

Structural Identifiability Analysis via Extended Observability and Decomposition

Alejandro F. Villaverde^{*,**,*} Antonio Barreiro^{**}
Antonis Papachristodoulou^{*}

^{*} *Department of Engineering Science, University of Oxford
Oxford OX1 3PJ, United Kingdom*

^{**} *Department of Systems & Control Engineering, University of Vigo
Vigo 36310, Spain*

^{***} *Current address: Bioprocess Engineering Group, IIM-CSIC
Vigo 36208, Spain (e-mail: afvillaverde@iim.csic.es)*

Abstract: Structural identifiability analysis of nonlinear dynamic models requires symbolic manipulations, whose computational cost rises very fast with problem size. This hampers the application of these techniques to the large models which are increasingly common in systems biology. Here we present a method to assess parametric identifiability based on the framework of nonlinear observability. Essentially, our method considers model parameters as particular cases of state variables with zero dynamics, and evaluates structural identifiability by calculating the rank of a generalized observability-identifiability matrix. If a model is unidentifiable as a whole, the method determines the identifiability of its individual parameters. For models whose size or complexity prevents the direct application of this procedure, an optimization approach is used to decompose them into tractable subsystems. We demonstrate the feasibility of this approach by applying it to three well-known case studies.

Keywords: structural identifiability, observability, parameter identification, model decomposition, optimization, systems biology

1. INTRODUCTION

Structural identifiability analysis determines whether the parameters in a model can be identified from knowledge of the system dynamics, observable functions, external stimuli, and initial conditions (Walter and Pronzato, 1997). The concept of structural identifiability was introduced by Bellman and Åström (1970). A wide variety of methods, reviewed e.g. in (Miao et al., 2011; Chiş et al., 2011b; Villaverde and Barreiro, 2016), have been proposed for assessing the structural identifiability of nonlinear dynamic models. We can distinguish between symbolic (Pohjanpalo, 1978; Walter and Lecourtier, 1982; Vajda et al., 1989; Ljung and Glad, 1994; Bellu et al., 2007; Balsa-Canto et al., 2010), semi-numerical (Sedoglavic, 2002; Karlsson et al., 2012; Stigter and Molenaar, 2016), and numerical approaches (Raue et al., 2009) for structural identifiability. Symbolic manipulations can provide exact (as opposed to probabilistic) results, and in some cases they can determine global (as opposed to local) identifiability. However, they quickly give rise to long expressions as the system size increases, and are hardly applicable to large-scale or even medium-size models (Miao et al., 2011; Chiş et al., 2011b; Grandjean et al., 2014). Numerical methods can be more efficient, although at the expense of generality. The two

approaches can be seen as complementary (Raue et al., 2014).

The existing methodologies have different strengths and weaknesses, and when choosing one it is necessary to take into account trade-offs between generality (some methods are only valid for certain types of systems, such as rational or polynomial expressions (Sedoglavic, 2002; Bellu et al., 2007; Karlsson et al., 2012; Merkt et al., 2015), while others are more generally applicable), computational cost (application of some methods to large-scale problems is infeasible), and level of detail of the results. For example, most methods not only classify the model as identifiable or unidentifiable as a whole, but also determine the identifiability of individual parameters. Additionally, other methods distinguish between local and global identifiability (Bellu et al., 2007; Chiş et al., 2011a), find symmetries between parameters (Yates et al., 2009; Merkt et al., 2015), or provide identifiable reparameterizations (Meshkat et al., 2014).

A conclusion from the aforementioned studies is that structural identifiability analysis is still a challenging task, particularly for large models. Hence it is seldom performed before undertaking parameter estimation, due to its complexity (Miao et al., 2011). In summary, despite recent advances there is still a need for structural identifiability methods that have the sufficient generality and efficiency to be applicable to the increasingly complex models being developed in the systems biology community.

* AFV acknowledges funding from the Galician government (Xunta de Galiza) through the I2C postdoctoral fellowship ED481B2014/133-0. This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 686282 (CanPathPro).

Here we present a method that approaches local structural identifiability as a generalized version of observability. A system is observable at time t_1 if it is possible to determine its state $x(t_1)$ from future measurements, that is, from $y(t)$ such that $t_1 < t < t_2$, where t_2 is a finite time. Tools for determining observability have been developed for both linear and nonlinear systems. If the model parameters p are considered as state variables with dynamics $\dot{p} = 0$, structural identifiability analysis can be recast as a generalization of observability analysis (Tunali and Tarn, 1987). By adopting this point of view it is possible to assess the structural identifiability of nonlinear systems of small and medium size (Chatzis et al., 2015). We show that larger models can be analysed by decomposing them into smaller and more tractable submodels, using an optimization algorithm. For models which are diagnosed as unidentifiable the method determines the subset of identifiable parameters. If all or some of the initial conditions of the state variables are unknown, their identifiability can also be assessed. The methodology proposed here is of broad applicability, since it is not restricted to systems in polynomial or rational form.

The structure of this paper is as follows: in section 2 we define mathematically the structural identifiability problem, and describe how its analysis can be recast as a generalization of nonlinear observability analysis. In section 3 we present a methodology based in this approach, and apply it to several case studies in section 4. Finally, in section 5 we provide some conclusions and guidelines for future work.

2. STRUCTURAL IDENTIFIABILITY AND OBSERVABILITY

2.1 Structural identifiability: definitions

Let us denote by M a general nonlinear model structure with the following dynamic equations:

$$\begin{aligned} \dot{x}(t) &= f[x(t, p), u(t), p] \\ y(t) &= g[x(t, p), p] \\ x_0 &= x(t_0, p) \end{aligned} \quad (1)$$

where f and g are nonlinear vector functions, $p \in \mathbb{R}^q$ is a real-valued vector of parameters, $u(t) \in \mathbb{R}^r$ is the input vector, $x(t) \in \mathbb{R}^n$ the state variable vector, and $y(t) \in \mathbb{R}^m$ the measurable output, also called observables vector.

Assuming that the model structure M is correct, that the data is noise-free, and that the inputs to the system can be chosen freely, it is always possible to choose an estimated parameter vector \hat{p} such that the model output $M(\hat{p})$ equals the one obtained by the true parameter vector, $M(p^*)$. If $\hat{p} = p^*$ this is obviously the case.

Parameter p_i is structurally globally (or uniquely) identifiable (s.g.i.) if, for almost any p^* ,

$$M(\hat{p}) = M(p^*) \Rightarrow \hat{p}_i = p_i^* \quad (2)$$

A model M is s.g.i. if all its parameters are s.g.i.

A parameter p_i is structurally locally identifiable (s.l.i.) if for almost any p^* there is a neighbourhood $V(p^*)$ such that

$$\hat{p} \in V(p^*) \text{ and } M(\hat{p}) = M(p^*) \Rightarrow \hat{p}_i = p_i^* \quad (3)$$

A model M is s.l.i. if all its parameters are s.l.i.

If equation (3) does not hold in any neighbourhood of p^* , parameter p_i is structurally unidentifiable (s.u.i.). A model M is s.u.i. if at least one of its parameters is s.u.i.

2.2 Nonlinear observability

Two states $x_0 \neq x_1$ are said to be distinguishable when there exists some input $u(t)$ such that $y(t, x_0, u(t)) \neq y(t, x_1, u(t))$, where $y(t, x_i, u(t))$ denotes the output function of the system for the input $u(t)$ and initial state $x_i (i = 0, 1)$. The system is said to be (locally) observable at x_0 if there exists a neighbourhood N of x_0 such that every other $x_1 \in N$ is distinguishable from x_0 . The concept of observability was initially developed for linear systems, and was soon extended to the nonlinear case with the use of Lie algebra (Hermann and Krener, 1977).

A way to extract information about the state $x(t)$ from the output $y(t)$ of a system given by equations (1) is to build the derivatives \dot{y}, \ddot{y}, \dots . In these differentiations, the so-called Lie derivatives of the output function appear. Given a smooth function $g(x)$ and a vector field $z(x)$, the Lie derivative of g with respect to z is:

$$L_z g = \frac{\partial}{\partial x} g(x) \cdot z(x), \quad (4)$$

where $\frac{\partial}{\partial x} g(x)$ is a row vector containing the partial derivatives of the smooth function $g(x)$. We are interested in the Lie derivative of g along f , which is defined as:

$$L_f g(x) = \frac{\partial g(x)}{\partial x} f(x, u) \quad (5)$$

For a generic system with n states and m outputs, $\frac{\partial}{\partial x} g(x)$ is a $m \times n$ matrix, and $L_f g(x) = \frac{\partial g(x)}{\partial x} f(x, u)$ is a $m \times 1$ column vector. The i^{th} order Lie derivatives are recursively defined as follows:

$$\begin{aligned} L_f^2 g(x) &= \frac{\partial L_f g(x)}{\partial x} f(x, u) \\ &\vdots \\ L_f^i g(x) &= \frac{\partial L_f^{i-1} g(x)}{\partial x} f(x, u) \end{aligned} \quad (6)$$

The observation space \mathbf{O} of (1) is the space of linear combinations (with constant coefficients) of functions of the form:

$$L_f^k g(x) = \overbrace{L_f \dots L_f}^k g(x), \quad k = 0, 1, 2, \dots \quad (7)$$

where $L_f^0 g(x) = g(x)$. For each x , let $\mathbf{dO}(x)$ denote the subspace of rows consisting of all $\frac{\partial}{\partial x} \alpha(x)$ with $\alpha(x) \in \mathbf{O}$.

Thus it is formed by the gradients of linear combinations of the Lie derivatives of g in (7). A sufficient condition for observability is as follows: given System (1) and a state x_0 , the system is locally observable around x_0 if the dimension of the rows subspace $\mathbf{dO}(x_0)$ is n . The previous condition is sufficient and “almost necessary”: if the system (1) is locally observable around *all* the states x , then the dimension of $\mathbf{dO}(x)$ is n for all the states belonging to an *open dense* subset of the state space, in the standard topology in \mathbb{R}^n , see (Vidyasagar, 2002).

Finally, stack n sub-matrices forming the nonlinear observability matrix:

$$\mathcal{O}(x) = \begin{pmatrix} \frac{\partial}{\partial x} g(x) \\ \frac{\partial}{\partial x} (L_f g(x)) \\ \frac{\partial}{\partial x} (L_f^2 g(x)) \\ \vdots \\ \frac{\partial}{\partial x} (L_f^{n-1} g(x)) \end{pmatrix} \quad (8)$$

Simplified sufficient condition for nonlinear observability: The system given by equations (1) is (locally) observable around x_0 if $\text{rank}(\mathcal{O}(x_0)) = n$.

2.3 Structural identifiability recast as observability

It is important to notice that the parameters p can be appended to the state, (x, p) , with trivial dynamics $\dot{p} = 0$, and, in this way, the identifiability problem can be easily recast in the framework of nonlinear observability (Tunali and Tarn, 1987). If we augment the state variable vector so as to include also the model parameters, $\tilde{x} = [x, p]^T$, we can write the resulting generalized observability-identifiability matrix, $\mathcal{O}_I(\tilde{x})$, as:

$$\mathcal{O}_I(\tilde{x}) = \begin{pmatrix} \frac{\partial}{\partial \tilde{x}} g(\tilde{x}) \\ \frac{\partial}{\partial \tilde{x}} (L_f g(\tilde{x})) \\ \frac{\partial}{\partial \tilde{x}} (L_f^2 g(\tilde{x})) \\ \vdots \\ \frac{\partial}{\partial \tilde{x}} (L_f^{n+q-1} g(\tilde{x})) \end{pmatrix} \quad (9)$$

Generalized observability-identifiability condition: the system given by equations (1) is (locally) observable and identifiable in a neighbourhood $N(\tilde{x}_0)$ of \tilde{x}_0 if $\text{rank}(\mathcal{O}_I(\tilde{x}_0)) = n + q$.

3. METHODOLOGY

The theory presented in Section 2 can be directly used to analyse structural identifiability by checking the generalized observability-identifiability condition: if \mathcal{O}_I is full rank, the system is observable and s.l.i.. However, in practice such analysis would often be incomplete and computationally inefficient: it would be incomplete because, if \mathcal{O}_I is

not full rank, no additional information is provided about which parameters are identifiable and which are not. It would be computationally inefficient (or even infeasible) because building \mathcal{O}_I and calculating its rank is a highly demanding, memory-consuming task, due to the symbolic calculation of the Lie derivatives and the subsequent rank. Hence in the following subsections we present an algorithmic procedure aimed at avoiding these shortcomings.

3.1 Assessing identifiability with a minimum size matrix

Let us first note that each of the $n+q$ sub-matrices stacked in the generalized observability-identifiability matrix of equation (9) has dimension $m \times (n+q)$, and the full matrix \mathcal{O}_I has dimensions $(m \cdot (n+q)) \times (n+q)$. Therefore it may not be necessary to calculate the $n+q-1$ Lie derivatives in order to test whether \mathcal{O}_I is full rank, since full rank may be achieved with a lower number of derivatives. The minimum number of Lie derivatives for which the matrix may be full rank is

$$n_d = \left\lceil \frac{n+q}{m} - 1 \right\rceil \quad (10)$$

that is, the smallest integer not less than $(n+q)/m - 1$, where n , q , and m are the numbers of states, parameters, and outputs, respectively.

The maximum number of Lie derivatives is also known a priori: derivatives of order higher than $n+q-1$ cannot increase the matrix rank (Anguelova, 2004). Having lower and upper bounds for the necessary Lie derivatives is an advantage of this methodology compared to, e.g., power series approaches, for which the maximum number of derivatives is in principle infinite (Chiş et al., 2011b).

The algorithm proposed here builds \mathcal{O}_I recursively. Once n_d is reached, addition of a new Lie derivative is followed by calculation of the rank. This process is repeated until the maximum number $n+q-1$ is reached, or until adding a new Lie derivative does not increase the matrix rank. At that point, if \mathcal{O}_I is full rank the corresponding model is observable and identifiable, as discussed in Section 2.3. If \mathcal{O}_I is not full rank, the algorithm proceeds to find identifiable subsets of parameters (which may include initial conditions, if considered unknown), as will be explained in Section 3.2.

3.2 Determining structurally identifiable parameter subsets

If deleting the i^{th} column of the generalized observability-identifiability matrix does not change its rank, then the corresponding i^{th} state (parameter) is non-observable (unidentifiable). This fact can be exploited to determine in a sequential procedure which parameters in an unidentifiable model are identifiable and which are not. After the matrix rank has been calculated and the model has been found to be unidentifiable, each of the columns in \mathcal{O}_I corresponding to a particular parameter is removed one by one and the rank is recalculated. In this way the identifiability of each of the parameters is evaluated. The same procedure can be carried out for the matrix columns corresponding to the states, in order to determine the identifiability of their initial conditions.

3.3 Decomposing large models to facilitate their analysis

The methodology described in the preceding subsections can be used to analyse the identifiability of whole models and, if the model is unidentifiable, of its parameters individually. However, since it relies heavily on symbolic operations, it may be computationally infeasible for large or complex models. It should be noted that the main limiting operations are:

- Obtaining high order Lie derivatives to build $\mathcal{O}_I(\tilde{x})$.
- Calculating the rank of the resulting $\mathcal{O}_I(\tilde{x})$.

The minimum number of derivatives necessary for building $\mathcal{O}_I(\tilde{x})$ is given by n_d as defined in equation (10). The limit of what is computationally possible is difficult to quantify a priori, since it depends on the model equations and the machine used in the calculations. As a rule of thumb, analyses involving $n_d \geq 10$ are infeasible except for very small models. As model size or complexity increases, this upper bound decreases; some examples will be shown in Section 4.

A solution is to decompose those models into smaller submodels whose analysis is possible computationally. Thus, we seek to decompose a model M into submodels $\{M_1, M_2, \dots\}$ which require few Lie derivatives for their analysis, that is, they have a small n_d . For each submodel M_i we select a subset of the states in M by performing a combinatorial optimization where we minimize n_d :

$$\min_{\mathbf{s}} n_d(\mathbf{s}) \quad (11)$$

where $\mathbf{s} = \{s_1, s_2, \dots, s_n\}$ is a binary vector of size n , whose entries $s_j \in \{0, 1\}$ denote inclusion ($s_j = 1$) or exclusion ($s_j = 0$) of the corresponding state.

The combinatorial optimization is performed with the Variable Neighbourhood Search metaheuristic (Mladenović and Hansen, 1997). We carry out n optimizations (one per state); in the j^{th} optimization we force $s_j = 1$, so that each state appears in at least one solution. This, in turn, guarantees that all the parameters will eventually be evaluated. A penalty term is included in the objective function to penalize solutions that have more states than a chosen maximum.

Note that the quantity n_d relates the size of a model (number of state variables and parameters, $n + q$) with how much measured it is (number of outputs, m). Thus, even a relatively small model can have a large n_d , which will make it more difficult to analyse than a larger model with more measured outputs.

3.4 Assessing identifiability of decomposed models

Let us clarify how we can conclude identifiability of a parameter from analysis of a submodel. As an example, consider M to be the model of *A. thaliana* that will be described by equations (15) in Section 4.3. Assume the optimization algorithm has defined a submodel M_{sub} by selecting two states, $x_{sub} = \{x_1, x_7\}$. The equations of M_{sub} are those equations in (15) that correspond to the states $\{x_1, x_7\}$, that is:

$$\begin{cases} \dot{x}_1 = n_1 \frac{x_6}{g_1 + x_6} - m_1 \frac{x_1}{k_1 + x_1} + q_1 x_7 u(t), \\ \dot{x}_7 = p_3 - m_7 \frac{x_7}{k_7 + x_7} - (p_3 + q_2 x_7) u(t), \\ x_1(0) = 0, x_7(0) = 0 \end{cases} \quad (12)$$

The outputs of M_{sub} are those outputs of M which are functions of at least one of the states in M_{sub} (in this example, $y_1 = x_1$). The parameters and inputs of M_{sub} are those present in equations (12): respectively, $\{n_1, g_1, m_1, k_1, q_1, p_3, m_7, k_7, q_2\}$ and u . Additionally, we must also include as parameters the states that do not belong to x_{sub} but appear in equations (12) or in y_{sub} (in this case, x_6). Thus in this example the submodel parameters would be $\{n_1, g_1, m_1, k_1, q_1, p_3, m_7, k_7, q_2, x_6\}$.

Note that, although the state x_6 appears in the equations (12), the dynamic equation of \dot{x}_6 is not included in the submodel; instead, x_6 is considered as an extra parameter. By including coupled states such as x_6 as parameters we are considering them as unknown and constant. In contrast, if they were included as inputs to the submodel, we would be implicitly assuming that they provide sufficient excitation for identification purposes. Thus, including them as parameters is a conservative assumption in terms of identifiability. Therefore, if a parameter is classified as identifiable in a submodel under these conditions, it will also be identifiable when considering the whole model.

3.5 An alternative to decomposition: building \mathcal{O}_I with less than n_d Lie derivatives

When the n_d of the full model is so high that it is not feasible to build \mathcal{O}_I , a solution is to decompose the model into smaller submodels as described in the preceding subsections. Another possibility is simply to build \mathcal{O}_I with $i < n_d$ derivatives. In this case we know that full rank cannot be achieved, so even if the model is identifiable we will not be able to determine it in this way. However, it may be possible to determine identifiability of at least *some* of the parameters. This procedure can be helpful exactly in the same circumstances as decomposition. In some cases one approach will be more successful than the other one, but both can be used to determine the identifiability of different parameters, and may therefore be complementary.

3.6 Obtaining more complete results by removing columns of identifiable parameters

In certain cases the aforementioned procedure can yield incomplete results, that is, it may fail to determine the (un)identifiability of some parameters. For example:

- (1) When, due to computational limitations, \mathcal{O}_I is calculated with less Lie derivatives than those needed to guarantee identifiability. In this case, it may happen that an \mathcal{O}_I calculated with more Lie derivatives would have a higher rank and reveal the identifiability of more parameters.
- (2) When using decomposition, a parameter may not be determined as identifiable if it is not tested with the appropriate combination of states and outputs. Imagine, for example, that identification of a particular parameter p_i requires observing two outputs, y_a and

y_b , but only one of them was included in the submodel used to evaluate the identifiability of p_i .

A way of reducing the conservative nature of the methodology under the aforementioned circumstances is to refine the solutions iteratively, by removing the columns of \mathcal{O}_I corresponding to those parameters that were already found to be identifiable in previous steps. This will lead to a smaller \mathcal{O}_I matrix, reducing the dimension of the problem, which may enable its analysis with the rank condition without resorting to decomposition. If that is the case, it will be possible to assess the identifiability of all the parameters. This manipulation is legitimate for two reasons: first, it yields the same \mathcal{O}_I matrix than would be obtained by using the original vector $\tilde{x} = [x, p]$ and removing the corresponding columns afterwards. Second, we know that the rank of the resulting matrix decreases with every removed column, because it corresponds to an identifiable parameter. Therefore it is possible to extract conclusions from the rank test of the reduced matrix.

In summary, if a model M is too large to calculate the rank of its identifiability matrix, it can be analysed as follows:

- (1) Decompose M into several submodels, S_i .
- (2) Analyse the resulting S_i submodels using the generalised observability-identifiability rank condition. If the array is not full rank, test the identifiability of each parameter separately by comparing the rank before and after removing its column.
- (3) Parameters found to be identifiable in a submodel S_i are identifiable in the whole model M .
- (4) Several decompositions can be tested, which may lead to complementary results.
- (5) Additionally, it may be possible to find identifiable parameters by checking the rank of a \mathcal{O}_I built with less than n_d Lie derivatives.
- (6) Decrease the size of \mathcal{O}_I by removing the parameters determined to be identifiable in the previous steps and calculate its rank if possible.

4. RESULTS

Here we test the proposed methodology on three problems used in the recent literature (Chiş et al., 2011b; Raue et al., 2014) to evaluate structural identifiability approaches. Calculations were carried out on a computer running Windows7 SP1 64bit, with an Intel processor at 3.40 GHz and 16 GB of RAM, using MATLAB R2015b.

4.1 JAK/STAT signalling pathway

The first case study is the IL13-Induced JAK/STAT signalling pathway model presented by Raia et al. (2011). This problem was later used by Raue et al. (2014) to benchmark three identifiability analysis methods. It has 10 states, 23 parameters, and 8 outputs, which are: $y_1 = x_1 + x_3 + x_4$, $y_2 = \theta_{18}(x_3 + x_4 + x_5 + 0.34 - x_{11})$, $y_3 = \theta_{19}(x_4 + x_5)$, $y_4 = \theta_{20}(2.8 - x_6)$, $y_5 = \theta_{21}x_{10}$, $y_6 = \theta_{17}\theta_{22}x_{10}/\theta_{11}$, $y_7 = x_{13}$, $y_8 = 165 - x_8$.

The model dynamics and initial conditions are given by the following equations:

$$\begin{cases} \dot{x}_1 = \theta_6 x_2 - \theta_5 x_1 - \frac{453 \theta_1 u_1 x_1}{200}, \\ \dot{x}_2 = \theta_5 x_1 - \theta_6 x_2, \\ \dot{x}_3 = \theta_2 x_3 \left(x_6 - \frac{14}{5} \right) + \frac{453 \theta_1 u_1 x_1}{200}, \\ \dot{x}_4 = -\theta_3 x_4 - \theta_2 x_3 \left(x_6 - \frac{14}{5} \right), \\ \dot{x}_5 = \theta_3 x_4 - \theta_4 x_5, \\ \dot{x}_6 = -91 \theta_8 \left(x_6 - \frac{14}{5} \right) - \frac{\theta_7 x_3 x_6}{\theta_{13} x_1 + 1} - \frac{\theta_7 x_4 x_6}{\theta_{13} x_{13} + 1}, \\ \dot{x}_8 = \theta_9 x_8 \left(x_6 - \frac{14}{5} \right) - 91 \theta_{10} (x_8 - 165), \\ \dot{x}_{10} = -\theta_{11} (x_8 - 165), \\ \dot{x}_{11} = -\frac{453 \theta_{12} u_1 x_{11}}{200}, \\ \dot{x}_{13} = \frac{\theta_{14} x_{10}}{\theta_{15} + x_{10}} - \theta_{16} x_{13}, \\ \mathbf{x}(0) = [1.3, \theta_{23}, 0, 0, 0, 2.8, 165, 0, 0.34, 0] \end{cases} \quad (13)$$

The JAK/STAT signalling pathway model can be completely analysed with the methodology outlined in Sections 3.1–3.2, without using decomposition. The results of this method coincide with those reported in (Raue et al., 2014), that is, there are five unidentifiable parameters, θ_{11} , θ_{15} , θ_{17} , θ_{21} , and θ_{22} .

4.2 Enzymatic oscillations

Our second case study is a classic model of oscillations in enzyme kinetics proposed by Goodwin (1965) and used by Chiş et al. (2011b) to benchmark several global structural identifiability methods. The model has 1 output ($y = x_1$), 3 states, and 8 parameters:

$$\begin{cases} \dot{x}_1 = -bx_1 + \frac{a}{A + x_3^{\sigma}}, \\ \dot{x}_2 = \alpha x_1 - \beta x_2, \\ \dot{x}_3 = \gamma x_2 - \delta x_3, \\ x_1(0) = 0.3617, x_2(0) = 0.9137, x_3(0) = 1.3934 \end{cases} \quad (14)$$

Despite being smaller, this system is more difficult to analyse than the JAK/STAT model, because it has only one observable. Hence none of the methods tested by Chiş et al. (2011b) managed to reach a conclusion due to computational limitations. The EAR method (Karlsson et al., 2012) cannot be applied either because it requires that the system is rational, which is not the case for this model due to the presence of the σ parameter as an exponent of x_3^{σ} in the dynamic equation of \dot{x}_1 .

In principle, our approach requires at least 10 Lie derivatives to build the identifiability matrix \mathcal{O}_I . With this number the subsequent rank calculation was too demanding computationally. However, using only 9 Lie derivatives, as suggested in in Section 3.5, it was possible to determine identifiability of four parameters: b, σ, β, δ . Then, removing these parameters from the model, as explained in Section 3.6, made it possible to calculate the rank of \mathcal{O}_I and, in this way, to classify the remaining parameters (a, A, α, γ) as unidentifiable.

4.3 Circadian clock in *Arabidopsis thaliana*

The third case study is the genetic network that controls the circadian clock in *A. thaliana* (Locke et al., 2005). This model has 2 outputs ($y_1 = x_1, y_2 = x_4$), 7 states, and 27 parameters:

$$\begin{cases} \dot{x}_1 = n_1 \frac{x_6}{g_1 + x_6} - m_1 \frac{x_1}{k_1 + x_1} + q_1 x_7 u(t), \\ \dot{x}_2 = p_1 x_1 - r_1 x_2 + r_2 x_3 - m_2 \frac{x_2}{k_2 + x_2}, \\ \dot{x}_3 = r_1 x_2 - r_2 x_3 - m_3 \frac{x_3}{k_3 + x_3}, \\ \dot{x}_4 = n_2 \frac{g_2^2}{g_2^2 + x_3^2} - m_4 \frac{x_4}{k_4 + x_4}, \\ \dot{x}_5 = p_2 x_4 - r_3 x_5 + r_4 x_6 - m_5 \frac{x_5}{k_5 + x_5}, \\ \dot{x}_6 = r_3 x_5 - r_4 x_6 - m_6 \frac{x_6}{k_6 + x_6}, \\ \dot{x}_7 = p_3 - m_7 \frac{x_7}{k_7 + x_7} - (p_3 + q_2 x_7) u(t) \\ x_i(0) = 0; i = 1, \dots, 7. \end{cases} \quad (15)$$

Chiş et al. (2011b) reported that the Taylor Series (TS) and Generating series (GS) approaches determine (global) identifiability of 2 and 5 parameters, respectively, while other methods fail to provide results. TS and GS yield partial results—i.e., do not classify certain parameters as (at least) locally identifiable—due to the computational complexity of the problem. The EAR method (Karlsson et al., 2012) classifies 11 parameters as (locally) identifiable.

This example illustrates the utility of decomposition approach: building $\mathcal{O}_I(\tilde{x})$ would require $n_d = 16$ Lie derivatives, which is computationally infeasible. Attempting to analyse the identifiability of individual parameters with fewer than n_d derivatives, as done for the model in Section 4.2, did not produce results in this case. However, by decomposing the model we classified 8 parameters as identifiable. Removing these parameters from the model decreased the number of derivatives to $n_d = 12$. Since this number was still too high, we built \mathcal{O}_I with less derivatives, which led to classifying an additional parameter as identifiable. Thus in this way a total of 9 parameters were determined to be identifiable (a subset of the 11 parameters classified by EAR); unidentifiability of the remaining ones could not be guaranteed.

5. CONCLUSIONS

Analysing the structural identifiability of nonlinear dynamic models is a challenging task, whose difficulty increases rapidly with model size. While a plethora of methods have been proposed for tackling this problem, no universally applicable solution exists yet. The methodology presented here studies structural identifiability as an extension of nonlinear observability, exploiting tools originally developed for observability analysis of nonlinear systems. It is based on checking a generalized observability-identifiability condition by calculating the rank of a matrix. In this way it is possible to handle very general nonlinear models; for example, it is not necessary that they are in polynomial or rational form. For structurally

unidentifiable models the method reports a list of identifiable parameters.

The generalized observability-identifiability condition can be directly checked even for problems of medium size, such as the JAK/STAT used by Raue et al. (2014) to benchmark identifiability methods. For models which are larger, more complex, or for which less observations are available, it may not be possible to perform this test directly due to computational limitations. In that case, the methodology can be used in one of two ways. The first is to try to obtain partial results by building the observability-identifiability matrix with less derivatives than required for full rank, as was shown for the case of the model of enzymatic oscillations (Goodwin, 1965). The second is to decompose the model into a set of submodels whose analysis is feasible. The decomposition is carried out with an optimization method, and is performed in such a way that identifiability of the parameters in any submodel guarantees identifiability in the whole model. The use of decomposition was demonstrated on a model of the circadian clock in *A. thaliana* (Locke et al., 2005).

ACKNOWLEDGEMENTS

We are grateful to Dhruva Raman and James Anderson for many helpful discussions.

REFERENCES

- Angelova, M. (2004). *Nonlinear observability and identifiability: General theory and a case study of a kinetic model for S. Cerevisiae*. Master’s thesis, Chalmers University of Technology and Göteborg University.
- Balsa-Canto, E., Alonso, A., and Banga, J.R. (2010). An iterative identification procedure for dynamic modeling of biochemical networks. *BMC Syst. Biol.*, 4, 11:1–11:18.
- Bellman, R. and Åström, K.J. (1970). On structural identifiability. *Math. Biosci.*, 7(3), 329–339.
- Bellu, G., Saccomani, M.P., Audoly, S., and D’Angio, L. (2007). DAISY: a new software tool to test global identifiability of biological and physiological systems. *Comput. Methods Programs Biomed.*, 88(1), 52–61.
- Chatzis, M.N., Chatzi, E.N., and Smyth, A.W. (2015). On the observability and identifiability of nonlinear structural and mechanical systems. *Struct. Contr. Health Monitor.*, 22(3), 574–593.
- Chiş, O., Banga, J.R., and Balsa-Canto, E. (2011a). Genssi: a software toolbox for structural identifiability analysis of biological models. *Bioinformatics*, 27(18), 2610–2611.
- Chiş, O.T., Banga, J.R., and Balsa-Canto, E. (2011b). Structural identifiability of systems biology models: a critical comparison of methods. *PLoS One*, 6(11), e27755.
- Goodwin, B.C. (1965). Oscillatory behavior in enzymatic control processes. *Adv. Enzyme Regul.*, 3, 425–437.
- Grandjean, T.R., Chappell, M.J., Yates, J.W., and Evans, N.D. (2014). Structural identifiability analyses of candidate models for in vitro pitavastatin hepatic uptake. *Comput. Methods Programs Biomed.*, 114(3), e60–e69.
- Hermann, R. and Krener, A.J. (1977). Nonlinear controllability and observability. *IEEE Trans. Autom. Control*, 22(5), 728–740.

- Karlsson, J., Anguelova, M., and Jirstrand, M. (2012). An efficient method for structural identifiability analysis of large dynamic systems. In *16th IFAC Symposium on System Identification*, volume 16, 941–946.
- Ljung, L. and Glad, T. (1994). On global identifiability for arbitrary model parametrizations. *Automatica*, 30(2), 265–276.
- Locke, J., Millar, A., and Turner, M. (2005). Modelling genetic networks with noisy and varied experimental data: the circadian clock in *Arabidopsis thaliana*. *J. Theor. Biol.*, 234(3), 383–393.
- Merkt, B., Timmer, J., and Kaschek, D. (2015). Higher-order Lie symmetries in identifiability and predictability analysis of dynamic models. *Phys. Rev. E*, 92(1), 012920.
- Meshkat, N., Kuo, C.E.z., and DiStefano III, J. (2014). On finding and using identifiable parameter combinations in nonlinear dynamic systems biology models and combos: A novel web implementation. *PLoS One*, 9(10).
- Miao, H., Xia, X., Perelson, A.S., and Wu, H. (2011). On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Rev.*, 53(1), 3–39.
- Mladenović, N. and Hansen, P. (1997). Variable neighborhood search. *Comput. Oper. Res.*, 24(11), 1097–1100.
- Pohjanpalo, H. (1978). System identifiability based on the power series expansion of the solution. *Math. Biosci.*, 41(1), 21–33.
- Raia, V., Schilling, M., Böhm, M., Hahn, B., Kowarsch, A., Raue, A., Sticht, C., Bohl, S., Saile, M., Möller, P., et al. (2011). Dynamic mathematical modeling of IL13-induced signaling in hodgkin and primary mediastinal b-cell lymphoma allows prediction of therapeutic targets. *Cancer Res.*, 71(3), 693–704.
- Raue, A., Karlsson, J., Saccomani, M.P., Jirstrand, M., and Timmer, J. (2014). Comparison of approaches for parameter identifiability analysis of biological systems. *Bioinformatics*, btt006.
- Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15), 1923–1929.
- Sedoglavic, A. (2002). A probabilistic algorithm to test local algebraic observability in polynomial time. *J. Symb. Comput.*, 33, 735–755.
- Stigter, J.D. and Molenaar, J. (2016). A fast algorithm to assess local structural identifiability. *Automatica*, 58, 118–124.
- Tunali, E.T. and Tarn, T.J. (1987). New results for identifiability of nonlinear systems. *IEEE Trans. Autom. Control*, 32(2), 146–154.
- Vajda, S., Godfrey, K.R., and Rabitz, H. (1989). Similarity transformation approach to identifiability analysis of nonlinear compartmental models. *Math. Biosci.*, 93(2), 217–248.
- Vidyasagar, M. (2002). *Nonlinear systems analysis*, volume 42. SIAM.
- Villaverde, A.F. and Barreiro, A. (2016). Identifiability of large nonlinear biochemical networks. *MATCH Commun. Math. Comput. Chem.*, 76(2), 259–296.
- Walter, E. and Pronzato, L. (1997). *Identification of parametric models from experimental data*. Communications and Control Engineering Series. Springer, London, UK.
- Walter, E. and Lecourtier, Y. (1982). Global approaches to identifiability testing for linear and nonlinear state space models. *Math. Comput. Simul.*, 24(6), 472–482.
- Yates, J.W., Evans, N.D., and Chappell, M.J. (2009). Structural identifiability analysis via symmetries of differential equations. *Automatica*, 45(11), 2585–2591.