16th IFAC Symposium on System Identification
The International Federation of Automatic Control
Brussels, Belgium. July 11-13, 2012

An Efficient Method for Structural Identifiability Analysis of Large Dynamic Systems ⋆

Johan Karlsson * **Milena Anguelova** **,⋆⋆ **Mats Jirstrand** *

* *Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Chalmers Science Park, SE-412 88 Göteborg Sweden (e-mail: {johkar, matsj}@fcc.chalmers.se).*
** *Imego AB, Arvid Hedvalls Backe 4, SE-400 14 Göteborg, Sweden.*

Abstract: Ordinary differential equation models often contain a large number of parameters that must be determined from measurements by parameter estimation. For a parameter estimation procedure to be successful, there must be a unique set of parameters that can have produced the measured data. This is not the case if a model is not structurally identifiable with the given set of outputs selected as measurements. We describe the implementation of a recent probabilistic semi-numerical method for testing local structural identifiability based on computing the rank of a numerically instantiated Jacobian matrix (observability/identifiability matrix). To obtain this, matrix parameters and initial conditions are specialized to random integer numbers, inputs are specialized to truncated random integer coefficient power series, and the corresponding output of the state space system is computed in terms of a truncated power series, which then is utilized to calculate the elements of a Jacobian matrix. To reduce the memory requirements and increase the speed of the computations all operations are done modulo a large prime number. The method has been extended to handle parametrized initial conditions and is demonstrated to be capable of handling systems in the order of a hundred state variables and equally many parameters on a standard desktop computer.

Keywords: Systems identification, Identifiability; Parameter estimation; Dynamic systems.

1. INTRODUCTION

In many applications, the parameters in models of dynamic systems are not directly measurable but only accessible indirectly through their impact on measured entities, which typically are time varying signals, *outputs*, responding to some applied perturbations, *inputs*, to the system under study. There are many parameter estimation methods, which, given a parametrized set of candidate models, a *model structure*, and measured input-output data, perform a numerical search to obtain good numerical values of the parameters. However, a fundamental question to be answered before such methods are invoked is if the model structure in question is identifiable. Structural identifiability is a property of a model structure that ensures that parameters can be uniquely (globally or locally) determined from knowledge of the input-output behavior of the system. It is not an uncommon situation that model structures obtained by physical or chemical modeling are unidentifiable, i.e., there is an infinite number of sets of parameter values that equally well describe the input-output data.

⋆ This work was supported by grants from the European Commission 7th Framework Programme (UNICELLSYS, grant No 201142 and CANCERSYS, grant No 223188) and the Swedish Foundation for Strategic Research through the Gothenburg Mathematical Modelling Centre.
⋆⋆This work was carried out while the author was affiliated with Fraunhofer-Chalmers Centre.

A large amount of literature has been devoted to the theoretical characterization of this subject, starting in Kalman [1961] for linear systems and in Hermann and Krener [1977] for the nonlinear case, and continuing until today, where Pohjanpalo [1978], Vajda et al. [1989], Walter [1987], Ollivier [1990], Ljung and Glad [1990], Diop and Fliess [1991], Evans and Chappell [2000], Audoly et al. [2001], Margaria et al. [2001], Sedoglavic [2002], Saccomani et al. [2003], Anguelova [2007], and Yates et al. [2009] are just a few references. Before the work of Sedoglavic [2002], the available methods for testing structural observability or identifiability of nonlinear systems relied on characteristic set or standard bases computation (Ollivier [1990], Diop and Fliess [1991], Ljung and Glad [1990], Audoly et al. [2001] and Margaria et al. [2001]) or the local state variable isomorphism approach (Vajda et al. [1989] and Evans and Chappell [2000]). The complexity in the number of variables and parameters of these methods grows too fast for them to be generally applicable to models of large dynamic systems. In Sedoglavic [2002], a probabilistic semi-numerical algorithm is presented for testing the local structural identifiability of a model, even for large models with a few hundred state variables and parameters.

We describe a Mathematica implementation of the probabilistic semi-numerical algorithm described in Sedoglavic [2002] and outline the main ideas behind the algorithm without the need for the reader to have extensive knowl-

978-3-902823-06-9/12/$20.00 © 2012 IFAC

10.3182/20120711-3-BE-2027.00381

edge in algebraic or differential algebraic theory. Furthermore, the algorithm is extended to cover more generally parametrized initial conditions. The performance of the implementation is demonstrated through a number of examples of large systems of biochemical reaction networks. The paper is organized as follows. In Section 2 we define the concepts of identifiability and observability to be used in this paper. Section 3 describes the algorithm for testing local structural identifiability of a given model structure and some notes on the implementation of the method in Mathematica is outlined in Section 4. Section 5 contains a number of examples and Section 6 concludes the paper and outlines some future work.

2. IDENTIFIABILITY AND OBSERVABILITY

Consider a parametrized class of models in state space form

$$\dot{x}(t) = f(x(t), u(t), \theta), \qquad x(0) = x^0(\theta) \tag{1}$$
$$y(t) = g(x(t), u(t), \theta) \tag{2}$$

where $x(t) \in \mathbb{R}^n$ is the *state*, $u(t) \in \mathbb{R}^m$ is the *input*, $\theta \in \mathbb{R}^d$ is the parameter vector, and $y(t) \in \mathbb{R}^p$ is the *output* of the system. Note the explicit dependence on parameters in the equation for the initial conditions. This includes both the case of unknown initial conditions independent from other parameters in the model and the case when some initial conditions are given as expressions of other model parameters. The latter is for example the case if one requires the system to be initialized in steady state. The methods presented in this paper requires the vector valued functions f, g, and x^0 to be rational functions of their arguments. This requirement is not as restrictive as it may sound, since it can be shown that any function, which in itself is the solution to an equation such as (1), can be handled through an extended state space approach, see Lindskog [1996].

We will now decribe one approach of how to arrive at a Jacobian matrix whose rank determines if a system is structurally identifiable or not. This is *not* the approach used in the algorithm of this paper but the description is included for comparative reasons.

Introducing the extended Lie-derivative operator, $\mathcal{L}_f$, along the vector field, f, by the expression

$$\mathcal{L}_f = \sum_{i=1}^{n} f_i \frac{\partial}{\partial x_i} + \sum_{i=0}^{\infty} u^{(i+1)} \frac{\partial}{\partial u^{(i)}} \tag{3}$$

and denoting k times repeated application of the operator with $\mathcal{L}_f^k$ the derivative with respect to time of any order of $x(t)$ or $y(t)$ evaluated at $t = 0$ can very compactly be expressed according to

$$x^{(j)}(0) = \mathcal{L}_f^j f(x, u, \theta)|_{x=x(0), u^{(k)}=u^{(k)}(0), k=0,\ldots,j} \tag{4}$$
$$y^{(j)}(0) = \mathcal{L}_f^j g(x, u, \theta)|_{x=x(0), u^{(k)}=u^{(k)}(0), k=0,\ldots,j} \tag{5}$$

Hence, this gives explicit expressions linking the initial values of the state variables with the initial value of a derivative with respect to time of any order of both the state and output of a system defined by (1) and (2).

Now, assume that the input-output behavior of a system is given. The right hand side of (5) is an expression in $x(0)$ and θ and the left hand side of (5) is a known quantity for any j (since $y(t)$ is assumed to be described by the solution to a system of differential equations and hence is well behaved and can be proven to have a Taylor series expansion). This can be put in vector equation form

$$\mathcal{Y} = \mathcal{Y}(x(0), \theta) \tag{6}$$

where $\mathcal{Y}$ is a column vector containing $y^{(j)}(0), j = 0, \ldots, \nu$, $\nu = n + d - 1$, and the dependence on $u^{(j)}(0), j = 0, \ldots, \nu$ has been absorbed into the notation of the ν-dimensional vector valued function on the right hand side. It can be shown that derivatives of order higher than this ν are algebraically dependent on lower order derivatives, see Sedoglavic [2002] and Anguelova [2007]. Utilizing the inverse function theorem, equation (6) can be uniquely solved (locally) for $x(0)$ and θ if and only if the Jacobian matrix

$$J(x(0), \theta) = \frac{\partial \mathcal{Y}(x, \theta)}{\partial (x, \theta)}\Big|_{x=x(0)} \tag{7}$$

has full rank. This is called the rank-test for structural identifiability, see Pohjanpalo [1978]. The entries of the above Jacobian is given by

$$\tfrac{\partial}{\partial x_i} \mathcal{L}_f^j g(x, u, \theta)|_{x=x(0), u^{(k)}=u^{(k)}(0), k=0,\ldots,j} \tag{8}$$
$$\tfrac{\partial}{\partial \theta_i} \mathcal{L}_f^j g(x, u, \theta)|_{x=x(0), u^{(k)}=u^{(k)}(0), k=0,\ldots,j} \tag{9}$$

In the case of parametrized initial conditions, the relation between the derivatives of the output at $t = 0$ and the parameters becomes

$$\mathcal{Y} = \mathcal{Y}(x^0(\theta), \theta) \tag{10}$$

and hence the Jacobian when $x(0)$ no longer is assumed to be independent of θ can be expressed as

$$J(\theta) = \frac{\partial \mathcal{Y}(x, \theta)}{\partial x} \frac{\partial x}{\partial \theta} + \frac{\partial \mathcal{Y}(x, \theta)}{\partial \theta}\Big|_{x=x^0(\theta)} \tag{11}$$

Consider also the n equations corresponding to the initial conditions

$$x(0) = x^0(\theta) \tag{12}$$

The parameters θ, as well as the model structure (1) and (2) is said to be *locally structurally identifiable* if for almost all values of $x(0)$ and θ and their corresponding $\mathcal{Y}$ (generated by the model structure, specific $x(0)$ and θ, and given input), the equation (6) has a locally unique solution $x(0)$ and θ.

The above definition is close to the definition of local algebraic observability in Sedoglavic [2002] but we have avoided the introduction of more advanced algebraic terminology.

The concept of identifiability is closely related to that of observability. Consider for a moment the parameter vector θ to be known. A system (1) is said to be *globally observable* if given the input-output behavior, the initial state, x^0, can be uniquely determined, see Hermann and Krener [1977]. Using the trivial parametrization $x(0) = \theta$ with no explicit dependence of f and g on θ the observability of the initial state can be included in the notion of structural identifiability of the model parameters.

3. THE ALGORITHM

A straightforward symbolic computation of the Jacobian matrix in terms of repeated Lie derivatives according to (5) followed by taking partial derivatives with respect to $x(0)$

and θ to obtain the symbolic form of the Jacobian and finally testing its generic rank suffers from a computational complexity problem. The rank test of the symbolic Jacobian could be replaced by a numeric rank test on an integer valued matrix after a random specialization of the initial state and parameters. However, the expression swell in symbolically obtaining higher order derivatives by repeated Lie-derivatives is huge and prevents the application of this approach to large systems.

There is however a way to directly compute the Jacobian for the random specializations mentioned above, without the need to actually carry out the repeated Lie derivatives. Instead one may directly calculate the entries of the specialized Jacobian via the computation of power series expansion of partial derivatives with respect to $x(0)$ and θ of the output. To cut computational complexity even more, all calculations could be carried out modulo a large prime number preventing the risk of switching to slow software arithmetics for large integers. This indirect way of computing the entries of the Jacobian, avoiding the prohibitive computational complexity for large systems, was first recognized by Sedoglavic [2002]. The use of random specializations and modulo calculations both introduce the risk of loosing rank regardless of the Jacobian's generic rank properties. However, the probability of this can be bounded from above and this upper bound decrease with the size of the modulus in such a way that for large prime modulus suitable for software implemention the probability becomes vanishingly small. Note that, because integers are used, the results are exact in the sense that it is structural identifiability (not practical identifiability) that is analyzed.

Let us introduce the notation $x(t; x^0, \theta)$ to indicate implicit dependence of a solution to system (1) on the initial state and parameters. The derivatives of the output with respect to the initial state and parameters are

$$\begin{aligned}\frac{d}{dx^0}y(t;x^0,\theta) &= \frac{d}{dx^0}g(x(t;x^0,\theta),u(t),\theta)\\ &= \frac{\partial g}{\partial x}\frac{\partial x}{\partial x^0}\Big|_{x=x(t;x^0,\theta)} \qquad (13)\\ \frac{d}{d\theta}y(t;x^0,\theta) &= \frac{d}{d\theta}g(x(t;x^0,\theta),u(t),\theta)\\ &= \frac{\partial g}{\partial x}\frac{\partial x}{\partial \theta}+\frac{\partial g}{\partial \theta}\Big|_{x=x(t;x^0,\theta)} \qquad (14)\end{aligned}$$

Here is an outline of the steps of the algorithm

(1) Generate specializations of parameters and initial conditions to random integer values.
(2) The input of the system is specialized to a truncated random integer coefficient power series.
(3) Truncated power series solutions of the state, $x = x(t; x^0, \theta)$ and the state sensitivity with respect to the initial state, $\frac{\partial x}{\partial x^0}$, and with respect to the parameters, $\frac{\partial x}{\partial \theta}$, are computed.
(4) The above truncated power series are inserted in the expression for the derivatives of the outputs with respect to the initial state (13) and with respect to the parameters (14), which results in truncated power series representation of the output derivatives.
(5) Identification of the coefficients of the truncated power series of the output derivatives with the coefficients (apart from a j! denominator) of a general Taylor expansion of the output derivatives around $t = 0$ gives the higher order time-derivatives of the output derivatives evaluated at $t = 0$, i.e., the entries of the specialized Jacobian matrix.
(6) Calculate the rank of the specialized Jacobian.

For efficiency reasons, the algorithm starts with truncating power series after rather few terms. The above steps may then need to be repeated, truncating the power series at a higher order term until the rank computation of the specialized Jacobian either reaches $n + d$, which shows local structural identifiability, or stabilizes on a smaller number indicating unidentifiability. Observe that for vector output systems ($p > 1$) the number of terms in the involved truncated power series may be significantly less than $n + d$. In practice the rank computation may be performed by calculating the null space of the Jacobian. A non-trivial null space gives information about which combinations of columns in the Jacobian are generically linearly dependent, i.e., which initial state variable values and parameters that cannot be independently solved for.

3.1 Rationally Parametrized Initial Conditions

In the original algorithm presented by Sedoglavic [2002], parameters are treated as state variables with zero time-derivatives. In this way the problem is turned into the framework of nonlinear observability, where an observable system means that all parameters and the initial values of the state variables can generically be uniquely determined from input-output data. However, this also means that initial values and parameters are inherently separated.

The model structure (1) and (2) permits parametrized initial conditions, which generalizes the above approach since initial values of state variables and other parameters in a model do not need to be independent. This is a common situation for models constrained to start in steady state, which puts the implicit constraint $0 = f(x^0, u(0), \theta)$ on the initial values. The approach taken in Sedoglavic [2002] is covered in our setting by letting the initial values be separate parameters independent of the rest of the model parameters.

We extend the algorithm to parametrized initial conditions by firstly, specializing the parameters to random integer values, which now also determines the initial state variable values. Secondly, in step (4) of the algorithm equation (14) is replaced by

$$\begin{aligned}\frac{d}{d\theta}y(t;x^0(\theta),\theta) &= \frac{d}{d\theta}g(x(t;x^0(\theta),\theta),u(t),\theta)\\ &= \frac{\partial g}{\partial x}\frac{\partial x}{\partial x^0}\frac{\partial x^0}{\partial \theta}+\frac{\partial g}{\partial x}\frac{\partial x}{\partial \theta}+\frac{\partial g}{\partial \theta}\Big|_{x=x(t;x^0(\theta),\theta)} \qquad (15)\end{aligned}$$

for calculating the truncated power series for the derivatives of the output with respect to the parameters, which gives the modified Jacobian matrix $\tilde{J}(x(0), \theta)$.

Now, in addition to equation (6), we have the algebraic equations for the initial values (12). This means that we must ask whether any of the null-vectors of $J(x(0), \theta)$ provide directions in which $x(0)$ and θ can be changed while keeping (12) fulfilled.

Each null-vector $\nu = (\nu_1, \ldots, \nu_{n+p})$ corresponds to a differentiation $\sigma = \nu_1 \frac{\partial}{\partial x_1} + \cdots + \nu_n \frac{\partial}{\partial x_n} + \nu_{n+1} \frac{\partial}{\partial \theta_1} + \cdots + \nu_{n+p} \frac{\partial}{\partial \theta_p}$. If this differentiation σ is such that it also gives zero when it is applied to the initial condition equations,

$$\sigma(x(0) - x^0(\theta))\big|_{x(0)=x^0(\theta)} = 0, \tag{16}$$

then σ represents combinations of the parameters θ which fulfill both (6) and (12). They give the same Taylor series for $y(t)$ around t_0 and θ is not locally structurally identifiable.

Algorithmically it works as follows. Form a vector $I = x(0) - x^0(\theta)$. Now, to complement $\tilde{J}(x(0), \theta)$, take the derivatives of I with respect to the state variables and parameters. Then, from the resulting matrix of derivatives, take those combinations of columns of I that gave corresponding null-vectors for $\tilde{J}(x(0), \theta)$, and evaluate the column combinations for the specialized numerical values of the parameters. These column combinations represent symmetries, relations between parameters, where a change in one parameter can be compensated by changes in other parameters without the ouputs being changed, possibly resulting in unidentifiability. Next, calculate the nullspace of the resulting matrix. If this space is empty, then the system is identifiable after all, because of the additional information in the initial value equations. If it is not empty, then there are combinations (as defined by the null-vectors) of the original symmetries to $\tilde{J}(x(0), \theta)$ such that the information in I does not solve the problem and make the system identifiable. Form those combinations, giving a matrix S, representing the final symmetries.

Identifiable entities are those where the corresponding rows in S consist entirely of zeros, since these entities are not involved in any final symmetries. Entities where the corresponding row in S has nonzero-elements are involved in some symmetry and are not identifiable.

A variation of the above approach would be to eliminate $x(0)$ from the beginning, which means that one ends up with equation (10). Its Jacobian matrix (11) does not depend on the parameters. Now, in step (1) only the parameters needs to be specialized to random integer values and in step (4) of the algorithm both equation (13) and (14) are replaced by (15) for calculating the truncated power series for the derivatives of the output with respect to the parameters.

3.2 Known Initial Conditions

In the case that some initial conditions have known numerical values we proceed as follows. Introduce a parameter for each known state variable initial value. In step (1) of the algorithm specialize the corresponding parameters to the known initial conditions. In this way the solvability of equation (10) for these special values of the initial state variable values is decided by the algorithm having the rest of the parameters take generical values.

4. IMPLEMENTATION

The algorithm has been implemented as a Mathematica package. The function syntax is

IdentifiabilityAnalysis [{*deqn*, *ic*}, *y*, *u*, *x*, *θ*, *t*]

where *deqn* is a list of differential equations, *ic* is a list of parametrized initial conditions (if omitted independently parametrized initial conditions are assumed), y is a list of output expressions, and u, x, and θ are lists of names of inputs, state variables, and parameters, respectively. The last argument, t, is the name of the independent variable. The answer is given in the form

IdentifiabilityAnalysisData[*Boolean*, <>]

where *Boolean* is TRUE or FALSE depending on if the system is structurally identifiable or not. The resulting IdentifiabilityAnalysisData[...] object *iad*, contains a number of properties, which can be extracted with the syntax *iad*["*property*"]. A list of available properties is given by *iad*["Properties"], which includes "IdentifiableParameters", "UnidentifiableParameters", "TranscendenceDegree", and "Jacobian".

Some of the most time consuming steps of the algorithm are parallelizable, which is implemented and can be turned on with a simple option to the function if the system on which Mathematica is runing has access to more than one CPU.

Using this algorithm, even large systems can usually be analyzed on a standard desktop computer in reasonable time. All of the following examples but one were executed on a laptop with 4GB RAM and a dual core 2.53GHz CPU.For the most extreme cases we switched to a computer with 8GB RAM and a quad core 2.66GHz CPU.

Further information about the Mathematica package and instructions on how to obtain it can be found at www.fcc.chalmers.se/sys/products/identifiabilityanalysis.

5. EXAMPLES

To demonstrate the power and scalability of the presented algorithm we will now analyze the local structural identifiability of a set of models of biochemical reaction networks. A common approach to model the dynamic behavior of such networks is to use the so called reaction rate equations. These are derived from the mass balances of the reaction network, i.e., what reactants and products there are and their kinetics. Graphically such networks can be depicted using the Systems Biology Graphical Notation (SBGN), see Le Novère et al. [2009] and Jansson and Jirstrand [2010]. Note that the main point of the examples is to demonstrate the size of systems and sets of outputs that can be handled. Practical biological concerns like which things can easily be measured, what does it mean that some parameters are not identifiable if not measured, and such questions, are not adressed in this paper.

Example *A model of the NF-κB pathway.*
The first example is the signaling network of the two-feedback-loop regulatory module of the NF-κB signaling pathway from Lipniacki et al. [2004]. Mathematically, the network is given by a system of differential equations describing the rate of change of concentrations of different molecules in or around the cell.

It turns out that this particular system is not very well connected and many things must be measured or known *a priori* for the system to be identifiable. For example, if the state variables are selected as admissible output

signals and all parameters are to be estimated, then it can be shown that there are five state variables that must be included among the outputs for the system to be identifiable. A local structural identifiability analysis of this system using our Mathematica implementation looks as follows.

In[1]:= **vars = Table[x_i, {i, 15}];**

In[2]:= **params = Table[θ_i, {i, 28}];**

In[3]:= sys = {$x_1'[t] == -\theta_1 x_1[t] x_2[t] + \frac{1}{333}(-\theta_{14} x_1[t] + \theta_{15} x_9[t])$,
$x_2'[t] == -\theta_1 x_1[t] x_2[t] + \frac{1}{333}\theta_{13} x_8[t]$,
$x_3'[t] == \theta_1 x_1[t] x_2[t] - \frac{1}{333}\theta_{11} x_3[t]$,
$x_4'[t] == \theta_3 + \theta_2 x_2[t] - \theta_4 x_4[t]$,
$x_5'[t] == \theta_6 + \theta_5 x_2[t] - \theta_7 x_5[t]$,
$x_6'[t] == \theta_9 + \theta_8 x_2[t] - \theta_{10} x_6[t]$,
$x_7'[t] == \frac{10}{16667}\theta_{11} x_3[t] - \theta_{21} x_7[t] + \theta_1 x_8[t] x_9[t] - \theta_{28} x_7[t] x_{11}[t]$,
$x_8'[t] == \theta_{21} x_7[t] - \frac{10}{16667}\theta_{13} x_8[t] - \theta_1 x_8[t] x_9[t] + \theta_{26} x_{15}[t]$,
$x_9'[t] == \theta_{18} x_5[t] - \theta_{23} x_9[t] - \theta_1 x_8[t] x_9[t] + \frac{10}{16667}(\theta_{14} x_1[t] - \theta_{15} x_9[t]) - \theta_{25} x_9[t] x_{11}[t]$,
$x_{10}'[t] == \theta_{27} x_4[t] - \theta_{24} x_{10}[t]$,
$x_{11}'[t] == -\theta_{12} x_{11}[t] - \theta_{16} x_{11}[t] - \theta_{28} x_7[t] x_{11}[t] - \theta_{25} x_9[t] x_{11}[t] - \theta_{19} x_{10}[t] x_{11}[t] + \theta_{17} x_{13}[t] + \theta_{22} x_{14}[t] + \theta_{26} x_{15}[t]$,
$x_{12}'[t] == \theta_{16} x_{11}[t] + \theta_{19} x_{10}[t] x_{11}[t] - \theta_{12} x_{12}[t]$,
$x_{13}'[t] == \theta_{25} x_9[t] x_{11}[t] - \theta_{17} x_{13}[t]$,
$x_{14}'[t] == \theta_{20} - \theta_{12} x_{14}[t] - \theta_{22} x_{14}[t]$,
$x_{15}'[t] == \theta_{28} x_7[t] x_{11}[t] - \theta_{26} x_{15}[t]$ };

In[4]:= **output = {$x_4[t], x_5[t], x_6[t], x_{10}[t], x_{12}[t]$};**

In[5]:= **res = IdentifiabilityAnalysis[sys, output, vars, params, t]**

Out[5]= IdentifiabilityAnalysisData[True, <>]

In[6]:= **output = {$x_4[t], x_5[t], x_6[t], x_{10}[t]$};**

In[7]:= **res = IdentifiabilityAnalysis[sys, output, vars, params, t]**

Out[7]= IdentifiabilityAnalysisData[False, <>]

In[8]:= **res["UnidentifiableParameters"]**

Out[8]= $\{\theta_1, \theta_2, \theta_5, \theta_8, \theta_{12}, \theta_{16}, \theta_{18}, \theta_{20}, \theta_{22}, \theta_{25}, \theta_{28}, x_1, x_2, x_3, x_7, x_8, x_9, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}$

As demonstrated above, skipping just one of the five ouputs results in a large number of entities becoming unidentifiable. The identifiability analysis of this system given one set of outputs takes only a few seconds on a standard laptop.

Example *The JAK-STAT signaling pathway.*
A model describing the JAK-STAT pathway (Yamada et al. [2003]) contains 31 state variables and 51 parameters and its differential equations are given below.

The runtime for the identifiability analysis depends on the number of output signals. If for example all state variables in this case are used as output signals (of course unrealistic in practice), the program finishes in a few seconds on a standard laptop. If only a few output signals are selected, it takes a couple of minutes for the program to return the result.

For local structural identifiability of this model, it can be shown to be enough to measure a few carefully selected output signals. For example, measuring $x_{10}[t]$ and $x_{31}[t]$ is enough, but if selecting other state variables as output signals then no set excluding $x_{31}[t]$ can be shown to give an identifiable system.

$$
\begin{aligned}
\dot{x}_1 &= -2\left(x_1^2\theta_1 - x_2\theta_2\right) - x_1x_4\theta_4 + x_6\theta_5 - x_1x_5\theta_7 + x_7\theta_8 \\
\dot{x}_2 &= x_1^2\theta_1 - x_2\theta_2 + x_3\theta_3 - x_2x_4\theta_9 + x_8\theta_{10} \\
\dot{x}_3 &= -x_3\theta_3 + x_{14}^2\theta_{23} - x_3\theta_{24} - x_3x_{16}\theta_{30} + x_{27}\theta_{31} \\
\dot{x}_4 &= -x_1x_4\theta_4 + x_6\theta_5 + x_6\theta_6 - x_2x_4\theta_9 + x_8\theta_{10} + x_8\theta_{11} \\
\dot{x}_5 &= x_6\theta_6 - x_1x_5\theta_7 + x_7\theta_8 - x_5\theta_{12} \\
\dot{x}_6 &= x_1x_4\theta_4 - x_6\theta_5 - x_6\theta_6 \\
\dot{x}_7 &= x_1x_5\theta_7 - x_7\theta_8 + x_8\theta_{11} \\
\dot{x}_8 &= x_2x_4\theta_9 - x_8\theta_{10} - x_8\theta_{11} \\
\dot{x}_9 &= x_5\theta_{12} - x_9x_{14}\theta_{16} + x_{15}\theta_{17} - x_9x_{22}\theta_{19} + \\
&\quad + x_{23}\theta_{20} - x_9x_{20}\theta_{21} + x_{21}\theta_{22} + x_{25}\theta_{29} + x_{26}\theta_{47} \\
\dot{x}_{10} &= -x_{10}\theta_{13} + \tfrac{x_2\theta_{14}}{x_2+\theta_{15}} \\
\dot{x}_{11} &= x_{10}\theta_{13} - x_{11}\theta_{50} \\
\dot{x}_{12} &= -x_{12}x_{29}\theta_{41} + x_{30}\theta_{42} \\
\dot{x}_{13} &= -x_{13}x_{20}\theta_{25} + x_{22}\theta_{26} + x_{26}\theta_{47} - x_{13}\theta_{48} + x_{11}\theta_{51} \\
\dot{x}_{14} &= -x_9x_{14}\theta_{16} + x_{15}\theta_{17} - 2\left(x_{14}^2\theta_{23} - x_3\theta_{24}\right) + x_{21}\theta_{27} + \\
&\quad - x_{14}x_{20}\theta_{33} + x_{24}\theta_{34} - x_{14}x_{16}\theta_{35} + x_{25}\theta_{36} \\
\dot{x}_{15} &= x_9x_{14}\theta_{16} - x_{15}\theta_{17} + x_{27}\theta_{28} \\
\dot{x}_{16} &= x_{27}\theta_{28} + x_{25}\theta_{29} - x_3x_{16}\theta_{30} + x_{27}\theta_{31} - x_{14}x_{16}\theta_{35} + x_{25}\theta_{36} \\
\dot{x}_{17} &= -x_{17}\theta_{18} + x_{18}x_{20}\theta_{39} - x_{17}\theta_{40} \\
\dot{x}_{18} &= x_{17}\theta_{18} - x_{18}x_{20}\theta_{39} + x_{17}\theta_{40} - x_{18}x_{23}\theta_{45} + x_{26}\theta_{46} + x_{26}\theta_{47} \\
\dot{x}_{19} &= x_{17}\theta_{18} - x_{19}\theta_{32} + x_{28}^2\theta_{37} - x_{19}\theta_{38} + x_{26}\theta_{47} \\
\dot{x}_{20} &= -x_9x_{20}\theta_{21} + x_{21}\theta_{22} - x_{13}x_{20}\theta_{25} + x_{22}\theta_{26} + x_{21}\theta_{27} \\
&\quad + x_{19}\theta_{32} - x_{14}x_{20}\theta_{33} + x_{24}\theta_{34} - x_{18}x_{20}\theta_{39} + x_{17}\theta_{40} \\
\dot{x}_{21} &= x_9x_{20}\theta_{21} - x_{21}\theta_{22} - x_{21}\theta_{27} \\
\dot{x}_{22} &= -x_9x_{22}\theta_{19} + x_{23}\theta_{20} + x_{13}x_{20}\theta_{25} - x_{22}\theta_{26} \\
\dot{x}_{23} &= x_9x_{22}\theta_{19} - x_{23}\theta_{20} - x_{18}x_{23}\theta_{45} + x_{26}\theta_{46} \\
\dot{x}_{24} &= x_{14}x_{20}\theta_{33} - x_{24}\theta_{34} \\
\dot{x}_{25} &= -x_{25}\theta_{29} + x_{14}x_{16}\theta_{35} - x_{25}\theta_{36} \\
\dot{x}_{26} &= x_{18}x_{23}\theta_{45} - x_{26}\theta_{46} - x_{26}\theta_{47} - x_{26}\theta_{49} \\
\dot{x}_{27} &= -x_{27}\theta_{28} + x_3x_{16}\theta_{30} - x_{27}\theta_{31} \\
\dot{x}_{28} &= -2\left(x_{28}^2\theta_{37} - x_{19}\theta_{38}\right) + x_{30}\theta_{43} - x_{28}\theta_{44} \\
\dot{x}_{29} &= -x_{12}x_{29}\theta_{41} + x_{30}\theta_{42} \\
\dot{x}_{30} &= x_{12}x_{29}\theta_{41} - x_{30}\theta_{42} - x_{30}\theta_{43} + x_{28}\theta_{44} \\
\dot{x}_{31} &= x_{26}\theta_{49}
\end{aligned}
$$

Example *The Ras Signaling Pathway* (Wolf et al. [2007]). The model of the Ras Signaling Pathway has 67 state variables and about 100 parameters. The network is highly connected and it turns out that it is enough to measure only one state variable to get local structural identifiability. This variable does not even have to be carefully selected (in fact, all variables we tested were ok as single outputs).

The runtime for the identifiability analysis again depends on the number of output signals. If all variables are used as output signals (again unlikely in practice), the program finishes in minutes, and in the extreme case of only one output signal, the analysis takes a few hours.

Example *A MAP Kinase Cascade.*
This is a very large model (Schoeberl et al. [2002]) with about 100 variables and 100 parameters. The time required for the analysis is highly dependent on the set of output signals selected for analysis. If, for example, all variables are selected as output signals, the algorithm finishes in half an hour. On the other extreme end, with only one output signal, much more must be calculated before reaching the result, so in this case the run takes about two days. This example shows that even systems with about 100

variables and 100 parameters can be analyzed for local structural identifiability on a standard desktop computer in reasonable time for any set of output functions. A careful analysis of what must be measured for the MAP system to be identifiable is not the aim of this paper but the topic of a paper in preparation, Anguelova et al. [2012].

6. CONCLUSIONS

In this paper we have demonstrated the feasibility of determining (local) structural identifiability for large scale dynamic systems using a semi-numerical probabilistic method. The Mathematica implementation takes the system equations, parametrized initial conditions, and outputs written in a natural syntax and computes a number of properties of interest of the system such as identifiability and the sets of identifiable or unidentifiable parameters. The computations, also on quite large systems, are carried out on a standard desktop computer in reasonable time.

ACKNOWLEDGEMENTS

We are grateful to Prof Alexandre Sedoglavic for providing us with source code and documentation for his `ObservabilityTest` program.

REFERENCES

M Anguelova. *Observability and identifiability of nonlinear systems with applications in biology.* PhD thesis, Chalmers University of Technology and Gothenburg University, Sweden, 2007.

M Anguelova, J Karlsson, and M Jirstrand. Minimal output sets for identifability. *Manuscript in preparation*, 2012.

S Audoly, G Bellu, L D'Angio, M P Saccomani, and C Cobelli. Global identifiability of nonlinear models of biological systems. *IEEE Trans Biomed Eng*, 48(1):55–65, 2001.

S Diop and M Fliess. On nonlinear observability. In *Proc First Europ Control Conf*, pages 152–157, 1991.

N.D. Evans and M.J. Chappell. Extensions to a procedure for generating locally identifiable reparameterisations of unidentifiable systems. *Mathematical Biosciences*, 168 (2):137–159, 2000.

R Hermann and A J Krener. Nonlinear controllability and observability. *IEEE Trans on Aut. Control*, 22(5):728–740, 1977.

A Jansson and M Jirstrand. Biochemical modeling with Systems Biology Graphical Notation. *Drug Discovery Today*, 15(9-10):365–370, 2010.

R E Kalman. On the general theory of control systems. In *Proc ICAC*, volume 1, pages 481–492, 1961.

N. Le Novère et al.. The Systems Biology Graphical Notation. *Nature Biotechnology*, 27:735–741, 2009.

P Lindskog. *Methods, Algorithms and Tools for System Identification Based on Prior Knowledge.* PhD Thesis no 436, Linköping University, May 1996.

T Lipniacki, P Paszek, A R Brasier, B Luxon, and Kimmel M. Mathematical model of nf-κb regulatory module. *J Theoretical Biology*, 228:195–215, 2004.

L Ljung and T Glad. Parametrization of nonlinear model structures as linear regressions. In *Proc 11th IFAC world congress*, pages 67–71, 1990.

G Margaria, E Riccomagno, M J Chappell, and H P Wynn. Differential algebra methods for the study of the structural identifiability of rational polynomial state-space models in the biosciences. *Math Biosci*, 174:1–26, 2001.

F Ollivier. *'Le problème de l'identifiabilité structurelle globale: approche théorique, méthodes effectives et bornes de complexité'.* PhD thesis, École polytechnique, France, 1990.

H Pohjanpalo. System identifiability based on the power series expansion of the solution. *Math Biosci*, 41(1-2): 21–33, 1978.

M.P. Saccomani, S. Audoly, and L. D'Angio. Parameter identifiability of nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.

B Schoeberl, C Eichler-Jonsson, E D Gilles, and G Muller. Computational modeling of the dynamics of the map kinase cascade activated by surface and internalized egf receptors. *Nat Biotech*, 20:370–375, 2002.

A Sedoglavic. A probabilistic algorithm to test local algebraic observability in polynomial time. *J Symbolic Comput*, 33:735–755, 2002.

S Vajda, KR Godfrey, and Rabitz H. Similarity transformation approach to identifiability analysis of nonlinear compartment models. *Math Biosci*, 93(2):271–248, 1989.

E D Walter. *Identifiability of parametric models.* Pergamon Books Ltd, 1987.

J Wolf, S Dronov, F Tobin, and I Goryanin. The impact of the regulatory design on the response of epidermal growth factor receptor-mediated signal transduction towards oncogenic mutations. *FEBS Journal*, 274(21): 5505–5517, 2007. ISSN 1742-4658.

S Yamada, S Shiono, A Joo, and A Yoshimura. Control mechanism of jak/stat signal transduction pathway. *FEBS Lett*, 534(1-3):190–196, 2003.

J.W.T. Yates, N.D. Evans, and M.J. Chappell. Structural identifiability analysis via symmetries of differential equations. *Automatica*, 45(11):2585–2591, 2009.