MECHANIKA
CZASOPISMO TECHNICZNE
TECHNICAL TRANSACTIONS
MECHANICS
WYDAWNICTWO POLITECHNIKI KRAKOWSKIEJ
1-M/2012
ZESZYT 5
ROK 109
ISSUE 5
YEAR 109

DIANA LÓPEZ, TILMAN BARZ, HARVEY ARELLANO-GARCIA,
GÜNTER WOZNY*, ADRIANA VILLEGAS, SILVIA OCHOA**

SUBSET SELECTION FOR IMPROVED PARAMETER IDENTIFICATION IN A BIO-ETHANOL PRODUCTION PROCESS

PODZBIÓR SŁUŻĄCY DO PRECYZYJNIEJSZEGO OKREŚLANIA PARAMETRÓW W PROCESIE WYTWARZANIA BIOETANOLU

Abstract

A systematic approach for system identification is applied to experimental data of ethanol production from cellulose. Special attention is given to the identification of model parameters, which can be reliably estimated from available measurements. For this purpose, an identifiable parameter subset selection algorithm for nonlinear least squares parameter estimation is used. The procedure determines the parameters whose effects are unique and have a strong effect on the predicted (measurement variables) output variables. The system is described by a generic process model for the simultaneous saccharification and fermentation including three enzyme-catalyzed reactions. The process model is clearly over-parameterized. By applying the subset selection approach the parameter space is reduced to a reasonable subset, whose estimated parameters are still able to predict the experimental data accurately.

Keywords: identifiability analysis, subset selection, least squares, bio-ethanol, bagasse, SSF

Streszczenie

Systematyczne podejście do identyfikacji systemu stosowane jest wraz z doświadczalnymi danymi dotyczącymi wytwarzania etanolu z celulozy. Szczególną uwagę zwraca się na określanie parametrów modelu, które można wiarygodnie oszacować na podstawie ogólnodostępnych pomiarów. W tym celu zastosowano algorytm podzbioru parametru identyfikowalnego służący do nieliniowego szacowania parametrów metodą najmniejszych kwadratów. Procedura ta określa parametry, które dają niepowtarzalne efekty i wywierają silny wpływ na przewidywane zmienne zdolności produkcyjnej (zmienne pomiarów). System ten opisywany jest przez rodzajowy model procesu jednoczesnego scukrzania i fermentacji, wliczając w to trzy reakcje katalizowane enzymowo. Model procesowy jest nadmiernie sparametryzowany. Przy zastosowaniu opisywanego podejścia dana przestrzeń zostaje ograniczona do uzasadnionego podzbioru, którego szacowane parametry pozwalają nadal celnie przewidywać dane doświadczalne.

Słowa kluczowe: analiza identyfikowalności, podzbiór, metoda najmniejszych kwadratów, bioetanol, pozostałość, scukrzanie i fermentacja

* MSc. Diana López, PhD. Eng. Tilman Barz, PhD. Eng. Harvey Arellano-Garcia, prof. PhD. Eng. Günter Wozny Chair of Process Dynamics and Operation, TU Berlin.

** MSc. Adriana Villegas, PhD. Eng. Silvia Ochoa, Research Group in Process Simulation, Design, Control and Optimization (SIDCOP), University of Antioquia, Medellín, Colombia.

1. Introduction

Parameter estimation in biochemical models often means the determination of a relatively high number of kinetic parameters compared to the number of measured process variables. Moreover, due to the nonlinearity of kinetic models the parameter estimation problem most likely contains multiple local minima. The solution of interest is the global minimum, which hopefully also provides the biologically most reasonable parameters. However, due to the existence of multiple minima good initial estimates of the parameters are crucial to ensure that the obtained solution is close to a physiologically reasonable minimum. When a solution is found for the estimation problem it should be checked how robust the minimum is by re-running the estimation routine with the new parameter set as initial guesses. Furthermore the robustness may be evaluated by starting from different but also physiologically reasonable initial values. However, correlation between model parameters constitutes an obstacle to determining a unique minimizing parameter set [1].

In this paper, a methodology for identifying kinetic parameters in structured growth models is presented. The methodology is applied to a case study where kinetic parameters of a bio-ethanol production process are estimated. The focus lies on the identifiability analysis for the determination of model parameters, $\theta \in R^{Np}$, which can be reliably estimated from available measurements. For this purpose, an identifiable parameter Subset Selection (SsS) algorithm for nonlinear least squares parameter estimation is used, which is based on the ill-conditioned parameter selection [2, 3]. By fixing the ill-conditioned parameters at prior estimates, a reduced-order and well-conditioned parameter estimation problem is then solved where the remaining parameters are determined. In the subset selection algorithm, the sensitivity matrix (S) of the least squares problem is considered and a Singular Value Decomposition (SVD) is applied as rank-revealing factorization [4] within the algorithm of [2, 3]. The procedure permits to determine the parameters whose effects are unique (linear independent parameter sensitivities) and have a strong effect on the predicted measurement variables.

2. Case study – bio-ethanol production

Experimental data used in this case study was taken from bio-ethanol production in a saccharification and fermentation (SSF) process (see [5]). Experiments were carried out in a two-liter fermentor with a working volume of 0.6 l. The initial suspension contained dry weight solid content of 20% (w/w) considering a content of cellulose of 67% (w/w). Prior to the SSF process, an enzymatic pre-hydrolysis of 12 h at 47°C was realized to allow for the build-up of fermentable glucose. A commercial cellulase preparation was used with an enzymatic load of 26 FPU/ gram of solid of GC 220 – Genencor and 17 UI/ gram of solid of β-glucosidase with activities of 104 FPU/ml and 439 IU/ml, respectively. The concentration of protein per ml of GC 220 was 109 mg/ml and 127 mg/ml of ß-glucosidase. After 12 h, 6 g of microorganism/l was added and the process was continued at 37°C until completing 50 h. The micro-organism was a commercially available *Saccharomyces cerevisiae*.

The SSF is described by a generic model taken from [6–8] which considers the four main influencing factors for the kinetics of SSF: cellulosic substrate, cellulase and β-glucosidase

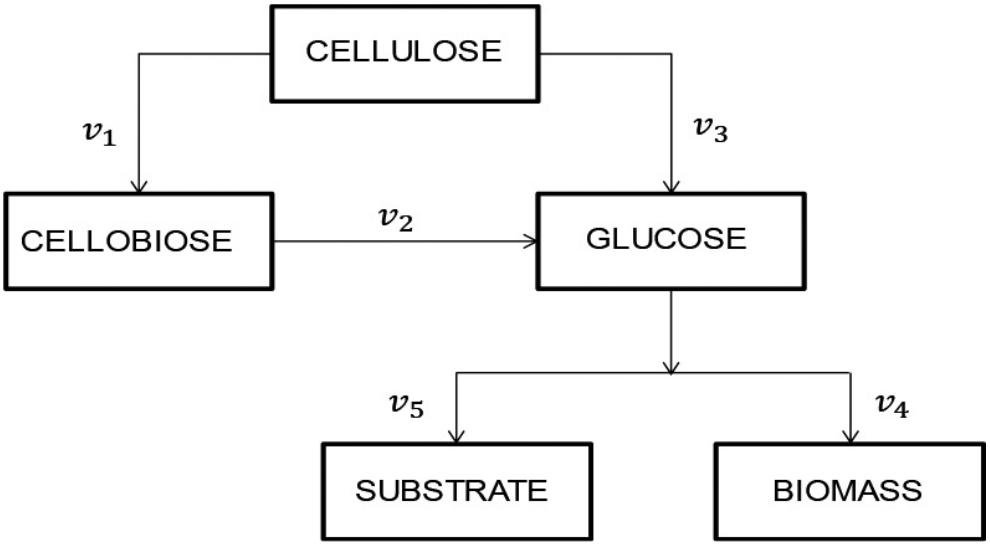

Fig. 1. Simplified reaction mechanisms in SSF processes [6]

Rys. 1. Uproszczone mechanizmy reakcji w procesach scukrzania i fermentacji [6]

enzyme system, substrate-enzyme interaction and enzyme-yeast interaction. The simplified reaction mechanisms are presented in Fig. 1, where cellulose is simultaneously hydrolyzed to cellobiose (v_1) and glucose (v_3), cellobiose is converted to glucose (v_2), and glucose is catabolized to ethanol, cell mass, and carbon dioxide by the fermentative microorganism. Yeast growth and glucose consumption rate are expressed by v_4 and v_5, respectively. The hydrolysis model is presented in Eq. 1–2, in which the effects of ethanol on cellulase are included [6]. For modeling glucose consumption and biomass formation, standard Monod kinetics were assumed, expanded to include ethanol inhibition on yeast [6]. Yeast growth rate and substrate consumption rate are in Eq. 3–4.

For reactions with cellulose as a substrate v_1 and v_3 (Eq. 1), the active amount of enzyme is assumed to be determined by enzyme adsorption onto the cellulose substrate (adsorption constant K_L). Inhibition by glucose to cellulase and ß-glucosidase was assumed through $K_{1,G}$ and $K_{2,G}$. For all three reactions, the zero-order rate constant is given as a function of temperature (activation energy E_a); in addition, all enzyme activity is assumed to be subject to thermal inactivation (K_D). The nature of the cellulose substrate is assumed to be conversion-dependent such that a recalcitrance constant *Krec* was used. On the other hand, ethanol inhibition is assumed to affect the rates of reactions v_1 and v_3 by inhibition constant $K_{1,EtOH}$ [6] and reaction (Eq. 2) by inhibition constant $K_{2,EtOH}$ [7].

For the cellulase adsorption to cellulose, within v_1 and v_2, it is not considered that substrate surface area is proportional to cellulose concentration [7], but it is considered as a constant lumped in the maximum specific rates of cellulose hydrolysis to cellobiose and glucose ($k_{max,1}$ and $k_{max,3}$, respectively). Thermal inactivation constant (K_D) follows an Arrhenius type relationship, $K_D(T) = A_D e^{-\Delta H/T}$.

$$v_i = \underbrace{\left(k_{\max,i}.\frac{C_E}{K_L + C_E}\right)}_{\text{Adsortion of Cellulase to Cellulose}} . \underbrace{\left(\frac{K_{1,G}}{K_{1,G} + C_G}\right)}_{\text{Inhibition by Glucose}} . \underbrace{\left(e^{-K_D(T).t}\right)}_{\text{Thermal inactivation of all enzyme activity}} \cdot \left(\frac{e^{\frac{-Ea}{RT}}}{e^{\frac{-Ea}{RT_{ref}}}}\right) . \underbrace{\left(\frac{K_{1,EtOH}}{K_{1,EtOH} + C_{EtOH}}\right)}_{\text{Inhibition of Cellulase by Ethanol}} . \underbrace{\left(e^{-Krec.\left(1-\frac{Cc}{Cc_0}\right)}\right)}_{\text{Inhibition by substrate recalcitrance}} \quad i = 1,3 \tag{1}$$

$$v_2 = \left(k_{\max,2}.e_g.e_T\right).C_{cb} \Bigg/ \underbrace{\left\{K_m\left(1+\frac{C_G}{K_{2,G}}\right)\right.}_{\text{Inhibition by Glucose}} + \underbrace{\left.C_{cb}\right\}}_{\text{Inhibition by Cellobiose}} . \underbrace{\left(e^{-K_D(T).t}\right)}_{\text{Thermal inactivation of all enzyme activity}} \cdot \left(\frac{e^{\frac{-Ea}{RT}}}{e^{\frac{-Ea}{RT_{ref}}}}\right) . \underbrace{\left(\frac{K_{2,\mathrm{EtOH}}}{K_{2,\mathrm{EtOH}} + C_{\mathrm{EtOH}}}\right)}_{\text{Inhibition by Ethanol}} \tag{2}$$

$$v_4 = \mu_{\max} . \underbrace{\frac{C_G}{K_G + C_G}}_{\text{Inhibition by Substrate}} . \underbrace{\left(\frac{K_{iy,EtOH}}{K_{iy,EtOH} + C_{EtOH}}\right)}_{\text{Inhibition by Product}} . C_x \tag{3}$$

$$\mathrm{v}_5 = \underbrace{-\frac{v_4}{Y_{xg}}}_{\text{Biomass production}} \quad \underbrace{-C_x.m_s}_{\text{Maintenance requirements}} \tag{4}$$

$$\begin{aligned} dC_C/dt &= -\left[v_1 + v_3\right] ; \qquad dC_{Cb}/dt = 1.056v_1 - v_2 ; \\ dC_G/dt &= 1.053v_2 + 1.111v_3 + v_5 ; \qquad dC_X/dt = v_4 \end{aligned} \tag{5}$$

$$C_{EtOH} = -0.511.\left[\begin{matrix} 1.111(C_c - C_{c0}) + 1.053.(C_{cb} - C_{cb0}) + \\ +(C_G - C_{G0}) + C_X - C_{X0} \end{matrix}\right] \tag{6}$$

$$dC_E/dt = -K_D C_E \tag{7}$$

3. Parameter identification: subset selection algorithm

The model parameter are determined solving the parameter estimation (PE) problem:

$$\hat{\theta} = \arg\min_{\theta}\left((Y - Y^m)^T (Y - Y^m)\right) \tag{8}$$

where $\hat{\theta}$ is an unbiased estimator containing the best currently available estimate of the true parameter vector θ^*, $Y^m \in R^{Ny.Nm}$ is the experimental data vector, Ny is the number of measurement variables and Nm is the discrete set of instances t_k when $y_i^m \in Y^m$ is measured; the collection of the predicted response variables $Y \in R^{Ny.Nm}$ are calculated in each t_k. The analysis of the identifiability of model parameters is done here based on the sensitivity matrix $S \in R^{Ny.Nm\ x\ Np}$:

$$S = \left[\partial y \Big/ \partial\theta \Big|_{t_1} \quad \partial y \Big/ \partial\theta \Big|_{t_2} \cdots \partial y \Big/ \partial\theta \Big|_{t_{Nm}} \right]^T \tag{9}$$

To account for significant differences in the magnitude of parameter values to be analyzed, the sensitivity matrix must be normalized, such that

$$s_{ij} = \left(\partial y_i / \partial\theta_j\right)\left(\max\left(\left|\theta_j\right|, \theta_{trsh}\right)\right) / \max\left(\left|y_i\right|, y_{trsh}\right)$$

where θ_{trsh} and y_{trsh} are the machine tolerance. All corresponding criteria for the SsS are adapted from the ill-conditioned parameter selection approach presented in [2, 3] and based on the analysis of the Hessian matrix $H_\theta = S^T S$. Generally, all parameters with low or non-existing sensitivities (columns of S with values equal or near to zero), or linearly dependent parameters are not identifiable. In both cases S is singular or "almost" singular from a numerical point of view. This situation is undesirable, because it reflects near indeterminacy in the parameter estimates, caused by having more parameters than can be reliably estimated from available measurements. Thus, applying the parameter subset selection, ill-conditioned parameters are fixed at prior estimates and reduced-order and well-conditioned PE is considered for the determination of the active parameters.

A rank-revealing factorization is done by the Singular Value Decomposition (SVD) of $S = U\Sigma V^T$, where $U \in R^{Ny.Nm\ x\ Ny.Nm}$ is a real or complex unitary matrix, $V^T \in R^{Np\ x\ Np}$ the conjugate transpose of V is a real or complex unitary matrix, and $\Sigma \in R^{Ny.Nm\ x\ Np}$ is a rectangular diagonal matrix with nonnegative real numbers on the diagonal. The diagonal entries $\Sigma_{i,i}$ are the singular values $\sigma_i > 0$ of matrix S such as $\sigma_1 > \sigma_2 > \ldots > \sigma_{Np}$. A criterion for the nearness to singularity of S is the condition number $\kappa(S) = \sigma_1/\sigma_{Np}$. A "very high" condition number of S indicates an almost singular sensitivity matrix. According to [3], an upper bound $\kappa^{max} \cong 1000$ is defined and parameter identifiability either of the original or a reduced problem is given, when $\kappa \le \kappa^{max}$ holds. Additionally, the Collinearity Index $\gamma = 1/\sigma_{Np}$ is considered as singularity measurement. γ equals one, if the columns of S are orthogonal and reaches infinity if the columns are linearly dependent. In [9] an empirically found threshold of $\gamma^{max} \cong 10$–15 has been named. Thus, if $\gamma > \gamma^{max}$ the corresponding parameter set is considered as poorly identifiable.

Steps within the SsS algorithm are: 1) For the current parameter set θ , compute the SVD of $S(\theta)$. 2) Evaluate singularity measurements based on the condition number $\kappa(S)$, and all sub-condition numbers $\kappa_j = \sigma_j/\sigma_{Np}$, with $j = 1, ..., Np\text{-}1$ for each σ_j available in the diagonal matrix S found in the above step. Calculate the collinearity index γ. If $\kappa(S) \le \kappa^{max}$ and $\gamma \le \gamma^{max}$ the parameter set is identifiable and the algorithm finishes, if not, go to the next step. 3) Determine r as set dimension of $\Gamma = \{\sigma_j \,|\, \kappa_j \le \kappa^{max}\}$, such that a maximum number of r singular values σ_j, with $j = 1,...,r$, are found, for which $\kappa_j \le \kappa^{max}$. 4) Determine a permutation matrix P by constructing a QR decomposition with column pivoting (QRP) for $S \in R^{Ny.Nm\ x\ Np}$ such that $SP = QR$, where $Q \in R^{Ny.Nm\ x\ Ny.Nm}$ is an orthogonal matrix, $R \in R^{Ny.Nm\ x\ Np}$ is an upper triangular matrix with decreasing diagonal elements and $P \in R^{Np\ x\ Np}$ is a permutation matrix which orders the columns of S according to linear independency, it means that the first columns of SP are the largest independent set of columns of S. 5) Use P to re-order the parameter vector θ according to $\tilde{\theta} = P^T\ \theta$. 6) Make the partition $\tilde{\theta} = [\tilde{\theta}^{(r)^t}, \tilde{\theta}^{(Np-r)^t}]^T$ with $\tilde{\theta}^{(r)}$ containing the first r elements of $\tilde{\theta}$. 7) Fix $\tilde{\theta}^{(Np-r)}$ to a priori estimate. 8) Solve reduced-order parameter estimation problem, considering $\tilde{\theta}^{(r)}$ only.

4. Determination of model parameters

Generally, the parameter estimation procedure is divided in six steps. First, initial guesses for the parameters have to be obtained either from literature or by performing simple model calculations using selected sets of experimental data. Secondly, identifiable subset selection of the parameters must be performed; thirdly, the parameters selected by parameter SsS procedure must be removed of the estimation problem by fixing them to appropriate values; fourthly, the new reduced parameter estimation problem must be run along with the new evaluation of the system identifiability (SsS algorithm). Fifthly, singularity measurements (κ and γ, see section 4) must be monitored to assure that the new reduced problem is well-conditioned; if the corresponding conditions are not fulfilled (thresholds are exceeded), the problem must be reduced again by fixing the ill-conditioned parameters found by SsS to the values calculated in the current optimization; this iterative process must be repeated until the singularity measurements do not exceed their corresponing thresholds. Finally, a statistical result analysis can be performed, e.g. analysis of the parameter accuracy by assessing their standard deviations using the covariance matrix of the estimates.

5. Results

Nonlinear regression in Eq. 8 was used based on the Levenberg-Marquardt least squares minimization algorithm, which is a hybrid of the Gauss-Newton and the steepest descent methods [7]. In Table 1 are the complete parameters for this differential algebraic equation system (DAE), where the parameter vector to be estimated is $\theta = [k_{\max,1}\ k_{\max,2}\ k_{\max,3}\ K_L\ K_{1,G}\ Krec\ Km\ K_{2,G}\ K_{iy,EtOH}\ m_s\ Yxg\ \mu_{\max}\ K_G\ K_{1,EtOH}\ K_{2,EtOH}]$ with $Np = 15$. The rest of parameters in Table 1 were maintained constants according to literature values in [6] taking to account these parameters did not exhibit changes according to previous sensitivity analysis. Measured variables in the experiment realized by [5] were Cellobiose (C_{cb}), Glucose (C_G) and Ethanol (C_{EtOH}) concentrations such that the experimental data vector was $Y^m = [C_{cb}{}^m\ C_G{}^m\ C_{EtOH}{}^m]$ with $Ny = 3$. $C_{cb}{}^m$, $C_G{}^m$, and $C_{EtOH}{}^m$ were sampled in time range of 0 to 50 h until having 17 measurements point ($Nm = 17$).

Following the procedure described in section 5, the first step was to find good initial parameter guesses for the nonlinear least squares algorithm. Four different initial estimates (*IE*) are depicted in Table 1; IE_1 and IE_2 make reference to literature parameters calculated by [6, 8] respectively; IE_3 is a set of parameter guesses, which were obtained by an independent consideration of the cellulose hydrolysis and glucose fermentation step and a separate solution of these problems along with subsequent re-optimizations until finding a stable initial estimate for future parameter estimations. IE_4 was created by changing the value of *Km* reported in [8] and taking into account that this parameter demonstrates the most sensitive parameter in all calculations done in this work. In Table 1 the objective function values (*OF*) obtained when running the parameter estimation from each *IE*. The best fit of the experimental data was found for IE_4 with $OF = 157$. Accordingly, all subsequent computations the values in IE_4 were used as initial guess.

As second step, identifiable subset selection of the parameters is performed using the normalized sensitivity matrix (see section 4). In Table 2, the columns “Estimated Parameter”

show parameter values that minimize Eq. 8, columns named "Sensitivity measure δ_j" show the Euclidean Norm of sensitivity matrix columns, and columns named "Identifiability Order" contain the order of the new parameter vector $\tilde{\theta}$ whose first *r* elements correspond to the identifiable parameter vector $\tilde{\theta}^{(r)}$. In Table 2, OPT_1 makes reference to the problem with the original parameter vector $\theta_1 = \theta$ and $Np_1 = 15$, with rank of sensitivity matrix $r_1 = 9$, objective function $OF_1 = 159.78$, condition number $\kappa_1 = 1.8 \times 10^7$ and collinearity index $\gamma_1 = 726.88$. From this is becomes clear that the estimation problem is ill-conditioned. despite of the good fitting of the experimental data (same values of OF_1 for all results OPT_1-OPT_3); strong correlation between the parameters of OPT_1 are indicated by high values of κ and γ. A SsS step gives 9 parameters $\theta^{1\,(r\,)}_{\ 1} = [k_{\max,1}\ k_{\max,2}\ K_L\ K_{1,G}\ Km\ K_{iy,Et\mathrm{OH}}\ Yxg\ K_{1,Et\mathrm{OH}}\ K_{2,Et\mathrm{OH}}]$ which are identifiable, the remaining 6 parameters $\theta^{1(\mathrm{Np\text{-}r}\,)}_{\ 1} = [k_{\max,3}\ Krec\ K_{2,G}\ m_s\ \mu_{\max}\ K_G]$ are discarded in the next step.

Table 1

Selection of the best Initial Estimate

PARAMETER		UNIT	IE_1	IE_2	IE_3	IE_4
1	$k_{\max,1}$	h^{-1}	0.0827	0.081	9.480	9.480
2	$k_{\max,2}$	$gU^{-1}h^{-1}$	0.00406	0.0108	8.161E-02	8.161E-02
3	$k_{\max,3}$	h^{-1}	0.0834	0.058	0.001	0.001
4	K_L	$FPUL^{-1}$	544.89	18.2	1386	1386
5	$K_{1,G}$	gL^{-1}	53.16	6.3	1486.1	1486.1
6	*Krec*	-	2.8*	2.8	1.719	1.719
7	Km	gL^{-1}	10.56	100**	1410	<u>10.56</u>
8	$K_{2,G}$	gL^{-1}	0.62	0.54	39.17	39.17
9	$K_{iy,Et\mathrm{OH}}$	gL^{-1}	50	50	55.19	55.19
10	m_s	h^{-1}	0	0.02	1.150×10^{-5}	1.150×10^{-5}
11	Yxg	gg^{-1}	0.113	0.11	1.809×10^{-4}	1.809×10^{-4}
12	$\mu_{\max}$	h^{-1}	0.19	0.25	6.914×10^{-2}	6.914×10^{-2}
13	K_G	gL^{-1}	0.000037	0.0252	15262.7	15262.7
14	$K_{1,Et\mathrm{OH}}$	gL^{-1}	50.35	95	17.43	17.43
15	$K_{2,Et\mathrm{OH}}$	gL^{-1}	500*	500*	31.8	31.8
OF			39490	446	159	157

* Parameter added to original model proposed by Authors

**Heuristic parameter value in order to overcome convergence problems

Thirdly, the 6 non-identifiable parameters from the last step were removed from parameter estimation problem by fixing them to values found in OPT_1. Fourthly, the new reduced estimation problem was run and again a SsS was performed; in Table 2 this new problem is referenced as OPT_2, in which $\theta^2 = \theta^{1\,(r\,)}_{\ 1}$, with $Np_2 = 9$, $r_2 = 7$, $OF_2 = 157.54$, $\kappa_2 = 8081$ and $\gamma_2 = 84.23$. The obtained reduction in the singularity measurements (κ_2 and γ_2) indicate that the current reduced problem is better conditioned than the original one (OPT_1) but still

there are two parameters which are not identifiable, indicated by the rank of the sensitivity matrix $r_2 = 7$ and since $\kappa_2 > \kappa^{max}$ and $\gamma_2 > \gamma^{max}$. The identifiable parameter vector for OPT_2 was $\theta^{2\,(r_2)} = [k_{max,1}\, k_{max,2}\, Km\, K_{iy,EtOH}\, Yxg\, K_{1,EtOH}\, K_{2,EtOH}]$, and the vector of parameters to fix was $\theta^{2\,(Np-r_2)} = [K_L K_{1,G}]$. The last reduced estimation problem and SsS is OPT_3, in which $\theta^3 = \theta^{2\,(r_2)}$, with $Np_3 = 7$, $r_3 = 7$, $OF_3 = 157.51$, $\kappa_3 = 560$ and $\gamma_3 = 9.03$. For this estimation, the condition number and the collinearity index do not exceed the defined thresholds, with $\kappa_3 < 1000$ and $\gamma_3 < 10$ and all parameters in θ^3 were identified $r_3 = Np_3$. At this point, the parameter estimation has been stopped and the obtained parameters were statistically analyzed by calculating confidence intervals using the covariance matrix of the estimates.

Table 2

Application of Subset Selection Algorithm

θ_j	Estimated Parameter $\hat{\theta}_j$				Sensitivity measure δ_j			Identifiability Order		
	OPT_1	OPT_2	OPT_3	OPT_4	OPT_1	OPT_2	OPT_3	OPT_1	OPT_2	OPT_3
$k_{max,1}$	9.51	11.07	11.08	11.08	1.19	1.13	34.38	1	7	7
$k_{max,2}$	0.2479	0.0773	0.0498	0.033	2.56	1.20	0.68	7	2	2
$k_{max,3}$	0.3218	F	F	F	0.09	-	-	15	-	-
K_L	5915	8121	F	F	0.36	0.43	-	5	8	-
$K_{1,G}$	386.2	286.9	F	F	0.12	0.15	-	8	9	-
Krec	0.1687	F	F	F	0.05	-	-	14	-	-
Km	895.7	286.7	184.4	111.5	8.02	26.43	12.81	6	1	4
$K_{2,G}$	5.389	F	F	F	0.54	-	-	13	-	-
$K_{iy,EtOH}$	72.37	74.09	74.26	74.31	0.23	0.22	0.22	4	6	3
m_s	$5.5\text{x}10^{-6}$	F	F	F	0.00	-	-	11	-	-
Yxg	$4\text{x}10^{-4}$	$4\text{x}10^{-4}$	$4\text{x}10^{-4}$	$4\text{x}10^{-4}$	0.70	0.69	0.69	9	3	1
μ_{max}	0.0533	F	F	F	0.71	-	-	12	-	-
K_G	5722	F	F	F	0.70	-	-	10	-	-
$K_{1,EtOH}$	14.66	14.92	14.92	14.93	0.51	0.47	27.35	2	4	6
$K_{2,EtOH}$	23.14	25.72	25.75	25.74	0.34	0.33	0.33	3	5	5

[F] Parameter fixed to previous optimum which was not estimated in current reduced parameter estimation problem

From the comparison with the results from the original problem (OPT_1) with $Np = 15$ parameters with the first and second reduced problems (OPT_2 and OPT_3) with $Np = 9$ and $Np = 7$, respectively, enormous improvements in the parameter accuracy, validated by reductions in confidence intervals, were observed. For further improvements of the estimation in OPT_3 (further reduction in *StDev* for each parameter), a last optimization run

(OPT_4) was performed, which used $k_{max,2} = 4x10^{-6}$ and $Km = 124$ as initial estimates along with the other values of IE_4. The parameters obtained from OPT_4 are considered to be the most accurate parameter estimates in this research, with an improvement in the maximum relative standard deviation of the most uncertain problem parameter $k_{max,2}$ and Km from 400% to 14%.

6. Conclucions

A systematic approach to parameter identifiable subset selection based on the sensitivity matrix, Singular Value Decomposition (SVD) as a rank-revealing factorization and QR decomposition with column pivoting (QRP) has been successfully applied to a biological system within a least square parameter estimation problems.

Besides the proper normalization of the sensitivity matrix, it is of importance to generate an appropriate initial parameter guess for estimation, which is sufficiently close to the optimal parameter set in order to provide a subset selection that does not differ significantly from the one based on the sensitivity matrix evaluated at the optimal estimate. If this is not possible, an iterative proceeding as discussed in this paper should be followed.

S y m b o l s

A_D – Type Arrhenius constant [h^{-1}]
C_c – Cellulose Concentration [gL^{-1}]
Ccb – Cellobiose Concentration [gL^{-1}]
C_E – Enzyme Concentration [$FPUL^{-1}$]
C_{EtOH} – Ethanol Concentration [gL^{-1}]
C_G – Glucose Concentration [gL^{-1}]
C_x – Yeast Concentration [gL^{-1}]
e_T – Total protein (cellulase and β-glucosidase) concentration per liter reaction volume [gL^{-1}]
e_g – β-glucosidase activity per g of protein in the enzyme preparation [IUg^{-1}]
E_a – Activation Energy for enzymatic activity [$Jmol^{-1}$]
K_D – Specific rate of cellulose [h^{-1}]
K_G – Glucose saturation constant for yeast [gL^{-1}]
$K_{1,EtOH}$ – Inhibition constant of cellulase by ethanol [gL^{-1}]
$K_{2,EtOH}$ – Inhibition constant of ß-glucosidase by ethanol [gL^{-1}]
$K_{iy,EtOH}$ – Inhibition constant of ethanol on yeast [gL^{-1}]
K_L – Langmuir adsorption constant (cellulase adsorption saturation constant [$FPUL^{-1}$]
$K_{1,G}$ – Inhibition constants of cellulase by glucose [gL^{-1}]
$K_{2,G}$ – Inhibition constants of β-glucosidase by glucose [gL^{-1}]
$k_{max,1}$ – Maximum specific rate of cellulose hydrolysis to cellobiose [h^{-1}]
$k_{max,2}$ – Specific rate of cellobiose hydrolysis to glucose [$gU^{-1}h^{-1}$]
$k_{max,3}$ – Maximum specific rate of cellulose hydrolysis to glucose [h^{-1}]

Km	–	Michaelis constant for β-glucosidase for cellobiose [gL^{-1}]
$Krec$	–	Recalcitrance constant [-]
m_s	–	Maintenance requirement for yeast [h^{-1}]
OF	–	Objective Function value of parameter estimation
v_1	–	Production rate of cellobiose from cellulose by cellulase [$gL^{-1}h^{-1}$]
v_2	–	Production rate of glucose from cellobiose by β-glucosidase [$gL^{-1}h^{-1}$]
v_3	–	Production rate of glucose from cellulose by cellulase [$gL^{-1}h^{-1}$]
v_4	–	Production rate of biomass [$gL^{-1}h^{-1}$]
v_5	–	Consumption rate of glucose by yeast [$gL^{-1}h^{-1}$]
Y_{xg}	–	Anaerobic yield of cell mass on glucose (yield coefficient of cell mass from glucose) [gg^{-1}]
ΔH	–	Deactivation Enthalpy [$Jmol^{-1}$]
μ_{max}	–	Maximum growth rate (maximum specific growth rate of the microorganism [h^{-1}]
Y^m	–	Experimental data vector
Y	–	Predicted response variables by model
Nm	–	Number of sample times
Np	–	Number of parameters
Ny	–	Number of measurement variables
S	–	Sensitivity matrix
H_θ	–	Hessian matrix
U	–	Real or complex unitary matrix of SVD
V	–	Real or complex unitary matrix of SVD
Σ	–	Rectangular diagonal matrix with nonnegative real numbers on the diagonal of SVD
r	–	Rank of the sensitivity matrix
Q	–	Orthogonal matrix of QR decomposition with column pivoting (QRP)
R	–	Upper triangular matrix with decreasing diagonal elements of QRP
P	–	Permutation matrix of QRP
σ_j	–	Singular value j
δ_j	–	Sensitivity measure of column j of sensitivity matrix
κ	–	Condition number of parameter estimation problem
κ^{max}	–	Condition number threshold to guarantee nonlinear dependence between parameters
γ	–	Collinearity index
γ^{max}	–	Collinearity index threshold to guarantee nonlinear dependence between parameters
Γ	–	Set of singular values $_j$ with $\kappa_j \leq \kappa^{max}$
θ	–	Parameter vector
$\theta^{(Np-r)}$	–	Unidentifiable parameter vector after SsS algorithm
$\theta^{(r)}$	–	Identifiable parameter vector after SsS algorithm

The authors express their gratitude to Ph.D. Mariana Peñuela, who provided the experimental information for the development of this study. Diana López acknowledges the generous support of German Academic Interchange Service for funding her Ph.D. research.

References

[1] Lei F., Jorgensen S.B., *Estimation of kinetic parameters in a structured yeast model using regularisation*, Journal of Biotechnology, **88** (3), 2001, 223-237.

[2] Vélez-Reyes M., Verghese G.C., *Subset selection in identification, and application to speed and parameter estimation for induction machines*, Proceedings of the 4th IEEE Conference on Control Applications, 1995, 991-997.

[3] Burth M., Verghese G.C., Vélez-Reyes M., *Subset selection for improved parameter estimation in on-line identification of a synchronous generator*, Power Systems, IEEE Transactions, **14**, 1999, 218-225.

[4] Grah A., *Entwicklung und Anwendung modularer Software zur Simulation und Parameterschätzung in gaskatalytischen Festbettreaktoren*, Ph.D. thesis, Martin Luther University Halle-Wittenberg 2004.

[5] Vásquez M.P., *Desenvolvimento de processo de hidrolise enzimática e fermentação simultâneas para a produção de etanol a partir de bagaço de cana-de-açúcar*, Ph.D. thesis, Universidade Federal Do Rio De Janeiro, Brazil 2007.

[6] Drissen R.E.T., Maas R.H.W., Tramper J., Beeftink H.H., *Modelling ethanol production from cellulose: separate hydrolysis and fermentation versus simultaneous saccharification and fermentation*, Biocatalysis and Biotransformation, **27** (1), 2008, 27-35.

[7] Philippidis G.P., Spindler D.D., Wyman C.E., *Mathematical modeling of cellulose conversion to ethanol by the simultaneous saccharification and fermentation process*, Applied Biochemistry and Biotechnology, **34** (1), 1992, 543-556.

[8] Philippidis G.P., Hatzis C., *Biochemical Engineering Analysis of Critical Process Factors in the Biomass-to-Ethanol Technology*, Biotechnology Progress, **13** (3), 1997, 222-231.

[9] Brun R., Kühni M., Siegrist H., Gujer W., Reichert P., *Practical identifiability of ASM2D parameters – systematic selection and tuning of parameter subsets*, Water Research, **36**, 2002, 411-412.