Mathematical Biosciences 282 (2016) 147–161

Contents lists available at ScienceDirect

Mathematical Biosciences

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

On the relationship between sloppiness and identifiability

Oana-Teodora Chis[a,b], Alejandro F. Villaverde[a], Julio R. Banga[a], Eva Balsa-Canto[a,1,*]

[a] Bioprocess Engineering Group, Institute of Marine Research (IIM-CSIC), Vigo 36208, Spain
[b] Technological Institute for Industrial Mathematics (ITMATI), Santiago de Compostela 15782, Spain

ARTICLE INFO

Article history:
Received 5 August 2016
Revised 21 October 2016
Accepted 23 October 2016
Available online 24 October 2016

Keywords:
Sloppiness
Identifiability
Parameter estimation
Optimal experimental design

ABSTRACT

Dynamic models of biochemical networks are often formulated as sets of non-linear ordinary differential equations, whose states are the concentrations or abundances of the network components. They typically have a large number of kinetic parameters, which must be determined by calibrating the model with experimental data. In recent years it has been suggested that dynamic systems biology models are universally sloppy, meaning that the values of some parameters can be perturbed by several orders of magnitude without causing significant changes in the model output. This observation has prompted calls for focusing on model predictions rather than on parameters. In this work we examine the concept of sloppiness, investigating its links with the long-established notions of structural and practical identifiability. By analysing a set of case studies we show that sloppiness is not equivalent to lack of identifiability, and that sloppy models can be identifiable. Thus, using sloppiness to draw conclusions about the possibility of estimating parameter values can be misleading. Instead, structural and practical identifiability analyses are better tools for assessing the confidence in parameter estimates. Furthermore, we show that, when designing new experiments to decrease parametric uncertainty, designs that optimize practical identifiability criteria are more informative than those that minimize sloppiness.

1. Introduction

Dynamic models of cellular processes describe the interactions among molecular entities – for example, proteins, transcripts or regulatory sites – that determine cellular behaviour. Such models typically consist of non-linear ordinary differential equations, whose state variables represent the concentrations of the network components and whose parameters characterize the reaction kinetics. Unfortunately, in most cases the parameter values are unknown, or only rough estimates are available. It is therefore necessary to calibrate the model using time-series experimental data [28]. The task of estimating the parameter values is an optimization problem, whose objective is to minimize a cost function that quantifies the differences between model predictions and experimental data. In dynamic models of biochemical systems this problem is typically characterized by limited observability, large number of parameters, and scarce, poor quality data [7]. As a consequence, its solution is in general challenging and computationally expensive, even with efficient optimization methods. In addition, data limitations often lead to great uncertainty in the parameter estimates [47,48].

During the last decade, several works [22,23,34,44–46,52] have introduced and elaborated the concept of *sloppiness*. The parameters of a dynamic model can be divided into stiff (those that can be determined with great certainty) and sloppy (those that can vary by orders of magnitude without influencing significantly the model output), although it is difficult to establish a clear cut-off between both categories. The sloppiness of a model is quantified from the distribution of the eigenvalues of its Fisher information matrix; a separation of more than 3 orders of magnitude in the eigenvalues qualifies a model as sloppy.

It has been claimed that dynamic systems biology models are *universally sloppy* [23], and therefore it is not possible to obtain accurate estimates of their parameters. This idea has been cited in many publications and, unfortunately, has sometimes led to misinterpretations. Since parameter estimation is often an arduous task in practice, it is tempting to use the notion of sloppiness to argue that it is not necessary nor possible to uniquely determine the parameter values, thus justifying that no further efforts are invested to it (see for example [16,19,31,39,51]). The suggestion that sloppiness is a universal – or, more precisely, ubiquitous – property of systems biology models has spurred a debate: should modellers desist from trying to estimate precise values for the parameters and, instead, focus on characterizing model *predictions*? Exploring

* Corresponding author.
E-mail address: ebalsa@iim.csic.es (E. Balsa-Canto).
[1] To whom correspondence should be addressed.

http://dx.doi.org/10.1016/j.mbs.2016.10.009
0025-5564/

this direction, Cedersund and coworkers [11–13] coined the term "core predictions" to denote specific model outcomes that can be uniquely determined, even if parameter values cannot. The parameter regions complying with core predictions can be found using optimization, at least for models of moderate size [11].

In [43] the origin of sloppiness was traced back to the structure of the sensitivity matrix, which contains the sensitivities of the model outputs with respect to the parameters. *Experimental design* was proposed as a way of reducing sloppiness, concluding that the intensity of the effect is highly dependent on the available data, thus challenging the universality of the property. The importance of the role played by experimental design in this task had been stressed e.g. in [2,32]. Finally, it is worth mentioning that sloppiness has sometimes been considered as an indication of biological robustness [21,30].

In this work we aim at clarifying the application and implications of sloppiness, to avoid certain misconceptions. We study the role played by model parameters using the well established framework of *identifiability*, which has a long history of application in dynamical systems [50], including biological models. Indeed, while the study of parameter identifiability has been present in the systems and control literature for decades, many methodological advances in the field have been motivated by biological applications, starting in the 1970s and 1980s [8,20,27,37] and continuing until the present day [49]. Identifiability-based concepts are rigorously defined and well understood, and they can be analysed with a large number of techniques. Hence it is of interest to clarify the connection between sloppiness and (lack of) identifiability [19,38,41], which, despite recent developments, is still incompletely understood. Here we study the relationship between sloppiness and identifiability from both structural and practical points of view. Using a set of case studies, we inquire to which extent is sloppiness determined by the model structure, and how it is influenced by the quantity and quality of experimental data. Then we explore optimal experiment design alternatives that reduce sloppiness and improve identifiability, and clarify the connections between both concepts. We conclude that identifiability analysis can be more insightful than sloppiness for characterizing the mapping between parameters and outputs.

2. Methods

We consider general nonlinear models of the form:

$$\Sigma(\mathbf{p}):\begin{cases}\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x},\mathbf{p})+\sum_{j=1}^{n_u}\mathbf{g}_j(\mathbf{x},\mathbf{p})\mathbf{u}_j,\\ \mathbf{y}=\mathbf{h}(\mathbf{x},\mathbf{p}),\ \mathbf{x}(t_0)=\mathbf{x}_0(\mathbf{p})\end{cases}\tag{1}$$

where $\mathbf{x}=(x_1,\dots,x_{n_x})\in\mathbf{R}^{n_x}$ is the state vector, $\mathbf{u}=(u_1,\dots,u_{n_u})\in\mathbf{R}^{n_u}$ a n_u-dimensional input (control) vector, and $\mathbf{y}=(y_1,\dots,y_{n_y})\in\mathbf{R}^{n_y}$ is the n_y-dimensional output (experimentally observed quantities). The vector of unknown parameters is denoted by $\mathbf{p}=(p_1,\dots,p_{n_p})\in\mathbf{P}$, and is assumed to belong to an open and connected subset of $\mathbf{R}^{n_p}$. The entries of $\mathbf{f}$, $\mathbf{g}=(\mathbf{g}_1,\dots,\mathbf{g}_{n_u})$ and $\mathbf{h}$ are analytic functions of their arguments. These functions and the initial conditions may depend on the parameter vector $\mathbf{p}\in\mathbf{P}$.

Note that the model in Eq. (1) is composed of two different elements: (i) a set of ordinary differential equations (ODEs), describing the system dynamics and (ii) the observation function, which relates states (typically concentrations or amounts) and measurements. In this work we consider that the mathematical structure of the system dynamics ($\mathbf{f}$, $\mathbf{g}$) can by no means be modified, whereas the mathematical structure of the observation function ($\mathbf{h}$) may eventually be modified by the experimental scheme (which, in fact, leads to a different model).

2.1. Parameter estimation

The above representation (1) is a sufficiently accurate mathematical description of the real system, i.e. the only uncertainty is represented by the vector of unknown parameters. This means that, in principle, there is a unique "true" value of the parameters, denoted by $\mathbf{p}^*=(p_1^*,\dots,p_{n_p}^*)$, which allows the model to reproduce a given data set and to predict the system behaviour. This vector $\mathbf{p}^*$ is computed by means of data fitting, i.e. by solving an optimization problem devoted to minimizing the log-likelihood function, which for Gaussian experimental noise reads:

$$\chi^2(\mathbf{p})=\sum_{e=1}^{n_e}\sum_{o=1}^{n_y}\sum_{s=1}^{n_s}\frac{[\mathbf{y}_{e,o,s}(\mathbf{p},t_s)-\tilde{\mathbf{y}}_{e,o,s}]^2}{\sigma_{e,o,s}^2},\tag{2}$$

where n_e is the number of experiments, n_y the number of observables for each experiment, and n_s the number of sampling times; $\mathbf{y}_{e,o,s}(\mathbf{p},t_s)$ denotes the output of the model (1) for the sampling time t_s under the experimental conditions e; $\tilde{\mathbf{y}}_{e,o,s}$ is the corresponding experimental data; and $\sigma_{e,o,s}^2$ is the variance of the measurement noise.

2.2. Structural identifiability

Structural identifiability analysis studies the possibility of finding a unique solution to the parameter estimation problem, assuming perfect experimental data (i.e. noise-free and continuous in time) [50]. A parameter p_i, $i=1,\dots,n_p$ is *structurally globally (or uniquely) identifiable* if for almost any $\mathbf{p}^*\in\mathbf{P}$, $\Sigma(\mathbf{p})=\Sigma(\mathbf{p}^*)\Rightarrow p_i=p_i^*$, whereas a parameter p_i, $i=1,\dots,n_p$ is *structurally locally identifiable* if for almost any $\mathbf{p}^*\in\mathbf{P}$ there exists a neighbourhood $\mathbf{V}(\mathbf{p}^*)$ such that $\mathbf{p}\in\mathbf{V}(\mathbf{p}^*)$ and $\Sigma(\mathbf{p})=\Sigma(\mathbf{p}^*)\Rightarrow p_i=p_i^*$.

In some cases, an unidentifiable parameter may be made identifiable by including more measured outputs in the observation function, $\mathbf{h}$. This modification leads to a new model with a different structure. In other cases, however, the model may be structurally unidentifiable even if all states are accessible to the experimentation, i.e. $\mathbf{y}=\mathbf{x}$. In this case it will not be possible to avoid the lack of identifiability.

Recent reviews [18,35] compare alternative methods to perform global structural identifiability analysis for nonlinear models; additionally, state of the art local methods are described in [15,49]. In this work we adopt the MATLAB based GenSSI toolbox [17], which combines the generating series approach with identifiability *tableaus*. The underlying idea of the generating series approach is that the observables $\mathbf{y}$ can be expanded in series with respect to time and inputs around a given time point (t_0), and that the uniqueness of the series coefficients guarantees the structural identifiability of the model. The series coefficients are computed by means of successive Lie derivative of $\mathbf{h}$ along the vector fields $\mathbf{f}$ and $\mathbf{g}$. The identifiability *tableaus* correspond to the Jacobian of the Lie derivatives with respect to the model parameters, and help to decide on global or local structural identifiability of the model [5].

2.3. Sloppiness

Parameter sloppiness can be quantified by means of the eigenvalues of the Hessian of the log-likelihood function (Eq. (2)) as evaluated in the optimal value of the parameters $\mathbf{p}^*$. The Fisher information matrix ($\mathcal{F}$) can be used as an approximation of the Hessian:

$$\mathcal{F}=E\left(\left[\frac{\partial\chi^2(\mathbf{p})}{\partial\mathbf{p}}\right]^T\left[\frac{\partial\chi^2(\mathbf{p})}{\partial\mathbf{p}}\right]\right),\tag{3}$$