Structural and practical identifiability of approximate metabolic network models

Sara Berthoumieux, Daniel Kahn, Hidde De Jong, Eugenio Cinquemani

Structural and practical identifiability of approximate metabolic network models

Sara Berthoumieux [*] **Daniel Kahn** [**] **Hidde de Jong** [*]
Eugenio Cinquemani [*,⋆]

[*] *INRIA Grenoble - Rhône-Alpes, Montbonnot, France*
[**] *Laboratoire de Biométrie et Biologie Evolutive, CNRS UMR 5558, Université Lyon 1, INRA, Villeurbanne, France*

Abstract:
Parameter estimation from experimental data is a crucial problem in quantitative modeling of biochemical reaction networks. An especially important issue, raised by the complexity of the models and the challenging nature of the experimental data, is parameter identifiability. Despite several approaches proposed in the systems biology literature, no agreement exists on the analysis of structural and practical identifiability, and the relations among the two. In this paper we propose a mathematical framework for the analysis of identifiability of metabolic network models, establish basic results and methods for the structural and practical identifiability analysis of the class of so-called linlog models, and discuss the results on the basis of an artificial example.

1. INTRODUCTION

Kinetic models of biochemical reaction systems usually contain a large number of parameters, many of which are difficult to measure in a direct way. This makes parameter estimation from experimental data a crucial problem for quantitative systems biology [Ashyraliyev et al., 2009, Crampin, 2006]. For all but the simplest systems, parameter estimation is a difficult problem. Models contain a large number of variables, whose dynamics evolve on different time-scales and are described by complex, nonlinear rate equations. Moreover, the available experimental data usually consist of noisy and incomplete measurements, obtained under different conditions and by means of heterogeneous experimental methods.

These difficulties have stimulated the use of approximate kinetic models, with rate equations following, for example, linear, piecewise-linear, linlog, or power-law kinetics [de Jong, 2002, Heijnen, 2005]. In general, these models have less parameters to estimate and, in many cases, parameter estimation can be reduced to linear or orthogonal regression [Nikerel et al., 2009]. Therefore, approximate kinetic models have been chosen and successfully used for the quantitative modeling of both metabolic networks and gene regulatory networks [Heijnen, 2005].

We are interested in how approximate kinetic models help addressing the key problem of the identifiability of biochemical networks in a principled and scalable way. We focus on the case where the structure of the model is fixed by a priori knowledge on the network (i.e. the chemical species considered and the reactions among them), and discuss the identifiability of the model parameters. That is, we are interested in the problem of unambiguously reconstructing the unknown parameter values from the observed network behavior.

A distinction is usually made between structural (or a priori) and practical (or a posteriori) identifiability [Ljung, 1999, Walter and Pronzato, 1997]. Structural identifiability is an intrinsic property of the model family, guaranteeing that unique parameter reconstruction would be possible from perfect observations of the system response to an arbitrarily rich set of inputs. Practical identifiability refers to the ability of estimating unknown parameter values from the available experimental data within a prespecified degree of accuracy. In classical control theory, this concept is essentially related to the notion of persistence of excitation (see e.g. Ljung [1999] for the case of linear dynamical systems). However, in a biological context, this notion requires at least some nontrivial adaptation, and no agreement on the concepts of structural and practical identifiability exists.

The aim of this paper is to discuss identifiability of kinetic models of metabolism from a rigorous theoretical standpoint. While most of our definitions are of general applicability, identifiability results will be developed primarily for approximate kinetic models known as linlog models [Visser and Heijnen, 2003], whose pseudo-linear form enables us to apply tools from linear algebra and estimation theory in a straightforward manner. Similar results can be derived easily for many other approximate kinetic model classes in pseudo-linear form, such as the linear, loglin and generalized mass-action kinetic models [Delgado and Liao, 1992, Hatzimanikatis and Bailey, 1997, Savageau, 1976].

The contributions of the paper are as follows. First of all, we provide precise definitions and links between structural and practical identifiability of kinetic network models. Then, based on this solid mathematical foundation, our analysis provides tools that are applicable to the currently available data sets, e.g., those obtained by means of recent high-throughput methods in biology [Ishii et al., 2007].

⋆ Corresponding author, e-mail: eugenio.cinquemani@inria.fr

This sets the stage for rigorous kinetic model reduction, which will be addressed separately in another paper.

The paper is organized as follows. In Section 2 we discuss kinetic models as well as linlog and related approximations. In Section 3 we discuss the notions of structural and practical identifiability, and provide readily applicable methods for the identifiability analysis of linlog models. Results are illustrated throughout by means of a small example. Conclusions and perspectives of the work are reported in Section 4. Proofs of theoretical results are rather straightforward and are omitted in the interest of space.

2. PARAMETER ESTIMATION IN LINLOG AND RELATED MODELING FRAMEWORKS

The dynamics of biochemical reaction networks are described by kinetic models having the form of systems of ordinary differential equations (ODEs) [Heinrich and Schuster, 1996]. In this paper we focus on kinetic models of metabolism, where the rate functions describe enzyme-catalyzed reactions. This leads to models of the general form:

$$\dot{x} = N \cdot v(x, u, e), \tag{1}$$

with $x(0) = x_0 \in \mathbb{R}^{n_x}_{\geq 0}$, where $x \in X \subseteq \mathbb{R}^{n_x}_{\geq 0}$ denotes the vector of (nonnegative) internal metabolite concentrations, $u \in U \subseteq \mathbb{R}^{n_u}_{\geq 0}$ the vector of external metabolite concentrations, $e \in E \subseteq \mathbb{R}^{m}_{>0}$ the vector of enzyme concentrations, and $v : \mathbb{R}^{n_x+n_u+m}_{\geq 0} \to V$, with $V \subseteq \mathbb{R}^m$, the vector of reaction rate functions. $N \in \mathbb{Z}^{n_x \times m}$ is a stoichiometry matrix.

Kinetic modeling formalisms differ in the choice of rate functions. Examples of classical functions in enzyme kinetics are the Michaelis-Menten, reversible Michaelis-Menten, and Monod-Wyman-Changeux rate laws [Heinrich and Schuster, 1996]. Approximate formalisms simplify the mathematical form of the rate laws, in particular the nonlinear dependency of the reaction rates on the metabolite concentrations. Moreover, they usually assume all reactions to follow the same simplified kinetic format, thus giving a uniform structure to the models. For one type of approximate kinetic formalism, so-called linear-logarithmic (linlog) models, below we show that the parameter estimation problem can be formulated as multiple linear regression. A brief discussion of how this can be equally done for a number of other well-known approximate formalisms is reported at the end of the section.

The linear-logarithmic (linlog) approximation [Heijnen, 2005, Visser and Heijnen, 2003] expresses the reaction rates as proportional to the enzyme concentrations and to a linear function of the logarithms of internal and external metabolite concentrations,

$$v(x, u, e) = \mathrm{diag}(e) \cdot \big(a + B^x \cdot \ln(x) + B^u \cdot \ln(u)\big), \tag{2}$$

where $\mathrm{diag}(e)$ is the square diagonal matrix with the elements of e on the diagonal, and the logarithm of a vector means the vector of logarithms of its elements. Concavity of the function with respect to metabolite concentrations, and the fact that B^x and B^u are in direct relationship with the so-called system elasticities [Heijnen, 2005, Heinrich and Schuster, 1996] make linlog modelling a very convenient (local) approximation of metabolic kinetics. For conciseness, in the sequel we shall often drop the dependence of v on (x, u, e) from the notation.

The identification of metabolic networks in the linlog formalism amounts to estimating the (generally unknown) parameters $a \in \mathbb{R}^m$, $B^x \in \mathbb{R}^{m \times n_x}$ and $B^u \in \mathbb{R}^{m \times n_u}$ from experimental data. In most experiments, concentrations of enzymes and external metabolites are under (partial) control of the experimentalist, and the concentrations of internal metabolites and metabolic fluxes are measured after the system has relaxed to the steady state

$$N \cdot v(x, u, e) = 0. \tag{3}$$

In accordance with this, we shall assume that, from each of $q \in \mathbb{N}$ experiments, the data are (noisy) measurements $(\tilde{v}^k, \tilde{x}^k, \tilde{u}^k, \tilde{e}^k)$ of (v^k, x^k, u^k, e^k), where the latter satisfy $v^k = v(x^k, u^k, e^k)$ and (3), with $k = 1, \ldots, q$. Clearly the restriction to steady-state measurements limits the informativity of the data and may affect the identifiability of the models, as will be apparent in later sections.

For the purpose of parameter estimation, it is convenient to rewrite (2) in the form of a regression model:

$$(v/e)^T = \begin{bmatrix}1 & \ln(x)^T & \ln(u)^T\end{bmatrix} \cdot [a \;\; B^x \;\; B^u]^T \tag{4}$$

where the ratio of two vectors (here v/e) denotes element-wise division. Let us use an upperbar to denote the mean of a quantity over its q experimental outcomes, for instance: $\overline{v/e} = (1/q)\sum_{k=1}^{q} v^k/e^k$. By the linearity of (4), it holds that

$$\overline{(v/e)}^T = \begin{bmatrix}1 & \overline{\ln(x)}^T & \overline{\ln(u)}^T\end{bmatrix} \cdot [a \;\; B^x \;\; B^u]^T. \tag{5}$$

This allows one to rewrite (4) as a mean-removed model

$$\left(\frac{v}{e} - \overline{\left(\frac{v}{e}\right)}\right)^T = \begin{bmatrix}\ln(x) - \overline{\ln(x)}\\ \ln(u) - \overline{\ln(u)}\end{bmatrix}^T \cdot \begin{bmatrix}(B^x)^T\\ (B^u)^T\end{bmatrix}. \tag{6}$$

We can now define the following estimation problem.

Problem 1. Given the data matrices

$$\underbrace{\begin{bmatrix}\left(\frac{\tilde{v}^1}{\tilde{e}^1} - \overline{\left(\frac{\tilde{v}}{\tilde{e}}\right)}\right)^T\\ \vdots\\ \left(\frac{\tilde{v}^q}{\tilde{e}^q} - \overline{\left(\frac{\tilde{v}}{\tilde{e}}\right)}\right)^T\end{bmatrix}}_{\triangleq \tilde{W}}, \quad \underbrace{\begin{bmatrix}\left(\ln(\tilde{x}^1) - \overline{\ln(\tilde{x})}\right)^T & \left(\ln(\tilde{u}^1) - \overline{\ln(\tilde{u})}\right)^T\\ \vdots & \vdots\\ \left(\ln(\tilde{x}^q) - \overline{\ln(\tilde{x})}\right)^T & \left(\ln(\tilde{u}^q) - \overline{\ln(\tilde{u})}\right)^T\end{bmatrix}}_{\triangleq \tilde{Y}}$$

find parameters $B \triangleq [B^x \;\; B^u]^T$ minimizing $|||\tilde{W} - \tilde{Y} \cdot B|||$, for some convenient (matrix) norm $|||\cdot|||$. We fix this to be the Frobenious norm.

Notice that the parameter vector a no longer appears in the regression problem, but that an estimate of it can be recovered from estimates of $B = [B^x \;\; B^u]^T$ via Eq. (5).

In practice, the experimental measurement error on the normalized fluxes v/e is often significantly larger than that on metabolite concentrations. In light of this [Berthoumieux et al., 2011] we consider the linear model $\tilde{W} = W + \varepsilon = Y \cdot B + \varepsilon$, where W and Y are the noiseless versions of $\tilde{W}$ and $\tilde{Y}$, respectively. $\varepsilon \in \mathbb{R}^{q \times m}$ is assumed to be a zero-mean Gaussian matrix with independent columns, each with q-dimensional positive definite covariance matrix $\mathrm{Var}(\varepsilon_{\cdot i}) = \Sigma_{\varepsilon_{\cdot i}}$, $i = 1, \ldots, m$. Thus, the problem is equivalent to the m least-squares problems of minimizing $||W_{\cdot i} - Y \cdot B_{\cdot i}||_{\Sigma_{\varepsilon_{\cdot i}}}$ over $B_{\cdot i}$, with $i = 1, \ldots, m$,

where $||\cdot||_{\Sigma_{\varepsilon\cdot i}}$ is the $\Sigma_{\varepsilon\cdot i}$-weighted L^2-norm (i.e., for $w \in \mathbb{R}^q$, $||w||_{\Sigma_{\varepsilon\cdot i}} = \sqrt{w^T \Sigma_{\varepsilon\cdot i}^{-1} w}$), and $W_{\cdot i}$ and $B_{\cdot i}$ are the ith columns of W and B, respectively. That is, the parameter estimation problem splits up into m smaller estimation problems, one for each reaction i, with $i = 1, \ldots, m$.

In turn, each reaction i depends only on a (known) subset of metabolites. In this case, the elements of $B_{\cdot i}$ corresponding to the metabolites not involved in the reaction can be set to zero, and the least-squares problem can be reduced accordingly. If $C(i) \subseteq \{1, \ldots, n_x + n_u\}$ denotes the indices of the relevant values in $B_{\cdot i}$, then the regression problem becomes

$$\min_{B_{C(i)i}} ||W_{\cdot i} - Y_{\cdot C(i)} \cdot B_{C(i)i}||_{\Sigma_{\varepsilon\cdot i}}, \tag{7}$$

where the meaning of $B_{C(i)i}$ follows from the above and $Y_{\cdot C(i)}$ is the submatrix of Y formed by the columns indexed by $C(i)$.

Similar parameter estimation problems can be formulated for other approximate modeling formalisms, such as models linear in metabolite concentrations [Delgado and Liao, 1992], log-lin models [Hatzimanikatis and Bailey, 1997], and generalized mass-action models [Savageau, 1976]. Thus, the results that will be discussed below for linlog models can be adapted straightforwardly to the approximate kinetic modeling formalisms mentioned above. On the contrary, most results are not applicable to formalisms where parameter estimation cannot be reduced to linear regression, such as classical Michaelis-Menten kinetics and convenience kinetics [Liebermeister and Klipp, 2006].

3. IDENTIFIABILITY OF LINLOG MODELS

The problem of identifiability refers to the ability to unambiguously extract parameter values of a model structure from experimental data. Here we focus on linlog models. We shall first discuss the problem from the perspective of structural identifiability. For practical purposes, this is equivalent to answering the question whether each parameter can be uniquely reconstructed from arbitrarily rich and errorless data sets. Structural identifiability forms the basis for studying practical identifiability, i.e. the ability to estimate parameter values from real data sets, which will be discussed further below.

The system, described by Equations (1)–(3), is parametrized by the parameter vector $p = [a \;\; B^x \;\; B^u] \in P \subseteq \mathbb{R}^{m\times(n_x+n_u+1)}$. For simplicity, let us make the standing assumption that $N\,\mathrm{diag}(e)B^x$ is invertible for all $p \in P$ and $e \in E$ (generalizations are possible). For varying values of p, this defines a class of models $\mathscr{M}_p$ mapping inputs $(e, u) \in E \times U$ to outputs $(J_p(e, u), x_p(e, u)) \in V \times X$, the latter being the steady-state values of v, x (in accordance with metabolic control analysis standards [Heinrich and Schuster, 1996], symbol J is used in place of v for steady state fluxes). More precisely, we have

$$\mathscr{M}_p : \; E \times U \to V \times X : \; (e, u) \mapsto \big(J_p(e, u), x_p(e, u)\big) \tag{8}$$

where, plugging (2) into (3) to express $\ln x_p(e, u)$, and writing e in place of $\mathrm{diag}(e)$ for shortness,

$$\begin{cases} J_p(e, u) = e \cdot (a + B^x \cdot \ln x_p(e, u) + B^u \cdot \ln u), \\ \ln x_p(e, u) = -(NeB^x)^{-1} \cdot Ne \cdot (a + B^u \cdot \ln u). \end{cases} \tag{9}$$

Since we are observing the system in steady state, inputs and outputs are not time-varying signals but just fixed vectors.

In agreement with Section 2, where the identification problem is split into the identification of each reaction separately, we look at the identifiability of the parameters of the generic ith reaction, and say that a model is identifiable if all its reactions are.

3.1 Identifiability from a structural perspective

We adapt the definition from Ljung [1999] to our context as follows.

Definition 1. A reaction i of model $\mathscr{M}_p$ is identifiable at p^* if there exists $I \subseteq E \times U$ such that, for all $p \in P$,

$$\big((J_p)_i, x_p\big)|_I = \big((J_{p^*})_i, x_{p^*}\big)|_I \Rightarrow p_i = p_i^*. \tag{10}$$

Here J_p and x_p are seen as vector functions of e and u, subscript "i" indicates the ith element (row) of the vector, and "$|_I$" indicates the restriction of the functions on I. Turned another way, a reaction i is considered identifiable for a particular model parametrization p^* if no $p \in P$ with $p_i \neq p_i^*$ exists such that the velocity of the reaction predicted by $\mathscr{M}_p$ and $\mathscr{M}_{p^*}$ from identical steady state concentrations is identical for all possible system inputs. Note that this definition is applicable to any form of the reaction rates (2) (provided suitable definition of the parameters p).

How can we apply Definition 1 to the analysis of identifiability of linlog models? The following proposition establishes the link between this definition and the uniqueness of the solution to Problem 1 (in the form (7)). Given the input set $I = \{(e^1, u^1), \cdots, (e^q, u^q)\}$ and a "true" parameter vector p^*, let J_*^k and x_*^k denote the outputs $J_{p^*}(e^k, u^k)$ and $x_{p^*}(e^k, u^k)$, respectively, with $k = 1, \ldots, q$.

Proposition 1. A reaction i of $\mathscr{M}_p$ is identifiable at p^* if and only if there exists $I \subseteq E \times U$ such that the solution of the equation $W_{\cdot i}^* = Y^* B_{\cdot i}$, with

$$W_{\cdot i}^* = \left[\left(\frac{J_*^1}{e^1} - \overline{\left(\frac{J_*}{e}\right)}\right)_i \cdots \left(\frac{J_*^q}{e^q} - \overline{\left(\frac{J_*}{e}\right)}\right)_i\right]^T,$$

$$Y^* = \begin{bmatrix} \ln x_*^1 - \overline{\ln x_*} & \cdots & \ln x_*^q - \overline{\ln x_*} \\ \ln u^1 - \overline{\ln u} & \cdots & \ln u^q - \overline{\ln u} \end{bmatrix}^T,$$

is unique in the parameters $B_{\cdot i} = ([B^x \;\; B^u]_{i\cdot})^T$. In turn, this happens if and only if $Y_{\cdot C(i)}^*$ (the submatrix of Y^* formed by the columns indexed by $C(i)$) is full column-rank.

Clearly this condition is equivalent to the uniqueness of the solution of (7), the difference being that the identifiable parameter vector p_i also contains the element a_i.

A standard approach for studying the rank of $Y_{\cdot C(i)}^*$ is the Singular Value Decomposition (SVD)

$$Y_{\cdot C(i)}^* = U \cdot \mathrm{diag}(s_1, s_2, ..., s_{n_b}) \cdot V^T \tag{11}$$

with $n_b = |C(i)|$, $U \in \mathbb{R}^{q\times n_b}$ and $V \in \mathbb{R}^{n_b\times n_b}$ orthonormal matrices and $s_1 \geq \cdots \geq s_{n_b} \geq 0$ the singular values of $Y_{\cdot C(i)}^*$. In the presence of dependencies between the columns, there exists an index r with $1 \leq r < n_b$ such

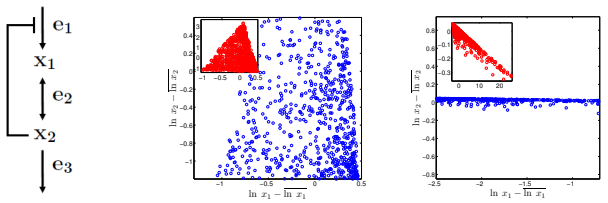
that $s_{r+1} = \cdots = s_{n_b} = 0$, and $Y^*_{\cdot C(i)}$ is of rank r. In order to illustrate the identifiability properties of a metabolic reaction, consider the following example.

Example 1. Consider the negative feedback network structure shown in Figure 1(a). The network includes $n = 2$ internal metabolites, no external metabolites, and $m = 3$ reactions (enzymes). The network structure and linlog model parameters are

$$N = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}, \quad a = \begin{bmatrix} a_1 \\ 0.0297 \\ 0.0296 \end{bmatrix}, \quad B^x = \begin{bmatrix} -0.0938 & B_{21} \\ 0.0286 & -0.0073 \\ 0 & 0.0287 \end{bmatrix},$$

(with $B^x = B^T$) where different values of $a_1 \in \mathbb{R}_{\geq 0}$ and $B_{21} \in \mathbb{R}_{\geq 0}$ (the coefficient that determines the strength of the feedback regulation) will be considered. For all values of the enzyme concentrations $e_i > 0$, with $i = 1, 2, 3$, and all $a_1, B_{21} \in \mathbb{R}_{\geq 0}$, the equation $Nv(x, e) = N \operatorname{diag}(e)(a + B^x \ln x) = 0$ yields a unique steady-state solution $\ln x = -(N \operatorname{diag}(e) B^x)^{-1} N \operatorname{diag}(e) a$. One first observation is that different values of a_1 and B_{21} may lead to very different properties of the matrix Y^* even when this remains full rank, i.e. the system is structurally identifiable. For $a_1 = 0.0297$ and two different values of B_{21}, scatter plots of the steady state solutions $\ln x$ from 1000 randomly generated samples of e are reported in Figures 1(b) ($B_{21} = -0.0073$, weeker feedback action) and 1(c) ($B_{21} = -7.2961$, stronger feedback action). In Figure 1(b), steady-state metabolite concentrations are spread over a two-dimensional region, while in Figure 1(c) they are essentially aligned along a one-dimensional line. SVD analysis reveals that, in the latter case, $s_2 \ll s_1$, which hints at an ill-conditioned estimation problem. This point will be further developed in the next example. Second, some pathological parameterizations may give rise to a nonidentifiable model. Indeed, if a is in the span of B^x, then the unique solution of $N \operatorname{diag}(e)(a + B^x \ln x) = 0$ corresponds to the value of $\ln x$ satisfying $B^x \ln x = -a$, i.e. it is independent of e. Thus, no matter the number of experiments q, the rank of Y^* is at most 1 (for $a = 0$, the solution is $\ln x = 0$, i.e. Y is the zero matrix which has rank 0). Thus, in the light of Proposition 1, the model is not identifiable. In our example, this is the case for $a_1 = -7.5491$ and $B_{21} = -7.2961$.

From the example above, it is clear that, for some specific values of the parameters, a model may be non-identifiable even if it is identifiable for non-pathological parametrizations. In the light of this, a convenient generalization of Definition 1 to the identifiability of a network structure can be obtained following Walter and Pronzato [1997], who state that a model in $\mathscr{M}_p$ is identifiable if it is identifiable at almost every $p^* \in P$ (in the sense of the Lebesgue measure on P). Whence, the negative feedback network structure of Example 1 is identifiable in the sense of Walter and Pronzato. The identifiability criterion of Definition 1 does not hold for the "rare" parameter combinations p^* such that $a \in \operatorname{span}(B^x)$. A second observation, following from the example above, is that our theoretical definition of identifiability is actually too restricted to be practically useful. If we look at Figure 1(c), we see that strong collinearities exist between the metabolite concentrations x_1 and x_2. This makes the informativity of the experiments limited. Again, we will analyze later the implications of this on the achievable parameter estimation performance.



(a) (b) (c)

Fig. 1. An example of a small metabolic network with negative feedback. (a) Structure of the network. (b)-(c) Steady-state metabolite concentrations from 1000 randomly generated enzyme concentrations, for (b) $a_1 = 0.0297$, $B_{21} = -0.0073$ and (c) $a_1 = 0.0297$, $B_{21} = -7.2961$ (see Example 1). Main figures are crops of the scatter plots of the data (blue markers) with equally scaled axes, emphasizing the shape of the data clouds. All generated data points are shown in the inset scatter plots (red markers), in arbitrary axes.

Moreover, the definition assumes that the measurements are not corrupted by noise, which is even less realistic. We therefore need to weaken the definition of identifiability in order to make it more suitable for applications to actual data on metabolism. While taking into account realistic assumptions on the experimental datasets, i.e., measurements available in a limited amount and affected by experimental error, this notion of identifiability should draw upon the theoretical notion of model identifiability discussed above.

3.2 Identifiability from a practical perspective

Let I be a fixed set of q inputs (external metabolites and enzyme concentrations), and let O be the set of the correponding system outputs (fluxes and steady-state concentrations of internal metabolites determined by $\mathscr{M}_{p^*}$). Consider the problem of estimating the parameters $B_{\cdot i}$ of reaction i given observations of I and O affected by measurement error. An *estimator* $\hat{B}_{\cdot i}$ of $B_{\cdot i}$ is a function of the observations of I and O, well-defined for every possible (a priori unknown) value of $p^* \in \mathbb{R}^{n_b}$ (compare [Ljung, 1999, §7.4]). Since, due to noise, the observations are stochastic variables, $\hat{B}_{\cdot i}$ is itself a stochastic variable. Therefore, one cannot hope to estimate $B_{\cdot i}$ exactly, but only within a certain degree of approximation, except possibly for a few "adverse" outcomes of the measurement error. In this spirit, we define identifiability in terms of the existence of an estimator satisfying prespecified statistical requirements. In doing this, we restrict attention to the nonzero entries of $B_{\cdot i}$, i.e., $B_{C(i)i}$. Let $\mathscr{B}_i \subset \mathbb{R}^{n_b}$ be a *bounded* neighbourhood of the origin, and let $\alpha \in (0, 1)$.

Definition 2. For a given $I \subseteq E \times U$, a reaction i of $\mathscr{M}_p$ is identifiable at p^* with uncertainty $\mathscr{B}_i$ and confidence level $1 - \alpha$ if there exists an estimator $\hat{B}_{C(i)i}$ such that

$$\mathbb{P}_{p^*}[\hat{B}_{C(i)i} - B_{C(i)i} \in \mathscr{B}_i] \geq 1 - \alpha, \tag{12}$$

where $\mathbb{P}_{p^*}$ is the probability measure induced by $\mathscr{M}_{p^*}$. [1]

[1] Strictly speaking, a better version of Definition 2 would require that condition (12) holds for all p^* within a sufficiently large subset of P. This would automatically rule out trivial definitions of $\hat{B}_{C(i)i}$

In this view, the experimentalist, or the modeler, sets the requirements (estimation accuracy and confidence level) that the estimates must fulfill in order to be useful, via the a priori specification of $\mathscr{B}_i$ and α. Then, the possibility of fulfilling (12), i.e. the practical identifiability of the model, depends on the system itself and on the richness of the input set I. In general, the larger the I, the tighter the requirements that one can fulfill (i.e. the smaller the values of $\mathscr{B}_i$ and α for which practical identifiability in the sense of Definition 2 holds). Note that this is conceptually different from what suggested by Raue et al. [2009], where the definition of practical identifiability only requires that the uncertainty on the parameter estimates (as defined via the profile likelihood) is bounded (though arbitrarily large) for a specific (not necessarily optimal) choice of the estimator. In an alternative view, one may start from a given input set I, and look for the choices of α and $\mathscr{B}_i$ that ensure satisfaction of (12). Here in turn, one may fix α and look for the $\mathscr{B}_i$ that makes (12) achievable, or fix the acceptable estimation uncertainty $\mathscr{B}_i$ and establish at what confidence level α this performance can be attained.

In all cases, the natural questions that arise are how Definition 2 can be verified in practice, how this notion of identifiability depends on the structural system identifiability discussed in the previous section, and what $\mathscr{B}_i$ may look like. To answer these questions one needs to further specify the properties of the data, i.e., the (stochastic) "measurement model". Recall from Section 2 the following assumptions. The metabolite concentrations forming Y are measured without error. For $i = 1, \dots, q$, the noisy versions $\tilde{W}_{\cdot i}$ of the $W_{\cdot i}$ obey the model

$$\tilde{W}_{\cdot i} = W_{\cdot i} + \varepsilon_{\cdot i} = Y_{\cdot C(i)} B_{C(i)i} + \varepsilon_{\cdot i}, \tag{13}$$

with $\varepsilon_{\cdot i} \sim \mathscr{N}(0, \Sigma_{\varepsilon_i})$ and $\Sigma_{\varepsilon_i} > 0$. The following proposition answers the questions above.

Proposition 2. If a reaction i of $\mathscr{M}_p$ is identifiable at p^* in the sense of Definition 1 then, for every $\alpha \in (0, 1)$, it is identifiable in the sense of Definition 2 with confidence level at least $1 - \alpha$ for any uncertainty set $\mathscr{B}_i$ containing the $(1 - \alpha)$-confidence ellipsoid of a zero-mean Gaussian distribution with variance $(Y_{\cdot C(i)}^T \Sigma_{\varepsilon_i}^{-1} Y_{\cdot C(i)})^{-1}$.

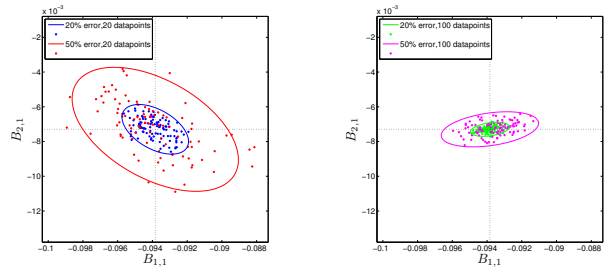
The proof relies on the use of minimum variance estimators, as dictated by standard results in linear estimation theory [Ljung, 1999, Appendix II]. From now on, estimation will be performed based on this type of estimator. Observe that $\mathscr{B}_i$ implicitly depends on the choice of inputs I via the structure of $W_{\cdot i}$ and $Y_{\cdot C(i)}$. Typically, the larger q, the smaller $\mathscr{B}_i$ for a fixed α. We argue that similar identifiability results can be derived even in cases where the noise is not Gaussian and metabolite measurements are affected by stochastic error, at the price of a much more complicated characterization of $\mathscr{B}_i$. Finally, one may speak about identifiability of the whole model, e.g. by requiring that each reaction is individually identifiable with a given confidence level α and uncertainty set $\mathscr{B}_i \in \mathbb{R}^{m \times n_b}$. Alternatively, one may require that all reactions be simultaneously identifiable with confidence level $1 - \alpha$ and a suitably defined joint uncertainty set.

There is a clear link between Definition 2 and singular values of the regression matrix $Y_{\cdot C(i)}$, i.e., the rank analysis of the previous section. Assume for simplicity that all error terms have the same variance $\Sigma_{\varepsilon_i} = \sigma^2 I$, with $\sigma > 0$. From Proposition 2, if $B_{C(i)i}$ is identifiable with uncertainty $\mathscr{B}_i$ and confidence $1 - \alpha$, then there exists an estimator (the minimum variance estimator) such that, with probability $1-\alpha$, the estimates of $B_{C(i)i}$ will lie in an ellipsoid centered at $B_{C(i)i}$ with axes length proportional to the square roots of the singular values of $\sigma^2 (Y_{\cdot C(i)}^T Y_{\cdot C(i)})^{-1}$, i.e., to $\sigma[s_1^{-1}, \dots, s_{n_b}^{-1}]$ (as before, s_l denotes the lth singular value of $Y_{\cdot C(i)}$, $l \in \{1, \dots, n_b\}$). If these eigenvalues are vastly different, the ellipsoid is skewed in the direction of the singular vectors of $Y_{\cdot C(i)}^T Y_{\cdot C(i)}$ associated to the smallest singular values of $Y_{\cdot C(i)}$. In this sense, if a data matrix $Y_{\cdot C(i)}$ is ill-conditioned ($s_1 \gg s_{n_b}$), i.e., some data vectors are nearly collinear, parameter $B_{C(i)i}$ appears as nonidentifiable even if $\mathrm{rank}(Y_{\cdot C(i)}) = n_b$ (that is, even if $s_{n_b} > 0$). In other words, this makes it possible to detect that, for some $r < n_b$, discrepancies between values of $s_{r+1}, \cdots, s_{n_b}$ and s_1 can make the estimates of $B_{C(i)i}$ solving regression poorly determined.

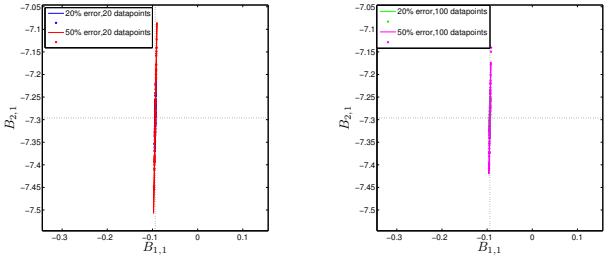
Example 2. To illustrate the implications for parameter identifiability of a poorly conditioned data matrix, consider the parameter estimation results from noisy and finite datasets for the two different identifiable parametrizations of the model of Example 1. As in Example 1, datapoints were simulated from random values of enzyme concentrations. Noise was added to W by drawing values from normal distributions with standard deviations proportional to the corresponding elements of W. Two different dataset sizes ($q = 20$ and $q = 100$) and two different noise levels (20% and 50%, meaning that 99% of the noise samples fall within 20% and 50% of the corresponding values in W) were tested, for a total of 4 experimental scenarios for each model parameterization. For each scenario, 100 datasets were simulated and the corresponding estimates drawn for reaction 1 are reported in the scatter plots of Figure 2(a) ($a_1 = 0.0297$ and $B_{21} = -0.0073$) and 2(b) ($a_1 = 0.0297$ and $B_{21} = -7.2961$).

An immediate observation is that the shape of the 95%-confidence ellipse of the parameter estimates is different for the two different model parameterizations. While estimation accuracy for B_{11} and B_{21} is comparable in the case of smaller feedback ($B_{21} = -0.0073$), the shape of the uncertainty ellipse becomes very skewed in the case of stronger feedback ($B_{21} = -7.2961$). In particular, in this case, estimation accuracy is much higher for B_{11} than for B_{21} regardless of the features of the dataset (note the change in the scale of the vertical axes of the plots from Figure 2(a) to Figure 2(b)). As apparent from the plots, larger and/or less noisy datasets improve estimation performance. However, ameliorating the estimation of B_{21} requires extremely large/high-quality datasets. In other words, estimating the model (i.e. all its parameters) accurately requires a significant increase in the experimental effort, even if most parameters are easy to estimate (see also Gutenkunst et al. [2007]). Following upon Example 1, it is also apparent that the skewed estimation uncertainty

such as $\hat{B}_{C(i)i} \triangleq B_{C(i)i}$ (which makes the reaction identifiable for any α and $\mathscr{B}_i$ but cannot be built without the knowledge of $B_{C(i)i}$ itself). Unfortunately, this is not a good choice in general, in that the uncertainty set $\mathscr{B}_i$ may severely depend on p^*, as we shall see later on in Example 2. Hence we stick to Definition 2 with the understanding that any such triviality is avoided.



(a)



(b)

Fig. 2. Estimates of the parameters $(B^x)_{1\cdot} = [\,B_{11}\ B_{21}\,]$, mediating the effects of x_1 and x_2 in reaction 1, from simulated steady-state data, for a linlog model of the network of Figure 1(a) with two different parametrizations: (a) $a_1 = 0.0297$, $B_{21} = -0.0073$; (b) $a_1 = 0.0297$, $B_{21} = -7.2961$. In each panel, scatter plots are reported for four different experimental scenarios specified by the couple (% noise level, number of datapoints q): $(20\%, 20)$ (blue), $(20\%, 100)$ (green), $(50\%, 20)$ (red), $(50\%, 100)$ (magenta). 95%-confidence ellipses are drawn for each scenario (solid lines). Reference parameter values are indicated by horizontal and vertical dotted lines, and are given in Example 1 (see Example 2 for other details).

is related to the poor conditioning of the data matrix Y in the case of stronger feedback (the shape of the ellipsoid is determined by the ratio of the singular values of Y). In terms of practical identifiability, assuming a modeler has set a maximum allowable uncertainty $\mathscr{B}_1$ for some confidence level α, it is clear that in this case the system will not be practically identifiable (even if the model is structurally identifiable at the given p^*), unless $\mathscr{B}_1$ is large enough, i.e. rather sloppy estimates are deemed acceptable.

4. CONCLUSIONS

In this paper we have discussed the notion of structural and practical identifiability for quantitative dynamical metabolic network models. Assuming a fixed model structure, we have introduced definitions of general applicability for the identifiability of the model parameters a priori and a posteriori from a given dataset. In the relevant case of approximate linlog modeling, using standard tools from linear algebra and estimation theory, we have linked the two concepts, provided conditions for structural and practical identifiability, and methods for testing these conditions in practice. Concepts and results were illustrated by a simple example. Current research directions include model reduction, as well as methods for handling missing data entries and noise on metabolite concentrations.

REFERENCES

M. Ashyraliyev, Y. Fomekong Nanfack, J.A. Kaandorp, and J.G. Blom. Systems biology: Parameter estimation for biochemical models. *FEBS J.*, 276(4):886–902, 2009.

S. Berthoumieux, M. Brilli, H. de Jong, D. Kahn, and E. Cinquemani. Identification of linlog models of metabolic networks from incomplete high-throughput datasets. *Bioinformatics*, 27(13):i186–i195, 2011.

E.J. Crampin. System identification challenges from systems biology. In *Proc. 14th IFAC Symp. Syst. Identif. (SYSID 2006)*, pages 81–93, Newcastle, Australia, 2006.

H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *J. Comput. Biol.*, 9(1):67–103, 2002.

X. Delgado and J.C. Liao. Metabolic control analysis using transient metabolite concentrations. *Biochem. J.*, 285: 965–72, 1992.

R.N. Gutenkunst, J.J. Waterfall, F.P. Casey, K.S. Brown, C.R. Myers, and J.P. Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLOS Computational Biology*, 3(10):e189, 2007.

V. Hatzimanikatis and J.E. Bailey. Effects of spatiotemporal variations on metabolic control: Approximate analysis using (log)linear kinetic models. *Biotechnol. Bioeng.*, 54(2):91–104, 1997.

J.J. Heijnen. Approximative kinetic formats used in metabolic network modeling. *Biotechnol. Bioeng.*, 91 (5):534–45, 2005.

R. Heinrich and S. Schuster. *The Regulation of Cellular Systems*. Chapman & Hall, 1996.

N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P.Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita. Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science*, 316(5824):593–7, 2007.

W. Liebermeister and E. Klipp. Bringing metabolic networks to life: Convenience rate law and thermodynamic constraints. *Theor. Biol. Med. Model.*, 3:41, 2006.

L. Ljung. *System identification, theory for the user*. Prentice Hall PTR, 1999.

I.E. Nikerel, W.A. van Winden, P.J.T. Verheijen, and J.J. Heijnen. Model reduction and *a priori* kinetic parameter identifiability analysis using metabolome time series for metabolic reaction networks with linlog kinetics. *Metab. Eng.*, 11(1):20–30, 2009.

A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–29, 2009.

M.A. Savageau. *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*. Addison-Wesley, 1976.

D. Visser and J.J. Heijnen. Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics. *Metab. Eng.*, 5(3):164–76, 2003.

E. Walter and L. Pronzato. *Identification of parametric models*. Springer, 1997.