Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 56-2 (2023) 9721–9726

On the Number of Reactions and Stoichiometry of Bioprocess Macroscopic Models: an Implicit Sparse Identification Approach

Guilherme A. Pimentel * **Laurent Dewasme** * **Alain Vande Wouwer** *

* *Systems, Estimation, Control and Optimization (SECO) , University of Mons, 7000 Mons, Belgium (e-mail: <guilherme.araujopimentel, laurent.dewasme, alain.vandewouwer>@umons.ac.be).*

Abstract: This paper presents a data-driven methodology to infer a macroscopic reaction scheme with stoichiometric parameters from a bioprocess database. The data sets consist of measurements of a few extracellular species, i.e., biomass, substrates, and products of interest. The proposed original procedure is based on implicit sparse identification. The methodology is illustrated with two case studies: (i) data generated by a two-step anaerobic digestion model and (ii) an experimental data set from the production of therapeutic proteins using mammalian cell cultures. Finally, the results of the latter application are compared with a standard data-driven algorithm, e.g., maximum-likelihood principal component analysis.

Keywords: macroscopic modeling, bioprocess, sparse identification, stoichiometry, number of reactions

1. INTRODUCTION

In the last decades, access to a large amount of data in different fields has guided the development of new approaches to infer dynamical models directly from measurement data. Several data-driven tools such as dynamic mode decomposition (DMD) and the Koopman operator (Schmid, 2022; Garcia-Tenorio et al., 2021), multilinear Gaussian processes (Wang et al., 2020), and nonnegative matrix decomposition (Pimentel et al., 2022) have been used to obtain models without predefined structures. Currently, attention has been drawn to parsimonious models that consider the lowest complexity required to describe the observed data, benefitting from being interpretable and preventing overfitting. Based on sparsity-promoting optimization, the sparse identification of nonlinear dynamics (SINDy) algorithm discovers dynamical models using a few structures from a predefined library of candidate functions (Brunton et al., 2016). This framework has been applied in many different research fields (see, for instance, Sorokina et al. (2016); Hoffmann et al. (2018); Loiseau et al. (2018)). Also, it has been extended to rational and implicit dynamics (Kaheman et al., 2020; Mangan et al., 2016), partial differential equations (Messenger and Bortz, 2021), and stochastic dynamics (Boninsegna et al., 2018), among many others.

In bioprocess modeling, Hoffmann et al. (2018) proposed a sparse identification tool combined with predefined reaction scheme libraries to numerically test all hypotheses and infer a consistent reaction scheme describing the process. Inferring bioreaction schemes from bioprocess data is also the goal of principal component analysis (PCA), considering a minimum of *a priori* knowledge of the process and providing the corresponding stoichiometry (Bernard and Bastin, 2005). An extension of this method, called maximum likelihood principal component analysis (MLPCA), has been exploited in Mailier et al. (2012) to account for higher levels of measurement noise.

In this study, we propose a new data-driven method to infer macroscopic reaction schemes, i.e., the number of macroscopic reactions and a relevant stoichiometric parameter basis, considering noisy measurements using the robust algorithm for parallel implicit sparse identification of nonlinear dynamics (SINDy-PI) proposed by Kaheman et al. (2020). In contrast with Hoffmann et al. (2018), the proposed methodology does not require a *a priori* library with the predefined reaction rate structures, nor any knowledge about the number of reactions. The procedure only requires basic knowledge of the process species interactions, as in most bioprocess modeling procedures (Hoffmann et al., 2018; Dewasme et al., 2017).

This paper is organized as follows. Section 2 details the macroscopic modeling approach and presents the data-driven techniques involved in the proposed method. Section 3 shows the results and analysis of two case studies - one based on simulation and one using an experimental dataset. The latter is compared with the application of MLPCA. The conclusion and future works are presented in Section 4.

2. MACROSCOPIC MODEL INFERENCE

2.1 Bioprocess Macroscopic Modeling

A macroscopic reaction scheme is a set of M reactions involving N key species, which are typically biomass, substrates, and products (Bastin and Dochain, 1990):

$$\sum_{i\in\mathcal{R}_j}(-k_{ij})\xi_i \xrightarrow{\varphi_j(\xi,\vartheta_j)} \sum_{i\in\mathcal{P}_j}(k_{ij})\xi_j \tag{1}$$

where $\mathcal{R}_j$ and $\mathcal{P}_j$ denote the sets of reactants and products, respectively, in the j^{th} reaction. k_{ij} are pseudo-stoichiometric coefficients while $\varphi_j(\xi,\vartheta_j)$ are the corresponding reaction rates, functions of ξ (reactant/product quantities or concentrations) and ϑ_j, the parameters of the rate kinetic structure.

10.1016/j.ifacol.2023.10.285

Applying mass balance to (1), the following ordinary differential equation system is obtained:

$$\frac{d\xi(t)}{dt} = K\varphi(\xi(t),\vartheta) + v(\xi(t),t), \tag{2}$$

where K is the pseudo-stoichiometric matrix, and $v(\xi(t),t)$ represents the transport term, including dilution effects, input feeds, and gaseous outflows. In most cases, the number of components N is larger than the number of reactions M so that the rank of the stoichiometric matrix K is assumed to be M. Since we have access to the process measurements $\xi(t)$ and inputs $v(\xi(t),\vartheta(t))$, we can implicitly identify the matrix K, which has linear relations with the state variable derivatives when the vector $\varphi(\xi(t),\vartheta)$ has an assumed unknown structure. To this end, we express (2) as

$$\frac{d\xi^{\star}(t)}{dt} = K\varphi(\xi(t),\vartheta), \tag{3}$$

where $\dot{\xi}^{\star}(t) = \dot{\xi}(t) - v(\xi(t),t)$, and $\dot{\xi}(t)$ denotes the time derivative of $\xi(t)$.

2.2 Parallel and Implicit Sparse Identification

Consider the following general dynamical nonlinear system

$$\frac{d\xi(t)}{dt} = f(\xi(t)), \tag{4}$$

where $\xi(t)$ is the state vector $\xi(t) = [\xi_1(t)\ \cdots\ \xi_N(t)]^T \in \mathbb{R}^N$, and the system dynamics $\dot{\xi}(t)$ is function $f(\xi(t))$. We assume that the system dynamics can have a sparse representation if the candidate library is

$$\Theta(\xi) = [\theta_1(\xi)\ \ \theta_2(\xi)\ \cdots\ \theta_w(\xi)], \tag{5}$$

where w is the library number of elements. Thus, each row equation may be written as

$$\frac{d\xi_k(t)}{dt} = f_k(\xi(t)) \approx \Theta(\xi)\Omega_k, \tag{6}$$

where Ω_k is a sparse vector, indicating which terms are active in the dynamics (Brunton et al., 2016).

To determine the nonzero entries of Ω_k through sparse regression based on trajectory data, the time-series data is arranged into a matrix $\Xi = [\xi(t_1),\xi(t_2)\ \cdots\ \xi(t_{n_s})]^T$, and the associated derivative matrix $\dot{\Xi} = \left[\dot{\xi}(t_1),\dot{\xi}(t_2)\ \cdots\ \dot{\xi}(t_{n_s})\right]^T$ is computed using appropriate numerical differentiation scheme.

It is now possible to describe the dynamical system using a model which is linear in the parameters and evaluated with the measured state trajectories:

$$\dot{\Xi} = \Theta(\Xi)\Omega. \tag{7}$$

Equation (7) might also involve derivatives of the state variables on the right-hand side, i.e., include a factor $\Theta(\Xi,\dot{\Xi})$. This situation will appear later on in this study (Section 2.3). To solve implicit model structures, Kaheman et al. (2020) proposed SINDy-PI, a constrained optimization formulation where each candidate function is tested individually in an implicit and parallel optimization. However, each of these individual equations may be combined into a single constrained system of equations

$$\Theta(\Xi,\dot{\Xi}) = \Theta(\Xi,\dot{\Xi})\Omega \quad \text{such that } \Omega_{yy} = 0, \tag{8}$$

where y is the number of columns. The constraint $\Omega_{yy} = 0$ forces the solution not to be the trivial one ($\Omega = I_{w\times w}$) and the optimization problem can be written as

$$\min_{\Omega} \|\Theta(\Xi,\dot{\Xi}) - \Theta(\Xi,\dot{\Xi})\Omega\|_2, \tag{9}$$
$$\text{s.t. } diag(\Omega) = 0, \text{and } |\Omega_{\{i,y\}}| < \lambda \text{ then } |\Omega_{\{i,y\}}| = 0,$$

where λ is a sparsity-promoting parameter. In this work, the sparsity is obtained using sequentially thresholded least squares, which iteratively computes a least-squares solution to minimize (9). Any element of Ω smaller than a threshold λ is set to zero and then (9) is solved again with these fixed zero elements. The sparsity parameter λ is a hyper-parameter, and each column equation may require a different parameter λ_y (Kaheman et al., 2020). To solve problem (9), we use CVX, a package for specifying and solving convex programs (Grant and Boyd, 2008, 2014).

Figure 1 presents the steps to obtain the model basis. First, the derivatives of the measurements are organized in the vector $\Theta(\Xi,\dot{\Xi})$, and a large value is given to λ. Then, the value of this parameter is decreased, and the fitting error

$$error = \frac{||\dot{\Xi} - \hat{\dot{\Xi}}||_2}{||\dot{\Xi}||_2}, \tag{10}$$

is analysed for each of the state derivatives, where $\hat{\dot{\Xi}}$ is the identified state derivative. A model candidate is obtained when the error is small, and the vector Ω is sparse. This procedure is repeated for each state variable.

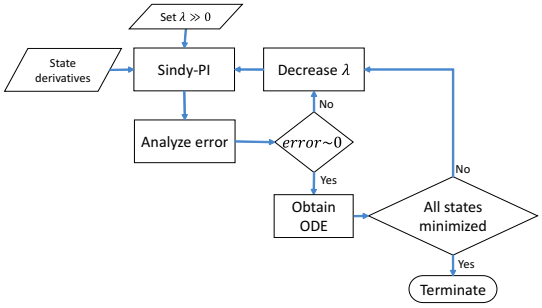


Fig. 1. Applied methodology flowchart.

2.3 Minimum Number of Reactions and Stoichiometry

The sparsity of Ω reveals the minimum number of macroscopic reactions represented by the maximum number of state variable derivatives required to reconstruct each trajectory. In addition, using the values identified in the columns of Ω, we implicitly extract the stoichiometric parameters of each macroscopic reaction.

To exemplify this procedure, a simple biomass growth on a glucose medium in fed-batch mode is considered (Bastin and Dochain, 1990). The modeling of this process is elementary, expressed by a single reaction that implies the consumption of glucose (G), the production of biomass (X), and lactate (L) as a byproduct.

$$G \xrightarrow{\varphi_1} k_{11}X + k_{31}L \tag{11}$$

Based on the reaction scheme, the corresponding mass balance equations can be written as follows:

$$\frac{dX}{dt} = k_{11}\varphi_1 - DX, \tag{12a}$$
$$\frac{dG}{dt} = -\varphi_1 - D(G - G_{in}), \tag{12b}$$
$$\frac{dL}{dt} = k_{31}\varphi_1 - DL, \tag{12c}$$

where the dilution rate is expressed by $D = F_{in}/V$ (F_{in} is the inlet flow rate and V the volume of the broth in the bioreactor) and the glucose concentration in the inflow by G_{in}. Biomass growth could be modeled by any suitable law, or combination of laws to represent, for instance, activation by glucose and inhibition by lactate, but the knowledge of this kinetic model is not required.

Applying (3), $\dot{X}^\star = (\dot{X} + DX)$, $\dot{G}^\star = (\dot{G} + D(G - G_{in}))$, $\dot{L}^\star = (\dot{L} + DL)$ and equation (12) can be rewritten as

$$\frac{dX^\star}{dt} = k_{11}\varphi_1, \tag{13a}$$
$$\frac{dG^\star}{dt} = -\varphi_1, \tag{13b}$$
$$\frac{dL^\star}{dt} = k_{31}\varphi_1. \tag{13c}$$

Equation (13) can be rewritten in function of its derivatives. For example, using $\varphi_1 = -\dot{G}^\star$,

$$\frac{dX^\star}{dt} = -k_{11}\frac{dG^\star}{dt}, \tag{14}$$
$$\frac{dL^\star}{dt} = -k_{31}\frac{dG^\star}{dt}, \tag{15}$$

or, using $\varphi_1 = \dot{L}^\star/k_{31}$,

$$\frac{dX^\star}{dt} = \frac{k_{11}}{k_{31}}\frac{dL^\star}{dt}, \tag{16}$$
$$\frac{dG^\star}{dt} = -\frac{1}{k_{31}}\frac{dL^\star}{dt}. \tag{17}$$

Other relations can be obtained from (13a) with $\varphi_1 = \dot{X}^\star/k_{11}$. All the linear relations in this example are functions of only one derivative, revealing that the measurement data can be expressed by only one reaction. The library matrix can be defined as $\Theta(\Xi, \dot{\Xi}) = [\dot{X}^\star \quad \dot{G}^\star \quad \dot{L}^\star]$ and the optimization problem (9) can be solved, resulting in the matrix Ω that is used to implicitly identify K using the measurement derivatives.

3. APPLICATIONS

This section presents two realistic applications to show the potential of the procedure to infer the number of macroscopic reactions and their stoichiometric parameters from numerical data. The first application considers simulated data, while the second is based on experimental data.

3.1 Case Study 1: Anaerobic Digestion (two reactions)

In this process, a microorganism consortium degrades the organic matter in the liquid phase to produce biogas, a mixture of methane and carbon dioxide. A two-step model (Antonelli et al., 2003), i.e., acidogenesis and methanogenesis, is considered in the following:

(a) Chemical oxygen demand (COD) consumption

$$k_{31}S_1 \xrightarrow{\varphi_1} X_1 + k_{41}S_2 \tag{18}$$

(b) Volatile fatty acid (VFA) consumption

$$k_{42}S_2 \xrightarrow{\varphi_2} X_2. \tag{19}$$

From the reaction scheme, the corresponding mass balance equations can be written as follows:

$$\frac{dX_1}{dt} = \varphi_1, \tag{20a}$$
$$\frac{dX_2}{dt} = \varphi_2, \tag{20b}$$
$$\frac{dS_1}{dt} = -k_{31}\varphi_1 \tag{20c}$$
$$\frac{dS_2}{dt} = k_{41}\varphi_1 - k_{42}\varphi_2 \tag{20d}$$

where X_1, X_2, S_1, and S_2 are the concentrations of acidogenic bacteria, methanogenic bacteria, chemical oxygen demand (COD) and volatile fatty acids (VFA), respectively. In addition, k_{31}, k_{41}, and k_{42} are the yield coefficients for COD degradation, VFA production, and consumption, respectively. Kinetics are modeled with two microbial growth rates

$$\varphi_1 = \mu_{max,1}\frac{S_1}{K_{S_1} + S_1}X_1, \quad \varphi_2 = \mu_{max,2}\frac{S_2}{K_{S_2} + S_2}X_2. \tag{21}$$

where $\mu_{max,i}$ are the maximum growth rates, and K_{S_i} are the half-saturation parameters, with $i = 1, 2$.

Using model (20) with parameter values presented in Table 1, a numerical dataset is generated and the measurements derivatives are gathered in the library vector as $\Theta(\Xi, \dot{\Xi}) = [\dot{X}_1 \quad \dot{X}_2 \quad \dot{S}_1 \quad \dot{S}_2]$.

Table 1. Anaerobic digestion parameter values: nominal model and estimated values.

Parameter	Simulation	Identification	Parameter	Simulation
k_{31}	0.31204	0.3120	$\mu_{max,1}$	0.42912
k_{41}	0.06277	0.0628	K_{S1}	2.6493
k_{42}	3.1473	3.1473		

Starting with a large value of λ, the procedure consists in decreasing this parameter until the error of one of the estimates is minimized. With $\lambda = 3$, $\mathbf{error}_{\dot{\mathbf{X}}_1} = \mathbf{1.0978 \times 10^{-16}}$, $error_{\dot{X}_2} = 1$, $error_{\dot{S}_1} = 1$ and $error_{\dot{S}_2} = 1$ are obtained. Figure 2a shows the error evolution of the estimates. From the dataset, the parsimonious implicit representation of $\dot{X}_1$ reads

$$\hat{\dot{\Xi}} = \underbrace{[\dot{X}_1 \quad \dot{X}_2 \quad \dot{S}_1 \quad \dot{S}_2]}_{\Theta(\Xi,\dot{\Xi})} \underbrace{\begin{bmatrix} \mathbf{0} & 0\ 0\ 0 \\ \mathbf{0} & 0\ 0\ 0 \\ \mathbf{-3.2047} & 0\ 0\ 0 \\ \mathbf{0} & 0\ 0\ 0 \end{bmatrix}}_{\Omega} \tag{22}$$

resulting in

$$\hat{\dot{X}}_1 = -3.2047\,\dot{S}_1. \tag{23}$$

The next step consists in reducing λ to 0.3 and $error_{\dot{X}_1} = 1.0978 \times 10^{-16}$, $error_{\dot{X}_2} = 0.37159$, $\mathbf{error}_{\dot{\mathbf{S}}_1} = \mathbf{9.1111 \times 10^{-17}}$ and $error_{\dot{S}_2} = 0.11887$ are obtained (see Figure 2b). From the sparse matrix Ω, $\dot{S}_1$ is obtained as

$$\hat{\dot{\Xi}} = \underbrace{[\dot{X}_1 \quad \dot{X}_2 \quad \dot{S}_1 \quad \dot{S}_2]}_{\Theta(\Xi,\dot{\Xi})} \underbrace{\begin{bmatrix} 0 & 0 & \mathbf{-0.31204} & 0 \\ 0 & 0 & \mathbf{0} & -3.0301 \\ -3.2047 & 0 & \mathbf{0} & 0 \\ 0 & -0.20484 & \mathbf{0} & 0 \end{bmatrix}}_{\Omega} \tag{24}$$

resulting in

$$\hat{\dot{S}}_1 = -0.3120\,\dot{X}_1. \tag{25}$$

Anew, λ is reduced to 0.055, delivering $error_{\dot{X}_1} = 9.0576 \times 10^{-11}$, $error_{\dot{X}_2} = 0.11645$, $error_{\dot{S}_1} = 9.1111 \times 10^{-17}$ and $\mathbf{error_{\dot{S}_2} = 2.923 \times 10^{-9}}$ (see Figure 2c), and $\dot{S}_2$ reads

$$\hat{\Xi} = [\dot{X}_1 \quad \dot{X}_2 \quad \dot{S}_1 \quad \dot{S}_2] \begin{bmatrix} 0 & 0 & -0.3120 & \mathbf{0.0628} \\ 0.4349 & 0 & 0 & \mathbf{-3.1473} \\ -3.1769 & 0 & 0 & \mathbf{0} \\ 0.1382 & -0.3177 & 0 & \mathbf{0} \end{bmatrix} \tag{26}$$

resulting in

$$\hat{\dot{S}}_2 = 0.0628\,\dot{X}_1 - 3.1473\dot{X}_2. \tag{27}$$

To find the missing state variable, λ is set to 0.018, providing $error_{\dot{X}_1} = 1.0878 \times 10^{-16}$, $\mathbf{error_{\dot{X}_2} = 9.553 \times 10^{-17}}$, $error_{\dot{S}_1} = 9.1111 \times 10^{-17}$ and $error_{\dot{S}_2} = 1.3139 \times 10^{-16}$ (see Figure 2d). The last sparse relation is

$$\hat{\Xi} = [\dot{X}_1 \quad \dot{X}_2 \quad \dot{S}_1 \quad \dot{S}_2] \begin{bmatrix} 0 & \mathbf{0.0199} & -0.3120 & 0.0628 \\ 0.4669 & \mathbf{0} & 0 & -3.1473 \\ -3.1749 & \mathbf{0} & 0 & 0 \\ 0.1483 & \mathbf{-0.3177} & 0 & 0 \end{bmatrix} \tag{28}$$

resulting in

$$\hat{\dot{X}}_2 = 0.0199\,\dot{X}_1 - 0.3177\,\dot{S}_2. \tag{29}$$

and the system is summed up as

$$\hat{\dot{X}}_1 = -3.2047\,\dot{S}_1, \qquad \lambda = 3, \tag{30a}$$

$$\hat{\dot{S}}_1 = -0.3120\,\dot{X}_1, \qquad \lambda = 0.15, \tag{30b}$$

$$\hat{\dot{X}}_2 = 0.0199\,\dot{X}_1 - 0.3177\dot{S}_2, \qquad \lambda = 0.018, \tag{30c}$$

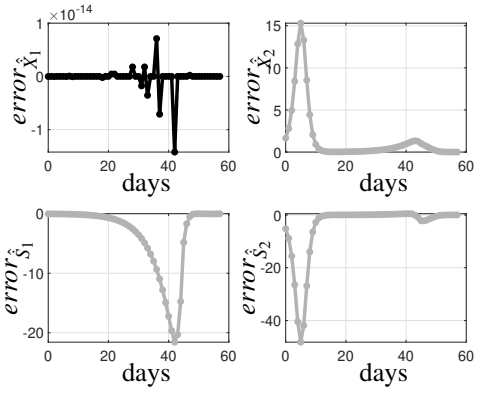
$$\hat{\dot{S}}_2 = 0.0628\,\dot{X}_1 - 3.1473\dot{X}_2, \qquad \lambda = 0.055. \tag{30d}$$

The minimal number of macroscopic reactions required to parsimoniously express the measurements data is two, as the maximum number of species concentration derivatives required to express each of the equations in (30) is two, i.e., $\hat{\dot{X}}_2(\dot{X}_1, \dot{S}_2)$ and $\hat{\dot{S}}_2(\dot{X}_1, \dot{X}_2)$, constituting φ_j with $j = 1, 2$.

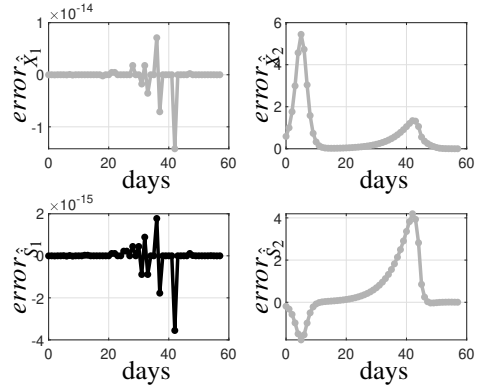
Combining the values of Ω with the process a priori knowledge, the values of the stoichiometric parameters of the macroscopic reactions can be inferred. In this example, we consider that reaction rate 1, φ_1, is responsible for the production of acidogenic bacteria X_1 and reaction rate 2, φ_2, is responsible for the production of methanogenic bacteria X_2, which is one plausible assumption. Then, considering only the signal in (30), we can deduce which metabolite is consumed or produced. Furthermore, from $\varphi_1 = \dot{X}_1$ and $\varphi_2 = \dot{X}_2$, and (30a) $1/k_{31}$ is found, as well as k_{31} from (30b) (see (20c)). Moreover, considering the aforementioned definition of φ_1 and φ_2, (30c) is rewritten as $\hat{\dot{S}}_2 = 0.0628\varphi_1 - 3.1473\varphi_2$ that is (20d). Finally, combining (30d) with $\varphi_1 = \dot{X}_1$ and $\hat{\dot{S}}_2$ results in $\hat{\dot{X}}_2 = 0.9999\varphi_2$, which is analogous to (20b).

3.2 Case Study 2: Hybridoma Cell Culture

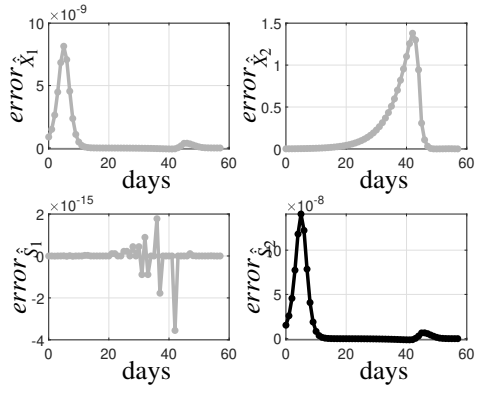
The second application considers a batch culture experiment of a hybridoma strain performed in 200 mL T-flasks. The substrate concentrations (glucose and glutamine) are set to prescribed values respectively, ranging between 6 and 7 g/L and 0.3 and 0.4 g/L. The culture time is approximately nine days. The measurements of viable biomass X_v, dead biomass X_d, glucose G,



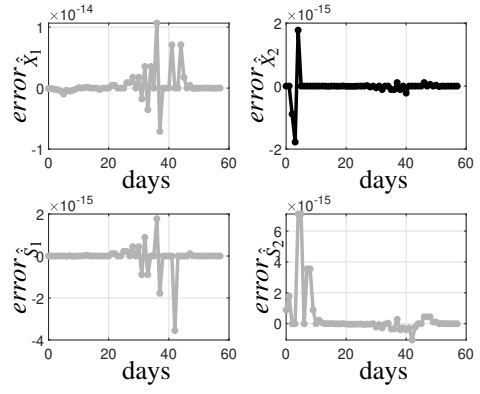
(a) $\lambda = 3$: $d\hat{X}_1/dt$ found, $error_{\hat{\dot{X}}_1} \approx 0$.



(b) $\lambda = 0.15$: $d\hat{S}_1/dt$ found, $error_{\hat{\dot{S}}_1} \approx 0$.



(c) $\lambda = 0.055$: $d\hat{S}_2/dt$ found, $error_{\hat{\dot{S}}_2} \approx 0$.



(d) $\lambda = 0.018$: $d\hat{X}_2/dt$ found, $error_{\hat{\dot{X}}_2} \approx 0$.

Fig. 2. Plots of the fitting errors for different values of λ.

glutamine G_n, lactate L, and monoclonal antibodies MAb are taken once every day. These data are first preprocessed using a smoothing spline (*spaps* from Matlab), and the corresponding derivatives are calculated to build the library $\Theta(\Xi,\dot{\Xi}) = [\dot{X}_v \quad \dot{X}_d \quad \dot{G} \quad \dot{G}_n \quad \dot{L} \quad \dot{MAb}]$.

Remark 1. The level of noise on the measurements and, in turn, the level of noise on their derivatives might, of course, significantly influence the results. Several numerical tools can be used to alleviate this potential issue, such as smoothing splines (Reinsch, 1967) or algorithms for computing derivatives with a total variation regularization approach (Chartrand, 2011).

The reduction procedure of λ until minimizing the errors of each of the estimates is carried out, and the following relations are obtained:

$$\hat{MAb} = 38.444\dot{X}_v + 42.4723\dot{X}_d \qquad \lambda = 20 \tag{31a}$$
$$\hat{\dot{G}} = 8.8332\dot{L} \qquad \lambda = 3 \tag{31b}$$
$$\hat{\dot{L}} = -9.0099\dot{G}_n \qquad \lambda = 2 \tag{31c}$$
$$\hat{\dot{X}}_d = -0.8721\dot{X}_v - 1.2954\dot{G}_n \qquad \lambda = 0.3 \tag{31d}$$
$$\hat{\dot{X}}_v = -1.0611\dot{X}_d - 1.4102\dot{G}_n \qquad \lambda = 0.15 \tag{31e}$$
$$\hat{\dot{G}}_n = 0.2589\dot{X}_d - 0.1282\dot{L} \qquad \lambda = 0.06 \tag{31f}$$

The inferred number of reactions is two, as the maximum number of species concentration derivatives involved in equations (31) is two, i.e., $\hat{\dot{G}}_n(\dot{X}_d,\dot{L})$. The following assumption is formulated to obtain the parameter values: X_v, MAb, and L are produced by the consumption of glucose and glutamine with a reaction rate φ_1. The second reaction involves biomass decay, degrading X_v into dead biomass X_d and releasing $MAbs$ with a reaction rate φ_2. Thus, the reaction scheme is expressed as follows:

(a) Substrate oxidation:

$$k_{31}G + k_{41}G_n \xrightarrow{\varphi_1} X_v + k_{51}L + k_{61}MAb, \tag{32}$$

(b) Biomass death

$$X_v \xrightarrow{\varphi_2} X_d + k_{62}MAb. \tag{33}$$

From the reaction scheme, the corresponding mass balance equations can be written as follows:

$$\frac{dX_v}{dt} = \varphi_1 - \varphi_2, \tag{34a}$$
$$\frac{dX_d}{dt} = \varphi_2, \tag{34b}$$
$$\frac{dG}{dt} = -k_{31}\varphi_1, \tag{34c}$$
$$\frac{dGn}{dt} = -k_{41}\varphi_1, \tag{34d}$$
$$\frac{dL}{dt} = k_{51}\varphi_1, \tag{34e}$$
$$\frac{dMAb}{dt} = k_{61}\varphi_1 + k_{62}\varphi_2. \tag{34f}$$

The stoichiometric parameters can be estimated using the values of Ω included in (31). We obtain k_{61} and k_{62}, see (34f), using relation (31a) with $\dot{X}_v = \varphi_1 - \varphi_2$ and $\dot{X}_d = \varphi_2$. The parameter k_{41}, see (34d), is obtained using (31e) with the same assumption for $\dot{X}_v$ and $\dot{X}_d$. The parameter k_{51}, see (34e), uses the relation (31c) with $\dot{G}_n = -k_{41}\varphi_1$. The last parameter k_{31}, see (34c), uses (31b) with $\dot{L} = k_{51}\varphi_1$, as k_{51} has already been identified. Table 2 reports all the parameter estimates.

Comparison with MLPCA: Maximum-likelihood principal component analysis (MLPCA) allows determining reaction schemes of increasing dimension p, explaining a noisy data set, minimizing a log-likelihood cost of the form:

$$J_p = \sum_{i=1}^{n_s} (\xi_{m_i} - \xi_{\theta,p,i})^T Q_i^{-1} (\xi_{m_i} - \xi_{\theta,p,i}), \tag{35}$$

where n_s is the number of measured observations, ξ_{m_i} is the experimental measurement vector, with an error covariance matrix Q_i and $\xi_{\theta,p,i}$ is its maximum-likelihood (ML) estimate by the reduced p-dimensional linear model (Mailier et al., 2012). Note that J_p decreases as p increases and should be smaller or equal to the log-likelihood cost J^* of the true nonlinear model, assumed to follow a chi-square distribution with $n_S \times N$ degrees of freedom (Mahalanobis, 1936). The number of reactions is selected as the smallest p such that the log-likelihood cost J_p is smaller or equal to the range of a $\chi^2_{n_s \times N}$-distributed random variable. Once the number of reactions is determined, the resulting N by p affine subspace basis $\hat{\rho}$ can be used to estimate a stoichiometric matrix $\hat{K}$, which is a linear combination of the basis vectors, i.e.,

$$\hat{K} = \hat{\rho} G, \tag{36}$$

with G as $p \times p$ regular matrix. For a complete estimation of the stoichiometry, p biological constraints have to be imposed in each column of $\hat{K}$.

Considering the experimental data, Figure 3 presents the log-likelihood of the cost function J_p for each p dimension. The two dashed gray lines represent the $\chi^2_{n_s \times N}$ quantiles at 99.9% and 0.1%. From the bar graphs in Figure 3, the minimum number of macroscopic reactions to explain the experimental data should be chosen as two since the cost function value for $p = 2$ is just below the 0.1% quantile (meaning that at least 99.9% of the information content of the data can be represented). Interestingly, the implicit sparse identification approach achieves the same minimal number of reactions.

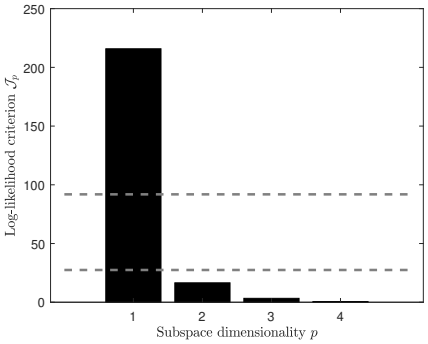


Fig. 3. MLPCA cost function experimental dataset.

Considering the subspace basis of dimension two ($p = 2$), (36) and assuming that biomass is generated by φ_1 and degraded by φ_2, the following stoichiometric matrix is obtained:

$$\hat{K} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \\ -5.1342 & \mathbf{-1.1856} \\ -0.70865 & \mathbf{-0.02952} \\ 5.1759 & \mathbf{2.7638} \\ 38.087 & 4.2769 \end{bmatrix} \tag{37}$$

where the rows represent the stoichiometric values for X_v, X_d, G, G_n, L, and MAb, respectively, and the column index indi-

cates the corresponding reaction. Table 2 presents the results of both approaches. Even if the deviation is 18% for k_{31}, the results are comparable. Moreover, the matrix $\hat{K}$ computation considers only two constraints per column, limiting the number of possible zero terms. Further constraints/assumptions could be applied to the MLPCA results to deepen the analysis. For instance, the values in bold font in (37) could be set to zero as they are smaller than the ones in the first column and do not make part of the predefined reaction scheme (33). This additional assumption is not required in the proposed approach as the sparsity of the results is the core of the method, improving their interpretation.

Table 2. Stoichiometric parameter from the experimental dataset.

Parameter	Implicit Sparse Ident	MLPCA
k_{31}	-6.26	-5.13
k_{41}	-0.729	-0.708
k_{51}	6.28	5.17
k_{61}	38.4	38.0
k_{62}	4.02	4.27

4. CONCLUSION

This paper presents an approach based on sparse identification to infer macroscopic reaction schemes and the corresponding stoichiometry from a process database. Two case studies are used to describe and validate the proposed method. An experimental data set compares the approach with the well-known maximum-likelihood principal component analysis, showing coherent results. Future work points toward selecting libraries to simplify the indirect identification of parameters for processes with larger chains of macroscopic reactions.

ACKNOWLEDGMENT

The authors acknowledge the support of the ProtoDrive project (convention *no. 2010119*) of the Win2Wal program of the Walloon Region (DGO6) and MabDrive project (convention *no. 1410085*), both achieved in collaboration with the CER Groupe. The scientific responsibility rests with its authors.

REFERENCES

Antonelli, R., Harmand, J., Steyer, J.P., and Astolfi, A. (2003). Set-point regulation of an anaerobic digestion process with bounded output feedback. *IEEE Transactions on Control Systems Technology*, 11(4), 495–504.

Bastin, G. and Dochain, D. (1990). *On-Line Estimation and Adaptive Control of Bioreactors*. Volume 1 of Process Measurement and Control, Elsevier: Amsterdam.

Bernard, O. and Bastin, G. (2005). On the estimation of the pseudo-stoichiometric matrix for macroscopic mass balance modelling of biotechnological processes. *Mathematical Biosciences*, 193(1), 51–77.

Boninsegna, L., Nüske, F., and Clementi, C. (2018). Sparse learning of stochastic dynamical equations. *The Journal of Chemical Physics*, 148, 241723.

Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15), 3932–3937.

Chartrand, R. (2011). Numerical differentiation of noisy, nonsmooth data. *International Scholarly Research Network - ISRN Applied Mathematics*, 2011, 1–11.

Dewasme, L., Côte, F., Filee, P., Hantson, A.L., and Vande Wouwer, A. (2017). Macroscopic dynamic modeling of sequential batch cultures of hybridoma cells: an experimental validation. *Bioengineering*, 4, 17, 1 – 20.

Garcia-Tenorio, C., Mojica-Nava, E., Sbarciog, M., and Vande Wouwer, A. (2021). Analysis of the ROA of an anaerobic digestion process via data-driven Koopman operator. *Nonlinear Engineering*, 10(1), 109–131.

Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura (eds.), *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, 95–110. Springer-Verlag Limited.

Grant, M. and Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1.

Hoffmann, M., Fröhner, C., and Noé, F. (2018). Reactive SINDy: Discovering governing reactions from concentration data. *The Journal of Chemical Physics*, 116.

Kaheman, K., Kutz, J.N., and Brunton, S.L. (2020). SINDy-PI: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics. *Proceedings of the Royal Society A*, 476 (2242).

Loiseau, J.C., Noack, B.R., and Brunton, S.L. (2018). Sparse reduced-order modelling: sensor-based dynamics to full-state estimation. *Journal of Fluid Mechanics*, 844, 459–490.

Mahalanobis, P. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, 2, 49–55.

Mailier, J., Remy, M., and Wouwer, A.V. (2012). Stoichiometric identification with maximum likelihood principal component analysis. *Journal of Mathematical Biology*, 67(4), 739–765.

Mangan, N.M., Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2(1), 52–63.

Messenger, D.A. and Bortz, D.M. (2021). Weak SINDy for partial differential equations. *Journal of Computational Physics*, 443, 110525.

Pimentel, G.A., Dewasme, L., and Vande Wouwer, A. (2022). Data-driven linear predictor based on maximum likelihood nonnegative matrix decomposition for batch cultures of hybridoma cells. *IFAC-PapersOnLine*, 55(7), 903–908. 13th IFAC Symposium on Dynamics and Control of Process Systems, including Biosystems DYCOPS 2022.

Reinsch, C. (1967). Smoothing by spline functions. *Numer. Math.*, 10, 177—-183.

Schmid, P.J. (2022). Dynamic mode decomposition and its variants. *Annual Review of Fluid Mechanics*, 54(1), 225–254.

Sorokina, M., Sygletos, S., and Turitsyn, S. (2016). Sparse identification for nonlinear optical communication systems: SINO method. *Optica Express*, 24(26), 30433–30443.

Wang, M., Risuleo, R.S., Jacobsen, E.W., Chotteau, V., and Hjalmarsson, H. (2020). Identification of nonlinear kinetics of macroscopic bio-reactions using multilinear Gaussian processes. *Computers & Chemical Engineering*, 133, 106671.