# On the identifiability of metabolic network models.

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

# On the Identifiability of Metabolic Network Models

Sara Berthoumieux[a], Matteo Brilli[b], Daniel Kahn[b], Hidde de Jong[a], Eugenio Cinquemani[a,*]

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

**Abstract**

A major problem for the identification of metabolic network models is parameter identifiability. The identifiability of a parameter consists in the possibility to unambiguously infer the parameter from the data. Identifiability problems may be due to the structure of the model, in particular implicit dependencies between the parameters, or to limitations in the quantity and quality of the available data. We address the detection and resolution of identifiability problems for a class of pseudo-linear models of metabolism, so-called linlog models. Linlog models have the advantage that parameter estimation reduces to linear or orthogonal regression, which facilitates the analysis of identifiability. We develop precise definitions of structural and practical identifiability, and clarify the fundamental relations between these concepts. In addition, we use Singular Value Decomposition (SVD) to detect identifiability problems and reduce the model to an identifiable approximation by a Principal Component Analysis (PCA) approach. The criterion is adapted to real data, which are frequently scarce, incomplete, and noisy. The test of the criterion on a model with simulated data shows that it is capable of correctly identifying the principal components of the data vector. The application to a state-of-the-art dataset on central carbon metabolism in *Escherichia coli* yields the surprising result that only 4 out of 31 reactions, and 37 out of 100 parameters, are identifiable. This underlines the practical importance of identifiability analysis and model reduction in the modeling of large-scale metabolic networks. Although our approach has been developed in the context of linlog models, it carries over to other pseudo-linear models and provides useful hints for the identifiability analysis of more general classes of nonlinear models of metabolism.

*Keywords:* Systems biology, Metabolic network modeling, Parameter estimation, Structural and practical identifiability, Principal Component Analysis, Singular Value Decomposition, *Escherichia coli* carbon metabolism

## 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; Jaqaman and Danuser, 2006; Chou and Voit, 2009). For all but the

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

simplest systems, parameter estimation of biological systems 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; Chou and Voit, 2009). In general, these models have fewer parameters to estimate and, in many cases, parameter estimation can be reduced to linear or orthogonal regression (Nikerel et al., 2009; Sands and Voit, 1996). Therefore, approximate kinetic models have been chosen and successfully used for the quantitative modeling of both metabolic networks and gene regulatory networks. In this paper, 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, the reactions among them and the possible regulatory interactions).

A major and often overlooked problem in parameter estimation is the identifiability of the model, that is, 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 (Ljung, 1999); unfortunately, limitations in the variety and quality of the data make this notion inapplicable for biological applications. In recent years, the topic of identifiability has gained considerable interest in the field of systems biology (Chen et al., 2010; Chis et al., 2011b; Nikerel et al., 2009; Raue et al., 2009, 2011; Gutenkunst et al., 2007; Nemcova, 2010; Srinath and Gunawan, 2010; Voit et al., 2006a) and several specialized software packages have been developed to support the modeler (Bellu et al., 2007; Chis et al., 2011a; Maiwald and Timmer, 2008). Despite these efforts, however, no common agreement on definitions and links between structural and practical identifiability exist to date.

The aim of this paper is to develop methods for the analysis of parameter identifiability of kinetic models of metabolism, and for the reduction of nonidentifiable models to identifiable approximations. These methods should have a solid mathematical foundation, but at the same time be applicable to practical problems and be scalable to currently available datasets, such as those obtained by means of recent high-throughput methods in biology (Ishii et al., 2007). While many 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 for many other approximate kinetic modeling formalisms in pseudo-linear form, such as the linear, loglin and generalized mass-action kinetic formats (Delgado and Liao, 1992; Hatzimanikatis and Bailey, 1997; Savageau, 1976). Moreover, estimated parameters of approximate kinetic models provide useful hints for the identification of more detailed nonlinear models.

The present paper builds upon and elaborates earlier work, in which we developed an identification method for approximate kinetic models in the case of incomplete data (Berthoumieux et al., 2011). More precisely, the main contributions of the present paper are threefold. First, we precisely define the notions of structural and practical identifiability of linearized kinetic models, drawing upon the systems identification literature. This conceptual clarification allows us to develop the relations between structural and practical identifiability in a fundamental way. Second, we show how model reduction using Singular Value Decomposition (SVD) (Jolliffe, 1986) provides a suitable theoretical framework for addressing identifiability problems. We discuss several different criteria for model reduction, based on the singular values returned by the SVD analysis, and we show to which extent these criteria are appropriate for dealing with actual biological datasets, which are typically scarce, noisy and incomplete. Third, we apply the methods for identifiability analysis and model reduction to both simulated data and a published dataset concerning central metabolism in *E. coli* (Ishii et al., 2007). These examples show that the mathematical tools developed in this paper are of practical utility for the estimation of parameters in metabolic network models, and beyond, from current high-throughput data sets.

For the readers'convenience, the notation adopted in the paper is summarized in Appendix A. To simplify the reading, all mathematical proofs are deferred to Appendix B.

## 2. Parameter estimation in linearized kinetic models

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:

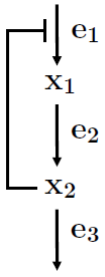
$$\dot{x} = N \cdot v(x, u, e), \tag{1}$$

with $x(0) = x_0 \in \mathbb{R}^{n_x}_{>0}$, where $x \in X \subseteq \mathbb{R}^{n_x}_{>0}$ denotes the vector of internal metabolite concentrations, $u \in U \subseteq \mathbb{R}^{n_u}_{>0}$ the vector of external metabolite concentrations, $e \in E \subseteq \mathbb{R}^{m}_{>0}$ the vector of enzyme concentrations, and $v : \mathbb{R}^{n_x+n_u+m}_{>0} \to V$, 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. They usually assume all reactions to follow the same simplified kinetic format, thus giving a uniform structure to the models. Moreover the parameter estimation problem for linearized kinetic models can be recast as multiple linear regression, as we will now show on linear-logarithmic (linlog) models.

The 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.

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




(a)

$$
\begin{aligned}
v_1 &= e_1 \cdot (a_1 + B^x_{11} \ln x_1 + B^x_{12} \ln x_2) \\
v_2 &= e_2 \cdot (a_2 + B^x_{21} \ln x_1 + B^x_{22} \ln x_2) \\
v_3 &= e_3 \cdot (a_3 \qquad\qquad\quad + B^x_{32} \ln x_2)
\end{aligned}
$$

(b)

Figure 1: (a) Structure of a small metabolic network with negative feedback. (b) Equations of the linlog model of the network

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. For conciseness, in the sequel we shall often drop the dependence of $v$ on $(x, u, e)$ from the notation.

**Example 1.** *Fig. 1(a) illustrates a prototype of a metabolic reaction network with negative feedback regulation. In terms of Eq. 1, we have* $x = [x_1\ x_2]^T$, $e = [e_1\ e_2\ e_3]^T$, $v = [v_1\ v_2\ v_3]^T$, *and*

$$N = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}.$$

*The linlog rate equations for this system are shown in Fig. 1(b). We will refer to this network as a running example illustrating the concepts introduced below.*

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:

$$\left(\frac{v}{e}\right)^T = [1\ \ln x^T\ \ln u^T] \cdot \begin{bmatrix} a^T \\ (B^x)^T \\ (B^u)^T \end{bmatrix}. \tag{4}$$

Note that $e$ is a vector of strictly positive elements, which enables the formulation of Eq. (4). By the linearity of (4), it holds that

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

This allows (4) to be reformulated 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 formulate our general 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* $B \triangleq \begin{bmatrix} B^x & B^u \end{bmatrix}^T$ *minimizing* $|||\tilde{W} - \tilde{Y} \cdot B|||$, *where* $||| \cdot |||$ *is a convenient matrix norm on* $\mathbb{R}^{q \times m}$.

To this avail we consider the probabilistic measurement error model:

$$\tilde{W} = W + \varepsilon \qquad \varepsilon = \begin{bmatrix} \varepsilon_1, \ldots, \varepsilon_m \end{bmatrix} \qquad \varepsilon_i \sim \mathcal{N}(0, \Sigma_{\varepsilon_i}) \tag{7}$$

$$\tilde{Y} = Y + \eta \qquad \eta = \begin{bmatrix} \eta_1, \ldots, \eta_{n_x+n_u} \end{bmatrix} \qquad \eta_j \sim \mathcal{N}(0, \Sigma_{\eta_j}) \tag{8}$$

with $\Sigma_{\varepsilon_i} = \sigma_i^2 I > 0$, $\Sigma_{\eta_j} = \nu^2 I \geq 0$, and $\varepsilon_i$, $\eta_j$ mutually independent for all $i = 1, \ldots, m$ and $j = 1, \ldots, n_x + n_u$.

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 $B = \begin{bmatrix} B^x & B^u \end{bmatrix}^T$ by way of Eq. (5). With the assumption above that measurement noise is independent across different reactions, it makes sense to separate the problem into the independent estimation of the parameter vector $B_i$ of each reaction $i$, with $i = 1, \ldots, m$.

**Example 2.** *Let* $W$ *and* $Y$ *denote the noiseless versions of* $\tilde{W}$ *and* $\tilde{Y}$*, respectively. Consider the case where* $\tilde{W} = W + \varepsilon = Y \cdot B + \varepsilon$*, i.e. the measurement error for the metabolite concentrations is negligible. Maximum likelihood estimation of* $B$ *amounts to maximize the probability (density function) of* $\tilde{W}$ *given* $Y$ *as a function of* $B$*. After simple computations and thanks to the independence assumptions on* $\varepsilon$*, one finds that the maximum likelihood estimate of* $B$ *is any solution of*

$$\min_B \ \frac{1}{2} \sum_{i=1}^{m} (\tilde{W}_i - Y B_i)^T \Sigma_{\varepsilon_i}^{-1} (\tilde{W}_i - Y B_i),$$

*wihch can be solved separately for every column of $B$ by solving, for $i = 1, \ldots, m$,*

$$\min_{B_i} \ (\tilde{W}_i - YB_i)^T \Sigma_{\varepsilon_i}^{-1} (\tilde{W}_i - YB_i) = ||\tilde{W}_i - YB_i||^2_{\Sigma_{\varepsilon_i}^{-1}}.$$

*Thus, defining $||\cdot||_i = ||\cdot||_{\Sigma_{\varepsilon_i}^{-1}}$ and*

$$|||\cdot||| : \mathbb{R}^{q \times m} \to \mathbb{R}_{\geq 0} : M \mapsto \sqrt{\begin{bmatrix} M_1 \\ \vdots \\ M_m \end{bmatrix}^T \begin{bmatrix} \Sigma_{\varepsilon_1}^{-1} & & \\ & \ddots & \\ & & \Sigma_{\varepsilon_m}^{-1} \end{bmatrix} \begin{bmatrix} M_1 \\ \vdots \\ M_m \end{bmatrix}},$$

*we see that Problem 1 is equivalent to the maximum likelihood estimation of $B$, which is in turn equivalent to separate maximum likelihood estimation of the parameters $B_i$ of each reaction $i$.*

We can now fully detail Problem 1 and express it as a series of estimation problems on individual reactions. In doing so we note that each reaction $i$ depends only on a known subset of metabolites $C(i) \subseteq \{1, \ldots, n_x + n_u\}$. Therefore, the entries of $B_i$ corresponding to metabolites that do not participate in reaction $i$ can be set to zero, and the least-squares problem can be reduced accordingly. We will address two cases, formalized by two alternative problem statements. The first we consider is a standard regression problem (Nikerel et al., 2009; Sands and Voit, 1996). In analogy with Example 2, it amounts to assuming negligible noise for metabolite concentrations.

**Problem 2.** *Given $Y$ and $\tilde{W}$ as in* (7)*, solve*

$$\min_{B_{C(i),i}} ||\tilde{W}_i - Y_{C(i)} \cdot B_{C(i),i}||_i \ , \qquad i = 1, \ldots, m \ . \tag{9}$$

The second case is more challenging and less commonly addressed in the literature. It corresponds to an errors-in-variables regression model (van Huffel and Vandewalle, 1991) and accounts explicitly for noise on the relative fluxes as well as on metabolite concentrations.

**Problem 3.** *Given $\tilde{Y}$ as in* (8) *and $\tilde{W}$ as in* (7)*, solve*

$$\min_{B_{C(i),i}} ||\tilde{W}_i - \tilde{Y}_{C(i)} \cdot B_{C(i),i}||_i \ , \qquad i = 1, \ldots, m \ . \tag{10}$$

From now on we will drop subscript $i$ from $||\cdot||_i$, the meaning being clear from the argument of the norm.

**Remark 1.** *A similar parameter estimation problem can be formulated for other pseudo-linear modeling formalisms. Models linear in metabolite concentrations (Delgado and Liao, 1992), loglin models (Hatzimanikatis and Bailey, 1997), and generalized mass-action models (Savageau, 1976) can be defined analogously to Eq. (4). This gives rise to, respectively,*

$$\left(\frac{v}{e}\right)^T = [1\ x^T\ u^T] \cdot \begin{bmatrix} a^T \\ (B^x)^T \\ (B^u)^T \end{bmatrix}, \tag{11}$$

$$(v)^T - \ln e^T = [1\ \ln x^T\ \ln u^T] \cdot \begin{bmatrix} a^T \\ (B^x)^T \\ (B^u)^T \end{bmatrix}, \tag{12}$$

$$\ln\left(\frac{v}{e}\right)^T = [1\ \ln x^T\ \ln u^T] \cdot \begin{bmatrix} a^T \\ (B^x)^T \\ (B^u)^T \end{bmatrix}. \tag{13}$$

*Notice that the modifications concern the way in which reaction rates and concentrations enter into the linear equations. The translation of these equations into variants of Problems 2–3, by removing the mean, is straightforward. In each case, we obtain a linear regression problem.*

Although below we illustrate the identifiability issues and reduction methods for the case of linlog models, it should be borne in mind that analogous results also apply to the other approximate kinetic modeling formalisms defined in Eq.s (11)–(13). However, they are not applicable to formalisms for which parameter estimation cannot be turned into linear regression, such as reversible Michaelis-Menten kinetics and convenience kinetics (Liebermeister and Klipp, 2006).

## 3. Identifiability of linlog and related 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 and investigate the identifiability of this model class along the lines of Berthoumieux et al. (2012). 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 an arbitrarily rich and errorless dataset. Structural identifiability forms the basis for studying practical identifiability, i.e. the ability to estimate parameter values from real datasets, which will be discussed further below.

The system, described by Eq.s (1)–(3), is parametrized by the parameter vector $p = [a\ B^x\ B^u]^T \in P \subseteq \mathbb{R}^{(1+n_x+n_u)\times m}$. Let $e$, $u$, $x$ and $v$ take values in the sets $E \subseteq \mathbb{R}^m_{>0}$, $U \subseteq \mathbb{R}^{n_u}_{>0}$, $X \subseteq \mathbb{R}^{n_x}_{>0}$ and $V \subseteq \mathbb{R}^m$, respectively. We assume that $e$ and $u$ are system inputs, i.e. independent variables whose values can be fixed at will. We make the following standing assumption.

**Assumption 1.** *For every $p \in P$, $e \in E$ and $u \in U$, the solution to the system of equations*

$$0 = Nv \tag{14a}$$

$$v = \mathrm{diag}(e) \cdot (a + B^x \cdot \ln x + B^u \cdot \ln u) \tag{14b}$$

*is unique in $x \in X$ and $v \in V$.*

This guarantees that, for every admissible parametrization and system input, a steady-state exists and is unique. In order for this steady-state to be observable experimentally, we also make the assumption that it is locally asymptotically stable. In accordance with the metabolic control theory literature (Heinrich and Schuster, 1996), fluxes $v$ at steady-state are denoted by $J$. We write $J_p(e,u)$ to emphasize dependence on inputs and model parameters.

For varying values of $p$, Assumption 1 enables us to express the linlog model with parameters $p$ as a map

$$\mathscr{M}_p:\ E\times U\to V\times X:\ (e,u)\mapsto \big(J_p(e,u),x_p(e,u)\big). \tag{15}$$

Assumption 1 is met, in particular, when matrix $N\operatorname{diag}(e)B^x$ is invertible. In this case, one may write the output $(J_p,x_p)=\mathscr{M}_p(e,u)$ as an explicit function of the input $(e,u)$,

$$J_p(e,u)=\operatorname{diag}(e)\cdot(a+B^x\cdot\ln x_p(e,u)+B^u\cdot\ln u), \tag{16}$$

$$\ln x_p(e,u)=-(N\operatorname{diag}(e)B^x)^{-1}\cdot N\operatorname{diag}(e)\cdot(a+B^u\cdot\ln u). \tag{17}$$

As can be easily verified, this requires that the stoichiometry matrix $N$ is full row rank, which is the case for systems with no mass conservation constraints (Heinrich and Schuster, 1996).

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 $i$th 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. Recall that $p_i$ is the $i$th column of $p$, i.e. the parameter vector for reaction $i$.

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

$$\big((J_p)_i,x_p\big)|_D=\big((J_{p^*})_i,x_{p^*}\big)|_D\Rightarrow p_i=p_i^*. \tag{18}$$

Here $(J_p,x_p)$ is seen as a function from $E\times U$ to $X\times V$, and "$|_D$" is its restriction to the input set $D$. In words, 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 predictions of $\mathscr{M}_p$ and $\mathscr{M}_{p^*}$ are identical over all possible input sets $D$. Note that this definition is applicable to any metabolic reaction model, provided suitable definition of the parameters of the model class. In particular, it applies to the linlog form of the reaction rates as well as to any other pseudo-linear form reviewed in Section 2.

How can we apply Def. 1 to the analysis of identifiability of linlog models? The following proposition establishes a link between this definition and the uniqueness of the solution to Problems 2–3. Given the input set $D=\{(e^1,u^1),\cdots,(e^q,u^q)\}$ 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 structurally identifiable at $p^*$ if and only if there exists $D = \{(e^1, u^1), \cdots, (e^q, u^q)\} \subseteq E \times U$ such that the solution of the equation $W_i^* = Y^* B_i^*$, with*

$$W_i^* = \begin{bmatrix} \left(\frac{J_*^1}{e^1} - \overline{\left(\frac{J_*}{e}\right)}\right)_i & \cdots & \left(\frac{J_*^q}{e^q} - \overline{\left(\frac{J_*}{e}\right)}\right)_i \end{bmatrix}^T,$$

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

*is unique in the parameters $B_i^* = \left([B^{x*}\ B^{u*}]^T\right)_i$.*

**Corollary 1.** *A reaction $i$ of $\mathscr{M}_p$ is structurally identifiable at $p^*$ if and only if there exists $D = \{(e^1, u^1), \cdots, (e^q, u^q)\} \subseteq E \times U$ such that $Y^*_{C(i)}$ is full column-rank.*

To complete the link with Section 2, notice that the matrices $Y^*$ and $W^*$ computed from the inputs and outputs of the model with true parameters $p^*$ coincide with the noiseless measurement matrices $Y$ and $W$, respectively.

It is clear that the rank condition of Corollary 1 can be fulfilled only for a number of experiments $q = |D|$ greater than or equal to the number of unknown parameters $n_i = |C(i)|$ of reaction $i$. The possibility to find $q \geq |C(i)|$ experiments making $Y^*_{C(i)}$ full column rank depends on the network model and parameters themselves. Indeed, in our framework, the experimentalist can impose different enzyme concentrations $e$ and inputs $u$, but the resulting metabolite concentrations are determined by the network. In other words, there is no full control of the regression matrix $Y^*$, which impairs the design of optimal experiments for parameter regression. We show this by a simple example.

**Example 3.** *Consider the negative feedback network structure shown in Fig. 1. Let us define the network parameter values*

$$a = \begin{bmatrix} a_1 \\ 0.0297 \\ 0.0296 \end{bmatrix}, \quad B^x = B^T = \begin{bmatrix} -0.0938 & B_{2,1} \\ 0.0286 & -0.0073 \\ 0 & 0.0287 \end{bmatrix},$$

*where different values of $a_1 \in \mathbb{R}$ and $B_{2,1} \in \mathbb{R}_{<0}$ (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 $a_1, B_{2,1}$, the equation $Nv(x, e) = N\,\mathrm{diag}(e)(a + B^x \ln x) = 0$ yields a unique solution $\ln x = -(N\,\mathrm{diag}(e)B^x)^{-1} N\,\mathrm{diag}(e)a$. This defines the unique steady-state of the system. Provided it is asymptotically stable, this gives us a steady-state of the system that can be observed experimentally. One first consideration is that different values of $a_1$ and $B_{2,1}$ 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 values of $B_{2,1}$ equal to $-0.0073$ (weaker feedback action) and $-7.2961$ (stronger feedback action), respectively, scatter plots of the steady-state solutions for $\ln x$ from 1000 randomly generated samples of $e$ are reported in Fig. 2. Steady-state metabolite concentrations in the case of weak feedback are spread similarly in all directions, while with stronger feedback they are essentially aligned along a one-dimensional line. Here the strong feedback exerted by metabolite $X_2$ on the production of $X_1$ induces a negative correlation between their concentrations, which may*

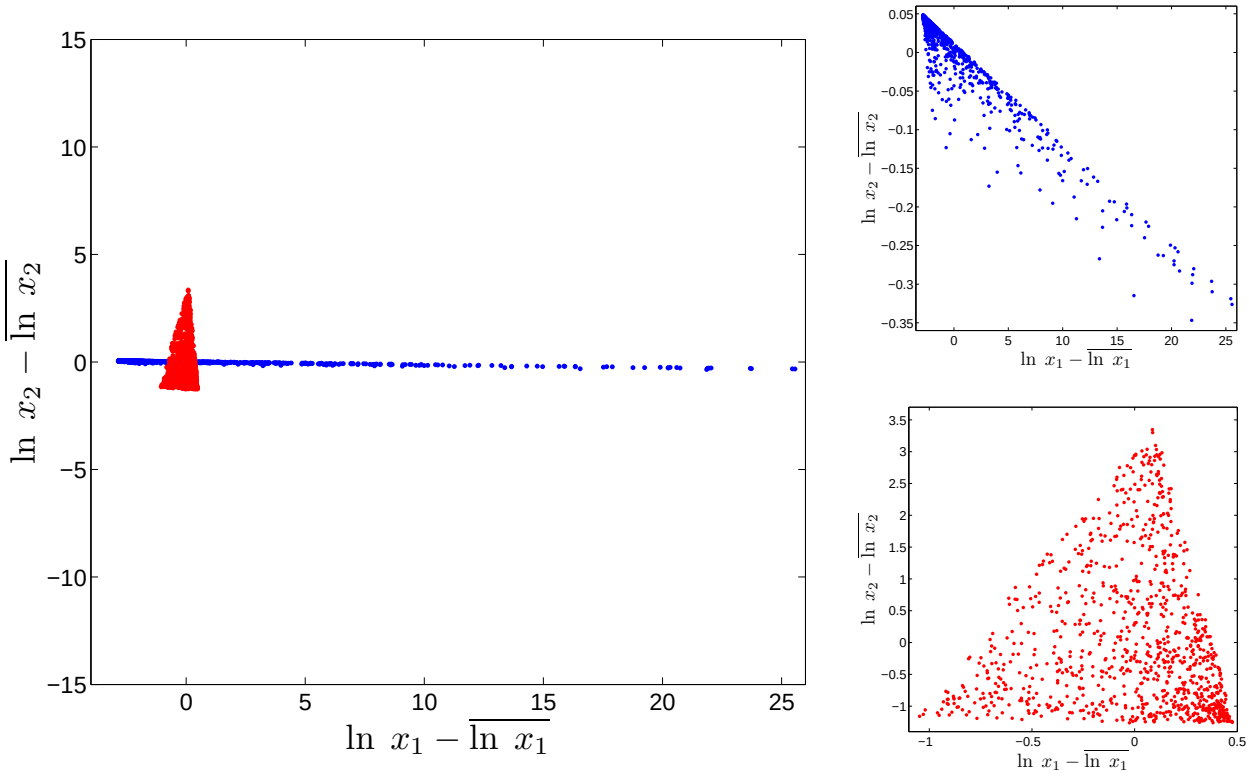



Figure 2: Left: Scatter plot of steady-state metabolite concentrations for 1000 randomly generated enzyme concentrations, for two different parametrizations of the model of Fig. 1 (see the text of Example 3 for more details). Red: Simulation for $a_1 = 0.0297$ and $B_{2,1} = -0.0073$ (weak feedback); Blue: Simulation for $a_1 = 0.0297$ and $B_{2,1} = -7.2961$ (strong feedback). Right: Individual zooms of the two datasets, with consistent coloring.

*result in an ill-conditioned estimation problem. In addition this strong feedback results in a near homeostasis of $X_2$ that may also impede identification. These points 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 $N\,\mathrm{diag}(e)(a + B^x \ln x) = 0$ always corresponds to the value of $\ln x$ satisfying $B^x \ln x = -a$, independently of the value of $e$. Thus, no matter the number of experiments $q$, the rank of $Y^*$ is at most 1 and the model is not identifiable. In our example, this is the case for $a_1 = -7.5491$ and $B_{2,1} = -7.2961$.*

From the example above it is clear that a reaction may be nonidentifiable for specific values of the models parameters even if it is identifiable for other parameterizations. In the light of this, a generalization of Def. 1 from reaction identifiability *at a parameter* $p^*$ to reaction identifiability *tout court* can be obtained following Walter and Pronzato (1997). Namely, we stipulate that a reaction (and by extension, a model) is identifiable if it is identifiable almost everywhere in $P$, i.e. at almost every parametrization $p^* \in P$ of $\mathscr{M}_p$. Here 'almost everywhere' and 'almost every' are interpreted in terms of a suitable (e.g. Lebesgue) measure on $P$. Hence, the negative feedback network structure of Example 3 is identifiable in the sense of Walter and Pronzato. The identifiability criterion of Def. 1 holds except for the "rare" parameter combinations $p^*$ such that $a \in \mathrm{span}(B^x)$.

A second observation, following from the example above, is that the mathematical conditions that the system must fulfill to be declared nonidentifiable are too strong to be useful in practice. If we look at Fig. 2, we see that strong collinearities exist between the metabolite concentrations $x_1$ and $x_2$. As a result, an unreasonably large number of experiments would be needed to resolve the effects of the two. Moreover, the definition of identifiability assumes that the measurements are not corrupted by noise, which is even less realistic. We therefore need to weaken our 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 $D$ be a fixed set of $q$ inputs (external metabolites and enzyme concentrations), and let $O$ be the set of the corresponding system outputs (fluxes and steady-state concentrations of internal metabolites determined by $\mathscr{M}_{p^*}$). Consider the problem of estimating the parameters $B_i^*$ of reaction $i$ given observations of $D$ and $O$ affected by measurement error. An *estimator* $\hat{B}_i$ of $B_i^*$ is a function of the observations of $D$ and $O$, well-defined for every possible (a priori unknown) value of $p^*$ (compare (Ljung, 1999, §7.4)). Since, due to noise, the observations are stochastic variables, $\hat{B}_i$ is itself a stochastic variable. Therefore, one cannot hope to estimate $B_i^*$ exactly, but only within a certain degree of approximation. 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 $B_i$, i.e., $B_{C(i),i}$. Let $\mathscr{B}_i \subset \mathbb{R}^{n_i}$ be a *bounded* neighbourhood of the origin, and let $\alpha \in (0, 1)$.

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

$$\mathbb{P}_{p^*}[\hat{B}_{C(i),i} - B^*_{C(i),i} \in \mathscr{B}_i] \geq 1-\alpha, \tag{19}$$

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

Note that this definition is conceptually different from the one suggested by Raue et al. (2009), where the definition of practical identifiability requires that the uncertainty on the parameter estimates (as defined via the profile likelihood) is bounded, but contrary to our definition, can be arbitrarily large. In addition, the definition in (Raue et al., 2009) is given in terms of a specific, not necessarily optimal choice of the estimator.

The point of view expressed by Def. 2 is that 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 $\alpha$. Then, the possibility of fulfilling (19), i.e. the practical identifiability of the model, depends on the system itself and on the richness of the input set $D$. In general, the larger the $D$, the tighter the requirements that one can fulfill (i.e. the smaller the values of $\mathscr{B}_i$ and $\alpha$ for which practical identifiability in the sense of Def. 2 holds).

From an alternative viewpoint, one may start from a given input set $D$, and look for the choices of $\alpha$ and $\mathscr{B}_i$ that ensure satisfaction of (19). Here in turn, one may fix $\alpha$ and look for the $\mathscr{B}_i$ that makes (19) achievable, or fix the acceptable estimation uncertainty $\mathscr{B}_i$ and establish at what confidence level $\alpha$ this performance can be attained.

In all of the above cases, the natural questions that arise are how Def. 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, the relation between observations and observed quantities must be specified. We refer to the measurement model introduced in Section 2. For the sake of simplicity, we assume for this section that $\nu = 0$, i.e. we address Problem 2. Problem 3 can be addressed with the same tools, but at the price of technical complications.

The following proposition answers the questions above.

**Proposition 2.** *If a reaction $i$ of $\mathscr{M}_p$ is structurally identifiable at $p^*$ in the sense of Def. 1 then, for every $\alpha \in (0,1)$, it is practically identifiable in the sense of Def. 2 with confidence level at least $1-\alpha$ for any uncertainty set $\mathscr{B}_i \supseteq \mathcal{E}_{\widehat{\Sigma}}(\alpha)$, where $\mathcal{E}_{\widehat{\Sigma}}(\alpha)$ denotes the $(1-\alpha)$-confidence ellipsoid of a zero-mean Gaussian distribution with variance $\widehat{\Sigma} = (Y^T_{C(i)}\Sigma^{-1}_{\varepsilon_i}Y_{C(i)})^{-1}$.*

The proof relies on the use of minimum variance estimators, as dictated by standard results in linear estimation theory (see (Ljung, 1999, Appendix II) and also Appendix B).

[1] Strictly speaking, a better version of Def. 2 would require that condition (19) holds for all $p^*$ in $P$. This would automatically rule out trivial definitions of $\hat{B}_{C(i),i}$ such as $\hat{B}_{C(i),i} \triangleq B^*_{C(i),i}$ (which makes the reaction identifiable for any $\alpha$ 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 estimation uncertainty may severely depend on $p^*$ itself, as we shall see later on in Example 4. For simplicity, here we stick to Def. 2 with the understanding that any such triviality is avoided.

From now on, estimation will be discussed and performed based on this type of estimator. Observe that $\mathcal{E}_{\widehat{\Sigma}}(\alpha)$, and hence the shape and size of the uncertainty regions $\mathscr{B}_i$ for which the model is practically identifiable, depends on the choice of inputs $D$. In particular, the noise on $W_i$ affects the covariance matrix $\widehat{\Sigma}$ by its statistics $\Sigma_{\varepsilon_i}$, while the contribution of the data $Y_{C(i)}$ is apparent. The number of data points $q$ enters the picture in terms of the size of the matrix $\Sigma_{\varepsilon_i}$ and the number of rows of $Y_{C(i)}$. Typically, the larger $q$, the smaller $\mathscr{B}_i$ can be for a fixed $\alpha$. 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 more complicated characterization of $\mathscr{B}_i$. Finally, one may speak about identifiability of the whole model, e.g. by requiring that each reaction $i$ is individually identifiable with a given confidence level $\alpha$ and uncertainty set $\mathscr{B}_i \in \mathbb{R}^{n_i}$. Alternatively, one may require that all reactions be simultaneously identifiable with confidence level $1-\alpha$ and a suitably defined joint uncertainty set.

A discussion of practical identifiability in terms of covariance matrix of a (linearized) parameter estimation problem also appears in (Srinath and Gunawan, 2010), in the context of power-law models. However, the discussion in Srinath and Gunawan (2010) is essentially limited to one particular choice of the admissible estimation uncertainty $\mathscr{B}_i$, namely the one ensuring that the sign of the parameter values is estimated correctly with probability $1-\alpha$.

A useful tool for better understanding Proposition 2 and the links between available data and practical identifiability is the Singular Value Decomposition (SVD) of a matrix (Jolliffe, 1986). The SVD of $Y_{C(i)}$ is given by

$$Y_{C(i)} = USV^T, \qquad S = \mathrm{diag}(s_1, s_2, ..., s_{n_i}), \tag{20}$$

with $U \in \mathbb{R}^{q\times n_i}$ and $V \in \mathbb{R}^{n_i\times n_i}$ orthonormal matrices and $s_1 \geq \cdots \geq s_{n_i} \geq 0$ the singular values of $Y_{C(i)}$. In the presence of dependencies between the columns, there exists an index $r$ with $1 \leq r < n_i$ such that $s_{r+1} = \ldots = s_{n_i} = 0$, and $Y_{C(i)}$ is of rank $r$. Based on this, the covariance matrix of Proposition 2 can be written as $\widehat{\Sigma} = (Y_{C(i)}^T \Sigma_{\varepsilon_i}^{-1} Y_{C(i)})^{-1} = VS^{-1}U^T\Sigma_{\varepsilon_i}US^{-1}V^T$. Using the assumption that $\Sigma_{\varepsilon_i} = \sigma_i^2 I$, the previous formula simplifies to

$$\widehat{\Sigma} = V\sigma_i^2 S^{-2} V^T, \tag{21}$$

which is (up to resorting of the entries) the SVD of $\widehat{\Sigma}$, with singular values $\sigma_i^2 s_1^{-2}, \ldots, \sigma_i^2 s_{n_i}^{-2}$. Multiplied by a factor $\lambda(\alpha)$ fixed by $\alpha$, the square roots of these values define the length of the axes of the confidence ellipsoid of Proposition 2. Now suppose that we seek parameter estimates that, with confidence $1-\alpha$, fall within a ball $\mathscr{B}_\delta = \{p : \ |p| < \delta\}$, for some $\delta > 0$. That is, all the entries of the parameter vector must be estimated with accuracy at least $\delta$. From Proposition 2 and Def. 2, reaction $i$ is practically identifiable if it is structurally identifiable and, for the given input set $D$, the ellipsoid $\mathcal{E}_{\widehat{\Sigma}}(\alpha)$ associated with (21) fits into $\mathscr{B}_\delta$, which happens if $\lambda(\alpha)\sigma_i/s_\ell < \delta$ for $\ell = 1, \ldots, n_i$. In turn, since $s_1 \geq \ldots \geq s_{n_i}$, this holds whenever $\lambda(\alpha)\sigma_i/s_{n_i} \leq \delta$, i.e. the smallest singular value of $Y_{C(i)}$ dictates the overall estimation performance.

If $Y_{C(i)}$ is ill-conditioned, i.e., some data vectors are nearly collinear, large discrepancies exist between its largest and its smallest singular values. Then, the condition $s_1 \gg s_{n_i}$ implies that, for practical identifiability, it must hold that $\lambda(\alpha)\sigma_i/s_1 \ll \delta$, i.e. a high accuracy in the estimation of the components of $p$ along direction $V_1$ is required. Thus, achieving an even mild

accuracy $\lambda(\alpha)\sigma_i/s_{n_i} \leq \delta$ along direction $V_{n_i}$ would generally require an unreasonable amount of experimental effort in terms of experimental replicas and/or measurement accuracy (see also Gutenkunst et al. (2007)). Note, however, that high accuracy along $V_1$ is solely needed to ensure that the less accurate estimates of the components of $p$ in direction $V_{n_i}$ be acceptable. In a sense, this hard requirement is an artifact of the problem statement: If we accept that certain components are just not relevant, the remaining part of the model is identifiable in practice with good accuracy and much less experimental effort. To quantify our discussion, let us look at a numerical example.

**Example 4.** *To illustrate the implications for parameter identifiability of a poorly conditioned data matrix, consider the estimation results from noisy and finite datasets for the two different identifiable parameterizations of the model of Example 3. As in the latter example, 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% of the standard deviation) were tested, resulting in a total of 4 experimental scenarios for each model parameterization. For each scenario, 100 datasets were simulated and the corresponding estimates for reaction 2 are reported in the scatter plots of Fig. 3(a) ($a_1 = 0.0297$ and $B_{2,1} = -0.0073$) and 3(b) ($a_1 = 0.0297$ and $B_{2,1} = -7.2961$).*

*An immediate observation is that the shape of the 95%-confidence ellipse of the parameter estimates is different for the two model parameterizations. While estimation accuracy for $B_{1,1}$ and $B_{2,1}$ is comparable in the case of weaker feedback ($B_{2,1} = -0.0073$), the shape of the uncertainty ellipse becomes very skewed in the case of stronger feedback ($B_{2,1} = -7.2961$).*

*In particular, in the latter case estimation accuracy is much higher for $B_{1,1}$ than for $B_{2,1}$ regardless of the features of the dataset because of the strong homeostasis on $x_2$ (note the change in scale of the vertical axes of the plots in Fig. 3(a) and Fig. 3(b)). The plots also show that larger and/or less noisy datasets improve estimation performance, as expected. However, in the case of $B_{2,1}$ in Fig. 3(b), this improvement is seen to require extremely large and high-quality datasets. In other words, accurately estimating all parameters of the model demands a significant increase in experimental effort, even when most individual parameters are easy to estimate.*

*Following upon Example 3, it is apparent that the skewed estimation uncertainty in part (b) 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 $\alpha$, 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.*

To summarize the main points of the section, we have discussed practical identifiability as a relative concept that depends on the parameter estimation uncertainty that is deemed acceptable. If this is compatible with the quality of the data (dataset size, amount of noise) and the dataset is sufficiently diverse (more independent components than unknown parameters), then practical identifiability follows from structural identifiability. On the other hand,

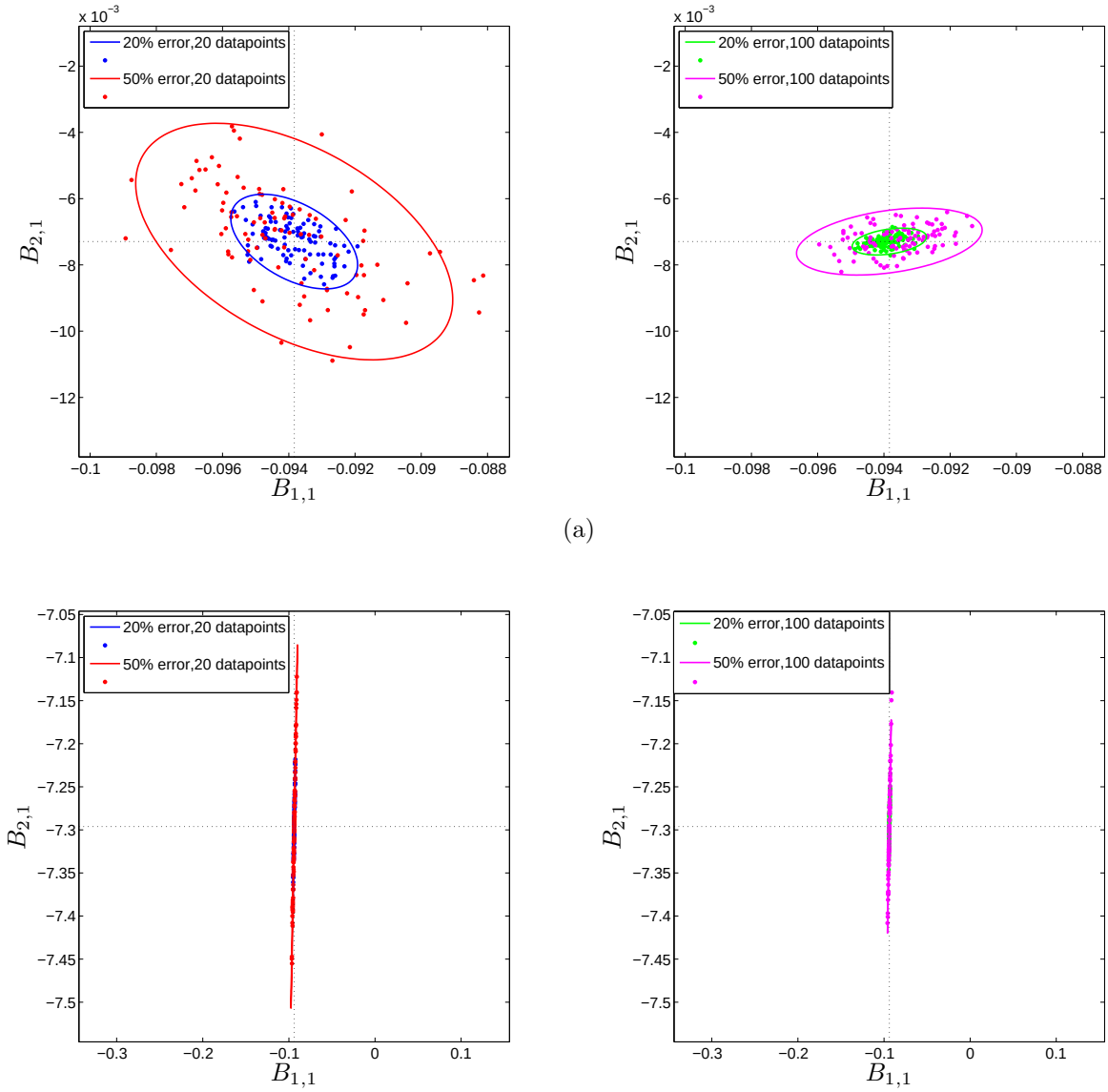



Figure 3: Estimates of parameters $B_{1,1}, B_{2,1}$ (first row of $B^x$) from metabolite concentrations, enzyme concentrations and flux measurements at steady-state, for a linlog model of the negative feedback network in Fig. 1. In each panel, scatter plots are reported for four different experimental scenarios: 20% noise and $q = 20$ datapoints (blue); 20% noise and $q = 100$ datapoints (green); 50% noise and $q = 20$ datapoints (red); 50% noise level and $q = 100$ datapoints (magenta). 95%-confidence ellipses are drawn for each scenario (dashed lines). Reference parameter values are indicated by the intersection of horizontal and vertical dotted lines. Refer to the text of Example 4 for additional details. True parameter values are given in Example 3. (a) Case $a_1 = 0.0297$ and $B_{2,1} = -0.0073$. (b) Case $a_1 = 0.0297$ and $B_{2,1} = -7.2961$.

if the experiments are not informative enough or the network itself implies heavy correlations among the data, we detect lack of practical identifiability from the existence of nearly zero singular values in the decomposition of the regression matrix. For metabolic systems this will occur when feedback regulation results in strong homeostasis or when metabolite concentrations are correlated, which could be due in general to steady-state constraints. A particularly relevant situation concerns reactions that operate close to equilibrium. Indeed in this situation mass-action law relates metabolite concentrations as follows:

$$\sum_j N_{j,i} \ln x_j \approx \ln K_i \tag{22}$$

where $K_i$ is the equilibrium constant of reaction $i$. This results in a quasi-dependency between the $\ln x_j$ that makes the reaction nonidentifiable in practice.

In addition to providing little information for the estimation of the model parameters, the smallest components of $Y_{C(i)}$ have negligible effect on the solution of the steady-state equations (3)–(6), i.e. in determining the system steady-state. In the next section, we will build upon this analysis to define a criterion for eliminating the components of $Y_{C(i)}$ associated with the smallest singular values and reduce the network model accordingly. This will make parameter estimation (regression) a well-conditioned problem for every single reaction while minimally affecting the quality of the model.

## 4. Reduction to identifiable models

It was shown in Example 4 that attempting to identify the parameters of a reaction that is not practically identifiable leads to an ill-posed estimation problem. That is, certain parameter combinations are practically indistinguishable from the data. Eliminating redundant components of the model parameters by Principal Component Analysis (PCA, see Jolliffe (1986)) is a way to ensure well-posedness of parameter estimation (i.e. practical identifiability) by a "minimal" approximation of the model.

The method applies as usual reaction by reaction. For any given reaction $i$, with $i = 1, \ldots, m$, we compute and manipulate the SVD of the matrix $\tilde{Y}_{C(i)}$, so as to get transformed data and a corresponding model with a reduced number of parameters that can be estimated reliably.

### *4.1. Identifiability analysis and model reduction by PCA*

We start by considering the case where the regression matrix is noiseless, i.e. $\tilde{Y}_{C(i)} = Y_{C(i)}$, and $\operatorname{rank} Y_{C(i)} = r$, with $r < n_i$. Notice that the latter is always the case for structurally nonidentifiable models. The extension of the method to practical identifiability (where $Y_{C(i)}$ is full column rank but ill-conditioned) and to noisy and incomplete data $\tilde{Y}$ will be discussed in the next sections.

Consider again the SVD $Y_{C(i)} = U \cdot \operatorname{diag}(s_1, s_2, ..., s_{n_i}) \cdot V^T$, with $s_1 \geq s_2 \geq \ldots \geq s_{n_i} \geq 0$. Since $r < n_i$, it holds that $s_1 \geq \ldots \geq s_r > 0$ and $s_{r+1} = \ldots = s_{n_i} = 0$. Then, $Y_{C(i)}$ has an $(n_i - r)$-dimensional kernel $K_Y$, given by $K_Y = \operatorname{range}(V_{r+1:n_i})$. For any $B_{C(i),i}$ and any $k_Y \in K_Y$, it holds that $Y_{C(i)} \cdot B_{C(i),i} = Y_{C(i)} \cdot (B_{C(i),i} + k_Y)$. For the purpose of identification, this means that $B_{C(i),i}$ cannot be uniquely reconstructed from the data. On the other hand, $\operatorname{range}(Y_{C(i)}) = \operatorname{range}(Y_{C(i)} V_{1:r})$, where $Y_{C(i)} V_{1:r}$ is full column rank. Then, for every $B_{C(i),i}$

there exists a unique $\breve{B}_i \in \mathbb{R}^{r\times 1}$ such that $Y_{C(i)} \cdot B_{C(i),i} = Y_{C(i)}V_{1:r} \cdot \breve{B}_i$. This suggests to modify the regression problem $W_i = Y_{C(i)} \cdot B_{C(i),i} + \varepsilon_i$ into

$$\begin{cases} W_i = \breve{Y}_i \cdot \breve{B}_i + \varepsilon_i \\ \breve{Y}_i = Y_{C(i)}V_{1:r} \end{cases} \tag{23}$$

which has a unique solution in $\breve{B}_i$, i.e. $\breve{B}_i$ is identifiable. We call (23) the reduced model and $\breve{B}_i$ the reduced parameter vector.

For a fixed outcome of the noise $\varepsilon_i$, from the unique solution $\breve{B}$ in the reduced parameter space one can infer a whole subspace of equivalent solutions in the original parameter space as $\{B_{C(i),i} = V_{1:r}\cdot\breve{B}_i + k_Y, \quad k_Y \in K_Y\}$. Thus, in general, a fixed solution $\breve{B}_i$ does not determine uniquely any of the parameters $B_{j,i}$ ($j$ being element of $C(i)$). However, depending on the structure of $V$, we may be able to isolate some parameters $B_{j,i}$ that can be reconstructed without ambiguity.

**Proposition 3.** *Let index $j$ be an element of $C(i)$, i.e., for some $\ell = 1, \ldots, n_i$, $j = C_\ell(i)$. Suppose that the entries of $V_{\ell,r+1:n_i}$ are all zero. If $\breve{B}_i$ is the (unique) solution to* (23)*, then $B_{j,i} = V_{\ell,1:r}\breve{B}_i$ is uniquely determined.*

A similar, but less general approach to separate identifiable from nonidentifiable parameters has been considered by (Nikerel et al., 2009).

*4.2. Model reduction put in practice*

In a real setting, as shown in Example 4, small nonzero values of $s_{r+1}, \ldots, s_{n_i}$ can also make the problem of estimating $B_{C(i),i}$ ill-conditioned, thus preventing practical identifiability. In addition, measurement error can make certain components of the data indistinguishable from noise. The idea here is to remove the components of the parameters that are poorly determined from the data, thus ensuring smaller estimation uncertainty and hence practical identifiability in a reduced parameter space. In order to develop and explain the rationale of our method, we will first reconsider model reduction in the setting of Problem 2 where the metabolite data are assumed noiseless, and then move on to the more realistic scenario of Problem 3 where metabolite data are noisy. In the remarks concluding the section, we will briefly discuss the application of the method to datasets with missing or corrupted entries (e.g. outliers) and its biological interpretation. We will then summarize the model reduction procedure in Section 4.3.

*The scenario of Problem 2.* Here $\tilde{Y}_{C(i)} = Y^*_{C(i)}$. One may consider the rank-$r$ approximation of the data matrix

$$Y^*_{C(i)} = U \cdot \mathrm{diag}(s_1, s_2, ..., s_{n_i}) \cdot V^T \simeq U \cdot \mathrm{diag}(s_1, s_2, ..., s_r, \underbrace{0, \ldots, 0}_{n_i - r}) \cdot V^T = \hat{Y}_i.$$

Following the previous section, for $\breve{Y}_i = \hat{Y}_i V_{1:r}$, consider the reduced model

$$\begin{cases} W_i = \breve{Y}_i \cdot \breve{B}_i + \varepsilon_i \\ \breve{Y}_i = \hat{Y}_i V_{1:r} \end{cases} . \tag{24}$$

with unique least-squares solution for the reduced parameter $\breve{B}_i$. With the same arguments as in Section 3.2, for $\Sigma_{\varepsilon_i} = \sigma_i^2 I$, one observes that the confidence ellipsoid associated with the estimate of $\breve{B}_i$ is determined by the matrix $\sigma_i^2 \left(\breve{Y}_i^T \breve{Y}_i\right)^{-1}$. In particular, the largest axis length, corresponding to the largest parameter estimation uncertainty, is proportional to $\sigma_i / s_r$, i.e. it has been reduced by a factor $s_r / s_{n_i}$ with respect to the original estimation problem. This analysis suggests a criterion for the choice of $r$ based on our definition of practical identifiability. Suppose that, with a given confidence $100 \cdot (1-\alpha)\%$, the admissible uncertainty $\mathscr{B}_i$ is a ball of radius $\delta$. Since the radii of the estimation error confidence ellipsoid are given by $\lambda(\alpha)\sigma_i / s_r \geq \ldots, \geq \lambda(\alpha)\sigma_i / s_1$ it suffices to choose $r$ as the minimum value for which $\lambda(\alpha)\sigma_i / s_r \leq \delta$ for the reduced model to be practically identifiable. If this holds for $r = n_i$, the full model is practically identifiable and needs no reduction.

*The scenario of Problem 3.* Here the noisy versions $\tilde{Y}$ of $Y$ are the available data. The idea is to remove from the problem not only the components that make estimation ill-conditioned, but also those components that are detrimental in that they are dominated by noise. To do this, let us look at the empirical covariance matrix of the data $\tilde{Y}_{C(i)}^T \tilde{Y}_{C(i)} / q$. From the approximation[2]

$$\tilde{Y}_{C(i)}^T \tilde{Y}_{C(i)} = Y_{C(i)}^T Y_{C(i)} + \eta_{C(i)}^T \eta_{C(i)} + Y_{C(i)}^T \eta_{C(i)} + \eta_{C(i)}^T Y_{C(i)} \simeq Y_{C(i)}^T Y_{C(i)} + q\Sigma_{\eta_{C(i)}},$$

where $\Sigma_{\eta_{C(i)}} = \nu^2 I$, it follows that

$$\tilde{Y}_{C(i)}^T \tilde{Y}_{C(i)} / q \simeq \left(U \operatorname{diag}(s_1, \ldots, s_{n_i}) V^T\right)^T \left(U \operatorname{diag}(s_1, \ldots, s_{n_i}) V^T\right) / q + \nu^2 I = V \left(\operatorname{diag}(s_1^2, \ldots, s_{n_i}^2) / q + \nu^2 I\right) V^T.$$

The expression after the last equality sign is clearly the SVD of $\tilde{Y}_{C(i)}^T \tilde{Y}_{C(i)} / q$, with singular values $\tilde{s}_\ell^2 = s_\ell^2 / q + \nu^2$, with $\ell = 1, \ldots, n_i$, composed of the signal contribution $s_j^2 / q$ and the noise contribution $\nu^2$. In the light of this, to remove the components of the data dominated by noise, we compute the (noisy) singular values $\tilde{s}_1^2 \geq \tilde{s}_2^2 \geq \ldots \geq \tilde{s}_{n_i}^2 \geq 0$ from the SVD of $\tilde{Y}_{C(i)}^T \tilde{Y}_{C(i)} / q$, draw estimates $\hat{s}_\ell^2$ of the true (noiseless) singular values $s_\ell^2$ by posing $\hat{s}_\ell^2 = \max(0, \tilde{s}_\ell^2 - \nu^2)$ for $\ell = 1, \ldots, n_i$, and define what we call the "effective rank" of the data matrix as follows.

**Definition 3.** *The effective rank of the data matrix is*

$$r = \max\{\ell : \ \hat{s}_\ell^2 \geq \nu^2, \ \ell = 1, \ldots, n_i\} \tag{25}$$

According to this definition, the effective rank indicates the number of independent components that can be safely distinguished in the data in that not blurred by noise. Notice that noise, by its very nature, tends to decorrelate all matrix entries. Following on the discussion for the scenario of Problem 2, this criterion may also be seen as implementing model reduction for practical identifiability, with a choice of the uncertainty region $\mathscr{B}_i$ depending on $\nu$, i.e. adapted to the presence of noise on metabolite data.

[2]This holds as an equality in the sense of expectation, and can also be motivated by asymptotic arguments as $q \to +\infty$.

An alternative approach to determine the effective rank of a data matrix, useful when $\nu$ is assumed small but not known with certainty, is to remove the components associated with the smallest singular values by setting $r$ so that a suitable proportion $\theta \in (0,1)$ of the total variance $\sum_{\ell=1}^{n_i} \tilde{s}_\ell^2$ of the data is retained (Berthoumieux et al., 2011). This gives rise to the following definition of effective rank.

**Definition 4.** *The* $100 \cdot \theta\%$*-variance effective rank of the data matrix is*

$$r = \min \left\{ r' : \sum_{\ell=1}^{r'} \tilde{s}_\ell^2 \geq \theta \cdot \sum_{\ell=1}^{n_i} \tilde{s}_\ell^2,\ r' = 1, \ldots, n_i \right\}. \tag{26}$$

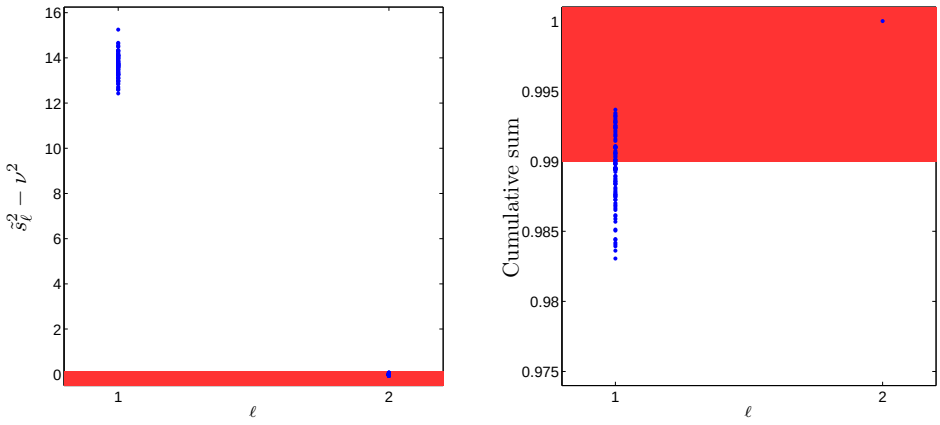
Different from the previous definition, effective rank is intended here simply as the number of components needed to express most of the data content. When applied to data with small noise, data components dominated by noise are also small and are hence excluded from the count. For large noise levels, this reasoning no longer applies. Note that precise knowledge of the noise variance $\nu^2$ is not required here, at the price of a rather uninformed choice of parameter $\theta$.

In both cases, after computing the effective rank $r$, the original model can be replaced by the reduced model (24), providing us with a well-behaved model for the subsequent identification of the system.

**Example 5.** *We have seen in Example 4 that the first reaction of the feedback model with* $a_1 = 0.0297$ *and* $B_{2,1} = -7.2961$ *is not practically identifiable and we want to see which approach to the choice of* $r$ *enables PCA to detect this property on a limited noisy dataset. In order to mimic available experimental data (Ishii et al., 2007), noise was added to the data matrix* $Y$ *by drawing* $q = 30$ *values from a normal distribution with standard deviation of* $0.4$*, corresponding approximately to* $40\%$ *noise for metabolite concentrations.* $100$ *datasets were generated in this way and PCA was performed on each of them. Two different criteria for the choice of the reduced model order* $r$ *were tested, based on Def. 3 and Def. 4, respectively. Fig. 4(a) shows the estimates of the squared singular values and the cutoff of* $\nu^2$ *proposed by Def. 3, while Fig. 4(b) shows the normalized cumulative sums of the squared singular values and the cutoff of* $0.99$ *proposed by Def. 4. The second squared singular value is always smaller than* $\nu^2$*, so that the model is correctly and consistently found nonidentifiable with the first definition, contrary to what is found with the second definition* $57$ *times out of* $100$*.*

The above example thus illustrates that the criterion for model reduction taking into account the noise level, when applicable, is more relevant.

**Remark 2.** *The data matrix* $\tilde{Y}_{C(i)}$ *may suffer from the lack of certain data entries, typically due to the removal of outliers or faults of the experimental machinery. A simple but wasteful option to recover a full data matrix for later use in a well-defined model reduction/parameter estimation problem is to discard those data points* $\tilde{Y}_{k,C(i)}$*, and the corresponding flux information* $\tilde{W}_{k,i}$*, suffering from the absence of some entry. In Berthoumieux et al. (2011), in the context of parameter estimation, we have proposed methods compensating for the missing entries by the use of statistical priors inferred from the available data. For the sake of model reduction, which requires in our approach the SVD of the data matrix, completion of the*




(a) (b)

Figure 4: Identifiability analysis for the feedback model in Fig. 1, with $a_1 = 0.0297$ and $B_{2,1} = -7.2961$. (a) Squared singular values for 100 data matrices $Y$ (see text of Example 5). The blue dots are the estimates $\tilde{s}_\ell^2 - \nu^2$ of the squared singular values for all datasets, and the red box covers the area below the cutoff of $\nu^2$ (Def. 3). (b) Normalized cumulative sum of the squared singular values $\tilde{s}_\ell^2$ for 100 data matrices $Y$. The red box displays the area above the cutoff $\theta = 0.99$ (Def. 4).

*data matrix by a suitable imputation method was suggested. Several imputation methods can be considered, still relying on statistics from the available data, such as multiple imputation or completion by the mean of the available metabolite data (see Berthoumieux et al. (2011) and references therein). As an appealing alternative we cite "minimal rank" SVD, which has been developed and applied in Brand (2002) for reduced-order modelling in computer vision.*

**Remark 3.** *The reduced model* (24) *is expressed in terms of parameters that are linear combinations of the original parameters. As a consequence, the results of identification may be difficult to interpret from a biological point of view. Proposition 3 suggests a way to isolate identifiable parameters in nonidentifiable reactions, and thus extract partial, but unambiguous information from the dataset. Unfortunately, the condition of the proposition is not usually verified in practice since, due to noise, the entries of the data kernel-generating matrix* $V_{r+1:n_i}$ *will never be exactly* 0*. In order to ease the interpretation of the results, one may relax this condition as follows. Bearing in mind that* $V_{r+1:n_i}$ *is composed of unit-*$\mathscr{L}^2$*-norm column vectors, we consider negligible all entries of* $V_{r+1:n_i}$ *whose square is below a threshold* $\rho^2$ *significantly smaller than* 1*. As we shall see, in several cases, this allows us to clarify the biological interpretation of the estimation results. Further study of the kernel-generating matrix* $V_{r+1:n_i}$ *would yield a theoretically more sophisticated criterion, but we will not pursue this discussion here.*

### *4.3. The overall procedure for identification analysis and model reduction*

Based on the discussion of the previous sections, here we summarize the procedure for obtaining a practically identifiable approximate kinetic model from noisy and incomplete datasets. The procedure is also summarized in Fig. 5. Given noisy steady-state metabolite data $\ln \tilde{x}^1, \ldots, \ln \tilde{x}^q$:

- Compute the data matrix $\tilde{Y}$;
- In case of missing entries, complete the matrix by a method of choice (multiple imputation, minimum rank completion etc.);
- For every reaction $i = 1, \ldots, m$:
  1. Extract from $\tilde{Y}$ the data submatrix $\tilde{Y}_{C(i)}$;
  2. Compute the SVD of the empirical data covariance matrix,
  $$\tilde{Y}_{C(i)}^T \tilde{Y}_{C(i)}/q = V \,\mathrm{diag}(\tilde{s}_1^2, \ldots, \tilde{s}_{n_i}^2) V^T;$$
  3. Compute the effective data rank $r = \max\{\ell : \ \tilde{s}_\ell^2 - \nu^2 \geq \nu^2\}$;
  4. Compute $\hat{Y}_i$, the data matrix obtained by discarding the $n_i - r$ smallest components, as $\hat{Y}_i = Y_{C(i)} \cdot \left[V_{1:r} \quad 0_{n_i \times (n_i - r)}\right] V^T$, $0_{n_i \times (n_i - r)}$ being the $n_i \times (n_i - r)$ null matrix;
  5. Return the reduced model $W_i = \breve{Y}_i \cdot \breve{B}_i + \varepsilon_i$, with $\breve{Y}_i = \hat{Y}_i V_{1:r}$.

## 5. Applications of the model identifiability and reduction approach

*5.1. Application to a network with simulated data*

In order to evaluate performance of our identifiability and model reduction procedure, we now discuss its application to a more realistic simulated network originally presented in (Visser and Heijnen, 2003) and depicted in Fig. 6(a). The network involves $n_x = 8$ internal and $n_u = 3$ external metabolites, participating in a total of $m = 8$ reactions. We developed a linlog model of the network based on the state and input vectors $x$ and $u$ whose entries are listed in Fig. 6(b). The parameter matrices of the model include 33 nonzero entries and are given by

$$a = \begin{bmatrix} -31.4 & 4.41 & 0.13 & 0.31 & 0.31 & 0.13 & -0.42 & 0.97 \end{bmatrix}^T,$$

$$B^x = \begin{bmatrix} -2.470 & -17.40 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.061 & -0.219 & 0 & 0 & 0 & 0.351 & 0 & -1.040 \\ 0 & 0.083 & -0.015 & 0 & 0 & 0 & 0 & -0.029 \\ 0 & 0 & 0.027 & -0.003 & 0 & -0.001 & 0 & 0.086 \\ 0 & 0 & 0 & 0.848 & 0 & 0 & 0 & 0 \\ 0 & 0.093 & 0 & 0 & 0 & 0 & -0.004 & -0.017 \\ 0 & 0 & 0 & 0 & -0.486 & -0.039 & 0.090 & 0.099 \\ 0 & 0 & 0 & 0 & 2.160 & 0 & 0 & 0 \end{bmatrix}, \quad B^u = \begin{bmatrix} 3.880 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -0.713 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1.560 \end{bmatrix}.$$

The values of $B^u$ were taken without changes from (Visser and Heijnen, 2003), while the values of $a$ and $B^x$ were adapted from the same paper. The stoichiometry matrix $N$, given in Eq. (27) below, is fixed by the ordering of the input and state vector entries and the scheme in Fig. 6(a). The row rank of this matrix is equal to 6, corresponding to 2 mass conservation constraints (see also Eq. (28) below). Following the analysis of Reder (1988), it is possible to factor the matrix into a link matrix $L$ expressing dependencies between concentrations and a reduced-order full-row rank matrix $\breve{N}$ corresponding to stoichiometries of independent

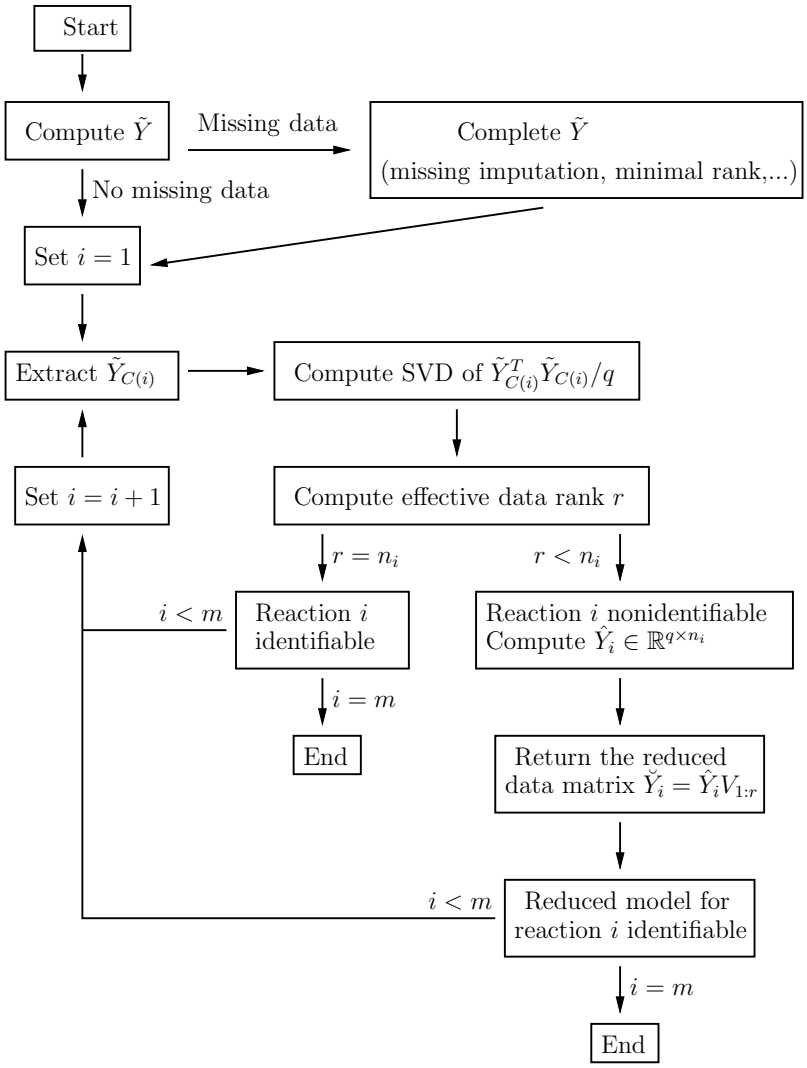
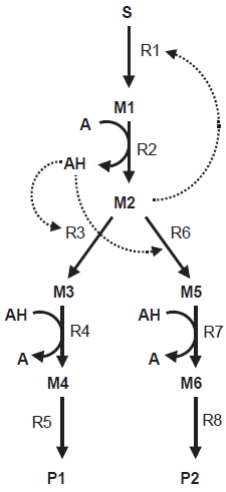



Figure 5: Overall procedure for identifiability analysis and model reduction.




(a)

| Internal metabolites | | External metabolites | |
|---|---|---|---|
| **Variable** | **Metabolite** | **Input** | **Metabolite** |
| $x_1$ | M1 | $u_1$ | S |
| $x_2$ | M2 | $u_2$ | P1 |
| $x_3$ | M3 | $u_3$ | P2 |
| $x_4$ | M4 | | |
| $x_5$ | M6 | | |
| $x_6$ | A | | |
| $x_7$ | M5 | | |
| $x_8$ | AH | | |

(b)

Figure 6: (a) A branched metabolic pathway with feedback from (Visser and Heijnen, 2003). All reactions are chemically reversible, the arrows represent the positive flux directions. Dashed lines represent allosteric interactions. (b) Model variables for internal and external metabolites.

metabolites. Factorization is non-unique. In our case, one such factorization gives

$$N = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & -1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & -1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & -1 & -1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{27}$$

$$= L \cdot \breve{N}.$$

With this factorization, the entries of $x_{1:6}$ (M1, M2, M3, M4, M6 and A) are treated as independent quantities, and determine the values of $x_7$ (M5) and $x_8$ (AH) via conservation of mass. That is, for some fixed constants $T_1, T_2 \in \mathbb{R}_{>0}$,

$$\begin{cases} \dot{x}_{1:6} = \breve{N}\,\mathrm{diag}(e)(a + B^x \cdot \ln x + B^u \cdot \ln u), \\ T_1 = x_2 + x_3 + x_6 + x_7, \\ T_2 = x_6 + x_8. \end{cases} \tag{28}$$

Notice that this reformulation allows one to compute the steady-state values of all system variables by setting the differential part to zero. In our case, the method is used to compute steady-state data as a function of enzyme and external metabolite concentrations.

To assess identification performance, we considered the scenario of (Visser and Heijnen, 2003), where the external metabolite concentrations are fixed to $u = \begin{bmatrix} 1 & 0.1 & 0.2 \end{bmatrix}^T$, $T_1 = 0.3$ and $T_2 = 0.1$. Since $u$ is fixed, the parameter matrix $B^u$ is obviously nonidentifiable, and the contributions of $a$ and $B^u \ln u$ are indistinguishable. To circumvent this issue, we define

| Reaction number | Number of parameters | Average effective rank | |
|---|---|---|---|
| | | Def. 3 | Def. 4 |
| R1 | 2 | 1 | 1.98 |
| R2 | 4 | 2 | 3.51 |
| R3 | 3 | 1.05 | 3 |
| R4 | 4 | 2.11 | 4 |
| R5 | 1 | 0.21 | 1 |
| R6 | 3 | 1.3 | 2.99 |
| R7 | 4 | 2.02 | 3.96 |
| R8 | 1 | 0.03 | 1 |

Table 1: Average effective rank computed for each reaction and with different definitions of $r$ over 100 datasets of the model of Fig. 6. The criterion of Def. 4 was computed choosing $\theta = 0.99$.

the lumped constant term $a' = a + B^u \ln u$ so as to obtain the modified linlog reaction rate model $v = \mathrm{diag}(e) \cdot (a' + B^x \ln x)$, and study the identifiability and reduction of the latter in terms of $a'$ and $B^x$.

Identifiability analysis and model reduction are performed in accordance with Section 4.3 on $R = 100$ randomly generated datasets $\tilde{Y}$ and $\tilde{W}$ with realistic statistical properties. Each dataset shares the same steady-state values $Y$ and $W$ computed for $q = 30$ different values of enzyme concentrations, generated once as in Example 5, and each differs in the randomly generated 40% noise corrupting the measurements. We generated the results from the application of both Def. 3 and Def. 4, with $\theta = 0.99$. The results are depicted in Fig. 7 and are also reported in Table 1.

First, we notice that for every reaction, the effective rank computed with the criterion of Def. 4 is higher than the one computed with the criterion of Def. 3. Thus, the latter criterion gives more conservative results, in the sense that, on average, fewer reactions are deemed identifiable. Except for reaction 2, application of Def. 4 returns the full size of the reaction for at least 96 out of 100 datasets. That is, in this case, the ability of the criterion to detect dependencies among data is very limited. This can be explained by the presence of a significant amount of noise on metabolite data, which tends to decorrelate the observations.

The criterion based on Def. 3 detects dependencies among the data for all reactions. Indeed, the effective rank determined by this criterion is consistently smaller and differs from the results of Def. 4 by an average of about 1 for reactions 1, 5 and 8, and of about 2 for the other reactions. This can be attributed to the compensation of noise in the computation of the singular value estimates $\hat{s}_\ell^2$ in (25), which relies upon and exploits the knowledge of the noise level $\nu$. This is apparent in Fig. 7, where the blue dots representing the singular value estimates drawn from each of the 100 datasets $\tilde{Y}$ are correctly concentrated around the true singular values from $Y$. Fig. 7 also clarifies the non-integer average rank values of Table 1 (notably for reactions 4, 5 and 6). This comes from the fact that the singular values lying close to the significance cutoff value $\nu^2$ are estimated above or below this threshold depending on the simulation run. For instance, the estimates of the second singular value of reaction 6 were smaller than $\nu^2$ in 70 of the 100 runs.

Finally, the results for reactions 5 and 8 reveal a fundamental difference between Def.s 3 and 4. The former method leads to the conclusion that no modeling is possible (the effective

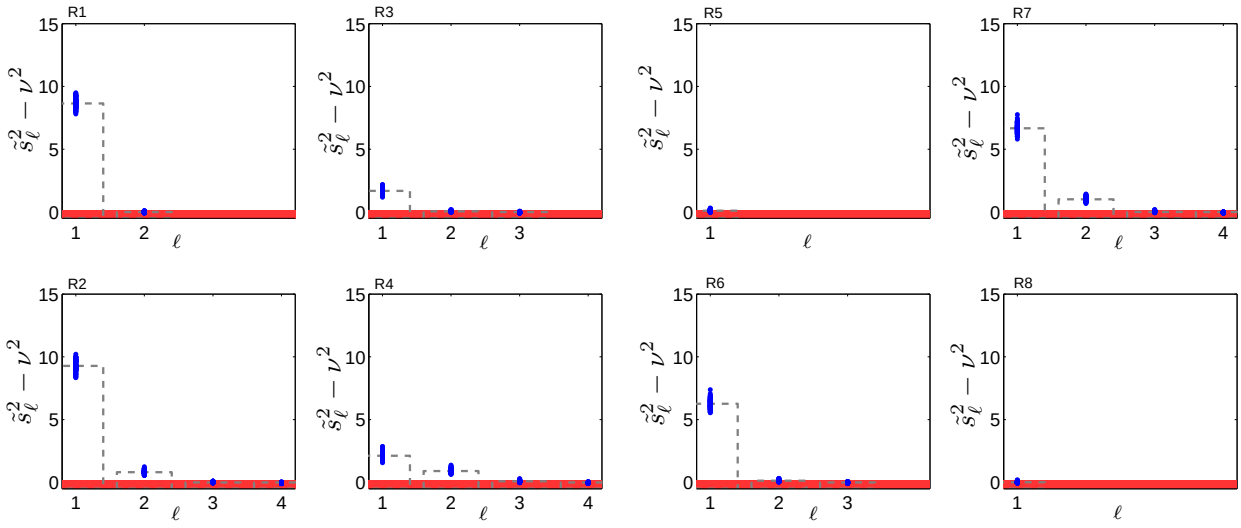



Figure 7: Singular value estimates $\hat{s}_\ell^2 = \tilde{s}_\ell^2 - \nu^2$ computed from 100 noisy datasets $\tilde{Y}$ of size $q = 30$ (blue dots) of the model of Fig. 6. The dashed bars correspond to the singular values $s_\ell^2/q$ from the noiseless dataset $Y$ of size $q = 30$. In red is the area below $\nu^2$; all singular values in this area are considered negligible.

rank is estimated to be zero in most runs) since the corresponding metabolite data are dominated by noise, whereas Def. 4 provides effective rank estimates that are by construction lower-bounded by one.

From the results of this section, it is clear that identifiability analysis and model reduction in the presence of noise should be performed on the basis of the effective rank of Def. 3, which outperforms Def. 4 and returns consistent results (Table 1 and Fig. 7). The actual application of the method and the effects of using Def. 3 in place of Def. 4 on a real dataset are discussed in the next section.

### *5.2. Application to central carbon metabolism of E.coli*

As a second example, we illustrate the application of our method to a complex network of biochemical reactions involved in carbon assimilation in the enterobacterium *Escherichia coli*. The network we consider gathers enzymes, metabolites and reactions that make up the bulk of central metabolism, including glycolysis, the pentose-phosphate pathway, the tricarboxylic acid cycle and anaplerotic reactions such as glyoxylate shunt and PEP-carboxylase (Fig. 8). The network has been studied for a long time from different perspectives, which makes it an ideal model system for our purpose. The structure of the *E. coli* carbon metabolism network is known in a rather precise way, its dynamics have been modeled by means of a variety of formalisms ((Bettenbrock et al., 2006; Kotte et al., 2010) and references therein), and recently a high-throughput dataset containing the required information for addressing Problem 3 has been published (Ishii et al., 2007).

We now investigate which reactions are identifiable following the criteria of Section 4, given the available experimental data and a linlog model of the network. From a methodological point of view, we are interested in analyzing the differences between the results

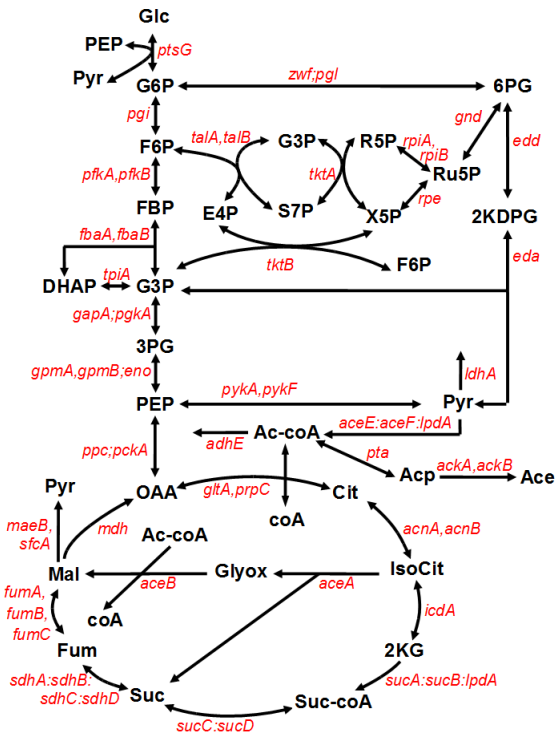



Figure 8: Scheme of *Escherichia coli* central carbon metabolism (Berthoumieux et al., 2011). The map shows metabolites (bold fonts) and genes (italic). Abbreviations of metabolites are glucose (Glc), glucose 6-phosphate (G6P), fructose 6-phosphate (F6P), fructose 1-6-biphosphate (FBP), dihydroxyacetone phosphate (DHAP), glyceraldehyde 3-phosphate (G3P), 3-phosphoglycerate (3PG), phosphoenolpyruvate (PEP), pyruvate (Pyr), 6-phosphogluconate (6PG), 2-keto-3-deoxy-6-phospho-gluconate (2KDPG), ribulose 5-phosphate (Ru5P), ribose 5-phosphate (R5P), xylulose 5-phosphate (X5P), sedoheptulose 7-phosphate (S7P), erythrose 4-phosphate (E4P), oxaloacetate (OAA), citrate (Cit), isocitrate (IsoCit), 2-keto-glutarate (2KG), succinate-CoA (Suc-coA), succinate (Suc), fumarate (Fum), malate (Mal), glyoxylate (Glyox), acetyl-CoA (Ac-coA), acetylphosphate (Acp) and acetate (Ace). Cofactors impacting the reactions are not shown: adenosine triphosphate (ATP), adenosine diphosphate (ADP), nicotinamide adenine dinucleotide phosphate (NADP) and its reduced form (NADPH), nicotinamide adenine dinucleotide (NAD) and its reduced form (NADH) and flavin adenine dinucleotide (FAD). The gene names are separated by a comma in the case of isoenzymes, by a colon for enzyme complexes, and by a semicolon when the enzymes catalyze reactions that have been lumped together in the model.

obtained with Def. 3 and those obtained with the method of Def. 4 used in (Berthoumieux et al., 2011). From a biological point of view, we wish to understand how much information is actually contained in a state-of-the-art dataset for the purpose of parameter estimation.

The dataset used for identification of the network in Fig. 8 was obtained by experiments with 24 single-gene deletions that were grown at a fixed dilution rate of $0.2\text{h}^{-1}$ in a glucose-limited chemostat, and with wild-type cells at 5 different dilution rates (Ishii et al., 2007). The authors collected data using multiple high-throughput techniques, in particular DNA microarray analysis and two-dimensional differential gel electrophoresis (2D-DIGE) for genes and proteins, capillary electrophoresis time-of-flight mass spectrometry (CE-TOFMS) for metabolites, and metabolic flux analysis. They thus obtained a steady-state dataset consisting of metabolite concentrations, mRNA and protein concentrations for the enzymes, and metabolic fluxes under 29 different experimental conditions. Therefore this dataset contains the information for setting up a parameter estimation problem as defined in Section 2.

We carried out the identifiability analysis for the linlog model developed in (Berthoumieux et al., 2011). This model is a translation of the reaction scheme of Fig. 8 into linlog rate equations (2). When certain metabolites could not be measured, preventing their inclusion in the model, we lumped together the reactions in which they are involved. In addition to the above model simplification, imposed by the available data, we added a phenomenological reaction to model biomass production. The resulting model has $n_x = 16$ internal metabolites, $n_u = 7$ external metabolites and measured cofactors, and $m = 31$ reactions (Berthoumieux et al., 2011).

A complication for determining the identifiability of reactions and finding a suitable model reduction is that the dataset contains a large amount of missing data. In particular, certain metabolites could not be measured in up to 80% of the experimental conditions (28% on average for all metabolites). Following Remark 2 of Section 4, we therefore completed the dataset by means of multiple imputation, generating 100 datasets to allow the computation of statistics and test the robustness of the results.

Table 2 summarizes the results of applying the reduction method of Section 4.3 to the model and the data. For each of the reactions in the model, we computed the average of the effective rank of the 100 completed datasets. The effective rank for the individual datasets was usually found to be the same (in at least 82 of the 100 datasets), which explains that the computed average values are close to integers. Remarkably, out of the 31 reactions in the model, only 4 were found to be fully identifiable: reactions 4, 5, 14, and 31. The first three reactions involve two metabolites: fructose 1-6-biphosphate (FBP) and dihydroxyacetone phosphate (DHAP) (reaction 4), DHAP and 3-phosphoglycerate (3PG) (reaction 5) and ribose 5-phosphate (R5P) and sedoheptulose 7-phosphate (S7P) (reaction 14). The identifiability of these reactions means that the method did not detect any dependencies between these pairs of metabolite concentrations. Reaction 31 involves a single metabolic variable acetyl-coA/coA (Ac-coA/coA). Among the remaining 27 nonidentifiable reactions, two reactions are known to operate close to equilibrium (reactions 2 and 7: Pgi and Gpm:Eno, respectively), which provides a robust rationale for their nonidentifiability whatever the physiological conditions that are explored in the dataset (cf. Eq. (22)). In the other nonidentifiable cases the effective rank is reduced by 1 (for 16 reactions), 2 (6 reactions), 3 (2 reactions), and 6 (1 reaction). The latter case concerns the growth-rate reaction, which has 11 variables. A striking observation on Tables 2 and 3 is therefore the

large number of nonidentifiable reactions and parameters.

| Reaction | Enzyme | Average effective rank | Full dimension | Reaction | Enzyme | Average effective rank | Full dimension |
|---|---|---|---|---|---|---|---|
| 1 | PtsG | 3 | 4 | 17 | GltA,PrpC | 2.97 | 4 |
| 2 | Pgi | 1 | 2 | 18 | AcnA,AcnB | 1 | 2 |
| 3 | PfkA,PfkB | 2.85 | 4 | 19 | IcdA | 1 | 3 |
| 4 | FbaA,FbaB | 2 | 2 | 20 | SucA:SucB:LpdA;SucC:SucD | 1 | 3 |
| 5 | TpiA | 2 | 2 | 21 | SdhA:SdhB:SdhC:SdhD | 1 | 3 |
| 6 | GapA;Pgk | 2.99 | 4 | 22 | FumA,FumB,FumC | 1 | 2 |
| 7 | GpmA,GpmB;Eno | 1 | 2 | 23 | Mdh | 2.97 | 4 |
| 8 | PykA,PykF | 2 | 4 | 24 | Ppc;PckA | 3 | 5 |
| 9 | AceE:AceF:LpdA | 1.99 | 3 | 25 | MaeB,SfcA | 2 | 5 |
| 10 | Zwf;Pgl | 1.98 | 3 | 26 | AceA;AceB | 1 | 3 |
| 11 | Gnd | 2 | 3 | 27 | $\mu$ | 4.94 | 11 |
| 12 | Rpe | 1 | 2 | 28 | Edd;Eda | 1 | 2 |
| 13 | RpiA,RpiB | 1.99 | 3 | 29 | Pta;AckA,AckB | 3 | 6 |
| 14 | TktA | 1.82 | 2 | 30 | LdhA | 1 | 2 |
| 15 | TalA,TalB | 1 | 2 | 31 | AdhE | 1 | 1 |
| 16 | TktB | 1.01 | 2 | | | | |

Table 2: Average effective rank computed for the reactions in the linlog model of *E. coli* central carbon metabolism, using the data of Ishii et al. (2007). SVD has been applied to $Y_{C(i)}$ for each reaction and singular values were discarded based on Def. 3. Identifiable reactions are shown in green. Reaction 27, labeled $\mu$, is a phenomenological reaction for biomass production.

In order to isolate identifiable parameters in nonidentifiable reactions, Proposition 3 proposes a criterion that has been relaxed in Remark 3 so as to make it applicable to noisy data. The approach is based on the choice of a threshold $\rho^2$ for neglecting components of the kernel-generating marix $V_{r+1:n_i}$ extracted from the data. In what follows, to set a ground for discussion, we set $\rho$ equal to 0.15. We verified that changes of this threshold within the range $\rho \in (0.1,\ 0.2)$ do not significantly alter the results as reported in Table 3. In this table, parameters that were diagnosed as being identifiable in more than 50% of the completed datasets are highlighted in green. From the 27 nonidentifiable reactions, no individual parameter could be unambiguously extracted in 8 cases (reactions 2, 7, 8, 12, 15, 16, 22 and 24). Of the 72 parameters in the remaining 19 reactions, 30 are identifiable in more than half of the datasets. In particular, we observe that all parameters associated to glucose (Glc), DHAP, Ac-coA/coA, 6-phosphogluconate (6PG), R5P, FAD and acetate (Ace) are identifiable in the sense that no significant dependencies with other metabolites could be detected in the experimental conditions of Ishii et al. (2007).

The results shown in Table 3 are different from those obtained in our earlier work (Berthoumieux et al., 2011), where we used a method based on Def. 4 with $\theta = 0.99$ instead of Def. 3. Indeed we previously found many more reactions to be identifiable (24 out of 31) although for most cases parameter estimates turned out to be unreliable because of large confidence intervals. Reactions 2 and 7 (Pgi and Gpm:Eno), that were classified as identifiable in Berthoumieux et al. (2011), are found here to be nonidentifiable in agreement with the fact that they operate close to equilibrium. Therefore, the results of the identifiability analysis in Berthoumieux et al. (2011) appear to be overly optimistic, i.e. overestimating the number of identifiable reactions because measurement errors on metabolite concentrations are not taken into account. Notwithstanding, the results of both analyses are consistent in the sense that all 7 reactions detected as nonidentifiable by means of Def. 4 also remain nonidentifiable according to Def. 3.

Table 3: Parameter matrix $[B^x\ B^u]$ and results of identifiability analysis and model reduction on the data of (Ishii et al., 2007) for the linlog model of *E. coli* central carbon metabolism. SVD has been applied to $Y_{C(i)}$ for each reaction and singular values were discarded based on Def. 3. Nonidentifiable parameters are shown in grey and identifiable ones, as well as identifiable reactions, in green. The percentages of cases for which the parameters were found identifiable are given. To avoid complications deriving from the presence of conserved moieties, some of the metabolites are modeled as ratios of metabolite concentrations, *e.g.* ATP/ADP. Reaction 27, labeled $\mu$, is a phenomenological reaction for biomass production. The last row indicates the percentage of missing data per metabolite. Abbreviations are as in Fig. 8.

| | **Enzyme** \ **Metabolite** | Glc | PEP | G6P | Pyr | F6P | FBP | DHAP | 3PG | Ac-coA / coA | 6PG | Ru5P | R5P | S7P | 2KG | Suc | Fum | Mal | ATP / ADP | Cit | NADPH / NADP | NADH / NAD | FAD | Ace |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | PtsG | 100% | 100% | 31% | 0 % | | | | | | | | | | | | | | | | | | | |
| 2 | Pgi | | | 0 % | | 0 % | | | | | | | | | | | | | | | | | | |
| 3 | PfkA,PfkB | | 70% | | | 2% | 63% | | | | | | | | | | | | 8% | | | | | |
| 4 | FbaA,FbaB | | | | | | 100% | 100% | | | | | | | | | | | | | | | | |
| 5 | TpiA | | | | | | | 100% | 100% | | | | | | | | | | | | | | | |
| 6 | GapA;Pgk | | | | | | 84% | | 6% | | | | | | | | | | 0% | | | 51% | | |
| 7 | GpmA,GpmB;Eno | | 0% | | | | | | 0% | | | | | | | | | | | | | | | |
| 8 | PykA,PykF | | 5% | | 0% | | 22% | | | | | | | | | | | | 0% | | | | | |
| 9 | AceE:AceF:LpdA | | | | 1% | | | | | 99% | | | | | | | | | | | | 94% | | |
| 10 | Zwf;Pgl | | | 70% | | | | | | | 100% | | | | | | | | | | 2% | | | |
| 11 | Gnd | | | | | | | | | | 100% | 100% | | | | | | | | | 0% | | | |
| 12 | Rpe | | | | | | | | | | | 0% | | 0% | | | | | | | | | | |
| 13 | RpiA,RpiB | | | 0% | | | | | | | | 0% | 66% | | | | | | | | | | | |
| 14 | TktA | | | | | | | | | | | | 82% | 82% | | | | | | | | | | |
| 15 | TalA,TalB | | | | | 0% | | | | | | | | 0% | | | | | | | | | | |
| 16 | TktB | | | | | 1% | | | | | | 1% | | | | | | | | | | | | |
| 17 | GltA,PrpC | | | | | | | | | 97% | | | | | 2% | | | | | 98% | | 97% | | |
| 18 | AcnA,AcnB | | | | | | | | | | | | | | 0% | | | | | 100% | | | | |
| 19 | IcdA | | | | | | | | | 100% | | | | | 0% | | | | | | 0% | | | |
| 20 | SucA:SucB:LpdA;SucC:SucD | | | | | | | | | | | | | | 0% | 0% | | | | | | 90% | | |
| 21 | SdhA:SdhB:SdhC:SdhD | | | | | | | | | | | | | | | 0% | 0% | | | | | | 62% | |
| 22 | FumA,FumB,FumC | | | | | | | | | | | | | | | | 0% | 0% | | | | | | |
| 23 | Mdh | | 87% | | | | | | | | | | | | | | | 0% | | 75% | | 35% | | |
| 24 | Ppc;PckA | | 19% | | | | 5% | | | | | | | | | | | 0% | 0% | 32% | | | | |
| 25 | MaeB,SfcA | | | | 0% | | | | | 99% | | | | | | | | 0% | | | 0% | 30% | | |
| 26 | AceA;AceB | | | | | | | | | 100% | | | | | | 0% | | 0% | | | | | | |
| 27 | $\mu$ | | 0% | 0% | 0% | 0% | | | 0% | 68% | | | 81% | | 0% | | | | 0% | | 0% | 27% | | |
| 28 | Edd;Eda | | | | 0% | | | | | | 100% | | | | | | | | | | | | | |
| 29 | Pta;AckA,AckB | | | | 0% | | | | | 70% | | | | | | | | | 0% | | 0% | 37% | | 100% |
| 30 | LdhA | | | | 0% | | | | | | | | | | | | | | | | | 95% | | |
| 31 | AdhE | | | | | | | | | 100% | | | | | | | | | | | | | | |
| | % Missing data | 3 | 17 | 0 | 48 | 7 | 34 | 59 | 10 | 3 | 72 | 3 | 38 | 3 | 59 | 3 | 14 | 14 | 0 | 62 | 79 | 79 | 17 | 17 |

## 6. Discussion

A major, but often overlooked problem in the identification of metabolic network models is the identifiability of the parameters, and hence of the model. Informally speaking, the identifiability of a model (parameter) consists in the possibility to unambiguously reconstruct the model (parameter) from the observed behavior of the system. Identifiability problems may reside in the very structure of the model, notably the occurrence of implicit dependencies between parameters. These dependencies might be due to an inappropriate model formulation, constraints on the kind of experimental perturbations that can be realized, and unobserved variables. In addition, practical identifiability problems may arise from limitations on the quality and quantity of available data, in particular the fact that experimental data in biology are frequently noisy, sparse and incomplete.

We have studied identifiability issues in the context of approximate kinetic modeling formalisms, notably linear-logarithmic (linlog) models. On the theoretical side, following the classical systems identification literature (Ljung, 1999; Walter and Pronzato, 1997), we have first precisely defined the notions of structural (a priori) identifiability (Def. 1) and practical (a posteriori) identifiability (Def. 2). The latter notion is obviously related to the former, in the sense that structural nonidentifiability entails practical nonidentifiability. However, Proposition 2 goes beyond by saying that identifiability in the theoretical sense may also imply identifiability in the practical sense, provided that the uncertainty on the parameters, as determined by the available dataset, remains within the desired accuracy bounds. Notice that practical identifiability is thus not an absolute notion, but rather conditional on the data properties and the required model precision.

A second methodological contribution of this paper is the development of theoretically sound and practically applicable methods for the detection of identifiability problems and the transformation of a nonidentifiable model to a reduced identifiable model. In particular, we have formulated criteria based on the SVD of the matrix of log-transformed and centered measurements of metabolite concentrations. These criteria define the effective rank of the data matrix, corresponding to the number of parameters that can be safely distinguished from the output. The criterion privileged in this paper (Def. 3), contrary to a criterion that we proposed in earlier work (Def. 4), takes into account the estimated variance of the noise. The flow chart in Fig. 5 gives a step-by-step procedure for identifiability analysis and model reduction.

The identifiability of models of biological systems is a topic that has been much studied in mathematical biology, and that has received renewed attention in the context of systems biology (see Chappell et al. (1990); Chis et al. (2011b); Cobelli and DiStefano (1980); Raue et al. (2011) for reviews). Systems of biochemical reactions, which have the general form of Eq. (3) at steady-state, have a number of peculiarities for identifiability analysis. When reaction rates, enzyme concentrations and metabolite concentrations are measured, the identification problem can be decomposed into subproblems for each of the individual reactions. Determining the identifiability of a model then reduces to checking the identifiability of the reactions. If in addition the reaction rates are expressed in terms of linlog or other pseudo-linear equations, identification becomes a linear or orthogonal regression problem, depending on whether noise on the metabolite concentrations is taken into account or not (Problem 2 and Problem 3, respectively). Identifiability analysis then amounts to checking for linear de-

pendencies in a transformed data matrix, which can be done using standard techniques from linear algebra and statistics. Related ideas starting from this general approach can be found in other recent work (Berthoumieux et al., 2011; Srinath and Gunawan, 2010; Nikerel et al., 2009). However, our development in the current paper is different in fundamental ways, such as the very definition of structural and practical identifiability, and the application of SVD to detect and resolve identifiability problems.

Although linlog models have been central in this paper, the results directly carry over to the other approximate formalisms mentioned in Section 2. In addition, they also bear on more general classes of nonlinear models of metabolic networks. Indeed the parameters in the mean-removed linlog model of Eq. (6) are proportional to elasticity coefficients that describe the sensitivities of reaction rates to changes in metabolite concentrations. If a reaction in a linlog model is nonidentifiable, this means that elasticity coefficients are not identifiable, therefore any other class of kinetic models is liable to encounter similar identifiability issues.

The approach for determining the identifiability of linlog models proposed in this paper has been tested on a network with simulated data and applied to a high-throughput data set for central carbon metabolism in *E. coli*. The use of simulated data has made it possible to demonstrate that, for typical sizes of the dataset and realistic noise levels, our approach is able to correctly identify the principal components of the parameter vector (Fig. 7). Surprisingly, the determination of the effective rank of the datasets for the different reactions in the *E. coli* metabolic network shows that only a small fraction of the reactions (4 out of 31) is fully identifiable from the data of Ishii et al. (2007). In addition, only 37 out of the total of 100 model parameters are individually identifiable. We note that these results are different from those reported in (Berthoumieux et al., 2011), due to the fact that here we take into account the metabolite concentration measurement error to decide whether a parameter associated with a principal component is negligible. The low numbers of identifiable reactions and parameters agree with those obtained with power-law models on other state-of-the-art datasets (Srinath and Gunawan, 2010). This further demonstrates the importance of a preliminary identifiability analysis when estimating parameters in metabolic network models.

The rank analysis carried out to determine the identifiability of a reaction also shows how the model can be reduced if the data does not allow the parameters of the model to be unambiguously determined. This reduction step yields minimal models in the form of Eq. (24), that have the advantage of being adapted to the informativeness of the dataset. A disadvantage of this approach, however, is that the parameters of the reduced model may be difficult to interpret from a biological point of view, as they do not generally determine the original parameters in a unique way, but rather define a linear subspace of the parameter space. Nevertheless, in some cases it may still be possible to identify some parameters of the original model (Proposition 3 and Remark 3). This criterion was shown to be useful in practice, as it allowed to uniquely identify 30 out of 93 parameters in the nonidentifiable reactions of the *E.coli* metabolic network model (Table 3).

If the parameters of the original, non-reduced model cannot be uniquely determined from the data, then additional experiments may be necessary. Generally speaking, experimental conditions that explore the range of possible behaviors of the network as much as possible improve identifiability. Given that experiments are usually carried out at steady-state, especially for metabolic flux measurements, the available datasets have a sampling bias that

may complicate parameter estimation. In particular, metabolic systems almost invariably contain highly evolved regulatory loops that may homeostatically buffer the concentrations of some metabolic pools (Bennett et al., 2009; Ishii et al., 2007). As a consequence, a range of different growth conditions and genetic backgrounds may lead to little variation in steady-state concentrations. Moreover, some metabolic pools may change in a correlated fashion against experimental perturbations, which also hinders identifiability. Finally reactions that normally operate close to equilibrium are intrinsically nonidentifiable, unless the experimentalist achieves to measure their rates further from equilibrium. The growing availability of time-series data (e.g., Voit et al. (2006b); Hardiman et al. (2007)), although more demanding from an experimental point of view, promises to relieve this problem.

**Acknowledgments**

This work was supported by the Agence Nationale de la Recherche under project MetaGenoReg (ANR-06-BYOS-0003).

# Appendices

## Appendix A. Notation and terminology

$\mathbb{R}$, $\mathbb{R}_{>0}$, $\mathbb{Z}$ and $\mathbb{N}$ denote the sets of real, positive real, integer and positive natural numbers, respectively. For an index $n \in \mathbb{N}$, $\mathbb{R}^n$ and $\mathbb{R}^n_{>0}$ denote the $n$ dimensional versions of $\mathbb{R}$ and $\mathbb{R}_{>0}$. $I$ denotes an identity matrix of dimension fixed by the context.
Let $M$ be any matrix. For two indices $i$ and $j$ and a vector of indices $C$ compatible with the dimensions of $M$, $M_i$ denotes the $i$th column of $M$, $M_C$ denotes the submatrix of $M$ formed by the columns of $M$ with indices $C$, $M_{ji}$ denotes the element of $M$ in row $j$ and column $i$, and $M_{j,C}$ denotes the row vector formed by the elements of $M$ in row $j$ and columns indexed by $C$. $[M]_{C,C}$ denotes the minor formed by the rows and columns indexed by $C$. When convenient, notation $M_{i:i'}$ with $i' \geq i$ is used instead of $M_C$ with $C = \left[i, i+1, \ldots, i'\right]$. For vectors, a subscript $i$ refers to the $i$th element of the vector.
For a square matrix $\Sigma$, $\Sigma > 0$ (resp. $\Sigma \geq 0$) means that $\Sigma$ is positive definite (resp. semidefinite). For a vector $\mu$ of suitable dimension, $\varepsilon \sim \mathcal{N}(\mu, \Sigma)$ means that $\varepsilon$ is a Gaussian random vector with mean $\mu$ and covariance matrix $\Sigma$. For a scalar quantity observed in Gaussian noise with standard deviation $\nu$, a noise level of $N\%$, with $N \in \mathbb{R}_{\geq 0}$, will stand for a value of $\nu$ such that $\nu$ is equal to $N\%$ of the value of that quantity (such that $\simeq 99\%$ of the noise outcomes fall within $\pm 3 \cdot N\%$ of the observed quantity).
For two vectors $v$ and $e$ of equal size, both $v/e$ and $\frac{v}{e}$ indicate the vector obtained by element-wise division. Given a vector sequence $v^1, \ldots, v^q$, $\bar{v}$ is the mean $(1/q)\sum_{k=1}^{q} v^k$. For vectors and sets, $|\cdot|$ denotes vector dimension and set cardinality, respectively. For a function $f : A \to B$, $f|_D$ indicates its restriction on $D \subseteq A$.

## Appendix B. Mathematical proofs

**Proposition 1.** *A reaction $i$ of $\mathscr{M}_p$ is structurally identifiable at $p^*$ if and only if there exists $D = \{(e^1, u^1), \cdots, (e^q, u^q)\} \subseteq E \times U$ such that the solution of the equation $W_i^* = Y^* B_i^*$,*

*with*

$$W_i^* = \left[ \left( \frac{J_*^1}{e^1} - \overline{\left(\frac{J_*}{e}\right)} \right)_i \quad \cdots \quad \left( \frac{J_*^q}{e^q} - \overline{\left(\frac{J_*}{e}\right)} \right)_i \right]^T,$$
$$Y^* = \begin{bmatrix} \ln x_*^1 - \overline{\ln x_*} & \cdots & \ln x_*^q - \overline{\ln x_*} \\ \ln u^1 - \overline{\ln u} & \cdots & \ln u^q - \overline{\ln u} \end{bmatrix}^T,$$

*is unique in the parameters* $B_i^* = \left( [B^{x*} \; B^{u*}]^T \right)_i$.

*Proof.* (If) Assume that, for a given $D \subseteq E \times U$, the solution of $W_i^* = Y^* B_i^*$ is unique. We need to prove that $\big((J_p)_i, x_p\big)|_D = \big((J_{p^*})_i, x_{p^*}\big)|_D$ implies $p_i = p_i^*$. For simplicity, here we drop index $i$ from subscripts.
Given any two parameters $p^* = \begin{bmatrix} a^* & B^{*T} \end{bmatrix}^T$ and $p = \begin{bmatrix} a & B^T \end{bmatrix}^T$, for which $\mathscr{M}_p : (e,u) \mapsto (J_p, x_p)$ and $\mathscr{M}_{p^*} : (e,u) \mapsto (J_{p^*}, x_{p^*})$, it holds by construction that $W = YB$ and $W^* = YB^*$. If $(J_p, x_p)|_D = (J_{p^*}, x_{p^*})|_D$, then it also holds that $Y = Y^*$ and $W = W^*$, therefore we can write $W^* = Y^* B$. Because the solution in $B$ of the latter is unique and one solution is $B^*$, it follows that $B = B^*$. To conclude that $p = p^*$, we are left with showing that $a = a^*$. This follows from

$$a^* = \overline{\left(\frac{J_*}{e}\right)} - \left[\overline{\frac{\ln x_*}{\ln u}}\right]^T \cdot B^* = \overline{\left(\frac{J}{e}\right)} - \left[\overline{\frac{\ln x}{\ln u}}\right]^T \cdot B = a.$$

(Only if) Here the hypothesis is that, for a given $D \subseteq E \times U$, $\big((J_p)_i, x_p\big)|_D = \big((J_{p^*})_i, x_{p^*}\big)|_D$ implies $p_i = p_i^*$, and we need to show that the solution in $B_i$ of $W_i^* = Y^* B_i$ is unique. For simplicity, we will again drop $i$ from the subscripts.
For the sake of contradiction, assume that $W^* = Y^* B$ admits distinct solutions. Since $B^*$ is a solution, all solutions are of the form $B = B^* + z$, with $z$ in the nontrivial kernel of $Y^*$. For any such $z$ we can write $Y^* B^* = Y^*(B^* + z)$, i.e., $\forall (e,u) \in D$,

$$\left[(\ln x_* - \overline{\ln x_*})^T \quad (\ln u_* - \overline{\ln u_*})^T\right] B^* = \left[(\ln x_* - \overline{\ln x_*})^T \quad (\ln u_* - \overline{\ln u_*})^T\right] (B^* + z). \quad \text{(B.1)}$$

Let $p^* = \begin{bmatrix} a^* & B^{*T} \end{bmatrix}^T$. For any $(e,u) \in D$, $J_* = J_{p^*}(e,u)$ and $x_* = x_{p^*}(e,u)$ are given by the solution of

$$\begin{aligned} 0 &= N J_{p^*}, \\ J_{p^*} &= \operatorname{diag}(e) \cdot (a^* + \begin{bmatrix} \ln x_*^T & \ln u_*^T \end{bmatrix} B^*) \end{aligned} \quad \text{(B.2)}$$

(which is unique by virtue of Assumption 1). Using (B.1), term $\begin{bmatrix} \ln x_*^T & \ln u_*^T \end{bmatrix} B^*$ can be rewritten as $\begin{bmatrix} \ln x_*^T & \ln u_*^T \end{bmatrix} (B^* + z) - \begin{bmatrix} \overline{\ln x_*^T} & \overline{\ln u_*^T} \end{bmatrix} z$. Replacing this into (B.2) yields

$$\begin{aligned} 0 &= N J_*, \\ J_* &= \operatorname{diag}(e) \cdot \Big( \underbrace{(a^* - \begin{bmatrix} \overline{\ln x_*^T} & \overline{\ln u_*^T} \end{bmatrix} z)}_{\triangleq a} + \begin{bmatrix} \ln x_*^T & \ln u_*^T \end{bmatrix} \underbrace{(B^* + z)}_{=B} \Big), \qquad \forall (e,u) \in D. \end{aligned}$$

From this we see that $p = \begin{bmatrix} a & B^T \end{bmatrix}^T$, with $a$ defined as above, is different from $p^*$ but is such that $\big(J_p(e,u), x_p(e,u)\big) = \big(J_{p^*}(e,u), x_{p^*}(e,u)\big)$ for all $(e,u) \in D$, which contradicts the hypothesis.

□

**Corollary 1.** *A reaction $i$ of $\mathscr{M}_p$ is structurally identifiable at $p^*$ if and only if there exists $D = \{(e^1, u^1), \cdots, (e^q, u^q)\} \subseteq E \times U$ such that $Y^*_{C(i)}$ is full column-rank.*

*Proof.* From Proposition 1, we know that identifiability is equivalent to the uniqueness of the solution in $B_i$ of $W_i^* = Y^* B_i$, i.e. of the solution in $B_{C(i),i}$ of $W_i^* = Y^*_{C(i)} B_{C(i),i}$ (the elements of $B_i$ not included in $B_{C(i),i}$ are set to zero by definition). Uniqueness holds if and only if $\ker(Y^*_{C(i)}) = \{0\}$, i.e. $Y^*_{C(i)}$ is full column-rank or equivalently $\text{rank}(Y^*_{C(i)}) = n_i$. □

**Proposition 2.** *If a reaction $i$ of $\mathscr{M}_p$ is structurally identifiable at $p^*$ in the sense of Def. 1 then, for every $\alpha \in (0,1)$, it is practically identifiable in the sense of Def. 2 with confidence level at least $1-\alpha$ for any uncertainty set $\mathscr{B}_i \supseteq \mathcal{E}_{\widehat{\Sigma}}(\alpha)$, where $\mathcal{E}_{\widehat{\Sigma}}(\alpha)$ denotes the $(1-\alpha)$-confidence ellipsoid of a zero-mean Gaussian distribution with variance $\widehat{\Sigma} = (Y^T_{C(i)} \Sigma^{-1}_{\varepsilon_i} Y_{C(i)})^{-1}$.*

*Proof.* The definition of $\mathscr{M}_p$ ensures that $W_i^* = Y^*_{C(i)} B^*_{C(i),i}$. Using the fact that $Y^* = Y$ and $W^* = W$, given a noisy dataset $\tilde{W}_i = W_i + \varepsilon_i$ and the errorless dataset $Y$, the regression problem becomes

$$\tilde{W}_i = Y_{C(i)} \cdot B^*_{C(i),i} + \varepsilon_i. \tag{B.3}$$

From Section 3.1, identifiability of $\mathscr{M}_p$ at $p^*$ for the given input set $D$ is equivalent to $Y^*_{C(i)} = Y_{C(i)}$ being full column-rank. Thus, if $D$ ensures structural identifiability of $\mathscr{M}_p$ at $p^*$, $Y_{C(i)}$ is full column-rank and the weighted pseudoinverse of $Y_{C(i)}$, defined as $Y^\dagger \triangleq \left(Y^T_{C(i)} \Sigma^{-1}_{\varepsilon_i} Y_{C(i)}\right)^{-1} Y^T \Sigma^{-1}_{\varepsilon_i}$, is well-defined. This enables us to define the minimum variance estimator of $B^*_{C(i),i}$, $\hat{B}_{C(i),i} = Y^\dagger W_i$. From the linearity of the estimator in the Gaussian noise $\varepsilon_i$, after simple calculations of first and second-order moments, one gets $\hat{B}_{C(i),i} \sim \mathcal{N}\left(B^*_{C(i),i}, \widehat{\Sigma}\right)$ (also compare (Ljung, 1999, Appendix II)). Thus, from the definition of $\mathscr{B}_i$, $\mathbb{P}_{p^*}[\hat{B}_{C(i),i} - B^*_{C(i),i} \in \mathscr{B}_i] \geq \mathbb{P}_{p^*}[\hat{B}_{C(i),i} - B^*_{C(i),i} \in \mathcal{E}_{\widehat{\Sigma}}(\alpha)] = 1 - \alpha$. □

## References


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

Bellu, G., Saccomani, M., Audoly, S., D'Angiò, L., 2007. DAISY: a new software tool to test global identifiability of biological and physiological systems. Comput. Methods Programs Biomed. 88 (1), 52–61.

Bennett, B., Kimball, E., Gao, M., Osterhout, R., Dien, S. V., Rabinowitz, J., 2009. Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. Nat. Chem. Biol. 5 (8), 593–9.

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

Berthoumieux, S., Kahn, D., de Jong, H., Cinquemani, E., 2012. Structural and practical identifiability of approximate metabolic network models. Proc. 16th IFAC Symp. System Identif. (SYSID 2012), to appear.

Bettenbrock, K., Fischer, S., Kremling, A., Jahreis, K., Sauter, T., Gilles, E., 2006. A quantitative approach to catabolite repression in *Escherichia coli*. J. Biol. Chem. 281 (5), 2578–84.

Brand, M., 2002. Incremental singular value decomposition of uncertain data with missing values. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (Eds.), Proc. 7th Eur. Conf. Comput. Vision (ECCV 2002). Vol. 2350 of Lecture Notes in Computer Science. Springer Verlag, pp. 707–20.

Chappell, M., Godfrey, K., Vajda, S., 1990. Global identifiability of the parameters of nonlinear systems with specified inputs: A comparison of methods. Math. Biosci. 102 (1), 41–73.

Chen, W., Nieper, M., Sorger, P., 2010. Classic and contemporary approaches to modeling biochemical reactions. Genes Dev. 24 (17), 1861–75.

Chis, O., Banga, J., Balsa-Canto, E., 2011a. GenSSI: a software toolbox for structural identifiability analysis of biological models. Bioinformatics 27 (18), 2610–1.

Chis, O., Banga, J., Balsa-Canto, E., 2011b. Structural identifiability of systems biology models: a critical comparison of methods. PLoS One 6 (11), e27755.

Chou, I.-C., Voit, E., 2009. Recent developments in parameter estimation and structure identification of biochemical and genomic systems. Math. Biosci. 219 (2), 57–83.

Cobelli, C., DiStefano, J., 1980. Parameter and structural identifiability concepts and ambiguities: A critical review and analysis. Am. J. Physiol. 239 (1), R7–24.

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

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

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

Gutenkunst, R., Waterfall, J., Casey, F., Brown, K., Myers, C., Sethna, J., 2007. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput. Biol. 3 (10), e189.

Hardiman, T., Lemuth, K., Reuss, M. K. M., Siemann-Herzberg, M., 2007. Topology of the global regulatory network of carbon limitation in *Escherichia coli*. J. Biotechnol. 132 (4), 359–374.

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

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

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

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

Jaqaman, K., Danuser, G., 2006. Linking data to models: data regression. Nat. Rev. Mol. Cell. Biol. 7 (11), 813–9.

Jolliffe, I., 1986. Principal Component Analysis. Springer-Verlag.

Kotte, O., Zaugg, J., Heinemann, M., 2010. Bacterial adaptation through distributed sensing of metabolic fluxes. Mol. Syst. Biol. 6, 355.

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

Ljung, L., 1999. System Identification, Theory for the User. Prentice Hall PTR.

Maiwald, T., Timmer, J., 2008. Dynamical modeling and multi-experiment fitting with PottersWheel. Bioinformatics 24 (18), 2037–43.

Nemcova, J., 2010. Structural identifiability of polynomial and rational systems. Math. Biosci. 223 (2), 83–96.

Nikerel, I., van Winden, W., Verheijen, P., Heijnen, J., 2009. 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.

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

Raue, A., Kreutz, C., Maiwald, T., Klingmüller, U., Timmer, J., 2011. Addressing parameter identifiability by model-based experimentation. IET Syst. Biol. 5 (2), 120–30.

Reder, C., 1988. Metabolic control theory: A structural approach. J. Theor. Biol. 135 (2), 175–201.

Sands, P., Voit, E., 1996. Flux-based estimation of parameters in S-systems. Ecol. Model. 93 (1-3), 75–88.

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

Srinath, S., Gunawan, R., 2010. Parameter identifiability of power-law biochemical system models. J. Biotechnol. 149 (3), 132–40.

van Huffel, S., Vandewalle, J., 1991. The Total Least Squares Problems: Computational Aspects and Analysis. SIAM, Philadelphia, PA.

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

Voit, E., Almeida, J., Marino, S., Lall, R., Goel, G., Neves, A., Santos, H., 2006a. Regulation of glycolysis in *Lactococcus lactis*: an unfinished systems biological case study. IET Syst. Biol. 153 (4), 286–98.

Voit, E., Neves, A., Santos, H., 2006b. The intricate side of systems biology. Proc. Natl. Acad. Sci. USA 103 (25), 9452–7.

Walter, E., Pronzato, L., 1997. Identification of Parametric Models. Springer.