Noname manuscript No.
(will be inserted by the editor)

Identifiability Analysis and Prediction Error Identification of Anaerobic Batch Bioreactors

L. Campestrini · D. Eckhard · R. Rui · A. S. Bazanella

Received: date / Accepted: date

Abstract This paper presents the identifiability analysis of a nonlinear model for a batch bioreactor and the estimation of the identifiable parameters within the prediction error framework. The output data of the experiment are the measurements of the methane gas generated by the process, during 37 days, and knowledge of the initial conditions is limited to the initial quantity of chemical oxygen demand. It is shown by the identifiability analysis that only three out of the eight model parameters can be identified with the available measurements and that identification of the remaining parameters would require further knowledge of the initial conditions. A prediction error algorithm is implemented for the estimation of the identifiable parameters.This algorithm is iterative, relies on the gradient of the prediction error, whose calculation is implemented recursively, and consists of a combination of two classic optimization methods: the conjugated gradient method and the Gauss-Newton method.

Keywords Identifiability · Nonlinear Identification · Prediction Error · Anaerobic Digestion · Batch Bioreactors

1 Introduction

Biological treatment of wastewater, along with the resulting production of biogas, plays a major role in the context of sustainable development. The biological treatment of waste water may be achieved by anaerobic digestion in bioreactors, especially when dealing with plant residues, food industry wastewater, animal wastes (Bernard et al. 2001). Anaerobic digestion presents several advantages with respect to aerobic treatment, among which the higher energy production and the lower sludge production are probably the most important ones (Bernard et al. 2001; Sbarciog et al. 2010; Antonelli et al. 2003). Obtaining dynamic models for anaerobic digestion processes is of great importance considering the conception, operation and optimization of bioreactors, what justifies the extensive studies in this field in the past decades (Bernard et al. 2001; Haag et al. 2003; Bogaerts and Vande Wouwer 2004).

Different models describing the anaerobic digestion are presented in the literature, from simple models like the one bacteria population model in (Andrews 1974), to quite complex ones, like the ADM1 model established by the IWA Anaerobic Digestion Modeling Task Group, which consists of 32 state variables (Batstone et al. 2002). Although in principle a complex model could represent the system more accurately, identification of its large number of parameters up to a sufficiently good accuracy can be a prohibitively hard task. For this reason, it is found in the literature wide acceptance of the representation of the anaerobic digestion process by moderate dimension models, with four to six state variables, which represent the acidogenesis and methanogenesis reactions (and ions balancing) (Bernard et al. 2001; Antonelli et al. 2003; Donoso-Bravo et al. 2011b). In this work, the main goal is the identification of the parameters in such models of moderate dimension, derived from mass-balance calculations.

Identification of the parameters of these phenomenological models presents a number of challenges. The question of whether or not it is possible to identify the model parameters from some appropriate experiment

L. Campestrini · D. Eckhard · R. Rui · A. S. Bazanella
Department of Electrical Engineering, Universidade Federal do Rio Grande do Sul,
Av. Osvaldo Aranha, 103, Porto Alegre, RS, Brazil
Tel.: +55-51-33084472
Fax: +55-51-33083129
E-mail: luciola@ece.ufrgs.br

(the identifiability issue) is challenging in itself (Margaria et al. 2001; Berthoumieux et al. 2012; Sedoglavic 2002; Karlsson et al. 2012). If the model is identifiable, then there is the issue of how to generate such "appropriate experiments" (the informativity issue). Finally, assuming that data from an appropriate experiment have been collected, the issue of using these data in such a way as to obtain the most accurate estimate of the parameters is also an open one, given the highly nonlinear nature of the models.

The parameter identification of mass balance models of continuous bioreactors has been performed in several different ways, as summarized in (Donoso-Bravo et al. 2011b). Often not all the model parameters can be identified, and this feature of partial identifiability in the different approaches has also been highlighted in (Donoso-Bravo et al. 2011b). It is common place to perform the parameter identification using steady state data only, like in the classical paper (Bernard et al. 2001). On one hand, this procedure is not feasible for batch reactors, in which there is no manipulated input to drive the reactor to a different steady state - usually there is not even any steady state condition possible. Indeed, batch reactors present different challenges than continuous reactors, and do not seem to have received the same attention in the past, even though they are equally important. On the other hand, disregarding the transient data and hence the information that they carry about the parameters' values will result in an estimation that is less accurate than could be achieved if the transient data were considered appropriately. The prediction error framework in system identification (Ljung 1999) is appropriate to do it, and its principles shall be applied in this paper. A least squares approach, which is reminiscent of prediction error methods, has been applied to linearized models in (Donoso-Bravo et al. 2011a).

In this paper an identifiability analysis of the mass balance model for a batch anaerobic digestion process is performed. The bioreactors under study and the mass balance models are described in Section 2. It is shown in Section 3 that identification of different parameters requires knowledge of different initial conditions. Then, as described in Section 4, the application of a prediction error method for the identification of the parameters in this class of models is proposed. This is a nonstandard prediction error problem, which becomes a highly nonconvex optimization problem. Accordingly, a dedicated optimization procedure had to be developed for its solution. This algorithm was presented firstly in (Campestrini et al. 2012). This optimization procedure uses simulations of the system model to generate estimates of the gradient of the prediction error, which is then used in gradient based iterations.

Simulation results under different scenarios, presented in Section 5.1, confirm the identifiability analysis and shed light into their interpretation, setting the stage for the practical application of the method. The simulations also show that convergence of the proposed prediction error method to the correct parameter values of the identifiable parameters is possible even under the practical circumstances of ignorance of the initial conditions. Last, but of course not least, Section 5.2.1 presents the experimental setup, which consists of one liter glass bottles (bioreactors) filled with substrate and inoculum, from which the amount of methane is measured. Data have been collected from two reactors during a 37 days experiment in which four output samples per day were measured. The proposed method is applied to these data, resulting in the identification of the parameters' values. Although values for all the parameters result from the procedure, the theoretical analysis shows that only three of them are meaningful. It is shown, not surprisingly, that even having identified only three parameters correctly, the model with the five incorrect parameter values produces reasonably accurate predictions of the output - that is, the remaining four parameters are not that relevant for the purpose of predicting the methane production.

2 Bioreactor Model

Models of high complexity are not adequate to be used in a control design (Bastin and Dochain 1990), since they result in complex controllers. Besides, due to the large amount of parameters to be estimated, its identification may be so costly and imprecise that the complexity of the model will not imply in an accurate description of the process. This motivates the use of models with moderate complexity, but still sufficiently detailed to describe the process' dynamics with sufficient precision, so that these models can be used in a control design, in order to obtain a controller that is as simple as possible. A mathematical model with these characteristics, acclaimed in the literature, is the model based on mass-balance, composed by four states and one output (Antonelli et al. 2003). The states are described by

$$\begin{cases} \dot{x}_1(t) = [\nu_1(S_1(t)) - \alpha D]x_1(t) \\ \dot{x}_2(t) = [\nu_2(S_2(t)) - \alpha D]x_2(t) \\ \dot{S}_1(t) = D(S_1^{in}(t) - S_1(t)) - k_1\nu_1(S_1(t))x_1(t) \\ \dot{S}_2(t) = D(S_2^{in}(t) - S_2(t)) + k_2\nu_1(S_1(t))x_1(t) \\ -k_3\nu_2(S_2(t))x_2(t), \end{cases} \tag{1}$$

where $x_1(t)$ (mg/L) is the concentration of acidogenic bacteria, $x_2(t)$ (mg/L) is the concentration of methanogenic bacteria, $S_1(t)$ (mg/L) is the concentration of chemical oxygen demand (COD), $S_2(t)$ (mmol/L) is the concentration of volatile fatty acids (VFA), $S_1^{in}(t)$ (mg/L) and $S_2^{in}(t)$ (mmol/L) are the influent concentrations of $S_1(t)$ and $S_2(t)$ respectively, $0 < \alpha \leq 1$ is a proportionality parameter of experimental determination, D (day^{-1}) is the dilution rate of the influents, k_1 (mg COD/mg x_1) is the yield coefficient for COD degradation, k_2 (mmol VFA/mg x_1) is the yield coefficient for fatty acid production, k_3 (mmol VFA/mg x_2) is the yield coefficient for fatty acid consumption.

The nonlinear behavior is given by the two specific microbial growth rates, $\nu_1(S_1(t))$ and $\nu_2(S_2(t))$, expressed by the Monod law

$$\nu_1(S_1(t)) = \mu_{m1}\frac{S_1(t)}{K_{S1}+S_1(t)} \tag{2}$$

$$\nu_2(S_2(t)) = \mu_{m2}\frac{S_2(t)}{K_{S2}+S_2(t)}, \tag{3}$$

where μ_{m1} (day^{-1}) is the maximum acidogenic biomass growth rate, μ_{m2} (day^{-1}) is the maximum methanogenic biomass growth rate, K_{S1}(mg/L) is the saturation parameter associated with $S_1(t)$ and K_{S2} (mmol/L) is the saturation parameter associated with $S_2(t)$.It is also common in the literature $\nu_2(S_2(t))$ expressed by the Haldane law, where there is an extra parameter which represents the VFA inhibition. However, when the substrate concentration is low, the VFA inhibition is not observed, and the Monod law is adequate to model the bacteria growth.

The output of the model is the methane flow rate $q_M(t)$, which is given by

$$q_M(t) = k_6\nu_2(S_2(t))x_2(t). \tag{4}$$

The bioreactor produces mainly two gases, methane and carbon dioxide. In order to obtain the methane production, which can be used as a source of energy, it is necessary to first measure the total gas flow and then evaluate the gas composition.

This model is able to represent both continuous and batch reactors. In a batch reactor, substrate and inoculum are inserted into the bioreactor only at the onset of the process, and for this reason are considered as initial conditions for the states related to the substrate and bacteria concentration, respectively. So, since there is no input for the system, $S_1^{in}(t) = S_2^{in}(t) = 0$ and $D = 0$, reducing the model (1) to

$$\begin{cases} \dot{x}_1(t) = \nu_1(S_1(t))x_1(t) \\ \dot{x}_2(t) = \nu_2(S_2(t))x_2(t) \\ \dot{S}_1(t) = -k_1\nu_1(S_1(t))x_1(t) \\ \dot{S}_2(t) = k_2\nu_1(S_1(t))x_1(t) - k_3\nu_2(S_2(t))x_2(t). \end{cases} \tag{5}$$

3 Identifiability Analysis

Consider the following class of deterministic continuous-time nonlinear model structures:

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t), \theta), \\ y(t,\theta) &= h(x(t), \theta) \end{aligned} \tag{6}$$

where $x(t) \in \mathbb{R}^n$ is the system's state, $u(t)$ and $y(t,\theta)$ are the system's input and output signals respectively, $f(\cdot,\cdot)$, $g(\cdot,\cdot)$ and $h(\cdot,\cdot)$ are given (i.e. known) analytical vector fields; $\theta \in \mathbb{R}^d$ is the parameter vector, whose estimation is the purpose of the identification procedure. The family of all models (6) generated by all $\theta \in \mathbb{R}^d$ is called the model class $\mathcal{M}$. In the same spirit as (Dasgupta et al. 1991; Karlsson et al. 2012) and (Ljung and Glad 1994), the following assumption on the input signal $u(.)$ is made.

Assumption 1 *The signal $u(t)$ is analytic and is such that the solution $x(t)$ of (6) is an analytic function.*

The virtue of this assumption is that knowing all derivatives of an analytic signal at some time is equivalent to knowing that signal everywhere, a property that is instrumental to the development of various identifiability tests.

Identifiability can be defined as follows.

Definition 1 (Identifiability) Consider the model (6) at a given parameter value θ_1. The model (6) is *locally identifiable at* θ_1 if there exists a $\delta > 0$ and a data set $z(.) \triangleq \{u(.), x_0\}$ such that, for all $\theta \in ||\theta - \theta_1|| \leq \delta$, the outputs of the model (6) with these two different parameter values θ and θ_1, both driven by the same data set, are identical (i.e. $y(t,\theta) = y(t,\theta_1) \;\; \forall t \geq 0$) only if $\theta = \theta_1$. The model (6) is *globally identifiable* at θ_1 if the same holds for all $\delta > 0$. The model is *structurally identifiable* if it is identifiable at almost all θ.

This definition relies on the possible existence of an appropriate data set $z(.)$ which allows to differentiate between different values of θ by measuring the output. Note that this may require knowledge of the initial conditions, as will be detailed shortly. Such a data set, when it exists, is called informative.

Definition 2 (Informativity) The data set $z(.) \triangleq \{u(.), x_0\}$ is locally informative at θ_1 for the model set (6) if there exists a $\delta > 0$ such that, for all $\theta \in ||\theta - \theta_1|| \leq \delta$, the outputs of the model (6) with these two different parameter values θ and θ_1, both driven by this same data set $z(.)$, are identical (i.e. $y(t,\theta) = \hat{y}(t,\theta_1) \;\; \forall t \geq 0$) only if $\theta = \theta_1$.

These definitions exhibit the two ingredients that are necessary for a meaningful identification: informativity, which is a property of the data set, and identifiability, which refers to the possible existence of an informative data set given a particular model structure, and thus is a property of the model structure. Whereas informativity depends on the true system, because it generates the data, identifiability is a property of the model structure (in particular, it does not depend on the true system itself or in the satisfaction of Assumption 1).

Notice that, in the case of the batch reactor, the informativity must be provided by a data set of the form $z(.) \triangleq \{0, x_0\}$, since the system does not allow the application of input signals. In the case of other reactors, where input signals may me applied, the initial conditions can be neglected if the data are collected after the initial transient dies out.

Identifiability test

Given a parametrized class of models in state space form (6), the j-th derivative with respect to time of any order of $y(t)$ evaluated at $t = 0$ can be expressed by

$$y^{(j)}(0) = \sum_{i=1}^{n} (f_i) \frac{\partial y^{(j-1)}}{\partial x_i}\bigg|_{x=x(0),u^{(k)}=u^{(k)}(0),k=1,\cdots,j}, \tag{7}$$

where f_i is the i_{th} element of the vector field $f(\cdot,\cdot)$ and likewise for g_i. This gives explicit expressions linking the initial values of the state variables with the initial value of a derivative with respect to time of any order of the output of the system. The right hand side of (7) is an expression in $x(0)$ and θ, and the left hand side is a measured quantity for any j. Taking derivatives up to $n + d$ (the state dimension plus the number of parameters), this expression can be written in the form

$$\mathcal{Y} = \mathcal{Y}(x(0), \theta), \tag{8}$$

where $\mathcal{Y}$ is a column vector containing $y^{(j)}(0), j = 0, \cdots, n + d - 1$, and the dependence on $u^{(j)}(0)$ has been absorbed into the notation of the vector valued function on the right hand side. Identifiability at some θ_1 is then equivalent to the existence of a data set such that the map $\mathcal{Y}(x(0), \theta)$ in the right hand side of (8) is a local diffeomorphism at θ_1 (Karlsson et al. 2012).This will be the case, according to the inverse function theorem, if and only if the Jacobian matrix

$$J_a(x_0, \theta_1) = \frac{\partial \mathcal{Y}(x(0), \theta)}{\partial(x, \theta)}\bigg|_{x=x_0,\theta=\theta_1} \tag{9}$$

has full rank.

Thus, local identifiability at a given θ_1 can be checked by calculating the rank of the Jacobian for θ_1 and some initial condition x_0. The model is identifiable if there exists an experiment (that is, an initial condition) such that the Jacobian is full rank. If it exists, such an initial condition provides an informative experiment. On the other hand, structural identifiability can be checked by calculating the rank of the Jacobian for a randomly generated parameter value of θ. When this structural identifiability test succeeds, that is, the corresponding Jacobian matrix is full rank, then the parameter value can be determined by solving equation (8).

However, the symbolic computation of the Jacobian matrix in terms of (8) suffers from a computational complexity problem and can become prohibitively costly even for modestly sized problems. In this paper the algorithm proposed in (Karlsson et al. 2012) is adopted, which reduces this computational burden dramatically via the computation of power series expansions of the partial derivatives of the system's output with respect to $x(0)$ and θ.

The results of the identifiability test allow to interpret the results of the identification procedure correctly. As will be shown later, knowledge of the states' initial conditions may be required for identification of the parameters. When the initial conditions are not known, only a subset of the total parameters of a model are correctly identified. Understanding this is necessary to interpret correctly the identification results.

4 Identification of the model

This section presents a dedicated method to identify the parameters of the batch reactor model (5), based on the prediction error method for linear systems, widely used in system identification. Since the bioreactor model is nonlinear and not in standard black-box form, standard identification software packages do not apply directly; it is necessary to adapt the prediction error method as will be shown in this section.

First of all, the model (5) needs to be converted into a discrete-time model. Using Euler's method, that is $\dot{x}(t) \approx \frac{x(k+1)-x(k)}{T}$, where T is the sampling period, yields the following model class

$$\mathcal{M} = \begin{cases} x_1(k+1,\theta) &= x_1(k,\theta) + Tv_1(k,\theta) \\ x_2(k+1,\theta) &= x_2(k,\theta) + Tv_2(k,\theta) \\ S_1(k+1,\theta) &= S_1(k,\theta) - T\theta_5 v_1(k,\theta) \\ S_2(k+1,\theta) &= S_2(k,\theta) + T\theta_6 v_1(k,\theta) \\ &\quad -T\theta_7 v_2(k,\theta), \end{cases} \tag{10}$$

where

$$\begin{cases} v_1(k,\theta) = \theta_1 \frac{S_1(k,\theta)x_1(k,\theta)}{S_1(k,\theta)+\theta_2} \\ v_2(k,\theta) = \theta_3 \frac{S_2(k,\theta)x_2(k,\theta)}{S_2(k,\theta)+\theta_4} \end{cases}$$

and the output system's predictor is given by

$$\hat{q}_M(k,\theta) = \theta_8 v_2(k,\theta). \tag{11}$$

The parameter vector θ has been defined as

$$\begin{aligned} \theta &= \left[\, \theta_1 \ldots \theta_8 \,\right] \\ &= \left[\, \mu_{m1}\ K_{S1}\ \mu_{m2}\ K_{S2}\ k_1\ k_2\ k_3\ k_6 \,\right]. \end{aligned} \tag{12}$$

Using only output data $q_M(k),\ k = 1, \ldots N$ and the states' initial conditions, the parameters of the model (10) will be estimated through the prediction error method (Ljung 1999). The prediction error estimate of the parameter vector $\hat{\theta}_N$ is given by

$$\hat{\theta}_N = \arg\min_{\theta} J(\theta) \tag{13}$$

where

$$J(\theta) \triangleq \frac{1}{N}\sum_{k=1}^{N} (\varepsilon(k,\theta))^2 \tag{14}$$

and $\varepsilon(k,\theta) = q_M(k) - \hat{q}_M(k,\theta)$.

Standard system identification solutions are not appropriate to solve this problem, since they are designed to estimate black-box linear or nonlinear models with standard structures, such as NARX, NARMAX or also neural networks for the nonlinear case (Aguirre and Jacome 1998; Sjöberg et al. 1994). The identification of the bioreactor model must be of a "grey-box" nature in order to obtain the system's physical parameters.

Concerning the identification of (10), the most common solutions found in the literature are obtained through the use of algorithms that estimate the cost function gradient numerically, without the use of an analytic expression of it (Donoso-Bravo et al. 2011b), which increases the computational cost. Thus, since classical optimization algorithms are based on the cost function gradient, an analytic expression of this function allows an improvement of the optimization process, making it faster than the usual solutions. The cost function gradient can be estimated as

$$\begin{aligned} \nabla J(\theta) &= -\frac{1}{N}\sum_{k=1}^{N} 2\varepsilon(k,\theta)\frac{\partial}{\partial\theta}\varepsilon(k,\theta) \\ &= -\frac{1}{N}\sum_{k=1}^{N} 2\varepsilon(k,\theta)\frac{\partial}{\partial\theta}\hat{q}_M(k,\theta) \\ &= -\frac{1}{N}\sum_{k=1}^{N} 2\varepsilon(k,\theta)(\theta_8\frac{\partial}{\partial\theta}v_2(k,\theta) + v_2(k,\theta)\frac{\partial}{\partial\theta}\theta_8). \end{aligned}$$

From now on, in order to shorten the analytical expressions and thus improve readability, the dependence on θ will be omitted. The term $\frac{\partial}{\partial\theta}v_2(k,\theta)$ can be estimated as

$$\begin{aligned} \frac{\partial}{\partial\theta}v_2(k) = {}& \frac{\partial}{\partial\theta}\theta_3\frac{S_2(k)x_2(k)}{S_2(k)+\theta_4} + \theta_3\left[-\frac{S_2(k)x_2(k)}{(S_2(k)+\theta_4)^2}\left(\frac{\partial}{\partial\theta}S_2(k)\right.\right. \\ & \left.\left.+\frac{\partial}{\partial\theta}\theta_4\right) + \frac{x_2(k)}{S_2(k)+\theta_4}\frac{\partial}{\partial\theta}S_2(k) + \frac{S_2(k)}{S_2(k)+\theta_4}\frac{\partial}{\partial\theta}x_2(k)\right], \end{aligned}$$

where the other partial derivatives are estimated as

$$\begin{aligned} &\frac{\partial}{\partial\theta}x_1(k+1) = \frac{\partial}{\partial\theta}x_1(k) + T\frac{\partial}{\partial\theta}v_1(k) \\ &\frac{\partial}{\partial\theta}x_2(k+1) = \frac{\partial}{\partial\theta}x_2(k) + T\frac{\partial}{\partial\theta}v_2(k) \\ &\frac{\partial}{\partial\theta}S_1(k+1) = \frac{\partial}{\partial\theta}S_1(k) - Tv_1(k)\frac{\partial}{\partial\theta}\theta_5 - T\theta_5\frac{\partial}{\partial\theta}v_1(k) \\ &\frac{\partial}{\partial\theta}S_2(k+1) = \frac{\partial}{\partial\theta}S_2(k) + Tv_1(k)\frac{\partial}{\partial\theta}\theta_6 + T\theta_6\frac{\partial}{\partial\theta}v_1(k) \\ &- Tv_2(k)\frac{\partial}{\partial\theta}\theta_7 - T\theta_7\frac{\partial}{\partial\theta}v_2(k) \\ &\frac{\partial}{\partial\theta}v_1(k) = \frac{\partial}{\partial\theta}\theta_1\frac{S_1(k)x_1(k)}{S_1(k)+\theta_2} + \theta_1\left[-\frac{S_1(k)x_1(k)}{(S_1(k)+\theta_2)^2}\left(\frac{\partial}{\partial\theta}S_1(k)\right.\right. \\ &\left.\left.+\frac{\partial}{\partial\theta}\theta_2\right) + \frac{x_1(k)}{S_1(k)+\theta_2}\frac{\partial}{\partial\theta}S_1(k) + \frac{S_1(k)}{S_1(k)+\theta_2}\frac{\partial}{\partial\theta}x_1(k)\right]. \end{aligned}$$

The optimization problem (13) is solved by an algorithm that consists in a series of conjugate gradient iterations followed by Gauss-Newton iterations (Campestrini et al. 2012) . It is known that the region of attraction of the conjugate gradient method is larger than the Gauss-Newton, while on the other hand the Gauss-Newton algorithm presents a higher convergence rate. These properties motivate the combination of these two algorithms that is applied here.

The conjugate gradient algorithm (Polak 1973) can be described as

$$\theta(i+1) = \theta(i) - \gamma(i)D(i,\theta),$$

where

$$D(i,\theta) = \nabla J(\theta) + \lambda D(i-1,\theta)$$

is the direction of the parameters update, $\gamma(i)$ is a parameter that determines the size of the iteration step and λ is a factor that states the variation rate of the algorithm's direction. When $\lambda = 0$ this algorithm reduces to the steepest descent method algorithm, whose direction tends to make the convergence rate to become very low. This happens mainly because the algorithm evolves in a zigzag path. The larger λ is, the lower is the variation of the algorithm's direction, which minimizes the zigzag effect. In this work the following parameters were used:

$$\lambda = 4 \tag{15}$$

$$\gamma(i) = \frac{\beta}{\|D(i,\theta)\|}, \tag{16}$$

where the β factor is incremented in 1% of its value at each iteration when the cost function value is decreased ($J(i+1) < J(i)$), and decremented in 1% in the cases where the cost increases ($J(i+1) \geq J(i)$). A minor variation was also implemented in the algorithm: if the cost increases in an iteration, we make $\theta(i+1) = \theta(i)$, so the algorithm returns to a known point with a lower cost.

When the convergence rate of the algorithm becomes too low, a switch to Gauss-Newton iterations is made, which are described as

$$\theta(i+1) = \theta(i) - \gamma R^{-1}(\theta(i))\nabla J(\theta(i)),$$

where γ is the step size of the iteration and

$$R = \sum_{k=1}^{N}(\frac{\partial}{\partial\theta}\hat{q}_M(k,\theta))(\frac{\partial}{\partial\theta}\hat{q}_M(k,\theta))^T.$$

Notice that the calculation of the gradient (which is necessary to estimate the parameter vector) needs $\hat{q}_M(k,\theta)$, which in turn needs the estimate of all states. So, when estimating $\theta(i+1)$ the simulation of the system is done using $\theta(i)$ and, in the end of the iterative procedure to estimate the parameter θ, an estimate of the state variables of the system are also obtained.

5 Results

5.1 Simulated results

In order to validate the method and the analysis proposed in this work, the identifiability test presented in Section 3 together with the identification algorithm presented in Section 4 are applied to a simulated bioreactor.

Suppose the real process is described by the model (10), where the real parameter vector θ_0 (Dochain 2008) is given by

$$\theta_0 = \begin{bmatrix} 1.2 \times 10^0 \; 7.1 \times 10^0 \; 7.4 \times 10^{-1} \; 9.28 \times 10^0 \\ 4.214 \times 10^1 \; 1.165 \times 10^2 \; 2.68 \times 10^2 \; 4.53 \times 10^2 \end{bmatrix}^T \quad (17)$$

and that the initial conditions are given by $x_1(0) = 0.2$ mg/L, $x_2(0) = 0.8$ mg/L, $S_1(0) = 9.5$ mg/L and $S_2(0) = 93$ mmol/L. The process' behavior is presented in Figure 1, where the top figures present the behavior of the states $x_1(k)$, $x_2(k)$, $S_1(k)$ and $S_2(k)$ and the bottom figures present the production of methane.

Structural identifiability of the model (1) will be determined through the method presented in Section 3. The method will be applied in three different scenarios, corresponding to different knowledge of the initial conditions. After testing structural identifiability in each scenario, the prediction error identification method described in Section 4 is applied to estimate the parameters. The identification is performed in each case with a set of $N = 72$ samples of the output, where it is collected during 3 days, every hour.

5.1.1 Knowledge of all initial conditions

Assuming that all four initial conditions $x_1(0)$, $x_2(0)$, $S_1(0)$ and $S_2(0)$ are known a priori, the parameter vector θ can be recovered from (8) if the rank of the Jacobian J_a is $d = 8$. The structural identifiability test described in Section 3 is applied to this scenario and results in a positive answer. That is, when all initial conditions are known, all the parameters $\theta_1 \ldots \theta_8$ can be identified.

Next, the prediction error identification method has been applied to the model (1) with the data presented in Fig 1. In order to do that, the parameter vector θ must be initialized. The following initialization was used

$$\theta(0) = \begin{bmatrix} 1.32 \times 10^0 \; 7.81 \times 10^0 \; 8.14 \times 10^{-1} \; 1.0208 \times 10^1 \\ 4.6354 \times 10^1 \; 1.2815 \times 10^2 \; 2.948 \times 10^2 \; 4.983 \times 10^2 \\ 4.983 \times 10^2 \end{bmatrix}^T. \quad (18)$$

After $100,000$ conjugate gradient iterations, $50,000$ Gauss-Newton iterations with $\gamma = 0.001$ and 50 Gauss-Newton iterations with $\gamma = 1$, the correct values of the parameters with 12 significant digits were obtained. This result corroborates the result of the identifiability test, since all the parameters are correctly estimated. Since the simulated data were noise free, there is no variance error in the estimate. In a practical situation, even when all the parameters of the system are identifiable, the estimate will present some variance error due to the unavoidable noise present in the measurements obtained from the real process.

5.1.2 Knowledge of no initial conditions

A totally different scenario would be the lack of knowledge of all initial conditions, that is, when one has access only to the output signal of the process and do not know the initial conditions of the states. Applying the identifiability test to this scenario, it is determined that the model is not structurally identifiable, and that the only parameters that are identifiable in the system are θ_1 and θ_3. So, even when the initial conditions of the states are unknown, it is still possible to identify the maximum biomass growth rates correctly for both bacteria.

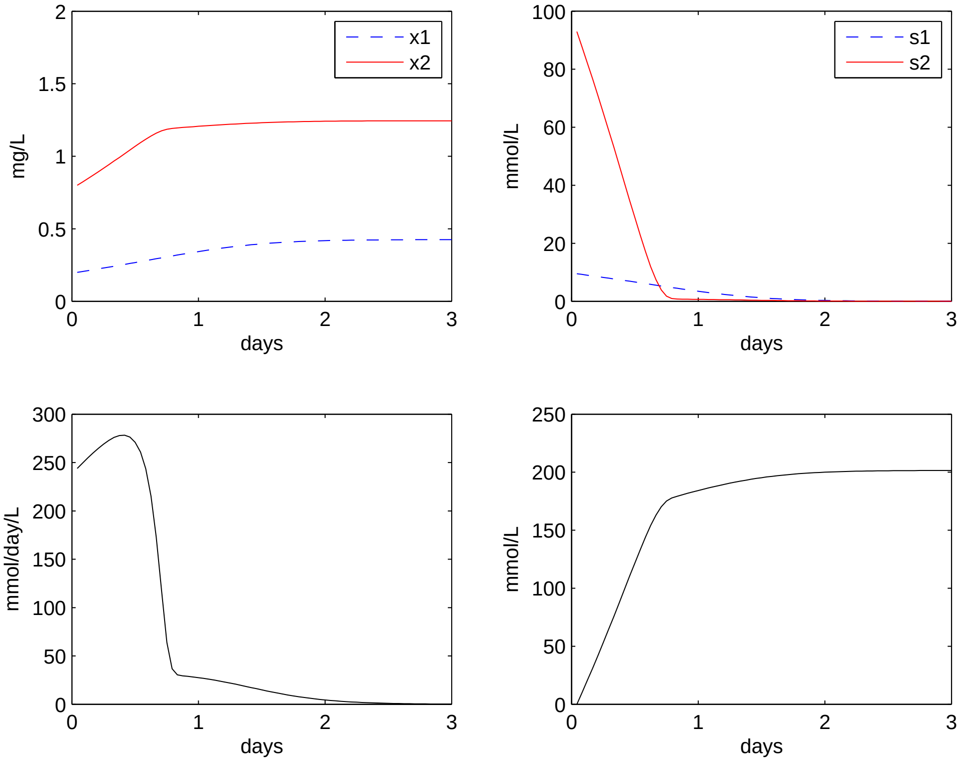

Fig. 1 Behavior of the simulated batch reactor, described by (10) and (17): top figures present the state variables while bottom figures present the production of methane (left figure) and the accumulated amount of methane (right figure).

To test this result, the identification of the system is performed, as presented in Section 4. The initial conditions are unknown, but they are necessary to initialize the identification algorithm, so the following randomly generated values were used $[x_1(0) \quad x_2(0) \quad S_1(0) \quad S2(0)]^T$ $= \left[3.7374\times10^{-2}\ 3.9181\times10^{-1}\ 4.2330\times10^{0}\ 6.0107\times10^{1}\right]^T$. The initial value of the parameter vector $\theta(0)$ is chosen as in (18).

After $100,000$ conjugate gradient iterations, $50,000$ Gauss-Newton iterations with $\gamma = 0.001$ and 50 Gauss-Newton iterations with $\gamma = 1$, the following estimate was obtained

$$\begin{aligned}\hat{\theta}_N \cong \big[\,&1.2000 \times 10^{0}\ 3.1636 \times 10^{0}\ 7.4000 \times 10^{-1}\\ &5.9977 \times 10^{0}\ 1.0048 \times 10^{2}\ 4.0292 \times 10^{2}\\ &3.5366 \times 10^{2}\ 9.2493 \times 10^{2}\,\big]^T .\end{aligned}$$

The identification result again corroborates the identifiability analysis, since θ_1 and θ_3 were estimated correctly and the estimates of the other parameters are incorrect. The prediction error cost $(J(\theta))$ has vanished, which indicates that the output of the estimated model correctly predicts the system's output - that is, $\hat{q}_M(k, \hat{\theta}_N) \approx q_M(k, \theta_0)\ \forall k$.

5.1.3 Knowledge of only $S_1(0)$

In the previous examples two extreme situations were presented: the case where all initial conditions are known (and it is possible to identify all the parameters of the chosen model) and the case where the initial condition are unknown (and in this case it is possible to identify only two parameters of the model). It is clear that the more information about the initial conditions is available, more parameters can be identified.

In the practical example, as it will be seen in the next section, only the information of $S_1(0)$ is available, the initial concentration of COD. Let us investigate what are the parameters that can be identified in this scenario. According to the identifiability analysis, if $S_1(0)$ is the only initial condition that is known, then it is possible to identify θ_1, θ_2 and θ_3. So the knowledge of $S_1(0)$ allows the correct estimate of θ_2, in addition to the estimation of θ_1 and θ_3 that could already be identified even without any knowledge of initial conditions.

The identification algorithm is applied, using the correct initial condition for $S_1(0)$ and random values for the other states initial conditions: $[x_1(0) \quad x_2(0) \quad S_1(0)$ $S_2(0)]^T = \big[1.4187 \times 10^{-1}\ 6.0374 \times 10^{-1}\ 9.5000 \times 10^{0}$

$6.3212 \times 10^1\big]^T$. The initialization of the parameter vector was again used as in (18). After $100,000$ conjugated gradient iterations, $50,000$ Gauss-Newton iterations with $\gamma = 0.001$ and 50 Gauss-Newton iterations with $\gamma = 1$, the following estimate was obtained

$$\hat{\theta}_N \cong \big[1.2000 \times 10^0 \; 7.1000 \times 10^0 \; 7.4000 \times 10^{-1} \\ 6.3076 \times 10^0 \; 5.9405 \times 10^1 \; 1.1162 \times 10^2 \\ 2.4137 \times 10^2 \; 6.0024 \times 10^2\big]^T .$$

Note that the correct result for only the three first components of the parameter vector is exactly what it has been foreseen using the identifiability analysis. Again, based on the value of the prediction error cost, which is very close to zero, it is seen that the model is able to predict the output of the system even though some parameters are incorrect.

Based on these results, the identification algorithm can now be applied to real data, collected from a real batch reactor.

5.2 Practical application

The biogas quantification equipments, used to collect the biogas produced by the bioreactors, are located in the Bioreactors Laboratory, in Centro Universitário UNIVATES, Brazil. The bioreactors are 1 liter glass bottles, as presented in the bottom picture of Figure 2. Each bioreactor gas production is measured by an equipment formed by a gas collector constituted by glass U-shaped tube, an optical sensor, a styrofoam ball and an electronic circuit that records the biogas flow and calculates the generated gas volume. A set of these equipments is presented in the top picture of Figure 2. The operating principle of the device is the fluid displacement, and the biogas quantification is performed when the biogas, as it fills the U-shaped tube, moves the fluid (water) down in one side and raises the fluid level in the opposite side, which is detected by the optical sensor, and the information is then sent to the electronic circuit. The generated biogas volume is determined by the combined gas law, which states that the ratio between the pressure-volume product and the temperature of a system remains constant. The experiments were realized in a bacteriological incubator adapted for this purpose, where it is possible to keep the temperature constantly at 35^oC inside the reactors, which is the indicated temperature for the biogas production (Bernard et al. 2001; Batstone et al. 2009; García-Ochoa et al. 1999). As can be seen in the bottom picture of Figure 2, the incubator contains several bioreactors, and data from two of them have been collected. Each bioreactor is filled with 420 mL of substrate and 180 mL of inoculum, mixed and homogenized, obtained from a large-scale biodigester. The substrate is formed by sludge from the Wastewater Treatment Station of Ecological Cooperative of Vale do Caí (Ecocitrus), in Rio Grande do Sul, Brazil. The qualification of the biogas produced, measured in percentage of methane, was performed from the injection of biogas at a specific sensor for measuring the concentration of methane, named Advanced Gasmitter. These measurements were realized 4 times a day (every 6 hours) during 37 days, thus generating a vector with $N = 148$ samples to be used in the identification procedure.

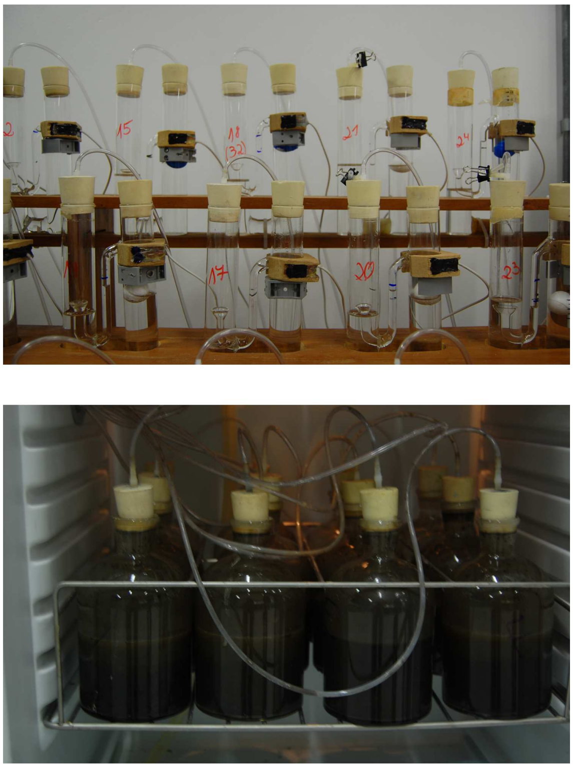

Fig. 2 Bioreactors and equipment used to collect gas.

5.2.1 Experimental results

From the experiments realized, one set of data was used to perform the identification of the bioreactor, and other set to validate the identified model. As stated above, $N = 148$ samples of the output of the system $q_M(k)$ were collected, and the known initial condition was $S_1(0) = 74$ mg/L, for both bioreactors. As seen in the identifiability analysis, knowing only $S_1(0)$ allows the correct identification of θ_1, θ_2 and θ_3. The measured output of one of the bioreactors is the continuous line presented in Figure 3. Using this set of data collected from one bioreactor, the identification algorithm presented in Section 4 was applied, where it was set $T = 1/4$ day. To set up the identification procedure, the

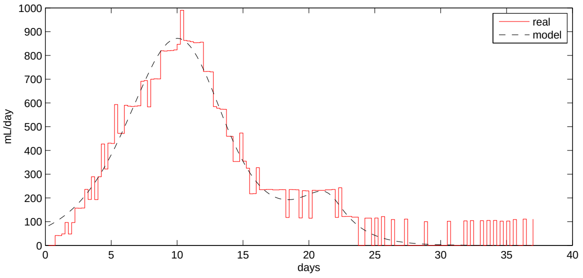

Fig. 3 CH_4 measurements obtained from the experimental data of one bioreactor compared with the output of the obtained model.

simulation of the behavior of the system given by the model (10) is needed. The only known quantity is the initial condition of the state $S_1(k)$, so the other initial conditions were chosen according to literature values (Dochain 2008) as

$$\begin{bmatrix} x_1(0) \\ x_2(0) \\ S_1(0) \\ S2(0) \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.8 \\ 74 \\ 93 \end{bmatrix}. \tag{19}$$

The initial values for the parameter vector were chosen as

$$\theta(0) \cong \big[4 \times 10^{-1} \; 2 \times 10^{1} \; 3 \times 10^{0} \; 6 \times 10^{2} \\ 4 \times 10^{-1} \; 7 \times 10^{-2} \; 3 \times 10^{0} \; 3 \times 10^{2} \big]^T. \tag{20}$$

After $1,000,000$ conjugated gradient iterations, $50,000$ Gauss-Newton iterations with $\gamma = 0.001$ and 50 Gauss-Newton iterations with $\gamma = 1$, the following estimate was obtained

$$\hat{\theta}_N \cong \big[4.2912 \times 10^{-1} \; 1.3065 \times 10^{1} \; 2.6493 \times 10^{0} \\ 5.7127 \times 10^{2} \; 3.1204 \times 10^{-1} \; 6.2776 \times 10^{-2} \\ 3.1473 \times 10^{0} \; 2.7862 \times 10^{2} \big]^T.$$

Hence, it can be said that for this bioreactor $\mu_{m1} = 0.42912$ day^{-1}, $K_{S1} = 13.065$ mg/L and $\mu_{m2} = 2.6493$ day^{-1}. The comparison between the real data and the estimated model is presented in Figure 3, where the cost (14) obtained in the estimation was $J = 3159.72$ mL^2/L^2/sample.

Because the data are very noisy, a Butterworth filter of fifth order was used to filter the data, with cut-off frequency $\omega_c = 0.6\pi$ rad/sample, in order to compare the response of the model with the real data. This comparison is presented in Figure 4 (bottom-left figure), and the cost calculated using this filtered signal is $J = 2323.39$ mL^2/L^2/sample. Also shown in this figure are the estimates of the states of the system (top figures), and the comparison between the accumulated amount of CH_4 of the real data and the estimate of the model (bottom-right figure). Notice that, despite the noise present in the measurements and the fact that most parameter values are not correctly estimated, the model is able to approximately represent the system's output. The validation of the obtained model was realized with the other set of data, taken from a second reactor. The comparison of the output of the model, using (19) as the initial conditions with the filtered data[1] was done regarding to the measured output of the second bioreactor; the cost is calculated as $J = 2676.66$ mL^2/L^2/sample. The cost obtained is slightly higher than the one obtained in the identification procedure, but the comparison between the real data and the model is similar to the one presented in Figure 4.

In this practical problem, only the COD measure at the initial of the experiment and the CH_4 measure during all the experiment were available. Although, in the literature, it is shown that it is also possible to measure the VFA using gas chromatography and the total biomass $x_1(t) + x_2(t)$, but not each biomass concentration separately. It is natural to think that with this knowledge in the beginning of the experiment, it would be possible to estimate more parameters. When performing the identifiability test (presented in Section 3) taking into account the knowledge of $S_1(0)$, $S_2(0)$, $x_1(0) + x_2(0)$ and only one parameter in the set $\{\theta_5,\ \theta_6,\ \theta_7,\ \theta_8\}$ (obtained from the literature or an *extra* experiment), it is possible to identify all parameters of the model. Another way of identifying all parameters is using the knowledge of $S_1(0)$, $S_2(0)$, one parameter in the set $\{\theta_5,\ \theta_6\}$, and one parameter in the set $\{\theta_7,\ \theta_8\}$. As future work, we search for these *extra* experiments that will allow the complete identification of the model.

6 Conclusions

The identification of the identifiable parameters in a four-state mass-balance model for a batch anaerobic digestion reactor, using a customized prediction error, was performed. The prediction error framework makes use of the transient data for the identification, which is a necessary feature in the identification of batch reactor models. The measurements available allowed the identification of three parameters, and an important message coming from the theoretical analysis is that it is impossible to identify the remaining parameters without extra knowledge about the initial conditions or previous knowledge of some of the remaining parameters.

[1] The data were filtered with the same Butterworth filter used in the previous case.

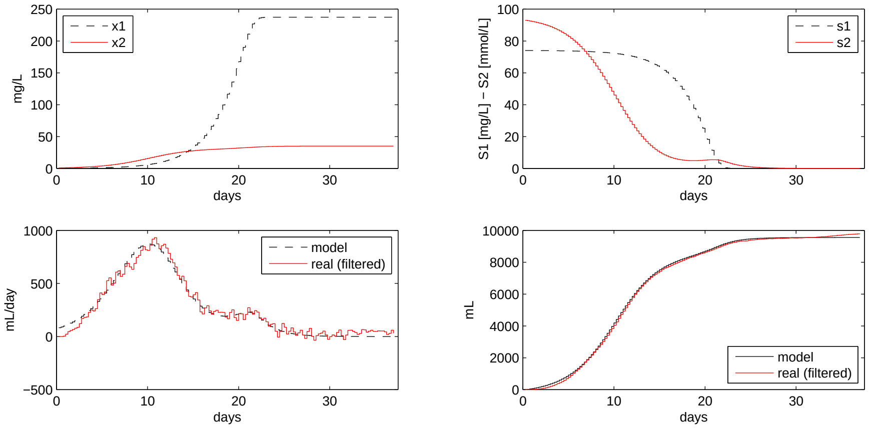

Fig. 4 Behavior of the model compared with the filtered output of the system. Top figures present the estimate of the states obtained in the end of the identification procedure; bottom figures present the estimate of the output of the model, compared with the filtered real data.

Thus, identification of additional parameters would require additional measurements.

Acknowledgments

The authors are grateful to Prof. Odorico Konrad from Centro Universitário UNIVATES for supplying the methane measurement samples and to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and CNPq/Brazil for financial support.

References

L.A. Aguirre, C.R.F. Jacome, Cluster analysis of NARMAX models for signal-dependent systems. Control Theory and Applications, IEE Proceedings - **145**(4), 409–414 (1998). doi:10.1049/ip-cta:19982112

J.F. Andrews, Dynamic models and control strategies for wastewater treatment processes. Water Research **8**(5), 261–289 (1974). doi:10.1016/0043-1354(74)90090-6

R. Antonelli, J. Harmand, J.-P. Steyer, A. Astolfi, Set-point regulation of an anaerobic digestion process with bounded output feedback. Control Systems Technology, IEEE Transactions on **11**(4), 495–504 (2003). doi:10.1109/TCST.2003.813376

G. Bastin, D. Dochain, *On-line Estimation and Adaptive Control of Bioreactors* (Elsevier, ???, 1990)

D.J. Batstone, S. Tait, D. Starrenburg, Estimation of hydrolysis parameters in full-scale anerobic digesters. Biotechnology and Bioengineering **102**(5), 1513–1520 (2009). doi:10.1002/bit.22163

D.J. Batstone, J. Keller, I. Angelidaki, S.V. Kalyuzhnyi, S.G. Pavlostathis, A. Rozzi, W.T.M. Sanders, H. Siegrist, V.A. Vavilin, The IWA anaerobic digestion model no 1 (ADM1). Water Science and Technology **45**(10), 65–73 (2002)

O. Bernard, Z. Hadj-Sadok, D. Dochain, A. Genovesi, J.P. Steyer, Dynamical model development and parameter identification for an anaerobic wastewater treatment process. Biotechnology and Bioengineering **75**(4), 424–38 (2001)

S. Berthoumieux, D. Kahn, H. de Jong, E. Cinquemani, Structural and practical identifiability of approximate metabolic network models, in *Preprints of the 16th IFAC Symposium on System Identification*, IFAC, Brussels, Belgium, 2012, pp. 1719–1724. IFAC

P. Bogaerts, A. Vande Wouwer, Parameter identification for state estimation-application to bioprocess software sensors. Chemical Engineering Science **59**, 2465–2476 (2004)

L. Campestrini, D. Eckhard, O. Konrad, A.S. Bazanella, Identificação não linear de um biorreator através da minimização do erro de predição, in *XIX Congresso Brasileiro de Automática*, vol. 1, Campina Grande, Paraiba, 2012, pp. 3066–3072

S. Dasgupta, M. Gevers, G. Bastin, G. Campion, L. Chen, Identifiability of scalar linearly parametrized polynomial systems, in *Proc. 9th IFAC Symposium on Identification and System Parameter Estimation*, vol. 1, Budapest, Hungary, 1991, pp. 374–378

D. Dochain, *Automatic Control of Bioprocess* (Wiley, Hoboken, NJ, USA, 2008)

A. Donoso-Bravo, S. Pérez-Elvira, E. Aymerich, F. Fdz-Polanco, Assessment of the influence of thermal pre-treatment time on the macromolecular composition and anaerobic biodegradability of sewage sludge. Bioresource Technology **102**(2), 660–666 (2011a). doi:10.1016/j.biortech.2010.08.035

A. Donoso-Bravo, J. Mailier, C. Aceves-Lara, C. Martin, J. Rodriguez, A. Vade Wouver, Model selection, identification and validation in anaerobic digestion: A review. Water Research **45**(17), 5347–5364 (2011b). doi:10.1016/j.watres.2011.08.059

F. García-Ochoa, V.E. Santos, L. Naval, E. Guardiola, B. López, Kinetic model for anaerobic digestion of livestock manure. Enzyme and Microbial Technology **25**(1-2), 55–60 (1999). doi:10.1016/S0141-0229(99)00014-9

J.E. Haag, A. Vande Wouwer, I. Queinnec, Macroscopic modelling and identification of an anaerobic waste treatment process. Chemical Engineering Science **58**(19), 4307–4316 (2003). doi:10.1016/S0009-2509(03)00272-0

J. Karlsson, M. Anguelova, M. Jirstrand, An Effcient Method for Structural Identifiability Analysis of Large Dynamic Systems, in *Preprints of the 16th IFAC Symposium on System Identification*, IFAC, Brussels, Belgium, 2012, pp. 941–946. IFAC

L. Ljung, *System Identification: Theory for the User*, 2nd edn. (Prentice-Hall, Englewood Cliffs, NJ, 1999)

L. Ljung, T. Glad, On global identifiability for arbitrary model parametrizations. Automatica **30**(2), 265–276 (1994). doi:10.1016/0005-1098(94)90029-9

G. Margaria, E. Riccomagno, M.J. Chappell, M.J. Wynn, Differential algebra methods for the study of the structural identifiability of rational function state-space models in the biosciences. Mathematical Biosciences **174**(1), 1–26 (2001). doi:10.1016/S0025-5564(01)00079-7

E. Polak, An historical survey of computational methods in optimal control. SIAM review **15**(2), 553–584 (1973)

M. Sbarciog, M. Loccufier, E. Noldus, Determination of appropriate operating strategies for anaerobic digestion systems. Biochemical Engineering Journal **51**, 180–188 (2010)

A. Sedoglavic, A probabilistic algorithm to test local algebraic observability in polynomial time. Journal of Symbolic Computation **55**(5), 735–755 (2002)

J. Sjöberg, H. Hjalmarsson, L. Ljung, Neural networks in system identification, in *Proc. 10th IFAC Symposium on System Identification (SYSID'94)*, vol. 2, Copenhagen, Denmark, 1994, pp. 49–72