Inference of chemical reaction networks using mixed integer linear programming

Mark J. Willis*, Moritz von Stosch

School of Chemical Engineering and Advanced Materials, University of Newcastle, Newcastle upon Tyne, NE1 7RU, United Kingdom

ABSTRACT

The manual determination of chemical reaction networks (CRN) and reaction rate equations is cumbersome and becomes workload prohibitive for large systems. In this paper, a framework is developed that allows an almost entirely automated recovery of sets of reactions comprising a CRN using experimental data. A global CRN structure is used describing all feasible chemical reactions between chemical species, i.e. a superstructure. Network search within this superstructure using mixed integer linear programming (MILP) is designed to promote sparse connectivity and can integrate known structural properties using linear constraints. The identification procedure is successfully demonstrated using simulated noisy data for linear CRNs comprising two to seven species (modelling networks that can comprise up to forty two reactions) and for batch operation of the nonlinear Van de Vusse reaction. A further case study using real experimental data from a biodiesel reaction is also provided.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

A major barrier in the transition from chemistry research to process development is that more quantitative information, regarding a chemical synthesis, is needed. For process development the over-riding concern is to mathematically characterise the route from reactants to products, which often occurs over several reaction steps, with the involvement of measurable intermediates. The objective is to develop stoichiometric and kinetic descriptions of chemical reactions as opposed to obtaining a detailed mechanistic understanding of a synthetic route. For multiple reaction systems, this is referred to as a chemical reaction network (CRN). A mathematical model of a CRN, written as a set of coupled non-linear ordinary differential equations (ODEs) describing the dynamic behaviour of the system, instantiates a CRN in commercial process simulation and optimization software. Software of this nature is required for numerous reasons including, accurate and economic plant design and process optimization (Maria, 2004) and so methods, tools and procedures for rapidly establishing a CRN using experimental data are desirable.

In particular, methods that can be applied to data obtained from reaction systems operating far away from chemical or biochemical equilibrium are of interest. This is because batch and semi-batch reactors – rather than continuous stirred tank reactors (CSTRs) – tend to be used in the fine chemical and pharmaceutical industries during the new chemical entity development lifecycle. Furthermore, the increased uptake of high throughput technologies e.g. automated robotic workstations for performing many experiments in parallel, coupled with improved sensor technology is likely to provide an increase in the quantity and quality of non-equilibrium experimental reaction data.

The work by Aris and Mah (1963) has been the basis for many stoichiometric and kinetic modelling studies aimed at CRN determination. One of the earliest advances being made by Bonvin and Rippin (1990) who proposed target factor analysis (TFA). This may be used to identify the number of linearly independent reactions. It may also be used to test suggested reaction stoichiometry is consistent with experimental data. Using TFA as a basis Brendel et al. (2006) and Bhatt et al. (2012) demonstrate an incremental identification strategy for CRNs. In their step-wise procedure, reaction stoichiometry (the CRN structure) is identified using TFA and then kinetic model identification strategies are used to determine the most appropriate ODE model (specifically, the structure and parameters of the rate laws). This decomposition of the problem allows for systematic development of a CRN. However, as stepwise methods are essentially local search operators they may produce sub-optimal solutions (local minima to the global optimization problem).

---

* Corresponding author.
E-mail addresses: mark.willis@ncl.ac.uk (M.J. Willis),
moritz.von-stosch@ncl.ac.uk (M.v. Stosch).

A contrasting approach is to parameterise a suitable ODE model structure directly from the observed data and to use the resulting model to infer network properties. Domain dependent knowledge can be exploited to narrow the network search space, by restricting the form of the ODE model used to explain the dynamic behaviour of the CRN. For instance, (bio) chemical reactions, occurring in well mixed, relatively dilute, homogeneous phases – such as may be found in controlled laboratory batch and fed-batch experiments – typically obey the law of mass action kinetics. This allows a class of physically interpretable ODE models with pseudo-linear properties to be formulated, enabling classical regression techniques to be applied to the CRN model search process. While Crampin et al. (2004), Searson et al. (2007), Srividhya et al. (2007), Burnham et al. (2008), Searson et al. (2012), Hii et al. (2014) demonstrate the potential of this approach the identification methods used fail to consistently predict an underlying network structure. Potential explanations for this are that, (a) Crampin et al. (2004), Srividhya et al. (2007), Burnham et al. (2008) use step-wise identification procedures which are susceptible to local minima, (b) while Searson et al. (2007), Searson et al. (2012), Hii et al. (2014) use evolutionary algorithms (EAs) structural constraints are difficult to incorporate into any EA (it is generally left to the objective function to manage and quantify possible structural infeasibility rather than directly imposing these constraints as part of network search). Therefore, an EA will perform poorly if the search space is highly constrained.

MILP (and variants such as mixed integer quadratic programming, integer linear programming etc.) has been used for many years for process synthesis, scheduling and control and a vast amount of literature are available, e.g. Grossmann (1985), Achenie and Biegler (1990), Raman and Grossmann (1991, 1992), Floudas and Lin (2005), Moro and Grossmann (2013) are a selection of examples of the literature in this area. To synthesize process flowsheets using MILP, a superstructure e.g. see Achenie and Biegler (1990) is often used. This is constructed to contain all possible alternatives of a potential process flowsheet of which the optimal solution belongs. The use of a superstructure for process synthesis has proved an effective tool in many application studies. Adopting this approach for CRN elucidation, defines a global model structure (which can be represented as a digraph) consistent with all possible kinetic rate terms arising from elementary chemical reactions. In effect this simplifies the difficult task of simultaneous structure and parameter estimation to one of just parameter estimation, where correct estimation of the parameters (links within the digraph) is, in principle, sufficient to deduce the structure of the underlying reaction network.

In this work, parameter estimation is achieved through minimization of the sum of the absolute errors (the L_1 norm) between measured and predicted species concentrations. In the absence of additional constraints on the structure of the system equations any identification strategy will normally over-fit the observed data with terms being included which model measurement noise rather than actual system dynamics. This would have negative effects on both the portability of the model (its ability to model different instances of the system) and the interpretability of the model (vital for network identification). Therefore, it is preferable to introduce additional information into the cost function in order to balance the trade-off between model complexity (in this case, the number of reactions) and how well the model fits the data. A number of regularization techniques are available, including ridge regression (Hoerl and Kennard, 1970) and the Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani (1996) e.g. see a review by Hesterberg et al. (2008). The method proposed in this paper is conceptually similar to these techniques however, a set of binary variables (associated with each of the parameters of the model) are used to perform regularization rather than the parameters themselves. The binary variables provide a normalised entropy
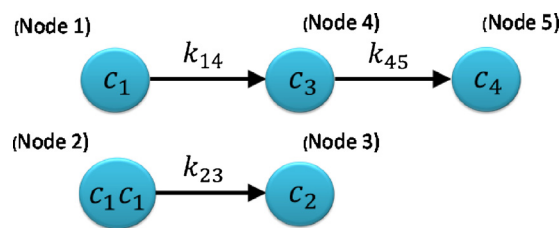

Fig. 1. A weighted digraph representing the Van de Vusse reaction with nodes $\boldsymbol{m} = \begin{bmatrix} c_1 & c_1^2 & c_2 & c_3 & c_4 \end{bmatrix}^T$ edges labelled with isothermal rate constants.
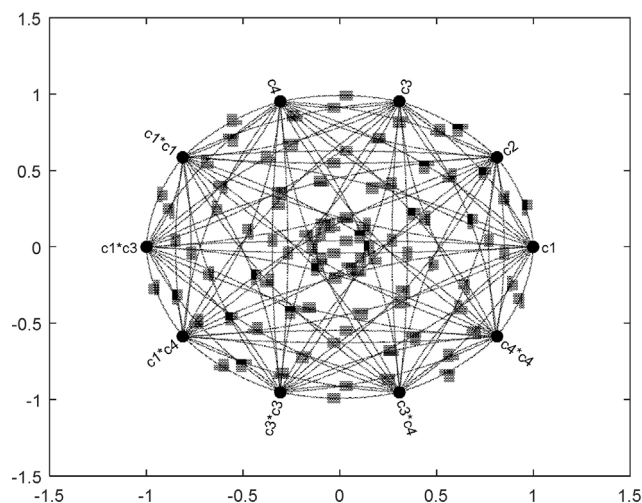

Fig. 2. An unweighted directed graph, representing the reaction superstructure for chemical reactions (1) (all possible reactions between the species—which admits ninety possible chemical reactions). The digraph was generated in MATLAB 2015b using the built in digraph command.

measure (independent of the magnitude of the regression parameters) and are directly related to the number of chemical reactions within the network. Whilst determination of these binary variables increases the number of parameters being identified as part of the optimization process, they are also used to indicate 'options' (e.g. between different reactions or species combinations) effectively turning 'on' or 'off' alternative solutions to an optimization problem. This provides a flexible identification framework that may be used to incorporate known information about the chemical species and reactions in the form of linear equality and inequality constraints. Furthermore, the MILP can be solved using fast, efficient and readily available commercial (and open source) software.

2. Chemical reaction networks

A CRN may be represented by a directed graph (or digraph). For example, Fig. 1 shows a weighted di-graph representation of the following CRN comprising four reactive species $x_1, \ldots, x_4$ involved in three reactions (known as the Van de Vusse reaction scheme),

$$2x_1 \rightarrow x_2 \tag{1}$$

$$x_1 \rightarrow x_3 \rightarrow x_4$$
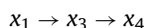
The graph is labelled with nodes, $\boldsymbol{m} = \begin{bmatrix} c_1 & c_1^2 & c_2 & c_3 & c_4 \end{bmatrix}^T$ and there are connections between the nodes which are referred to as edges (or links) in the graph. The links have a specified direction—as indicated by the arrows. The nodes are associated with the concentration of chemical species and are defined using