Parameter estimation combined with model reduction techniques for identifiability analysis of biological models

DOI:
10.1016/B978-0-443-15274-0.50167-0

Antonis Kokossis, Michael C. Georgiadis, Efstratios N. Pistikopoulos (Eds.)
PROCEEDINGS OF THE 33rd European Symposium on Computer Aided Process Engineering (ESCAPE33), June 18-21, 2023, Athens, Greece

Parameter estimation combined with model reduction techniques for identifiability analysis of biological models

Michael Binns,[a] Alessandro Usai,[b] Constantinos Theodoropoulos,[b]

[a]*Department of Chemical and Biochemical Engineering, Dongguk University-Seoul, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, Republic of Korea*
[b]*Department of Chemical Engineering, Biochemical and Bioprocess Engineering Group, The University of Manchester, Manchester M13 9PL, United Kingdom*
k.theodoropoulos@manchester.ac.uk

Abstract

Parameter estimation is typically used as part of model development to determine the values of unknown parameters. However, depending on the model complexity the number of parameters can also vary. High complexity models have large numbers of parameters requiring more computational effort to determine them and are also prone to overfitting. Low complexity models have smaller numbers of parameters but may have reduced accuracy. Based on available experimental data cross-validation can be used to compare different complexity models and determine the most appropriate complexity (James et al., 2013). Alternatively, it is possible to look at the identifiability of parameters based on experimental data which considers the sensitivity and correlation between parameters. Both these types of methodologies can be used to reduce the complexity of models such that insensitive and/or dependent/correlated can be removed or re-estimated and an alternative set of parameters can be computed. In this work both types of methods are explored with examples. Cross-validation combined with a Least Absolute Shrinkage and Selection Operator (LASSO) regularisation method is used to reduce the complexity of linear empirical equations for predicting the performance of downdraft biomass gasification (Binns and Ayub, 2021). Sensitivity and identifiability methods utilizing the Fischer Information Matrix (FIM) are used to reduce the complexity of a nonlinear system of partial integral differential equations describing a population balance model for microalgae cultivation (Usai et al., 2022). Application of these methods allows the number of parameters to be reduced depending on the tolerance and/or accuracy required.

Keywords: Model reduction, Parameter estimation, Optimisation, Identifiability, LASSO.

1. Introduction

Model development typically starts from some knowledge derived from existing experimental data and the equations which have previously been used to model similar systems. Based on this starting point a model might be suggested which includes all known physical variables and parameters which affect the system outputs. This approach might lead to a complex model for which there are a large number of unknown parameters. For this reason model reduction and identifiability methods can be used to reduce the complexity (Baker et al., 2015). In this study we compare two different

approaches for model reduction combined with parameter estimation: LASSO based regularization (James et al., 2013; Binns and Ayub, 2021) and a sensitivity and identifiability analysis utilizing the Fischer information matrix and multi-objective optimisation to reduce correlations and improve identifiability.

2. Parameter estimation methods

Parameter estimation involves minimising or reducing the difference between model predictions and experimental values by changing the values of parameters in the models. This is typically achieved by minimising a least-squares type function such as equation 1 where y_z is the experimental value, $\hat{y}_z$ is the corresponding model prediction and N is the number of data points.

$$RSS = \sum_{z=1}^{N} (y_z - \hat{y}_z)^2 \qquad (1)$$

The number of parameters to be estimated depends on the complexity of the terms used in the models to predict $\hat{y}_z$.

2.1. Parameter elimination

Reducing the complexity of models may be possible through the elimination of certain parameters which are either insensitive or do not affect the model outputs. There are potentially three ways by which parameters can be eliminated

- Set parameter to 0
- Set parameter to 1
- Set parameter to a fixed value

In cases where a fixed value is used this is defined using existing knowledge of the system. Alternatively setting values to 0 or 1 will typically allow the model equations to be written in a simpler form.

2.2. LASSO regularisation

$$RSS = \sum_{z=1}^{N} (y_z - \hat{y}_z)^2 + \lambda \sum_{i=1}^{n} |\beta_i| \qquad (2)$$

Regularisation is the methodology where the objective function from equation 1 is modified to include a second set of terms including the sum of the values of n parameters, β, multiplied by a factor, λ. When this function is minimised the values of parameters are also minimised and if the absolute values of β are used (as shown in equation 2) the effect is that a number of parameters are set equal to zero (James et al., 2013). Varying the value of λ will change the number of parameters set to zero. A very small λ value leads to the same solution obtained by minimising equation 1, a very large value λ will lead to a solution where all the parameters are set to zero. Additionally, this value will also affect the model accuracy, so cross-validation is required to find the most appropriate value of λ. In particular cross-validation involves dividing the training set into a number of sections or folds. One of the folds is selected for testing and the remainder are used for training. Sequentially repeating this for every fold will then lead to an overall cross validation mean square error. This cross validation is then repeated for every value of , λ being considered giving a range of different model solutions with varying accuracy and varying numbers of non-zero parameters.

2.3. Sensitivity and identifiability analysis

2.3.1. Sensitivity based parameter reduction

As mentioned in section 2.1 a parameter might be eliminated by setting its value to either zero, one or some pre-defined fixed value. Starting from some existing full model containing possibly a very large set of parameters the change in model output (y) resulting from and a small change in each parameter (P) will give sensitivity values as shown in equation 3. Then sequentially considering the elimination of each parameter by setting it to some fixed value and checking if the resulting model error exceeds some specified tolerance leads to a reduction of the number of parameters.

$$S_p = \frac{dy}{dP} \tag{3}$$

2.3.2. Identifiability analysis and correlations

After reducing the number of parameters either through the regularisation (as described in section 2.2) or utilizing sensitivities and a sequential removal method (as described in section 2.3.1) it is useful to consider the identifiability of the remaining parameters. Based on the available experimental data identifiability analysis asks which of those parameters can be uniquely identified. Those without a unique solution may have a range of acceptable values or it may not be possible to identify a reasonable value. Starting from the sensitivities it is possible to calculate the Fischer Information Matrix (FIM) which uses the local sensitivities around a fitted set of parameters (S_p) together with the weighted variance $(\hat{\Sigma})$.

$$FIM = \sum_{j=1}^{N_{exp}} \sum_{i=1}^{N_{sample}} \left(S_p(i,j)\right)^T \hat{\Sigma}^{-1} \left(S_p(i,j)\right) \tag{4}$$

The inverse of this FIM gives a lower bound for the covariance matrix which can be used to calculate the correlation matrix (Stoica and Ng, 1998; Baker et al., 2015). The following equations (5-8) are suggested in order to minimise correlations between parameters. From the correlation matrix $(\boldsymbol{\rho})$ it is possible to calculate an overall correlation coefficient as shown in equation 5. Subsequently from this correlation matrix the most highly correlated parameter (with the highest I_j) is chosen as a pivot parameter and other parameters which are strongly correlated with this are defined by linear relations as in equation 6.

$$I_j = \sum_{i=1}^{N_{parameters}} \rho_{ij} \tag{5}$$

$$P_i = C_{ij} P_j \tag{6}$$

$$Z_1 = min \left(\frac{y(C_{ij}P) - y(\hat{P})}{y(\hat{P})} \right)^2 \tag{7}$$

$$Z_2 = min\ I_j(C_{ij}) \tag{8}$$

To reduce the magnitude of correlations present the correlation values can be optimised (subject to any bounds on the final parameter values from the model) by minimising Z_2 as defined in equation 8. A second objective is given in equation 7 which is the deviation in model output between using the original parameters and those modified by changing

the correlation values. Hence, this is a multi-objective optimisation which should give a pareto curve of possible solutions. If parameters change without affecting model results this suggests they are not identifiable and can be eliminated or set to constant values.

3. Case studies

3.1. Empirical models for biomass gasification

Biomass gasification models should predict the outlet gas composition based on a number of input variables. In particular the input biomass composition $(Ultimate\ analysis: \%C, \%H, \%O, \%Ash)$, moisture content (MC), equivalence ratio (ER) and gasification temperature (T_{gas}) are expected to affect the output gas composition. For the gasifier considered here additional inputs are also available including grate rotation speed (Gr), gas fan speed (Fs), Wet bulk biomass density $(bulk)$ and biomass void percent $(void)$.

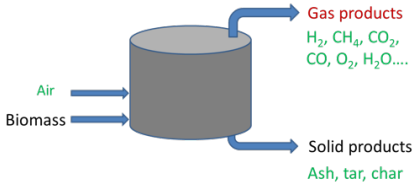

Figure 1. Block diagram representing a biomass gasifier.

Although more complex models are possible, for example based on knowledge of reaction kinetics, in this study simpler empirical models are considered relating the inputs with the outputs through linear and quadratic expressions. In this case the experimental data of Chee (1987) is used including 34 data points with inputs and outputs defined in Table 1. This is the same case investigated by Binns and Ayub (2021) who focused on finding the minimum cross-validation error model for each output. Here we consider the ranges of solutions which can be found with varying λ, the accuracy and simplicity of models obtained. For this purpose we consider the modelling of CO_2 volume percentage in the produced gas with a range of possible solutions shown in table 1 and figure 2. As can be seen from figure 2 and table 2 the model accuracy varies with changing value of λ. For lower values all eleven parameters are non-zero but the cross-validation error is highest. The lowest cross-validation error is achieved in a linear model with 7 non-zero parameters and at high values of λ all the parameters are zero and the model reduces to a constant value with a single fixed value.

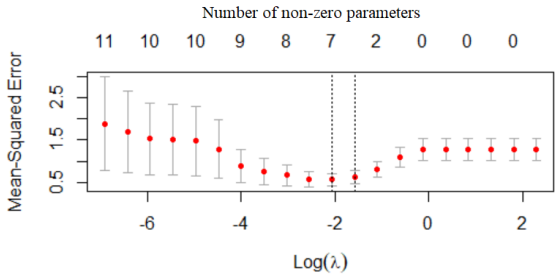

Figure 2. Cross-validation mean square error for CO_2 using a linear model with varying λ. The number of non-zero parameters excludes a constant fixed parameter (β_0).

Table 2. Parameter reduction using LASSO regularisation (Binns and Ayub, 2021)

$\lambda\ value$	Cross validation error	Expression
$\lambda = 0.001$	1.8851	$CO_2(\%) = \beta_0 + \beta_1 Tgas + \beta_2 ER + \beta_3 MC + \beta_4 H + \beta_5 O + \beta_6 C + \beta_7 Ash + \beta_8 Gr + \beta_9 Fs + \beta_{10} Bulk + \beta_{11} void$
$\lambda = 0.2069$	0.6402	$CO_2(\%) = \beta_0 + \beta_2 ER + \beta_3 MC + \beta_5 O + \beta_7 Ash + \beta_8 Gr + \beta_{10} Bulk + \beta_{11} void$
$\lambda = 0.3360$	0.8053	$CO_2(\%) = \beta_0 + \beta_2 ER + \beta_3 MC$
$\lambda = 1.4384$	1.2909	$CO_2(\%) = \beta_0$

3.2. Population balance models for microalgae cultivation

The population balance model considered here is for the growth of the microalgae *Haematococcus pluvialis* from the model developed by Usai et al. (2022). This model describes the growth (G), birth (B), disappearance of mother cells (M) and cell lysis (D) giving the overall population balance in equation 9.

$$\frac{\partial \Psi_V}{\partial t} + G(v) = B(v) - M(v) - D(v) \qquad (9)$$

where Ψ_V is the density distribution function of cells in the system.
The full model including expressions for each of the terms in equation 9 requires 34 parameters which are fitted to experimental data (Usai et al. 2022). It can be shown that applying sensitivity based parameter reduction 10 parameters can be eliminated with a tolerance of 0.05. Increasing the tolerance would allow more parameters to be eliminated (similar to the way increasing λ reduces the number of parameters).

Table 3. Parameter reduction using sensitivity approach (Usai et al., 2022b)

$Tolerance$	Parameters eliminated	Total parameters remaining
0.025	5	29
0.05	10	24
0.1	11	23

After the reduction of parameters, the sensitivity values can also be used to calculate the Fischer information matrix and subsequent covariance and correlation matrices. This identifies a number of strong correlations including existing relations between Monod kinetic parameters. If the objective function in equation 8 is minimised the magnitude of most correlations can be reduced by approximately 35%.

4. Discussion and conclusions

Two methods for model reduction based parameter estimation have been evaluated with two case studies and both are able to reduce complexity of the resulting models. A comparison of the two methods can be seen in table 4. The main advantage of the LASSO regularisation method is that it can start from a very complex model and directly reduce the number of parameters as part of the parameter fitting while the sensitivity and identifiability approach generally starts from a full fitted complex model before reducing

model complexity. However, the sensitivity and identifiability approach can more directly control the accuracy of the reduced model through the setting of a tolerance. It also offers more flexibility in terms of how to eliminate parameters, although possibly the LASSO regularisation could be modified so it can eliminate parameters is the same way. The main advantage of the sensitivity and identifiability methods are that they identify potential correlations between parameters and optimisation methods can be used to reduce the correlation values and potentially improve the identifiability of the model.

Table 4. Comparison of model reduction based parameter estimation methods

Issue	LASSO regularisation (Binns and Ayub, 2021)	Sensitivity and Identifiability analysis
Complex initial model	Can identify reduced models starting from a very complex model $\begin{pmatrix} N_{parameters} \gg N_{datapoints} \\ N_{reduced\ parameters} < N_{datapoints} \end{pmatrix}$	Requires full potentially complex model with fitted parameters as starting point
Accuracy vs. simplicity control	Control using λ factor (accuracy or model error must be calculated as an extra step)	Control using model tolerance
Parameter elimination	Set eliminated parameters to zero	Set eliminated parameters to either zero, one, or to specified values
Correlations and identifiability	Does not test for correlations. Reduces some statistical identifiability issues by reducing the number of parameters compared to the number of data points	Identifies potential correlations Can reduce correlation values to improve identifiability

References

S.M. Baker, C.H. Poskar, F. Schreiber, B.H. Junker, 2015, A unified framework for estimating parameters of kinetic biological models, BMC Bioinformatics, 16, 104

M. Binns, H.M.U. Ayub, 2021, Model reduction applied to empirical models for biomass gasification in downdraft gasifiers, Sustainability, 13, 12191

C.S. Chee, 1987, The air gasification of wood chips in a downdraft gasifier, Master's thesis, Kansas State University, Manhattan, USA.

G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning with applications in R, 2nd ed., Springer, New York, 2013, 59-126

P. Stoica, B.C. Ng, 1998, On the Cramer-rao bound under parametric constraints, IEEE Signal Procesing Letters, 5, 177-179.

A. Usai, J.K. Pittman, C. Theodoropoulos, 2022, A multiscale modelling approach for haematococcus pluvvialis cultivation under different environmental conditions, Biotechnology Reports, 36, e00771