Journal of Process Control 134 (2024) 103148

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Data-driven inference of bioprocess models: A low-rank matrix approximation approach

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

Systems, Estimation, Control and Optimization Group (SECO), University of Mons, Mons, 7000, Belgium

ARTICLE INFO

Keywords:
Mathematical modeling
Estimation
Low-rank matrix approximation
Successive projection algorithm
Biotechnology
Hybridoma cell cultures

ABSTRACT

Following the recent advent of Process Analytical Technologies, dataset production has undergone significant leverage. In this new abundance of data, isolating meaningful, informative content is critical for process dynamic modeling. This paper proposes a data-driven algorithm based on low-rank matrix approximation, the so-called successive projection algorithm, to retrieve a minimal set of macroscopic reactions, the corresponding stoichiometry, and a consistent kinetic model structure from the measurements of the trajectories of the species concentrations during cultures in a bioreactor. The proposed method is successfully validated in simulation, considering a case study related to monoclonal antibody (MAb) production with hybridoma cell cultures.

1. Introduction

In the past few decades, Process Analytical Technologies (PAT) have contributed to improving bioprocess production yields and quality attributes [1], partly due to the progress in techniques involving process modeling, monitoring, and control. In addition, new achievements in online, in-line, at-line, and off-line measurement protocols are responsible for gathering large amounts of process data that allow advanced process modeling and analysis, targeting the development of digital twins.

One of the fundamental modeling tools for designing bioprocess monitoring and control is the concept of macroscopic reaction schemes, introduced in the late eighties by Bastin and Dochain [2]. It has been successfully applied for decades in several fields, such as wastewater treatment [3–5], anaerobic digestion [6–8], mammalian cell cultures for biopharmaceuticals [9–12] or microalgae cultures [13–16]. A macroscopic representation of a bioprocess consists of lumping complex reactions and metabolite interconnections in parsimonious relations to derive a set of mass balance equations capturing the primary dynamics. The design of a macroscopic model requires first the inference of a global structure and, in the second step, parameter estimation.

This paper focuses on the first objective, which can be decomposed into three tasks: (i) extracting the number of macroscopic reactions, (ii) inferring the corresponding stoichiometry, and (iii) designing an appropriate structural kinetic model. Many different approaches could be used to solve these problems, from reducing complex metabolic networks [17,18] to machine learning of black-box structures such as neural networks [19,20]. Halfway, hybrid models combine data-driven techniques with mechanistic process knowledge. For instance, Rogers et al. [21] integrate physical knowledge into machine-learning solutions to model complex biochemical processes. Since the kinetic structure is considered the most uncertain part of a biochemical model, hybrid modeling suggests combining mass balance differential equation systems with neural-network-based kinetics [22–24]. For an extensive review on machine learning-based optimization and control for bioprocesses, see [25]. However, these hybrid methods still consider significant metabolic a priori knowledge.

Recently, new approaches have been proposed to identify the underlying structure of a dataset and extract meaningful information to guide the design of dynamical models with minimum complexity. These models benefit from being interpretable and tend to generalize and prevent overfitting [26]. Examples of these approaches are the extended dynamic mode decomposition (EDMD) [27,28], multilinear Gaussian processes [12,29], sparse matrix decomposition [30], and sparse identification of nonlinear dynamics [26,31,32].

Other simple and powerful methods to retrieve valuable information from data can also be considered, such as linear dimensionality reduction (LDR) techniques, which are equivalent to low-rank matrix approximations (LRMA). The latter represents each data point of a dataset as a linear combination of a small number of subspace basis elements. Linear dimensionality reduction methods are a cornerstone of high dimensional data analysis due to their simple geometric interpretation and attractive computational properties. These methods capture features of interest from the data, such as covariance,

* Corresponding author.
E-mail addresses: guilherme.araujopimentel@umons.ac.be (G.A. Pimentel), laurent.dewasme@umons.ac.be (L. Dewasme), alain.vandewouwer@umons.ac.be (A. Vande Wouwer).

https://doi.org/10.1016/j.jprocont.2023.103148
Received 21 May 2023; Received in revised form 27 November 2023; Accepted 11 December 2023

dynamical structure, the correlation between data sets, input–output relationships, as well as margins between data classes [33]. Among the LDR techniques, principal component analysis (PCA) and independent component analysis (ICA) are popular methods for simple linear transformations. Considering minimal a priori process knowledge, PCA and maximum likelihood PCA (MLPCA) have been used successfully to infer the number of reactions and extract information about stoichiometry [34–36]. Besides, LRMA is a workhorse in numerical linear algebra, with singular value decomposition (SVD) being a central technique. LRMA is closely related to eigenvalue decomposition and factorization, such as Cholesky, QR, and LU, to cite a few [37].

In a previous work of the authors [30], a class of LRMA called nonnegative matrix decomposition is used to design a data-driven linear predictor for batch cultures of hybridoma cells. In this study, the nonnegativity constraint imposed on the input and mixing matrices is relaxed, enhancing stoichiometry estimation and revealing hidden kinetic model structures from the dataset. The main contribution of this paper is, therefore, to propose data-driven tools based on low-rank matrix approximation, and especially a successive projection algorithm [38,39], to obtain the process stoichiometry and guidelines for selecting the kinetic model structure. As it will become obvious in the following, the differentiation of the time evolution of the species concentrations is required in the course of the procedure and a discussion of some of the readily available numerical approaches is also provided. The whole procedure is tested using a case study related to the production of monoclonal antibodies.

The paper is structured as follows: First, Section 2 introduces the problem statement. Next, Section 3 presents low-rank matrix approximation methods, and Section 4 proposes a procedure for extracting the number of reactions as well as information on the stoichiometry and kinetics illustrated by a simple numerical example. In addition, the numerical differentiation of noisy signals is discussed, together with the continuation of the simple numerical example. Finally, Section 5 applies the method to a simulation case study related to a hybridoma cell culture process, and the main conclusions and perspectives are discussed in Section 6.

2. Problem statement

The metabolism of a microorganism can be macroscopically described by a scheme of M reactions involving N key species, which are typically biomass, substrates, and products [2]:

$$\sum_{i\in\mathscr{R}_j} k_{i,j}\xi_i \xrightarrow{\varphi_j(\xi,\vartheta_j)} \sum_{l\in\mathscr{P}_j} k_{l,j}\xi_l \tag{1}$$

where $\mathscr{R}_j$ and $\mathscr{P}_j$ denote the sets of reactants and products, respectively, in the jth reaction. $k_{i,j}$ and $k_{l,j}$ are the pseudo-stoichiometric coefficients related to the ith and lth species while $\varphi_j(\xi,\vartheta_j)$ is the reaction rate, function of the vector of species ξ (gathering reactants and products) considered as state variables and ϑ_j, the kinetic parameter vector.

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

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

where K is the pseudo-stoichiometric matrix, and $v(\xi(t),t)$ represents the transportation term, including dilution effects, input feeds, and gaseous outflows. The number of components N is usually larger than the number of reactions M, which are assumed independent, and the rank of the stoichiometric matrix K is, therefore, M. Assuming that the component concentrations of interest (the state variables) and the bioreactor inputs (inflows and outflows in liquid and gaseous forms) can be measured, the original mass-balance system (2) can be reformulated as a transport-free equation of the form [36]:

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

where $\dot\xi^\star(t) = \dot\xi(t) - v(\xi(t),t)$, and $\dot\xi(t)$ denotes the time derivative of $\xi(t)$. As stated in [36],

Property 1. the transport-free vector evolves within a linear affine subspace defined by the column-subspace (or range) of K. This affine subspace is independent of the reaction rates and its dimensionality is equal to the reaction number M.

This implies that the matrix K can be estimated, even when the vector $\varphi(\xi(t),\vartheta)$ has an unknown structure if the evolution of the transport-free vector can be measured or estimated.

In the following, a data-driven approach will be developed to infer the reaction stoichiometry and kinetics from the measurements of the component concentrations and bioreactor inflows and outflows (leading to the estimation of transport-free derivatives). To this end, the transport-free system (3) will be cast into a low-rank matrix approximation of the form $Y^T = H^T\cdot W^T$, where the input matrix Y^T, which collects the estimated transport-free derivatives $\frac{d\xi^\star(t)}{dt}$, is factorized into a mixing matrix H^T that will be the used to infer the number of reactions and stoichiometry and W^T, which will lead to the appropriate kinetics. Before embarking on the algorithm details, concepts of near-separable matrix factorization and successive projection algorithms are first introduced in the next section.

3. Matrix factorization methods

Linear dimensionality reduction (LDR) represents each data point as a linear combination of a small number of subspace basis elements. Mathematically, the problem can be expressed as follows: given a data set of m measurements of n species $y_j\in\mathbb{R}^m$ $(1\le j\le n)$, use LDR to search for a small number r of basis vectors $w_k\in\mathbb{R}^m$ $(1\le k\le r)$ such that each data point is approximated by a linear combination of these basis vectors:

$$y_j \approx \sum_{k=1}^{r} w_k h_{kj}, \text{ for all } j=1,\dots,n \tag{4}$$

where h_{kj} are scalar coefficients. Note that LDR is equivalent to low-rank matrix approximation (LRMA) formulated as follows:

$$\underbrace{[y_1,y_2,\dots,y_n]}_{Y\in\mathbb{R}^{m\times n}} \approx \underbrace{[w_1,w_2,\dots,w_r]}_{W\in\mathbb{R}^{m\times r}} \cdot \underbrace{[h_1,h_2,\dots,h_n]}_{H\in\mathbb{R}^{r\times n}}, \tag{5}$$

where each column of the matrix $Y\in\mathbb{R}^{m\times n}$ stands for one species j, that is, $Y(:,j)=y_j$ for $1\le j\le n$. Each column of the matrix $W\in\mathbb{R}^{m\times r}$ is a basis element, that is, $W(:,k)=w_k$ for $1\le k\le r$, and each column of the mixing matrix $H\in\mathbb{R}^{r\times n}$ contains the coordinates of a data point $Y(:,j)$ in the basis W, that is, $H(:,j)=h_j$ for $1\le j\le n$ [37].

Hence, LDR provides a rank-r approximation $W\cdot H$ of Y, and each data point is mapped into the basis W using the corresponding column of H:

$$y_j \approx W h_j, \quad \text{for all} \quad j=1,\dots,n. \tag{6}$$

Typically, the corresponding subspace basis dimension r is much smaller than the dimension n, and the number of data points m, that is, $r\ll\min(m,n)$. In order to compute W and H, given Y and r, a cost criterion to be minimized can be established, for instance, from the sum of squares of the residual $Y-WH$, e.g., using the following squared Frobenius norm:

$$\|Y-WH\|_F^2 = \sum_{i,j}(Y-WH)_{ij}^2. \tag{7}$$

In this work, LRMA techniques are used to find the minimal macroscopic reaction scheme (with the minimum number of reactions, rank(K)) required to represent a specific dataset, infer the corresponding stoichiometry K, and reveal an appropriate kinetic structure (i.e., $\varphi(t)=\mu(\xi)X$), as in the following relation:

$$\underbrace{\frac{d\xi^{\star}(t)}{dt}}_{} = \underbrace{K}_{} \quad \underbrace{\varphi(\xi(t),\vartheta)}_{} \tag{8}$$

$$Y^T_{n\times m} = H^T_{n\times r} \quad W^T_{r\times m} \tag{9}$$

where m is the number of collected samples (observations), n is the number of process species, which, in this study, are essentially extracellular measurements (i.e., biomass, metabolites, and substrate concentrations), and r is the number of biochemical reactions.

Considering Eq. (8), the following assumption must hold in order to minimize the criterion (7).

Assumption 1. The matrix $\frac{d\xi^{\star}(t)}{dt}$, also denoted $Y^T_{n\times m}$, can be obtained from the process measurement vector $\xi(t)$ by numerical differentiation.

The first step towards the solution of our problem is the idea of separable matrix factorization, which was first introduced as a subclass of nonnegative matrix factorization [40], and which states that if the input matrix Y has the form $Y = WH$, for H to be separable requires that all unit vectors are hidden among the columns of H.

Separability of H is equivalent to assuming an index set κ such that $W = Y(:,\kappa)$, i.e., the elements of the vector κ represent the column indexes of W, which appear as columns of Y [37]. Thus, the proposed method aims to recover from the input matrix Y the index vector set κ of dimension r and a matrix $H \in \mathbb{R}^{r\times n}$ with $Y = Y(:,\kappa)H$, which is equivalent to

$$Y = W\begin{bmatrix} I_r & \bar{H}\end{bmatrix}\Pi, \tag{10}$$

where I_r is the $r-by-r$ identity matrix, $\bar{H} \in \mathbb{R}^{r\times(n-r)}$ is a nonnegative matrix and Π is a permutation. This permutation factor allows the change in the column order of both Y and H matrices (see Eq. (6)) that results in having the first r columns of Y correspond to the columns of W. This allows a simplification in the implementation of the proposed algorithm.

The matrix separability proposed by [40] aims to identify the number r of columns of Y, and its indexes (vector κ), to reconstruct Y perfectly from the given $Y(:,\kappa)H$ [38]. It is essential to highlight that it is unreasonable to reconstruct Y perfectly in practice, as input matrix Y rarely admits an exact, separable matrix factorization decomposition, mainly because of noise and model structural misfit. Therefore, in practice, it is more reasonable to consider separable matrix factorization problems in the presence of noise, referred to as Near-Separable Matrix Factorization in the literature.

Near-Separable Matrix Factorization has been successful in many different applications, such as document classification, blind source separation, video compression, image classification, and hyperspectral unmixing [37], and can be stated as follows: given the noisy r-separable matrix $\tilde{Y} = WH + \mathcal{N} \in \mathbb{R}^{m\times n}$ where $\mathcal{N}$ is the noise, $W \in \mathbb{R}^{m\times r}$, $H = [I_r \ \bar{H}]\Pi$ where Π is a permutation, recover approximately the columns of W among the columns of $\tilde{Y}$. This statement is the foundation of the successive projection algorithm (SPA) used to solve the near-separable matrix factorization problem.

4. Model feature extraction based on a modified successive projection algorithm

In order to retrieve the minimal set of macroscopic reactions, the corresponding stoichiometry, and a consistent kinetic model structure from noisy measurements, a Successive Projection Algorithm (SPA) is used. SPA is a simple but fast and robust recursive algorithm, first introduced in [39], that has attracted increasing interest in many different communities in the past ten years [41]. In this study, a distinct projection formulation from [39] is used and the nonnegative restrictions imposed by [38] are relaxed.

The algorithm (see Algorithm 1) first requires an input matrix $\tilde{Y}$, the number r of basis vectors to be tested, and a specific ending flag when the residual norm is below a threshold $\epsilon_{relative}$. Before the extraction of the indexes of the vector κ, the scaling of the matrices could be carried out as (this step is not explicitly accounted for in Algorithm 1)

$$1 = \sum_i Y(i,j) = \sum_i \sum_k W(i,k)H(k,j) = \sum_k H(k,j)\sum_i W(i,k) = \sum_k H(k,j). \tag{11}$$

As highlighted by [37], the column normalization increases the algorithm robustness with respect to noise since the sum of the entries of each column of the normalized matrices Y and W is one, and $Y(:,j) = WH(:,j)$ for all j. Therefore, the sum of the entries of each column of H must also be one, for all j.

At each step, the column of the input matrix $\tilde{Y}$ with maximum ℓ_2 norm is selected (line 3) creating the vector κ (line 4). The matrix of the residual R is then updated by projecting each column onto the orthogonal complement of the columns selected so far (line 5). The computation of the residuals R uses H_r, which is the best approximation of a mixing matrix considering $W = \tilde{Y}(:,\kappa)$ (line 6). The algorithm stops when $k > r$ or $\text{argmax}_\kappa \|R(:,\kappa)\|_2 \le \epsilon_{relative} \cdot \text{argmax}_\kappa \|\tilde{Y}(:,\kappa)\|_2$ (line 2). Note that for the first iteration we have $\|R(:,\kappa)\|_2 = \|\tilde{Y}\|_2$.

Algorithm 1 Successive Projection Algorithm

Input: Near-separable matrix $\tilde{Y} = WH + \mathcal{N} \in \mathbb{R}^{m\times n}$, and the number r of columns to be extracted.

Output: Set of indices κ such that $\tilde{Y}(:,\kappa) \approx W$ up to permutation and mixing matrix H_r.

1: Let $R = \tilde{Y}$, $\kappa = \{\}$, $k = 1$.
2: **while** $\text{argmax}_\kappa \|R(:,\kappa)\|_2 / \text{argmax}_\kappa \|\tilde{Y}(:,\kappa)\|_2 > \epsilon_{relative}$ and $k \le r$ **do**
3: $p = armax_j \|R_{:j}\|_2$.
4: $\kappa = \kappa \cup \{p\}$.
5: $R(:,j) = \tilde{Y}(:,j) - \tilde{Y}(:,\kappa)H_r(:,j)$ for all j, where
6: $H_r(:,j) = \underset{h_j}{\text{argmin}} \|\tilde{Y}(:,j) - \tilde{Y}(:,\kappa)h_j\|_2$;
7: $k = k + 1$.
8: **end while**

The advantages of SPA, highlighted in [38,39,41], are (i) the algorithm execution time, corresponding to $2mnr + \mathcal{O}(mr)$ operations for a dense input matrix $\tilde{Y}$ and $\mathcal{O}(r\ nnz(\tilde{Y}))$ operations for a sparse input matrix $\tilde{Y}$, where $nnz(\tilde{Y})$ is the number of nonzero entries in $\tilde{Y}$, (ii) the choice of the number of columns of $\tilde{Y}$ being the sole parameter to tune. However, the main issue of this approach is its sensitivity to outliers, which can be minimized by preprocessing the input data matrix.

4.1. Unveiling bioprocess models

SPA will be used repeatedly for increasing values of parameter r, and the following cost function will be computed:

$$J_r = \epsilon_r\ Q^{-1}\ \epsilon_r^T, \tag{12}$$

where $\epsilon_r = \sum_{i=1}^{n} |Y - W_r H_r|$, with $W_r = Y(:,\kappa)$ and H_r is computed by the SPA algorithm for the predefined r-dimension. Q is a diagonal scaling matrix whose diagonal elements are the maximum absolute values from each column of Y. Note that J_r is a decreasing function of r, i.e., the larger the number of reactions r, the smaller the fitting error norm and J_r). It is always smaller or equal to the log-likelihood cost $J^{\star}$, known to have a chi-square distribution with $N \cdot n$ degrees of freedom [42], where N is the number of components. Thus, the number of macroscopic reactions is chosen as the smallest r such that the fitting error is smaller or equal to the range of a $\chi_{n\cdot N}$-distributed random variable.

The next step is to infer the pseudo-stoichiometric matrix $\hat{K}$ using the mixing matrix H_r^T obtained by SPA. It can be defined as a linear combination of the basis vectors, i.e.,

$$\hat{K} = H_r^T \Omega, \tag{13}$$

where Ω is a r by r regular matrix imposing a maximum of r constraints per column of $\hat{K}$. These constraints confer biological consistency using

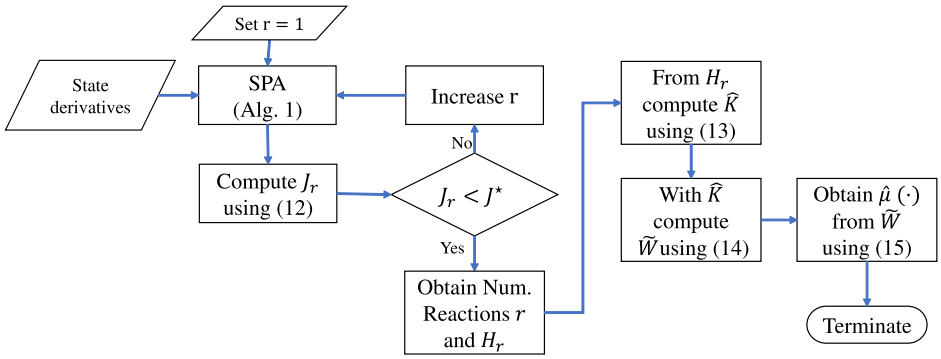


Fig. 1. Flowchart of the procedure to extract the reaction number, stoichiometry and information on the kinetics.

a priori knowledge about cell metabolism. They allow, for instance, eliminating specific reactants or products from one of the reactions or normalizing a reaction with respect to a specific compound [11]. Note that the matrix $W_r = Y(:,\kappa)$, obtained by SPA, corresponds to the best estimate considering the mixing matrix H_r. However, due to the linear transformation (13), matrix W_r should, in general, be recomputed considering the stoichiometric matrix $\hat{K}$, which is a constrained linear combination of the basis vectors of H_r, by the following optimization:

$$\arg\min_{\tilde{W}>=0} \|Y - \tilde{W}\hat{K}^T\|_F^2, \tag{14}$$

which is a special case of (7), where the matrix $\hat{K}^T$ is fixed.

The new matrix $\tilde{W}$ contains the time evolution of the reaction rates $\varphi(\xi(t),\vartheta)$, and the growth rate signal $\hat{\mu}(\cdot)$ can be obtained by

$$\hat{\mu}(\cdot) = \frac{\tilde{W}}{X}. \tag{15}$$

It is now possible to summarize the approach in the flowchart of Fig. 1.

4.2. Illustrative example

A simple application considering cell growth on glucose in batch mode [2] is now considered to illustrate and develop the numerical procedure. The reaction involves the consumption of a substrate (glucose, G) and the production of biomass (X) combined with the release of a byproduct (lactate, L). Therefore, the following macroscopic reaction is the underlying mechanism:

$$G \xrightarrow{\varphi(X,G)} k_X X + k_L L, \tag{16}$$

where k_X and k_L are the biomass and lactate pseudo-stoichiometric coefficients, respectively, and $\varphi(X,G)$, is the reaction rate, function of the available biomass and substrate concentrations. Applying mass balance to (16) yields the following ordinary differential equation system that describes the time evolution of the concentrations X, G, and L:

$$\frac{dX}{dt} = k_X\varphi(X,G), \tag{17a}$$
$$\frac{dG}{dt} = -\varphi(X,G), \tag{17b}$$
$$\frac{dL}{dt} = k_L\varphi(X,G). \tag{17c}$$

The specific growth rate and the reaction rate are respectively defined as

$$\mu(G) = \mu_{max}\frac{G}{(K_G+G)}, \tag{18}$$

$$\varphi(\cdot) = \mu(\cdot)X, \tag{19}$$

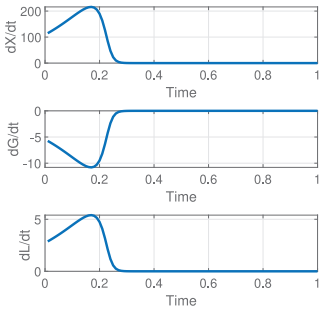
where K_G is the half-saturation constant, and μ_{max} the maximum specific growth rate.

A simulation considering $k_X = 20$, $k_L = 0.5$, $\mu_{max} = 0.3$ d^{-1}, and $K_G = 0.2$ g/l is performed to generate a dataset containing the time evolution of the derivatives of the concentration trajectories that can be stored into the input matrix Y. Note that this information is assumed to be noise-free, an assumption that will be relaxed later on in this study. The first step is obtaining the minimum number of macroscopic reactions explaining the data by using the procedure explained in Fig. 1. Therefore, we compute the 99.9% quantile of $\chi_{n\cdot N}$, resulting in $J^\star = 760.66$. Next, we assume that the dataset can be represented by only one reaction, $r = 1$, and compute (12), which gives $J_1 = 3.7\times10^{-30}$. This simple test concludes that the 1-reaction macroscopic scheme is a good model candidate, as $J_1 < J^\star$. This means that in this application, where there is no measurement noise and no noise amplification by computing time derivatives, we obtain the exact factorization of matrix Y.

Given $Y = \begin{bmatrix} dX/dt & dG/dt & dL/dt \end{bmatrix}$ and $r = 1$, the next step consists of computing the pseudo-stoichiometric matrix $\hat{K}$ using biologically inspired constraints and matrix H_r obtained by SPA.

Considering the hypothetically true system, Fig. 2 illustrates the relations between concentration derivatives, stoichiometric matrix, and kinetic law.

$$\frac{d\xi}{dt} = K \cdot \varphi(\xi),$$



$$= \begin{bmatrix} k_X \\ -1 \\ k_L \end{bmatrix} \cdot \varphi(\xi).$$

Fig. 2. Relation between concentration derivatives, stoichiometric matrix, and kinetic law.

The output of the SPA algorithm reveals $\kappa = [1]$, thus $W_r = Y(:,\kappa) = dX/dt$ and the mixing matrix $H_r = \begin{bmatrix} 1.0000 & -0.0500 & 0.0250 \end{bmatrix}$, which can be illustrated by the relations presented in Fig. 3.

It is easy to see the importance of the separability assumption in this example. The growth rate $\varphi(\xi)$, represented by matrix W_r^T, is indeed an image of the derivative of the biomass concentration dX/dt, that can be used to compute the growth rate signal $\mu(G)$. In this simple case the stoichiometry is normalized with respect to G (see (16)), and $\hat{K}$ can be retrieved as

$$\hat{K} = \left[\frac{H_r}{|H_r(1,2)|}\right]^T = \begin{bmatrix} 20.000 \\ -1.0000 \\ 0.5000 \end{bmatrix}, \tag{20}$$

resulting in the same stoichiometric values as the nominal ones. Note that the same result is obtained using Eq. (13).

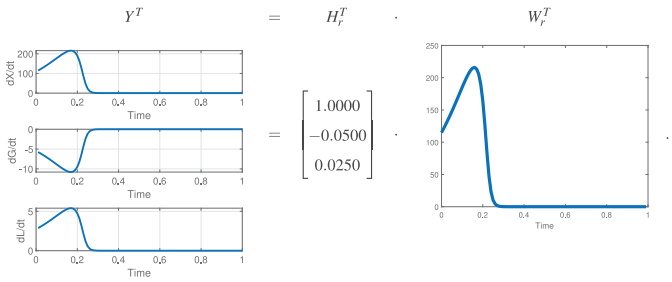


Fig. 3. Outcomes of the SPA algorithm.

Remark 1. In simple cases, such as the example under consideration, $\hat{K}$ can be obtained by the normalization of H_r considering one of its elements, and the optimization (14) is not required. The estimation of $\varphi(\xi(t), \vartheta)$ can be obtained by the normalization of W_r considering the same element of H_r as in Eq. (21).

$$\hat{\mu}(\cdot) = \frac{W_r \cdot |H_r(1,2)|}{X} \equiv \frac{\tilde{W}}{X}, \tag{21}$$

as shown in Fig. 4, which represents the Monod law as defined by Eq. (18).

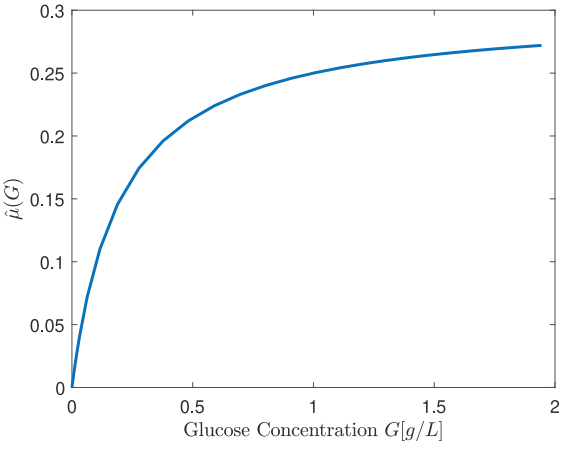


Fig. 4. Monod law estimation computed by Eq. (21).

Remark 2. In Eq. (21), the division operation symbol is applied between two vectors of respective dimensions $1 \times m$. In the present case, an element-by-element division is considered.

4.3. Differentiating noisy signals

In real-life applications, the computation of the stoichiometric matrix and kinetics has to face the noise present in the measurements. Calculating derivatives of concentration time evolutions is a crucial step of the procedure. However, differentiating experimental data can be a challenge due to measurement sparsity and noise [43].

This challenge can be effectively addressed in two directions. First, as discussed in the introduction, the introduction of PAT in many bioprocess industries is reducing the traditional scarcity of data that was observed in the past. Second, the use of recent regularization techniques that impose smoothness on the signals might provide the required derivative estimates.

The topic of numerical differentiation of noisy signals has attracted considerable attention in the literature. According to [44], numerical differentiation methods based on filtering, Tikhonov regularization, and smoothing splines have all demonstrated success in different scenarios. Other alternatives include the use of sliding mode techniques and Levant's robust differentiators [45]. In [46], generalized super twisting observers are developed while [47] proposes arbitrary-order fixed-time differentiators. In [48], the use of state observers is highlighted for obtaining time derivatives when a model of the system and the noise characteristics are known. This restriction can be alleviated using the constant acceleration forward–backward Kalman smoother, which can be developed considering a simple model and noise covariances [49].

In this study, the total-variation regularized difference method (`TVRegDiff`) developed in [50] is considered to compute the measurement derivatives. There are mainly two key benefits related to this method. First, it effectively decreases noise by reducing the total variation. Secondly, it does not suppress jump discontinuities, making it an ideal choice for detecting corners and edges in noisy data while also computing discontinuous derivatives [50]. One of the drawbacks of this method is that it also requires hyperparameter tuning and causes aliasing, so there is the need to trim the ends of the time series generated by each initial condition [26,51,52]. This phenomenon typically also happens in interpolation methods where the error tends to be larger at the boundaries of the interpolating interval since the second derivatives of the approximating functions are assumed to vanish there [53].

In order to illustrate some features of the differentiation methods, Figs. 5 and 6 present the results of the application of `TVRegDiff` and a smoothing spline (Matlab function `spaps` [54]) to a set of experimental data (trajectory of lactate measurements in a cell culture).

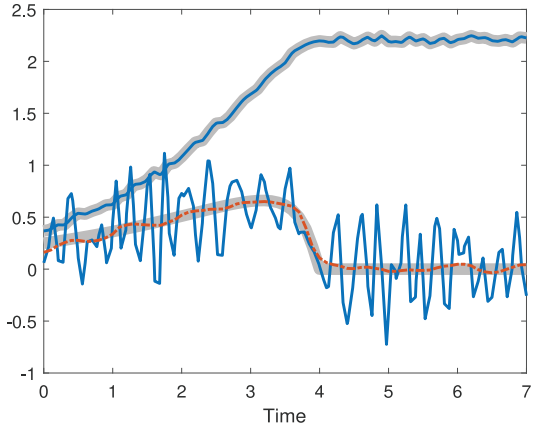


Fig. 5. Using splines with moderate smoothing: measurement signals (top), and their derivatives (bottom). Thick-gray lines are the ground truth values, blue lines are the smoothing spline (Matlab function `spaps`), and red dashed-line is the `TVRegDiff` derivatives.

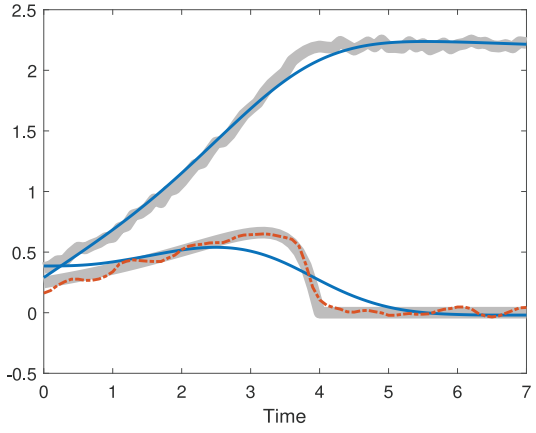


Fig. 6. Splines with stronger smoothing: measurement signals (top) and their derivatives (bottom). Thick-gray lines are the ground truth values, blue lines are the smoothing spline (Matlab function `spaps`), and red dashed-line is the `TVRegDiff` derivatives.

Table 1
Simulation parameters, identified values, and their standard deviations for the one-reaction case study.

Monod kinetics			Haldane kinetics			Contois kinetics		
Stoichiometric parameters			Stoichiometric parameters			Stoichiometric parameters		
	Value	Identified		Value	Identified		Value	Identified
k_X	20.0	(20.39 ± 0.234)	k_X	15.0	(14.96 ± 0.387)	k_X	25.0	(24.73 ± 0.601)
k_L	0.50	(0.501 ± 0.005)	k_L	0.8	(0.794 ± 0.011)	k_L	6.0	(5.813 ± 0.129)
Kinetic parameters			Kinetic parameters			Kinetic parameters		
μ_{max}	0.3 d^{-1}	–	μ_{max}	0.3 d^{-1}	–	μ_{max}	0.3 d^{-1}	–
K_G	0.5 g/l	–	K_G	1 g/l	–	K_C	0.5 g/X	–
			K_I	0.06 g^2/l^2	–			

When splines are used to smooth data, two extreme situations can occur: (1) the smoothing spline follows the data realization including the noise, or (2) the smoothing spline cuts off much of the data variability including part of the process dynamics. Fig. 5 shows that moderate data smoothing will result in derivatives with large variations. In contrast, Fig. 6 shows that strong spline smoothing may result in a loss of information about the derivatives. Conversely, the derivatives obtained by `TVRegDiff` are quite close to the true values, even when the derivatives present an abrupt behavior around time 4.

It is important to stress that depending on the signal-to-noise ratio, satisfactory results can be obtained by the direct use of smoothing splines, as in [18,55–57] to name a few.

In this study, we first use moderate smoothing splines in order to resample the signal and have enough data points for the rest of the procedure. Then, we apply `TVRegDiff` to compute derivatives of the resampled signal and trim the ends of the time series generated by each initial condition (first and last 10%), to improve accuracy.

The illustrative example of Section 4.2 is now revisited considering noisy measurements. In addition, three alternative models for the specific growth rate are considered:

$$\mu_M(G) = \mu_{max}\frac{G}{(K_G + G)}, \tag{22a}$$

$$\mu_H(G) = \mu_{max}\frac{G}{\left(K_G + G + \frac{G^2}{K_I}\right)}, \tag{22b}$$

$$\mu_C(X, G) = \mu_{max}\frac{G}{(K_C X + G)}, \tag{22c}$$

with the corresponding reaction rate:

$$\varphi_i(\cdot) = \mu_i(\cdot)X, \tag{23}$$

where $i = \{M,\ H,\ C\}$ are indices respectively standing for Monod, Haldane, and Contois laws, K_G is the half-saturation constant, K_I the inhibition constant, K_C is the Contois constant and μ_{max} the maximum specific growth rate.

A dataset is generated in simulation using model (17) with a sampling period $t_s = 0.01d$ (around 15 min) for a batch lasting one day. A Monte Carlo analysis is carried out to quantify the uncertainties in the model feature extraction (stoichiometry and kinetics) when measurements are affected by noise. To this end, 100 different realizations of Gaussian noises with a standard deviation of 1% are considered for each scenario (i.e., Monod, Haldane, Contois). To obtain the measurement derivatives, smoothing splines are first applied and resampled (three times more), then, `TVRegDiff` is used. The first and last 10% of the time series are trimmed to improve the overall accuracy.

Given $\bar{Y} = \begin{bmatrix} dX/dt & dG/dt & dL/dt \end{bmatrix}$ and $r = 1$, the pseudo-stoichiometric matrix K is computed along the same lines as in Section 4.2 using biologically inspired constraints and matrix H_r obtained by SPA. Table 1 presents the stoichiometry identification with the corresponding standard deviations, which are quite satisfactory results. The growth rate evolutions as a function of glucose are shown in Figs. 7–9, respectively.

It is apparent in Figs. 7 and 8 that the variance is higher for large glucose concentrations. These concentrations occur at the start of the culture, and this variability is associated with the less accurate determination of the signal derivatives towards the ends of the time interval under consideration. However, on the whole, the kinetic laws are fairly well reconstructed.

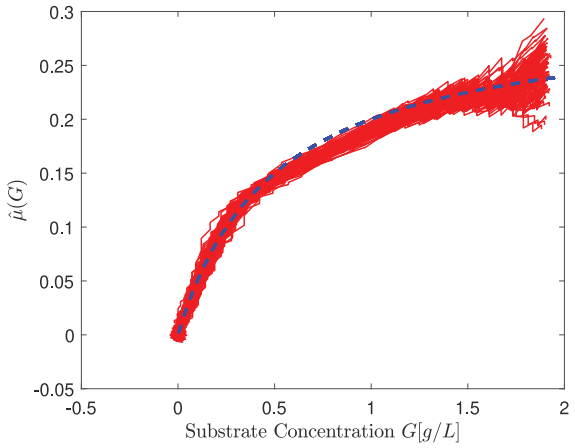


Fig. 7. Monod law estimation. The dashed blue line presents the nominal model. The red lines show the estimations for the 100 Gaussian noise realizations.

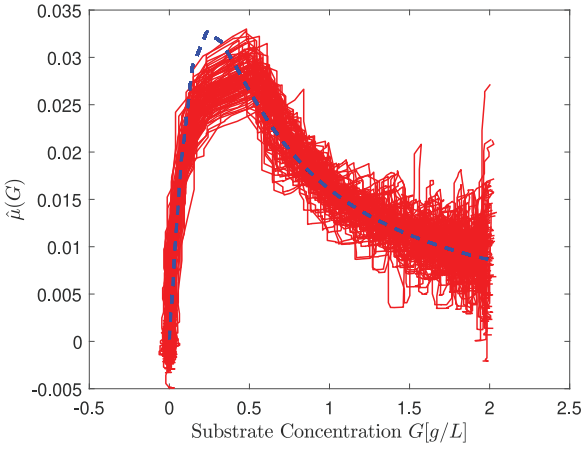


Fig. 8. Haldane Law estimation. The dashed blue line presents the nominal model. The red lines show the estimations for the 100 Gaussian noise realizations.

5. Case study: Hybridoma cell catabolism

In this section, the proposed method is tested in a more realistic scenario. The model presented in Dewasme et al. [11] is used to emulate a culture of hybridoma cells. Three macroscopic reactions are considered:

(a) Substrate oxidation:

$$k_{31}G + k_{41}Gn \xrightarrow{\varphi_1} X_v + k_{61}MAb, \tag{24}$$

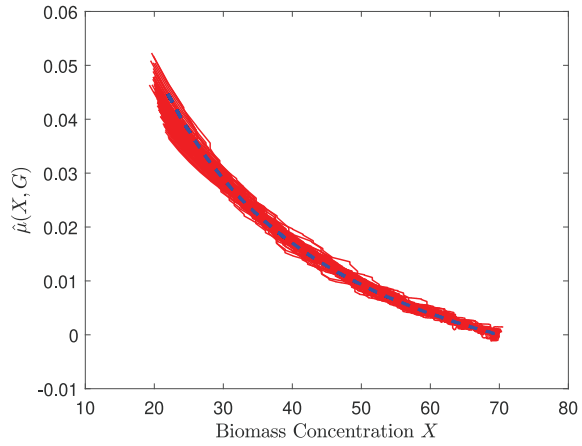


Fig. 9. Contois model estimation. The dashed blue line presents the nominal model. The red lines show the estimations for the 100 Gaussian noise realizations.

(b) Substrate overflow:

$$k_{32}G + k_{42}Gn \xrightarrow{\varphi_2} X_v + k_{52}L, \tag{25}$$

(c) Biomass death

$$X_v \xrightarrow{\varphi_3} X_d + k_{63}MAb, \tag{26}$$

where X_v, X_d, G, Gn, L, and MAb are the concentrations of viable biomass, dead biomass, glucose, glutamine, lactate, and monoclonal antibodies (MAb), respectively. φ_j is the jth reaction rate and k_{ij} is the stoichiometric coefficient of the ith compound in the jth reaction.

Applying mass balance to the above macroscopic reactions yields the following differential equation system:

$$\frac{dX_v}{dt} = \varphi_1 + \varphi_2 - \varphi_3, \tag{27a}$$
$$\frac{dX_d}{dt} = \varphi_3, \tag{27b}$$
$$\frac{dG}{dt} = -k_{31}\varphi_1 - k_{32}\varphi_2, \tag{27c}$$
$$\frac{dGn}{dt} = -k_{41}\varphi_1 - k_{42}\varphi_2, \tag{27d}$$
$$\frac{dL}{dt} = k_{52}\varphi_2, \tag{27e}$$
$$\frac{dMAb}{dt} = k_{61}\varphi_1 + k_{63}\varphi_3, \tag{27f}$$

where the specific reaction rates represent the overflow metabolism and cell decay, according to [9]:

$$\mu_1 = \min(\mu_G, \mu_{Gmax}), \tag{28a}$$
$$\mu_2 = \max(0, (\mu_G - \mu_{Gmax})), \tag{28b}$$
$$\mu_3 = \mu_{dmax}\frac{K_{Gd}}{K_{Gd}+G}\frac{K_{Gnd}}{K_{Gnd}+Gn}, \tag{28c}$$

where

$$\mu_G = \mu_{max1}\frac{Gn}{K_{Gn}+G_n}, \quad \mu_{Gmax} = \mu_{max2}, \text{ and} \tag{29}$$

$$\varphi_i = \mu_i \cdot X_v, \qquad i = \{1, 2, 3\}. \tag{30}$$

A dataset is generated by simulation of model (27) with the parameters of Table 2, considering a sampling time of $t_s = 0.1d$, and the addition of Gaussian noise on the measurements, with a relative standard deviation of 0.5%.

Table 2
Simulation parameters [11].

Parameters	Values	Parameters	Values
μ_{max1}	0.484 d^{-1}	k_{31}	3.12
μ_{max2}	0.319 d^{-1}	k_{32}	15.2
K_{Gn}	0.0089 g/l	k_{41}	0.624
K_{Gd}	1.58 g/l	k_{42}	1.22
K_{Gnd}	1.33 g/l	k_{52}	23.9
μ_{dmax}	0.866 d^{-1}	k_{61}	43.5
K_G	0.100 g/l	k_{63}	14.2
$X_v(0)$	0.100 cells/ml	$X_d(0)$	0.0151 cells/ml
$G(0)$	5.99 g/l	$Gn(0)$	0.303 g/L
$L(0)$	0.360 g/l	$MAb(0)$	6.53 µg/ml

5.1. Selection of the subspace dimension

The number of reactions is obtained by analyzing the values of J_r computed by (12), as we can see in the flowchart in Fig. 1. Thus, the number of macroscopic reactions is chosen as the smallest r such that the fitting error is smaller or equal to the range of a $\chi_{n\cdot N}$-distributed random variable, see Fig. 10.

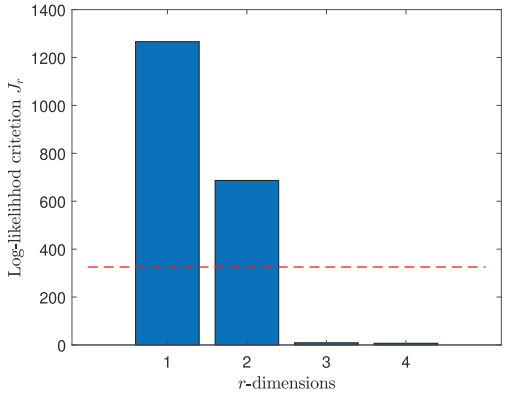


Fig. 10. Results of J_r for each r-dimension.

The histogram of Fig. 10 shows that a 3-reaction macroscopic scheme is the best model candidate since the corresponding log-likelihood is smaller than the 99.9% quantile represented by the red dashed line. This matches the (unknown) number of reactions of the process emulator.

5.2. Inference of the stoichiometry and kinetics

Using the process model (which is unknown to the user of the data-driven tool, but useful to interpret the results in this illustrative example), Fig. 11 shows the relations between the derivatives of the concentration trajectories, the stoichiometric matrix, and the kinetic laws.

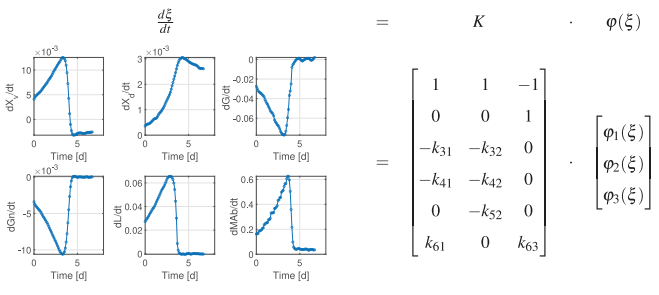


Fig. 11. Case Study. Relation between concentration derivatives, stoichiometric matrix, and kinetic law.

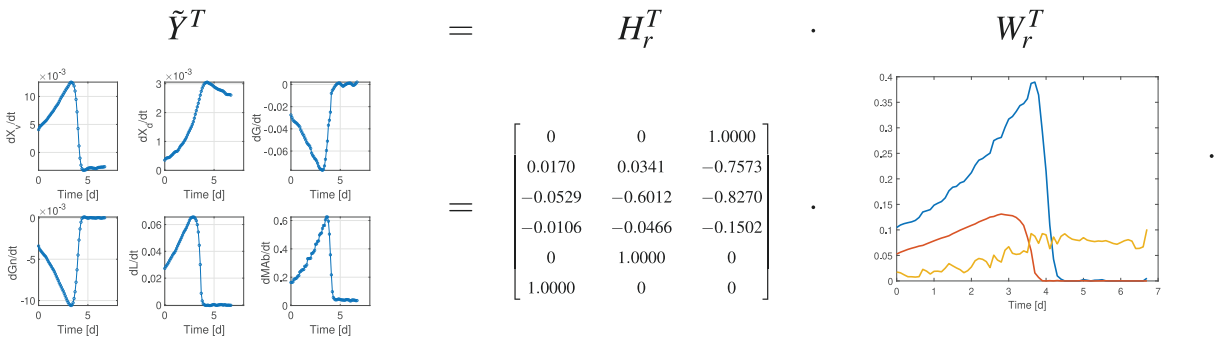


Fig. 12. Case Study. Outcomes of the SPA algorithm.

In the data-driven procedure, the derivatives $\tilde{Y} = [dX_v/dt \quad dX_d/dt \quad dG/dt \quad dGn/dt \quad dL/dt \quad dMAb/dt]$ are used (they are obtained by a combination of spline smoothing and TVD differentiation), $r = 3$ is adopted according to the results of the previous subsection, and SPA provides $\kappa = \begin{bmatrix}6 & 5 & 1\end{bmatrix}$, with $W_r = Y(:,\kappa)$ and H_r as presented in Fig. 12.

The factorization of the matrix $\tilde{Y}$ yields a matrix H_r^T emphasized with unitary values in rows 1, 5 and 6.

The reaction stoichiometry is now normalized with respect to the viable biomass, and $\hat{K}$ is computed according to (13), with the assumption that there is a reaction that is associated with biomass death and that MAbs are produced only in the growth and death reactions, to yield

$$\hat{K} = \begin{bmatrix} \mathbf{1.0000} & \mathbf{1.0000} & \mathbf{-1.0000} \\ \mathbf{0} & \mathbf{0} & \mathbf{1.0000} \\ -3.1335 & -14.3622 & 0.0012 \\ -0.6212 & -1.1954 & -0.0001 \\ 0 & 22.4159 & 0 \\ 44.3676 & \mathbf{0} & 14.1840 \end{bmatrix}, \tag{31}$$

where the bold values are the biologically inspired constraints imposed on the computations of $\hat{K}$. The values of $\hat{k}_{ij}$ are quite close to the nominal parameters presented in Table 2.

It is now required to recompute the W_r matrix to obtain the $\varphi_i(\cdot)$ values. To this end, (14) is solved with $\hat{K}$, resulting in a new W matrix, called $\bar{W}$. From relation (30), the specific reaction rate profiles are computed as

$$\hat{\mu}(\cdot) = \begin{bmatrix} \frac{\bar{W}(:,1)}{X_v} & \frac{\bar{W}(:,2)}{X_v} & \frac{\bar{W}(:,3)}{X_v} \end{bmatrix}. \tag{32}$$

Fig. 13 shows $\hat{\mu}_i(\cdot)$ with $i = \{1,2,3\}$ as functions of Gn. The estimated evolutions offer the possibility to extract candidate kinetic structures. The blue line represents $\mu_1(\cdot)$, which could be described by a Monod law. Indeed, an activation/saturation behavior is observed (see (30) and Table 2). The red line shows the evolution of $\mu_2(\cdot)$, starting with a slow increase, followed anew by a Monod profile. One way to express this behavior is to consider a discontinuous switching function, as in (28b). Alternatively, combining continuous rate expressions with inhibition factors could also be used, as already proposed in [58,59]. Lastly, the yellow curve corresponds to $\mu_3(\cdot)$, where the inhibition profile is evident and consistent with (28c).

This case study demonstrates the usefulness of the proposed procedure in extracting information about stoichiometry and kinetics from datasets of the evolution of the culture species. The most critical step in the procedure is the computation of the numerical differentiation of these signals whose quality can greatly affect the results.

5.3. Sensitivity of SPA to measurement noise

To assess the robustness of the SPA algorithm to noise, the derivatives computed directly with the simulation model (27) are corrupted with 1000 different Gaussian noise realizations with relative standard deviation ranging from 0.01 to 1.5%.

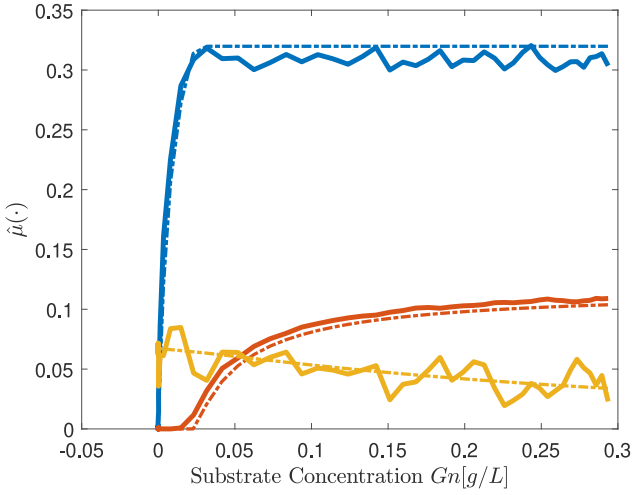


Fig. 13. Evolutions of the specific reaction rates as functions of the substrate Gn. Dashed lines stand for the nominal model values. Solid lines are the approximations from the noisy dataset. Blue, red, and yellow are respectively related to $\mu_1(\cdot)$, $\mu_2(\cdot)$, and $\mu_3(\cdot)$.

Fig. 14 presents the κ values computed by SPA, also called the cardinality of $W_r = Y(:,\kappa)$, where the reference values are $\kappa = \begin{bmatrix}6 & 5 & 1\end{bmatrix}$. The method extracts the correct structure for $\sigma = [0.01\ 1]\%$. When noise levels exceed 1%, the algorithm extracts another model structure corresponding to a different κ. The deviation occurs for 9 ‰ of the cases with $\sigma = 1\%$ and 61% with $\sigma = 1.5\%$. This can be considered as a worst-case scenario since no denoising methods are used (no smoothing spline or other filtering methods). In addition, Fig. 15 presents the relative error of the stoichiometric parameters, defined as

$$\epsilon_{\hat{k}_{ij}} = \frac{k_{ij} - \hat{k}_{ij}}{k_{ij}}, \tag{33}$$

where i and j are the desired indices of the elements of the identified stoichiometric matrix $\hat{K}$. As expected, the estimation is accurate at low noise levels. However, the parameter estimates deteriorate for $\sigma = 1\%$ and $\sigma = 1.5\%$, resulting in significant relative errors.

6. Conclusions

A bioprocess data-driven modeling strategy is proposed using a Successive Projection Algorithm. This data-driven tool reveals the minimal set of macroscopic reactions, computes the corresponding stoichiometry, and simplifies the choice of an adequate kinetic structure. A case study considering Hybridoma Cell cultures is used to validate the method. Future work focuses on alternative numerical differentiation

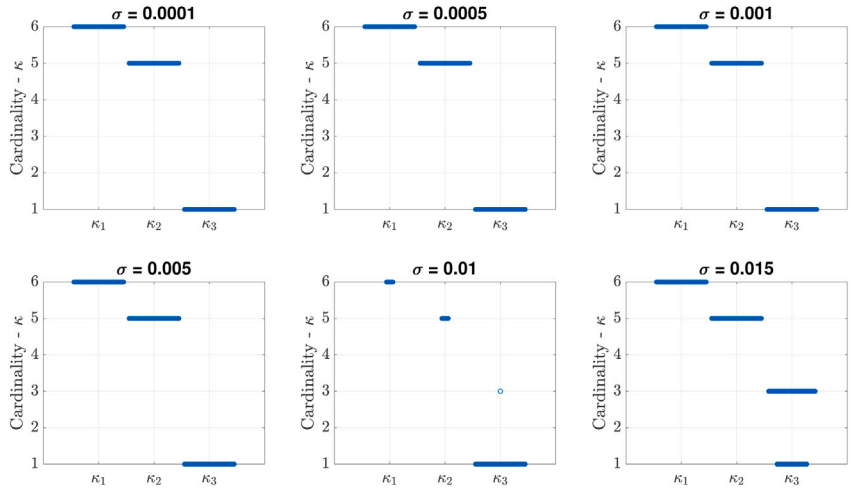


Fig. 14. Values of κ for each noise magnitude considering 1000 noise realizations for each σ.

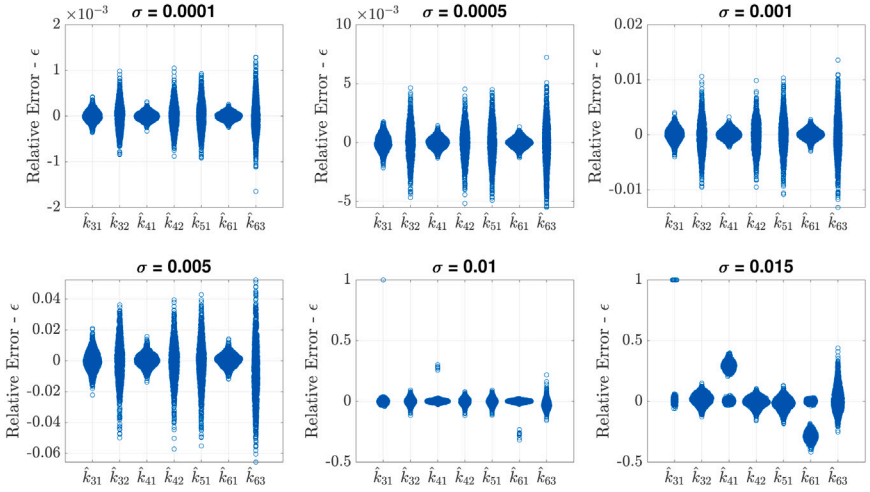


Fig. 15. Relative error of the stoichiometric parameters for each noise magnitude considering 1000 noise realizations for each σ.

schemes in order to enhance robustness to measurement noise and on the development of systematic data-driven methods for the identification of the species involved in the reaction kinetics and the selection of the most likely kinetic model structure.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

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

[1] G. Gerzon, Y. Sheng, M. Kirkitadze, Process analytical technologies – advances in bioprocess integration and future perspectives, J. Pharm. Biomed. Anal. 207 (2021).

[2] G. Bastin, D. Dochain, On-Line Estimation and Adaptive Control of Bioreactors, in: Process Measurement and Control, vol. 1, Elsevier, Amesterdam, 1990.

[3] P.A. Vanrolleghem, Dynamical Modelling & Estimation in Wastewater Treatment Processes, IWA publishing, 2001.

[4] R. Antonelli, J. Harmand, J.-P. Steyer, A. Astolfi, Set-point regulation of an anaerobic digestion process with bounded output feedback, IEEE Trans. Control Syst. Technol. 11 (4) (2003) 495–504.

[5] G.A. Pimentel, A. Vande Wouwer, A. Rapaport, J. Harmand, Modeling of submerged membrane bioreactors with a view to control, in: 11th IWA Conference on Instrumentation Control and Automation, ICA2013, 2013.

[6] O. Bernard, Z. Hadj-Sadok, D. Dochain, A. Genovesi, J.P. Steyer, Dynamical model development and parameter identification for an anaerobic wastewater treatment process, Biotechnol. Bioeng. 75 (4) (2001) 424–438.

[7] M. Sbarciog, M. Loccufier, E. Noldus, Determination of appropriate operating strategies for anaerobic digestion systems, Biochem. Eng. J. 51 (2010) 180–188.

[8] A. Henrotin, A.-L. Hantson, L. Dewasme, Dynamic modeling and parameter estimation of biomethane production from microalgae co-digestion, Bioprocess Biosyst. Eng. 46 (1) (2023) 129–146.

[9] Z. Amribt, H. Niu, P. Bogaerts, Macroscopic modelling of overflow metabolism and model based optimization of hybridoma cell fed-batch cultures, Biochem. Eng. J. 70 (2013) 196–209.

[10] B. Ben Yahia, L. Malphettes, E. Heinzle, Macroscopic modeling of mammalian cell growth and metabolism, Appl. Microbiol. Biotechnol. 99 (17) (2015) 7009–7024.

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

[12] L. Dewasme, M. Mäkinen, V. Chotteau, Practical data-driven modeling and robust predictive control of mammalian cell fed-batch process, Comput. Chem. Eng. 171 (2023) 108164.

[13] O. Bernard, Hurdles and challenges for modelling and control of microalgae for CO_2 mitigation and biofuel production, J. Process Control 21 (10) (2011) 1378–1389.

[14] O. Bernard, F. Mairet, B. Chachuat, Modelling of microalgae culture systems with applications to control and optimization, Microalgae Biotechnol. (2016) 59–87.

[15] D. Coutinho, A. Vargas, C. Feudjio, M. Benavides, A. Vande Wouwer, A robust approach to the design of super-twisting observers - application to monitoring microalgae cultures in photo-bioreactors, Comput. Chem. Eng. 121 (2019) 46–56.

[16] F.A. Gorrini, J.M. Zamudio Lara, S.I. Biagiola, J.L. Figueroa, H. Hernández Escoto, A.-L. Hantson, A. Vande Wouwer, Experimental study of substrate limitation and light acclimation in cultures of the microalgae *Scenedesmus obliquus*- Parameter identification and model predictive control, Processes 8 (12) (2020) 1551.

[17] H.Æ. Oddsdóttir, E. Hagrot, V. Chotteau, A. Forsgren, On dynamically generating relevant elementary flux modes in a metabolic network using optimization, J. Math. Biol. 71 (2015) 903–920.

[18] M. Maton, P. Bogaerts, A. Vande Wouwer, A systematic elementary flux mode selection procedure for deriving macroscopic bioreaction models from metabolic networks, J. Process Control 118 (2022) 170–184.

[19] A.L. Oliveira, Biotechnology, big data and artificial intelligence, Biotechnol. J. 14 (8) (2019) 1–6.

[20] L. Dewasme, Neural network-based software sensors for the estimation of key components in brewery wastewater anaerobic digester: An experimental validation, Water Sci. Technol. 80 (10) (2019) 1975–1985.

[21] A.W. Rogers, I.O.S. Cardenas, E.A. Del Rio-Chanona, D. Zhang, Investigating physics-informed neural networks for bioprocess hybrid model construction, in: A.C. Kokossis, M.C. Georgiadis, E. Pistikopoulos (Eds.), 33rd European Symposium on Computer Aided Process Engineering, in: Computer Aided Chemical Engineering, vol. 52, Elsevier, 2023, pp. 83–88.

[22] A. Vande Wouwer, C. Renotte, P. Bogaerts, Biological reaction modeling using radial basis function networks, Comput. Chem. Eng. 28 (11) (2004) 2157–2164.

[23] D. Lee, A. Jayaraman, J.S. Kwon, Development of a hybrid model for a partially known intracellular signaling pathway through correction term estimation and neural network modeling, PLoS Comput. Biol. 16 (12) (2020) 1–31.

[24] M. Maton, P. Bogaerts, A. Vande Wouwer, Hybrid dynamic models of bioprocesses based on elementary flux modes and multilayer perceptrons, Processes 10 (10) (2022) 2084.

[25] P.P. Mondal, A. Galodha, V.K. Verma, V. Singh, P.L. Show, M.K. Awasthi, B. Lall, S. Anees, K. Pollmann, R. Jain, Review on machine learning-based bioprocess optimization, monitoring, and control systems, Bioresour. Technol. 370 (2023) 128523.

[26] S.L. Brunton, J.L. Proctor, J.N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. 113 (15) (2016) 3932–3937.

[27] P.J. Schmid, Dynamic mode decomposition and its variants, Annu. Rev. Fluid Mech. 54 (1) (2022) 225–254.

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

[29] M. Wang, R.S. Risuleo, E.W. Jacobsen, V. Chotteau, H. Hjalmarsson, Identification of nonlinear kinetics of macroscopic bio-reactions using multilinear Gaussian processes, Comput. Chem. Eng. 133 (2020) 106671.

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

[31] M. Hoffmann, C. Fröhner, F. Noé, Reactive SINDy: Discovering governing reactions from concentration data, J. Chem. Phys. 116 (2018).

[32] G.A. Pimentel, L. Dewasme, A. Vande Wouwer, On the number of reactions and stoichiometry of bioprocess macroscopic models: An implicit sparse identification approach, IFAC-PapersOnLine 56 (2) (2023) 9721–9726, 22nd World Congress of the International Federation of Automatic Control, IFAC World Congress 2023.

[33] J.P. Cunningham, Z. Ghahramani, Linear dimensionality reduction: Survey, insights, and generalizations, J. Mach. Learn. Res. 16 (2015) 2859–2900.

[34] P.D. Wentzell, D.T. Andrews, D.C. Hamilton, K. Faber, B.R. Kowalski, Maximum likelihood principal component analysis, J. Chemom. 11 (1997) 339–366.

[35] O. Bernard, G. Bastin, On the estimation of the pseudo stoichiometric matrix for macroscopic mass balance modelling of biotechnological processes, Math. Biosci. 193 (2005) 51–77.

[36] J. Mailier, M. Remy, A. Vande Wouwer, Stoichiometric identification with maximum likelihood principal component analysis, J. Math. Biol. 67 (4) (2012) 739–765.

[37] N. Gillis, Nonnegative Matrix Factorization, SIAM, 2020.

[38] N. Gillis, Successive nonnegative projection algorithm for robust nonnegative blind source separation, SIAM J. Imaging Sci. 7 (2) (2014) 1420–1450.

[39] U. Araújo, B. Saldanha, R. Galvão, H. Yoneyama, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemometr. Intell. Lab. Syst. 57 (2) (2001) 65–73.

[40] S. Arora, R. Ge, R. Kannan, A. Moitra, Computing a nonnegative matrix factorization – provably, in: Proceedings of the 44th Symposium on Theory of Computing, STOC12, 2012, pp. 145–162.

[41] S.F.C. Soares, A.A. Gomes, A.R.G. Filho, M.C.U. Araujo, R.K.H. Galvão, The successive projections algorithm, Trends Anal. Chem. 42 (2013) 84–98.

[42] P.C. Mahalanobis, On the generalized distance in statistics, in: Proceedings of the National Academy of Sciences, India, vol. 2, 1936, pp. 49–55.

[43] C.S. Vertis, N.M.C. Oliveira, F.P. Bernardo, Macroscopic dynamic modeling of metabolic shift to lactate consumption of mammalian cell batch cultures, in: Constrained Smoothing of Experimental Data in the Identification of Kinetic Models, 2016.

[44] J.M. Varah, A spline least squares method for numerical parameter estimation in differential equations, SIAM J. Sci. Stat. Comput. 3 (1) (1982) 28–46.

[45] A. Levant, Robust exact differentiation via sliding mode technique, Automatica 34 (3) (1998) 379–384.

[46] M.T. Angulo, J.A. Moreno, L. Fridman, The differentiation error of noisy signals using the generalized super-twisting differentiator, in: Proceedings of the IEEE Conference on Decision and Control, 2012, pp. 7383–7388.

[47] J.A. Moreno, Arbitrary-order fixed-time differentiators, IEEE Trans. Automat. Control 67 (3) (2022) 1543–1549.

[48] F. Van Breugel, J.N. Kutz, B.W. Brunton, Numerical differentiation of noisy data: A unifying multi-objective optimization framework, IEEE Access 8 (2020) 196865–196877.

[49] J.L. Crassidis, J.L. Junkins, Optimal Estimation of Dynamic Systems, Chapman & Hall / CRC, 2nd ed., 2012.

[50] R. Chartrand, Numerical differentiation of noisy, nonsmooth data, ISRN Appl. Math. 2011 (2011) 1–11.

[51] K. Kaheman, J.N. Kutz, S.L. Brunton, SINDy-PI: A robust algorithm for parallel implicit sparse identification of nonlinear dynamics, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 476 (2242) (2020).

[52] N.M. Mangan, S.L. Brunton, J.L. Proctor, J.N. Kutz, Inferring biological networks by sparse identification of nonlinear dynamics, IEEE Trans. Mol. Biol. Multi-Scale Commun. 2 (1) (2016) 52–63.

[53] C. Huang, Boundary corrected cubic smoothing splines, J. Stat. Comput. Simul. 70 (2) (2001) 107–121.

[54] C. Reinsch, Smoothing by spline functions, Numer. Math. 10 (1967) 177–183.

[55] A. Grosfils, A.V. Wouwer, P. Bogaerts, On a general model structure for macroscopic biological reaction rates, J. Biotech. 130 (3) (2007) 253–264.

[56] P. Bogaerts, J. Castillo, R. Hanus, A general mathematical modelling technique for bioprocesses in engineering applications, Syst. Anal. Modelling Simul. 35 (1999) 87–113.

[57] A. Richelle, P. Bogaerts, Systematic methodology for bioprocess model identification based on generalized kinetic functions, Biochem. Eng. J. 100 (2015) 41–49.

[58] P.C. Segura, R. Wattiez, A.V. Wouwer, B. Leroy, L. Dewasme, Dynamic modeling of *Rhodospirillum rubrum* PHA production triggered by redox stress during VFA photoheterotrophic assimilations, J. Biotech. 360 (2022) 45–54.

[59] G.A. Pimentel, L. Dewasme, F.N. Santos-Navarro, A. Boes, F. Côte, P. Filée, A. Vande Wouwer, Macroscopic dynamic modeling of metabolic shift to lactate consumption of mammalian cell batch cultures, in: 2023 9th International Conference on Control, Decision and Information Technologies (CoDIT), 2023, pp. 1–6.