Structural vs Practical Identifiability in System Biology

Maria Pia Saccomani

Department of Information Engineering, University of Padova,
Via Gradenigo6a, 35131 Padova, Italy
mariapia.saccomani@unipd.it
http://www.dei.unipd.it

Abstract. The relevance of model identifiability in system biology is well known. The aim of this paper is to discuss and possibly clarify the role and differences of two different methodologies of testing identifiability of models described by differential equations. One is called *structural identifiability* and does not require to use data collected in the experiment which, instead, *data-based approaches* do. The two methods are compared and evaluated in the identifiability analysis of a much quoted biological model, the core model of erythropoietin (Epo) and Epo receptor (EpoR) interaction.

Keywords: system biology, differential equations model, structural identifiability, practical identifiability, parameter estimation, experiment design.

1 Introduction

Mathematical modeling of biological systems is becoming a standard approach to investigate complex dynamic, non-linear interaction mechanisms in cellular processes, like signal transduction pathways and metabolic networks [7]. These mechanisms are often modeled by ordinary differential equations involving parameters such as reaction rates. For example, the Michaelis-Menten equation is often used to describe the internal structure of the biochemistry of the system, assuming that diffusion is fast compared to reaction rates. The system parameters contain key information but in general they can only be measured indirectly as it is usually not possible to measure directly the dynamics of every portion of the system. The recovery of parameter values can then only be approached indirectly as a parameter estimation problem starting from external, input-output measurements [18]. In this context, the first question is whether the parameters of the model can be determined, at least for suitable input functions, assuming that all observable variables are error free. This is the property called *a priori* or *structural identifiability* of the model. It is a property of the model alone and of course depends on how it is parameterized. Structural identifiability can (and should) in principle be checked before collecting experimental data. If the postulated model is not structurally identifiable, the parameter estimates which

could, nevertheless, be obtained by some numerical optimization algorithms, will be totally unreliable and random. Obviously, although necessary, structural identifiability is not sufficient to guarantee an accurate identification of the model parameters from real input/output data.
Different methods have been proposed to check structural identifiability of models described by linear and nonlinear differential equations. Some approaches can test *global* (structural) identifiability and provide conclusions about identifiability holding for the whole parameter space [8, 18, 4, 9, 11, 1, 16]. Specifically, structural global identifiability guarantees the possibility of uniquely determining the model parameters from input-output data, under ideal conditions irrespective of the admissible parameter values. Nevertheless, often a weaker property of *local* (structural) identifiability about some specified parameter value, may be sufficient in practice.
It should be stressed that identifiability depends also on the experimental conditions. More precisely, for a fixed model structure and measurement schedule, identifiability does in general depend on the class of admissible input functions acting on the system. Input functions which do not "excite" the system properly may render some parameters invisible from the external output. Structural identifiability analysis is performed under the assumption that the input is *persistently exciting*, see [9, 16] for a precise definition of this condition. Of course the admissible inputs class must contains such persistently exciting functions.
A concept of *practical or data-based identifiability* has also been proposed in the literature [13, 7, 12]. Given a dynamical model described by

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta}) \qquad \mathbf{x}(t_o) = \mathbf{x}_0 \tag{1}$$

$$\mathbf{y}(t) = h((\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta}) + \boldsymbol{\varepsilon}(t) \ := \hat{\mathbf{y}}(t, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}(t) \tag{2}$$

with state $\mathbf{x}(t) \in \mathbb{R}^n$, input $\mathbf{u}(t) \in \mathbb{R}^q$, output $\mathbf{y}(t) \in \mathbb{R}^m$, random measurement noise $\boldsymbol{\varepsilon}(t) \in \mathbb{R}^m$, and unknown parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$, assuming a finite set of N input-output measurements are available, form the average weighted square *prediction error*

$$V_N(\boldsymbol{\theta}) := \frac{1}{N} \sum_{k=1}^{N} [\mathbf{y}(t_k) - \hat{\mathbf{y}}(t_k, \boldsymbol{\theta})]^\top Q_k, [\mathbf{y}(t_k) - \hat{\mathbf{y}}(t_k, \boldsymbol{\theta})] \tag{3}$$

where Q_k are positive semidefinite weights. One says that the system (or the parameter $\boldsymbol{\theta}$) is *practically identifiable* if $V_N(\boldsymbol{\theta})$ has a unique minimum; in other words there is a *unique minimum prediction error estimate*

$$\hat{\boldsymbol{\theta}}_N = \text{Arg} \min_{\boldsymbol{\theta}} V_N(\boldsymbol{\theta}) \tag{4}$$

compare [10]. If the error terms $\boldsymbol{\varepsilon}(t)$ are assumed to be Gaussian the function $V_N(\boldsymbol{\theta})$ is essentially the *likelihood function* of the experiment. The relation between structural identifiability and uniqueness of the minimum has been discussed in detail in the engineering literature, see e.g. Ljung's book or [17]. It can be proven that the former is equivalent to the latter only under additional

assumptions on the data (ergodicity) and only in the limit when the sample size N tends to infinity. This equivalence cannot be guaranteed for finite data records.
The goal of this paper is to discuss and compare the two different methods. It should be said however that the comparison can only be made on the issue of *uniqueness* of the parameters obtainable from the admissible input-output data. Practical identifiability, is based on actual experimental data and consists of procedures based on the analysis of the minima of a likelihood-type function. These methods in general compute numerical parameter estimates. Structural identifiability tests are instead run purely on the model equations and do not provide numerical parameter estimates based on real data. The computation of estimates based on real data, which makes sense only if the model is structurally identifiable, is left to a successive optimization algorithm which is a conceptually different issue.

2 Structural vs Practical Identifiability

Data-based methods seem to be the only choice when collected data, perhaps obtained in a unrepeatable or very expensive experiment, are already at hand. Then the experimenter may want to check if the model parameters can in fact be recovered uniquely from the given data. However practical identifiability tests based on a specific data set cannot give exact answers about structural identifiability and one should therefore resort to heuristics and extensive simulations in order to produce a representative data set to figure out the shape of the function $V_N(\boldsymbol{\theta})$ and verify the presence of a unique minimum experimentally [13]. There are some caveats and some important consequences:

1. If the model happens to be structurally non-identifiable then it is also practically non-identifiable. In fact, if the parameters of the postulated model are not a priori identifiable, then there is no way that the parameters could be uniquely identified in a real-life situation, with a fixed observed input function (perhaps not even sufficiently exciting), when noise in the data is inevitably present and possibly with insufficient data length.
   In principle, for a large family of models, structural identifiability or non-identifiability can be checked by suitable mathematical procedures directly on the model, without the need of collecting experimental data. This may avoid waste of resources for doing useless experiments, given the high costs, not only in economic terms, of biological experiments.
2. If the model is structurally identifiable, it may nevertheless turn out to be practically non-identifiable. However, only by first checking structural identifiability it is possible to know for sure if the problem lays on the experimental data or on the model structure.
3. If the model turns out to be practically non-identifiable, it may be very hard or impossible to assess the causes from practical identifiability tests. The fact may be due to structural non-identifiability or to the paucity the experimental data or to an imprecise reconstruction due to noise of the locus of

the minima of the function $V_N(\boldsymbol{\theta})$. In general practical identifiability tests can hardly suggest alternative experimental strategies to follow in order to obtain identifiability of the model.
This instead may be revealed, in analytic terms, by structural methods. These methods can provide guidelines to simplify the model structure or indicate, before performing the real experiment, when more information (measured outputs) is needed to guarantee unique identifiability [1]. In fact, *a priori* structural identifiability analysis can be helpful in the design of the experiment. Since when dealing with biological/physiological systems, severe constraints exist on experiment design, it is of great interest to check for *minimality* (i.e. non redundancy) of the input-output configuration. There are available tools for checking structural identifiability by which one can also check if the number of inputs and outputs are necessary and sufficient to guarantee a priori unique identifiability [15].

4. Some structural approaches allow to distinguish between global and local identifiability. Traditional tests based on computing the rank of a matrix (e.g the Hessian or Fisher information matrix) at a point [14] are essentially local. Even if it is true that in many applications it may be sufficient to work in a parameter neighborhood specified by experimental data, there are many biological models, where, say, two distinct values of a parameter can discriminate a pathological from a normal state. In these cases a local analysis may be insufficient [5].
5. Finally, one should also be aware of the limitations of the analytic procedures for checking structural identifiability. When the model is very complex, with complicated nonlinearities and a large number of states or unknown parameters, and/or few measurement equations, most algorithms take a very long time to terminate or may even not terminate at all due to computational complexity problems. Some algorithms based on differential algebra, like the one employed in the example below, require the model differential equations to be of polynomial or rational form.

3 A Model of Erythropoietin (EPO) Receptor

In this section we consider a recently proposed dynamic model, [2], addressing the nonlinear processes of ligand-receptor (Epo-EpoR) interaction and trafficking kinetics. In particular, the core model is a development of a previous published model [7] describing the endocytosis of the erythropoietin receptor, that is the process of engulfing substances outside the cell with a membrane and transporting them into cytoplasm. Six species are incorporated in the model, x_i $i = 1, ..., 6$ being the relative concentrations, and all interactions are modeled by mass-action kinetics. The detailed description of the biochemical processes underlying the EPO endocytosis is reported in the referenced paper.
The core model is described by the following nonlinear system of ordinary dif-

ferential equations:

$$\begin{cases} \dot{x}_1(t) = b_{max}k_1 - k_1x_1 - k_{on}x_1x_2 + k_{on}k_Dx_3 + k_{ex}x_4 \\ \dot{x}_2(t) = -k_{on}x_1x_2 + k_{on}k_Dx_3 + k_{ex}x_4 \\ \dot{x}_3(t) = k_{on}x_1x_2 - k_{on}k_Dx_3 - k_ex_3 \\ \dot{x}_4(t) = k_ex_3 - (k_{ex} + k_{dl} + k_{de})x_4 \\ \dot{x}_5(t) = k_{dl}x_4 \\ \dot{x}_6(t) = k_{de}x_4 \\ y_1(t) = x_2 + x_6 \\ y_2(t) = x_3 \\ y_3(t) = x_4 + x_5 \end{cases} \tag{5}$$

The initial conditions are assumed to be zero, except for $x_1(0) = b_{max}$ and $x_2(0) = ic_2$ which need to be estimated from the experimental data. $\boldsymbol{\theta} = [k_1, k_{on}, k_D, k_{ex}, k_e, k_{dl}, k_{de}, b_{max}]$ is the unknown parameter vector, and $\mathbf{y} = [y_1\ y_2\ y_3\,]$ is the measured output of the model.
Before checking identifiability of the model, it is convenient to check some mathematical properties in terms of system and control theory. One can first observe that the model is not in minimal form. In fact, by looking at the system equations (5), some state variables appear to be dependent. One can simplify the model by observing that, for example, $\dot{x}_6 = (k_{de}/k_{dl})\dot{x}_5$ which, integrated by using the known initial conditions, gives: $x_6 = (k_{de}/k_{dl})x_5$ and by substituting x_6 where it appears (only in the first measurement equation). In a similar way, the variable x_3 can be eliminated. In this way the model can be described by only four differential equations, as reported in the input file below. This is done not only for the sake of mathematical simplification, but in order to satisfy system theoretic properties, such as minimality (and accessibility) [16] in absence of which spurious identifiability results may follow. These structural properties must be always investigated beforehand.

4 Identifiability of the Erythropoietin Receptor Model

The question to be addressed is whether the unknown parameter vector $\boldsymbol{\theta}$ in the above (simplified) model is globally identifiable from the experiment. In the recent literature, the practical identifiability of the model (5) has been analyzed with data-based methods based on statistical criteria. In particular, in [13] the profile likelihood [12] approach is used, based on the idea of detecting flatness of the likelihood function $V_N(\boldsymbol{\theta})$ by exploring the parameter space in the direction of least increase in the objective function for each parameter component. The authors actually examine a more complex model of the model (5) described above. The profile likelihood method allows them to study the behavior of the function around a nominal parameter value (see for ex. Fig 3,4,5 of [13]). As mentioned by the authors "a structural nonidentifiability can be visualized by a *perfectly flat valley that is infinitely extended along the corresponding functional relation.* In theory however this flatness does detect non-identifiability only under the assumptions of ergodicity of the data and number of experimental data tending

to infinity. To approximate this ideal situation a very extensive simulation is needed by generating a large number of artificial samples. On the other hand, to check practical identifiability, that is to visualize numerically the "relative flat valley infinitely extended", a threshold has to be assessed. Thus the result is intrinsically approximate as it depends on the choice of the threshold value and may still provide results depending on the particular set of experimental data. In essence, some judgement has to be exercised to properly interpret the results of the simulations.
Anyway, by applying this data-based method, the original model turns out to be non-identifiable. By looking at the shape of the flat valley of the likelihood function, the authors cleverly establish that some parameters satisfy an algebraic equation and therefore cannot be identifiable (called structurally non-identifiable in the paper [13]) while the others are found to be (practically) identifiable except for one (k_{ex}) where the minimum is so flat to be declared "practically non-identifiable". They conclude that the structural non-identifiability is a result of missing information about absolute concentration in the experimental setup. To resolve this structural non-identifiability they enrich the experiment, so as a scale factor parameter becomes known, and to resolve the practical non-identifiability of k_{ex} a new measurement equation is added. In this way the authors define the identifiable model (5).
A techniques for estimating the equation describing a possible locus of minima is described in [7]. The algorithm is based on a non parametric nonlinear regression technique which however can only reveal a very specific functional form of dependence among the parameters. Unless it is a priori known that the model parameters may be only related by a GAM relation, the method does not seem to be able to test global identifiability of the model.
We shall now describe a structural identifiability test based on differential algebra and on the software DAISY (Differential Algebra for Identifiability of SYstems) [3]. This a priori analysis seems to be done here for the first time. The reader is referred to [1, 16] for a detailed documentation of the theory behind the software tool and to [3] for the algorithm. The underlying algorithm permits to eliminate the non-observed state variables from the system and to find the *input-output relation*: a set of polynomial differential equations involving only the variables (u, y) and thus describing all input-output pairs satisfying the original dynamic system. The coefficients of the input-output relation provide a set of (nonlinear) algebraic functions in the unknown $\boldsymbol{\theta}$, These functions form the *exhaustive summary* of the model and can be easily extracted. Identifiability is tested by checking injectivity of the exhaustive summary function with respect to the parameter $\boldsymbol{\theta}$. By applying a computer algebra algorithm; i.e. the Buchberger algorithm, it is possible to compute a Gröbner basis of the system which shows if the parameters satisfy algebraic relations or have one and only one solution, in which case *the model is globally identifiable.*
DAISY checks the global identifiability of the original complex model [14] with the original two measurement equations. In 2-3 seconds, it is found that all the model parameters are globally identifiable except the same five that were found

to be non-identifiable in [13]. This result actually shows that, just by looking at the Gröbner basis computed analytically by the algorithm, it is sufficient to know just one of the five non-identifiable parameters (not necessarily the scale factor parameter) to make the model globally identifiable. This allows for different choices in the design of the experiment, where many constraints exist especially in the biological experimental setup.
The structural test guarantees that, in fact, parameter k_{ex} *is identifiable*. This result reveals that the practical non-identifiability of k_{ex} found in [13] is therefore due only to data problems.
In this case, our approach to structural identifiability has proven in an analytical way the results obtained in [13], and has provided some additional information helpful for the experiment design.
In practice, to check the global identifiability of this model with DAISY, the user has to write the input file in a given format. In the following the input file for the model (5) with the simplification above presented to eliminate redundancy is reported: *Input File of DAISY*

```
WRITE "CORE MODEL (simplified) Becker et al. SCIENCE 2010 Suppl.
Mat. pg.17, with y1 and y2."$
% B_ IS THE VARIABLE VECTOR
B_:={y1,y2,x3,x4,x5,x1}$
FOR EACH EL_ IN B_ DO DEPEND EL_,T$
%B1_ IS THE UNKNOWN PARAMETER VECTOR
B1_:={k1,kon,kD,kex,ke,kdl,kde,bmax,ic2}$
%NUMBER OF STATES
NX_:=4$
%NUMBER OF OUTPUTS
NY_:=2$
%MODEL EQUATIONS
c_:={df(x1,t)=bmax*k1-k1*x1-kon*x1*(-x3-x4-x5-(kde/kdl)*x5+ic2)+
              kon*kon*kD*x3+kex*x4,
    df(x3,t)=kon*x1*(-x3-x4-x5-(kde/kdl)*x5+ic2)-kon*kon*kD*x3-ke*x3,
    df(x4,t)=ke*x3-(kex+kdl+kde)*x4,
    df(x5,t)=kdl*x4,
    y1=-x3-x4-x5+ic2,
    y2=x3,
%    y3=x4+x5}$
SEED_:=70$
DAISY()$
%VALUES OF INITIAL CONDITIONS ARE GIVEN
IC_:={x1=bmax,x2=ic2,x4=0,x5=0}$
CONDINIZ()$
END$
```

Due to space limitations the output file is not reported here but the reader can directly run the above input file and see that DAISY provides the required

structural identifiability answer in just 2-3 seconds. This computer algebra tool does not require expertise on mathematical modeling by the experimenter.

5 Conclusions

The goal of this paper is to make the researcher in system biology aware of the relevance of checking identifiability of the dynamic model under study and to show the differences between structural and practical (data-based) identifiability studies. We have discussed benefits and pitfalls of the two approaches by providing a practical example of a system biology model. A differential algebra based software tool able to check structural global identifiability in a fully automatic way, called DAISY [3] has been used.

References

1. Audoly,S., Bellu,G., D'Angiò,L., Saccomani,M.P., Cobelli,C.: Global Identifiability of Nonlinear Models of Biological Systems. IEEE Trans. Biomed. Eng. 48, 1, 55–65 (2001)
2. Becker,V., Shilling,M., Bachmann,J., Baumann,U., Raue,A., Maiwald,T., Timmer,J., Klingmller,U.: Covering Abroad Dynamic Range: Information Processing at the Erythropoietin Receptor. Science 328, 1404–1408 (2010)
3. Bellu,G., Saccomani,M.P., Audoly,S., D'Angiò, L.: DAISY: A New Software Tool to Test Global Identifiability of Biological and Physiological Systems. Comp. Meth. Prog. Biom. 88, 52–61 (2007)
4. Chapman,M.J., Godfrey,K.R., Chappell,M.J., Evans,N.D.: Structural Identifiability of Non-linear Systems Using Linear/non-linear Splitting. Int. J. Control 76, 3, 209–216 (2003)
5. Cobelli, C., Saccomani, M.P.: Unappreciation of a Priori Identifiability in Software Packages Causes Ambiguities in Numerical Estimates. Letter to the Editor. Am. J. Physiol. 21, E1058–E1059 (1990)
6. Hastie,T., Tibshirani,R.: Generalized Additive Models. Statistical Science 1, 297–318 (1986)
7. Hengl,S., Kreutz,C., Timmer,J., Maiwald,T.: Data-based Identifiability Analysis of Non-Linear Dynamical Models. Bioinformatics 23, 19, 2612–2618 (2007)
8. Joly-Blanchard,G., Denis-Vidal,L.: Some Remarks about Identifiability of Controlled and Uncontrolled Nonlinear Systems. Automatica 34, 1151–1152 (1998)
9. Ljung,L., Glad,S.T.: On Global Identifiability for Arbitrary Model Parameterizations. Automatica 30, 2, 265–276 (1994)
10. Ljung,L.: System Identification - Theory For the User, 2nd ed, PTR Prentice Hall, Upper Saddle River, N.J., (1999)
11. Ollivier, F.: Le Problème de l'Identifiabilité Structurelle Globale: Étude Théorique, Méthodes Effectives et Bornes de Complexité. Thèse de Doctorat en Science, École Polytéchnique, Paris, France (1990)
12. Raue,A., Kreutz,C., Maiwald,T., Bachmann,J., Shilling,M., Klingmller,U., Timmer,J.: Structural and Practical Identifiability Analysis of Partially Observed Dynamical Models by Exploiting the Profile Likelihood. Bioinformatics 25, 1923–1929 (2009)

13. Raue,A., Becker,V., Klingmller,U., Timmer,J.: Identifiability and Observability Analysis for Experimental Design in Nonlinear Dynamical Models. Chaos 20, 045105 (2010)
14. Raue,A., Kreutz,C., Maiwald,T., Klingmller,U., Timmer,J.: Addressing Parameter Identifiability by Model-based Experimentation. IET Syst. Biol. 5, 2, 120–130 (2011)
15. Saccomani, M.P., Cobelli,C.: Qualitative Experiment Design in Physiological System Identification. IEEE Control Systems. 12, 6, 18–23 (1992)
16. Saccomani,M.P., Audoly,S., D'Angiò L.: Parameter Identifiability of Nonlinear Systems: the Role of Initial Conditions. Automatica 39, 619–632 (2004)
17. Soderstrom, T., Stoica, P.: System identification. Prentice Hall International Series in Systems and Control Engineering. Publisher Prentice-Hall (1994)
18. Walter,E., Lecourtier,Y.: Global Approaches to Identifiability Testing for Linear and Nonlinear State Space Models. Math. and Comput. in Simul. 24, 472–482 (1992)