# On the estimation of the pseudo-stoichiometric matrix for macroscopic mass balance modelling of biotechnological processes

Olivier Bernard, Georges Bastin

# On the estimation of the pseudo-stoichiometric matrix for macroscopic mass balance modelling of biotechnological processes

Olivier Bernard [a,*], Georges Bastin [b]

[a] *INRIA-COMORE; BP93, 06902 Sophia-Antipolis Cedex, France*

[b] *UCL-CESAME, av. G. Lemaître 4-6, 1348 Louvain-La-Neuve, Belgium*

**Abstract**

In this paper we propose a methodology to determine the structure of the pseudo-stoichiometric coefficient matrix $\mathcal{K}$ in a macroscopic mass balance based model. The first step consists in estimating the minimal number of reactions that must be taken into account to represent the main mass transfer within the bioreactor. This provides the dimension of $\mathcal{K}$. Then we discuss the identifiability of the components of $\mathcal{K}$ and we propose a method to estimate their values. Finally we present a method to select among a set of possible macroscopic reaction networks those which are in agreement with the available measurements. These methods are illustrated with 3 examples: real data of the growth and biotransformation of the filamentous fungi *Pycnoporus cinnabarinus*, real data of an anaerobic digester involving a bacterial consortium degrading a mixture of organic substrates and a process of lipase production from olive oil by *Candida rugosa*.

*Key words:* Modelling, Nonlinear systems, Biotechnology, Identification, Identifiability, Validation.

## 1 Introduction and motivation

For a long time the macroscopic perspective of modelling of biotechnological processes has proved to be effi cient for solving many bioengineering problems [1,2]. In this perspective the total cell mass in the reactor is considered as a "black-box" (see e.g. [3], chapter 4) for the conversion of initial substrates into fi nal products. The

* Corresponding author

*Email addresses:* olivier.bernard@inria.fr (Olivier Bernard), bastin@auto.ucl.ac.be (Georges Bastin).

transformations are described by a small set of macroscopic (overall) reactions that lump together the many intracellular metabolic reactions of the various involved microbial species. Such macroscopic reaction networks represent the main mass transfers throughout the system and directly connect initial substrates to final products without describing the intracellular behaviour. On this basis, dynamical mass balance models can then be established. They rely on the category of the so-called "unstructured" models in the standard terminology reported e.g. in [4]. The goal of macroscopic modelling is clearly to derive simple dynamical models which have proved of great interest in bioengineering for the design of on-line algorithms for process monitoring, control and optimisation [2,5].

This paper clearly relies on this macroscopic perspective and our goal is to describe an approach for the determination of the minimal number of macroscopic reactions that should be involved in a mass balanced model in order to ensure constant pseudo-stoichiometric coefficients and to represent the main mass transfers within the system.

To motivate our objectives, let us consider a simple biotechnological process which could be represented by the simplest reaction network made of a single overall reaction :

$$\sum_{i=1}^{n_S} a_i S_i \longrightarrow X + \sum_{i=1}^{n_P} b_i P_i$$

where $S_i$, $P_j$ denote the substrates and products, $X$ is the total biomass (possibly made of multiple microbial species) and $a_i$, $b_i$ are the pseudo stoichiometric coefficients. In this case, as explained in [3], the pseudo-stoichiometric coefficients $a_i$ and $b_i$ are exactly the yield coefficients that can be directly computed from experimental data.

The main problem with a single overall reaction is obviously that it is often not able to describe accurately the process all along a transient operation with the same constant yield coefficients. A classical and efficient way to overcome this difficulty, without relying on complicated intracellular metabolic descriptions is to consider a network of some macroscopic reactions which involve only the initial substrates and the final products (without the intracellular species) but which are able to describe the process accurately with constant stoichiometric coefficients.

Whenever more than one macroscopic reaction are considered it should be emphasised that there is no longer an equivalence between the yield coefficients and the pseudo-stoichiometric coefficients.

We are thus interested in macroscopic reaction networks achieving a tradeoff between simplicity for process monitoring and control applications, and accuracy in order to match available experimental data.

In many applications, especially for cultures involving only a single or a few microbial species, it is clear that a detailed description of the metabolic and biosynthetic pathways may be available. In such a case, a macroscopic reaction network may be simply obtained by network reduction, i.e. by lumping or aggregating elementary metabolic reactions together (for instance by using the technique of "elementary flux modes" as described in [6]). It must however be noticed that the lumping of reactions corresponding to competing pathways may induce the appearance of pseudo-stoichiometric coefficients that are a-priori unspecified and have to be calibrated from the experimental data. Hence the issue of the pseudo-stoichiometric parameter estimation that we address in this paper may be a relevant issue even in the case where the metabolism is perfectly known.

The following simple illustration clarifies this point. Consider a culture of *E. coli* with anaerobic production of both lactate ($L$) and acetate ($A$) from glucose ($G$). It is obvious that there are two metabolic pathways that can be summarised by the two following reactions (stoichiometry in moles) :

$$G \longrightarrow 2L$$

$$G \longrightarrow 3A$$

These two reactions have perfectly known stoichiometric coefficients (2 and 3). Suppose now that these two reactions are aggregated into a single reaction. It is clear that any lumped reaction of the form:

$$G \longrightarrow 2\alpha L + 3(1-\alpha)A$$

is valid for any value $0 \leq \alpha \leq 1$. In fact $\alpha$ represents the fraction of glucose going through the first pathway and $(1-\alpha)$ the fraction going through the second pathway. The parameter $\alpha$ can be interpreted as a pseudo-stoichiometric coefficient which is *a priori* unspecified but can be calibrated from the data if measurements of the three species ($G, L, A$) in the culture medium are available. For instance, from the data reported in [7] with strain TC4 growing in anaerobic conditions we have $\alpha = 0.82$.

Depending on the assumptions made to lump the metabolic pathways, several macroscopic networks can be obtained. For example some pathways may or may not be neglected leading to different macroscopic reaction networks. In the previous simple example this would correspond e.g. to choosing $\alpha = 1$ and thus neglecting the acetate production. In section 4.2 we present a more complicated example where 3 possible lumped networks are *a priori* considered for the aerobic growth of the fungus *Pycnoporus cinnabarinus*.

It is worth noting that proposing a reduced macroscopic reaction network can be very difficult for some complex cases. For instance anaerobic wastewater treatment

bioreactors involve more than 140 coexisting microbial species [8] and many different complex substrates whose proportion varies with time.

Once the $n_r$ macroscopic reactions have been assumed involving $n_a$ biological or chemical species (microorganisms, substrates, metabolites, enzymes... ), the dynamical behaviour of the stirred tank bioreactor can be described by a general mass-balance model of the following form (see e.g. [2]):

$$\frac{dc_a}{dt} = \kappa\, r(c_a, d) + D(c_{ain} - c_a) - Q(c_a, d), \tag{1}$$

In this model, the vector $c_a = (c_{a1}, c_{a2}, \ldots, c_{an_a})^T$ is made-up of the concentrations of the various species inside the liquid medium. The term $c_{ain}$ represents the influent concentrations. The matrix $D$ is the dilution rate matrix representing the hydraulics mechanisms (inflows and outflows and possible retention) associated with the various species in the reactor. The exchange of matter in gaseous form between the surrounding and the reaction medium is represented by the gaseous flow rate $Q(c_a, d)$.

The term $\kappa\, r(c_a, d)$ represents the biological and biochemical conversions in the reactor (per unit of time) according to the underlying macroscopic reaction network. The $(n_a \times n_r)$ matrix $\kappa$ is a constant pseudo stoichiometric coeffi cient matrix. The term $r(c_a, d) = (r_1(c_a, d), r_2(c_a, d), \ldots, r_{n_r}(c_a, d))^T$ is a vector of reaction rates (or conversion rates). $Q(c_a, d)$ and $r(c_a, d)$ are supposed to depend on the state $c_a$ and on external environmental factors (represented by $d$) such as temperature, light, aeration rate, etc.

Matrix $\kappa$ is associated with the assumed macroscopic reaction network and plays a key role in the mass balance modelling. Each line of the matrix corresponds to one (bio)chemical species involved in the process. Each column of the matrix corresponds to a (bio) chemical reaction between some of the species. A positive entry $k_{ij}$ means that the $i^{\text{th}}$ species is a product of the $j^{\text{th}}$ reaction while a negative entry $k_{ij}$ means that it is a reactant or a substrate of the reaction. If $k_{ij} = 0$, the $i^{\text{th}}$ species is not involved in the $j^{\text{th}}$ reaction.

The objective of this paper is to propose a method to guide the user in the identifi - cation of the entries of pseudo-stoichiometric matrix $\kappa$. It is worth noting that the determination of matrix $\kappa$ is a problem equivalent to that of determining the structure of the reaction network. The usual approach dedicated to the determination of reaction networks relies on the linearisation of the dynamics around a reference solution [9,10]. Here, in the spirit of [11,12], we use linear algebraic properties to exploit the structure of the system (equation (1)) and our arguments are not based on any linearisation. As a consequence we are not limited to steady state data and we can exploit all the available measurements, even when associated to transient states.

We will show how to use a set of available data consisting of measurements of $c_a$, $Q$, $D$ and $c_{ain}$ at sampling instants, to determine the size of the matrix $\mathcal{K}$ (i.e. the number of reactions that must be taken into account) and to address the problem of the identification and validation of its coefficients.

Note that it is quite rare for bioprocesses that all the involved variables are measured (sometimes it is even unclear which variables are involved). For this reason we will focus on the estimation of $K$ the submatrix of $\mathcal{K}$ associated with the available measurements $c_m$.

We stress the fact that the methodology that we discuss is the first modelling stage. The second stage in the modelling, which is not discussed here, would consist in determining the reaction rates as functions of the state variables. This second problem is difficult and suffers as well from a lack of tools to assist the modeller. But this delicate step can be avoided for a large number of applications, where the knowledge of the mass balance (*i.e* matrix $\mathcal{K}$) is sufficient to design controllers or observers [2].

The paper will address the three following problems:

- How many reactions (i.e. how many columns for matrix $\mathcal{K}$) must be taken into account to reproduce the available data set ?
- Which reactions must be taken into account ? Which are the most plausible macroscopic reaction networks ?
- What are the values of the pseudo-stoichiometric coefficients ?

We will successively consider these three problems, without any *a priori* knowledge on the reaction rates $r(c_a, d)$. The approaches will be illustrated with three examples of significant complexity: real data of the growth and biotransformation of the filamentous fungi *Pycnoporus cinnabarinus*, real data of an anaerobic digester involving a bacterial consortium degrading a mixture of organic substrates and a process of lipase production from olive oil by *Candida rugosa*.

## 2 Determination of the minimum number of reactions

### *2.1 Statement of the problem*

In this section, we address the first problem, consisting in determining $n_r$ the minimum number of reactions to explain the observed dynamics of the fermenter. We assume that we measure a subset $c_m$ of $n_m$ components of $c_a$ that are involved in the system (*i.e.* which present significant variations along time). Indeed the measurements of the other state components (denoted $c_u$) may not be available, but

we assume however that we measure more variables than the number of reactions: $n_m > n_r$. If these components have a gaseous phase, we assume that the associated gaseous flow rates $Q(c_m, c_u, d)$ are measured.

The equation associated with $c_m$ is thus:

$$\frac{dc_m}{dt} = K\, r(c_m, c_u, d) + D(c_{min} - c_m) - Q(c_m, c_u, d), \tag{2}$$

The matrices $K$ and $Q$ are submatrices of $\kappa$ and $Q$, respectively, associated with $c_m$. Note now that $K$ is a rectangular matrix with more rows than columns. In the expression of the mass balance model (2), only the term $K\, r(c_m, c_u, d)$ needs to be mathematically expressed.

### *2.2 Theoretical determination of* $\dim(Im\ (\mathrm{K}))$

Let us integrate equation (2) between 2 time instants $t - T$ and $t$ ($T$ denotes the considered time window):

$$\begin{aligned} c_m(t) - c_m(t-T) - \textstyle\int_{t-T}^{t} D(c_{min}(\tau) - c_m(\tau)) + Q(c_m(\tau), d(\tau))d\tau \\ = K \textstyle\int_{t-T}^{t} r(c_a(\tau), d(\tau))d\tau \end{aligned} \tag{3}$$

Let us denote:

$$u_m(t) = c_m(t) - c_m(t-T) - \int_{t-T}^{t} D(c_{min}(\tau) - c_m(\tau)) + Q(c_m(\tau), d(\tau))d\tau \tag{4}$$

and

$$w_r(t) = \int_{t-T}^{t} r(c_a(\tau), d(\tau))d\tau$$

Equation (3) can then be rewritten:

$$u_m(t) = K\, w_r(t) \tag{5}$$

The vector $u_m(t)$ can be estimated along time from the available measurements. The value of the integral in (4) can be computed *e.g.* with a trapeze approximation.

In order to improve the cleaning of the data (noise reduction and diminution of autocorrelation) it may be useful to apply any linear scalar filter (*i.e.* any combination of integration, differentiation and delay $\theta$) to Equation (2) leading to a linear relationship of the same type as (5):

$$u(t) = K\, w(t) \tag{6}$$

where $u(t)$ and $w(t)$ denote respectively the signal derived from $u_m(t)$ and $w_r(t)$ after filtering. The moving average (3) that we have presented for sake of simplicity is of course only one example of such a filtering.

Now the question of the dimension of matrix $K$ can be formulated as the *determination of the dimension of the image of* $K$ or in other words, of the dimension of the space where $u(t)$ lives. Note that we assume $K$ to be a full rank matrix. Otherwise, it would mean that the same dynamical behaviour for $u(t)$ could be obtained with a matrix $K$ of lower dimension.

Determining the dimension of the $u(t)$ space is a classical problem in statistical analysis. It can be solved by a principal component analysis (see e.g. [13]) that determines the dimension of the vector space spanned by the vectors $k_i$, rows of $K$. In order to reach this objective, we consider $N$ time instants $t_1, \ldots, t_N$ (we choose $N > n_m$). The way to select these time instants will be discussed in the sequel. We build then the $n_m \times N$ matrix $U$ made of $N$ vectors $u(t)$ at these time instants:

$$U = (u(t_1), \ldots, u(t_N))$$

We will also consider the associated matrix of reaction rates, which is unknown:

$$W = (w(t_1), \ldots, w(t_N))$$

We assume that matrix $W$ has full rank. It means that the reactions are independent (none of the reaction rates can be written as a linear combination of the others).

**Property 1** *For a matrix $K$ of rank $n_r$, if $W$ has full rank, then the $n_m \times n_m$ matrix $M = UU^T = KWW^TK^T$ has rank $n_r$. Since it is a symmetric matrix, it can be written:*

$$M = P^T \Sigma P$$

*where $P$ is an orthogonal matrix ($P^T P = I$) and*

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & & \dots & 0 \\ 0 & \sigma_2 & 0 & & 0 \\ \vdots & & \ddots & & \\ & & \sigma_{n_r} & & \\ & & & 0 & \\ & & & & \ddots \ \vdots \\ 0 & & \dots & & 0 \end{pmatrix}$$

*with $\sigma_{i-1} \geq \sigma_i > 0$ for $i \in \{2, ..., n_r\}$.*

This property is a direct application of the singular decomposition theorem [14]. Since $\mathrm{rank}\,(M) = \mathrm{rank}\,(KW) = \mathrm{rank}\,(K) = \mathrm{rank}\,(\Sigma) = n_r$, it provides the result.

Now from a theoretical point of view it is possible to determine the number of reactions in the macroscopic reaction network: it corresponds to the number of non zero singular values of $UU^T$. This theoretical approach must however be adapted in the real case where the available measurements are discrete data points perturbed by a noise.

### *2.3 Practical determination of the number of reactions*

In the reality, the ideal case presented in the previous paragraph is perturbed for four main reasons:

- The macroscopic reaction network that we are looking for results from lumping of chemical or biochemical reactions which can be very complex. The "true" matrix $K$ is probably a very large matrix. The reactions which are fast or of low magnitude can be considered as perturbations of a dominant reaction network. It is this central (perturbed) macroscopic reaction network that we want to estimate.
- The measurements are corrupted by noise. This noise can be very important, especially for the measurement of biological quantities for which reliable sensors are rarely available.
- The measurements are performed on a discrete basis. Moreover they are rarely all available exactly at the same time instant $t_i$, and therefore they must be interpolated if we need a state estimate $c_m(t_i)$ at $N$ time instants $t_i$ in order to build vector $U$.

- In order to estimate $u(t)$ in equation (4) we need to compute the approximate value of an integral. This may generate additional perturbations.

### *2.3.1 Data processing: interpolation and smoothing*

The data collected on a biotechnological process often result from various sampling strategies carried out with various devices. As a consequence the data are seldom sampled simultaneously. In order to apply the proposed transformations vector $U$ has to be computed with values of the state variables at the same time instants $t_i$. A large number of tools are available in the literature to interpolate and smooth the data. We suggest here to use spline functions [15] which will at the same time interpolate and smooth a signal. The trade-off between interpolation and smoothing can be chosen by the user.

In the sequel we assume therefore that the set of measurements is available at the (irregular) time instants $\tau_j$ (depending on the variable), and that after a smoothing and interpolation process all the variable estimates are available at the time instants $t_i$.

We hypothesise that the estimates $c_m(t_i)$ are of reasonably good quality and in particular that the sampling frequency is in adequation with the time constants of the process.

### *2.3.2 Data normalisation*

To avoid conditioning problems and to give the same weighting to all the state variables, we normalise each component $u_i$ of the vector $u$ as follows:

$$\tilde{u}_i(t_j) = \frac{u_i(t_j) - a(u_i)}{\sqrt{N} s(u_i)}$$

where $a(u_i)$ is the average value of the $u_i(t_k)$ for $k \in \{1..N\}$, and $s(u_i)$ their standard deviation.

### *2.3.3 Conclusion for the determination of the minimal number of reactions*

In the reality, the noises due to model approximations, measurement errors or interpolation perturb the analysis. Therefore in practice there are no zero eigenvalues for the matrix $M = UU^T$.

The question is then to determine the number of eigenvectors that must be taken into account in order to represent a reasonable approximation of the data $u(t)$. To

solve this problem, let us remark that the eigenvalues $\sigma_i$ of $M$ correspond to the variance associated with the corresponding eigenvector (inertia axis) [13].

The method will then consist in selecting the $n_r$ first principal axis which represent a total variance larger than a fixed threshold.

For instance, in the next example, we have fixed a threshold (depending on the information available on noise measurements) at 95% of the variance. This led to the selection of 6 axes, and therefore $n_r = 5$.

**Remark:** if $\mathrm{rank}\,(M) = n_m$ it means that $\mathrm{rank}\,(K) \geq n_m$. In such a case we cannot estimate $n_r$ and measurements of additional variables are requested to apply the proposed method.

Finally, Figure 1 summarises the full procedure to compute the minimal number of reactions that are to be considered in order to reproduce an experimental data set.

### *2.4 Example 1: vanillin production by the filamentous fungi Pycnoporus cinnabarinus*

We consider here the production of vanillin from vanillic acid by the filamentous fungus *Pycnoporus cinnabarinus* [16]. The species involved in this biotransformation process are the carbon sources (maltose and glucose), the nitrogen source (ammonium), oxygen, carbon dioxide, fungal biomass and phenolic compounds (vanillic acid, vanillin, vanillic alcohol and methoxyhydroquinone). This results from a complicated set of reactions [17], most of them being ill known.

The process generally proceeds in two steps. In a first step (which generally lasts the first 3 days), the fungus uses the available substrates (nitrogen, maltose and glucose) to grow. The growth is aerobic, and therefore oxygen is consumed and $CO_2$ produced.

In a second step, the biosynthesis is triggered with addition of cellobiose 2 hours before continuous addition of vanillic acid. Then the fungus transforms the vanillic acid either in methoxyhydroquinone, or in vanillin. In this last case, vanillin can also be degraded into vanillic alcohol.

For illustration purpose, Figure 2 presents the typical evolution of some of the key variables during the fermentation. The figure presents also the splines used to smooth and interpolate the data in order to build the vector $u(t)$ made of the 10 measured species. The data set consists in 9 experiments which have been resampled to get 4 time instant $t_i$ per day. Finally, 619 data points $u(t_i)$ were considered.

Figure 3 represents the cumulated variance associated with the number of reactions.

Four reactions are sufficient to explain 80% of the observed variance. Five reactions explain 95% of the total variance. This analysis motivated the structure of the model presented in [18].

### *2.5 Example 2: an anaerobic digestion process for wastewater treatment*

In this second example we study a more complicated case where the biotechnological process is an anaerobic digester used for wastewater treatment [1]. The anaerobic process involves a complex consortium of bacteria degrading a mixture of substrates. Figure 4 presents a schematic overview of the degradation pathway from the set of macromolecules (proteins, carbohydrates, lipids, etc...) up to methane, carbon dioxide and hydrogen. Obviously this general overview lumps together a large number of simpler reactions involving single substrates. It turns out that more than 140 bacterial species can be found in the considered anaerobic digester [8].

The considered process for experimental data generation is a pilot-scale up-flow anaerobic fixed bed reactor processing raw industrial distillery wastewater [19]. The experimental results [20] were obtained for a period of 70 days over a wide range of experimental conditions (see Figure 7). The available daily measurements consist in organic carbon measured by the so-called soluble chemical oxygen demand (COD), the total volatile fatty acids (VFA) the volatile suspended solids (VSS), the total alkalinity and the dissolved inorganic carbon. The data set also contains a series of measurements of $CH_4$ and $CO_2$ flow rates.

The proposed method was applied to the available data set, including periods of biomass inhibition by an excess of VFA [20]. The obtained variance distribution is represented in Figure 5. It is worth noting that despite the apparent complexity of the process, a reaction network involving only 1 reaction (and thus one biomass) represents 83.2% of the variability. With 2 reactions, 97.8% of the variability are represented, which justified the choice of the very simple model presented in [20].

## 3 Validation of a macroscopic reaction network

### *3.1 Statement of the problem*

In the previous section we have shown how to determine the number of reactions which must be considered in order to explain the available data. Let us now assume that a plausible reaction network, with this number of reactions, is postulated with the aim of describing the process. In this section, we shall now show how such a candidate reaction network can be validated from the data.

One additional difficulty in comparing a reaction network, via its stoichiometric matrix $K$, with a set of data is that some pseudo-stoichiometric coefficients may be *a priori* unknown. We shall propose a method which will allow at the same time to test the validity of the macroscopic reaction network and to identify the missing pseudo-stoichiometric coefficients.

### *3.2 Finding the left kernel of the pseudo-stoichiometric matrix*

Let us consider a vector $\lambda \in Ker\ K^T$:

$$\lambda^T K = 0$$

Assume moreover that $\lambda$ is normalised such that one of its components $\lambda_i$ is 1: $\lambda_{i_0} = 1$

Now let us consider the scalar quantity $\lambda^T u(t)$. From equation (5), it satisfies at any time $t$:

$$\lambda^T u(t) = 0$$

In other words, we have:

$$u_{i_0}(t) = -\sum_{j \neq i_0} \lambda_j u_j(t) \tag{7}$$

This means that the $u_j$ are linked by a linear relation. The immediate idea that one can have is to check whether relationship (7) is in adequation with the data. This can be done by performing a linear regression between $u_{i_0}$ and the $u_j$.

Nevertheless, we have to keep in mind that the $u_j$ are *a priori* not independent, since they may be related by other relationships associated with other left kernel vectors of $K$. In particular, we have seen that $\mathrm{rank}\ (U) = n_r$, and thus a regression (7) cannot involve more than $n_r + 1$ independent terms $u_j$.

We will therefore select the vectors $\lambda$ of the left kernel that imply only independent $u_j$ in (7).

It is worth noting that the vector $\lambda$ involves three kinds of components:

(1) entries which are structurally zero
(2) entries that have an *a priori* known non-zero value (either 1 for the normalising component, see above, or a known value related to the stoichiometry of the reaction network).

(3) entries which are unknown because they depend on unknown coeffi cients of the pseudo-stoichiometric matrix. These entries have to be estimated from the data.

Remark that Equation (7) states a conservation between the variables $u_i$. This conservation is directly connected to the notion of reaction invariants [21,22].

**Definition 1** *We say that a set* $\{u_{i_1} \dots u_{i_k}\}$ *is associated with a left kernel vector* $\lambda$ *if* $\lambda_j = 0$ *for all the indices* $j \notin \{i_1 \dots i_k\}$. *We say that* $\lambda$ *is associated with the* $k \times n_r$ *submatrix* $\tilde{K}$ *which is the submatrix made of the rows* $i_1 \dots i_k$ *of* $K$. *Finally we call* $\tilde{\lambda}$ *the vector obtained by removing all the zeros entries in* $\lambda$. *The dimension of* $\tilde{\lambda}$ *(namely* $k$*) is called the regression dimension associated with* $\lambda$ *and denoted* $d(\lambda)$, *the number of unknown components of* $\lambda$ *is denoted* $d_u(\lambda)$.

We have therefore $\tilde{\lambda}^T \tilde{K} = 0$, and $\tilde{\lambda}$ has no zero component. Then, $\sum_{i_k} \lambda_{i_k} u_{i_k}(t) = 0$.

Note that, due to the normalisation of $\lambda$, we have $d_u(\lambda) \leq d(\lambda) - 1$.

**Definition 2** *We say that a left kernel vector* $\lambda$ *is* **sound** *if its associated* $d(\lambda) \times n_r$ *matrix* $\tilde{K}$ *does not contain itself any* $k \times n_r$ *submatrix (*$k < d(\lambda)$*) whose rank is not full or — equivalently — if* $\dim(Ker\ \tilde{K}^T) = 1$.

**Remark:** For a sound vector $\lambda$ we have $d(\lambda) \leq n_r + 1$

Indeed, if it has $k \geq n_r + 2$ non zero entries, then its associated submatrix $\tilde{K}$ is a $k \times n_r$ submatrix whose left kernel is at least of dimension 2.

### *3.3 Example 3: a process of lipase production from olive oil by Candida rugosa*

Let us consider the example of the competitive growth on two substrates [23] which could represent the production of lipase from olive oil by *Candida rugosa*. Here the microorganism is supposed to grow on two substrates that are produced by the hydrolysis of the primary organic substrate made of several molecules (mainly triglycerides). We assume the following 3-step reaction network:

- Hydrolysis:

$$k_1 S_1 \ + \ E \ \longrightarrow \ S_2 \ + \ k_2 S_3 \ + E$$

- Growth on $S_2$:

$$k_3 S_2 \ + \ k_4 O \ \longrightarrow \ X \ + \ k_5 P$$

- Growth on $S_3$:

$$k_6 S_3 \ + \ k_8 O \ \longrightarrow \ X \ + k_7 E \ + \ k_9 P$$

where $S_1$ is the primary substrate (olive oil, made of various compounds), $S_2$ (glycerol) and $S_3$ (fatty acids) are the secondary substrates, $E$ is the enzyme (lipase), $X$ the biomass (*Candida rugosa*), $O$ the dissolved oxygen and $P$ the dissolved $CO_2$. We assume that all the biochemical species are measured, except $S_1$.

The associated pseudo-stoichiometric matrix $\mathcal{K}$ and the state vector $c_a$ are therefore:

$$\mathcal{K} = \begin{pmatrix} -k_1 & 0 & 0 \\ 1 & -k_3 & 0 \\ k_2 & 0 & -k_6 \\ 0 & 0 & k_7 \\ 0 & 1 & 1 \\ 0 & -k_4 & -k_8 \\ 0 & k_5 & k_9 \end{pmatrix}, \quad c_a = \begin{pmatrix} S_1 \\ S_2 \\ S_3 \\ E \\ X \\ O \\ P \end{pmatrix}$$

Since $S_1$ is not measured, we will focus on the state $c_m$ associated with the submatrix $K$:

$$K = \begin{pmatrix} 1 & -k_3 & 0 \\ k_2 & 0 & -k_6 \\ 0 & 0 & k_7 \\ 0 & 1 & 1 \\ 0 & -k_4 & -k_8 \\ 0 & k_5 & k_9 \end{pmatrix}, \quad c_m = \begin{pmatrix} S_2 \\ S_3 \\ E \\ X \\ O \\ P \end{pmatrix}, \quad c_u = \begin{pmatrix} S_1 \end{pmatrix}$$

Now the following vector belongs to the kernel of matrix $K^T$:

$$\lambda^1 = \begin{pmatrix} 0 \\ 0 \\ \frac{k_5 - k_9}{k_7} \\ -k_5 \\ 0 \\ 1 \end{pmatrix}$$

We have $d(\lambda^1) = 3$ and $d_u(\lambda^1) = 2$. It is associated with the rank-2 submatrix $\tilde{K}_1$ and to the vector $\tilde{\lambda}^1$:

$$\tilde{K}_1 = \begin{pmatrix} 0 & 0 & k_7 \\ 0 & 1 & 1 \\ 0 & k_5 & k_9 \end{pmatrix}, \quad \tilde{\lambda}_1 = \begin{pmatrix} \frac{k_5 - k_9}{k_7} \\ -k_5 \\ 1 \end{pmatrix}$$

which is sound since the 3 possible $2 \times 3$ submatrices are of full rank.

Thus $u_4, u_5$ and $u_7$ are associated with $\lambda^1$, and related by the following relation:

$$u_7(t) = k_5 u_5(t) + \frac{k_9 - k_5}{k_7} u_4(t) \tag{8}$$

Now the kernel of matrix $K^T$ is spanned by the 2 other sound vectors:

$$\lambda^2 = \begin{pmatrix} 0 \\ 0 \\ \frac{k_8 - k_4}{k_7} \\ k_4 \\ 1 \\ 0 \end{pmatrix}, \quad \lambda^3 = \begin{pmatrix} -k_2 \\ 1 \\ \frac{k_3 k_2 + k_6}{k_7} \\ -k_3 k_2 \\ 0 \\ 0 \end{pmatrix}$$

Obviously, we have $d(\lambda^2) = 3$, $d_u(\lambda^2) = 2$, $d(\lambda^3) = 4$, $d_u(\lambda^3) = 3$,

### *3.4 Regressions associated with a set of $u_i$*

**Property 2** *A vector $\lambda$ associated with a set $\{u_{i_1} \dots u_{i_k}\}$ is sound if and only if the $u_i$ are not related by any other linear relation.*

**Proof:** Indeed, it is clear that it is not possible to have another relation between $n_r + 1$ different $u_i$, otherwise this relation would be associated with a second kernel vector $\lambda'$, meaning then that the kernel of $\tilde{K}^T$ is at least of dimension 2, and thus $\lambda$ would not be sound.

**Property 3** *Let us consider a sound kernel vector $\lambda$ of $K^T$, associated with $\tilde{\lambda}$ and to a set $\{u_{i_j},\ i_j \in \{i_1..i_{d(\lambda)}\}\}$. Moreover, let us denote by S the set of indices $j$ such that $\tilde{\lambda}_j$ is known. Then the following cost criterion:*

$$J(\alpha) = \sum_{t=t_1}^{t_N} (\sum_{j \in S} \tilde{\lambda}_j u_{i_j}(t) - \sum_{j \notin S} \alpha_j u_{i_j}(t))^2 \tag{9}$$

*admits a unique minimum, of zero value, obtained for $\alpha_j = \tilde{\lambda}_j$ (for any $j \notin S$).*

It is worth noting that minimising $J(\alpha)$ is exactly a linear regression problem.

### *3.5 Validation of the kernel of $K^T$ with the available data*

Now the validation will consist in verifying that $J(\alpha)$ (Equation 9) can be correctly minimised, or in other words, that the regression between $v = \sum_{j \in S} \tilde{\lambda}_j u_{i_j}$ and the $u_{i_j}$ ($j \notin S$) is signifi cant.

This analysis must be performed on all the sound kernel vectors $\lambda^i$ of $K^T$. In order to maximise the quality of the regression, the $u_{i_j}$ associated with $\lambda^i$ ($j \notin S$) and $v$ should in practice span a space of dimension $d_u(\lambda^i)$. So we perform a principal component analysis for the matrix

$$U_i = \begin{pmatrix} v(t_1) & \dots & v(t_N) \\ u_{j_1}(t_1) & \dots & u_{j_1}(t_N) \\ & \vdots & \\ u_{j_{d_u(\lambda^i)}}(t_1) & \dots & u_{j_{d_u(\lambda^i)}}(t_N) \end{pmatrix}$$

where the index $j_i$ correspond to the unknown elements of $\tilde{\lambda}^i$. The eigenvalues of $U_i U_i^T$ represent the total variance $\sigma_{ij}$ associated with the $jth$ principal axis. We then sort the singular values so that $\sigma_1 \geq \ldots \geq \sigma_{d_u(\lambda^i)} \geq \sigma_{d_u(\lambda^i)+1}$. Let us recall that, in principle, $\sigma_{d_u(\lambda^i)} > 0$ and $\sigma_{d_u(\lambda^i)+1} = 0$.

We consider the following criterion (reminiscent to the conditioning number) which assesses the balance of the variance along the axis:

$$B(\lambda^i) = \frac{\sigma_1(\lambda^i)}{\sigma_{d_u}(\lambda^i)}$$

With this criterion, we can now order the kernel vectors as follows:

- We first sort the kernel vectors $\lambda^i$ by sets of constant regression dimension $d_u(\lambda^i)$.
- Within the sets of constant regression dimension $d_u(\lambda^i)$, we sort the $\lambda^i$ by increasing index of associated variance balance $B(\lambda^i)$.

**Definition 3** *The basis made of the first $n_m - n_r$ independent vectors $\lambda^i$ is called the sound kernel basis.*

### *3.6 Identifiability of the pseudo-stoichiometric coefficients*

The question that we want to discuss in this section is to determine whether it is possible to determine the set of pseudo-stoichiometric coefficients $k_i$ from the values of $\lambda_i$ identified from the set of regressions given by equations (7). This identifiability property when the reaction rates $r(c_a, d)$ are unknown is referred to as C-identifiability in [11].

The answer to the C-identifiability question can be found in [11]. A version of this Theorem is recalled here in the considered framework of full rank matrices $K$:

**Theorem 1 (Chen & Bastin 1995)** *Let $K$ be an $n_m \times n_r$ full rank matrix with $n_m > n_r$. The unknown elements of the $j^{\text{th}}$ column of $K$ are C-identifiable if and only if there exists a nonsingular partition $(K_a,\ K_b)$, where $K_a$ is a full rank submatrix $n_r \times n_r$ which does not contain any unknown element in its $j^{\text{th}}$ column.*

We propose here a broader sufficient condition for the C-identifiability:

**Theorem 2** *Let $K$ be an $n_m \times n_r$ full rank matrix with $n_m > n_r$. The unknown element $k_{ij}$ of $K$ is identifiable if there exists a $k \times n_r$ full rank submatrix $K_a$, with $k \leq n_r$, which does not contain any unknown element on its $j^{\text{th}}$ column such that*

*the* $(k+1) \times n_r$ *submatrix of* $K$*:*

$$R = \begin{pmatrix} K_a \\ K_{bi} \end{pmatrix}$$

*is not full rank, i.e.* $\mathrm{rank}\ (R) < k+1$*, where* $K_{bi}$ *is the* $i^{\mathrm{th}}$ *line of* $K$*.*

**Proof:** see Appendix A.

**Remark:** This criterion, although it is more complicated than the one proposed in [11], allows to check the C-identifiability for each element of $K$ separately and not only for the columns.

Let us consider the following matrix $K$:

$$K = \begin{pmatrix} k_{11} & 1 \\ 1 & 0 \\ k_{31} & 0 \end{pmatrix} \tag{10}$$

Theorem 1 states that the first column of $K$ is not C-identifiable, since it is not possible to find a $2 \times 2$ submatrix $K_a$ which do not contain any unknown element in its first column. Now if we apply Theorem 2, we can use the following submatrices:

$$K_a = \begin{pmatrix} 1 & 0 \end{pmatrix},\ K_b = \begin{pmatrix} k_{31} & 0 \end{pmatrix}$$

Then $R$ is of rank 1, and verifies the condition $k+1 = 2 > \mathrm{rank}\ (R)$, it follows that $k_{31}$ is C-identifiable. It is now clear that $k_{11}$ is not C-identifiable, otherwise the first column of $K$ would be C-identifiable.

Remark that the analysis of the kernel of matrix $K^T$ also provides a criterion to test the identifiability of the $k_{ij}$. Even if this criterion is less convenient, it will give some hints on the practical identifiability, as we will see in the next Property.

**Property 4** *The pseudo-stoichiometric coefficient* $k_{ij}$ *is C-identifiable if and only if it can be computed from a combination of sound kernel vectors.*

In the previous example of equation (10), the sound kernel basis of $K^T$ was

$$\tilde{\lambda} = (0,\ -k_{31},\ 1)^T$$

It follows that $k_{31}$ is C-identifiable and that $k_{11}$ is not C-identifiable.

### *3.7 Identification of the pseudo-stoichiometric coefficients and final validation*

Now, once we know that the pseudo-stoichiometric coefficients are identifiable, we can estimate their value from experimental data using Property 4. For this, we will use the regression associated with the sound kernel basis of $K^T$ given by equation (7). The statistical significance of the correlation will allow to test from the data whether the vectors $\tilde{\lambda}^i$ are in the kernel of $K^T$ or not.

The final validation will consist in checking that the pseudo-stoichiometric coefficients are all positive. This test must be performed with regards to the uncertainty obtained from the linear regression (7). Indeed, because of the uncertainty obtained on the estimates for the $\lambda_i$, the $k_i$ may have a negative value, but with a confidence interval intersecting the positive domain.

The overall approach leading to the validation of an assumed macroscopic reaction network is summarised in Figure 6.

### *3.8 Improving the method*

A fermentation is often composed of several phases. In each phase, some reactions are not triggered. Therefore it is generally possible to find time intervals $]T_k, T_{k+1}[$ for which $r_j = 0$ for some $j$. In the same way, the concentration of some components may remain constant during certain periods of time.

This is for example the case in a reaction where the primary (associated with growth) and the secondary metabolisms are successively activated. During the first stage only growth takes place: no biotransformation appears since no precursor is added. During the secondary metabolism phase, the growth is inhibited and the microorganism concentrates on the bioproduction of a metabolite.

During these periods of time $]T_k, T_{k+1}[$, the system is then characterised by an index $j_0$ such that $r_{j_0} = 0$. System (1) is then equivalent to the following system:

$$\frac{dc_m}{dt} \;=\; \bar{K}\,\bar{r}(c_m, d) \;+ D(c_{min} - c_m) - Q(c_m, d), \tag{11}$$

where the matrix $\bar{K}$ is extracted from $K$ by removing the columns of $K$ corresponding to the index $j_0$.

Finally on these time intervals, the study of system (1) can be simplified by studying (11).

### *3.9 Application to the process of lipase production*

Let us consider the example of the competitive growth on 2 substrates. Let us assume that substrates $S_2$ and $S_3$ can directly be obtained from another bioreactor where the enzyme has been purified and directly added to $S_1$ without the biomass. We will then consider such an experiment where the secondary substrates $S_2$ and $S_3$ are directly added. Therefore, for all these experiments we will have $r_1 = 0$. The problem reduces thus to find the kernel of the submatrix $\bar{K}$ obtained after removing the first column of $K$:

$$\bar{K} = \begin{pmatrix} -k_3 & 0 \\ 0 & -k_6 \\ 0 & k_7 \\ 1 & 1 \\ -k_4 & -k_8 \\ k_5 & k_9 \end{pmatrix}$$

The kernel of $\bar{K}^T$ is spanned by the following sound vectors:

$$\bar{\lambda}^1 = \begin{pmatrix} 0 \\ \frac{k_7}{k_6} \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \ \bar{\lambda}^2 = \begin{pmatrix} \frac{k_8 - k_4}{k_3} \\ 0 \\ 0 \\ k_8 \\ 1 \\ 0 \end{pmatrix}, \ \bar{\lambda}^3 = \begin{pmatrix} -\frac{k_7}{k_3} \\ 0 \\ 1 \\ -k_7 \\ 0 \\ 0 \end{pmatrix}, \ \bar{\lambda}^4 = \begin{pmatrix} \frac{k_5 - k_9}{k_3} \\ 0 \\ 0 \\ -k_9 \\ 0 \\ 1 \end{pmatrix}$$

The regression dimension are $d(\bar{\lambda}^1) = 2, \ d_u(\bar{\lambda}^1) = 1$ and $d(\bar{\lambda}^i) = 3, \ d_u(\bar{\lambda}^i) = 2$ for $i > 1$. Note that $\bar{K}$ is associated with regressions of lower dimension than $K$ implying less unknown coefficients $\lambda_i^j$. It will therefore provide more reliable results (with the same amount of data), which will be easier to validate.

### *3.10 Application to the anaerobic digestion process*

Based on the results presented in paragraph 2.5, a reaction network relying on 2 main steps was assumed to summarise the main mass transfer within the digester. In the first reaction (acidification) the organic matter $S_1$ is degraded into VFA ($S_2$)

and $CO_2$ by a consortium of acidogenic bacteria ($X_1$). Then the VFA are degraded into $CH_4$ and $CO_2$ by methanogenic bacteria ($X_2$). The reaction network is thus the following:

- Acidogenesis (with reaction rate $r_1$):

$$k_1 S_1 \longrightarrow X_1 + k_2\, S_2 + k_4 \mathrm{CO}_2 \tag{12}$$

- Methanogenesis (with reaction rate $r_2$):

$$k_3 S_2 \longrightarrow X_2 + k_5 \mathrm{CO}_2 + k_6 \mathrm{CH}_4 \tag{13}$$

Let us denote by $C$ the total dissolved inorganic compounds (mainly $CO_2$ and bicarbonate). The associated pseudo-stoichiometric matrix $\kappa$ and the state vector $c_a$ are therefore:

$$\kappa = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -k_1 & 0 \\ k_2 & -k_3 \\ k_4 & k_5 \\ 0 & k_6 \end{pmatrix}, \quad c_a = \begin{pmatrix} X_1 \\ X_2 \\ S_1 \\ S_2 \\ C \\ CH_4 \end{pmatrix}$$

Since $X_1$ and $X_2$ (which represent a broad variety of species) are not measured, we will focus on the state $c_m$ associated with the submatrix $K$:

$$K = \begin{pmatrix} -k_1 & 0 \\ k_2 & -k_3 \\ k_4 & k_5 \\ 0 & k_6 \end{pmatrix}, \quad c_m = \begin{pmatrix} S_1 \\ S_2 \\ C \\ CH_4 \end{pmatrix}, \quad c_u = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

Theorem 1 shows that matrix $K$ is clearly not C-identifiable from the available data. Thus we normalised it, so that the rate of $S_1$ consumption and the rate of $CH_4$ production are now unitary, leading to matrix $\bar{K}$:

$$\bar{K} = \begin{pmatrix} -1 & 0 \\ \frac{k_2}{k_1} & -\frac{k_3}{k_6} \\ \frac{k_4}{k_1} & \frac{k_5}{k_6} \\ 0 & 1 \end{pmatrix}$$

Theorem 1 proves then that $\bar{K}$ is C-identifiable, choosing $K_a = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$. Thus the fractions $\frac{k_2}{k_1}$, $\frac{k_3}{k_6}$, $\frac{k_4}{k_1}$ and $\frac{k_5}{k_6}$ can be identified. To estimate the pseudo-stoichiometric coefficients $k_1$ to $k_6$ biomass measurements are required.

This mass balance modelling was used in [20]. The methane whose solubility is low was assumed to stay at low constant concentration, and the methane gaseous flow rate was assumed to be directly related to the methane bacterial production rate.

Finally, in [20] the kinetic rate modelling for $r_1$ and $r_2$ was performed by using a Monod type kinetics for the growth of acidogenic bacteria and Haldane kinetics for the methanogenesis. Then the kinetic parameter were estimated, leading to the results presented in Figure 7 for more than 70 days of experiments with various influent concentrations and various dilution rates. For more details see [20].

## 4 Comparison between several macroscopic reaction networks

### *4.1 Statement of the problem*

Once the number of reactions $n_r$ to be taken into account has been identified, the next step consists in selecting the set of reactions which are supposed to represent the main mass transfer in the fermenter. In general, several hypotheses can be stated with respect to the available knowledge.

We assume therefore that a set of $q$ plausible macroscopic reaction networks with $q$ associated pseudo-stoichiometric matrices $K_i$ are postulated by the user. It may *e.g.* be the result of automatic determination procedures, like those presented in [24,25]. The aim of this section is to determine how to select among these $q$ hypotheses those who provide a pseudo-stoichiometric matrix in agreement with the available data. Remark however that, in most cases, $q$ is a small number since there are only a few possible macroscopic reaction networks.

The method consists therefore in testing each matrix $K_i$ by using the methodology exposed in Section (3.7) and then to select the models which pass the validation tests.

The proposed methodology will be presented through a real life case study: the modelling of the growth of the filamentous fungus *Pycnoporus cinnabarinus*

We focus here on experimental phases were only aerobic growth of the fungus *Pycnoporus cinnabarinus* takes place. From a preliminary analysis of the available measurements, it turns out that 2 reactions are necessary to explain the observed data (representing 97% of the variance).

The aerobic growth of the fungal biomass ($X$) from a carbon source (glucose $G$ and maltose $M$) and a nitrogen source ($N$) can *a priori* be reasonably represented by the 3 following reactions networks:

- Network 1:
  The fungus is growing on maltose, glucose and nitrogen, and it can transform maltose into glucose in a first step:

$$\begin{aligned} & M \xrightarrow{r_1(.)} 2\,G \\ & k_1\,N \;+\; k_2\,G \;+\; k_3\,M \xrightarrow{r_2(.)} X \end{aligned} \tag{14}$$

- Network 2:
  The fungus is growing only on glucose and nitrogen, and it transforms maltose into glucose in a first step:

$$\begin{aligned} & M \xrightarrow{r_1(.)} 2\,G \\ & k_1\,N \;+\; k_2\,G \xrightarrow{r_2(.)} X \end{aligned} \tag{15}$$

- Network 3:
  The fungus can grow either on glucose and nitrogen or on maltose and nitrogen. In this second case glucose is produced.

$$\begin{aligned} & k_4\,N \;+\; k_5\,G \xrightarrow{r_1(.)} X \\ & k_1\,N \;+\; k_2\,M \xrightarrow{r_2(.)} X \;+\; k_3\,G \end{aligned} \tag{16}$$

The pseudo-stoichiometric matrices associated with (14), (15) and (16) are then respectively:

$$K_1 = \begin{pmatrix} 0 & -k_1 \\ -1 & -k_3 \\ 2 & -k_2 \\ 0 & 1 \end{pmatrix} \tag{17}$$

$$K_2 = \begin{pmatrix} 0 & -k_1 \\ -1 & 0 \\ 2 & -k_2 \\ 0 & 1 \end{pmatrix} \tag{18}$$

$$K_3 = \begin{pmatrix} -k_1 & -k_4 \\ -k_2 & 0 \\ k_3 & -k_5 \\ 1 & 1 \end{pmatrix} \tag{19}$$

Using the method presented in section (3.7) we give in Table 1 the sound kernel vectors and the corresponding regressions which are associated with these three pseudo-stoichiometric matrices.

| PS Matrix | Sound kernel basis of $K^T$ | Regressions | $B(\lambda^i)$ |
|---|---|---|---|
| $K_1$ | $\lambda_1^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ k_1 \end{pmatrix}$, $\lambda_2^1 = \begin{pmatrix} 0 \\ 2 \\ 1 \\ 2k_3 + k_2 \end{pmatrix}$ | $u_1 = -e_1^{1+} u_4$<br>$2u_2 + u_3 = -e_3^{1+} u_4$ | $B(\lambda_1^1) = 1$<br>$B(\lambda_2^1) = 1$ |
| $K_2$ | $\lambda_1^2 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ k_1 \end{pmatrix}$, $\lambda_2^2 = \begin{pmatrix} 0 \\ 2 \\ 1 \\ k_2 \end{pmatrix}$ | $u_1 = -e_1^{2+} u_4$<br>$2u_2 + u_3 = -e_3^{2+} u_4$ | $B(\lambda_1^2) = 1$<br>$B(\lambda_2^2) = 1$ |
| $K_3$ | $\lambda_1^3 = \begin{pmatrix} 0 \\ \frac{k_5+k_3}{k_2} \\ 1 \\ k_5 \end{pmatrix}$, $\lambda_2^3 = \begin{pmatrix} 1 \\ \frac{k_4-k_1}{k_2} \\ 0 \\ k_4 \end{pmatrix}$ | $u_3 = -e_1^{3+} u_2 - e_2^{3+} u_4$<br>$u_1 = e_3^3 u_2 - e_4^{3+} u_4$ | $B(\lambda_1^3) = 4.48$<br>$B(\lambda_2^3) = 6.29$ |

Table 1
Kernel vectors and regressions associated with the pseudo-stoichiometric matrices for each of the considered reaction network for the growth of *Pycnoporus cinnabarinus* on ammonium, maltose and glucose. The real $e_i^{j+}$ are positive, the $e_i^j$ can be of any sign.

The regression coeffi cients computed from 70 data points coming from 9 different experiments are presented in Table 2. The confi dence intervals for the parameters

have been estimated using a Student distribution with a 5% threshold and the significance of the regression has been tested.

| RN | Parameter | min | max | Positivity | Significance | Conclusion |
|---|---|---|---|---|---|---|
| 1 | $e_1^{1+}$ | 0.41 | 0.79 | YES | YES | $k_1 \in [0.41,\ 0.79]$ |
| | $e_3^{1+}$ | 1.4 | 1.77 | YES | YES | $k_2 + 2k_3 \in [1.4,\ 1.77]$ |
| 2 | $e_1^{2+}$ | 0.41 | 0.79 | YES | YES | $k_1 \in [0.41,\ 0.79]$ |
| | $e_3^{2+}$ | 1.4 | 1.78 | YES | YES | $k_2 \in [1.4,\ 1.78]$ |
| 3 | $e_1^{3+}$ | 0.72 | 1.1 | YES | YES | $\frac{k_5+k_3}{k_2} \in [0.72,\ 1.1]$ |
| | $e_2^{3+}$ | 1.40 | 1.78 | YES | YES | $k_5 \in [1.40,\ 1.78]$ |
| | $e_3^{3}$ | 0.93 | 1.28 | / | NO | $\frac{k_4-k_1}{k_2} \in [0.93,\ 1.28]$ |
| | $e_4^{3+}$ | -0.45 | -0.11 | NO | NO | $k_4 \in [-0.45,\ -0.11]$ |

Table 2
Estimation of intervals for parameter values and significance of the regressions (threshold 5%) associated with each of the reaction networks (RN).

From Table 2 we conclude immediately that network 3 is invalidated by the data. The coefficients associated with networks 1 and 2 have the correct signs, and therefore only these two networks are in agreement with the data and will be kept. Note that it is not possible to distinguish between network 1 and network 2. However the parameters $k_2$ and $k_3$ in network 1 are not identifiable, and thus network 2 would be preferred. Nevertheless, if network 1 should be kept for some reasons, the value (of at least one) of the (unidentifiable) parameters $k_2$ and $k_3$ should be selected, in such a way that $k_2 + 2k_3$ belongs to the confidence interval from Table 2.

## 5 Conclusion

Modelling of bioprocesses is known to be a difficult issue since there does not exist universal validated laws on which the model can rely as in other fields like mechanics (fundamental equations of mechanics), electronics (ohm law), etc.

We have presented here only the first stage of the macroscopic modelling, *i.e.* the mass balance modelling involving the reaction network through matrix $\mathcal{K}$. The second stage would now consist in estimating the reaction rates $r(c_m)$ with respect to the biochemical species in the system. This step is far from being trivial since the kinetics can be very sensitive to many factors, leading to high parametric uncertainties in the mathematical expressions. The reader can refer to [18,20] for details on this second step of kinetic determination, with example of model simulation and validation both for the vanillin production process (example 1) and for the anaerobic treatment plant (example 2).

The key point in the mass balance approach is to use linear algebra to uncouple the linear part of the model driven by matrix $\mathcal{K}$ from the nonlinear and unknown part of the model related to the microbial kinetics ($r(c_m)$). It is worth noting that some algorithms aiming at e.g. process monitoring or controlling can be based only on the mass balance part [2]. After algebraic operation the effect of the unknown $r(c_m)$ is eliminated, limiting the uncertainty associated with variability of the biological processes. However the resulting algorithms turn out to be very sensitive to the pseudo-stoichiometric matrix. Validation of this matrix and improvement of its identifi cation is therefore a key issue for biotechnological processes.

**Acknowledgement:** The authors are grateful to Benoit Chachuat for his comments and corrections. This work has been carried out with the support provided by the European commission, Information Society Technologies programme, Key action I Systems & Services for the Citizen, contract TELEMAC number IST-2000-28256. It also presents research results of the Belgian Programme on Inter-University Poles of Attraction initiated by the Belgian State, Prime Minister's office for Science, Technology and Culture. The scientific responsibility rests with its authors.

## References

[1] J. E. Bailey, D. F. Ollis, Biochemical engineering fundamentals, McGraw-Hill, 1986.

[2] G. Bastin, D. Dochain, On-line estimation and adaptive control of bioreactors, Elsevier, 1990.

[3] G. Stephanopoulos, A. Aristidou, J. Nielsen, Metabolic Engineering, Elsevier Science, 1998.

[4] J. E. Bailey, Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities, Biotechnology Progress 14 (1998) 8–20.

[5] G. Bastin, J. VanImpe, Nonlinear and adaptive control in biotechnology: a tutorial, European Journal of Control 1 (1) (1995) 1–37.

[6] A. Provost, G. Bastin, Dynamic metabolic modelling under the balanced growth condition, J. Process Contr. 14 (2004) 717–728.

[7] L. O. Ingram, T. Conway, Expression of different levels of ethanologenic enzymes from *zymomonas mobilis* in recombinant strains of *escherichia coli*, Appl. Environ. Microbiol. 54 (1988) 397–404.

[8] C. Delbès, R. Moletta, J.-J. Godon, Bacterial and archaeal 16s rdna and 16s rrna dynamics during an acetate crisis in an anaerobic digestor ecosystem, FEMS Microbiology Ecology 35 (2001) 19–26.

[9] M. Eiswirth, A. Freund, J. Ross, Mechanistic classification of chemical oscillators and the role of species, Vol. 80 of Advances in Chemical Physics, Wiley, New-York, 1991, Ch. 1, pp. 127–199.

[10] T. Chevalier, I. Schreiber, J. Ross, Toward a systematic determination of complex reaction mechanisms, J. Phys. Chem 97 (1993) 6776 – 6787.

[11] L. Chen, G. Bastin, Structural identifiability of the yield coefficients in bioprocess models when the reaction rates are unknown, Math. Biosciences 132 (1996) 35–67.

[12] O. Bernard, G. Bastin, Structural identification of nonlinear mathematical models for bioprocesses, in: Proceedings of the Nonlinear Control Systems Symposium, Enschede, July 1-3, 1998, pp. 449–454.

[13] R. Johnson, D. Wichern, Applied multivariate statistical analysis, 4th ed., Prentice-Hall, 1998.

[14] R. A. Horn, C. R. Johnson, Matrix analysis, Cambridge University Press, Cambridge MA, 1993.

[15] C. D. Boor, Applied mathematical sciences, in: Springer (Ed.), A Practical guide to splines, Wiley, New York, 1978, p. 392.

[16] B. Falconnier, C. Lapierre, L. Lesage-Meessen, G. Yonnet, P. Brunerie, B. Colonna-Ceccaldi, G. Corrieu, M. Asther, Vanillin as a product of ferulic acid biotransformation by the white-rot fungus *Pycnoporus cinnabarinus* I-937: identification of metabolic pathways, J. Biotechnol 37 (1994) 123–132.

[17] L. Lesage-Meessen, M. Delattre, M. Haon, J. Thibault, B. Colonna-Ceccaldi, P. Brunerie, M. Asther, A two-step bioconversion process for vanillin production from ferulic acid combining *Aspergillus niger* and *Pycnoporus cinnabarinus*, J. Biotechnol 50 (1996) 107–113.

[18] O. Bernard, G. Bastin, C. Stentelaire, L. Lesage-Meessen, M. Asther, Mass balance modelling of vanillin production from vanillic acid by cultures of the fungus *Pycnoporus cinnabarinus* in bioreactors, Biotech. Bioeng (1999) 558–571.

[19] J. P. Steyer, J. C. Bouvier, T. Conte, P. Gras, P. Sousbie, Evaluation of a four year experience with a fully instrumented anaerobic digestion process, Water Science and Technology 45 (2002) 495–502.

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

[21] M. Fjeld, On a pitfall in stability analysis of chemical reactions, Chem. Engin. Science 23 (1968) 565–573.

[22] O. Asbjornsen, M. Fjeld, Response modes of continuous stirred tank reactors, Chem. Engin. Science 25 (1970) 1627–1636.

[23] P. Serra, J. del Rio, J. Robust, M. Poch, C. Sola, A.Cheruy, A model for lipase production by *Candida rugosa*, Bioprocess Engineering 8 (1992) 145–150.

[24] P. Bogaerts, A. Vande Wouwer, Sytematic generation of identifiable macroscopic reaction schemes, in: Proceedings of the 8th IFAC Conference on Computer Applications in Biotechnology (CAB8), Montral, Canada, 2001.

[25] O. Bernard, G. Bastin, Identification of reaction schemes for bioprocesses: determination of an incompletely known yield matrix, in: Proceedings of ECC03, Cambridge, UK, 2003.

## Appendix A: proof of Theorem 2

We first demonstrate the following Property.

**Property 5** *Let us consider a set* $\{u_{i_1} \dots u_{i_k}\}$ *associated with a left kernel vector* $\lambda$*, and with a matrix* $\tilde{K}$*:* $\sum_{j=1}^{k} \lambda_{i_j} u_{i_j}(t) = 0$*. If* $\lambda$ *is not sound, then any submatrix* $\tilde{K}'$ *obtained from* $\tilde{K}$ *by removing the* $l^{th}$ *raw is associated with a subset of* $\{u_{i_1} \dots u_{i_k},\ i_j \neq l\}$*, i.e.:* $\sum_{\substack{j=1 \\ j\neq l}}^{k} \lambda'_{i_j} u_{i_j}(t) = 0$ *(for* $l \in \{i_1, \dots, i_k\}$*).*

**Proof:** if $\lambda$ is not sound, it means that $\dim(Ker\ \tilde{K}^T) > 1$. Therefore there exists at least 2 different vectors $\lambda^1$ and $\lambda^2$ such that $\sum_{j=1}^{k} \lambda^q_{i_j} u_{i_j}(t) = 0$ for $q \in \{1, 2\}$. If the $l^{th}$ component of $\lambda^1$ or $\lambda^2$ contains a zero, then we have the result. Otherwise, for $\lambda^1_l \lambda^2_l \neq 0$, we have

$$\sum_{j=1}^{k} \lambda^1_{i_j} u_{i_j}(t) - \frac{\lambda_{i_l}}{\lambda_{i_2}} \sum_{j=1}^{k} \lambda^2_{i_j} u_{i_j}(t) = 0$$

showing that the vector $\tilde{\lambda}'$ whose components are $\lambda^1_{i_j} - \frac{\lambda_{i_l}}{\lambda_{i_2}} \lambda^2_{i_j}$ for $i_j \in \{i_1, \dots, i_k\ \ i_j \neq l\}$ is associated with the matrix $\tilde{K}'$ obtained from $\tilde{K}$ by removing the $l^{th}$ raw.

Now we can prove Theorem 2.

**Proof of Theorem 2:**

If $k + 1 > \mathrm{rank}\,(R)$, then $\dim(Ker\ R^T) > 0$, there exists a kernel vector $\lambda = \begin{pmatrix} \lambda_a \\ \lambda_{bi} \end{pmatrix}$ such that $\lambda^T R = 0$.

We have therefore $\lambda_a^T K_a + \lambda_{bi} K_{bi} = 0$.

Since $K_a$ is a $k \times n_r$ full rank matrix with $k \leq n_r$, then $\dim Ker\ K_a^T = 0$, and thus $\lambda_{bi}$ cannot be zero.

If $\lambda$ is not sound *i.e.* $\dim Ker\ R^T > 1$. We must then consider the sound vector $\tilde{\lambda}$ associated with the submatrix $\begin{pmatrix} \tilde{K}_a \\ K_{bi} \end{pmatrix}$, where $\tilde{K}_a$ is extracted from $K_a$ according to Property 5.

The sound vector $\tilde{\lambda}$, verifies: $\tilde{\lambda}_a^T \tilde{K}_a + \lambda_{bi} K_{bi} = 0$

Let us remark that it is a matrix equality, and let us consider the $j^{\text{th}}$ column of this matrix equation:

$$\tilde{\lambda}_a^T \tilde{K}_{ai} + \lambda_{bi} k_{ij} = 0$$

where $\tilde{K}_{ai}$ is the $i^{\text{th}}$ column of $\tilde{K}_a$.

As we saw in Section 3.5, the coefficients of the sound kernel vector $\tilde{\lambda}$ can be identified from a linear regression. Therefore, $k_{ij}$ can be computed as follows:

$$k_{ij} = -\frac{\tilde{\lambda}_a^T \tilde{K}_{ai}}{\lambda_{bi}}$$

**Nomenclature**

| | |
|---|---|
| $a_i$ | yield coefficients for substrate |
| $b_i$ | yield coefficients for product |
| $c_a$ | vector of all the state variables |
| $c_{ain}$ | vector of influent concentrations |
| $c_m$ | vector of measured variables |
| $c_{min}$ | vector of influent concentrations |
| $c_u$ | vector of unmeasured variables |
| $d(\lambda)$ | number of components in $\lambda$ |
| $d_u(\lambda)$ | number of unknown components in $\lambda$ |
| $d$ | vector of external environmental factors |
| $D$ | dilution rate |
| $E$ | enzyme |
| $K$ | pseudo-stoichiometric matrix |
| $k_{ij}$ | entries of matrix $K$ |
| $n_r$ | number of reactions in the reaction network |
| $n_m$ | number of measured variables |
| $n_a$ | number of state variables |
| $P$ | product |
| $Q$ | vector of gaseous flow rates |
| $r(c_a, d)$ | vector of reaction rates |
| $S$ | substrate |
| $X$ | biomass |
| $\lambda$ | left kernel vector of $K$ |
| $Q$ | vector of gaseous flow rates |
| $\mathcal{K}$ | pseudo-stoichiometric matrix |

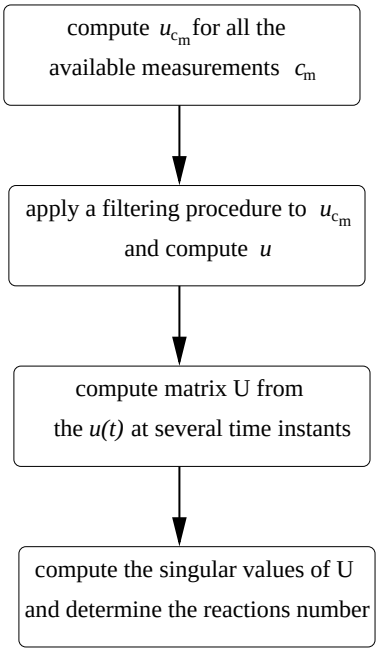


Fig. 1. Scheme of the procedure to compute the minimal number of reactions that are to be considered in order to reproduce an experimental data set.

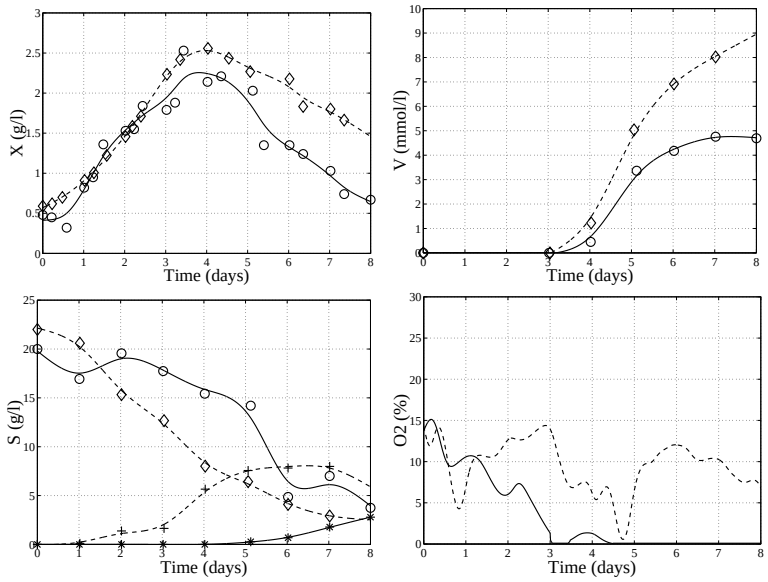


Fig. 2. Measurements of biomass ($X$), vanillin ($V$), maltose and glucose ($M$ and $G$) and oxygen ($O$) for two experiments of vanillin bioproduction by *P. cinnabarinus*. The continuous lines are the smoothing splines [18].

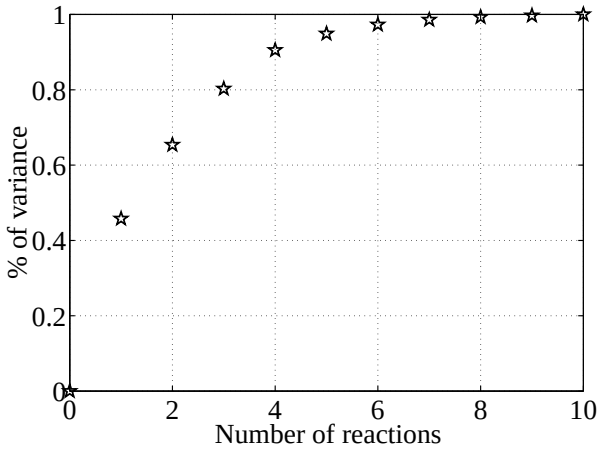


Fig. 3. Total variance explained with respect to the number of reactions for the process of vanillin bioproduction by the filamentous fungi *Pycnoporus cinnabarinus*.

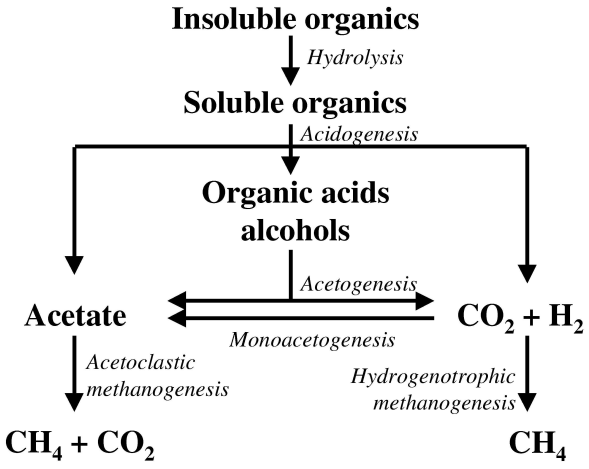


Fig. 4. Schematic overview of the anaerobic digestion reaction network.

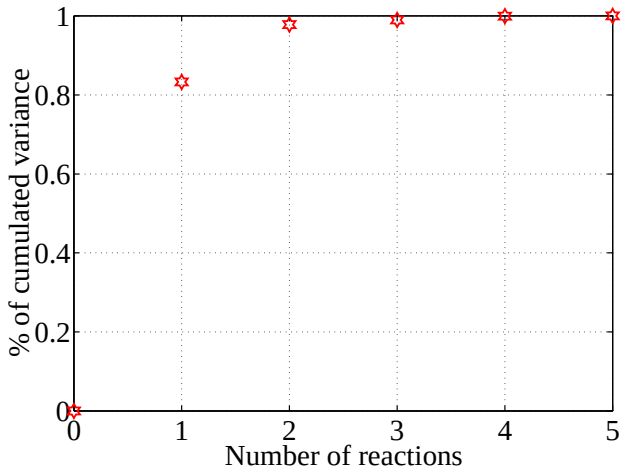


Fig. 5. Cumulated variance with respect to the number of reactions for 70 days of experiments (see [20]) on an anaerobic digester.

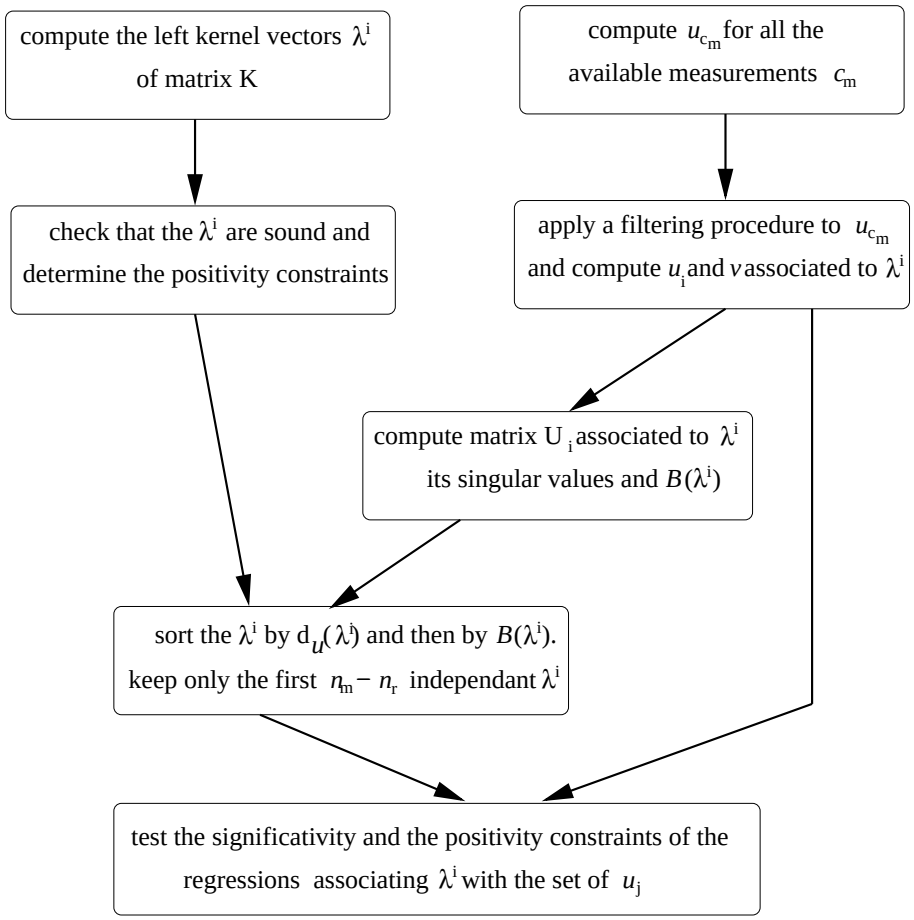


Fig. 6. Scheme of the procedure to validate a macroscopic reaction network described by matrix $K$.

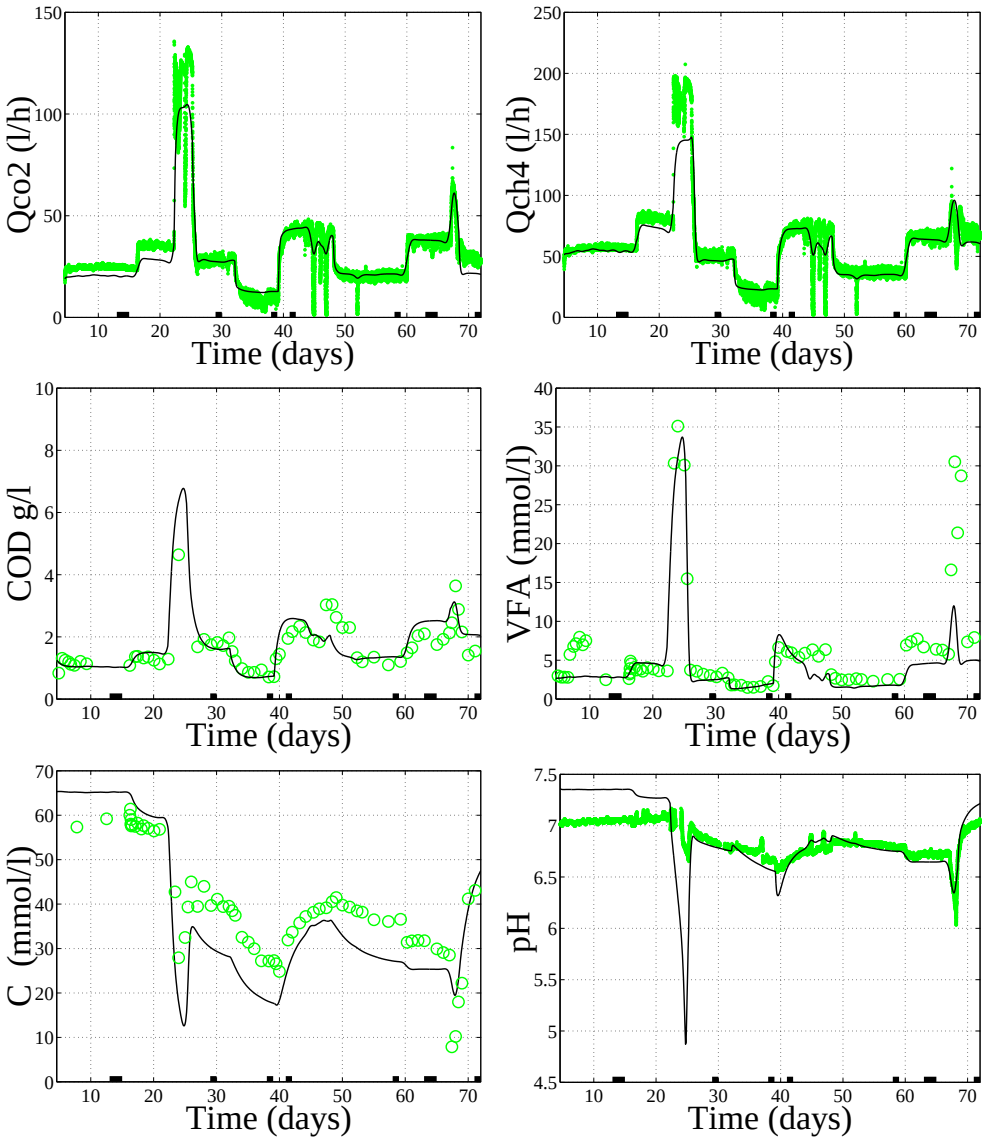


Fig. 7. Comparison between simulation results and measurements in a fixed bed anaerobic digester for the methane and $CO_2$ gaseous fbw rates, pH, COD, VFA and total inorganic carbon (from [20]).