17th European Symposium on Computer Aided Process Engineeesing – ESCAPE17
V. Plesu and P.S. Agachi (Editors)

Systematic qualitative experimental design based upon identifiability analysis

Florin Paul Davidescu,[a] Henrik Madsen,[b] Sten Bay Jørgensen[a]

[a]*CAPEC, Technical University of Denmark , DK-2800 Kgs. Lyngby, Denmark, fpd@kt.dtu.dk, sbj@kt.dtu.dk*
[b]*Informatics and Mathematical Modelling, Technical University of Denmark , DK-2800 Kgs. Lyngby, Denmark, hm@imm.dtu.dk*

Abstract

Parameter identifiability constitutes an important issue for development of quantitative dynamic process models from experimental data. The focus of this paper is on the presentation of a methodology for parameter identifiability, which can be systematically applied for qualitative experimental design. An enzymatic reaction network is used as a case study to illustrate the methodological aspects.

Keywords

Parameter identifiability, Lie algebra, qualitative experimental design

1. Introduction

There is an increasing interest in producing complex intermediates and expensive fine chemicals in the pharmaceutical industry using biochemical synthesis. To develop a purely enzymatic synthesis for complex molecules from inexpensive substrates, large reaction networks are necessary. One way to achieve such a functional network is by using a System of Biotransformations (SBT). The SBT is based on a microorganism's metabolic network containing the synthesis path including cofactor regeneration reactions down to the desired product, which most often is an intermediate in the metabolic network. Hence expression of the enzymes catalyzing reactions from this intermediate are

turned off prior to the extraction i.e the genes are knocked-out. The SBT is used as cell free extract in the production phase, combining the easy handling of a viable culture with the advantages of *in vitro* bio-transformations [1]. The general goal of this study is to identify the bottlenecks of the SBT, to describe them qualitatively and subsequently to optimize the productivity of the reaction network. The workhorse of the de-bottlenecking and optimization process is a model describing the biochemical reaction network with good long-term prediction properties. Since the system under investigation is a system of high dynamics and complexity it is not realistic to develop a "perfect model" from first principle engineering methods. Thus in this work a grey-box stochastic model development framework [2] is used to develop a state space model. For this purpose it is necessary to assess the parameters identifiability. After introduction and demonstration of the identifiability method, a systematic qualitative experimental design methodology is introduced based upon application of the identifiability analysis. The latter aspect constitutes the main contribution of this paper.

2. Problem Statement, background

Once a model has been formulated it is necessary to assess the identifiability of the parameters and to determine which should be the inputs and the outputs of the experiments in order to be able to make all the model parameters identifiable [3]. Concerning identifiability analysis then a parameter set can be globally identifiable, which means that there is only one unique solution; or locally identifiable, which means that there are more than one solution; or simply unidentifiable, in which case there is no solution for this parameter set. For nonlinear models, there are only very few methods for assessing parameter identifiability. The first tool to be mentioned is the local or global (multi local) sensitivity analysis together with the co-linearity indexes [6], which is one of the widely used methods and the second method is an optimization-based approach, proposed recently by [4]. The theoretical-based methods are applicable only to some classes of models. The first methods listed are based on differential algebra and Grobner basis; the second methods are based on Taylor or generating series [1,5]. The advantages of the theoretical methods, are, first that they constitute a definitive test and, second, are unaffected by scaling. However they are limited to some classes of models and, they can only be applied for relatively small problems. The method used in this contribution is the last method described above, based on generating series. The application of the method is limited however to the deterministic part of the model and assuming the data are error free. Model parameter identifiability is investigated first, and then qualitative experimental design is applied.

3. Identifiability analysis

This section describes briefly, the theoretical identifiability method based on generating series and Lie algebra [5]. For batch models the method is similar to the Taylor based method, but for fed-batch models the generating series-based method is simpler. In the given model structure $M(\theta)$ in equation (1), the f_0 functions are represented by the part of the right hand-side of the equations which is not affected by the feed flow-rate; while the part affected by the flow-rate are represented by the f_i functions.

$$\begin{cases} \dfrac{dx(t)}{dt} = f_0\left(x(t),\theta\right) + \sum_{i=1}^{m} u_i(t)\cdot f_i\left(x(t),\theta\right), \qquad x(0) = x_0(\theta) \\ y(t,\theta) = h\left(x(t),\theta\right) \end{cases} \tag{1}$$

For a given model as in equation 1 a series expansion can be written based on the Lie derivatives so that the model output can be expanded in series with respect to inputs and time [5] around an initial time denoted by sub index $_0$.

$$\begin{aligned} & y(t,\theta)|_0 = h|_0 + L_{f0}h|_0 + L_{f0}L_{f0}h|_0 + \ldots \underbrace{L_{f0}\ldots L_{f0}h}_{n \text{ times derivation}}|_0 + \\ & \sum_{i=1}^{m} u_i(t)\cdot L_{fi}L_{f0}h|_0 + \sum_{i=1}^{m} u_i(t)\cdot \sum_{j=0}^{d} L_{fi}L_{fj}L_{f0}h|_0 + \sum_{i=1}^{m} u_i(t)\cdot \sum_{j=0}^{d}\sum_{k=0}^{d} L_{fi}L_{fj}L_{fk}L_{f0}h|_0 + \\ & \sum_{i=1}^{m} u_i(t)\cdot \sum_{j=0}^{d}\sum_{k=0}^{d}\ldots\sum_{l=0}^{d} L_{fi}\ \underbrace{L_{fj}L_{fk}\ldots L_{fl}}_{n-1 \text{ times derivation}}\ L_{f0}h|_0 \end{aligned} \tag{2}$$

The index m represents the number of inputs and d is the maximum order of derivation. In the equation above $|_0$ indicates evaluation at $t=0$ in $h|_0=h(x(t),\theta)|_0$ and the terms $L_{fi}L_{fj}\ldots L_{fn}L_{f0}h(x(t),\ \theta)|_0$. Note that $L_{fi}h$ is the Lie derivative of h along the vector field f_i given by equation (3).

$$L_{fi}h\left(x(t),\theta\right) = \sum_{j=1}^{n} f_{i,j}\left(x(t),\theta\right)\cdot \tfrac{\partial}{\partial x_j} h\left(x(t),\theta\right) \tag{3}$$

In equation (3), n is the number of states and the index j is the j[th] element of the f_i function vector. Let $s(\theta)$ be the coefficients of the expansion given in equation 2, which are in fact combinations of Lie derivatives. The identity $M(\hat{\theta}) \equiv M(\tilde{\theta})$ translates into $s(\hat{\theta}) \equiv s(\tilde{\theta})$. One can therefore test the identifiability of $M(*)$ by calculating the number of solutions for $\hat{\theta}$ of the set of equations $s(\hat{\theta}) \equiv s(\tilde{\theta})$ [5]. The uniqueness of the solution is determined by solving $s(\hat{\theta}) = 0$. If the underlying system of equations has one solution then the parameter set is theoretically globally identifiable; if there are more solutions then the parameters at best may be only locally identifiable and if the system has no

solution the parameters are unidentifiable. Once the Lie brackets have been computed e.g., the $s(\hat{\theta})$ then a system of algebraic equations is obtained. The elements in the parameter vector θ are the initial model parameters and the initial values of the states. It should be mentioned that if two parameters e.g. a and b are unidentifiable then a combination of the two, e.g. the sum or the ratio could be identifiable. As discussed above, the model under investigation is an ode representation of an enzymatic reaction network, which has been obtained, and improved using the grey-box stochastic methodology mentioned in the introductory section [2]. The model equations are given below. Initially two species are measured, thus there are two h equations. The analysis will first focus on the batch model i.e. the f_0 functions. Since there are only two states, which are measured, only some of the parameters may be identifiable. The model investigated is given below:

$$
\begin{aligned}
&\frac{dc_{GL}}{dt} = -r_{1\max}\cdot\frac{c_{GL}}{c_{GL}+K_{11}} + \frac{F}{V}\cdot\left(c_{GLf} - c_{GL}\right)\\
&\frac{dc_{F16B}}{dt} = r_{1\max}\cdot\frac{c_{GL}}{c_{GL}+K_{11}} - r_{2\max}\cdot\frac{c_{F16B}}{c_{F16B}+K_{21}+\frac{c_{DHAP}\cdot c_{G3P}}{K_{22}}} - \frac{F}{V}\cdot c_{F16B}\\
&\frac{dc_{DHAP}}{dt} = r_{2\max}\cdot\frac{c_{F16B}}{c_{F16B}+K_{21}+\frac{c_{DHAP}\cdot c_{G3P}}{K_{22}}} - r_{6\max}\cdot c_{DHAP} - \frac{F}{V}\cdot c_{DHAP}\\
&\frac{dc_{G3P}}{dt} = r_{2\max}\cdot\frac{c_{F16B}}{c_{F16B}+K_{21}+\frac{c_{DHAP}\cdot c_{G3P}}{K_{22}}} - r_{3\max}\cdot c_{G3P}\cdot c_{NAD} - \frac{F}{V}\cdot c_{G3P}\\
&\frac{dc_{PYR}}{dt} = r_{3\max}\cdot c_{G3P}\cdot c_{NAD} - r_{4\max}\cdot c_{PYR} - \frac{F}{V}\cdot c_{PYR}\\
&\frac{dc_{LAC}}{dt} = r_{4\max}\cdot c_{PYR} - \frac{F}{V}\cdot c_{LAC}\\
&\frac{dc_{ATP}}{dt} = -2\cdot r_{1\max}\cdot\frac{c_{GL}}{c_{GL}+K_{11}} + 2\cdot r_{3\max}\cdot c_{G3P}\cdot c_{NAD} - r_{5\max}\cdot c_{ATP} - \frac{F}{V}\cdot c_{ATP}\\
&\frac{dc_{NAD}}{dt} = -r_{3\max}\cdot c_{G3P}\cdot c_{NAD} + r_{4\max}\cdot c_{PYR} - \frac{F}{V}\cdot c_{NAD}\\
&\frac{dc_{PO_4}}{dt} = -r_{3\max}\cdot c_{G3P}\cdot c_{NAD} + r_{5\max}\cdot c_{ATP} - \frac{F}{V}\cdot c_{PO_4}\\
&h_1 = c_{GL}\\
&h_2 = c_{DHAP}
\end{aligned}
\qquad (4)
$$

In the first order Lie derivatives there are 6 parameters. In order to obtain a determined system of equations it is necessary to compute two more of the higher order derivatives. Once the derivatives are computed, two more parameters appear in the higher order terms and therefore two more Lie derivatives have to be computed, thus the Lie derivatives up to the fourth order have been used for this analysis. The method has a drawback in the sense that there is no way to know a-priori how many Lie derivatives need to be computed in order to obtain a determined system of equations. The following nine parameters r_{1max}, K_{21}, K_{22}, r_{6max}, r_{2max}, K_{11}, r_{3max}, r_{4max}, c_{GL0} appear in the Lie brackets equations. The system of equations has been solved for the parameters and multiple solutions were found, meaning that the parameter set is only locally identifiable. In the next step, all subsets with seven parameters of the original 9 were investigated. The analysis led to the same conclusion that the

reduced parameter set is only locally identifiable since multiple solutions were found. When further reducing to 6, all the combinations also proved to be only locally identifiable. The Lie brackets for the first measurement depends only on c_{GL}, r_{1max}, and K_{11} and thus including extra terms will be redundant, therefore only one Lie bracket corresponding to the first measurement, is included. This observation led to an even more simplified set of only 5 equations. The system of equations has been solved for all the combinations of 5 parameters obtained from the full parameter set. The analysis led to the same conclusion as before. Finally, only 4 parameter subsets were considered. This time, for several parameters subsets a unique solution was found, thus several particular combinations with four parameters are globally identifiable.

4. Qualitative experimental design

The generating series and Lie algebra introduced above can also be used in a systematic manner to determine which inputs should be perturbed and which outputs should be measured in order to be able to identify more if not all the model parameters. Here only a single input variable is considered, i.e., the feed flow-rate of the main reactant glucose, which appear in the f_1 functions, thus the operation mode is changed from batch to fed-batch. It is considered initially that the same compounds are being measured, thus, the measurements equations are the same as in equation (3). For this model, besides the first term, f_0, in the series, a first term in the right hand sum occurs, due to the presence of one input, F to the reactor. For this case it is possible to compute mixed terms of the Lie derivatives in order to include the influence of the input on the process. The following Lie derivatives have been computed for the analysis preceding the qualitative experimental design: $L_{f1}L_{f0}(h_i)$, $L_{f1}L_{f1}L_{f0}(h_i)$, $L_{f0}L_{f0}(h_i)$, $L_{f1}L_{f0}L_{f0}(h_i)$, $L_{f1}L_{f1}L_{f0}L_{f0}(h_i)$ where i indicate the measurement equations h_i. The first mixed term for the first measurement has been discarded from the analysis because it is redundant. The system of equations formed by the remaining six equations has been solved for all possible combinations of 6 parameters. Among the parameters to be investigated the initial concentration of glucose c_{GL} has been included as well as the flow-rate of glucose. The glucose feed concentration c_{GLfeed} is a known parameter and the feed flow-rate F is the process input. The result of the investigation showed that combinations of 6 parameters were globally identifiable. In order to render more parameters identifiable more measurements needs to be included in the analysis. Since ATP is involved in many of the balances it was decided to include it first. After ATP has been included among the h functions, the same Lie brackets were computed as for the other measurements. In the search for a system of equations that would make more parameters identifiable, four combinations each with seven parameters were found to be identifiable. It was noticed that, in particular r_{4max} and r_{5max} could not be identified simultaneously. The balance for Lactate is related only to the r_{4max} parameter thus indicating that a measurement of Lactate could

improve identifiability. Hence a measurement of Lactate was included in the analysis. Then, two combinations of eight parameters can be uniquely identified, and furthermore this time the two parameters r_{4max} and r_{5max} can be identifiable together with other parameters. As an alternative to the Lactate measurement the Pyruvate measurement can be considered, and the same information obtained. Out of the total 10 parameters there are two parameters, which could not be identified, i.e. r_{6max} and K_{22}. These two parameters are related to the balances for DHAP and F16B and G3P. A measurement for DHAP is already included in the original measurement set; therefore in the next step F16B is included as an extra measurement. Subsequently, the parameter K_{22} became identifiable. All the attempts to find a system of equations formed with the various equations to obtain a unique solution for all the ten parameters were unfruitful since r_{6max} remains un-identifiable. Since r_{6max} does not influence the balances for the rest of the compounds, the last remaining option is to consider simultaneously the measurements of F16B and G3P. In the last step of the analysis therefore the two measurements were included simultaneously. The search for a combination of equations that would give a unique solution for all the parameters still was unfruitful. The parameter r_{6max} remained unidentifiable even if all the states would be measured. To render r_{6max} identifiable the product of the reaction degrading DHAP should be considered. Thus after this analysis, it is concluded that five measured species, GL, DHAP, ATP, LAC, F16B or GL, DHAP, ATP, PYR, G3P would give optimum information in terms of minimum measured outputs and maximum number of identifiable parameters.

5. Conclusions and future work

Enzymatic reaction network models have been investigated for theoretical identifiability using a general method based upon generating series. The theoretical identifiablity analysis is subsequently used as a basis for quantitative experimental design. It is demonstrated that additional parameters can be identified by changing the operation mode to fed-batch. Furthermore the effect of inclusion of additional measurements upon parameter identifiability is clearly demonstrated.

References

[1]Michael Schümperli, Matthias Heinemann, Anne Kümmel and Sven Panke, in prep. 2006
[2]Niels Rode Kristensen, Henrik Madsen and Sten Bay Jørgensen, Comp. and Chem. Eng., 28 (2004), 1431-1449
[3]Walter Eric, Pronzato Luc, Automatica, 26, No 2, (1990), 195-213
[4]Asprey S.P., Macchietto, Comp. and Chem Eng., 24 (2000), 1261-1267
[5]Walter Eric, Pronzato Luc, Mathematics and Computers in Simulation, 42, (1996)
[6]Brun R., Reichert P., Künsch H. R. Water Resour. Res., 37 (4), 1015-1030