Delft University of Technology

Delft Center for Systems and Control

Technical report 24-008

An identification algorithm of switched Box-Jenkins systems in the presence of bounded disturbances: An approach for approximating complex biological wastewater treatment models*

A. Moradvandi, E. Abraham, A. Goudjil, B. De Schutter, and R.E.F. Lindeboom

If you want to cite this report, please use the following reference instead:

A. Moradvandi, E. Abraham, A. Goudjil, B. De Schutter, and R.E.F. Lindeboom, "An identification algorithm of switched Box-Jenkins systems in the presence of bounded disturbances: An approach for approximating complex biological wastewater treatment models," *Journal of Water Process Engineering*, vol. 60, p. 105202, Apr. 2024. doi:10.1016/j.jwpe.2024.105202

Delft Center for Systems and Control
Delft University of Technology
Mekelweg 2, 2628 CD Delft
The Netherlands
phone: +31-15-278.24.73 (secretary)
URL: https://www.dcsc.tudelft.nl

* This report can also be downloaded via https://pub.bartdeschutter.org/abs/24_008

An Identification Algorithm of Switched Box-Jenkins Systems in the Presence of Bounded Disturbances: An Approach for Approximating Complex Biological Wastewater Treatment Models

Ali Moradvandi[a,b,*], Edo Abraham[a], Abdelhak Goudjil[c], Bart De Schutter[b], Ralph E. F. Lindeboom[a]

[a] *Department of Water Management,Delft University of Technology, Mekelweg 5, 2628 CD, Delft, The Netherlands*
[b] *Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, The Netherlands*
[c] *IPSA Institute of Polytechnic Science and Aeronautics, 63 Bd de Brandebourg Bis, 94200, Ivry-sur-Seine, France*

Abstract

This paper focuses on the development of linear Switched Box–Jenkins (SBJ) models for approximating complex dynamical models of biological wastewater treatment processes. We discuss the adaptation of these processes to the SBJ framework, showing the model's generality and flexibility as a class of switched systems that can offer accurate predictions for complex and nonlinear dynamics. This approach of modeling enables real-time data reconciliation from experiments and allows the design of model-based control strategies previously inaccessible with conventional complex wastewater treatment models. Through the extension of the Outer Bounding Ellipsoids (OBEs) algorithm, the paper introduces an online two-stage parameter identification algorithm that effectively handles bounded disturbances for SBJ models. Using the OBE method relaxes the stochastic assumptions on disturbances, which may not be satisfied in practice, particularly for biological and environmental fluctuations. The proposed decomposed OBE algorithm separately identifies the switching patterns and parameters of linear submodels, conducting parameter identification in two distinct phases for input/output and disturbance/output submodels. The efficacy of this approach is shown via simulation results validated against both ADM1 and PBM, demonstrating the proposed algorithm's capability to accurately predict outputs from different bio-process models.

Keywords: Bioprocess modeling, System identification, Data-driven

*Corresponding author. Email: a.moradvandi@tudelft.nl. Tel: +31(0)152782069 (A. Moradvandi)

modeling, Switched systems, Wastewater treatment modeling

1. Introduction

Hybrid (switched) dynamical systems capture interconnected continuous and discrete behaviors, serving to model processes with non-smooth behaviors or to approximate systems with high-order nonlinearities. Biological treatment processes are described by interconnected and competing bio- and physico-chemical reactions for substrate consumption and growth of different trophic groups within a microbial community, resulting in nonlinear model behaviors. This type of complex nonlinear behavior can be simplified in terms of modeling using hybrid systems. Switched systems, as a well-known class of hybrid systems, consist of a switching pattern (or mode) and a finite number of values (countable state variables) that coordinates with corresponding continuous and linear subsystems (or submodels) (Lauer and Bloch, 2018).

Hybrid system identification methods, as a tool to find a switched system to approximate a highly nonlinear biological treatment model, involve two steps: 1) estimating the parameters of the submodels, and 2) determining the switching patterns. Furthermore, hybrid system identification methods as a data-driven modeling approach avoid the complexity inherent in mechanistic modeling of input-output relations. Moreover, using a set of linear models to approximate a nonlinear dynamic of a biological treatment process not only is straightforward to implement in comparison with Neural Networks but also holds significant accuracy in comparison with non-switched systems.

The input-output model complexity ranges from relatively simple Auto-Regressive eXogenous (ARX) models to more complex general Box-Jenkins (BJ) models. Input-output models consist of two parts, i.e. auto-regressive (depending on the previous forecasts) and moving-average (depending on the error of previous forecasts). Box–Jenkins (BJ) models have the advantage of describing stochastic systems in a more general and flexible way, since they include the output error model (Ding et al., 2010), the output error moving average model (Wang, 2011), and the output error autoregressive model (Wang et al., 2010) as special cases. Moreover, switched finite impulse response, SFIR (Liu et al., 2021) switched autoregressive exogenous, SARX (Du et al., 2018), switched autoregressive and moving-average, SARMAX (Hojjatinia et al., 2020), switched output error, SOE (Goudjil et al., 2017), and error-in-variable SARX, EIV-SARX (Ozbay et al., 2019) models can be mathematically considered as subclasses of a switched Box-Jenkins (SBJ) model. In other words, the mentioned simple model structures can be driven with simplification of a BJ model.

The BJ structure, also, has been widely and effectively used for time series prediction due to its generality and efficiency in prediction (Box et al., 2015). As summarized in Table 1, some biological processes have been modeled by switched systems in the literature. The foundation of the submodels in these papers is ARX. The identification problem has been addressed using different approaches in these articles, including optimization-based methods by Hartmann et al. (2015) and Song et al. (2020), likelihood-based methods by Chen et al. (2020a,b), clustering-based methods by Wang et al. (2020), and Outer Bounding Ellipsoid (OBE) methods by Yahya et al. (2020). Since all these papers deal with ARX models, the identification approaches cannot be directly extended for the more general SBJ models.

In addition to the base model (parametrization), selecting a suitable algorithm for solving the identification problem is an integral part of hybrid system identification that should be developed based on the selected base model (Moradvandi et al., 2023). The approaches are classified into optimization-based techniques (Bianchi et al., 2021), clustering-based methods (Mazzoleni et al., 2021), likelihood-based methods (Chen et al., 2020a), algebraic methods (Hojjatinia et al., 2020), and Outer Bounding Ellipsoid (OBE) methods (Yahya et al., 2020; Goudjil et al., 2023). Comprehensive reviews of these techniques have been done by (Garulli et al., 2012; Moradvandi et al., 2023). The selection of an appropriate method depends on factors such as parametrization, available knowledge of the system, and the computational burden associated with the model. Optimization-based algorithms are the most commonly used, and they have recently been combined with other approaches such as clustering and classical algebraic methods (Wang et al., 2020; Du et al., 2020).

To select an approach, practical aspects of a biological treatment process should also be taken into account. The behavior of a biological process can be affected by random and unpredictable factors. Common examples are meteorological fluctuations and influent concentration perturbations. Under these situations, Piga et al. (2020a) showed that stochastic modeling can be an option. However, the assumption of a statistical consideration for disturbances or noises may not always be justified due to an unknown probability distribution or modeling mismatch (Goudjil et al., 2023). On the other hands, the stochastic assumption requires precise distribution information and employs a sequence of representative scenarios, which is hard to be satisfied in real-world applications. Alternatively, the assumption of bounded disturbances is considered less stringent and therefore a pragmatic solution.

Amongst the mentioned hybrid system identification methods, the OBE method is one of the methods that has the advantage of not requiring any stochastic noise assumption. Furthermore, since the basis of the OBE algorithm is matrix manipulation, the OBE algorithm is not only computationally efficient, but also well-suited for analyzing large datasets (Goudjil et al.,

2023). This method has been developed for hybrid systems parametrized by SARX (Goudjil et al., 2016), SOE (Goudjil et al., 2017), and piecewise affine ARX (PWARX) (Yahya et al., 2020) models, not yet for the general models such as SBJ. The OBE algorithm encompasses two stages: (1) the procedure of assigning data by considering both the residual error and an upper bound for the estimation error of all the submodels, and (2) utilizing Recursive Least Squares (RLS) simultaneously to update the parameters of the active submodel in each time step (Goudjil et al., 2023).

Motivated by the importance of BJ models, particularly for biological treatment processes as well as the current trend of extending other methods for SBJ models (Piga et al., 2020b; Chai et al., 2020), this paper addresses the extension of OBE algorithms to SBJ systems. For this purpose, auxiliary model identification and decomposition techniques, which have been discussed for non-switched systems by Ding and Duan (2013), are adapted to the considered switched structure and the OBE framework. This adaptation deals with lack of availability of internal signals within the BJ structure. Inspired by the work done by Chai et al. (2020), the underlying principle involves the decomposition of a BJ system into two parts (the autoregressive part and the moving-average part), followed by the auxiliary model identification approach to determine the parameters of each part and the internal signals simultaneously. Therefore, a reformulation of the two-stage OBE algorithm based on adaptation of the decomposed technique is addressed in this study, and the active submodel detection and the parameter identification procedures are developed based on a decomposed OBE objective function for SBJ models.

The primary aim of the present work is, therefore, to develop the OBE algorithm for SBJ models. To achieve this objective, we present a mathematical exposition by adapting the decomposition technique to switched systems in order to formulate the identification problem posed by SBJ systems within the OBE framework. Furthermore, the approximation of biological treatment processes represented by complex mathematical models, is explored within the framework SBJ models by validating the proposed algorithm for Anaerobic Digestion Model 1 (ADM1) and Purple Bacteria Model (PBM). Through a comprehensive numerical assessment and interpretation, this research sheds light on the potential applications of the SBJ modeling approach, contributing valuable insights into real-time data reconciliation and control strategies of biological treatment processes.

The paper is organized as follows. Materials and methods (Section 2) include the formulation of the identification problem in Section 2.1 and the OBE identification procedure in Section 2.2. Section 3 presents results and discussions. Formulating of biological models in the form of SBJ is discussed in this section, and the aforementioned case studies of biological wastewater treatment models are also analyzed. Limitations of the proposed method are discussed in Section 4, and in the last section, conclusions are drawn.

2. Materials and methods

2.1. Problem formulation

A switched discrete-time linear system parameterized by a BJ model is represented as follows:

$$y_k = \frac{B(q^{-1}, \theta_{z_k})}{A(q^{-1}, \theta_{z_k})} u_k + \frac{C(q^{-1}, \theta_{z_k})}{D(q^{-1}, \theta_{z_k})} v_k \tag{1}$$

where $y_k \in \mathbb{R}$, $u_k \in \mathbb{R}$, and $v_k \in \mathbb{R}$ denote the system output, the system input, and the disturbance (noise). Moreover, $A(q^{-1}, \theta_{z_k})$, $B(q^{-1}, \theta_{z_k})$, $C(q^{-1}, \theta_{z_k})$, and $D(q^{-1}, \theta_{z_k})$ are the linear filters. The discrete state, $z_k \in \{1, ..., m\}$ indicates the active mode of m number of parameterized submodels or modes at time step k. If we assume at time step k, the i-th mode is active, i.e. $z_k = i$, the linear filters that are rational functions of the time shift operator q^{-1} (i.e. $q^{-d} x_k = x_{k-d}$ for $d \in \mathbb{Z}$), can be written as follows:

$$\frac{B(q^{-1}, \theta_i)}{A(q^{-1}, \theta_i)} = \frac{b_{i1} q^{-1} + ... + b_{in_b} q^{-n_b}}{1 + a_{i1} q^{-1} + ... + a_{in_a} q^{-n_a}}, \tag{2a}$$

$$\frac{C(q^{-1}, \theta_i)}{D(q^{-1}, \theta_i)} = \frac{1 + c_{i1} q^{-1} + ... + c_{in_c} q^{-n_c}}{1 + d_{i1} q^{-1} + ... + d_{in_d} q^{-n_d}}, \tag{2b}$$

where n_a, n_b, n_c, and n_d are the orders of the filters $(A(\cdot), B(\cdot), C(\cdot), D(\cdot))$ respectively, and the vectors of parameters can be expressed as

$$\theta_{1i} = [a_{i1}, ..., a_{in_a}, b_{i1}, ..., b_{in_b}]^T \in \mathbb{R}^{n_a + n_b}, \tag{3a}$$

$$\theta_{2i} = [c_{i1}, ..., c_{in_c}, d_{i1}, ..., d_{in_d}]^T \in \mathbb{R}^{n_c + n_d}, \tag{3b}$$

$$\theta_i = [\theta_{1i}^T, \theta_{2i}^T]^T \in \mathbb{R}^{n_a + n_b + n_c + n_d}, \tag{3c}$$

The block diagram of the switched BJ system is depicted in Fig. 1. According to the block diagram, the two auxiliary variables x_k and w_k can be written as follows:

$$x_k = (1 - A(q^{-1}, \theta_{z_k})) x_k + B(q^{-1}, \theta_{z_k}) u_k = \phi_k^T \theta_{1, z_k}, \tag{4a}$$

$$w_k = (1 - C(q^{-1}, \theta_{z_k})) w_k + D(q^{-1}, \theta_{z_k}) v_k = \psi_k^T \theta_{2, z_k} + v_k, \tag{4b}$$

where ϕ_k and ψ_k are the regressor vectors:

$$\phi_k = [-x_{k-1}, ..., -x_{k-n_a}, u_{k-1}, ..., u_{k-n_b}]^T \in \mathbb{R}^{n_a + n_b}, \tag{5a}$$

$$\psi_k = [-w_{k-1}, ..., -w_{k-n_c}, v_{k-1}, ..., v_{k-n_d}]^T \in \mathbb{R}^{n_c + n_d}. \tag{5b}$$

Therefore, the model (1) can be rewritten as

$$\begin{aligned} y_k &= \phi_k^T \theta_{1, z_k} + \psi_k^T \theta_{2, z_k} + v_k \\ &= \Phi_k^T \theta_{z_k} + v_k. \end{aligned} \tag{6}$$

where $\Phi_k = [\phi_k^T, \psi_k^T]^T \in \mathbb{R}^{n_a+n_b+n_c+n_s d}$,

The decomposition technique is a tool that is used to deal with two-stage identification procedure (Ding and Duan, 2013). In this study, we want to formulate it for switched systems. An intermediate variable is defined as

$$\varpi_k = y_k - \psi_k^T \theta_{2,z_k} \tag{7}$$

and the main system in (6) can be decomposed into two subsystems as follows:

$$\varpi_k = \phi_k^T \theta_{1,z_k} + v_k \tag{8a}$$
$$w_k = \psi_k^T \theta_{2,z_k} + v_k, \tag{8b}$$
$$\varpi_k - \phi_k^T \theta_{1,z_k} = w_k - \psi_k^T \theta_{2,z_k} = v_k, \tag{8c}$$

and they can be rewritten as

$$\varpi_k = y_k - \phi_k^T \theta_{1,z_k} \tag{9a}$$
$$w_k = y_k - \psi_k^T \theta_{2,z_k}, \tag{9b}$$

These decomposed functions will be utilized in the parameter identification stage later on. The identification objective should be defined in order to estimate the discrete state, z_k, and the parameter vectors, $\theta_{z_k}, z_k = 1, ..., m$, given a collection of input and output observations. If the estimations of the discrete state and the parameter vectors are defined as $\hat{z}_k$, $\hat{\theta}_{1,\hat{z}_k}$, and $\hat{\theta}_{2,\hat{z}_k}$, they should satisfy

$$|y_k - \Phi_k^T \hat{\theta}_{\hat{z}_k}| \leq \delta, \forall k \tag{10a}$$
$$|y_k - \phi_k^T \hat{\theta}_{1,\hat{z}_k} - \psi_k^T \hat{\theta}_{2,\hat{z}_k}| \leq \delta, \forall k \tag{10b}$$

where δ is an upper bound of v_k, i.e. $|v_k| \leq \delta, \forall k$. The objective can also be expressed according to (8c). The representation of the objective for the decomposed form of the switched system will be used to derive the parameter identification procedure in the next section.

To apply the OBE algorithm for the defined objective and to derive the estimation procedure of the discrete state, the system represented by (6), should be extended in the following format. If we assume that at time step

k the submodel i is active, then it can be written as

$$\begin{cases} y_k = \phi_k^T\theta_{11} + \psi_k^T\theta_{21} + v_k + \phi_k^T(\theta_{1i} - \theta_{11}) + \psi_k^T(\theta_{2i} - \theta_{21}) \\ y_k = \phi_k^T\theta_{12} + \psi_k^T\theta_{22} + v_k + \phi_k^T(\theta_{1i} - \theta_{12}) + \psi_k^T(\theta_{2i} - \theta_{22}) \\ \quad \vdots \\ y_k = \phi_k^T\theta_{1i} + \psi_k^T\theta_{2i} + v_k \\ \quad \vdots \\ y_k = \phi_k^T\theta_{1m} + \psi_k^T\theta_{2m} + v_k + \phi_k^T(\theta_{1i} - \theta_{1m}) + \psi_k^T(\theta_{2i} - \theta_{2m}) \end{cases} \tag{11}$$

By defining the extended parameter vectors, $\Theta_1 \in \mathbb{R}^{(n_a+n_b)m\times 1}$ and $\Theta_2 \in \mathbb{R}^{(n_c+n_d)m\times 1}$, the extended noise vector, $V_k \in \mathbb{R}^{m\times 1}$, and the extended output vector, $Y_k \in \mathbb{R}^{m\times 1}$, as follows, the system (11) can be rewritten as follows:

$$\Theta_1 = [\theta_{11}, ..., \theta_{1m}]^T \tag{12a}$$

$$\Theta_2 = [\theta_{21}, ..., \theta_{2m}]^T \tag{12b}$$

$$Y_k = [y_k, ..., y_k]^T \tag{12c}$$

$$V_{k,z_k=i} = \begin{bmatrix} v_k + \phi_k^T(\theta_{1i} - \theta_{11}) + \psi_k^T(\theta_{2i} - \theta_{21}) \\ \vdots \\ v_k \\ \vdots \\ v_k + \phi_k^T(\theta_{1i} - \theta_{1m}) + \psi_k^T(\theta_{2i} - \theta_{2m}) \end{bmatrix} \tag{12d}$$

$$Y_k = \overline{\phi}_k^T\Theta_1 + \overline{\psi}_k^T\Theta_2 + V_{k,z_k} \tag{12e}$$

where $\overline{\phi} = I_m \otimes \phi$ and $\overline{\psi} = I_m \otimes \psi$, in which $\otimes$ and I_N denote the Kronecker product and the identity matrix of order m, respectively. If the estimations of z_k, and the parameter vectors, Θ_1 and Θ_2 are denoted by $\hat{z}_k$, $\hat{\Theta}_1$, and $\hat{\Theta}_2$, respectively, (12e) can be rewritten as

$$V_{k,\hat{z}_k} = Y_k - \overline{\phi}_k^T\hat{\Theta}_1 - \overline{\psi}_k^T\hat{\Theta}_2 \tag{13}$$

Therefore, if we define $\nu_k(j)$ as the j-th element of V_{k,z_k}, tanking (13) into account, the problem objective (10b) can be redefined as follows:

$$|\nu_k(\hat{z}_k)| \leq \delta, \forall k \tag{14}$$

where $\hat{z}_k$ can be any integer values between 1 and m at time step k.

2.2. Identification algorithm

The OBE method is a technique used in conventional identification algorithms to estimate the parameters of a model within a given set of constraints, where the feasible region (the set of possible solutions) is bounded.

Using this technique for switched systems allows computing the ellipsoid bounds for all the submodels and finding the active one that fits inside the assigned ellipsoid bound. The proposed identification algorithm is based on two stages, i.e. we first estimate the discrete state (the switching pattern), then the parameter vectors, in a repetitive manner for each time step. The parameter vector estimation is also derived based on the decomposition technique in two stages, i.e. the parameter vector is primarily updated, then we estimate internal signals for next steps. To derive the algorithm, the estimates of the parameter vectors at time step k are denoted by $\hat{\Theta}_{1,k}$ and $\hat{\Theta}_{2,k}$. The a priori and the a posteriori predictors of Y_k can be written w.r.t. (12e), respectively, as

$$\begin{cases} Y_{k/k-1} = \overline{\phi}_k^T \hat{\Theta}_{1,k-1} + \overline{\psi}_k^T \hat{\Theta}_{2,k-1} \\ Y_{k/k} = \overline{\phi}_k^T \hat{\Theta}_{1,k} + \overline{\psi}_k^T \hat{\Theta}_{2,k} \end{cases} \tag{15}$$

Then a priori prediction error can be defined as follows:

$$V_{k/k-1} = Y_k - Y_{k/k-1} = Y_k - \overline{\phi}_k^T \hat{\Theta}_{1,k-1} - \overline{\psi}_k^T \hat{\Theta}_{2,k-1} \tag{16}$$

Therefore, the two-stage OBE algorithm can be described as follows:

Step 1 (estimation of $\hat{z}_k$): The first step estimates the discrete state, i.e. $\hat{z}_k$ based on the smallest element of the vector $V_{k/k-1}$ that can be expressed by $\varrho_k = |\nu_{k/k-1}(\hat{z}_k)|$, in which $\hat{z}_k \in \{1, ..., m\}$ is the detected active mode at time step k.

Step 2 (estimations of $\hat{\Theta}_1$ and $\hat{\Theta}_2$): The second step is to identify the defined parameter vectors, i.e. $\hat{\Theta}_1$ and $\hat{\Theta}_2$. This step is derived based on the decomposition technique. According to the decomposed model written by (8c), the objective functions to derive a Recursive Least Square (RLS) minimization for the decomposed model can be defined as follows:

$$J_1(\theta_{1,z_k}) := \sum_{j=1}^{k} \left(\varpi_k - \phi_k^T \theta_{1,z_k}\right)^2 \tag{17a}$$

$$J_2(\theta_{2,z_k}) := \sum_{j=1}^{k} \left(w_k - \psi_k^T \theta_{2,z_k}\right)^2 \tag{17b}$$

where $J_1 = J_2$ according to (8c). Assuming the i-th mode is active at time step k ($\hat{z}_k = i$), the update laws for the estimates of the parameters, i.e. $\hat{\theta}_{1i}$ and $\hat{\theta}_{2i}$ can be written as a result of the RLS minimization as follows:

$$\hat{\theta}_{1i,k} = \hat{\theta}_{1i,k-1} + L_{1,k}\left[y_k - \psi_k^T \hat{\theta}_{2i,k-1} - \phi_k^T \hat{\theta}_{1i,k-1}\right], \tag{18a}$$

$$\hat{\theta}_{2i,k} = \hat{\theta}_{2i,k-1} + L_{2,k}\left[y_k - \phi_k^T \hat{\theta}_{1i,k-1} - \psi_k^T \hat{\theta}_{2i,k-1}\right], \tag{18b}$$

where

$$L_{1,k} = P_{1,k-1}\phi_k \left[1 + \phi_k^T P_{1,k-1}\phi_k\right]^{-1}, \tag{19a}$$

$$L_{2,k} = P_{2,k-1}\psi_k \left[1 + \psi_k^T P_{2,k-1}\psi_k\right]^{-1}, \tag{19b}$$

and

$$P_{1,k} = \left[I_{n_a+n_b} - L_{1,k}\phi_k^T\right] P_{1,k-1}, \tag{20a}$$

$$P_{2,k} = \left[I_{n_c+n_d} - L_{2,k}\psi_k^T\right] P_{2,k-1}, \tag{20b}$$

Now, the solution of the decomposed RLS formulated above for the i-th mode can be extended for all m number of submodels to be able to apply the OBE algorithm. This is done considering the definitions of Θ_1 and Θ_2 expressed by (12a) and (12b). The extended matrices, $\overline{\phi}$ and $\overline{\psi}$, should also be used as defined by the Kronecker product of an identity matrix of the order m to ϕ and ψ stated in (5a) and (5b). To be able to update only the parameters of the active submodel, a symmetric matrix is defined such that the values of all the elements are zero except for the one element corresponding to the identified active submodel (Goudjil et al., 2023). Because we are using the decomposition technique in this paper, we define two matrices - one for the autoregressive part, $\Upsilon_{1,k} \in \mathbb{R}^{m\times m}$, and the other one for the moving average part, $\Upsilon_{2,k} \in \mathbb{R}^{m\times m}$:

$$\Upsilon_{1,k} = \begin{cases} \left(\overline{\phi}_k^T P_{1,k-1}\overline{\phi}_k\right)^{-1}(\Lambda_k - I_m); \\ \qquad \text{if } \overline{\phi}_k^T P_{1,k-1}\overline{\phi}_k \succ 0 \text{ and } \varrho_k > \delta \\ \mathbf{0}_{m\times m}; \qquad \text{else} \end{cases} \tag{21a}$$

$$\Upsilon_{2,k} = \begin{cases} \left(\overline{\psi}_k^T P_{2,k-1}\overline{\psi}_k\right)^{-1}(\Lambda_k - I_m); \\ \qquad \text{if } \overline{\psi}_k^T P_{2,k-1}\overline{\psi}_k \succ 0 \text{ and } \varrho_k > \delta \\ \mathbf{0}_{m\times m}; \qquad \text{else} \end{cases} \tag{21b}$$

in which $\Lambda_k \in \mathbb{R}^{m\times m}$ denotes the identity matrix at time step k, where the $\hat{z}_k$-th element on the diagonal is $\frac{\varrho_k}{\delta}$. Therefore, the parameters of the active submodel are updated, when the error of the output, ϱ_k, is not within the assigned ellipsoid bound, δ. The update gain is $\frac{\varrho_k}{\delta}$ in matrix Λ_k. On the other side, the adaptation is frozen when $\varrho_k \leq \delta$.

Considering the discussions above, the equations (18a)-(20b) can be reformulated for the extended version as follows:

$$\hat{\Theta}_{1,k} = \hat{\Theta}_{1,k-1} + L_{1,k}\left[Y_k - \overline{\psi}_k^T \hat{\Theta}_{2,k-1} - \overline{\phi}_k^T \hat{\Theta}_{1,k-1}\right] \tag{22a}$$

$$\hat{\Theta}_{2,k} = \hat{\Theta}_{2,k-1} + L_{2,k}\left[Y_k - \overline{\phi}_k^T \hat{\Theta}_{1,k-1} - \overline{\psi}_k^T \hat{\Theta}_{2,k-1}\right] \tag{22b}$$

$$L_{1,k} = \frac{1}{2} P_{1,k-1}\overline{\phi}_k \Upsilon_{1,k}\left[I_m + \overline{\phi}_k^T P_{1,k-1}\overline{\phi}_k \Upsilon_{1,k}\right]^{-1} \tag{22c}$$

$$L_{2,k} = \frac{1}{2} P_{2,k-1}\overline{\psi}_k \Upsilon_{2,k}\left[I_m + \overline{\psi}_k^T P_{2,k-1}\overline{\psi}_k \Upsilon_{2,k}\right]^{-1} \tag{22d}$$

$$P_{1,k} = \left[I_{m\times(n_a+n_b)} - L_{1,k}\overline{\phi}_k^T\right] P_{1,k-1} \tag{22e}$$

$$P_{2,k} = \left[I_{m\times(n_c+n_d)} - L_{2,k}\overline{\psi}_k^T\right] P_{2,k-1} \tag{22f}$$

The introduction of the factor $\frac{1}{2}$ in (22c) and (22d) allows us to prove the objective we defined in (14), which comes later.

Remark 1. *It should be noted that individual update equations (18a) and (18b) are written based on this assumption that the system stays in one mode in two consecutive time instants $k-1$ and k. After the extension and defining $\Upsilon_{1,k}$ and $\Upsilon_{2,k}$, it is not the case for the extended update equations (22a) and (22b), since if the mode is changed from $k-1$ to k, the corresponding elements on diagonal of matrices $\Upsilon_{1,k}$ and $\Upsilon_{2,k}$ are also changed to the associated active mode to be updated at time step k and other submodels remain frozen for the update process until they are detected active and the procedure continues.*

The inner variables x and w and the variable v within the definition of ϕ_k, (5a), and ψ_k, (5b) and their extended corresponding matrices, i.e. $\overline{\phi}_k$ and $\overline{\psi}_k$ are unknown, which the estimates of these variables, i.e. $\hat{x}$, $\hat{w}$, and $\hat{v}$ can be replaced (Ding and Duan, 2013) as follows:

$$\hat{x}_k = X_k(\hat{z}_k) \tag{23a}$$

$$\hat{w}_k = y_k - \hat{x} \tag{23b}$$

$$\hat{v}_k = \hat{w}_k - W_k(\hat{z}_k) \tag{23c}$$

where $X_k = \overline{\phi}_k^T \hat{\Theta}_{1,k}$ and $W_k = \overline{\psi}_k^T \hat{\Theta}_{2,k}$ are the estimates of the unknown signals for all the submodels. If we assume the detected active submodel at time step k is i, i.e. $\hat{z}_k = i$, the i-th element of the vectors X_k and W_k should be used for the calculation of $\hat{x}_k$ and $\hat{v}_k$, respectively, as stated in (23a) and (23c). Considering the explained procedure, the two-stage decomposed OBE algorithm can be summarized in Algorithm .

Remark 2. *It can be shown that the objective defined in (14) is satisfied at each time step by implementing the proposed two-stage decomposed OBE*

Algorithm Two-stage decomposed OBE algorithm

Initialize: $P_{1,0} = p_0 I_{m\times(n_a+n_b)}$, $P_{2,0} = p_0 I_{m\times(n_c+n_d)}$,
  $\hat{\Theta}_{1,0}$ and $\hat{\Theta}_{2,0}$ randomly initialized,
  $\hat{x}_k = \hat{w}_k = \hat{v}_k = 0\ \forall k \leq 0$
for $k = 1$ **do**
  step 1: detect the active submodel $\hat{z}_k$
    Receive u_k and y_k
    Form $\overline{\phi}_k = \phi_k \otimes I_m$ and $\overline{\psi}_k = \psi_k \otimes I_m$
      based on (5a) and (5b)
    Compute $\nu_{k/k-1}$ as (16)
    Compute $\hat{z}_k = \underset{j=1,\ldots,m}{\arg\min} |\nu_{k/k-1}(j)|$
    Compute $\varrho_k = |\nu_{k/k-1}(\hat{z}_k)|$
  step 2: estimate the parameters vectors $\hat{\Theta}_{1,k}$ and $\hat{\Theta}_{2,k}$
    Compute $\Upsilon_{1,k}$ and $\Upsilon_{2,k}$ as (21a) and (21b)
    Compute $L_{1,k}$, $L_{2,k}$, $P_{1,k}$, and $P_{2,k}$
      as (22c)-(22f)
    Update $\hat{\Theta}_{1,k}$ and $\hat{\Theta}_{2,k}$ as (22a) and (22b)
    Compute $\hat{x}_k$, $\hat{w}_k$, and $\hat{v}_k$ as (23a)-(23c)
  $\boldsymbol{k = k + 1}$
end for

algorithm. The a posteriori prediction error, i.e. $V_{k/k}$, can be written according to (16) as follows:

$$V_{k/k} = V_{k/k-1} - (\overline{\phi}_k^T L_{1,k} + \overline{\psi}_k^T L_{2,k}) V_{k/k-1} \tag{24}$$

Using the definitions of $L_{1,k}$ and $L_{2,k}$ as stated in (22c) and (22d) in (24) yields

$$\begin{aligned} V_{k/k} = V_{k/k-1}(I_m - \frac{1}{2}\overline{\phi}_k^T P_{1,k-1}\overline{\phi}_k \Upsilon_{1,k}[I_m + \overline{\phi}_k^T P_{1,k-1}\overline{\phi}_k \Upsilon_{1,k}]^{-1} \\ - \frac{1}{2}\overline{\psi}_k^T P_{2,k-1}\overline{\psi}_k \Upsilon_{2,k}[I_N + \overline{\psi}_k^T P_{2,k-1}\overline{\psi}_k \Upsilon_{2,k}]^{-1}) \end{aligned} \tag{25}$$

If the persistent excitation conditions (Ljung, 1999) are satisfied, i.e. $\overline{\phi}_k^T P_{1,k-1}\overline{\phi}_k \succ 0$, and $\overline{\psi}_k^T P_{2,k-1}\overline{\psi}_k \succ 0$, according to the expressions of $\Upsilon_{1,k}$ and $\Upsilon_{2,k}$ stated by (21a) and (21b), we have

- *either $\varrho_k \leq \delta$: $\Upsilon_{1,k}$ and $\Upsilon_{2,k}$ become zero and (25) can be rewritten element-wise as follows for the detected active submodel:*

$$|\nu_{k/k}(\hat{z}_k)| = |\nu_{k/k-1}(\hat{z}_k)| \tag{26}$$

  which yields

$$|\nu_{k/k}(\hat{z}_k)| \leq \delta \tag{27}$$

- *or $\varrho_k > \delta$: by substituting $\Upsilon_{1,k}$ and $\Upsilon_{2,k}$ in (25) yields*

$$V_{k/k} = V_{k/k-1}\left(I_m - \frac{1}{2}\left(I_m - \Lambda_k^{-1}\right) - \frac{1}{2}\left(I_m - \Lambda_k^{-1}\right)\right) \tag{28}$$

which can be rewritten element-wise for the detected active submodel as follows:

$$\nu_{k/k}(\hat{z}_k) = \Lambda_k^{-1}(\hat{z}_k)\nu_{k/k-1}(\hat{z}_k) \tag{29}$$

where $\Lambda_k^{-1}(\hat{z}_k)$ denotes $\hat{z}_k$-th element of matrix Λ_k^{-1} and since $|\Lambda_k^{-1}(\hat{z}_k)| = \frac{\delta}{\varrho_k}$, it gives

$$|\nu_{k/k}(\hat{z}_k)| = \delta \tag{30}$$

Therefore, considering the two cases that can happen at each time step and according to (27) and (30), (14) is proved.

3. Results and discussions

3.1. Numerical example

A numerical example is considered to assess the accuracy of the prediction using the proposed identification algorithm. The dynamics of this example as a two-mode SBJ system are provided in Table 2. To satisfy the persistent excitation, the input sequence is generated randomly within the range of $[-1, 1]$. The lower and upper bounds of the noise sequence are considered -0.08 and 0.08, respectively. Therefore, δ as the upper bound of the noise can be taken any value as larger as 0.08, which it is set to 0.1 in this example. To reach and stay within the assigned bounds, 500 samples of the system are produced and given to the proposed algorithm for the purpose of prediction. The results are plotted for the last 100 samples. As depicted in Fig. 2 (a), the estimated output is capable to track the real output within the specified range. Fig. 2 shows the prediction output and errors and the detection of the switching time instants. Switching instants have been also detected accurately, except at a few steps. To assess the performance of the algorithm, the *FIT* index is considered, which is the percentage fitting error between the true output, y, and the estimated output, $\hat{y}$, which is 95.2 for the last 100 samples and 88.4 for all the samples.

Remark 3. *A few factors can impact the performance and the accuracy of the proposed algorithm. The value of δ that comes from the main constraint of the objective, is one of the major parameters. If it is chosen close to the bound of the system noise, it can numerically destabilize the prediction, while by selecting it too big, the accuracy is deteriorated. The other important factor is the forgetting procedure. The forgetting procedure is used to reduce the weight of past data and to avoid the matrices $P_{1,k}$ and $P_{2,k}$ from approaching zero, as this can affect the accuracy. Therefore, resetting the parameters $P_{1,k}$ and $P_{2,k}$ in a periodic time interval can affect the accuracy of the prediction, which should be taken into account.*

3.2. Biological wastewater treatment processes

A key question in modeling of biological wastewater treatment processes is which modeling approach to choose. Using first principal knowledge to mechanistically drive a model is one of the common and well-known approach. Mechanistic models rely on chemical and biochemical insights and experimental studies, yet they can suffer model mismatch due to potential inaccuracies, occasional perturbations, and varying operational scenarios. Input-output modeling enables an alternative, since it is a data-driven approach. These models can be utilized as prediction models of model-based control systems like model-predictive control, even with the lack of poor interpretability in some cases.

Within input-output modeling approaches, switched system identification is worth exploring, particularly for approximating (highly-)nonlinear complex biological processes. As discussed in the introduction, a few limited real-world applications have been modeled by using simple switched system structures like SARX. Therefore, in this study, we open up a new window for further exploration of input-output switched system identification for the purpose of predictive modeling of biological treatment processes.

For approximating a complex process in the form of input-output models, a critical question arises:"how do we select influential inputs and their corresponding influenced outputs?" Upon this selection, inputs can be categorized as main inputs and disturbances. Taking (1) into account, main inputs are denoted as u, and disturbances as v. By identifying parameters related to their dynamics, represented by $A(.)$, $B(.)$, $C(.)$, and $D(.)$, the relationship between outputs and inputs/disturbances is modeled in a data-driven framework. This paper sheds light on applications to be modeled using general SBJ models by illustrating this via two examples. Depending on the application, some simple structures would suffice for modeling of the process (Hartmann et al., 2015; Yahya et al., 2020; Wang et al., 2020; Chen et al., 2020a). For other cases, more complex structures may be needed.

In this section, we explore the implementation of the proposed prediction method through two wastewater treatment processes; anaerobic fermentation in a continuous stirred-tank reactor (CSTR) and microbial growth of purple phototrophic bacteria (PPB) in a raceway-pond reactor acting as sequencing batch reactor (SBR). Anaerobic fermentation in CSTR is chosen to discuss the importance of using a SBJ model for such a complex bioprocess widely-used in various operational scenarios. PPB biomass cultivation in an SBR is also selected not only because of dynamic complexity, but also for assessment of a potential application of the proposed algorithm in sequencing batch conditions. Moreover, the coupled anaerobic fermentation and purple bacteria raceway-pond reactors for the growth of PPB biomass is a resource recovery process, which has been designed as a pilot plant in *SARASWATI2.0* project.

Anaerobic fermentation in CSTR: Anaerobic digestion is a multistage complex biological process for converting biodegradable organic matter into biogas through volatile fatty acid (VFA) intermediates in the absence of oxygen (Batstone et al., 2002b; Anukam et al., 2019). This process can be represented by comprehensive mechanistic models such as ADM1, with a high-degree nonlinearity and stiffness (Batstone et al., 2002a). The model, however, is bio- and physio- chemical-structured for the purposes of *process design and understanding*, but it is computationally expensive to use for the purposes of *predictive models* (Kil et al., 2017; Ghanavati et al., 2021). Its differential-algebraic equation sets consist of time-varying parameters, multiple variables with intricate interconnections, monod-type kinetics, inhibition functions, and competitive uptakes, which are the reasons for the nonlinear behavior. Furthermore, significant fluctuations in both inflow and the composition of incoming wastewater, that do reflect real-world behaviors, perturb both liquid and gas phases characteristics. Input-output system identification for such a typical nonlinear biological model in the framework of switched systems and BJ structure is worth investigating, and as far as authors are aware is reported in literature for the first time in this study.

It is challenging to select input and output variables of the process. As mentioned, output variables can be a function of different variables. As an example, the output to be predicted is chosen *acetate* as the process is fermentation and *acetate* is expected to be the main product of the anaerobic fermentation process. Moreover, prediction of *acetate* is worth considering due to its critical role, especially when the anaerobic digestion is designed for operation in a wider range (Wainaina et al., 2019). From a practical point of view, the most influential while easily being manipulating input on production of VFAs is the input flow rate. The flow rate affects the hydraulic retention time, and is one of the most feasible manipulators in terms of process control in practice. However, as mentioned earlier, producing *acetate* does not depend only on inflow. Considering the mechanistic equation describing the dynamic of *acetate* in the ADM1 model (Batstone et al., 2002b), its function can be expressed as follows:

$$S_{\text{acetate}} = f(q, X_{\text{lipid}}, X_{\text{protein}}, X_{\text{charbohydrate}}, S_{\text{suger}}, S_{\text{amino acid}}, S_{\text{fatty acid}}, ...) \tag{31}$$

where S_i and X_i stand for soluble and particulate concentrations of material i, respectively, q denotes and inflow rate. The composition of the influent is considered as disturbance to the process. In practice, the process is usually designed around a specific operating point by monitoring various bioreactor operating parameters. However, perturbations like sudden influent concentration changes may happen any time during operation, playing a role as a disturbance. Therefore, the input-output relations can be represented by a BJ model. It means that disturbances can be integrated in modeling

with independent dynamics, which is biologically explainable due to different mechanistic effects between the input and the disturbance to the output. The dynamic between the input flow rate and *acetate* is completely different from the dynamic between other variables and *acetate* as described in the ADM1 model (Batstone et al., 2002b). Therefore, considering the schematic of a BJ structure as depicted in Fig. 1, dynamics of the disturbance is not the same as dynamics for input.

The three main components i.e. carbohydrate, protein, and lipid represents the influent characteristics, which can be considered as the disturbance. They highly impact the process output and are the potential perturbations due to lack of online measurement. Now, the schematic of the process can be drawn in Fig. 3. The nominal operating condition as given in (Batstone et al., 2002a) are considered to generate the dataset, while the reactor environment (the initial conditions) is considered to be acidified at the start-up phase. To explore a wide domain of operation, the process is excited by the input flow rate produced by a pseudo random input signal depicted in Fig. 4 (a). The nominal values for carbohydrate, protein, and lipid are 5, 20, and 5 kgCODm^{-3}, respectively, while for fluctuation purposes, a random deviation from the nominal values in a range of $[-0.5, 0.5]$ is assumed. Therefore, the process output deviates from its designated nominal value, as shown in Fig. 4 (b).

Considering the modeling structure explained above, the proposed algorithm is implemented to identify a parametric SBJ model, given the dataset generated from complex ADM1 model. A few design parameters, therefore, should be assigned. It should be noted that the process is not hybrid by its intrinsic nature and the algorithm is used to capture the dynamics within the designed operating space by a set of linear systems for simplicity for the purpose of prediction, not interpretation. The orders of the SBJ system, therefore, are assigned as *one* for all n_a, n_b, n_c, and n_d. While the higher order may result in higher accuracy, but no amelioration is observed when the complexity is increased. The bound of the disturbance, δ, should be set equal to or bigger than 0.05 due to the assigned range for the disturbance. The process dynamics can be captured accurately ($FIT \simeq 95$) by adjusting the two major design parameters for different number of submodels. It is highlighted in *Remark 1* that the value of δ and the forgetting factor play important role for the numerical stability as well as the output accuracy. The effects of these aforementioned parameters on prediction accuracy are investigated in Table 3.

A comparison with the conventional two-stage BJ system identification (Ding and Duan, 2013) is also made to explore the priority of using a SJB system instead of a non-switched system. The system orders are chosen the same for the both conventional BJ and SBJ systems. The number of modes and the ellipsoid bound for the SBJ system are assigned to 4 and 0.05, respectively. The initial values and other required parameters are set

similarly. For the forgetting factor, a period of 60 days is chosen for this particular application. This setting suffices the need for accurate prediction with the desire for a reasonable rate of convergence. Generally, the proposed SBJ system identification algorithm outperforms the conventional BJ system identification method. The accuracy of the identified SBJ model is better during the whole of the operation and particularly the start-up as shown in Fig. 5 (a). The OBE algorithm forces the system to stay within the assigned bound by jumping to other mode, while the conventional BJ system cannot keep the output error in the range accurately. As can be seen in Figs 5 (a) and (b), the spikes occur, when the direction of the response output is changed, which can be compensated by going to the other submodels in the SBJ system to keep the accuracy within the assigned bound.

Remark 4. *The anaerobic digestion process is not hybrid by its nature, but a highly nonlinear system. Approximation of the dynamics by using a SBJ model with the OBE algorithm has an advantage of capturing input-output relations with a limited number of linear submodels jumping among each other with a desired bound of accuracy in terms of prediction error. Moreover, the other advantage of using BJ structure is identifying different parameters for the moving average part, which is explainable because of different dynamical function of disturbance to output from mechanistic modeling point of view. Furthermore, the type of disturbance as it comes from a nonlinear dynamics in the real system cannot be fitted easily to the conventional stochastic assumption that is relaxed by proposing the developed OBE algorithm.*

Growth of PPB biomass in an SBR: Purple phototrophic bacteria (PPB) as a group of microbes for resource recovery from wastewater can be cultivated by cost-effective raceway-pond bioreactors (Alloul et al., 2023a). A mechanistic model for PPB in raceway bioreactors has been proposed by Alloul et al. (2023b), known as the Purple Bacteria Model (PBM). This type of bioprocesses, i.e. sequencing batch, is selected to assess modeling in the SBJ framework with the proposed OBE algorithm. The cyclic nature of sequencing batch bioreactor operation is regularly applied in conventional wastewater treatment, like for example in aerobic granular sludge technology.

Besides hydraulic and sludge retention times, light also plays a critical role in growth of PPB. In a raceway-pond bioreactor, control over light, more specifically solar radiation, is not practically feasible, due to various hour-by-hour, day-by-day, and seasonal fluctuations. It should be, therefore, considered as a potential disturbance, especially for modeling of an open reactor. Furthermore, distribution of solar radiation is barely representable by the common distribution functions. For instance, illumination durations and radiation angles at a single day are not independent of subsequent days, which may violate the independence assumption required for

probability distributions. It is, therefore, another motivation to employ the OBE algorithm for approximation of the process dynamics, since it is not subject to any assumptions for disturbances.

The dynamics of PPB in raceway reactors are also highly nonlinear (Alloul et al., 2023b). If the production of PPB is selected as an output to be predicted, flow rate that determines feeding of each sequence is considered as input, while solar irradiation fluctuation that deviates the process from the nominal operating is considered as disturbance. The schematic of an SBJ structure is depicted in Fig. 6. Considering the mechanistic model proposed by Alloul et al. (2023b), PPB production is the function of a wide range of variables with different dynamics. Therefore, defining the problem of approximating this bioreactor in the frame of BJ model is reasonable, due to different dynamics for the input and the disturbance.

To run the PBM model, the following conditions are considered; the sequential batch is designed to feed the reactor once a day at the midnight; influent filling and the effluent extraction are set at midnight, while feeding rate is set to one fourth of the volume per hydraulic retention time; the paddlewheel is considered working only during the light condition. Other operational parameters are set to the default values of the PBM (Alloul et al., 2023b).

The solar radiation is subject to fluctuation. Light intensity is depicted in Fig. 7 (a) from day 21 to 42, when the process reaches steady state. It can be observed, finding a probability distribution is subject to some simplifications that may not be reliable. Therefore, the OBE algorithm that is not subject to probability of disturbance is practically and theoretically more reasonable.

The deviation from nominal process operation with light variation as a disturbance to the operation is shown in Fig. 7 (b) between day 21 to 42. The output to be predicted is considered purple bacteria produced from the three photoheterotrophic, anaerobic and aerobic chemoheterotrophic pathways. The proposed algorithm is implemented, given the dataset produced. Since the effect of ellipsoid bound and number of modes were investigated in the previous case study, and the same results were observed, the detected switching patterns and its interpretations are explored in this case study.

The orders of the estimated SBJ system are assigned as *one* for all n_a, n_b, n_c, and n_d. The bound of the ellipsoid, δ, the number of modes, and the forgetting period are set to 0.25, 2, and 60 h, respectively and the process behavior is acceptably approximated as depicted in Fig. 8. Moreover, the switching patterns are shown in Fig. 8 (b). As can be seen, the time of being in mode one is much longer than mode two. If only the subsystem one is active for prediction, the ellipsoid bound constraint is violated, as shown in Fig. 8; sub-figures (a) and (b). In other words, using second mode assists the prediction process to stay within the bound.

Remark 5. *Instants of jumping can be explained based on process operating conditions that they occurred around time of extraction, when the light goes off. As described above, biomass removal happens every* 24 h*, and it is replaced by new influent. PPB are produced photoheterotrophically, aerobic and anaerobic chemoheterotrophiccally. As the reactor is an open system, the amount of PPB grown anaerobic chemoheterotrophiccally is negligible, while photoheterotrophic growth is the major metabolic growth pathway of PPB, which steadily increases when exposed to solar radiation and decreases when no light is available. A sudden decrease happens on the time extraction, and it is also affected negatively because of the absence of light availability. Therefore, the algorithm needs to switch to keep the accuracy within the assigned bound. In other words, this biomass withdrawal is behaving like a hybrid feature in this example that the algorithm is capable of capturing it.*

4. Limitations of the proposed approach and further work

This paper illustrates how SBJ models can be formulated for biological wastewater treatment process models by analyzing two ADM1 and PBM models. Depending on the application, some simple structures would suffice for process modeling (Hartmann et al., 2015; Yahya et al., 2020; Wang et al., 2020; Chen et al., 2020a). For other cases, more complex structures like SBJ may be more meaningful, as different dynamics could be fitted to represent the relation between disturbances and outputs.

The identification algorithm used does not require an assumption on statistical distribution for disturbances, and only has the less strict assumption that they are bounded. Nonetheless, the proposed method is built upon an approach that needs a few design parameters influencing the accuracy of prediction. These parameters discussed in Remark 3 can be determined through trial and error simulations. Moreover, preprocessing of a dataset for some cases may be required to avoid numerical issues.

As a future research, the algorithm can be extended for processes that require a multiple inputs and multiple outputs system representation. Parametrizing the switching domain in the form of polyhedral partitions for better interpretation of switching behavior may also be considered as another extension, specially for biological wastewater treatment processes

5. Conclusion

In this paper, the application of switched Box-Jenkins systems is investigated in the context of modeling biological treatment processes, using two widely-utilized complex models for benchmarking model performance, i.e. ADM1 and PBM. An identification method is introduced by extending the OBE identification algorithm for switched Box-Jenkins models. The algorithm builds upon the standard OBE approach as its foundation, eliminating

the need for the assumption that a probability distribution of disturbances exists and relying solely on the assumption of bounded disturbances. This feature is particularly valuable in practical scenarios of treatment processes, where such distributions might not even be available due to unpredictable fluctuations. To tackle the mathematical challenges arising from the SBJ structure and its inner signals, we employ a decomposition technique. The resulting algorithm is recursive, enabling real-time data processing. This attribute makes it well-suited for systems dealing with extensive data volumes. The results underscore the algorithm's capacity to yield accurate predictions, thereby highlighting its potential for real-world implementation for biological treatment processes.

6. CRediT authorship contribution statement

Ali Moradvandi: Conceptualization, Methodology, Software, Validation, Writing - Original Draft, **Edo Abraham:** Supervision, Funding acquisition, Writing - Review & Editing, **Abdelhak Goudjil:** Methodology, Software, Writing - Review & Editing, **Bart De Schutter:** Supervision, Funding acquisition, Writing - Review & Editing, **Ralph E. F. Lindeboom:** Supervision, Project administration, Funding acquisition, Writing - Review & Editing.

Acknowledgments

The authors kindly acknowledge a research grant of the European Horizon 2020 Framework Programme (grant agreement number 821427) under the Project SARASWATI2.0.

References

Alloul, A., Blansaer, N., Segura, P.C., Wattiez, R., Vlaeminck, S.E., Leroy, B., 2023a. Dehazing redox homeostasis to foster purple bacteria biotechnology. Trends in Biotechnology 41, 106–119. doi:10.1016/j.tibtech.2022.06.010.

Alloul, A., Moradvandi, A., Daniel Puyol, R.M., Gardella, G., Vlaeminck, S.E., Schutter, B.D., Abraham, E., Lindeboom, R.E.F., Weissbrodt, D.G., 2023b. A novel mechanistic modelling approach for microbial selection dynamics: towards improved design and control of raceway reactors for purple bacteria. Submitted to Bioresource Technology doi:10.5281/zenodo.8277031.

Anukam, A., Mohammadi, A., Naqvi, M., Granström, K., 2019. A review of the chemistry of anaerobic digestion: Methods of accelerating and optimizing process efficiency. Processes 7, 504. doi:10.3390/pr7080504.

Batstone, D., Keller, J., Angelidaki, I., Kalyuzhnyi, S., Pavlostathis, S., Rozzi, A., Sanders, W., Siegrist, H., Vavilin, V., 2002a. The IWA anaerobic digestion model no 1 (ADM1). Water Science and Technology 45, 65–73.

Batstone, D.J., Keller, J., Angelidaki, I., Kalyuzhnyi, S., Pavlostathis, S., Rozzi, A., Sanders, W., Siegrist, H.a., Vavilin, V., 2002b. The iwa anaerobic digestion model no 1 (adm1). Water Science and technology 45, 65–73.

Bianchi, F., Breschi, V., Piga, D., Piroddi, L., 2021. Model structure selection for switched NARX system identification: A randomized approach. Automatica 125, 109415. doi:10.1016/j.automatica.2020.109415.

Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time series analysis: forecasting and control. John Wiley & Sons.

Chai, X., Wang, H., Ji, X., Wang, L., 2020. Identification of switched linear systems based on expectation-maximization and bayesian algorithms. Transactions of the Institute of Measurement and Control 43, 412–420. doi:10.1177/0142331220960249.

Chen, X., Zhao, S., Liu, F., 2020a. Identification of jump Markov autoregressive exogenous systems with missing measurements. Journal of the Franklin Institute 357, 3498–3523. doi:10.1016/j.jfranklin.2019.03.041.

Chen, X., Zhao, S., Liu, F., 2020b. Online identification of time-delay jump Markov autoregressive exogenous systems with recursive expectation-maximization algorithm. International Journal of Adaptive Control and Signal Processing 34, 407–426. doi:10.1002/acs.3089.

Ding, F., Duan, H., 2013. Two-stage parameter estimation algorithms for box–jenkins systems. IET Signal Processing 7, 646–654. doi:10.1049/iet-spr.2012.0183.

Ding, F., Liu, P.X., Liu, G., 2010. Gradient based and least-squares based iterative identification methods for OE and OEMA systems. Digital Signal Processing 20, 664–677. doi:10.1016/j.dsp.2009.10.012.

Du, Y., Liu, F., Qiu, J., Buss, M., 2020. A semi-supervised learning approach for identification of piecewise affine systems. IEEE Transactions on Circuits and Systems I: Regular Papers 67, 3521–3532. doi:10.1109/TCSI.2020.2991645.

Du, Z., Balzano, L., Ozay, N., 2018. A robust algorithm for online switched system identification. IFAC-PapersOnLine 51, 293–298. doi:10.1016/j.ifacol.2018.09.150.

Garulli, A., Paoletti, S., Vicino, A., 2012. A survey on switched and piecewise affine system identification. IFAC Proceedings Volumes 45, 344–355. doi:10.3182/20120711-3-BE-2027.00332.

Ghanavati, M.A., Vafa, E., Shahrokhi, M., 2021. Control of an anaerobic bioreactor using a fuzzy supervisory controller. Journal of Process Control 103, 87–99. doi:10.1016/j.jprocont.2021.05.010.

Goudjil, A., Pouliquen, M., Pigeon, E., Gehan, O., 2016. A real-time identification algorithm for switched linear systems with bounded noise, in: 2016 European Control Conference, pp. 2626–2631. doi:10.1109/ECC.2016.7810686.

Goudjil, A., Pouliquen, M., Pigeon, E., Gehan, O., 2023. On-line outer bounding ellipsoid algorithm for clustering of hyperplanes in the presence of bounded noise. Cluster Computing doi:10.1007/s10586-023-03978-z.

Goudjil, A., Pouliquen, M., Pigeon, E., Gehan, O., Targui, B., 2017. Recursive output error identification algorithm for switched linear systems with bounded noise. IFAC-PapersOnLine 50, 14112–14117. doi:10.1109/CDC40024.2019.9029570.

Hartmann, A., Lemos, J.M., Costa, R.S., Xavier, J., Vinga, S., 2015. Identification of switched ARX models via convex optimization and expectation maximization. Journal of Process Control 28, 9–16. doi:10.1016/j.jprocont.2015.02.003.

Hojjatinia, S., Lagoa, C.M., Dabbene, F., 2020. Identification of switched autoregressive exogenous systems from large noisy datasets. International Journal of Robust and Nonlinear Control 30, 5777–5801. doi:10.1002/rnc.4968.

Kil, H., Li, D., Xi, Y., Li, J., 2017. Model predictive control with on-line model identification for anaerobic digestion processes. Biochemical Engineering Journal 128, 63–75. doi:10.1016/j.bej.2017.08.004.

Lauer, F., Bloch, G., 2018. Hybrid system identification: Theory and algorithms for learning switching models. Cham, Switzerland: Springer 478.

Liu, X., Yang, X., Yu, M., 2021. Identification of switched FIR systems with random missing outputs: A variational Bayesian approach. Journal of the Franklin Institute 358, 1136–1151. doi:10.1016/j.jfranklin.2020.10.046.

Ljung, L., 1999. System Identification - Theory for the User. Prentice-Hall.

Mazzoleni, M., Breschi, V., Formentin, S., 2021. Piecewise nonlinear regression with data augmentation. IFAC-PapersOnLine 54, 421–426. doi:10.1016/j.ifacol.2021.08.396.

Moradvandi, A., Lindeboom, R.E.F., Abraham, E., Schutter, B.D., 2023. Models and methods for hybrid system identification: a systematic survey. Accepted by IFAC-PapersOnLine .

Ozbay, B., Camps, O., Sznaier, M., 2019. Efficient identification of error-in-variables switched systems via a sum-of-squares polynomial based subspace clustering method, in: 2019 IEEE 58th Conference on Decision and Control, pp. 3429–3434. doi:10.1109/CDC40024.2019.9029570.

Piga, D., Bemporad, A., Benavoli, A., 2020a. Rao-blackwellized sampling for batch and recursive Bayesian inference of piecewise affine models. Automatica 117, 109002. doi:10.1016/j.automatica.2020.109002.

Piga, D., Breschi, V., Bemporad, A., 2020b. Estimation of jump box-jenkins models. Automatica 120, 109126. doi:10.1016/j.automatica.2020.109126.

Song, C., Wang, J., Ma, X., Zhao, J., 2020. A PWA model identification method based on optimal operating region partition with the output-error minimization for nonlinear systems. Journal of Process Control 88, 1–9. doi:10.1016/j.jprocont.2020.01.011.

Wainaina, S., Lukitawesa, Awasthi, M.K., Taherzadeh, M.J., 2019. Bioengineering of anaerobic digestion for volatile fatty acids, hydrogen or methane production: A critical review. Bioengineered 10, 437–458. doi:10.1080/21655979.2019.1673937.

Wang, D., 2011. Least squares-based recursive and iterative estimation for output error moving average systems using data filtering. IET Control Theory & Applications 5, 1648–1657. doi:10.1049/iet-cta.2010.0416.

Wang, D., Chu, Y., Yang, G., Ding, F., 2010. Auxiliary model based recursive generalized least squares parameter estimation for hammerstein OEAR systems. Mathematical and Computer Modelling 52, 309–317. doi:10.1016/j.mcm.2010.03.002.

Wang, J., Song, C., Zhao, J., Xu, Z., 2020. A PWA model identification method for nonlinear systems using hierarchical clustering based on the gap metric. Computers and Chemical Engineering 138, 106838. doi:10.1016/j.compchemeng.2020.106838.

Yahya, O., Lassoued, Z., Abderrahim, K., 2020. Identification of PWARX model based on outer bounding ellipsoid algorithm, in: 2020 20th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering, pp. 173–178. doi:10.1109/STA50679.2020.9329313.

Table 1: Applications of biological processes modeled by different hybrid systems.

Application	Hybrid model	Method	Reference
pH neutralization process	PWARX	Clustering-based	Wang et al. (2020)
Diauxic bacterial growth	SARX	Optimization-based	Hartmann et al. (2015)
CSTR with exothermic reaction	PWARX	Clustering-based	Song et al. (2020)
Continuous fermentation reactor	SARX	Likelihood-based	Chen et al. (2020a)
	Delay-SARX	Likelihood-based	Chen et al. (2020b)
Transesterification reactor	PWARX	OBE	Yahya et al. (2020)

Table 2: Dynamics of the numerical example; a two-mode SBJ system.

Subsystem dynamics	Subsystem 1	Subsystem 2
$A(q^{-1}, \theta_{1 \text{ or } 2})$	$1 + 0.45q^{-1} - 0.2q^{-2}$	$1 - 0.15q^{-1} + 0.35q^{-2}$
$B(q^{-1}, \theta_{1 \text{ or } 2})$	$-0.4 + 0.95q^{-1}$	$-0.5 + 1.15q^{-1}$
$C(q^{-1}, \theta_{1 \text{ or } 2})$	$1 + 0.64q^{-1}$	$1 - 0.36q^{-1}$
$D(q^{-1}, \theta_{1 \text{ or } 2})$	$1 - 0.32q^{-1}$	$1 - 0.50q^{-1}$

Table 3: Prediction accuracy of anaerobic fermentation process under different scenarios based on the proposed output prediction algorithm.

Number of modes (N)	Ellipsoid bound (δ)	Period of forgetting factor (day)	Accuracy (FIT)
2	0.2	40	94.9157
3	0.1	50	95.7483
4	0.05	60	96.7509
5	0.05	50	97.1826

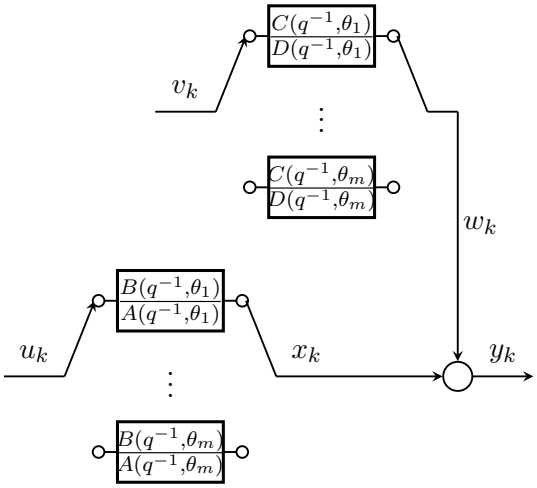
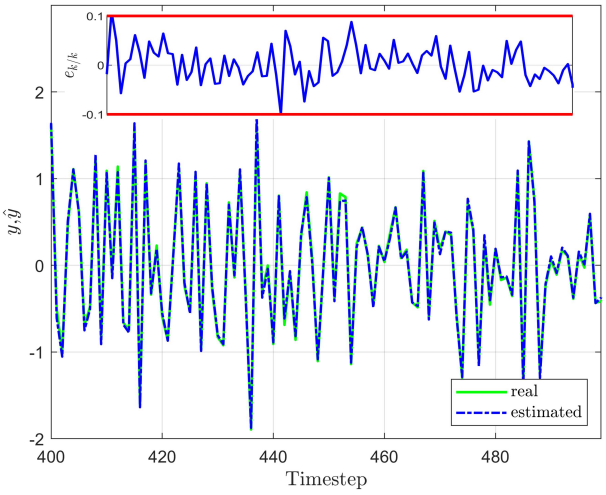
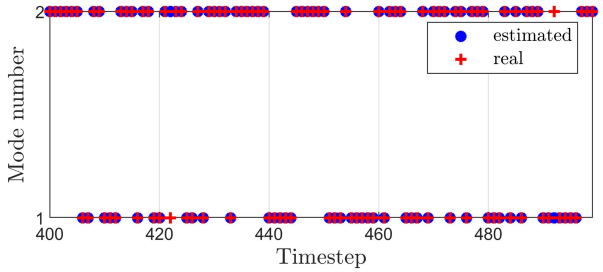


Figure 1: Schematization of the switched BJ system for m number of mode.



(a) the real system output, y, and the predicted system output, $\hat{y}$ based on the estimated SBJ system. The inner figure shows the posteriori prediction error



(b) Detection of the switching sequences of the SBJ.

Figure 2: Numerical example simulation.

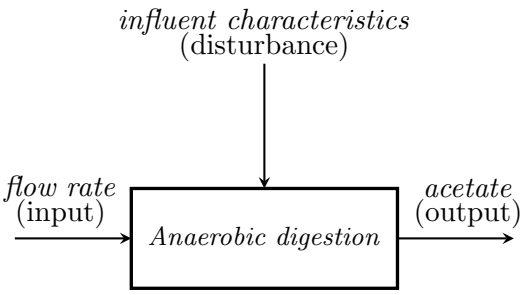
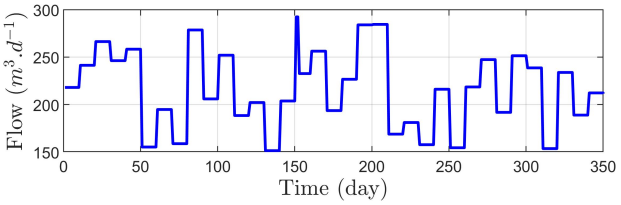
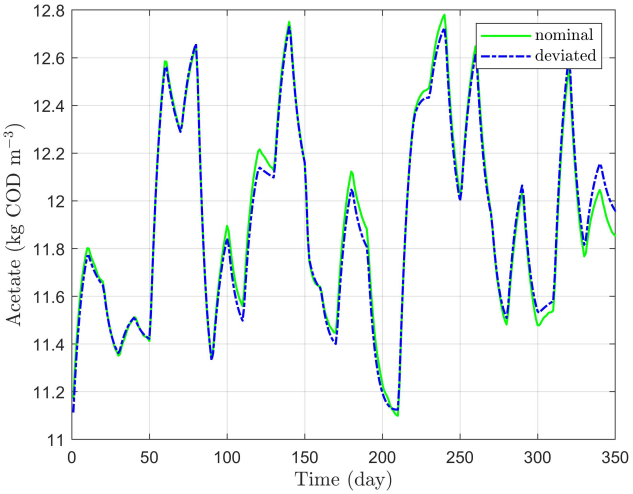


Figure 3: Simplified schematization of the anaerobic fermentation process for the purpose of estimation with a SBJ system.

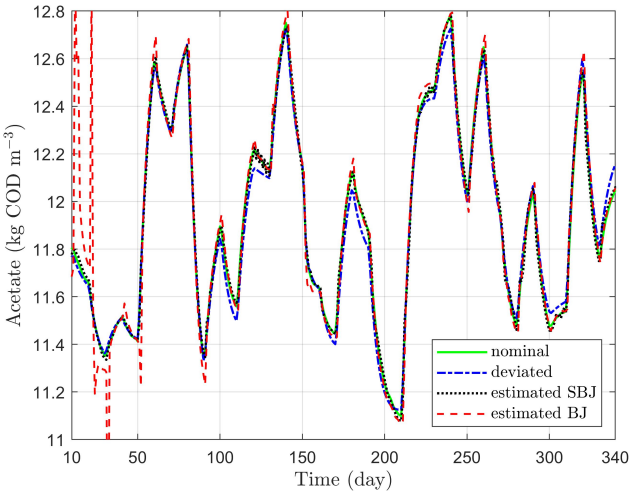


(a) The input flow rate produced by a pseudo random input signal used for identification process.

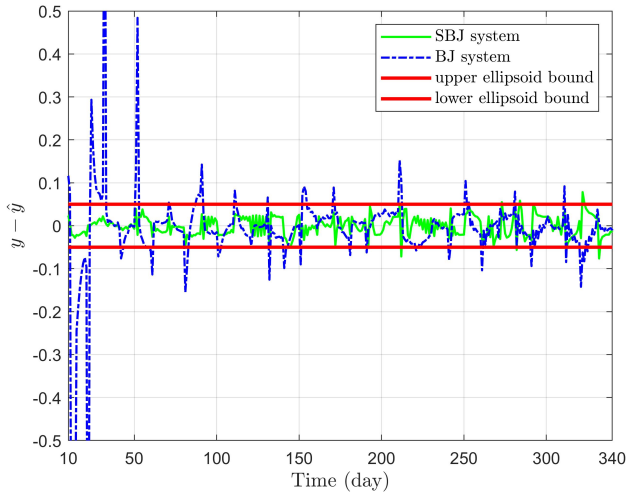


(b) The process output (acetate) in nominal operating condition (green line) and deviated (blue line) by random perturbation within the main components used for identification process.

Figure 4: Input (flow rate) and output (acetate) of the anaerobic fermentation process.



(a) The process output (acetate) in nominal operating condition (green line) and deviated (blue line) and its output prediction by a SBJ system (black line) and a BJ system (red line).



(b) The error comparison between the output of the estimated SBJ system (green line) and the output of the estimated BJ system (blue line).

Figure 5: Prediction performance of the proposed identification algorithm on the anaerobic fermentation process.

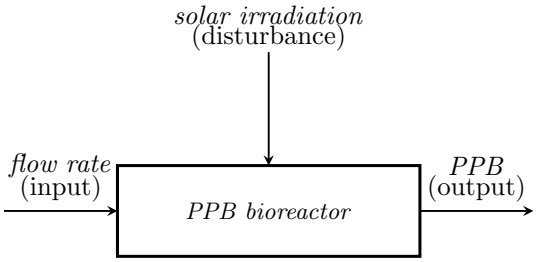
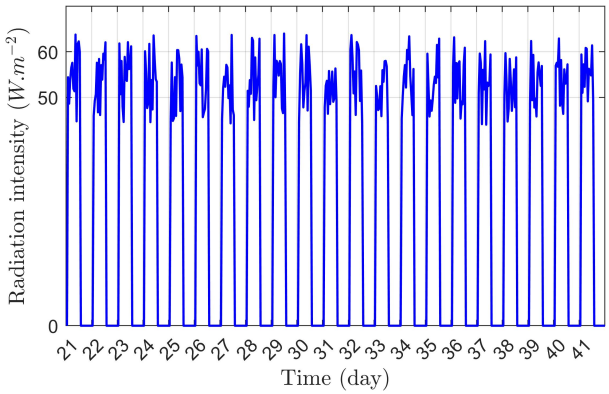
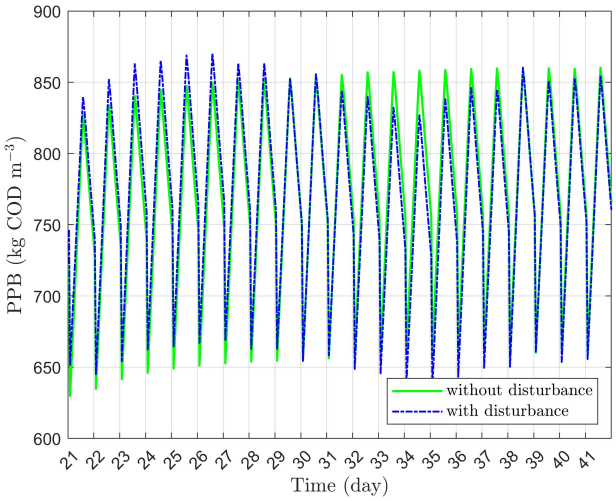


Figure 6: Simplified schematization of the purple bacteria raceway-pond photobioreactor process for the purpose of estimation with a switched BJ system.

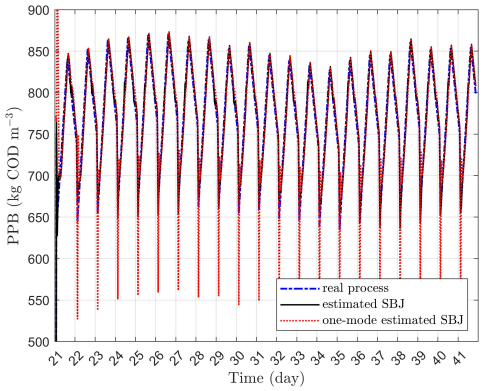


(a) The solar radiation fluctuation over a 24-hour period, with zero radiation occurring for 12 hours followed by non-zero radiation for the next 12 hours each day.

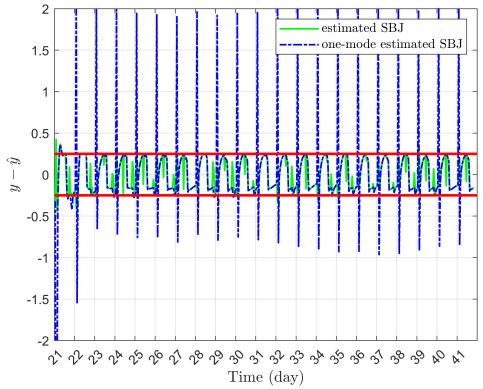


(b) the PPB production for the nominal-designed process (green line) and deviated PPB by disturbance caused by light intensity fluctuation (blue line).

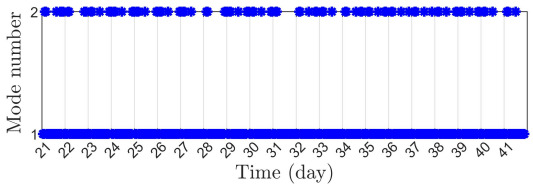
Figure 7: Implemented disturbance (solar radiation) and observed output (PPB) of the raceway-pond photobioreactor.



(a) The process output (blue line) and its output prediction by a SBJ system (black line) and the same estimated system with one mode, where more frequently mode is considered than the other one (red line).



(b) The error comparison between the output of the estimated SBJ system (green line) and the output of the estimated SBJ system with one mode, where that mode occurred more frequently is considered than the other (blue line).



(c) The switching patterns (mode occurrence) of two-mode estimated SBJ system.

Figure 8: Prediction performance of the proposed identification algorithm on the PPB photobioreactor.