Mathematical Biosciences 239 (2012) 139–153

Contents lists available at SciVerse ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Minimal output sets for identifiability

Milena Anguelova [a,*,1], Johan Karlsson [b,2], Mats Jirstrand [b,2]

[a] Imego AB, P.O. Box 53071, SE-400 14 Gothenburg, Sweden
[b] Fraunhofer-Chalmers Centre, Chalmers Science Park, SE-412 88 Gothenburg, Sweden

ARTICLE INFO

Article history:
Received 17 August 2011
Received in revised form 27 April 2012
Accepted 30 April 2012
Available online 16 May 2012

Keywords:
Observability
Structural identifiability
Nonlinear systems
Lie point symmetries
Experimental design
Signalling pathways

ABSTRACT

Ordinary differential equation models in biology 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 uniquely structurally identifiable with the given set of outputs selected as measurements. In designing an experiment for the purpose of parameter estimation, given a set of feasible but resource-consuming measurements, it is useful to know which ones must be included in order to obtain an identifiable system, or whether the system is unidentifiable from the feasible measurement set.

We have developed an algorithm that, from a user-provided set of variables and parameters or functions of them assumed to be measurable or known, determines all subsets that when used as outputs give a locally structurally identifiable system and are such that any output set for which the system is structurally identifiable must contain at least one of the calculated subsets.

The algorithm has been implemented in Mathematica and shown to be feasible and efficient. We have successfully applied it in the analysis of large signalling pathway models from the literature.

1. Introduction

Mathematical models of the chemical reactions in living cells often consist of systems of ordinary differential equations describing the reaction rates. In many applications, the parameters in these models are not directly measurable but only accessible indirectly through their impact on measured entities, which typically are time varying output signals responding to some applied perturbation inputs to the system under study. There are many parameter estimation methods, which, given a parameterized 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 parameter values that equally well describe the input-output data.

A large amount of literature has been devoted to the theoretical characterization of identifiability, starting already in Kalman's work [1] for linear systems and in [2] for the nonlinear case, and continuing until today, where [3–15] are just a few references. Before the work of Sedoglavic [12], the available methods for testing structural observability/identifiability of nonlinear systems relied on characteristic set or standard bases computation [6,8,7,10,11] or the local state variable isomorphism approach [4,9,16]. The complexity in the number of variables and parameters of these methods grows too fast for them to be applicable to many realistic biological models. In [12], a probabilistic seminumerical algorithm is presented which finally allows for local structural identifiability to be analyzed even for metabolic and signalling models with a few hundred variables and parameters.

Structural identifiability, or rather the lack of it, is strongly related to the existence of symmetries in the system which leave the measured output invariant, i.e., the existence of variable and parameter transformations that leave the output function(s) unchanged. In [17,18], an algorithm is proposed for the calculation of polynomial Lie point symmetries of systems of rational ordinary differential equations. The existence of these methods for identifiability analysis and symmetry calculation makes the problem of designing an identifiable system, discussed here, tractable.

Any system can be made structurally identifiable by choosing enough variables or functions of them to be measured. At the same time, in order to decrease time and effort it is often desired to make as few measurements as possible. Given a set of feasible, but

* Corresponding author. Tel.: +46 31 7501834.
E-mail addresses: milena.anguelova@imego.com (M. Anguelova), johan.karlsson@fcc.chalmers.se (J. Karlsson), mats.jirstrand@fcc.chalmers.se (M. Jirstrand).

[1] This work was carried out while the author was affiliated with the Fraunhofer–Chalmers Centre.

[2] Tel.: +46 31 7724000; fax: +46 31 7724260.

0025-5564/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.mbs.2012.04.005

resource-consuming measurements for the purpose of parameter estimation, it is therefore helpful to be able to find small subsets of these, that still allow the unique determination of all parameters. In this paper, we focus on the problem of determining minimal sets of variables or functions of them that when used as outputs, make the model parameters and initial conditions structurally identifiable. This problem has not received much attention previously, possibly due to the difficulty of testing structural identifiability in the first place. The only reference we have been able to find is [19], where an iterative majorization-minimization algorithm is proposed for choosing a minimal set of state variables that when chosen as output ensures identifiability. Its computational complexity scales badly with the size of the problem ([19]) and does not allow for the treatment of large ODE models. In this paper, we propose an algorithm, the feasibility and efficiency of which has been shown through a Mathematica implementation. Based on model structure, the algorithm determines all *minimal output sets* where variables and parameters or functions of them are selected from a user-provided set of expressions assumed to be measurable or known so that a locally structurally identifiable system is obtained and such that any output set for which the system is structurally identifiable contains at least one of the calculated sets as a subset. The measurements are assumed to be perfect noise-free time series since identifiability with this set-up is a necessary condition for identifiability from realistic experimental data. The algorithm with its Mathematica implementation has been successfully applied in the analysis of, for example, large-scale signalling pathway models.

2. Models, identifiability and problem formulation

The biochemical models considered in this paper are systems of differential equations, representing mass balances for each biochemical entity. They can normally be written in the form

$$\begin{cases} \dot{x} = f(t, x, u, \theta) \\ x(t_0) = x^0(\theta) \end{cases} \tag{1}$$

where $x \in \mathbb{R}^n$ denotes the intracellular concentrations of the biochemical entities and $\dot{x}$ denotes the time derivative of x. The variable $u \in \mathbb{R}^r$ denotes the exogenous input to the system. The symbol $\theta \in \mathbb{R}^p$ denotes the vector of parameters in the rate expressions or expressions for initial conditions of the model. In this paper we assume that all elements of the vector-valued function f are rational functions of their arguments which is a common case for biochemical models. At some places in the paper the dependence on t will not be explicitly written for the sake of brevity. The initial conditions of the system may sometimes have known numerical values, or they may be given as expressions of parameters. If no information is given about the initial conditions we shall assume that the $x_i(t_0)$:s are unknown parameters to be determined from the outputs.

The property of (local) structural identifiability guarantees the (local) uniqueness of the parameters θ for a given input-output structure corresponding to a set of measurements for the purpose of parameter estimation. This set of measurements usually consists of some metabolic concentrations and/or fluxes or combinations of these, denoted by the vector y. The latter can then be written as a vector-valued function $g \in \mathbb{R}^m$ of the variables in x and the parameters θ:

$$y = g(x, \theta), \tag{2}$$

or

$$y(t) = g(x(t), \theta), \tag{3}$$

depending on the context.

For almost all values of $x(t_0)$ and θ, the function $y(t)$ is assumed to be analytic in an open interval around t_0 and its Taylor series converges to the function value at each point in the interval. The time derivatives of $y(t)$ at t_0 are given by $\frac{dy_i^{(j)}}{dt}|_{t=t_0} = L_f^j g_i$, where $L_f := \frac{\partial}{\partial t}|_{t=t_0} + \sum_{i=1}^{n} f_i \frac{\partial}{\partial x_i} + \sum_{j\in\mathbb{N}} \sum_{i=1}^{r} u_i^{(j+1)} \frac{\partial}{\partial u_i^{(j)}}$ is the so-called Lie-derivative. The expressions $L_f^j g = (L_f^j g_1, \ldots, L_f^j g_m)$ are functions of $t_0, x(t_0)$ and θ. With t_0 assumed to be fixed, consider the $(n+p)\times m$ equations for $x(t_0)$ and θ:

$$\begin{aligned} y(t_0) &= g(x(t_0), \theta) \\ \dot{y}(t_0) &= L_f^1 g(x(t_0), \theta) \\ &\vdots \\ y^{(n+p-1)}(t_0) &= L_f^{(n+p-1)} g(x(t_0), \theta) \end{aligned} \tag{4}$$

It can be shown (see [11,12,15]) that any further $L_f^j g, j \geqslant n+p$ are dependent on the previous ones and so there is no need for further equations based on the output derivatives to be considered. Consider also the n equations corresponding to the initial conditions:

$$x(t_0) = x^0(\theta). \tag{5}$$

We will define the system (1) to be *locally structurally identifiable* if there is a locally unique solution to Eqs. (4) and (5) for almost all values of $x(t_0)$ and θ. This is the property referred to throughout the paper.

The above definition corresponds to local algebraic observability as defined in [12] but we have avoided the introduction of more advanced algebraic terminology.

Local structurally identifiability still allows for a countable sets of parameters and initial conditions to produce exactly the same output, that is, it only guarantees uniqueness of the parameters and initial conditions within an open interval. However, it is clear that local structurally identifiability is a necessary condition for any parameter and state estimation procedure to be successful and ensuring this property is therefore an important first step in experimental design.

In the case of unknown initial conditions, both algebraic (as in [12]) and differential-geometric (as in [2]) points of view lead to the same test for structural identifiability, see [15], the so-called rank test. The initial state values are important for the structural identifiability analysis, as shown in [13]. In our case, allowing for parameterized initial conditions, the rank test is modified in the following way. Instead of calculating rank, a basis for the nullspace of the matrix

$$\left[\begin{array}{ccccccc} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} & \frac{\partial g_1}{\partial \theta_1} & \cdots & \frac{\partial g_1}{\partial \theta_p} \\ \vdots & & & & & \\ \frac{\partial g_m}{\partial x_1} & \cdots & \frac{\partial g_m}{\partial x_n} & \frac{\partial g_m}{\partial \theta_1} & \cdots & \frac{\partial g_m}{\partial \theta_p} \\ \frac{\partial L_f g_1}{\partial x_1} & \cdots & \frac{\partial L_f g_1}{\partial x_n} & \frac{\partial L_f g_1}{\partial \theta_1} & \cdots & \frac{\partial L_f g_1}{\partial \theta_p} \\ \vdots & & & & & \\ \frac{\partial L_f g_m}{\partial x_1} & \cdots & \frac{\partial L_f g_m}{\partial x_n} & \frac{\partial L_f g_m}{\partial \theta_1} & \cdots & \frac{\partial L_f g_m}{\partial \theta_p} \\ \vdots & & & & & \\ \frac{\partial L_f^{n+p-1} g_1}{\partial x_1} & \cdots & \frac{\partial L_f^{n+p-1} g_1}{\partial x_p} & \frac{\partial L_f^{n+p-1} g_1}{\partial \theta_1} & \cdots & \frac{\partial L_f^{n+p-1} g_1}{\partial \theta_p} \\ \vdots & & & & & \\ \frac{\partial L_f^{n+p-1} g_m}{\partial x_1} & \cdots & \frac{\partial L_f^{n+p-1} g_m}{\partial x_n} & \frac{\partial L_f^{n+p-1} g_m}{\partial \theta_1} & \cdots & \frac{\partial L_f^{n+p-1} g_m}{\partial \theta_p} \end{array}\right]\Bigg|_{x(t_0)=x^0(\theta)} \tag{6}$$

is calculated. Then, we analyze whether any of the matrix' null-vectors provide directions in which $x(t_0)$ and θ can be changed while keeping (5) fulfilled. Each null-vector $v = (v_1, \ldots, v_{n+p})$ of the above matrix corresponds to a derivation $\sigma = v_1 \frac{\partial}{\partial x_1} + \cdots + v_n \frac{\partial}{\partial x_n} + v_{n+1} \frac{\partial}{\partial \theta_1} + \cdots + v_{n+p} \frac{\partial}{\partial \theta_p}$. If this derivation σ is such that

$$\sigma(x(t_0) - x^0(\theta))|_{x(t_0)=x^0(\theta)} = 0, \tag{7}$$