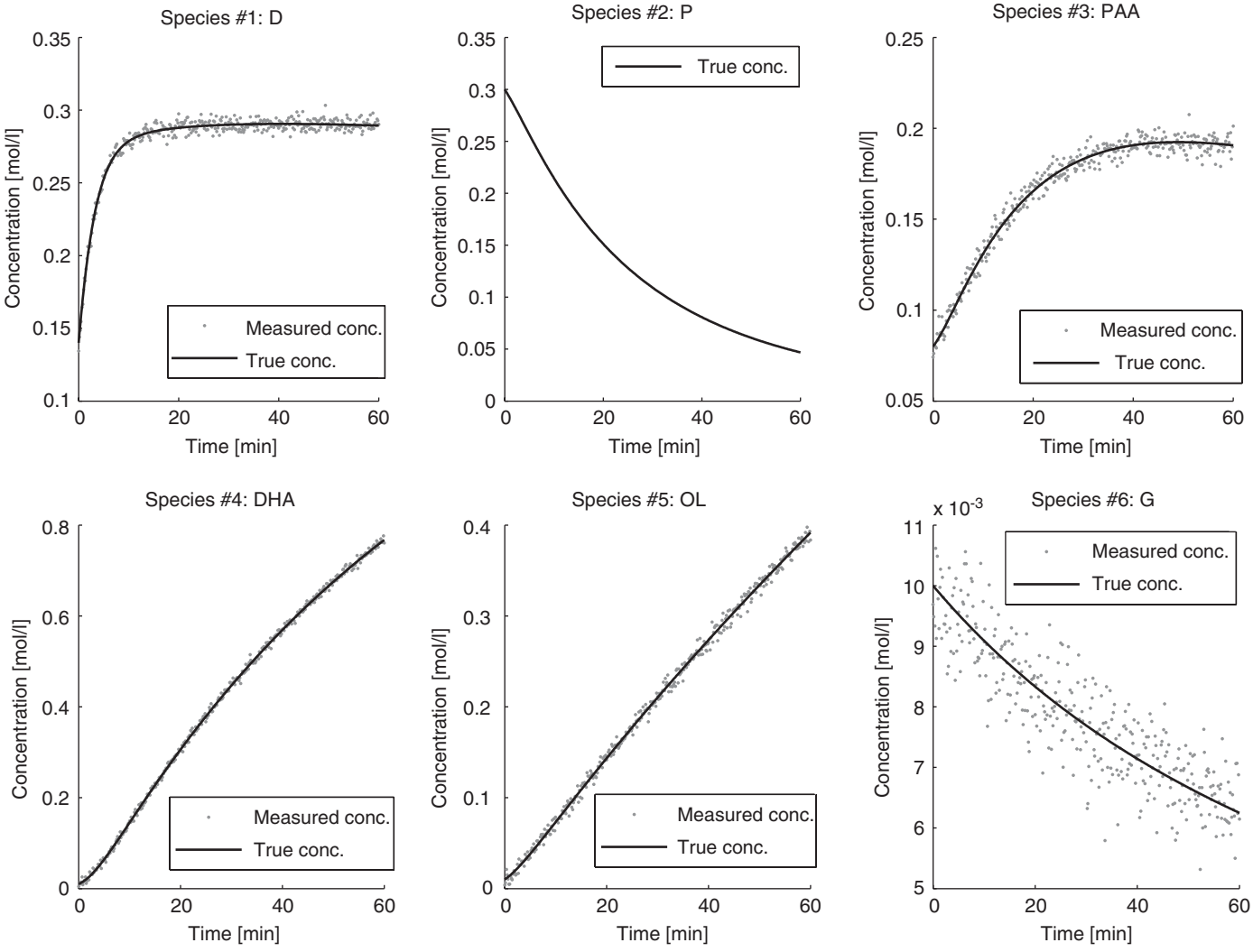

Fig. 3. True and noisy concentrations.

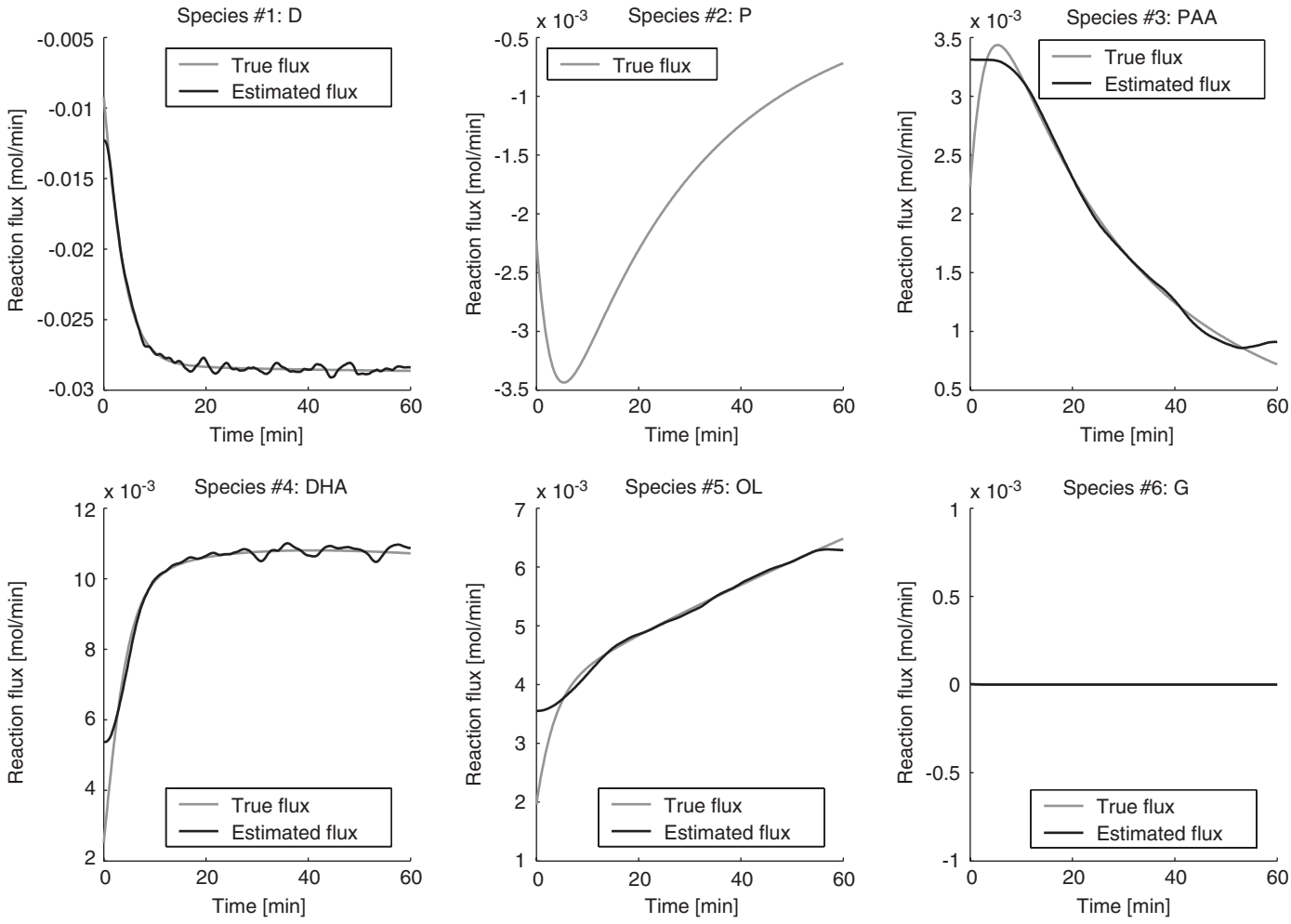

Fig. 4. True and estimated reaction fluxes.

resulting stoichiometric matrix of the reaction network reads

$$\mathbf{N} = \begin{bmatrix} -1 & -1 & +1 & 0 & 0 & 0 \\ -2 & 0 & 0 & +1 & 0 & 0 \\ -1 & 0 & 0 & 0 & +1 & 0 \end{bmatrix}. \tag{72}$$

The $n_R \times n_{S_m}$ stoichiometric matrix $\mathbf{N}_\mathrm{m}$ (32) is

$$\mathbf{N}_\mathrm{m} = \begin{bmatrix} -1 & +1 & 0 & 0 & 0 \\ -2 & 0 & +1 & 0 & 0 \\ -1 & 0 & 0 & +1 & 0 \end{bmatrix}. \tag{73}$$

With this stoichiometric matrix, reaction rates are estimated from (35). Since $\mathrm{rank}(\mathbf{N}_\mathrm{m}) = n_R$, all rates can be identified from the reaction fluxes present. With more species measured (5) than independent reactions occurring in system (3), a least-squares reconciliation problem results which reduces the errors in the rates. The resulting, time-variant reaction rates are depicted in Fig. 5, together with the true rates for comparison.

Estimates $\hat{c}_i(t)$, $i \in \mathscr{S}_\mathrm{m}$, are obtained using smoothing splines with GCV to select the regularization parameter. To identify the unknown parameters in the set of kinetic model candidates, an estimate of the concentration trajectory of species P needs to be available. With $\mathrm{rank}(\mathbf{N}_\mathrm{m}) = n_R$, $\hat{c}_\mathrm{P}$ can be obtained from (55) using the known initial concentration $c_{\mathrm{P}0}$. The estimate shows a close fit to the true concentration c_P, with a relative deviation of 0.83%.

For the description of reaction kinetics, a suitable model can now be selected from the set of model candidates available for each accepted reaction (Table 3), together with the unknown model parameters. To this end, for each reaction, the available model candidates are fitted to the estimates of the concentrations and rates, both available as a function of time, according to (27). Some representative fits are plotted in Fig. 6 for the first reaction (64a). Here, candidate 8 (cf. Table 3) can be best fitted to the estimated reaction rate and is identified as the most suitable kinetic law from the set of candidates.

From the data available, the following kinetics were identified:

$$r_a(t) = k_a c_\mathrm{P}(t) c_\mathrm{D}(t) c_\mathrm{K}(t), \tag{74a}$$

$$r_b(t) = k_b c_\mathrm{D}^2(t) c_\mathrm{K}(t), \tag{74b}$$

$$r_c(t) = k_c c_\mathrm{D}(t), \tag{74c}$$

with parameters $k_a = 0.0523\,1^2/(\mathrm{mol}^2\ \mathrm{min})$, $k_b = 0.1279\,1^2/(\mathrm{mol}^2\ \mathrm{min})$ and $k_c = 0.0281\ 1/\mathrm{min}$. For all three reactions, the kinetic laws correspond to the correct model structures, i.e.,

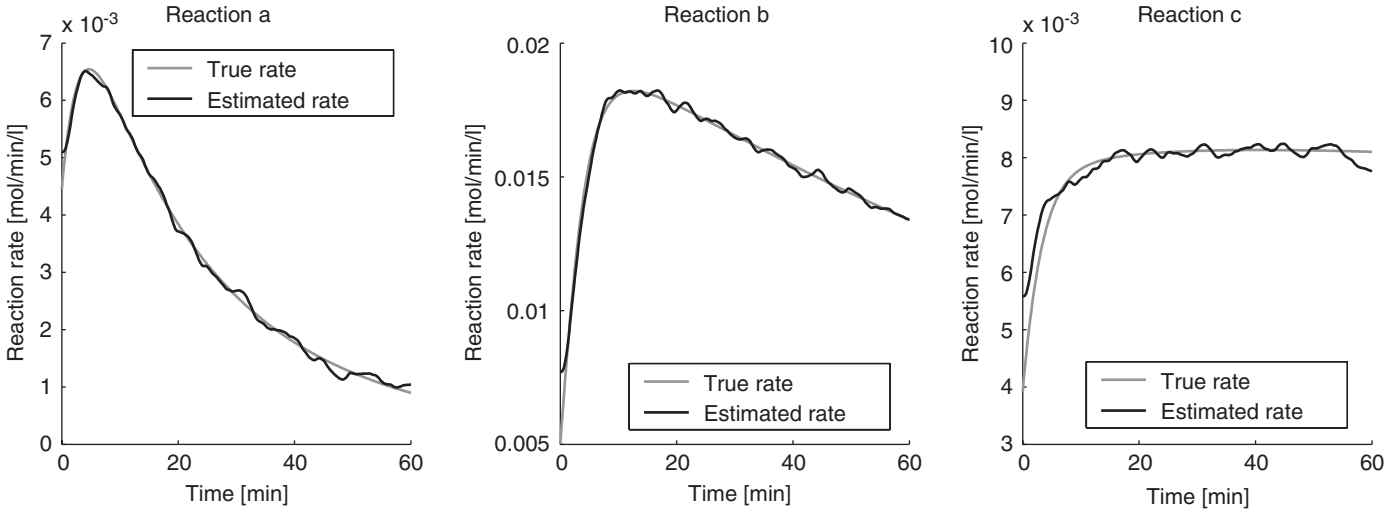

Fig. 5. True and estimated reaction rates.

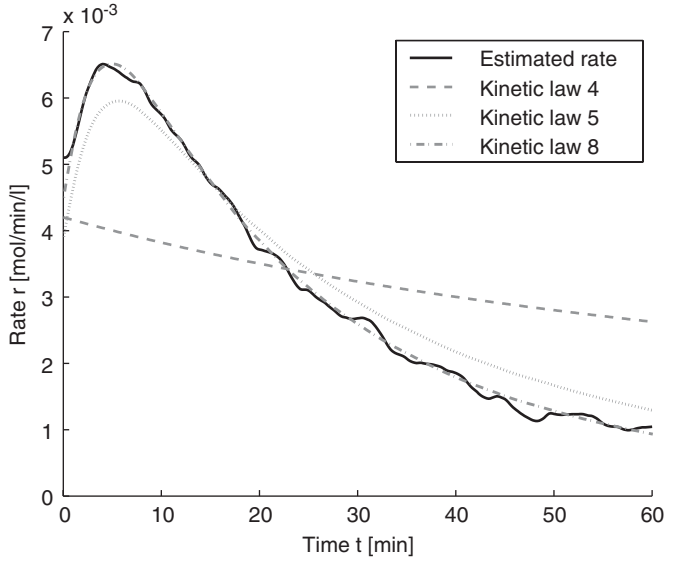

Fig. 6. Fit to the kinetic laws.

the models used for data simulation. The estimated parameter values come out to be very close to the parameters taken for simulation. Identification of the system using the proposed incremental procedure requires 42 s on a standard PC (1.5 GHz).

For comparison, a simultaneous identification was applied to the data given, requiring dynamic parameter estimation for each combination of kinetic models and subsequent model discrimination. The simultaneous procedure correctly identifies the number of reactions and the corresponding kinetics. The reaction parameters are calculated as $k_a = 0.0532\,\mathrm{l}^2/(\mathrm{mol}^2\ \mathrm{min})$, $k_b = 0.1281\,\mathrm{l}^2/(\mathrm{mol}^2\ \mathrm{min})$ and $k_c = 0.0280\ 1/\mathrm{min}$, giving a slightly better fit compared to the incremental identification results. However, the computational cost is excessive, lying in the order of 122,000 s or 34 h. Using incremental identification, an excellent approximation is calculated in only a fraction of time. The advantage in computational cost is especially thrilling for multiple reactions with a high number of candidate kinetics.

If the data quality is poor, i.e., high noise on the data and/or infrequent data, a single experiment may be insufficient to identify the reaction system. In this case, incremental identification can be fully integrated in iterative experiment planning using experimental design techniques (see e.g. Asprey and Macchietto (2000)). To obtain statistically optimal parameters and corresponding parameter confidence intervals, a subsequent simultaneous parameter estimation on the model obtained from incremental identification can be performed with good starting values. In the case considered, such correction of parameters only requires an additional 8.2 s of computing time.

6. Conclusions

An incremental approach has been presented for the identification of complex reaction systems. The problem is decomposed into a sequence of simple subproblems, allowing stepwise identification of model parts. This way, the number, stoichiometries and kinetics of the occurring reactions can be determined efficiently.

Maximum decoupling of the physical phenomena is achieved for arbitrary complexity of the reaction system. As system dynamics are fully covered in the flux estimation, they can be omitted in the following, thereby generating purely algebraic regression and structure discrimination problems. Much fewer model candidates are required in each step due to problem decoupling. The simplicity of the individual subproblems leads to drastically reduced calculation times. Moreover, this renders the method more robust compared to conventional simultaneous parameter estimation, e.g. by avoiding the difficulties in choosing suitable parameter (re-)initialization. The approach allows efficient use of a priori knowledge and, in each step, novel information is determined. This helps gain physical insight and choose models for the following step. If no feasible model structure can be set up, the previously calculated

Table 4
Comparison between simultaneous and incremental identification

Approach	Incremental	Simultaneous
Problem complexity	Decoupled sub-problems • Linear flux estimation • Algebraic regression problems ⇒ Low computational effort	Full problem • Dynamic parameter estimation required for each candidate ⇒ High computational effort
Initialization and convergence	• Robustness due to low subproblem complexity	• Suitable (re-)initialization may be difficult
Support of modeling process	• Incremental testing of submodels • Physical insight gained • Flexible use of physically founded and data-driven submodels	• Only lumped effect of model assumptions seen in output • Data-driven models require specific training algorithms
Data resolution	• Designed for high-resolution data • Sufficient data density required	• Designed for low resolution
Accuracy	• Parameter values biased but good approximation	• Statistically optimal estimates
Applicability	• Arbitrary problem complexity • Only partial system identification achieved in certain cases	• Arbitrary problem complexity

quantities may serve as input to some data-driven function approximation technique (Brendel et al., 2003; Brendel, 2005).

The concept is particularly efficient for the computationally expensive discrimination between model candidates. In these cases, substantial savings in computational effort can be experienced, especially for multiple reaction systems with a large number of kinetic model candidates. Note, however, that the estimated parameters may contain a bias, introduced by the solution of an infinite dimensional estimation problem involved in the initial flux estimation and propagated through the subsequent steps. Hence, the approach requires a sufficient amount of "good" data, since the bias introduced decreases with the number of data and a better signal-to-noise ratio. Subsequent parameter correction ensures statistically optimal parameters and allows the calculation of confidence intervals. In some problem settings, the incremental procedure allows only partial identification of the system, i.e., identification of a subset of reaction kinetics, whereas formulation of a simultaneous identification problem can lead to full identification. In such cases, incremental identification can be pursued first for the identifiable kinetics, benefitting from the advantages of the approach. The remaining unknown submodels are then identified (if applicable) using a simultaneous formulation of the already reduced identification problem. In this sense, the incremental identification approach does not aim at replacing conventional simultaneous model identification, but rather represents a valuable complement. The main characteristics of the incremental and simultaneous identification strategies are summarized in Table 4.

The focus of the paper has been set on the identification of homogeneous isothermal reaction systems, but the concept can be extended to the identification of non-isothermal and multi-phase systems including mass transport between phases (Brendel, 2005). The versatility of the approach produces a powerful model identification framework for constructing both physically motivated and possibly hybrid models, depending on the available knowledge.

Efficient interplay between incremental and simultaneous identification is currently investigated. Further work will also deal with the reduction of bias in the flux estimates and the calculation of reliable error estimates, representing important steps towards enhanced prediction accuracies.

Acknowledgments

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center (SFB 540) "Model-based experimental analysis of kinetic phenomena in fluid multiphase reactive systems".

References

Abramovich, F., Silverman, B.W., 1998. Wavelet decomposition approaches to statistical inverse problems. Biometrika 85 (1), 115–129.

Akaike, H., 1974. A new look at statistical model identification. IEEE Transactions on Automatic Control 19, 716–722.

Alsmeyer, F., Marquardt, W., Olf, G., 2002. A new method for phase equilibrium measurements in reacting mixtures. Fluid Phase Equilibria 203, 31–51.

Amrhein, M., Srinivasan, B., Bonvin, D., 1999. Target factor analysis of reaction data: use of data pre-treatment and reaction–invariant relationships. Chemical Engineering Science 54, 579–591.

Asprey, S.P., Macchietto, S., 2000. Statistical tools for optimal dynamic model building. Computers in Chemical Engineering 24, 1261–1267.

Bard, Y., 1974. Nonlinear Parameter Estimation. Academic Press, New York.

Bardow, A., Marquardt, W., 2004a. Identification of diffusive transport by means of an incremental approach. Computers in Chemical Engineering 28 (5), 585–595.

Bardow, A., Marquardt, W., 2004b. Incremental and simultaneous identification of reaction kinetics: methods and comparison. Chemical Engineering Science 59 (13), 2673–2684.

Bardow, A., Marquardt, W., Göke, V., Koß, H.-J., Lucas, K., 2003. Model-based measurement of diffusion using Raman spectroscopy. A.I.Ch.E. Journal 49 (2), 323–334.

Binder, T., Blank, L., Dahmen, W., Marquardt, W., 2002. On the regularization of dynamic data reconciliation problems. Journal of Processing and Control 12 (4), 557–567.

Björck, A., 1996. Numerical Methods for Least Squares Problems. SIAM, Philadelphia.

Bonvin, D., Rippin, D.W.T., 1990. Target factor analysis for the identification of stoichiometric models. Chemical Engineering Science 45 (12), 3417–3426.

Brendel, M., 2005. Incremental identification of complex reaction systems. Ph.D. Thesis, RWTH Aachen.

Brendel, M., Mhamdi, A., Bonvin, D., Marquardt, W., 2003. An incremental approach for the identification of reaction kinetics. In: Preprints of the Seventh IFAC Symposium on Advanced Control of Chemical Processes, ADCHEM 2003, HongKong, pp. 177–182.

Chang, J.-S., Hung, B.-C., 2002. Optimization of batch polymerization reactors using neural-network rate-function models. Industrial and Engineering Chemistry Research 11, 2716–2727.

Connors, K.A., 1990. Chemical Kinetics: The Study of Reaction Rates in Solution. VCH publishers, New York.

Craven, P., Wahba, G., 1979. Smoothing noisy data with spline functions. Numerical Mathematics 31, 377–403.

Engl, H.W., Hanke, M., Neubauer, A., 1996. Regularization of Inverse Problems. Kluwer Academic Publishers, Dordrecht.

Froment, G.F., Bischoff, K.B., 1990. Chemical Reactor Analysis and Design. Wiley, New York.

Gill, P.E., Murray, W., 1978. Numerically stable methods for quadratic programming. Mathematical Programming 14, 349–372.

Hansen, P.C., 1998. Rank-deficient and Discrete Ill-posed Problems. SIAM, Philadelphia.

Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, Cambridge.

Holland, D.H., Rayford, G.A., 1989. Fundamentals of Chemical Reaction Engineering. Prentice-Hall, New Jersey.

MacKay, D.J.C., 1992. Bayesian interpolation. Neural Computation 4 (3), 415–447.

Malinowski, E.R., 1991. Factor Analysis in Chemistry. Wiley, New York.

Marquardt, W., 2002. Adaptivity in process systems modeling. In: Grievink, J., van Schijndel, J. (Eds.), Proceedings of the European Symposium on Computer Aided Process Engineering (ESCAPE-12), Elsevier, The Hague, The Netherlands, pp. 42–56.

Mhamdi, A., Marquardt, W., 1999. An inversion approach to the estimation of reaction rates in chemical reactors. In: Proceedings of the European Control Conference (ECC'99), Karlsruhe, Germany, paper F1004-1.

Mhamdi, A., Marquardt, W., 2003. Estimation of reaction rates by nonlinear system inversion. In: Preprints of the Seventh IFAC Symposium on Advanced Control of Chemical Processes, ADCHEM 2003. HongKong, pp. 171–176.

Pothen, A., Fan, C.-J., 1990. Computing the block triangular form of a sparse matrix. ACM Transactions in Mathematical Software 16 (4), 303–324.

Ruppen, D., 1994. A contribution to the implementation of adaptive optimal operation for discontinuous chemical reactors. Ph.D. Thesis, ETH Zuerich.

Ruppen, D., Bonvin, D., Rippin, D.W.T., 1998. Implementation of adaptive optimal operation for a semi-batch reaction system. Computers in Chemical Engineering 22 (1–2), 185–199.

Stewart, W.E., Shon, Y., Box, G.E.P., 1998. Discrimination and goodness of fit of multiresponse mechanistic models. A.I.Ch.E. Journal 44 (6), 1404–1412.

Stoer, J., 1971. On the numerical solution of constrained least-squares problems. SIAM Journal of Numerical Analysis 8, 382–411.

Tholudur, A., Ramirez, W.F., 1999. Neural-network modeling and optimization of induced foreign protein production. A.I.Ch.E. Journal 45 (8), 1660–1670.

Van Lith, P.F., Betlem, B.H.L., Roffel, B., 2002. A structured modeling approach for dynamic hybrid fuzzy first-principles models. Journal of Processing and Control 12, 605–615.

Verheijen, P.J.T., 2003. Model selection: an overview of practices in chemical engineering. In: Asprey, S.P., Macchietto, S. (Eds.), Dynamic Model Development: Methods, Theory and Applications. Elsevier, Amsterdam, pp. 85–104.

Yeow, Y.L., Wickramasinghe, S.R., Han, B., Leong, Y.-K., 2003. A new method of processing the time–concentration data of reaction kinetics. Chemical Engineering Science 58, 3601–3610.

Yeow, Y.L., Pokethitiyook, P., Cheah, M.Y., Dang, H.D.T., Law, C.K.P., 2004. An alternative way of analyzing the progress curves of enzyme-catalyzed reactions. Biochemical Engineering Journal 21, 1–10.